AUTOMATISATION DE LA DÉCOUVERTE DE RELATIONS SÉMANTIQUES ENTRE CONTENUS CIRCULANT SUR LES DISPOSITIFS SOCIOTECHNIQUES

Thèse présentée comme exigence partielle
du doctorat en informatique cognitive

Par Sylvain Rocheleau

Avril 2019

http://r-libre.teluq.ca/1660

RÉSUMÉ

Les dispositifs sociotechniques tels que les réseaux sociaux (LinkedIn, Facebook, Google +), les sites de microblogues (Twitter, Identi.ca), les plateformes de partage de contenus (Digg, StumbleUpon, Buzznet) ou d'hébergement de contenus générés par les utilisateurs (Academia.edu, Youtube, DeviantArt, Flickr) génèrent un océan de données et de métadonnées. Plusieurs de ces plateformes s'en remettent partiellement aux utilisateurs lorsque vient le temps d'opérer une classification de leurs contenus. La méthode souvent privilégiée est celle qui permet l'ajout de mots-clés par les utilisateurs à leur propre contenu ou à celui des autres.

Nous soutiendrons dans cette thèse qu'à partir des folksonomies qui se créent par le regroupement de mots-clés autour d'une thématique, on peut extraire un champ lexical présentant une cohésion sémantique parmi les mots-clés présentant une forte cooccurrence entre eux. Nous démontrerons ensuite que l'identification de sous-thématiques est possible par la découverte de liens sémantiques de type *partie-tout* entre mots-clés – mots-clics sur Twitter – faisant partie d'un même champ lexical.

Notre plus grand défi consiste à établir ces liens sémantiques en l'absence de ressources linguistiques telles que des dictionnaires ou des thésaurus puisque souvent les mots-clics utilisés par les utilisateurs sont des néologismes, des abréviations ou encore des mots composés. Ce défi est présent sur Twitter, notre objet d'expérimentation, mais également sur d’autres dispositifs sociotechniques.

Dans le cas de Twitter, l'identification de sous-thématiques nous a permis de mettre en place un prototype de suivi de discussions autour de thématiques de discussions afin d'observer et de capter leur évolution dans le temps. Notre démarche comprend trois étapes soit la création de folksonomies par le regroupement de mots-clics, la création de champs lexicaux entre mots-clics affichant une forte cooccurrence entre eux et, finalement, l'attribution de relations sémantiques à ces derniers à partir de

calculs statistiques.

Les applications de la découverte de liens sémantiques entre mots-clés appartenant à une folksonomie sont vastes et pourraient être utiles entre autres à l'expansion de requêtes en recherche d'informations, à la création d'ontologies, à la mise en place d'outils de suggestions de contenu ou encore à l'organisation d'archives.

Mots-clés : champ lexical, apprentissage automatique, folksonomie, hashtag, mot clic, mot-clé, Twitter, web sémantique, web social, réseaux sociaux, dispositif sociotechnique

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1 :
PROBLÉMATIQUE

Dans ce chapitre, nous abordons en premier lieu le point d'ancrage de notre problématique, soit l'analyse de contenus circulant dans l'espace public. Nous voyons également comment cette notion d'espace public peut s'articuler avec celles du web social et du web sémantique pour mettre en lumière la problématique à laquelle nous nous sommes intéressé.

Nous explicitons ensuite pourquoi nous avons retenu une approche d'apprentissage automatique faisant l'emploi de la logique inductive ainsi que les hypothèses que nous avons formulées afin de répondre à notre problématique.

1.1 Mise en contexte

Dans son livre *The Great Good Place* (1999), Oldenburg remarque que l'humain a besoin de trois types d'espaces de vie distincts: un espace pour sa vie privée, un lieu de travail et un ou plusieurs lieux de rencontres tels que des cafés, des parcs, des lieux de cultes, etc. Habermas, père de la théorie de l'espace public (1962), s'est intéressé de près à ces lieux de rencontres qu'il considéra en tant qu'espaces discursifs. Il étudia plusieurs versions de ces espaces publics en France, en Grande-Bretagne et en Allemagne et bien qu'il remarqua à la fois des différences dans la constitution des publics ou de leur taille, dans les façons de délibérer ou dans le climat des différents débats, il nota surtout l'importance de ces lieux dans l'organisation des débats publics. Hauser résume bien la définition habermassienne de l'espace public : « a discursive space in which individuals and groups congregate to discuss matters of mutual interest and, where possible, to reach a common judgment. » (1998).

Sur le web, cet espace discursif s'élargit, se virtualise et les limites physiques des cafés, des parcs de récréation, ou des lieux de cultes s'estompent, permettant des

échanges sur des thématiques diverses par des personnes appartenant à des communautés d'intérêts. Dans cette recherche, nous nous intéressons au dispositif sociotechnique Twitter pour lequel nous constatons une analogie avec le lieu de rencontre communautaire d'Oldenburg au sens où il s'agit d'un « endroit » où les gens se rassemblent pour y tenir des discussions et aussi avec l'espace public de Habermas puisqu'on retrouve sur cette plateforme un environnement discursif correspondant à plusieurs des dynamiques observées par celui-ci.

Si Twitter était un café, des dizaines de millions de personnes y viendraient quotidiennement pour y discuter et y prendre leur dose de caféine. Il est facile d'imaginer le bruit généré par autant de discussions ayant lieu dans un même espace et surtout l'énorme diversité de thématiques ainsi que le transfert des locuteurs d'une table à l'autre. L'objectif de cette recherche est d'automatiser la découverte de sous-thématiques reliées à une thématique principale discutée sur Twitter dans le but de recueillir un corpus de gazouillis plus complet.

Outre son intérêt pour l'informatique cognitive, car nous développons une approche d'apprentissage automatique de liens sémantiques, cette recherche représente un apport aux sciences sociales et tout particulièrement aux sciences de la communication. En effet, plusieurs chercheurs s'intéressent à l'analyse de discours dans l'espace public et, bien que les technologies du web rendent possible la captation de ces discours, les outils ne sont pas toujours au rendez-vous. La problématique de ce projet a d'ailleurs émergé de la Faculté de communication de l'UQAM dans le cadre d'une recherche à laquelle j'ai participé.

1.2 Espace public, web social et web sémantique

La problématique de cette recherche s'articule autour des concepts d'espace public, de web social et de web sémantique. Précisons d'abord notre utilisation de la notion d'espace public. Au sens large, l'espace public comprend l'ensemble des espaces

discursifs où des personnes se rassemblent pour discuter de sujets d'intérêt commun. Selon Habermas (1962), les consensus qui émergent dans l'espace public forment ce qu'on appelle l'opinion publique. Celle-ci serait le résultat de discussions rationnelles au sujet de l'intérêt commun au cours desquelles les citoyens échangent des informations et des points de vue. Déjà au 18e siècle, période sur laquelle s'est penchée Habermas, on observe une étroite relation entre l'espace public et la presse de l'époque qui à la fois informe les débats et s'en nourrit.

Il est important de souligner que l'espace public « naissant » que décrit Habermas est constitué d'une petite élite scolarisée, bien nantie et homogène, à la fois sur le plan social et culturel, ainsi que d'une presse à faible tirage. Cet espace public « bourgeois » a progressivement gagné d'autres classes de la société sous l'impulsion notamment, de l'alphabétisation, du développement du salariat, du suffrage universel, de l'élargissement du champ politique et de l'institutionnalisation des services publics (Wolton, 1992).

Cet élargissement est aussi intrinsèquement lié aux nombreuses mutations de la presse écrite (le premier média de masse), suivi de l'avènement des médias électroniques, puis des médias numériques. Pour rendre compte de cette évolution de l'espace public dit « bourgeois » depuis le 18e siècle, certains parlent d'ailleurs d'espace public « de masse » (Dahlgren et al., 1994). D'autres, pour souligner son interdépendance avec les médias, proposent l'expression d'espace public médiatisé, car « il est fonctionnellement et normativement indissociable du rôle des médias » (Wolton, 1992).

L'espace public suscite un vaste intérêt pour les sciences sociales et tout particulièrement en communication. Au fil des décennies, ce dernier s'est d'ailleurs complexifié. Selon Fraser (2003), on se doit d'envisager une pluralité d'espaces publics qui « fonctionnent comme des espaces de repli et de regroupement » et

« comme des bases et des terrains d'essai pour des activités d’agitation dirigées vers les publics plus larges ». L’élargissement dont nous faisons état est donc également traversé d’une dynamique d’atomisation de laquelle émerge cette pluralité d’espaces publics.

Toutefois, chacun de ces espaces publics n’entretient pas la même relation avec les médias de masse. Les médias de masse informent et se nourrissent de certains débats provenant de certains espaces publics alors que d’autres demeurent sous silence. Cette réalité a amené certains groupes à former des médias dits alternatifs afin de médiatiser leurs préoccupations et de les diffuser vers de plus larges publics. Par exemple, les journaux militants, les radios et les télévisions communautaires et désormais les blogues et autres dispositifs sociotechniques sont tour à tour utilisés dans le but de discuter de sujets d'intérêt commun. Les dispositifs sociotechniques que nous étudions présentent la particularité de servir à la fois d’espaces discursifs où des débats peuvent être menés et des décisions prises, mais aussi, à l’instar des médias de masse malgré un rayonnement moindre, d’informer les débats et de s’en nourrir. L’évolution de l’espace public, qu’il soit conjugué au singulier ou au pluriel, continue d’être intimement liée à celle des médias et les deux tendent vers une complexification et une redéfinition constante.

Depuis l’avènement du web, et plus particulièrement du web social, une autre notion s'est complexifiée soit celle de discussions privées « vs » discussions publiques. La distinction entre ces deux sphères s'opérait plus facilement à l'époque des médias traditionnels : la discussion de cuisine était privée, l'entrevue télévisuelle était publique.

Depuis une vingtaine d'années, on voit apparaitre toute une série de dispositifs sociotechniques qui brouillent la frontière entre sphère privée et sphère publique. Les deux gagnent d'ailleurs à être perçues comme étant interconnectées et devant être

pensées sur un continuum. Une discussion de cuisine est considérée plus privée qu'une discussion dans un espace de *chat* qui est plus privée qu'un échange sur Facebook qui est plus privé qu'un échange de gazouillis sur Twitter, si on exclut bien sûr qu'une discussion de cuisine puisse être captée et mise en ligne sur Youtube ou qu'une capture d'écran d'un échange sur Facebook soit publiée sur Twitter !

Le brouillage entre sphères privée et publique s'accompagne d'une impression d'anonymat de la part de plusieurs usagés. Cachés derrière leurs écrans et oubliant que leurs activités en ligne sont archivées, certains usagers profitent de la dématérialisation de leurs échanges pour libérer leur colère et leurs frustrations ou pour provoquer ces dernières chez d'autres usagers. Ces comportements ont donné lieu à l'apparition de nouveaux termes pour décrire ces phénomènes comme la cyberintimidation (*cyberbullying*) et l'apparition des *trolls*, ces semeurs de zizanie sur les dispositifs sociotechniques. Ils ont bien sûr depuis longtemps leur pendant hors ligne, mais leurs incursions dans l'espace public, qui selon l'idéal habermassien est formé de « discussions rationnelles au sujet de l'intérêt commun », menacent l'atteinte d'une opinion publique éclairée. La ligne est parfois mince entre la liberté d'expression et la diffamation, Twitter, un dispositif sociotechnique où la plupart des messages circulent à la vue de tous, est particulièrement vulnérable à ce genre de discours haineux (Bellmore *et al.*, 2015). L'entreprise a dit avoir apporter des changements à ses mécanismes visant à retirer les gazouillis haineux ou à fermer les comptes d'usagers qui les propagent, mais continue d'être critiquée pour son inaction face à plusieurs messages à caractère racial.

D'autre part, non seulement la discussion privée entre souvent dans le domaine public, mais de plus, de par la nature des dispositifs sociotechniques, elle laisse des traces. Elle est archivée et prête à être analysée à l'aide d'une panoplie d'outils informatiques, notamment ceux développés par l'informatique cognitive. Karsenty nous indique que « Les technologies sont en train de changer toutes les étapes de la

recherche et le potentiel est énorme parce qu'aujourd'hui, les humains qui évoluent en société laissent désormais beaucoup de traces. On est passé d'une pénurie à une surabondance de données. » (Déglise, 2010)

Cette abondance de données se chiffre. Selon une étude du GlobalWebIndex[1] issue d'un sondage auprès de 170 000 internautes américains, 1.72 heure par jour serait dédiée à l'usage des réseaux sociaux soit 28 % du temps en ligne. De plus, cette étude distingue les plateformes de réseaux sociaux des plateformes de microblogues tels que Twitter et indique que 0.81 heure par jour supplémentaire, soit 13 % du temps en ligne est passé sur ces autres dispositifs sociotechniques. L'entreprise Domo, spécialisée en Big Data, estimait qu'à chaque minute s'ajoutait en moyenne 100 heures de vidéos sur Youtube, 216 302 partages sur de photos sur Facebook, 456 000 gazouillis sur Twitter et 46 740 nouvelles photos sur Instagram. Ces statistiques évoluent rapidement et leur mise à jour est difficile, mais il n'en demeure pas moins qu'elles nous donnent un aperçu de l'immense volume de contenus et de métadonnées générées dans une seule minute. Comme le soutient Karsenty, le potentiel de ces technologies est énorme, mais il pose d'énormes défis d'organisation, de classification, d'interopérabilité et d'interprétation des données. Bref, la communauté du web sémantique n'a pas la tâche facile...

Le web sémantique offre quant à lui des langages généraux tels que RDFS (Ressource Description Framework Schema) ou OWL (Ontology Web Language) permettant de décrire des ontologies. Par exemple, l'ontologie FOAF (*Friend of a friend*) décrit les personnes, leurs liens, ce qu'elles créent et ce qu'elles font, l'ontologie SIOC (*Semantically-Interlinked Online Communities*) décrit l'information contenue explicitement et implicitement dans les dispositifs sociotechniques, le langage SKOS (Simple Knowledge Organization System) sert à la représentation de thésaurus et à

[1] Voir https://www.globalwebindex.net/blog/daily-time-spent-on-social-networks-rises-to-1-72-hours

établir des relations entre ontologies alors que OWL (Web Ontology Language) fut mis en place pour définir des associations plus complexes de ressources ainsi que les propriétés de leurs classes respectives. Certains développements d'ontologies se sont concentrés plus spécifiquement sur les folksonomies et les réseaux sociaux tels que MOAT (Meaning of a Tag) pour la désambigüisation des mots-clés (Passant et Laublet, 2008) ou encore SCOT (Social Semantic Cloud of Tags), un projet visant à mieux représenter la structure et la sémantique des données du web social afin de les partager et les réutiliser dans divers services (Kim et al., 2007).

À l'aide de ces outils de représentation, un certain nombre de propositions d'utilisation de méthodes d'analyse des dispositifs sociotechniques sont suggérées afin d'y extraire des informations et des connaissances, telles que la construction de folksonomies ou la détection de communautés d'intérêts. Par exemple, Mika (2005) détecte des champs sémantiques et des communautés d'intérêts en se servant de folksonomies et en employant la théorie des graphes, Paolillo et Wright (2006) identifient eux aussi des communautés d'intérêts, mais à partir d'une base d'annotations FOAF alors que d'autres chercheurs (Anyanwu, Maduko et Sheth, 2007) (Kochut et Janik, 2007) modifient des outils SPARQL (SPARQL Protocol and RDF Query Language) afin d'extraire des chemins entre des ressources sémantiquement liées dans les graphes RDF. Ce type de recherches contribue ainsi à donner une base en vue d'une représentation et d'une analyse sémantique des dispositifs sociotechniques.

Résumons notre articulation entre espace public, web social et web sémantique. L'espace public, en se définissant comme lieu où les discussions sur des sujets d'intérêts communs s'articulent, occupe une place importante dans la compréhension des enjeux sociaux, culturels, politiques et économiques d'une société. L'avènement du web social permet quant à lui de capter à grande échelle une partie de ces discussions, car il rétrécit la distance entre sphères privée et publique. Le web

sémantique quant à lui tente avec ses outils et ses formalismes d'analyser ces discussions, de mieux les représenter et, ultimement, d'en tirer des connaissances qui pourront à leur tour alimenter des réflexions en cours dans l'espace public.

1.3 Twitter comme terrain d’expérimentation

Nous avons fait le choix d'étudier plus attentivement le web social et Twitter en particulier comme terrain d’observation d’espaces publics. Le choix de Twitter s'est imposé à plusieurs égards. D'abord parce qu'il présente plusieurs similarités avec les cafés de France, de Grande-Bretagne et d'Allemagne qui ont servi de modèle au concept habermassien d'espace public, c'est-à-dire qu'il représente un espace discursif. En fait, pour être plus précis à ce sujet il faudrait souligner que, tout comme dans un café, tout ce qui s’y discute ne participe pas *de facto* à la construction d’un espace discursif au sens habermassien, c’est-à-dire avec une volonté de débattre et d’atteindre certains consensus sur des questions d’intérêt commun. En effet, Twitter peut aussi être le théâtre du trivial et un puissant rediffuseur de photos de chats. Donc, si on poursuit notre analogie avec les cafés européens du 18^{e} siècle, on se doit de distinguer le lieu public de la notion d’espace public. Le café est un lieu public, tout comme Twitter, mais il n’est pas l’espace public. Toutefois, et c’est là où l’analogie s’estompe, en virtualisant le lieu public, les dispositifs sociotechniques comme Twitter deviennent le support des échanges se substituant ainsi à la parole des individus. Ils gagnent ainsi une fonction inédite, celle de pouvoir enregistrer, archiver et restituer l’intégralité de toutes discussions ayant lieu sur leurs plateformes.

De plus, contrairement toujours au café, certains de ces dispositifs permettent une médiatisation des débats les rapprochant ainsi de la sphère publique. La dynamique observée par Habermas d’aller-retour entre l’antichambre des discussions de cafés et la presse écrite s’en retrouve ainsi aplatie : la discussion est médiatisée en permanence et instantanément. Bien sûr, lorsque des individus considèrent que

certains échanges méritent un peu plus de discrétion, ils opteront pour des lieux plus privés qu'ils soient physiques ou virtuels. Il n'en demeure pas moins que ces dispositifs, Twitter au premier chef à cause de son caractère public et de sa forte adoption, deviennent des lieux privilégiés pour observer les espaces publics.

Certains pourront arguer que la limite de 140 caractères ne permet guère une argumentation très élaborée, mais ce serait oublier les moyens de contournement développés par les usagers de la plateforme ainsi que ceux introduits par l'équipe de Twitter au fil des années. Ainsi, aux 140 caractères originels, il est désormais possible d'ajouter des URLs pointant vers des ressources web comme des blogues, des articles, des forums de discussions, etc.., d'inclure des images (dont certaines contiennent des textes de plus de 140 caractères!) ou encore des vidéos.

Twitter propose également des éléments de typologie particuliers issus d'usages ayant émergé sur le web. Par exemple, l'utilisation de la typologie *@nom_utilisateur*, inspirée de l'*Internet Relay Chat* (IRC), permet d'interpeler directement un utilisateur de la plateforme. De plus, la forme *#mot-clé* transforme ce dernier en mot-clic (*hashtag*). Cette typologie est apparue au début de 2008 dans le but de favoriser le suivi de gazouillis (tweets) sur une thématique précise sans nécessairement être abonné à tous les utilisateurs gazouillant sur ce sujet (Huang, Thornton et Efthimiadis, 2010). Donc, en utilisant l'outil de recherche sur l'interface de Twitter ou en se connectant à son API, un utilisateur a la possibilité d'obtenir tous les messages à propos d'un mot-clic donné.

L'ajout de mots-clics est une pratique qui a émergé dans la twittosphère et qui est maintenant tenue pour acquise. Surtout, elle constitue une caractéristique importante de ce dispositif sociotechnique lorsque vient le temps d'analyser son contenu, sachant qu'une grande partie de ce qui est produit sur le web est rarement accompagnée de métadonnées. Dans le cas de contenus proposant des métadonnées (plus

particulièrement des balises et des mots-clés [*tags*]), il est par ailleurs fréquent que ces dernières soient ajoutées à postériori, par les utilisateurs (via des services comme Delicious, Reddit, Zotero, Evernote, etc.) ou par des outils qui automatisent l'ajout de métadonnées (comme AlchemyAPI, Semantria, OpenCalais ou TextRazor). À l'inverse, la publication de gazouillis sur Twitter et l'ajout de métadonnées comme des mots-clics ou des noms d'utilisateurs sont faits de façon simultanée et non pas à postériori et, la plupart du temps, par des humains[2] et non par des outils d'ajout de métadonnées automatiques.

Afin de tirer avantage de cette simultanéité dans la publication de contenus et de balises sous forme de mots-clics, Starbird et Stamberg (2010) ont d'ailleurs proposé un type de syntaxe afin d'améliorer l'efficacité de la diffusion d'informations en temps de crise. Selon eux, l'utilisation de plusieurs mots-clics dans un même gazouillis favoriserait la classification de l'information et améliorerait la capacité de réaction des organisations autant en ligne que hors ligne.

On sait en outre que les discussions sur Twitter évoluent très rapidement et que de nouveaux mots-clics pour les identifier apparaissent continuellement. Il faut aussi prendre en compte que des mots-clics sont parfois modifiés par un leadeur d'une communauté ou par des membres de cette communauté étant parvenus à un consensus, lequel est parfois implicite ou explicite. Or, ces deux facteurs (l'apparition de nouveaux mots-clics et la modification de mots-clics existants) complexifient la cueillette de gazouillis à propos d'une discussion autour d'une thématique ayant cours dans l'espace public. Par exemple, lors de nos recherches préliminaires en 2013, nous avons constaté que, lors de la cueillette de gazouillis provenant de la communauté canadienne-française hors Québec, celle-ci avait conçu plusieurs mots-clics qui

---

[2] Les outils de publication automatique (*Twitterbots*) permettent de publier du contenu automatiquement. Il s'agit d'une pratique croissante dans les grandes organisations. Twitter évalue à 14 % le nombre d'utilisateurs qui seraient des *Twitterbots*.

suivent tous la même règle de nomenclature (Millette et Rocheleau, 2014). Il s'agit de mots-clics de quatre lettres dont les deux premières sont « fr » pour « francophone » suivies de deux lettres identifiant une province. Cette pratique permet aux usagers de situer géographiquement leurs messages à l'intérieur de la communauté canadienne-française et on les retrouve souvent avec le mot-clic #frcan. Par exemple, la communauté albertaine utilise #frab, le Manitoba utilise #frmb alors que les Canadiens français de l'Ontario utilisent ou plutôt utilisaient #fron. En effet, la communauté francophone de l'Ontario a réalisé que le mot-clic #fron était également utilisé par une autre communauté se situant au Brésil et que leurs messages s'y retrouvaient amalgamés dans les résultats de recherche. Il est important de noter que l'interface de recherche avancée, tout comme l'API, permettent de spécifier des critères comme la langue et que par conséquent il demeurait possible de suivre les discussions utilisant le mot-clic #fron exclusivement en français. Néanmoins, la communauté francophone d'Ontario a préféré opter pour un changement de mot-clic et utilise désormais #onfr. Évidemment, nous avons découvert ce changement de mot-clic *à postériori* et notre corpus en a ainsi souffert.

Sans surprise, nous avons aussi observé l'irruption de nouveaux mots-clics dont certains sont devenus d'usage courant alors que d'autres sont disparus après quelques semaines, voire quelques jours. Ce fut le cas notamment de mots-clics identifiant des festivals comme celui du cinéma francophone de Toronto (#cinefranco) ou encore des mots-clics identifiant des enjeux sociopolitiques comme #FusionDesConseilsScolaires. La présence de ces mots-clics dans notre corpus s’observe quand ils sont utilisés aux côtés de mots-clics déjà récupérés via l’API. C’est le cas si, par exemple, notre requête vers l’API contient le mot-clic #fron et que certains gazouillis contiennent à la fois #fron et #cinefranco. Toutefois, cela implique que les gazouillis contenant seulement le mot-clic #cinefranco ne sont pas récupérés via l’API. Par conséquent, dans un tel cas, comme le mot-clic #cinefranco n’est pas

récupéré via l'API, une partie de la discussion à propos de la sous-thématique « cinéma francophone de Toronto » nous échappe. Comme cette sous-thématique fait partie de la thématique plus large de la francophonie canadienne, le corpus à propos de cette dernière est donc incomplet.

La question qui anime ce projet de recherche peut être formulée ainsi :

> Comment automatiser la découverte de mots-clics nécessaires au suivi d'une discussion ayant cours dans l'espace public à propos d'une thématique donnée?

1.4 Hypothèse de recherche et approche d'apprentissage sélectionnée

> Nous posons comme hypothèse qu'un raisonnement inductif appliqué à l'analyse de mots-clics cooccurrents à un mot clic thématique dans une discussion sur Twitter permet l'extraction de relations sémantiques de type *partieDe* entre ces mots-clics.

Cette thèse propose ainsi une approche d'apprentissage automatique permettant de capter plus efficacement l'ensemble d'une discussion sur Twitter par un ajout dynamique de mots-clics. Le raisonnement inductif est à la base des algorithmes et du prototype informatique que nous avons développé et constitue notre approche centrale pour l'analyse des données et l'extraction de connaissances.

Ce type d'apprentissage s'améliore en fonction de la quantité et de la qualité des données disponibles. Lors de tests comparant divers algorithmes ayant pour fonction d'induire des inférences à partir de corpus de textes, il a été observé à maintes reprises (Nikam et Mulla, 2014; Abu Abbas, 2008; Liu et al., 2003) que, bien que la qualité des inférences différait d'un algorithme à un autre, le facteur le plus important était la quantité de données sur lesquelles les algorithmes pouvaient travailler. Par exemple, un algorithme performant moins bien qu'un autre sur un corpus de 10 000 entrées

améliorera sa performance si on lui soumet un corpus plus important. Cette caractéristique du raisonnement inductif et des inférences qui peuvent en être tirées a d'ailleurs été tenue en compte (p.128) et figure également parmi les limites que nous avons identifiées (p.155).

Rappelons par ailleurs qu'une conclusion issue d'un raisonnement inductif n'est pas nécessairement vraie, car les hypothèses sous-jacentes sont seulement probables et non certaines. Voilà pourquoi nous dédions une large portion de cette thèse à définir notre cadre théorique, à la fois à l'aide de la littérature et de nos expérimentations préliminaires, dans le but de nous assurer du fort degré de probabilité de notre hypothèse principale et de celles qui en découlent. Nous avons également attaché une grande importance à la validation de nos résultats en faisant appel à deux évaluateurs externes.

Cette hypothèse prend appui sur trois concepts principaux : [1] la création de folksonomies par le regroupement de mots-clics, [2] l'émergence de champs lexicaux entre mots-clics affichant une forte cooccurrence entre eux et [3] l'attribution de relations sémantiques entre des mots-clics à l'aide de calculs statistiques.

Dans le chapitre sur le cadre théorique, une section sera consacrée à chacun de ces trois concepts où nous expliciterons les ancrages théoriques pertinents à l'élaboration de l'hypothèse principale auxquels se joindront des exemples de données issues de corpus recueillis entre 2013 et 2015 lors de nos expérimentations préliminaires.

CHAPITRE 2 :
OUTILS, CONCEPTS ET DÉFINITIONS

Certains termes plus ou moins spécialisés décrivant soit des outils soit des concepts sont fréquemment mobilisés tout au long de cette thèse. Avant d'aborder le cadre théorique, nous proposons quelques définitions précisant notre usage de ces termes.

2.1 Les outils

2.1.1 Vocabulaire SCOT

Nous opterons donc pour un vocabulaire ontologique offrant un niveau d'abstraction supérieure qui rendra possible l'attribution de relations à tous les mots-clics d'un champ lexical sans exception. On peut schématiser les classes avec les prédicats et les propriétés qui leur sont rattachés de la façon suivante :

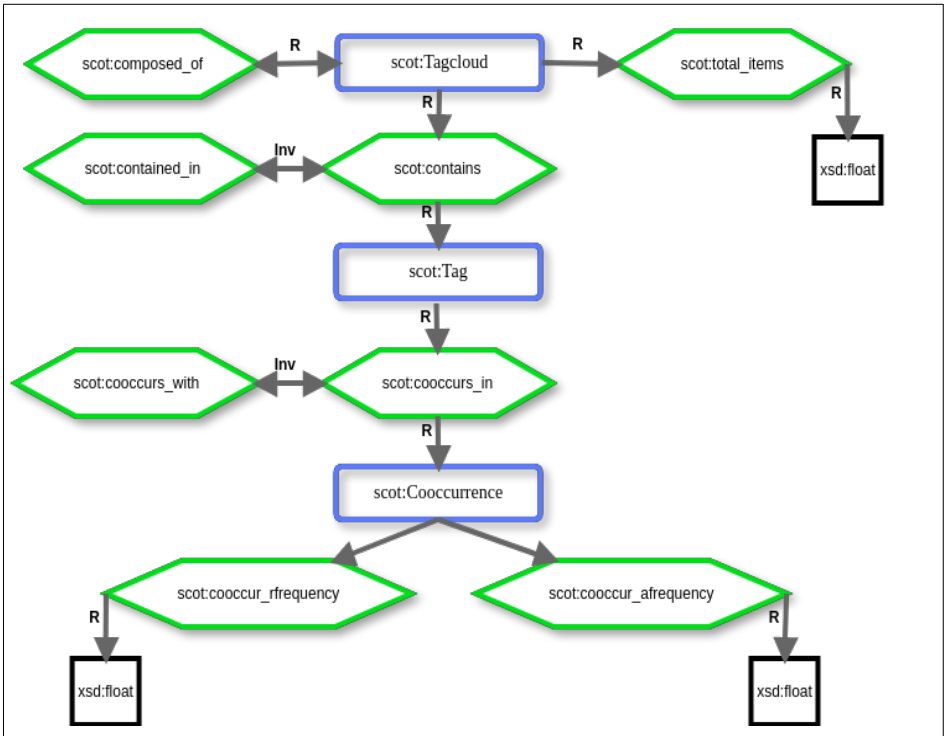


Figure 2.1: Schéma du vocabulaire ontologique SCOT

Le vocabulaire ontologique SCOT (Social Semantic Cloud of Tags) a pour objectif de faciliter la modélisation de l'étiquetage social. Ce vocabulaire n'est pas très complexe, mais comme il est peu connu il convient d'en présenter les éléments qui nous ont été utiles.

Ce langage est composé de trois classes soit « *scot:Tagcloud* », « *scot:Tag* » et « *scot:Cooccurrence* ».

La classe « *scot:Tagcloud* »[3] est composée d'ensembles de mots-clés (des mots-clics dans le cas de Twitter) et de métadonnées à propos de ceux-ci.

Les prédicats et propriétés de cette classe que nous avons utilisé sont les suivants :

- *scot:contains* : relation entre un ensemble de mots-clics de la classe scot:Tagcloud et un mot-clic de la classe scot:Tag qui appartient à cet ensemble. (inverse de scot:contained_in dans la classe « scot:Tag»).
  Ex : {#a,#b, #c} *scot:contains* #a
- scot:composed_of : relation entre deux ensembles où l'un est un sous-ensemble de l'autre, comme par exemple un champ lexical composé à partir d'une folksonomie, le premier étant un sous-ensemble du deuxième.
  Ex : {#a,#b} *scot:composed_of {#a, #b, #c}*
- *scot:total_items* : propriété de données prenant comme valeur le nombre d'items faisant partie d'un ensemble de mots-clics.

La classe « *scot:Tag* » regroupe tous les mots-clics décrivant les ressources présentes dans un « *scot:Tagcloud* ». De nombreux prédicats et propriétés y sont rattachés, voici ceux qui sont pertinents à nos travaux :

[3] Sauf lorsqu'il sera explicitement question de cette classe, nous préférerons l'expression « ensemble de mots-clics » plutôt que « tag cloud ».

- *scot:cooccurs_in* : relation de cooccurrence entre un mot-clic et un ensemble de mots-clics (inverse de scot:cooccurs_with dans la classe « *scot:Cooccurrence* »)
  Ex : {#a} *scot:cooccurs_in {#a, #b, #c}*
- *scot:contained_in :* relation entre un mot-clic ou un sous-ensemble de mots-clics contenu dans un ensemble. (inverse de scot:contains dans la classe « scot:Cooccurrence »)
  Ex : {#a} *scot:contained_in* {#a, #b, #c}
- *scot:synonym* : propriété de synonymie entre deux mots-clics

La classe « *scot:Cooccurrence* » regroupe les ensembles de mot-clics cooccurrents et permet de quantifier le niveau de cooccurrence entre ces derniers.

Les prédicats et propriétés qui y sont rattachés sont :

- *scot:cooccurs_with* : relation de cooccurrence entre un ensemble de mots-clics et un mot-clic (inverse de scot:cooccurs_in dans la classe « *scot:Tag*»).
  Ex : {#a, #b, #c} scot:cooccurs_with {#a}
- *scot:cooccur_afrequency* : propriété indiquant le nombre de fois où une cooccurrence est observable.
- *scot:cooccur_rfrequency :* propriété indiquant un ratio de la fréquence d'un mot-clic ou plusieurs mots-clics dans un ensemble de mots-clics. C'est le quotient de scot:cooccur_afrequency sur scot:total_items.

On peut schématiser les classes avec les prédicats et les propriétés qui leur sont rattachés de la façon suivante :

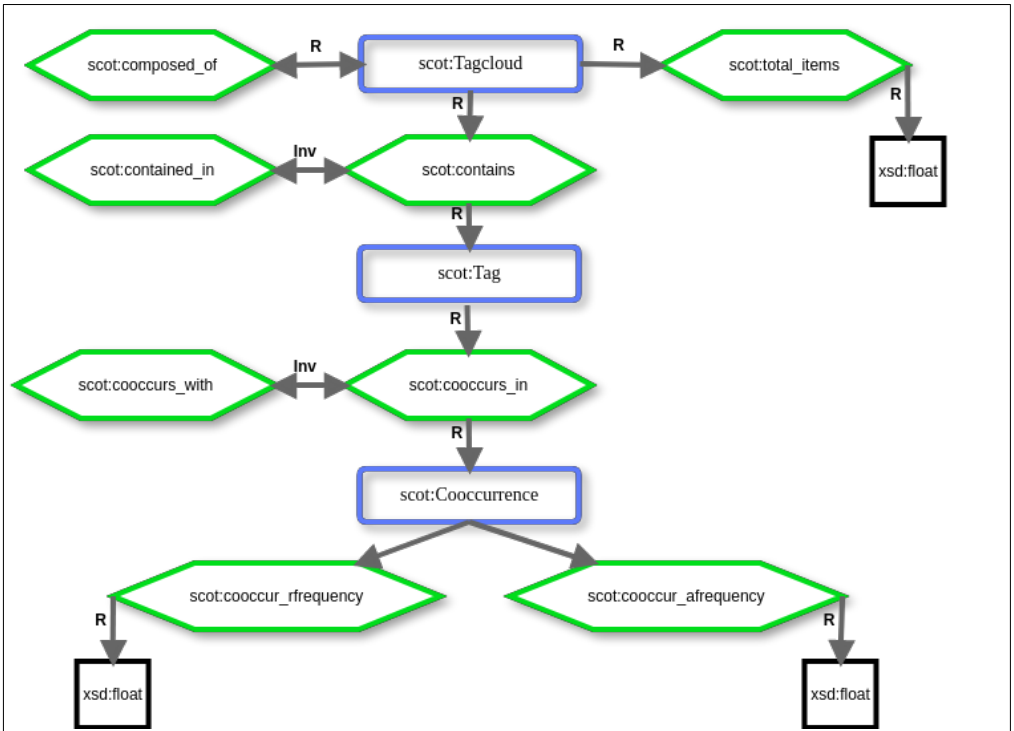


Figure 2.2: Schéma du vocabulaire ontologique SCOT

2.1.2 *La Streaming* API et la *Search* API

La cueillette de gazouillis est réalisée à partir de deux API offertes par Twitter : la *Streaming* API et la *Search* API.

Streaming API

La plupart du temps, lorsqu'on utilise une API REST, on ouvre une connexion, une requête est envoyée, des données sont reçues, puis la connexion est fermée. Les *Streaming* API, dont celle de Twitter, ont un fonctionnement différent au sens où la connexion n'est pas ouverte puis fermée, elle est permanente. Des corpus de gazouillis recueillis sur des périodes de plusieurs mois nécessitent donc une connexion ininterrompue à la *Streaming* API de Twitter pendant toute la durée de la

cueillette. Une fois la connexion établie, on envoie à la *Streaming* API une liste de mots-clics pour lesquels on souhaite obtenir des gazouillis. Une liste de mot-clics tel que [#polQc, #PQ#, #PLQ) recueillerait tous les gazouillis contenant au moins un de ces mots-clics. Ces gazouillis seront reçus presque instantanément suite à leur publication et nous les sauvegardons dans une base de données MySQL.

***Search* API**

La *Search API* de Twitter a un fonctionnement plus classique. Elle reçoit en entrée une requête formée d'un mot-clic et elle nous retourne des gazouillis qui ont employé ce dernier dans la dernière semaine.

Cette API comporte deux limites importantes. Premièrement, les gazouillis retournés ne représentent qu'un faible échantillon des gazouillis archivés par Twitter. La limite maximale de gazouillis retournés est de 500 et la fouille n'est exécutée que pour les 7 derniers jours. Deuxièmement, par souci d'économies de ses ressources informatiques, Twitter accorde un court laps de temps à l'exécution d'une requête envoyée à cette API. Ainsi, même si plus de 500 gazouillis ont été publiés dans les 7 derniers jours pour un mot-clic donné, la *Search API* en retourne beaucoup moins si le laps de temps accordée à l'exécution de la requête est dépassé. Il s'agit d'une situation que nous avons pu observer à plusieurs reprises.

Les impacts de ces limitations de la *Search* API sont discutées au Chapitre 8 : limites et apports dans la section Restrictions de l'accès aux archives de Twitter (p.155).

2.2 Les concepts

2.2.1 Thématique

Nous utilisons ce terme en référence à une discussion sur Twitter. Nous l'utilisons de façon large autant pour désigner des thématiques reliées à des émissions de télévision qu'à des enjeux sociaux. Une thématique regroupe généralement plusieurs sous-

thématiques. Par exemple, la thématique de la politique québécoise regroupe des sous-thématiques comme des élections, des scandales politiques, des projets de loi, des chefs de parti, etc.

2.2.2 Mot-clic

Un mot-clic, traduction de l'anglais *hashtag*, est une suite de caractères précédée du caractère dièse (#). En cliquant sur un mot-clic on retrouve tous les gazouillis de la discussion à propos de celui-ci. Plusieurs mots-clics sont symboliques d'une thématique de discussion comme #polQc pour la politique québécoise et #cdnpoli pour la politique canadienne. Une définition plus détaillée discutant à la fois des différentes formes de mots-clics et des processus cognitifs qui sont à l'œuvre dans leur sélection par les usagers est proposée dans la section Anatomie des mot-clés du Chapitre 3 : Cadre théorique (p.37)

2.2.3 Mot-clic thématique

Un mot-clic thématique est un mot-clic représentant une thématique. Par exemple, la thématique de la politique québécoise a comme mot-clic thématique #polQc. Lorsque nous démarrons une cueillette afin d'obtenir des gazouillis en provenance de la *Streaming* API, nous utilisons toujours un mot-clic thématique.

Dans certains cas, plus d'un mot-clic thématique sont utilisés. Par exemple, lors d'une compétition sportive, nous utilisons les mots-clics désignant chacune des deux équipes. Il s'agit de cas relativement rares où la thématique étant éphémère, il y a peu de chance qu'un mot-clic thématique émerge de la twittosphère. La *Streaming* API recueillant tous les gazouillis pour chacun des mots-clics, il faut être prudent lorsqu'on choisit plus d'un mot-clic thématique. En effet, on pourrait être tenté de capter la discussion sur le vin français et utilisant comme mots-clics thématiques #vin et #France, mais plutôt que de recueillir toute la discussion à propos du vin français

on recueillerait toute la discussion à propos du vin d’une part, et à propos de la France, d’autre part.

Nous avons expérimenté des façons d’assister l’utilisateur dans le choix du mot-clic thématique. Il y a bien sûr les méthodes d'observation ou d'observation participante où le sens commun peut nous guider efficacement vers l'identification d'un mot-clic thématique, mais cette méthode nécessite du temps en ressources humaines et s'adapte mal dans le temps. Un défi évoqué par l'entreprise Seevibes, où un stage a été effectué, consistait à déterminer le mot-clic thématique de chacune des émissions de télévision télédiffusées dans les marchés canadien et français. La solution recherchée devait être entièrement automatisée et ne nécessiter aucune supervision.

Illustrons ce défi avec quelques exemples. Dans le cas de l'émission « Top Chef », le mot-clic thématique est #topchef. On peut l'obtenir en prenant le titre de l'émission et en retirant un espace. Méthode séduisante, mais comportant trop d'exceptions : parfois le mot-clic thématique est un acronyme comme pour « Tout le monde en parle » (#tlmep), parfois c'est un mélange d'acronymes et de mots entiers comme pour « *Games of Thrones* » (#GofT), dans certains cas il est constamment modifié comme pour les rencontres sportives opposant des équipes différentes à chaque match. Une méthode consistant à utiliser en entrée le titre d'une émission pour ensuite y appliquer un certain nombre de règles qui auraient permis de déduire un mot-clic thématique ne semblait pas être très prometteuse.

Nous nous sommes posé la question sous un angle différent en nous demandant s’il existerait un leadeur susceptible d'imposer un mot-clic thématique pour une émission de télévision. La plupart des émissions de télévision disposent d'un compte Twitter et, sans surprise, on y retrouve des gazouillis qui sont presque exclusivement à propos de leur émission. Les usages quant à l'utilisation de mots-clics diffèrent d'une émission de télévision à une autre. En effet, certains en utilisent plusieurs alors que d'autres

n'en utilisent qu'un seul, mais une constante est observable : un de ces mots-clics revient plus souvent que les autres. Il a donc été envisagé de le considérer comme mot-clic thématique de l'émission. Nous avons donc testé avec succès une méthode simple consistant à récupérer un échantillon de mots-clics à partir du compte officiel d'une émission de télévision afin d'en extraire le mot-clic ayant la plus forte occurrence. Cette approche s'est avérée efficace dans d'autres situations où le compte officiel était celui d'un parti politique, d'une équipe sportive, d'un groupe militant ou d'un réseau de hobbyistes.

2.2.4 Discussion

Une discussion sur Twitter est comprise comme l’ensemble des gazouillis à propos d’une thématique. Une discussion ayant comme thématique la politique québécoise se définit comme l’ensemble des gazouillis contenant le mot-clic thématique #polQc et contenant aussi possiblement d’autres mots-clics représentant des sous-thématiques comme #PLQ, #PQ, #assnat, etc.

2.2.5 Cooccurrence

La cooccurrence est la présence simultanée de deux ou de plusieurs mots-clics. Nous l’observons à l’échelle d’un ensemble de gazouillis en procédant à un décompte de tous les mots-clics présents. Par exemple, dans un ensemble de plusieurs milliers de gazouillis à propos de la politique québécoise, on mesure le nombre de cooccurrence entre le mot-clic thématique et chacun des mots-clics présents dans l’ensemble.

À plus petite échelle, on observe aussi la cooccurrence à l’intérieur d’un gazouillis lorsqu’un utilisateur utilise plus d’un mot-clic.

La notion de cooccurrence est mobilisée à diverses occasions dans cette thèse et la courte définition produite ici est bonifiée lorsque nous présentons les trois concepts piliers du cadre théorique soit :

- Les folksonomies (p.33)
- Les champs lexicaux (p.42)
- Les calculs de similarités sémantiques (p.57)

2.2.6 Requête en cours

La requête en cours est un concept que nous avons développé pour désigner un ensemble de mots-clics pour lequel on veut obtenir des gazouillis en provenance de la *Streaming* API. Elle contient toujours au moins le ou les mots-clics thématiques. Les autres mots-clics constituant la requête en cours proviennent des raisonnements effectués par nos algorithmes et représentent des sous-thématiques. Par exemple, pour capter la discussion à propos de la thématique de la politique québécoise, le mot-clic thématique #polQc ferait toujours partie de la requête en cours. D'autres mots-clics pourrait faire partie de la requête en cours pendant plusieurs mois (#PKP), plusieurs semaines (#loi59) ou seulement quelques heures (#manifencours).

Notre concept de requête en cours en est une d'ensemble dynamique mis à jour toutes les 15 minutes afin de s'adapter à l'évolution d'une thématique dans le temps.

2.2.7 Mot-clic candidat

Un mot-clic candidat est un mot-clic qui sera évalué afin de vérifier s'il appartient à une sous-thématique de la thématique principale. S'il appartient à une sous-thématique de la thématique principale, le candidat sera « promu » en tant que mot-clic faisant partie de la requête en cours. Notre façon de déterminer quels mots-clics deviennent des candidats est explicitée dans le cadre théorique dans la section Champs lexicaux émergeant d'une folksonomie (p.42).

CHAPITRE 3 :
CADRE THÉORIQUE

Ce chapitre est composé de trois sections qui s'imbriquent les unes dans les autres. Nous décrivons d'abord comment se forme une folksonomie par regroupement de mots-clics issus des processus cognitifs de catégorisation des usagers de dispositifs sociotechniques. Nous explicitons ensuite de quelle manière un champ lexical peut être construit à partir des mots-clics les plus fréquemment utilisés dans une folksonomie. Puis, nous nous intéressons aux types de relations sémantiques qu'il est possible d'observer entre mots-clics partageant un même champ lexical dans le but de démontrer comment à l'aide de calculs statistiques effectués sur des données de cooccurrence entre mots-clics il devient possible d'automatiser la découverte de certaines de ces relations sémantiques.

3.1 Les folksonomies

3.1.1 De la nature des folksonomies

Le terme *folksonomie* est un néologisme attribué à Thomas Vander Wal, qui l'aurait utilisé pour la première fois dans un forum de discussion portant sur l'architecture de l'information (Smith, 2004). Il est formé de *folk* (populaire) et de taxonomie, terme lui-même composé du grec *taxis* (placement, classement) et de *nomos* (loi, règle). Une folksonomie est le résultat d'un vocabulaire et d'une classification produits par le regroupement de mot-clés cooccurrents qui sont employés par des personnes souvent non expertes pour décrire diverses ressources, alors que le terme « étiquetage social » (*social tagging*) peut être conçu comme le contexte sociotechnique dans lequel s'opère la création de mots-clés par les utilisateurs (Ding et coll., 2009).

Au sein d'un dispositif sociotechnique, une communauté d'utilisateurs négocie de façon tacite ou implicite l'usage de termes comme des mots-clés ainsi que leurs

significations ce qui favorise la stabilisation de vocabulaires communs pour décrire un domaine (Fu, 2008). Dans leur étude portant sur *del.icio.us*, Huberman et Golder (2005) suggèrent que les premiers mots-clés attribués à une ressource web correspondent souvent à des catégories du niveau de base, notion à laquelle nous nous sommes intéressé (p.34). Ils observent également que la distribution de mots-clés se stabilise rapidement, suggérant ainsi une forme de consensus parmi les utilisateurs de cette plateforme.

Dye (2006) identifie deux grands types de folksonomies : des folksonomies générales (*broad folksonomies*) dans lesquelles une communauté d'utilisateurs assigne des mots-clés à une même ressource, comme sur la plateforme *del.icio.us*, et des folksonomies particulières (*narrow folksonomies*) dans lesquelles un utilisateur assigne des mots-clés à un contenu qu'il vient de créer, comme sur la plateforme Twitter.

L'intérêt et la force des folksonomies sont liés à l'effet communautaire et au regroupement de mots-clés, ou autrement dit, par la mesure de la cooccurrence. D’ailleurs, la classification d'une ressource s'obtient souvent en retenant les mots-clés les plus souvent utilisés par les utilisateurs. Le regroupement de mots-clés cooccurrents (*clustering*) est ainsi un mécanisme qui permet d'observer l'émergence d'acceptions communes à partir des actions individuelles d'attribution de mots-clés (Trant, 2009; Magnuson, 2013).

3.1.2 Théorie du prototype et processus cognitifs de catégorisation

S'il est important de comprendre la mécanique sous-jacente à la création de folksonomies, il semble également opportun de poser la question du ou des processus cognitifs en amont qui entrent en action dans le choix des mots-clés. Comment un utilisateur procède-t-il pour choisir un mot-clé particulier? Quels processus mentaux sont mobilisés?

Il semble que très souvent l'exercice menant aux choix de mots-clés pour décrire une ressource emprunte l'un des processus cognitifs fondamentaux, soit la catégorisation (Rosch et coll., 1976; Weigend, Wiener et Pedersen 1999; Posch et coll., 2013). En termes généraux, la catégorisation peut être comprise comme une activité consistant à placer des objets ou des concepts dans des catégories en fonction de leur similarité. Il s'agit d'une fonction cognitive complexe qui peut faire appel à divers types de raisonnements incluant très souvent l'analogie, mais aussi la déduction, l'induction ou l'abduction pour ne nommer que ceux-ci. L'approche classique, que l'on doit à Aristote, présente la catégorisation comme la création d'entités distinctes qui sont définies par des caractéristiques communes (Peterson, 2006). Dans cette approche, ces caractéristiques communes fournissent des conditions nécessaires et suffisantes pour donner un sens à une catégorie. Il s'agit d'une vue hiérarchique dans laquelle, par exemple, les mammifères se retrouveraient dans la catégorie des animaux et le dauphin dans la catégorie des mammifères.

Les travaux en sciences cognitives de Rosch dans les années 1970 ont proposé que la catégorisation puisse aussi relever d'un processus fondé sur la recherche de prototypes correspondant le mieux à une classe. Par exemple, dans une de leurs expérimentations Rosch et coll. (1975) ont demandé à des étudiants de noter sur une échelle de 1 à 7 différents objets se retrouvant dans une maison qu'ils considéraient comme appartenant à la classe « meuble ». Alors que « chaise » et « canapé » ont obtenu les scores les plus forts, « four », « réfrigérateur » ou encore « téléphone » se sont retrouvés en bas de liste. Dans le cas de ces derniers, on peut simplement noter qu'en effet, ces trois éléments ne sont pas des meubles. Par contre, cette classification devient intéressante pour des éléments comme le tabouret ou la commode qui sont effectivement des meubles, mais qui se sont aussi retrouvés au bas de la liste. Rosch se sert de cette observation, que l'on pourrait considérer à juste titre comme un regroupement des catégorisations faites par les participants telles qu'elles se font dans

le cadre de la création de folksonomies, pour faire valoir que « chaise » et « canapé » sont les prototypes représentant le mieux la catégorie « meuble ».

De plus, la théorie du prototype introduit la notion de « niveau de base » dans la catégorisation. Le niveau de base ne serait ni le niveau le plus abstrait ni le niveau le plus précis d'une catégorie : « *Categorizations which humans make of the concrete world are not arbitrary but highly determined. In taxonomies of concrete objects, there is one level of abstraction at which the most basic category cuts are made. Basic categories are those which carry the most information, possess the highest category cue validity, and are, thus, the most differentiated from one another.* » (Rosch et coll., 1976). Elle reprend d'ailleurs l'exemple de la chaise pour expliciter son propos. Il existe des sous-catégories de chaises, la chaise de bureau, de cuisine, etc., mais à ce niveau subordonné il est difficile d'ajouter des caractéristiques significatives qui viendraient s'ajouter au niveau de base. De même, à un niveau d'abstraction plus élevé, tel que la catégorie meuble, les similarités de concepts sont difficiles à repérer. On peut facilement s'imaginer ou dessiner une chaise, mais plus difficilement un meuble, car le concept semble trop abstrait.

Au moment d'attribuer un mot-clé à une ressource, moment qui par ailleurs est généralement court, la notion du niveau de base dans la catégorisation joue un rôle important ; Rosch parle d'ailleurs de *basic level advantage* (*Ibid.*). Ley et Seitlinger (2010) explicitent son importance ainsi : *« In human communication, the basic level has an important role as it contains categories that are most easily retrieved from memory and have a high degree of information value in describing objects. Among many others, an advantage for the basic level has been shown when people verify the categories of pictures of objects, or in a free naming paradigm.* » (p.2).

La théorie du prototype et la notion du niveau de base dans la catégorisation sont particulièrement utiles à la compréhension de la formation des folksonomies. Elles

démontrent qu'une folksonomie est bien plus qu'un regroupement de mots-clés. C'est la zone d'intersection des processus cognitifs de catégorisation effectués par un ensemble d'utilisateurs. Bien sûr, l'application de ces processus diffère d'un utilisateur à l'autre, mais leur regroupement permet justement d'en extraire les éléments communs (Hotho et coll., 2006; Trant, 2009).

3.1.3 Anatomie des mot-clés

Parmi la panoplie d'usages qui ont émergé dans le sillon du web social, l'étiquetage (*tagging*) est l'un des phénomènes les plus intéressants. Alors que l'une des idées maitresses du web social est de laisser jouer aux utilisateurs un rôle de producteur de contenus, le marquage franchit une étape de plus en laissant à ces derniers le soin de classifier tous ces contenus. Par l'ajout de simples mots-clés à leur contenu et à celui des autres, ils peuvent décider quelles métadonnées peuvent être reliées à un contenu. Ces mots-clés peuvent être utilisés à plusieurs niveaux et désigner parfois une thématique du contenu, une opinion à propos de celui-ci, une note personnelle, etc. (Huberman et Golder, 2005). Le marquage des utilisateurs, souvent la seule information disponible à propos d'un objet du web, pose divers défis pour le web sémantique sur lesquels nous nous pencherons.

Le processus cognitif de catégorisation permet de comprendre un peu mieux comment s'opère le choix des mots-clés, mais il n'est pas le seul à intervenir dans leurs sélections par les usagers. Tout d'abord, rappelons que les mots-clés se présentent sous plusieurs formes, car la plupart du temps l'utilisateur dispose d'une complète liberté pour les déterminer. Il peut choisir plusieurs mots-clés pour décrire une seule ressource, choisir des termes issus du langage courant (cancer), agglutiner plusieurs mots (breastcancer), former des phrases (PrayForMamaSwift), faire des contractions de plusieurs termes (polQc) ou encore utiliser des abréviations (PQ) : « Ces items peuvent être catégorisés avec n'importe quel mot définissant une relation

entre la ressource en ligne et un concept issu de l'esprit de l'usager. Un nombre quasi infini de mots peuvent ainsi être choisis, dont quelques-uns sont issus de représentations évidentes tandis que d'autres ont peu de signification en dehors du contexte de l'auteur du tag. » (Guy et Tonkin, 2006)

Les mots-clés formant les folksonomies relèvent généralement de l'une des trois catégories suivantes (Sen et coll., 2006) :

1. **Mots-clés personnels** : Ils ont souvent comme seul utilisateur final son auteur. Ils sont utilisés pour l'organisation de bibliothèques ou de collections personnelles. Exemple : #mesrecettes.
2. **Mots-clés subjectifs** : Ils expriment une opinion à propos d'une ressource. Exemple : #cool.
3. **Mots-clés factuels** : Ils identifient des faits reliés aux ressources comme des personnes, des endroits, des concepts, des thématiques, etc. Exemple : #Coderre.

Plusieurs facteurs peuvent influencer le choix d'un mot-clé; parmi ceux-ci on retrouve bien sûr différents processus cognitifs de catégorisation selon les individus, mais également les biais personnels liés aux préférences et aux références culturelles de chacun, l'influence du groupe d'utilisateurs ayant attribué des mots-clés aux mêmes ressources, sans oublier les algorithmes de suggestions de mots-clés qui accentuent souvent l'influence du groupe, mais qui peuvent également introduire d'autres biais amenés par les concepteurs du dispositif sociotechnique (*Ibid*).

Une recherche en cours sur le web illustre bien les biais personnels et culturels liés aux préférences et aux références de chacun. Dans une interface créée par Alan McConchie (2015), les utilisateurs sont invités à choisir le terme qu'ils utilisent pour nommer la catégorie « boissons gazeuses ». Sans le spécifier, on leur demande en fait

de trouver le prototype de cette classe. La recherche s'adresse aux locuteurs anglophones des États-Unis et ils doivent spécifier l'État où ils ont appris l'anglais.

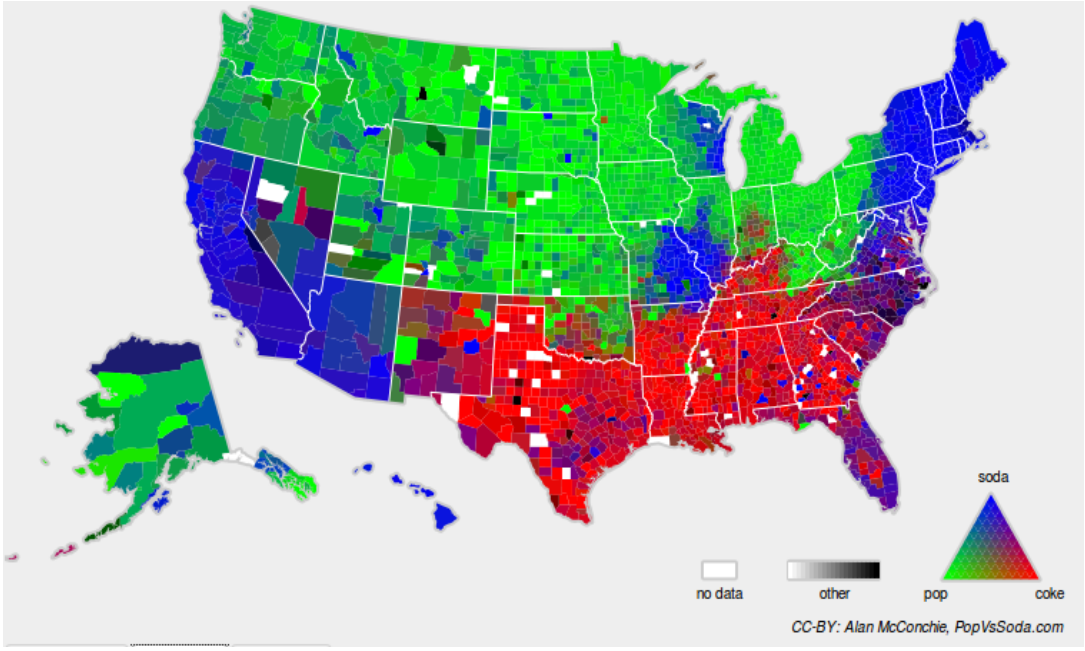


Figure 3.1: Répartition des termes utilisés pour désigner le prototype de la classe « boisson gazeuse »

Source : Pop vs Soda (http://popvssoda.com/)

En fonction des régions, trois prototypes de classe différents émergent soient *pop, soda* et *coke.* Ainsi, les états sudistes ont tendance à utiliser le terme « *coke* » pour désigner une boisson gazeuse, alors que les états du nord utilisent plutôt « *pop* » et que le terme « *soda* » est plutôt favorisé dans le nord-est (New-York), le sud-ouest (Californie), à la frontière de l'Illinois et du Missouri ainsi qu'à l'est du Wisconsin. Plusieurs angles d'analyses peuvent être développés à partir de ces données, deux de ceux-ci semblent particulièrement pertinents au regard de la création et de l'utilisation de folksonomies.

Premièrement, cette étude nous rappelle l'importance du facteur culturel, même lorsque le terrain de recherche ne se limite qu'à un seul pays. Les États-Unis sont un

bel exemple de diversité culturelle, mais des spécificités régionales quant à l'emploi de certains mots s'observent aussi dans les autres pays. On n'a qu'à penser aux différences régionales entre Montréal et une région comme le Témiscamingue ou encore entre le Sud et le Nord de la France pour s'en convaincre.

Deuxièmement, on doit se demander comment ces différences dans l'identification d'un terme générique influencent la construction d'une folksonomie en ligne et également poser la question des différences d'utilisation à l'oral et à l'écrit. Par exemple, en recueillant des corpus générés par des requêtes à la S*treaming* API de Twitter on remarque que c'est le mot-clic #soda qui est largement favorisé pour désigner une boisson gazeuse. La polysémie du mot-clic #pop qui, sur Twitter du moins, est plutôt utilisé pour étiqueter des ressources en lien avec la musique populaire, et le lien direct entre #coke et l'entreprise Coca-Cola ont possiblement favorisé l'émergence d'une utilisation commune du mot-clic #soda, malgré qu'il ne soit pas le terme le plus utilisé par une grande partie de la population américaine.

L'un des commentaires des participants à ce sondage illustre bien d'ailleurs l'influence du groupe dans le choix des termes : « *I use either one depending on who I'm talking to. Soda, for some reason, seems more formal, pop seems like a slang term from where I grew up in Indiana. I think I started using Soda when I moved to Chicago. When our family lived in Hannibal Missouri, the other kids made fun of us when we said pop* ». Cet autre commentaire illustre aussi la différence entre le choix des termes à l'oral et à l'écrit : « *If I use Coke it means Coca-cola but no other soft drink. There, I just used soft drink and I didn't even realize it. I never SAY soft drink, I must just use it when I write.* »

Cet exemple simple à propos du choix de termes pour désigner une boisson gazeuse aux États-Unis nous permet d'introduire certaines des limites liées à la construction des folksonomies. À la base, Peterson (2006) relève qu'en empruntant une voie qui

s'éloigne de la catégorisation aristotélicienne et en optant pour un certain relativisme culturel et un constructivisme épistémologique, cela permet de contourner les limites de la classification traditionnelle. Toutefois, parce qu'elle se nourrit des opinions et des perceptions de plusieurs individus, une telle approche prête le flanc à des incohérences ou des contradictions où, par exemple, une ressource pourrait être à la fois à propos de A et pas à propos de A.

Dans le même ordre d'idées, Mathes (2004) souligne que même si l'émergence de sens est observable à l'intérieur de folksonomies, il y émerge également des ambiguïtés où des utilisateurs utilisent les mêmes mots-clés à des fins différentes ou encore à l'opposé utilisent des termes synonymes pour décrire un même concept sans qu'il y ait au préalable de véritables mécanismes de détection de synonymes. Par exemple, il note que lors de ses recherches sur la plateforme *del.icio.us* le mot-clé « *filtering* » était utilisé pour décrire ces différentes ressources :

1. *Last.FM - Your personal music network - Personalized online radio station*
2. *InfoWorld: Collaborative knowledge gardening*
3. *Wired 12.10: The Long Tail*
4. *Oh My God It Burns! “ Practical Applications of the Philosopher’s stone. For drunks. Brita filter makes bad vodka into good vodka*
5. *Introduction to Bayesian Filtering*

L'utilisation d'acronymes est également sujette à la création d’ambiguïtés à l'intérieur de folksonomies s'étendant à plusieurs domaines. Mathes (2004) relève que l'emploi du mot-clé « ANT » par certains utilisateurs dans le domaine de la sociologie est un acronyme signifiant « Actor Network Theory ». Par contre, dans le domaine informatique il est plutôt associé au projet de logiciel libre Apache Ant, alors qu'en sciences de la nature il fait référence à la fourmi.

La folksonomie hérite donc des problèmes courants que l'on retrouve dans les vocabulaires non supervisés c'est-à-dire l’ambiguïté, la polysémie, la synonymie et la

difficulté de classer les connaissances en fonction de domaines d'applications. C'est exactement le défi auquel on fait face sur Twitter, une sphère largement dominée par l'actualité avec des flux de millions de gazouillis à l'heure et pour laquelle il est impossible d'opérer une classification classique.

3.2 Champs lexicaux émergeant d’une folksonomie

La folksonomie permet d'observer l'émergence d'une nomenclature commune obtenue par le regroupement de classifications de ressources cooccurrentes. Nous allons maintenant nous intéresser à ces ensembles et à leurs propriétés lexicales et sémantiques. Plus précisément, nous serons attentifs à l'émergence de champs lexicaux soit l'association de termes qui appartiennent à une même thématique.

Un champ lexical est formé de mots d'une même famille, de synonymes ou d'autres mots qui ont un rapport étroit avec une thématique (Pablo, 2000). Dans le cas qui nous intéresse, notre attention se situe au plan de la formation de champs lexicaux autour des mots-clics de la requête en cours. Cette notion de champ lexical composé de mots-clics cooccurrents nous est d'abord apparue par un concours de circonstances. Lors d’une recherche portant sur la quête de visibilité de la communauté canadienne-française sur Twitter, nous avons noté que parmi les mots-clics présentant une forte cooccurrence, il semblait se dégager une cohésion thématique. Une première piste vers l'émergence d'un champ lexical se dégageait ainsi.

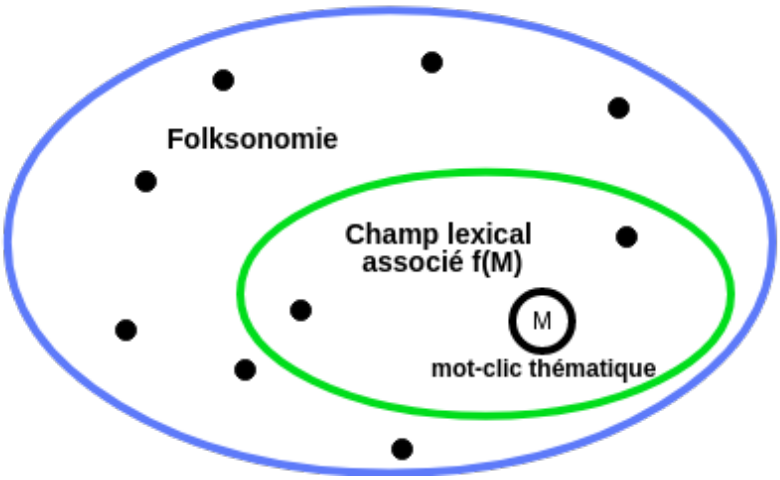


Figure 3.2: Délimitation d'un champ lexical à l'intérieur d'une folksonomie

De plus, il se trouvait quelques mots-clics qu'il aurait été souhaitable d'ajouter à notre requête. Ces mots-clics, nous semblait-il, appartenaient non seulement à une thématique commune, mais correspondaient à une sous-thématique. Ces premières observations ne nous permettaient pas de conclure à de quelconques relations sémantiques entre mots-clics, mais attirèrent néanmoins notre attention.

3.2.1 Formation de champs lexicaux

Afin de valider nos observations initiales dans la communauté canadienne-française, des échantillons supplémentaires de gazouillis à propos de thématiques variées ont été récupérés. Nous avons ainsi été en mesure de vérifier si nos premières observations étaient également valides pour des corpus aussi divers que des matchs de soccer en France, des élections au Québec ou encore des émissions de télévision d'ici ou d'ailleurs.

Pour chacun des échantillons de gazouillis recueillis, à l'exception d'un sur lequel nous reviendrons (p.51), une cohésion thématique parmi les mots-clics présentant une

forte cooccurrence a été observée. Par exemple, dans le cas du match de soccer opposant l'AS Monaco à l'En Avant de Guingamp[4], nous avons remarqué une forte cooccurrence de mots-clics autour des noms des deux équipes, des joueurs vedettes, du diffuseur du match (France 2) et de la Coupe de France, une compétition de soccer au sein de laquelle les deux équipes s'affrontaient.

Tableau 3.1: Champ lexical du match AS Monaco 'vs' En Avant de Guingamp

Mots-clics thématiques : #AS_Monaco, #EAGuingamp						
Rang	**Mot-clic**	**Pourc. de cooccurrence**		**Rang**	**Mot-clic**	**Pourc. de cooccurrence**
1	CDF	16,4%		9	guingamp	2,7%
2	FCBRMA	13,7%		10	RT	2,5%
3	teamEAG	9,8%		11	GoMonaco	2,4%
4	coupedefrance	6,8%		12	rmclive	2,2%
5	DagheMunegu	4,1%		13	France2	2,0%
6	EAG	3,7%		14	ASMonaco	1,9%
7	Berbatov	3,7%		15	ASMFC	1,5%
8	Monaco	3,1%		16	barcareal	1,5%
				17	Clasico	1,4%

Dans le corpus de gazouillis à propos des élections québécoises de 2014, à partir du mot-clic thématique #qc2014, nous avons constaté une forte cooccurrence de mots-clics autour des partis politiques, des chefs de partis, de l'Assemblée nationale et d'actualités à propos entre autres d'une manifestation en cours et de l'annonce de l'élection d'un gouvernement libéral majoritaire.

[4] Dans le cas d’un match opposant deux équipes sportives, nous avons utilisé deux mots-clics thématiques désignant chacune des équipes.

Tableau 3.2: Champ lexical durant les élection québécoises de 2014

Mot-clic thématique : #qc2014						
Rang	**Mot-clic**	**Pourc. de cooccurrence**		**Rang**	**Mot-clic**	**Pourc. de cooccurrence**
1	polqc	17,2%		15	qcpoli	1,6%
2	PQ	15,9%		16	rcqc	1,4%
3	PLQ	14,8%		17	rcma	1,4%
4	caq	9,3%		18	CEIC	1,4%
5	assnat	7,6%		19	BazzoTV	1,4%
6	electionsqc2014	7,6%		20	votehuff	1,2%
7	manifencours	5,1%		21	qcvotes	1,2%
8	qs	4,8%		22	Couillard	1,1%
9	polqc2014	3,4%		23	cdnpoli	1,1%
10	Québec	2,8%		24	TVA	1,0%
11	PKP	2,4%		25	Marois	0,9%
12	PLQmajoritaire	2,3%		26	Montréal	0,9%
13	charte	2,3%		27	VRAIESAFFAIRES	0,8%
14	quebec2014	1,8%		28	electionquebec	0,8%
				29	byebyePauline	0,6%

Du côté des émissions de variétés, nous avons récolté des gazouillis durant l'émission franco-belge de téléréalité culinaire « Top Chef » et à partir du mot-clic thématique #topchef nous avons observé une forte cooccurrence avec des mots-clics au sujet des participants, des chefs invités et de la chaine diffusant l'émission.

Tableau 3.3: Champ lexical de l'émission Top Chef

Rang	**Mot-clic**	**Pourc. De cooccurrence**		**Rang**	**Mot-clic**	**Pourc. de Cooccurrence**
1	topchef2015	22,9%		10	buche	2,7%
2	TeamGG	14,6%		11	Gibier	2,1%
3	Kevin	5,6%		12	xavier	1,7%
4	Olivier	4,4%		13	TeamKévin	1,3%
5	m6	4,2%		14	ConseilDexpert	1,2%

Rang	Mot-clic	Pourc. De cooccurrence		Rang	Mot-clic	Pourc. de Cooccurrence
6	Blaguea1franc	3,3%		15	broadchurch	1,1%
7	blaguepourrie	3,1%		16	JFPiège	1,1%
8	CroisonsLes	2,8%		17	teamxavier	0,8%
9	teamolivier	2,7%		18	qotd	0,8%

Cette cohésion thématique se dégageant de plusieurs corpus donne un socle plus solide aux premières observations et rejoint les conclusions de plusieurs recherches : « Under the corpus-based approach, word relationships are often derived from their co-occurrence distribution in a corpus » (Church et Hanks, 1989; Hindle, 1990; Grefenstette, 1992; cités dans Jiang et Conrath, 1997).

Le terme «relationships» dans le cas d'un mot-clic avec son ou ses mots-clics thématiques suggère une approche statistique, laquelle a été développée dans la section Relations sémantiques émergeant d'un champ lexical (p.53). En effet, même si la cooccurrence entre deux termes exprimée en valeur absolue peut apparaître simpliste, la force de cette dernière exprimée par un degré d'intensité, un indice de cooccurrence, nous dirige vers une hypothèse où cette cooccurrence deviendrait significative d'un point de vue statistique et potentiellement lexical et sématique. Ainsi, selon la force de l'indice de cooccurrence, il semble envisageable de faire le passage entre mots-clics cooccurrents et mots-clics qui sont des « corrélats » au sens de deux termes cooccurrents qui ont une relation de sens (Bourion, 2001).

Ce passage entre un indice de cooccurrence, une relation statistique entre deux mots-clics, et un corrélat, une relation de sens, recèle un potentiel danger de confusion de définition épistémologique entre différents plans de l'analyse (Martinez, 2003). On passe en effet du constat statistique, la cooccurrence, à une signification lexicale de cette relation. Nous tenterons de démontrer que la cooccurrence porte en elle-même un potentiel important pour la création de champs lexicaux puis éventuellement pour

l’identification de relations sémantiques. Toutefois, nous distinguons ce qui relève de la description formelle d’un corpus de gazouillis par l’usage de la statistique et de l’informatique et le sens ou la relation de sens toujours négociable.

Suite aux tests décrits ci-dessus, nous sommes en mesure d'observer que les mots-clics ayant une forte cooccurrence entre eux tendent effectivement à former un champ lexical. Il faut toutefois rester prudent face à cette observation, car s’il y a une appartenance claire à un même champ lexical lorsque la cooccurrence est à son plus fort, la cohésion thématique entre mots-clics s'estompe graduellement lorsque la cooccurrence s'affaiblit.

3.2.2 Courbe de distribution et loi de Zipf-Mandelbrot

S’il y a bien corrélation entre la formation d’un champ lexical et une forte cooccurrence entre mots-clics, il devient donc essentiel de bâtir un modèle capable de distinguer une forte cooccurrence d’une faible. Or, dans la littérature sur les folksonomies, on remarque que la distribution de mots-clés attribués par des usagers correspond souvent à une loi de puissance (Mathes, 2004 ; Cattuto, Loreto et Pietronero, 2007 ; Guy et Tonkin, 2006). L'examen de nos corpus de gazouillis a permis de constater que la distribution de mots-clics suivait le même modèle. Plus précisément, dans nos corpus la loi de puissance qui s'observe est celle de Zipf-Mandelbrot. C'est-à-dire que l'occurrence d'un mot-clic dans un corpus décroit de façon exponentielle en fonction de son rang dans l'ordre des fréquences. Cette observation correspond à ce qu’on retrouve dans la littérature sur l’analyse des fréquences de mots sur des corpus provenant de Twitter (Amati *et al.*, 2012) et c’est d’ailleurs un type de distribution que l’on retrouve fréquemment dans des corpus à base de textes.

Cette loi a été formulée par George Zipf (1949). Dans ces études de linguistique, il étudia la fréquence des mots dans différentes langues. En faisant ce décompte des

fréquences de mots sur l'œuvre d'Ulysse de James Joyce, il observa que le mot le plus souvent utilisé l'est deux fois plus que le deuxième, trois fois plus que le troisième, etc. Cette type de répartition de la fréquence d'un mot en fonction de son rang s'est vérifiée dans de nombreux corpus.

Les raisons pour lesquelles un tel type de distribution s'observe dans toutes les langues sont toujours discutées. Zipf lui-même (1949), proposa une adaptation de la loi du moindre effort stipulant que l'humain tente de communiquer de façon efficace tout en minimisant son effort et donc le nombre de mots employés pour s'exprimer.

Selon Zipf, dans un corpus donné, la fréquence d'occurrences d'un mot est liée à son rang dans l'ordre des fréquences par une loi de la forme $f(n) = \frac{K}{n}$ où K est une constante (Caron, 2004). Par exemple, si K=5000 et si le mot le plus souvent employé dans un corpus a 5000 occurrences, le dixième mot aura approximativement 500 occurrences et le centième 50. Mandelbrot proposa une amélioration de ce modèle en introduisant des paramètres supplémentaires afin de mieux correspondre aux données du langage naturel. Nous avons pu constater à partir de tests préliminaires que nos données correspondaient à cette logique de distribution.

Une deuxième façon de s'assurer que la cooccurrence des mots-clics correspond à cette loi de puissance est d'appliquer un logarithme naturel (népérien) à la fois au nombre d'occurrences et au rang d'un mot-clic et d'en tirer un graphique en deux dimensions. C'est ce que nous avons entrepris en nous servant des données recueillies sur Twitter lors de notre recherche sur la communauté canadienne-française. Dans ce corpus, 659 mots-clics distincts ont été recueillis. Le mot-clic qui y figure le plus souvent apparaît 935 fois, le mot-clic qui occupe le dixième rang y apparaît 71 fois alors que celui qu'on retrouve au 100ème rang apparaît 6 fois conformément à ce que l'on pourrait s'attendre dans ce type de distribution.

Selon cette loi de puissance, en mettant la valeur du logarithme naturel du rang d'un mot-clic (ln(rang)) sur l'axe des X et le logarithme naturel du nombre d'occurrences de mots-clics (ln(occurrence)) sur l'axe des Y, nous obtenons une droite avec une pente près de -1 caractérisée par un léger coude dans le haut de la courbe. C'est précisément le type de courbe que nous avons obtenue pour le corpus sur la communauté canadienne-française ainsi que dans d'autres corpus ayant fait partie de nos analyses préliminaires.

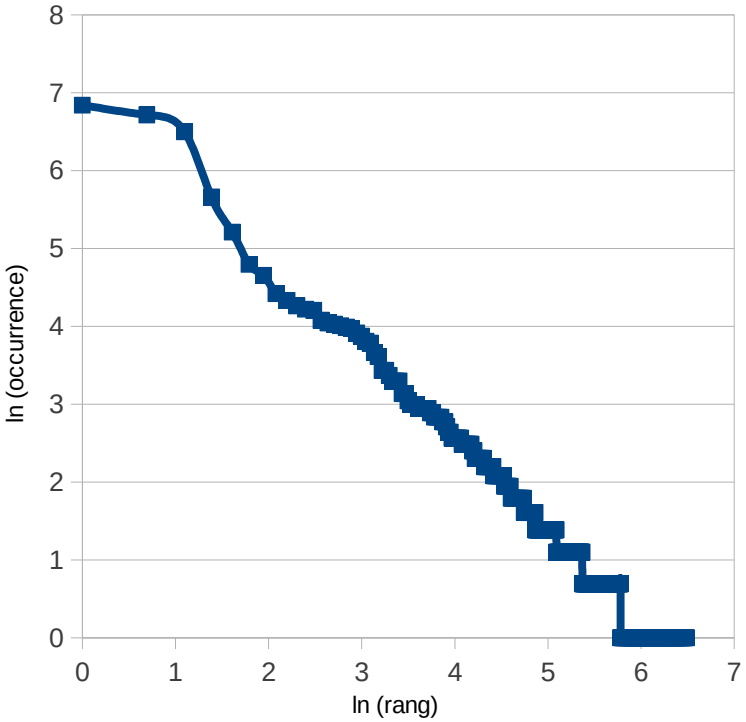


Figure 3.3: Courbe de distribution de Zipf-Mandelbrot appliquée au corpus de gazouillis de la communauté canadienne-française

Dans certains cas, lorsque la durée de la cueillette est très courte ou que le mot-clic thématique génère très peu de gazouillis de plus petits corpus sont récoltés et ces

derniers semblent alors s’éloigner d’une distribution zipfienne.

Un tel un modèle de distribution applicable à la cooccurrence des mots-clics de nos corpus nous fournit les outils nécessaires pour déterminer le point à partir duquel la cooccurrence d'un mot-clic est suffisamment forte pour qu'il soit considéré comme faisant partie du champ lexical. La question qui se pose toutefois est de savoir où se trouve ce point.

Nos expérimentations montrent que ce point en haut duquel un mot-clic devrait faire partie du champ lexical et en bas duquel il devrait en être exclu est imprécis. Il existe bien un lien directe entre une forte cooccurrence et la formation d’un champ lexical, la difficulté réside dans la mesure de la force de la cooccurrence et dans l’identification d’un point où elle n’est plus assez forte. Quelques avenues ont été explorées comme la sélection d’un pourcentage de mots-clics, mais cette dernière s’est avérée décevante. L’approche retenue fut celle où nous avons comparé l’évolution du logarithme naturel (népérien) du nombre d’occurrences à celle du rang d'un mot-clic.

Dans la figure qui suit, les courbes obtenues par le logarithme naturel du rang d'un mot-clic (ln(rang)) et le logarithme naturel du nombre d'occurrences de mots-clics (ln(occurrence)), illustrent bien la nature exponentielle de ces fonctions logarithmiques. La jonction de ces deux courbes où In(rang) et In(occurrence) sont égaux indique la valeur sous laquelle la cooccurrence est trop faible pour qu'un mot-clic soit considéré dans le champ lexical.

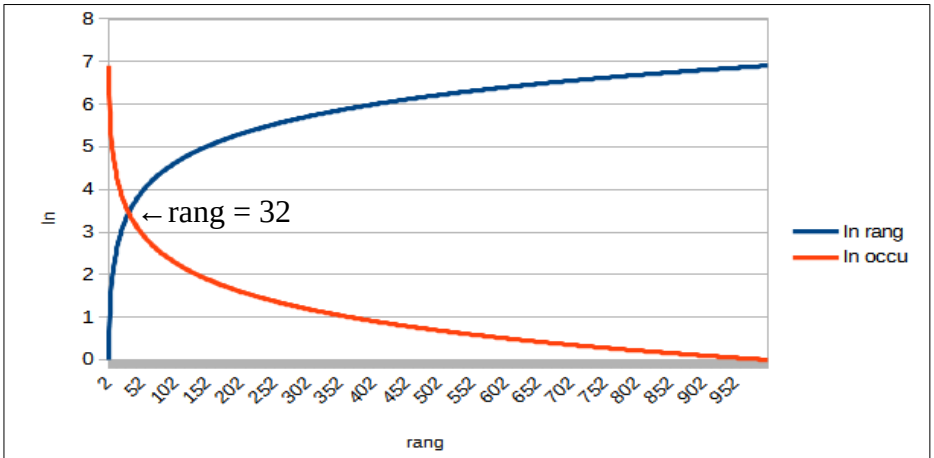


Figure 3.4: Jonction des courbes ln(rang) et ln(occurrence)

Par exemple, comme l'illustre la figure 3.4, sur un total de 1000 mots-clics, suivant la loi de Zipf-Mandelbrot, les valeurs de ln(occurrence) sont inférieures à ln(rang) à partir du rang #32. Ainsi, les 32 mots-clics ayant la plus forte cooccurrence seraient identifiés comme candidats potentiels à ajouter à la requête en cours. La loi de Zipf-Mandelbrot étant exponentielle, le nombre de candidats à tester croitra moins rapidement que le nombre total de gazouillis et de mots-clics. Par exemple, pour un nombre total de 10 000 gazouillis seulement les 100 premiers mots-clics sont sélectionnés pour faire partie du champ lexical.

Cette piste de solution pour déterminer un point en haut duquel la cooccurrence d'un mot-clics est suffisamment forte pour l'inclure dans le champ lexical demeure toutefois expérimentale, mais elle fut retenue suite à nos observations et une validation par des évaluateurs externes. Elle donne de bons résultats (p.98) et constitue un outil efficace pour former des champs lexicaux à partir de mots-clics. Nous la retenons aussi pour sa capacité à s'adapter à des corpus de différentes tailles.

3.2.3 Influence de l'entropie d'un mot-clic dans la formation d'un champ lexical

Une exception survenue lors de l'analyse de nos premiers corpus nous a permis de

constater que la cooccurrence de mots-clics ne participait pas toujours à la construction d'un champ lexical cohérent. En effet, on assiste parfois à la création de plusieurs champs lexicaux dont les contours et limites sont difficiles à tracer. En utilisant comme mot-clic thématique #canada, nous avons constaté des cooccurrences fortes avec des thèmes et des sujets ayant aussi peu de liens entre eux que des offres d'emplois (#Jobs, #TweetMyJob), des références à l'islam (#letter4u), des références géographiques (#USA, #Alberta, #Cuba, #CostaRica) et des références à des produits de consommation (#Luxury, #Shoes). S'ajoute à cette diversité thématique, le facteur de temporalité (p.67) qui parfois fait en sorte qu'un mot-clic peut être associé à d'autres thématiques par les usagers en fonction de l'actualité comme cela pourrait être le cas pour #canada lors de la tenue d'un championnat du monde de hockey.

De façon générale, lors de la construction d'un champ lexical à partir de folksonomies, l'un des prérequis à l'ajout de termes significatifs est que ces derniers soient utilisés dans le même contexte par différents utilisateurs (Laniado et Mika, 2010). Toutefois, certains termes ont des champs sémantiques si larges qu'il devient difficile de les contextualiser. Ce n'est pourtant pas le cas pour le mot « Canada », qui, en tant que nom propre décrivant un pays, a un champ sémantique restreint.

Par contre, il semblerait que dans le cadre de son utilisation sur Twitter l'usage de ce mot-clic serve à discuter de thématiques très différentes les unes des autres. On dira d'un mot-clic qu'il a une forte entropie lorsqu'il est utilisé dans de multiples contextes (Laniado et Mika, 2010). Par conséquent, son utilisation dans la création d'un champ lexical devient problématique. Dans le cas du mot-clic thématique #Canada, on peut distinguer au moins quatre thématiques différentes (emploi, islam, géographie et produits de consommation) qui le rendent non pertinent pour le type de suivi de discussions que nous souhaitons mettre en place. En résumé, l'approche préconisée s'avère fonctionnelle lorsque la cueillette démarre avec un mot-clic thématique ayant une entropie relativement faible ou moyenne. Souvent, le niveau de

cette entropie n’est observable qu’en faisant des tests de cueillette et en observant l’homogénéité du champ lexical.

3.3 Relations sémantiques émergeant d’un champ lexical

L'hypothèse voulant que la formation de champs lexicaux soit observable lorsqu'il y a une forte cooccurrence entre mots-clics s'est avérée dans les corpus que nous avons analysés. La question subséquente porte sur la possibilité d'identifier des mots-clics représentant des sous-thématiques de la thématique principale.

3.3.1 Sous-thématiques et relations *partieDe*

Sur le plan sémantique, nous définissons une sous-thématique comme étant une *partieDe* la thématique principale. Nous cherchons donc à découvrir des liens de type partie-tout qui unissent certains mots-clics présents dans un champ lexical. À ce sujet, Kadlec (2010) indique qu'à l'intérieur des champs lexicaux on retrouve des termes dont les champs sémantiques se recoupent (s*emantic relatedness*). C’est ce degré d’appartenance (relatedness) qui détermine pour nous si un mot-clic représente une sous-thématique ou non. Notre définition de la relation *partieDe* en est donc une d’inclusion thématique liant des sous-thématiques à une thématique principale.

Nous établissons cette relation *partieDe* comme étant antisymétrique, réflexive et transitive. Prenons par exemple la politique québécoise pour illustrer chacune des propriétés de notre relation. On prendra comme mots-clic thématique, #polQC, et comme mots-clics faisant partie de sous-thématiques, #PQ et #JFLisée. Il est important de rappeler que la ou les relations que l’on établit ne sont pas entre les objets ou les personnes, mais plutôt entre les thématiques de discussions liées à un mot-clic. Ainsi, la relation entre #JFLisée et le #PQ, n’est pas une relation entre le chef et son parti politique, mais plutôt une relation entre les discussions à propos de l’un et l’autre.

Pour l'ensemble de mots-clics MC = {(#polQc, #PQ, #JFLisée)}, la relation R *partieDe* pourrait être définie ainsi : R = { (#JFLisée, #PQ), (#PQ, #polQc)}

Cette relation est antisymétrique, car si la discussion à propos du #PQ est une partie de la discussion à propos de #polQc, alors la discussion à propos de #polQc n'est pas une partie de la discussion à propos du #PQ. Elle est réflexive, car chacune des discussions est une partie d'elle-même. Puis, finalement elle est transitive puisque si la discussion à propos de #JFLisée est une partie de la discussion à propos du #PQ et que la discussion à propos du #PQ est une partie de la discussion à propos de #polQc, alors la discussion à propos de #JFLisée est une partie de la discussion à propos de #polQc.

Notre définition de la relation *partieDe* entre une sous-thématique et une thématique correspond à la fois aux trois axiomes de la méréologie classique (Hovda, 2009) et à une utilisation courante de la relation *partOf* telle que définie dans l'ontologie OWL. Chaque fois que nous utilisons le prédicat RDF *(Resource Description Framewor*k)[5] *owl:partOf*, nous le faisons en référence à la définition de la relation que nous venons de produire.

Un prédicat RDF est un type de relation que l'on peut appliquer entre un sujet et son objet. Il est ainsi au centre d'un triplet RDF sous la forme « Sujet, Prédicat, Objet ». Donc, dans notre exemple précédent, deux triplets seraient ainsi formés :

- « #JFLisée → owl:partOf → #PQ »
- « #PQ → owl:partOf → #polQc »

À cause de la propriété de transitivité de cette relation, on peut également inférer un troisième triplet :

- « #JFLisée → owl:partOf → #polQc»

[5] Description provenant du groupe de travail sur les normes RDF du W3C: http://www.w3.org/RDF/

Dans le but de vérifier la présence de telles relations *partieDe,* une analyse manuelle a été effectuée à partir d'une dizaine de corpus de gazouillis portant sur des thématiques variées traitant de sport, de culture ou de politique. Pour chacun des corpus, les dix mots-clics présentant la plus forte cooccurrence ont été extraits et une définition de ces mots-clics a été obtenue parfois à l'aide d'un dictionnaire ou d'un service web de définition de mots-clics, mais, dans la majorité des cas, en lisant nous-mêmes les gazouillis. On a ensuite tenté de vérifier la présence d'une relation *partieDe* et de lui attribuer le prédicat *owl:partOf*, le cas échéant.

Dans le tableau qui suit, le triplet RDF est formé d'un **sujet** qui est une discussion à propos d'un mot-clic du champ lexical, du **prédicat** *owl:partOf* lorsqu'une relation *partieDe* peut être attribuée, et en troisième lieu, d'un **objet** qui est la discussion recueillie par les mots-clics de la requête en cours à propos de l'émission Top Chef. Par exemple, pour le mot-clic #Etchebest un triplet serait ainsi formé :

- #Etchebest → *owl :partOf* → «requête en cours »

Le **sujet** est la discussion à propos de #Etchebest, le **prédicat** est *owl:partOf* et l'**objet** est la discussion recueillie par la requête en cours. Si l'attribution du prédicat *owl :partOf* n'est pas possible, nous ne cherchons pas à en attribuer un autre, car notre objectif est strictement de découvrir ou non la présence de sous-thématiques.

Tableau 3.4: Émission Top Chef: champ lexical et relations partieDe

Mot-clic thématique : #topchef		
Mot-clic du champ lexical	**Prédicat RDF**	**Description**
topchef2014	*owl :partOf*	Édition 2014 de l'émission « Top Chef »
Etchebest	*owl :partOf*	Philippe Etchebest, chef cuisinier français
M6	-	Chaine de télévision généraliste française
Steven	*owl :partOf*	Steven Ramon, participant à « Top Chef »

Mot-clic thématique : #topchef		
Bachelor	-	Émission de télévision
GameOfThrones	-	Émission de télévision
alexis	*owl :partOf*	Alexis Braconnier, participant à « Top Chef »
Telecheck	-	Application social TV développée par Orange
Noemie	*owl :partOf*	Noémie Honiat, participante à « Top Chef »
MasterChef	-	Émission de téléréalité

Ce tableau montre que le prédicat *owl:partOf* est fréquemment observable entre les discussions à propos d'un mots-clics du champ lexical et la discussion portant sur la thématique principale recueillie par la requête en cours. Par exemple, #topchef2014, soit la discussion à propos de l'édition 2014 de l'émission est une partie de #topchef, une discussion sur la série, la discussion à propos de #Noemie est également une partie de la discussion à propos de #topchef comme le sont les discussions à propos des participants #alexis et #Steven ainsi que celle sur le chef cuisinier #Etchebest.

Dans le cas du corpus de gazouillis et de mots-clics recueilli pour l'émission « Top Chef », des prédicats RDF plus spécifiques auraient pu être employés tels que « #topchef2014 → *foaf:currentProjet* → #topchef » ou « #Etchebest → *foaf:member* → #topchef » ou encore « #M6 → *dcterms:publisher* → #topchef », mais, encore une fois, notre motivation étant de découvrir la présence du prédicat *owl:partOf,* synonyme pour nous d'une sous-thématique, la découverte d'autres prédicats présente peu d'intérêt.

Nous avons répété cette expérimentation visant à vérifier la présence de relations *partieDe* de façon plus exhaustive sur plusieurs corpus lors de la validation effectuée par des évaluateurs externes. Elle s'est avérée positive au sens où des relations *partieDe* ont pu être observable en grand nombre dans chacun des corpus. Les

résultats de ces expérimentations peuvent être consultés à l'Annexe C (p.179)

Il est donc possible d'observer des relations *partieDe* à l'intérieur d'un champ lexical, mais encore faut-il savoir ce que ces mots-clics signifient pour pouvoir établir des relations. Le défi de taille qui se pose dans le cas des mots-clics est qu'on ne dispose pas de ressources linguistiques pouvant les définir. En effet, notre matière première est constituée de termes formés par une série de lettres pour lesquels aucun dictionnaire, thésaurus, ontologie ou autre outil sémantique ne peut apporter de définitions et encore moins d'informations quant aux concepts, domaines d'applications ou thématiques sous-jacentes. Bien sûr, il y a quelques exceptions lorsque le mot-clic est un calque d'un nom commun (#Shoe) ou d'un nom propre (#RémiGaillard) et quelques services web tels que « tagdef.com » ou « hashtags.org » font appel à la production participative (*crowd sourcing*) afin de donner des définitions aux mots-clics, mais seule une infime partie de ceux-ci y est définie. Le champ sémantique d'un mot-clic, soit l'ensemble des sens qu'il peut avoir selon le contexte, est donc une variable la plupart du temps inconnue dans la twittosphère ou, pour être plus précis, les utilisateurs possèdent une connaissance relative du champ sémantique d'un mot-clic qu'ils utilisent, mais cette dernière est rarement documentée.

Il y a donc pénurie d'outils lexico-sémantiques permettant de définir la thématique de la discussion liée à un mot-clic et, conséquemment, il devient difficile d'identifier des liens de type *partieDe* entre les mots-clics d'un champ lexical. Nous avons produit manuellement des définitions pour les mots-clics de plusieurs champs lexicaux à des fins d'expérimentations et de validation pour vérifier la présence de relations *partieDe*. Maintenant que cette présence s'avère, notre travail consiste à tenter de les identifier de manière automatique sans avoir à définir leurs significations.

3.3.2 Calculs de similarité sémantique

La relation *partieDe* que nous venons de décrire peut également être schématisée en

utilisant le concepts d'intersection (*overlap*) de la théorie des ensembles. Prenons comme premier ensemble la discussion à propos de l'émission TopChef et comme deuxième ensemble la discussion à propos de #topchef2014.

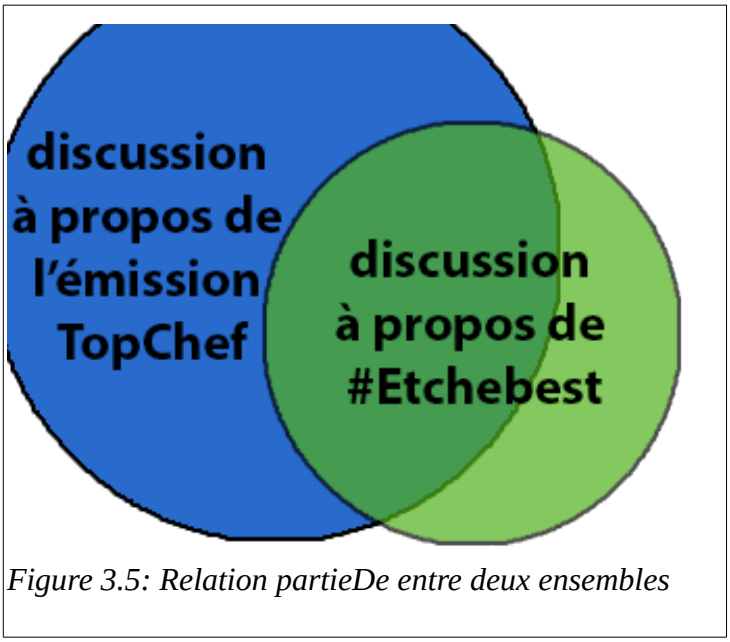


Figure 3.5: Relation partieDe entre deux ensembles

L'ensemble « discussion à propos de l'émission TopChef » est constitué de gazouillis contenant au moins un ou plusieurs mots-clics de la requête en cours alors que l'ensemble « discussion à propos de #topchef2014 » est constitué de gazouillis contenant au moins le mot-clic #topchef2014. La zone d'intersection constitue un sous-ensemble où les mots-clics de la requête en cours et #topchef2014 se retrouvent en cooccurrence dans les mêmes gazouillis. Cette schématisation de la relation *partieDe* permet d'illustrer une forme d'inclusion partielle. On pourrait la traduire en disant qu'une majorité de gazouillis de la discussion à propos de #topchef2014 contiennent un ou plusieurs des mots-clics de la requête en cours, car la zone d'intersection est plus grande que la zone de disjonction. C'est donc aussi dire qu'une discussion à propos de #topchef2014 est plus souvent une discussion à propos de

l'émission Top Chef qu'à propos de tout autre thématique.

Nous tissons ainsi un lien entre le niveau de cooccurrence des mots-clics appartenant à des ensembles de discussions et leur degré d'appartenance à une thématique. Autrement dit, nous estimons que plus cette cooccurrence est élevée, plus la probabilité de découvrir une relation *partieDe* l'est également.

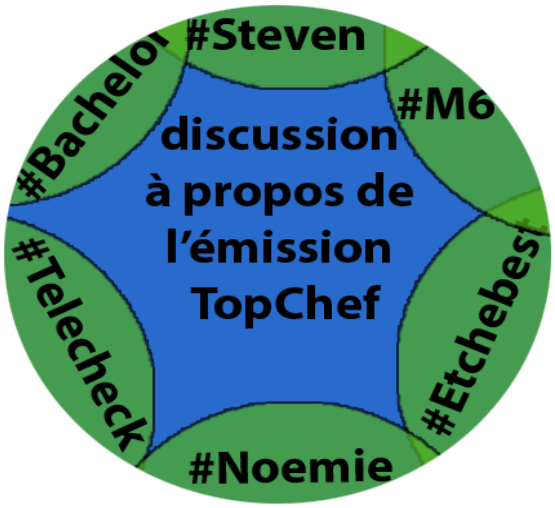


Figure 3.6: Ensembles partiels de mots-clics candidats (mot-clic thématique : #topchef)

Rappelons toutefois qu'une cueillette de gazouillis en provenance de la *Streaming API* s'effectue à partir de la requête en cours. Comme l'illustre la figure ci-dessus, le seul ensemble de gazouillis qui est donc complet est celui des mots-clics de la requête en cours. L'ensemble des gazouillis de la discussion à propos de #Etchebest ou des autres mots-clics du champ lexical n'est donc que partiel et nous est visible seulement parce que ces mots-clics sont cooccurrents avec un ou plusieurs des mots-clics de la requête en cours. L'ensemble des gazouillis de la twittosphère contenant un mot-clic

du champ lexical, mais ne contenant pas un mot-clic de la requête en cours nous est donc inconnu.

En terme d'ensembles, la figure ci-dessus n'est qu'une transposition du champs lexical, c'est-à-dire une représentation de la cooccurrence des mots-clics candidats avec ceux de la requête en cours comme nous l'avons déjà calculé dans la section sur la formation des champs lexicaux (p.43). Pour chacun de ces mots-clics candidats, il nous manque donc une partie de l'ensemble de la discussion qui y est associée permettant d'évaluer son intersection avec l'ensemble de la discussion à propos de l'émission Top Chef.

Pour pallier à ce manque d'information, nous avons mis en place ce que nous avons appelé un indicateur sémantique de cooccurrence inversée afin d'évaluer le degré d'appartenance de chacun des mots-clics du champ lexical à la thématique. Pour ce faire, nous créons des ensembles distincts composés de gazouillis à propos de chacun des mots-clics candidats à partir de la *Search API* pour ensuite y mesurer la cooccurrence des mots-clics de la requête en cours dans chacun de ces ensembles.

Par exemples, dans la figure 3.5 (p.58) pour réussir à vérifier le niveau de cooccurrence entre la discussion à propos de #topchef2014 et la discussion à propos de l'émission TopChef (#topchef), on calcule l'indicateur sémantique de cooccurrence inversée en recueillant sur la *Search API* des gazouillis de la discussion à propos de #topchef2014, et, pour chacun des gazouillis, on vérifie la présence ou non de l'un des mots-clics de la requête en cours. En recueillant un ensemble de la discussion à propos de #topchef2014 sur la *Search* API, il est donc possible de mesurer le niveau de cooccurrence inversée, c'est-à-dire celle des mots-clics de la requête en cours avec celle du mot-clic candidat (#topchef2014) .

Prenons un exemple plus complexe, où la requête en cours serait composée de deux mots-clics, #topchef et #topchef2014, et où #Etchebest ferait partie du champ lexical.

Nous cherchons à vérifier si la discussion à propos de #Etchebest est une *partieDe* la thématique principale. Pour ce faire, nous calculons un indicateur sémantique de cooccurrence inversée avec le candidat #Etchebest. Ce calcul va comme suit : dans un ensemble de 100 gazouillis recueillis sur la *Search API* pour #Etchebest, si 85 gazouillis contiennent le mot-clic #topchef et/ou #topchef2014, l'indicateur sémantique de cooccurrence inversée sera alors de 0,85 (85/100). Notons que ce qui nous importe lorsqu'on effectue ce calcul est de savoir si oui ou non un gazouillis contient un mot-clic de la requête en cours. C'est donc un test binaire de type VRAI ou FAUX. Si un gazouillis contient un mot-clic de la requête en cours, ce gazouillis fait partie de la discussion à propos de la thématique principale et la valeur totale de la cooccurrence inversée sera augmentée de +1. Si, par exemple, deux mots-clics de la requête en cours se retrouvent dans un même gazouillis, la valeur de la cooccurrence inversée ne sera pas augmentée de +2, car il n'y a pas de logique à considérer que ce gazouillis fait deux fois plus partie de la discussion à propos de la thématique qu'un gazouillis où ne se trouve qu'un seul mot-clic de la requête en cours.

La logique du calcul de cet indicateur s'exprime ainsi avec le vocabulaire SCOT :

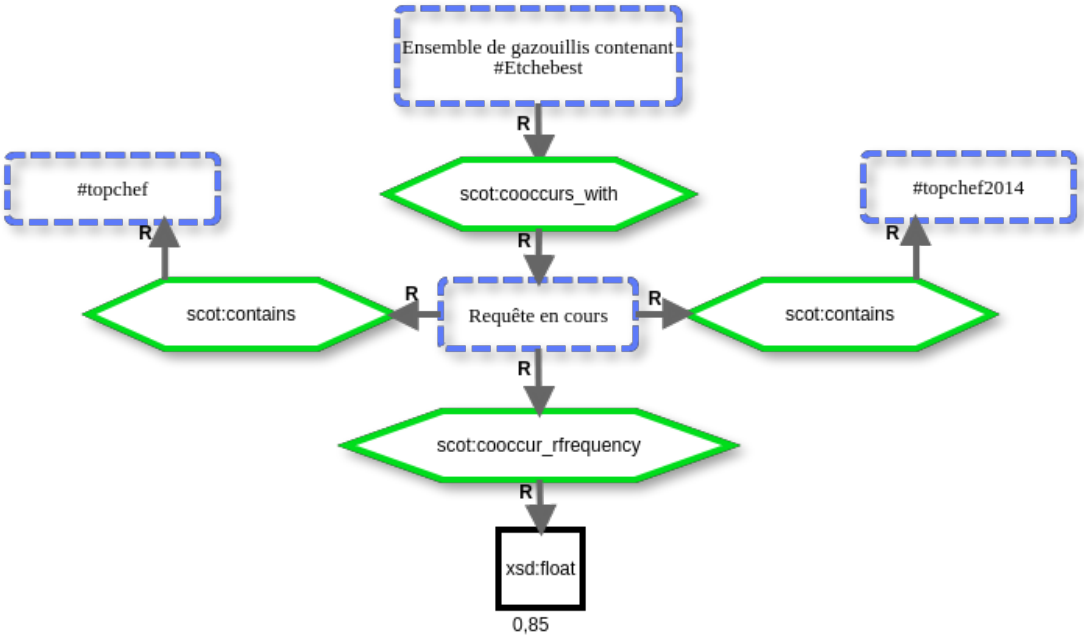


Figure 3.7: Calculs de similarité sémantique exprimés en langage SCOT

Qu'est-ce que nous apprend ce calcul d'indicateur sémantique de cooccurrence inversée? Dans le cas précis du mot-clic #Etchebest, il nous apprend que 85 % du temps, les utilisateurs qui gazouillent à propos de #Etchebest le font en référence directe à l'un des mots-clics de la thématique ou d'une sous-thématique de cette émission. Si la discussion à propos de #Etchebest porte 85 % du temps sur la thématique de l'émission, nous avançons que ce mot-clic représente en fait une sous-thématique de la thématique principale et que par conséquent la discussion à propos de #Etchebest est une *partieDe* la discussion à propos de l'émission « Top Chef », justifiant ainsi l'ajout dudit candidat à la requête en cours. Par contre, si l'échantillon avait été de 1000 et que la cooccurrence se serait maintenu à 85, cela aurait signifié que la discussion à propos de #Etchebest portait principalement sur autre chose que la discussion à propos de l'émission « Top Chef ». Cette hypothèse doit évidemment être vérifiée et fait l'objet d'une validation sur plusieurs corpus dans le Chapitre 6 :

Expérimentations du prototype et validation des résultats (p. 96)

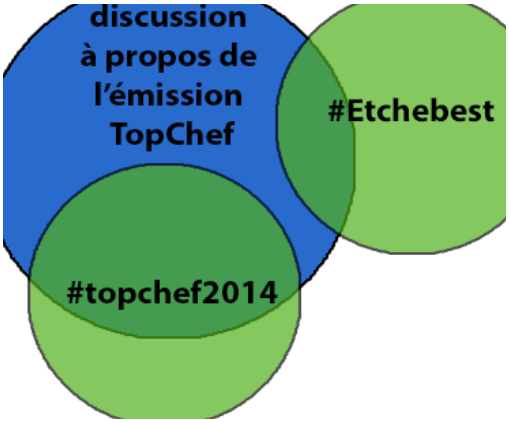
L’indicateur sémantique de cooccurrence inversée se présente sous la forme d'une variable numérique qui peut prendre une valeur entre 0 et 1 et non sous la forme d'un prédicat de type *owl:partOf* qui serait applicable entre un mot-clic candidat et les mots-clics de la requête en cours. La valeur de cet indicateur doit donc être interprétée afin de déterminer à quel moment elle est suffisamment forte pour conclure à une relation *partieDe.*

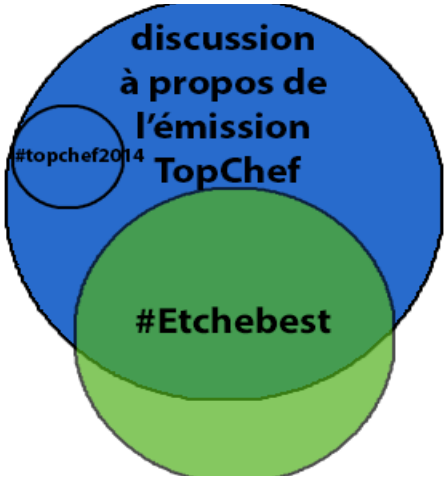
Dans l'analyse de nos échantillons, nous observons que plusieurs candidats possédant un indicateur sémantique de cooccurrence inversée se situant entre 0,3 et 0,5, faisant donc partie « moyennement » de l’ensemble de la discussion récoltée par la requête en cours, méritent parfois d'être sélectionnés. Toutefois, à ce degré d'appartenance, on retrouve aussi plusieurs mots-clics candidats ne méritant pas d'être sélectionnés. Il s'agit d'une zone grise où l'on retrouve à la fois des candidats pouvant être sélectionnés et d'autres ne devant pas l'être. À un tel degré d'incertitude, une décision prise par un humain chargé de départager les candidats à sélectionner pourrait être la bienvenue, mais voyons comment on peut augmenter ce degré de certitude au fil des itérations.

Dans notre exemple du mot-clic candidat #Etchebest et de la requête en cours composée de #topchef et de #topchef2014, nous indiquions que l'indicateur sémantique de cooccurrence inversée de #Etchebest était de 0,85. Toutefois, en début de cueillette, lorsque la requête en cours n’était composée que de #topchef, l'indicateur sémantique de cooccurrence inversée de #Etchebest était de 0,45. Ce candidat n'aurait pas été sélectionné puisque, en deçà de 0,5, son degré d'appartenance aurait été jugé trop faible. Par contre, au fil des itérations précédentes le candidat #topchef2014 a obtenu un indicateur sémantique de cooccurrence inversée suffisamment élevé pour être sélectionné dans la requête en cours. À l’itération qui a

suivi, la requête en cours n'était plus seulement composée de #topchef, mais aussi de #topchef2014. L'indicateur sémantique de cooccurrence inversée calculé avec ces deux mots-clics plutôt que seulement avec #topchef a fait en sorte que le candidat #Etchebest passe de la zone d'incertitude où l'indicateur sémantique de cooccurrence inversée se situait entre 0,3 et 0,5 à la zone où le degré d'appartenance est suffisamment élevé pour lui attribuer une relation *partOf* et procéder à son ajout à la requête en cours. Le tableau qui suit illustre ces différentes itérations menant à l'inclusion du mot-clic candidat #Etchebest à la requête en cours.

Tableau 3.5: Itérations menant à l'inclusion d'un mot-clic à la requête en cours

	**Itération #1 :** **Requête en cours = #topchef** L'indicateur sémantique de cooccurrence inversée est suffisamment élevé pour attribuer une relation *partieDe* à #topchef2014 et l'inclure dans la requête en cours, mais pas #Etchebest.

	**Itération #2 :** **Requête en cours = #topchef + topchef2014** L'indicateur sémantique de cooccurrence inversée est suffisamment élevé pour attribuer une relation *partieDe* à #Etchebest et l'inclure dans la requête en cours.
	**Itération #3 :** **Requête en cours = #topchef + topchef2014 + #Etchebest** La requête en cours étant maintenant composée de #topchef, topchef2014 et #Etchebest, les calculs d'indicateurs sémantiques de cooccurrence inversée pour les autres candidats se feront désormais avec ces 3 mots-clics.

Nous avons testé l'efficacité des calculs d'indicateurs sémantiques de cooccurrence inversée en reprenant les mots-clics du corpus de « Top Chef » (p.55) auxquels nous avions attribué manuellement des prédicats RDF et en comparant la valeur de leur indicateur sémantique de cooccurrence inversée.

Tableau 3.6: Corrélation entre prédicats RDF et indicateurs sémantiques de cooccurrence inversée

Mot-clic du champ lexical	**Prédicat RDF**	**Indicateur sémantique de cooccurrence inversée** (*scot:cooccur_rfrequency*)
topchef2014	*owl :partOf*	0,6408
Etchebest	*owl :partOf*	0,8452
M6	-	0,2343
Steven	*owl :partOf*	0,7345
Bachelor	-	0,0319
GameOfThrones	-	0,0382
alexis	*owl :partOf*	0,5568
Telecheck	-	0,2672
Noemie	*owl :partOf*	0,5426
MasterChef	-	0,1613

On observe dans ce tableau que tous les mots-clics du champ lexical auxquels on a attribué le prédicat *owl :partOf* ont un indicateur sémantique de cooccurrence inversée supérieur à 50 % et que tous les mots-clics auxquels nous n'avons pas attribué le prédicat *owl :partOf* ont des indicateurs inférieurs à 50 %.

Cette corrélation entre un indicateur sémantique de cooccurrence inversée au-delà de 0,5 et une présence d'un relation *partieDe* a été vérifiée et validée empiriquement sur plusieurs corpus. Les tableaux faisant état de cette vérification peuvent être consultés à l'Annexe C : Comparatif entre les décisions de l'algorithme et les classifications des évaluateurs sur les relations « partie de ». Le Chapitre 6 : Expérimentations du prototype et validation des résultats (p.96) propose quant à lui une analyse détaillée de cette corrélation entre l'indicateur sémantique de cooccurrence inversée et la relation *partieDe*.

Ajoutons que, dans ce tableau, une valeur de l'indicateur en deçà de 50 % est peu significative du point de vue sémantique, car on ne peut en conclure aucun prédicat.

En effet, un candidat pourrait être un mot-clic de la même catégorie comme l'émission #Masterchef qui n'a pas de lien de type *partie-tout* ou encore n'avoir aucun lien sémantique comme #Telecheck. L'indicateur sémantique de cooccurrence inversée est donc utilisé comme un calcul produisant une valeur permettant de tester si le candidat a une relation de type *partieDe* ou non. L'objectif étant de découvrir les candidats ayant ce type de relation et de rejeter les autres, il y a peu d'intérêt à développer une logique et une méthode pour attribuer des relations sémantiques aux candidats rejetés.

En résumé, l'indicateur sémantique de cooccurrence permet d'identifier les mots-clics présentant une forte cooccurrence afin de les sélectionner en tant que candidats, alors que l'indicateur sémantique de cooccurrence inversée permet de vérifier si une relation de type *partie-tout* existe afin d'ajouter le candidat à la requête en cours.

3.4 Relations sémantiques et temporalité

Il circule au sein de la twittosphère des milliards de messages hebdomadairement (Twitter Inc., 2015). L'actualité, le sport, la vie des gens célèbres, les nouveautés technologiques, les recettes de sauces spaghettis, les enjeux sociaux, les activités culturelles, tout y est discuté. Dans cet univers où les thématiques émergent, se dissolvent, reviennent au gré des saisons ou perdurent dans le temps, les mots, et particulièrement les mots-clics pour les décrire sont en constante évolution.

Ainsi, dans le champ lexical de la thématique « corruption », on pourrait voir apparaitre un parti politique mis sous examen à cause de financement illégal, puis ce parti pourrait en disparaitre suite à de nouvelles informations. Il en va de même pour un joueur de hockey vedette comme P.K. Subban qui est passé du champ lexical des Canadiens de Montréal à celui des Predators de Nashville lors de son échange ou encore du mot-clic #impôt qui apparait en cooccurrence avec #polQC à tous les débuts de printemps.

Or, même s’il a été vérifié que les folksonomies tendaient naturellement vers une stabilité sur le plan des mots-clés décrivant une ressource (voir De la nature des folksonomies p.33), la nature particulière de Twitter fait en sorte que pour plusieurs thématiques, ce mouvement vers une stabilité des termes est sans cesse perturbé par le flot de nouvelles informations en circulation dans l'espace public à propos de cette thématique. Dans cette grande twittosphère, théâtre de nombreux jeux de langage, au sens wittgensteinien (Hintikka et Hintikka, 1991), les définitions ostensives d'un terme n'existent pas. C'est plutôt leurs usages et leurs contextes d'utilisation qui leur donnent un sens, qui relient le langage aux actions et à la vie humaine. Le sens des mots et leur usage sont en mouvance perpétuelle et Twitter semble être un endroit idéal pour observer cette réalité.

Cette réalité, aussi riche soit-elle, ajoute un niveau de complexité à nos investigations, car il faut tenir compte que les folksonomies, les champs lexicaux, les relations sémantiques ainsi que les calculs d'indicateurs sémantiques sont l'expression de réalités en mouvement et ne reflètent que des vérités partielles et temporaires. Afin de s’ajuster à cette réalité en mouvement, les folksonomies, les champs lexicaux ainsi que les calculs de similarités sémantiques doivent être constamment réévalués, et ce, plusieurs fois par heure. On s’assure ainsi que les mots-clics de la requête en cours reflètent en tout temps les relations sémantiques à l’intérieur d’une thématique données. Une attention particulière est portée à ce défi dans l’élaboration de notre prototype afin d’ajuster nos inférences plusieurs fois par heure.

CHAPITRE 4 :
ALGORITHMES

Ce chapitre présente la mise en œuvre de trois algorithmes correspondant aux trois concepts principaux élaborés dans le cadre théorique soit la création de folksonomies par le regroupement de mots-clics, l'émergence de champs lexicaux composés de mots-clics affichant une forte cooccurrence et l'attribution de relations sémantiques à ces derniers à partir du calcul de l'indicateur sémantique de cooccurrence inversée.

Dans un premier temps, nous présentons une articulation des algorithmes issus de nos trois concepts. Ensuite, pour chacun des trois concepts, nous décrivons comment les algorithmes utilisent le raisonnement inductif appliqué à la fouille de texte et nous décrivons l'opérationnalisation de ces algorithmes à l'aide de diagrammes d'activités.

4.1 Interactions entre les concepts du cadre théorique dans le processus de sélection de mots-clics

L'objectif visé étant la sélection de mots-clics représentant des sous-thématiques, nous allons d'abord illustrer par un schéma simple comment chacun des trois concepts participe à ce processus de sélection.

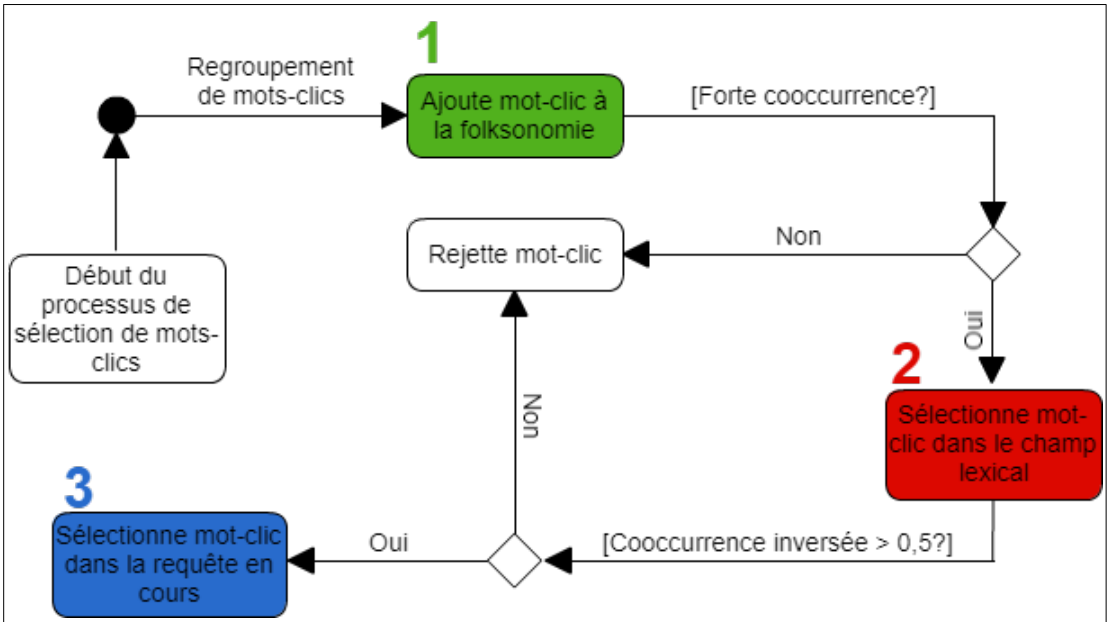


Figure 4.1: Diagramme d'activités du processus de sélection des mots-clics

Le processus de sélection de mots-clics illustré ci-dessus représente un cycle qui doit être répété 4 fois par heure pour tenir compte de l'évolution d'une thématique dans le temps. Le premier algorithme effectue un regroupement des mots-clics contenus dans les gazouillis de la base de données afin de construire une folksonomie. Cet algorithme se charge également de classer les mots-clics par ordre décroissant en fonction de leur cooccurrence dans le corpus.

Le deuxième algorithme s'appuie sur la loi de Zipf-Mandelbrot et sélectionne les mots-clics présentant une forte cooccurrence afin de former un champ lexical et rejette par conséquent les mots-clics ayant une faible cooccurrence.

Les mots-clics du champ lexical deviennent ensuite des « candidats » pour lesquels le troisième algorithme calcule pour chacun un indicateur sémantique de cooccurrence inversée. Si celui-ci est supérieur à 0,5, le mot-clic est sélectionné et ajouté à la requête en cours. Cette mise à jour de la requête en cours est communiquée à la

Streaming API de laquelle on recevra tout nouveau gazouillis contenant l'un des mots-clics de la requête en cours. Le processus de sélection est alors terminé et redémarrera son cycle 15 minutes plus tard.

4.2 Définition des algorithmes

Chacun des algorithmes à l'œuvre dans ces étapes sera maintenant décrit en détail. Pour chacun des trois algorithmes, on décrit le raisonnement à l'œuvre, ses processus d'inférences suivis d'un diagramme d'activités dans lequel nous mettrons l'accent sur les traitements.

4.2.1 Algorithme de création de folksonomies

Le raisonnement inductif appliqué à la création de folksonomies est le suivant :

1. Une folksonomie est le résultat d'un regroupement de mot-clics cooccurrents dans un ensemble de gazouillis.
2. Tous les gazouillis contiennent des mots-clics.
3. Tous les mots-clics font partie de la folksonomie.

Il s'agit d'un raisonnement simple qui permet d'inclure tout mot-clic compris dans un gazouillis dans une folksonomie pour une thématique et un intervalle de temps donné. La création d'une folksonomie bénéficie également d'une information supplémentaire liée à la cooccurrence des mots-clics qui la composent. C'est donc à cette étape que nous procèdons au regroupement des mots-clics et à une classification en ordre décroissant, du mot-clic ayant le plus forte cooccurrence à celui qui a la plus faible. Les mots-clics contenus dans un gazouillis étant sauvegardés dans une table leur étant dédiée, ce traitement s'effectue rapidement par une requête SQL (voir Annexe E : Requêtes SQL p.186) et permet dans un premier temps d'observer la formation d'une folksonomie autour d'une thématique pour un intervalle de temps donné.

Reprenons maintenant le Diagramme d'activités du processus de sélection des mots-clics (p.70) afin de détailler le fonctionnement de ce premier algorithme.

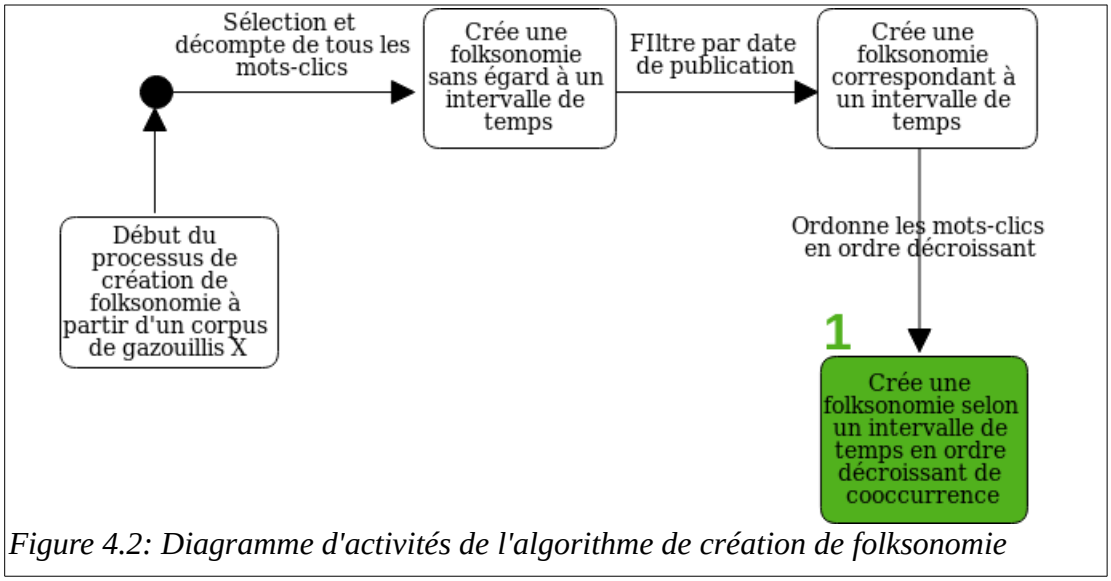


Figure 4.2: Diagramme d'activités de l'algorithme de création de folksonomie

4.2.2 Algorithme de création de champs lexicaux

Tel qu'explicité dans la section sur les Champs lexicaux émergeant d'une folksonomie (p.42), le résultat du regroupement des processus cognitifs de catégorisation accomplis par les utilisateurs (folksonomie) fait émerger un champ lexical formé des mots-clics ayant une forte cooccurrence avec le ou les mots-clics thématiques.

Lors de tests sur des échantillons de plusieurs dizaines de milliers de gazouillis, la liste de mots-clics faisant partie de la folksonomie croît très rapidement et le temps de calcul des indicateurs sémantiques devient très long et est sujet au dépassement des limites permises par Twitter sur son API. De plus, nous avons observé que l'ajout de mots-clics qui sont au bas de la liste – ayant donc une cooccurrence faible - n'a que

peu d'influence sur la quantité de gazouillis récoltés. D'ailleurs, nous observons que ces mots-clics sont la plus part du temps considérablement éloignés du champ lexical. Pour ces raisons, nous avons optimisé la liste de candidats en appliquant le modèle de distribution de Zipf-Mandelbrot qui permet de ne garder que les mots-clics ayant une forte cooccurrence. À cette étape on tente de calculer combien de mots-clics forment le champ lexical à partir du nombre d'occurrences du mot-clic le plus fréquent.

Le raisonnement inductif appliqué à la création d'un champ lexical est le suivant :

1. Un champ lexical est formé de mots-clics ayant une forte cooccurrence (P).
2. La loi de Zipf-Mandelbrot permet de sélectionner les mots-clics ayant une forte cooccurrence.
3. Les mots-clics sélectionnés font partie du champs lexical.

L'ensemble de mots-clics du champ lexical est un sous-ensemble extrait de la folksonomie. Nous avons déjà vu que la propriété P (possède une forte cooccurrence) est obtenue en utilisant la loi de Zipf-Mandelbrot (p.47), mais rappelons tout de même brièvement ses principes. La fréquence d'un terme dans un corpus est liée à son rang. Par exemple, le deuxième terme apparait environ deux fois moins souvent que le premier et le dixième apparait 10 fois moins souvent. Nous avons également vu que tant que la proposition ln(occurrence) > ln(rang) est vraie nous considérons que la cooccurrence est forte (voir Jonction des courbes ln(rang) et ln(occurrence) à la page 51). On peut également simplifier cette proposition en disant que tant que l'occurrence d'un mot-clic est plus élevée que son rang la cooccurrence est forte. C'est d'ailleurs cette proposition simplifiée que nous avons retenue pour nos calculs computationnels.

Du moment que l'on connait la valeur de l'occurrence du mot-clic apparaissant le plus souvent dans un corpus, on peut calculer la valeur de l'occurrence du second mot-clic

ayant la plus forte cooccurrence en la divisant par son rang (2) et ainsi de suite jusqu'à ce que la valeur du rang soit égale ou supérieure à la valeur de l'occurrence. Cette logique est illustrée dans cette simulation où le mot-clic le plus utilisé apparait 200 fois:

Tableau 4.1: Application de la loi de Zipf pour déterminer la forte occurrence

Rang	**Occurrence**
1	200
2	100
3	67
4	50
5	40
6	33
7	29
8	25
9	22
10	20
11	18
12	17
13	15
14	14

Les nombres en rouge indiquent le moment où notre condition n'est plus respectée. On déduirait donc pour ce corpus que seuls les 13 premiers mots-clics seraient retenus pour former le champ lexical. Pour effectuer ce calcul, on doit bien sûr disposer de la valeur de l’occurrence du mot-clic thématique laquelle s'obtient par une requête SQL (voir Annexe E : Requêtes SQL p.186).

Reprenons une autre fois le Diagramme d'activités du processus de sélection des mots-clics (p.70) afin de développer cette deuxième activité du processus de sélection

de mots-clics soit la sélection de mots-clics ayant une forte cooccurrence.

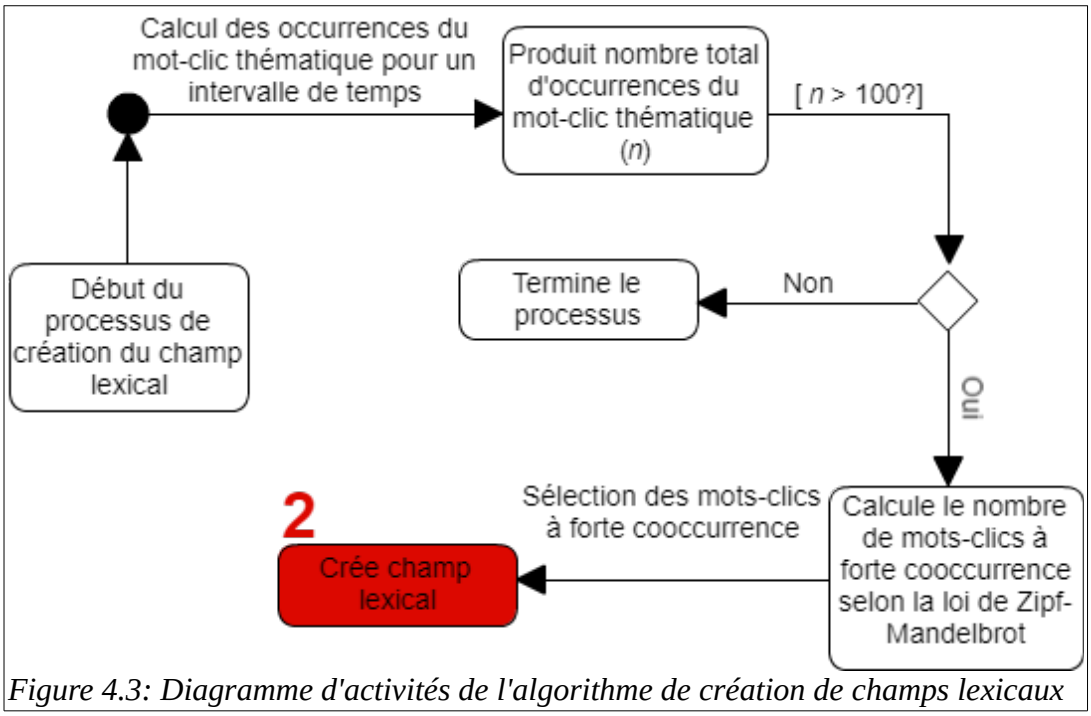


Figure 4.3: Diagramme d'activités de l'algorithme de création de champs lexicaux

On remarque dans ce diagramme que l'algorithme évalue une condition d'arrêt. En effet, lorsque le nombre d'occurrences du mot-clic est inférieur à 100, l'algorithme s'arrête et donne un message d'arrêt du processus. Il est assez rare que moins de 100 gazouillis soient publiés à propos d'une thématique donnée, mais lorsque c'est le cas, il vaut mieux arrêter le processus, car les inférences risquent de produire des résultats erronés faute de données suffisantes. En effet, comme pour tout algorithme tentant de faire des inférences inductives à partir d'un corpus de texte, en deçà d'un minimum de textes, la précision en souffre. Il existe deux raisons distinctes expliquant qu'un nombre minimum de gazouillis ne soit pas atteint. La première et la plus courante est que la discussion faisant usage du mot-clic thématique évolue au fil de la journée ou au fil de la semaine et connait parfois des baisses de trafic. Un cas de fluctuation quotidienne a par exemple été observé pour le mot-clic #polQc qui, n'étant pas une

thématique planétaire discutée sur plusieurs fuseaux horaires, voit le nombre de gazouillis à son sujet diminuer durant la nuit. On a également observé des diminutions, cette fois-ci hebdomadaires, pour des émissions de télévision. Pour ces deux cas de baisses de trafic durant certaines périodes du jour ou de la semaine, aucune modification des paramètres de l'algorithme n'est nécessaire. Même si le minimum de données requis n'est pas atteint, la cueillette des gazouillis du mot-clic thématique se poursuit et quand le trafic de gazouillis augmente suffisamment, l'algorithme reprend ses inférences automatiquement. La deuxième raison expliquant que le minimum de 100 gazouillis/heure ne soit pas atteint est tout simplement que la thématique choisie est très marginale. Nous changeons alors les paramètres de l'algorithme pour que ses inférences se basent sur un corpus des 24 dernières heures. Ceci diminue bien sûr la réactivité de l'algorithme, car l'émergence d'une nouvelle sous-thématique sera évaluée sur une plus longue période de temps. Toutefois, la qualité des inférences est maintenue.

En résumé, le raisonnement que l'algorithme induit à cette étape est qu'un mot-clic sera inclus dans le champ lexical en autant que la valeur du logarithme de son occurrence soit plus grande que celle du logarithme de son rang. En terme logique, on l'exprime comme suit : la cooccurrence est forte si et seulement si ln(occurrence) > ln(rang).

4.2.3 Algorithme de mise à jour de la requête en cours

Contrairement aux deux premiers algorithmes, celui-ci n'effectue pas ses inférences à partir de requêtes SQL, mais plutôt à partir de données recueillies de la *Search API* de Twitter. Comme données de départ, il reprend les mots-clics du champ lexical qu'il traite comme une liste de candidats dont les indicateurs sémantiques de cooccurrence inversée doivent être calculés. Optionnellement, si la cueillette s'effectue dans une langue en particulier, il utilise ce paramètre dans ses requêtes à la *Search* API.

Le raisonnement inductif appliqué à la mise à jour d’une requête en cours est le suivant :

1. Une requête en cours est formée par des mots-clics représentant une thématique ou une sous thématique.
2. Une forte cooccurrence inversée d'un mot-clic candidat (> 0.5) indique une relation partieDe avec la thématique ou une sous thématique.
3. Tous les candidats ayant une cooccurrence inversée supérieure à > 0.5 font *partieDe* la requête en cours.

Reprenons une dernière fois le Diagramme d'activités du processus de sélection des mots-clics, mais cette fois-ci en développant sur le calcul de cooccurrence inversée qui mène à la découverte de mots-clics représentant des sous thématiques et, ultimement, à la mise à jour de la requête en cours envoyée à la *Streaming* API.

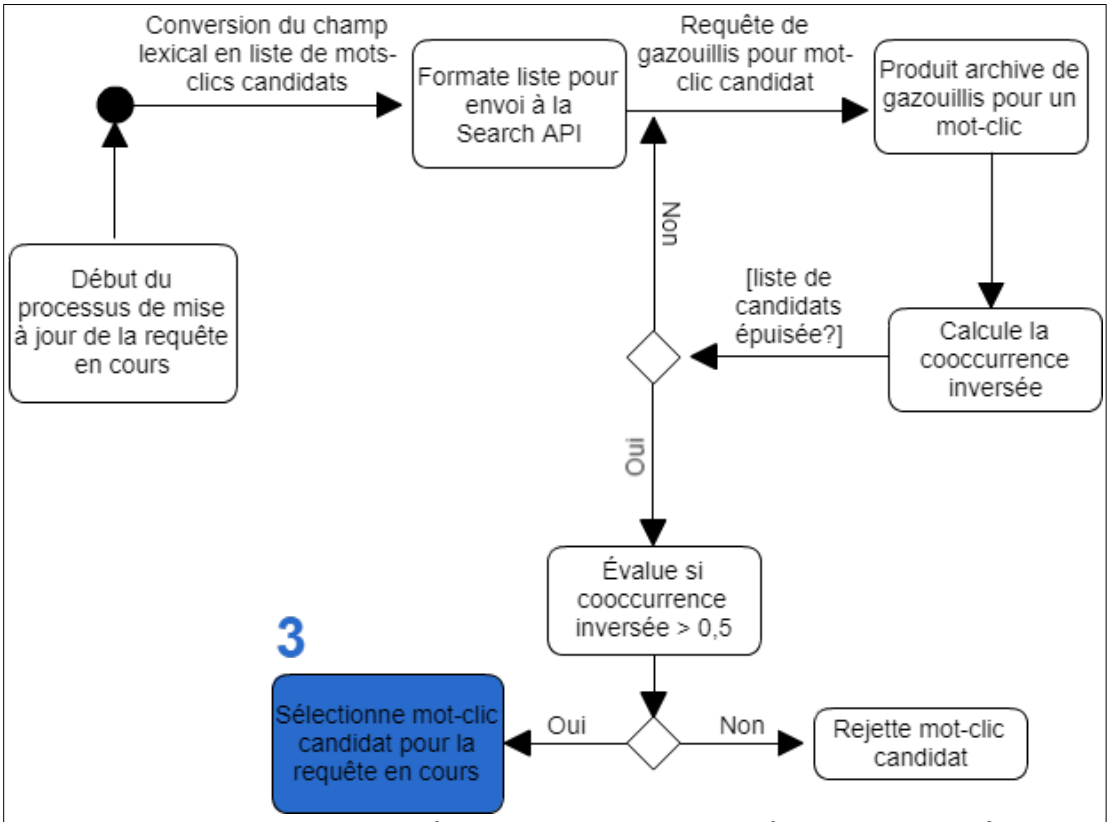


Figure 4.4: Diagramme d'activités de l'algorithme de mise à jour de la requête en cours

Le calcul des indicateurs sémantiques de cooccurrence inversée est effectué suite au regroupement des mots-clics (folksonomie) et à la sélection de ceux ayant la plus forte cooccurrence (champ lexical). Chacun des mots-clics du champ lexical est envoyé à la *Search* API afin de recueillir les gazouillis en faisant usage dans la dernière heure. Selon la popularité du mot-clic et la disponibilité des ressources informatiques de Twitter, un nombre de gazouillis entre 0 et 500 est retourné. Pour des thématiques plus pointues, comme le *data mining (#datamining)* ou ayant lieu dans des lieux où la population est plus petite, comme la thématique sur la politique québécoise (#polQc), le faible nombre de gazouillis rend parfois le calcul des

indicateurs sémantiques de cooccurrence inversée plus difficiles. Cette question sera discutée dans la section Limites (p.155).

Comme on peut le voir dans ce diagramme, une boucle est créée afin d'effectuer des calculs de cooccurrence inversée sur chacun des candidats de la liste. Durant l'activité « Calcule la cooccurrence inversée », chacun des gazouillis provenant de la *Search* API est ouvert et le champ contenant les mots-clics est converti en liste. Bien sûr, le mot-clic candidat est retiré de cette liste pour ne garder que les mots-clics cooccurrents avec ce dernier, sinon, tous les candidats déjà sélectionnés dans la requête en cours obtiendraient à tout coup des indicateurs sémantiques de cooccurrence inversée de 1. Cette liste est ensuite comparée à la liste des mots-clics de la requête en cours précédente et si un ou plusieurs mots-clics sont communs aux deux listes, c'est qu'il y a cooccurrence et l'algorithme additionnera une valeur de +1 à une variable (*inverse_cooccurrence*). À la fin de cette activité, la valeur numérique stockée dans cette variable sera divisée par le nombre total de gazouillis recueillis pour ce mot-clic candidat et un ratio sera obtenu. Ce ratio, situé entre 0 et 1, nous donne la valeur de l'indicateur sémantique de cooccurrence inversée.

CHAPITRE 5 :
MÉTHODOLOGIE

Notre méthodologie poursuit deux objectifs soit celui de construire un prototype mettant en œuvre les algorithmes issus de notre cadre théorique et celui de mettre en place un cadre expérimental capable de valider les résultats produits par le prototype et sa capacité à répondre à notre problématique.

Nous présentons d'abord notre approche pour le forage de données, pour ensuite décrire le développement de notre prototype et finalement présenter notre cadre expérimental.

5.1 Forage de données

Les méthodes s'intéressant à l'extraction de connaissances sur de vastes corpus de textes ont connu un important essor dans les dernières années et le vocable employé pour les décrire reflète cette évolution rapide provenant de diverses disciplines. Des termes comme l'exploration de données, la fouille de données, le forage de données, le *data mining*, le *knowledge discovery* ou encore la fouille de texte sont tour à tour utilisés pour nommer des concepts aux différences subtiles partageant plusieurs processus semblables (Cohen et Hunter, 2008).

Dans la littérature concernant les méthodologies de ces concepts voisins, on retrouve plusieurs efforts de systématisation des processus (Vidhya et Aghila, 2010; Oussalah et al., 2013; Mathiak et Eckstein, 2004). Selon les disciplines, le type de données à traiter ainsi que les connaissances que l'on souhaite extraire les processus varient et s'adaptent aux objets de recherche. Par exemple, dans les disciplines bio-médicales, la visualisation constitue souvent une étape incontournable alors que pour d'autres disciplines, comme la linguistique, elle s'avère plutôt optionnelle.

Pour répondre à notre problématique et mettre en œuvre les algorithmes issus de notre

cadre théorique, nous nous sommes inspiré de la méthode de *text-mining* de Vidhya et Aghila (2010) et en avons dégagé 4 étapes :

- Cueillette d'informations (Document retrieval)
- Prétraitement (pre-process parsing, stemming)
- Extraction d'informations (information retrieval)
- Extraction de connaissances (knowledge retrieval)

Ces étapes que nous détaillons ici ont fait l'objet de divers tests visant à valider la méthodologie, les résultats des inférences et la cohérence des corpus recueillis vis-à-vis de la thématique.

5.1.1 Cueillette d'informations

La cueillette d'informations est réalisée à partir des deux API offertes par Twitter : la *Streaming* API et la *Search* API. La *Streaming* API permet de recueillir en temps réel tous les gazouillis publiés contenant un ou plusieurs mots-clics de la requête en cours alors que la *Search* API permet un accès partiel aux archives de gazouillis.

La *Streaming* API est utilisée pour recueillir tous les gazouillis contenant les mots-clics de la requête en cours. Les gazouillis recueillis à partir de cette API sont stockés dans une base de données relationnelle (MySQL).

La *Search API* de Twitter est quant à elle utilisée pour faire des fouilles dans les archives de gazouillis. Ces gazouillis et leurs métadonnées servent exclusivement aux calculs et ne sont pas sauvegardées; seuls les résultats des inférences le sont.

5.1.2 Prétraitement

Le prétraitement des données fait référence au nettoyage de ces dernières. Il arrive souvent que suite à une cueillette d'informations certaines données soient incomplètes

ou même erronées. Le prétraitement est alors nécessaire pour assurer une validité et une cohérence au corpus avant les étapes de forage, d'indexation ou d'analyse (Nahm et Mooney, 2002).

Dans le cas des données provenant des API de Twitter, le prétraitement est assez simple, car il consiste à analyser (*parse*) chacun des gazouillis et ses métadonnées afin de les enregistrer correctement dans la base de données. Les données sont déjà structurées et aucun nettoyage n'est requis mis à part la mise en minuscule de tous les mots-clics pour faciliter leur regroupement. La *Streaming* API renvoie des gazouillis en format JSON (JavaScript Object Notation) qui sont facilement intégrable dans une base de données.

5.1.3 Extraction d’information

L’extraction d’information, typiquement d’une base de données comme dans notre cas, permet d’identifier des documents ou des parties de documents pertinents pour un objet de recherche donné (Manning *et al.*, 2008). Pour nous, il s’agit d’extraire les mots-clics d’un corpus ainsi que l’information relative à leur cooccurrence.

C'est donc à cette étape que l'on procède au regroupement de mots-clics pour former une folksonomie. Les mots-clics contenus dans un gazouillis étant sauvegardés dans une table leur étant dédiée, ce traitement s'effectue rapidement par une requête SQL et permet d'observer la formation d'une folksonomie autour d'une thématique et d’ordonner ces mots-clics en fonction de leur cooccurrence.

5.1.4 Extraction de connaissances

Dans la base de données, beaucoup d'informations sont recueillies, extraites et analysées pour produire essentiellement les deux types de connaissances suivants : un champ lexical autour d'une thématique et un indicateur sémantique de cooccurrence cooccurrence inversée permettant d’évaluer la présence d’une relation *partieDe*.

Autant les fondements théoriques associés à ces connaissances que les algorithmes qui permettent de les produire ont déjà été explicités. On récapitulera donc dans cette section en illustrant les connaissances extraites par un tableau représentant un champ lexical accompagné de l'évaluation de la relation *partieDe*. Ce tableau est issu de la base de données MySQL et illustre un des résultats des inférences produites toutes les 15 minutes. Il est issu d'une cueillette de la discussion à propos de la thématique des mégadonnées.

Tableau 5.1: Exemple d'extraction de connaissance

Mot-clic thématique : #BigData				
1	2	3	4	5
Candidat	**Cooccurrence inversée (brute)**	**Gazouillis recueillis pour le candidat**	**Indice de cooccurrence inversée**	**PartieDe?**
YARN	0	294	0.0000	0
Marketing	1	456	0.0022	0
InternetOfThings	32	271	0.1181	0
datascience	45	105	0.4286	0
datamining	61	77	0.7922	1
cloud	24	413	0.0581	0
venturecapital	14	101	0.1386	0
abdsc	1	4	0.2500	0
TC	4	256	0.0156	0
IOT	41	310	0.1323	0
Hadoop	53	91	0.5824	1
FuturePerfectVentures	8	11	0.7273	1

La colonne 1 contient les mots-clics candidats et constitue la première connaissance extraite par ce système soit le champ lexical. La colonne 2 présente le nombre de cooccurrences des mots-clics de la requête en cours (#BigData, #datamining, #Hadoop et #FuturePerfectVentures) avec le mot-clic candidat. Vient ensuite le

nombre de gazouillis récupérés par la *Search* API (colonne 3), l'indicateur sémantique de cooccurrence inversée (colonne 4) puis une valeur booléenne qui lorsque vraie signifie que le candidat est une *partieDe* et qu'il sera ajouté à la requête en cours (colonne 5).

5.2 Développement du prototype

5.2.1 Cas d'utilisation

Notre cas d'utilisation est relativement simple, la complexité se trouvant plutôt dans la problématique de la découverte de mots-clics représentant des sous-thématiques. Typiquement, un utilisateur choisit un ou plusieurs mots-clics thématique qu'il insère dans la base de données. Cette insertion peut être faite par le biais de l'interface graphique de *phpMyAdmin*, par un simple script d'insertion SQL ou à l'aide de l'interface graphique que nous présentons dans la prochaine section. Ensuite, la cueillette de gazouillis en provenance de la *Streaming* API démarre automatiquement et le processus de sélection de mots-clics est lancé à intervalles réguliers (p.70). La fin de la cueillette est décidée par l'utilisateur et ce dernier n'a qu'à supprimer le ou les mots-clics thématiques qu'il a inséré précédemment dans la base de données pour arrêter le processus. Il peut ensuite disposer de son corpus comme bon lui semble en l'exportant en CSV vers une autre application ou en conduisant ses analyses directement dans MySQL.

5.2.2 Interface graphique

Une interface graphique a été conçue pour permettre à un usager sans connaissance technique d'effectuer toutes les tâches nécessaires au suivi de discussions sur Twitter à l'aide de notre prototype.

Notre cas d'utilisation débute avec l'ajout d'un mot-clic thématique. Ce dernier est souvent connu de l'utilisateur, mais nous proposons un outil capable d'identifier le

mot-clic le plus souvent utilisé par un usager Twitter. Tel que décrit dans la section sur les mots-clics thématiques (p.29), le mot-clic le plus souvent utilisé par le leadeur d'une communauté est souvent représentatif d'une thématique.

Figure 5.1: Extraction d'un mot-clic thématique à partir d'un compte d'usager

Sans la figure qui précède, l'utilisateur [1] entre un compte d'usager Twitter, [2] le prototype récupère un échantillon de gazouillis de la *Search* API, puis finalement [3] le mot-clic le plus souvent utilisé par l'usager Twitter est affiché. Si l'utilisateur de notre prototype juge pertinent de choisir ce mot-clic comme mot-clic thématique, il démarre la cueillette en cliquant sur « OK ».

Dans les minutes qui suivront le démarrage de la cueillette, le prototype commencera à faire des inférences lui permettant d'extraire un champ lexical puis des relations sémantiques de type *partieDe*. L'utilisateur de notre prototype peut observer le résultat de ces inférences réalisées toutes les 15 minutes par le biais de l'interface sous un onglet nommé « Tableau de bord ».

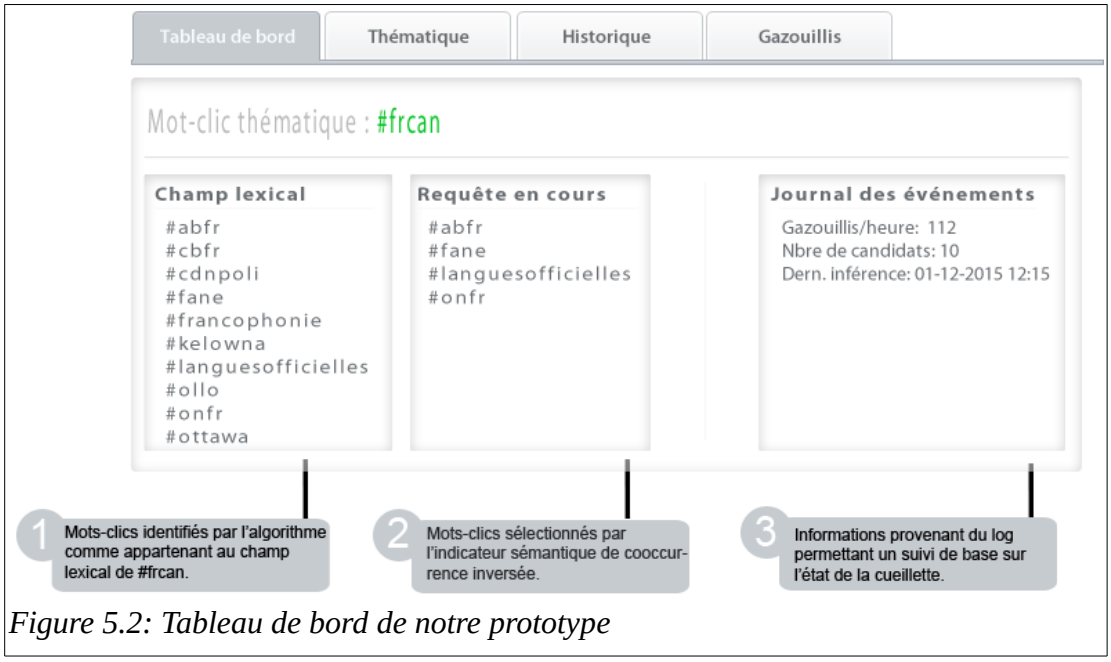
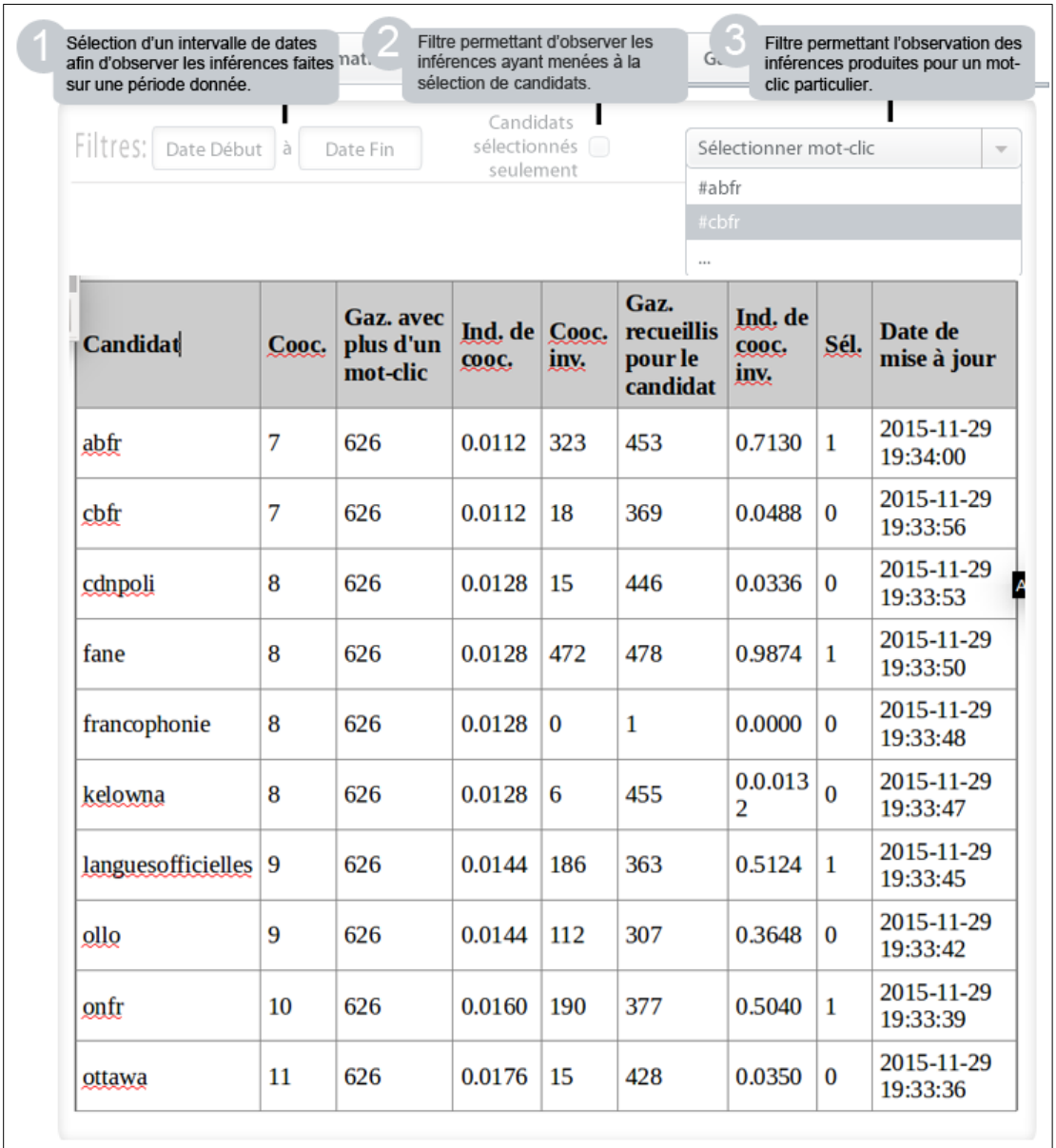


Figure 5.2: Tableau de bord de notre prototype

Le tableau de bord est utile pour [1] vérifier le dernier état du champ lexical [2] et de la requête en cours et [3] s'assurer du bon fonctionnement de notre prototype.

À des fins de contrôle de qualité, on aura souvent besoin de faire un retour dans le temps afin d'observer les inférences faites par l'algorithme et ses divers calculs d'indicateurs sémantiques de cooccurrence et de cooccurrence inversée. Pour ce faire, une partie de l'interface nous permet d'avoir accès à l'historique des inférences ayant menées à la construction du champ lexical et de la requête en cours.



Candidat	Cooc.	Gaz. avec plus d'un mot-clic	Ind. de cooc.	Cooc. inv.	Gaz. recueillis pour le candidat	Ind. de cooc. inv.	Sél.	Date de mise à jour
abfr	7	626	0.0112	323	453	0.7130	1	2015-11-29 19:34:00
cbfr	7	626	0.0112	18	369	0.0488	0	2015-11-29 19:33:56
cdnpoli	8	626	0.0128	15	446	0.0336	0	2015-11-29 19:33:53
fane	8	626	0.0128	472	478	0.9874	1	2015-11-29 19:33:50
francophonie	8	626	0.0128	0	1	0.0000	0	2015-11-29 19:33:48
kelowna	8	626	0.0128	6	455	0.0.0132	0	2015-11-29 19:33:47
languesofficielles	9	626	0.0144	186	363	0.5124	1	2015-11-29 19:33:45
ollo	9	626	0.0144	112	307	0.3648	0	2015-11-29 19:33:42
onfr	10	626	0.0160	190	377	0.5040	1	2015-11-29 19:33:39
ottawa	11	626	0.0176	15	428	0.0350	0	2015-11-29 19:33:36

Figure 5.3: Historique des inférences

Sous cet onglet l'utilisateur a donc accès à l'intégralité des inférences effectuées pour chacun des mots-clics candidats. Pour faciliter la navigation, il dispose de [1] quelques filtres permettant soit de sélectionner une période, [2] d'isoler les mots-clics candidats ayant été sélectionnés à un moment pour faire partie de la requête en cours

ou encore [3] d'obtenir toutes les inférences effectuées pour un mot-clic candidat particulier.

Le dernier élément d'interface propose une liste des gazouillis sauvegardés dans la base de données en provenance de la *Streaming* API. Elle donne à l'utilisateur un aperçu du résultat de sa cueillette au fur et à mesure qu'elle se déroule lui permettant de corriger le tir si la cueillette ne correspond pas à ses attentes. De plus, on a conçu une fonction d'exportation du corpus en CSV pour faciliter les analyses de contenus dans d'autres logiciels comme Excel, Calc ou encore Nvivo.

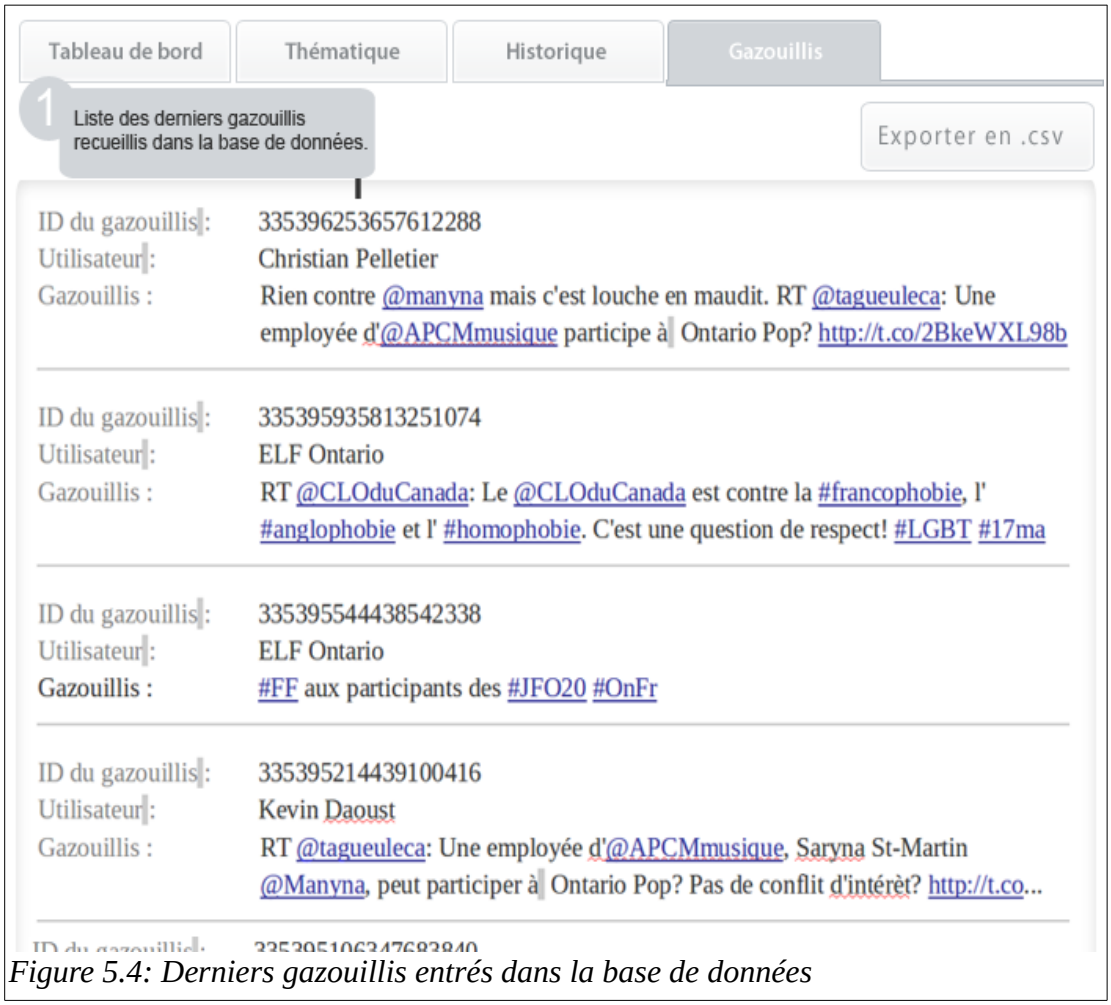


Figure 5.4: Derniers gazouillis entrés dans la base de données

5.2.3 Diagramme de flux

Notre diagramme de flux illustre le fonctionnement interne de notre prototype. Deux processus y fonctionnent en parallèle : le premier est un processus continu qui établit une connexion permanente avec la *Streaming* API de Twitter afin d'obtenir des gazouillis correspondant aux mots-clics de la requête en cours, alors que le second est un processus lancé à intervalles de 15 minutes qui est chargé de la découverte des mots-clics représentant des sous-thématiques. À ces deux processus correspondent deux modules soit un module de cueillette de gazouillis utilisant la bibliothèque PHP *140dev* et un module de sélection de mots-clics en Ruby développé dans le cadre de cette thèse afin d'implémenter nos trois algorithmes.

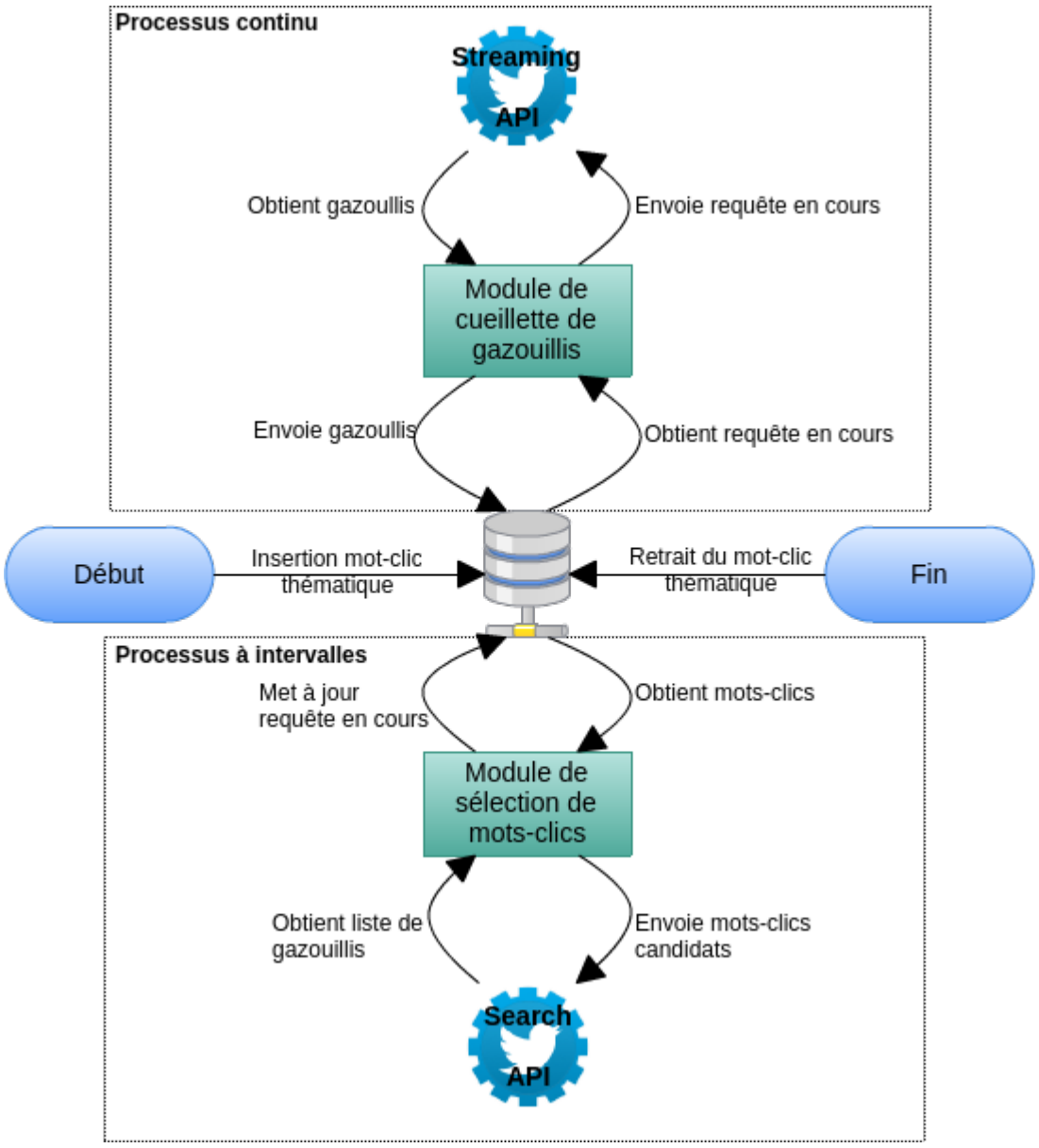


Figure 5.5: Diagramme de flux de notre prototype final

Tel que décrit dans notre cas d'utilisation, notre prototype se met en marche lorsqu'un ou plusieurs mots-clics thématiques sont insérés dans la base de données. Ces mots-clics constituent la première requête en cours. Suite à cet ajout, le module de

cueillette de gazouillis communique la requête en cours à la *Streaming* API et il établit une connexion permanente avec la *Streaming* API de Twitter. De cette API, il obtient les nouveaux gazouillis contenant un mot-clic de la requête en cours au fur et à mesure que les utilisateurs les publient. Ce processus continu, une fois démarré, s'effectue en boucle durant toute la durée de la cueillette.

Le deuxième module qui est en charge de la sélection de mots-clics opère en parallèle à intervalles réguliers. Dans ce module sont implémentés les trois algorithmes que nous avons présentés. Ce module obtient donc en entrée des mots-clics de gazouillis récoltés par le module de cueillette, crée ensuite une folksonomie puis un champ lexical et, finalement, le troisième algorithme envoie chacun des mots-clics du champ lexical à la *Search* API et obtient une liste de gazouillis dont les mots-clics seront utilisés pour les calculs sémantiques de cooccurrence inversée. Ce processus se termine par la mise à jour de la requête en cours avec l'ajout potentiel de mots-clics candidats représentant des sous-thématiques. Tel que mentionné dans la section Relations sémantiques et temporalité (p.67), les thématiques ainsi que les mots-clics s'y référant évoluent rapidement sur Twitter. Pour tenir compte de ce dynamisme, les calculs d'indicateurs sémantiques sont effectués toutes les 15 minutes.

Une fois que la mise à jour de la requête en cours est effectuée et sauvegardée dans la base de données, le module de cueillette de gazouillis les récupère et obtient des gazouillis y correspondant en provenance de la *Streaming* API. Ces deux processus interagiront ainsi pour toute la durée d'une cueillette qui peut parfois s'échelonner sur des mois. Le retrait du mot-clic thématique signale la fin du processus de cueillette et le processus de sélection de mots-clics s'arrêtera également.

5.2.4 Technologies utilisées

La phase de développement fut effectuée lors d'un stage Mitacs chez Seevibes, une entreprise en démarrage. L'entreprise partageait la même problématique que nous en

ce qui a trait à l'automatisation de la découverte de mots-clics. L'entreprise développant tous ses systèmes avec le langage de programmation Ruby, c'est donc avec ce dernier que le prototype a été développé. Toutefois, puisque la bibliothèque PHP *140dev* répondait à nos besoins en termes de cueillette de gazouillis en provenance de la *Streaming* API, nous l'avons conservée pour le prototype en la modifiant légèrement.

Voici notre diagramme de déploiement de notre prototype :

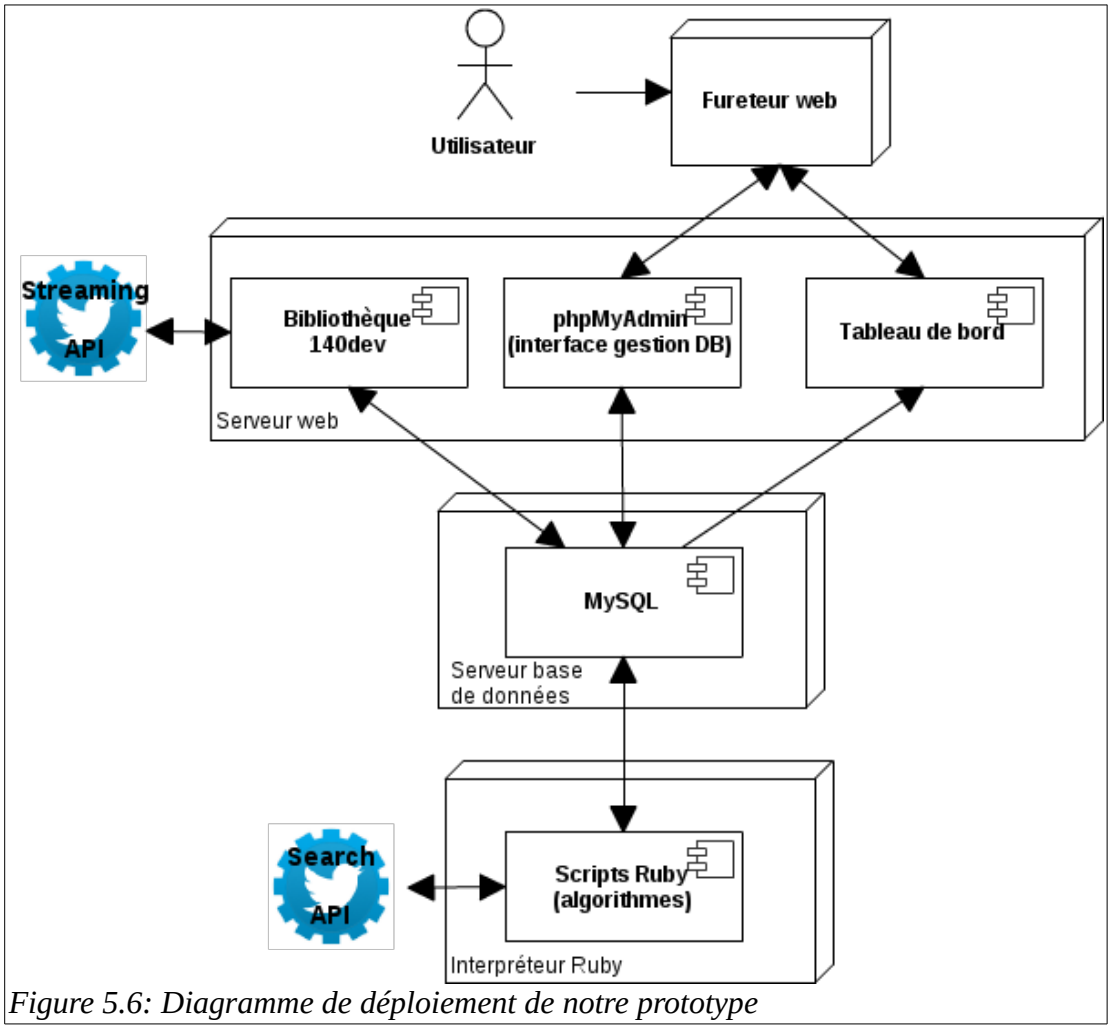


Figure 5.6: Diagramme de déploiement de notre prototype

Tel que nous l'illustrons dans ce diagramme de déploiement, notre modèle de prototypage comprend un serveur web dans lequel on retrouve la bibliothèque *140dev* pour la collecte de données en temps réel en provenance de la *Streaming* API, une interface de gestion MySQL reposant sur *phpMyAdmin* et un tableau de bord développé en PHP permettant aux usagers de faire un monitorage en ligne de la cueillette, des inférences des algorithmes et des gazouillis recueillis. En deuxième lieu, on retrouve un serveur de base de données MySQL qui recueille des données et les rend disponibles aux différentes composantes de notre prototype, puis finalement, notre code Ruby qui communique avec la *Search* API et MySQL afin d'exécuter les algorithmes.

5.3 Cadre expérimental

Deux cadres expérimentaux ont été mis en place, le premier répondant au besoin de l'entreprise Seevibes lors du stage et le deuxième, plus rigoureux, pour les besoins de la présente thèse.

Seevibes se spécialisant dans la télévision sociale, le premier cadre expérimental avait pour finalité de démontrer que l'automatisation de la découverte de mots-clics pour des discussions à propos d'émissions de télévision était possible. Notre découverte de mots-clics fut comparée aux mots-clics trouvés manuellement par le personnel de Seevibes et fut jugée suffisamment satisfaisante pour que l'entreprise souhaite breveter notre prototype.

Notre deuxième cadre expérimental a visé une évaluation plus complète des inférences produites pour la création de champs lexicaux et les calculs d'indicateurs sémantiques de cooccurrence inversée et s'est penché sur des méthodes s'assurant de la cohérence du corpus en regard à une thématique de départ. La méthodologie de validation a été appliquée sur 10 expérimentations que nous avons menées et l'évaluation fut effectuée par deux évaluateurs sans lien avec cette recherche.

Ces 10 expérimentations ont conduit à la production de jeux de données comportant chacun un corpus de gazouillis, des folksonomies, des champs lexicaux et des requêtes en cours.

Pour chacun des 10 jeux de données produit par notre prototype, trois niveaux de validation sont proposés soit [1] la création de champs lexicaux, [2] les calculs d'indicateurs sémantiques de cooccurrence inversée et [3] la cohérence du corpus. La méthodologie de validation a été appliquée de la même manière pour chacune des 10 expérimentations que nous avons menées. Cette validation a été effectuée par deux évaluateurs n'ayant aucun lien avec ce projet de recherche qui ont reçu une formation et un document contenant les directives à suivre pour réaliser la validation de nos résultats (Annexe A : Directives aux évaluateurs (p.165). Les mots-clics thématiques utilisés pour réaliser la cueillette sont les suivants :

Tableau 5.2: Mots-clics thématiques utilisés pour la cueillette d'échantillons

Mot-clic thématique	**Description de la thématique**
#BacktoTheFuture	Les célébrations du 30e anniversaire du film Back To The Future
#cop21	Conférence COP21 sur le Climat à Paris
#BigData	Le BigData
#cdnpoli	La politique canadienne
#DonDuSang	Polémique en France autour d'une politique d'exclusion des homosexuels concernant le don de leur sang
#fn	Le parti français Front National
#polQc	La politique québécoise
#PrayForMamaSwift	La vague de soutien pour la mère de Taylor Swift suite à son cancer
#PSG	L'équipe de football Paris Saint-Germain durant un match contre Chelsea
#Syria	Le conflit en Syrie

Notre choix de mots-clics a été motivé par un objectif de diversité dans les thématiques (sports, culture, technologie, politique, environnement), dans les langues

(français et anglais[6]) et dans la durée de vie des mots-clics. En effet, certains mots-clics sont encore largement utilisés (#cdnpoli), alors que d'autres sont disparus (#PrayForMamaSwift) ou le sens qu'on leur attribue a évolué (#DonDuSang).

Il est important de mentionner que tout notre processus de validation constitue une forme de rétro-ingénierie des processus à l'œuvre dans la création de folksonomies, de champs lexicaux, puis de relations de type *partieDe*. En effet, tel que nous l'avons explicité dans le chapitre dédié aux folksonomies (p.33), ces dernières reposent sur le regroupement de mots-clés et ceux-ci sont déterminés par les utilisateurs suite à un processus cognitif de catégorisation. Ainsi, nous avons défini les folksonomies en tant que « zone d'intersection des processus cognitifs de catégorisation » de plusieurs utilisateurs. En demandant aux évaluateurs de valider les résultats des inférences produites par notre prototype, nous tentons évidemment de mesurer la performance des inférences de nos algorithmes, mais, ce faisant, nous faisons appel à leur propre processus de catégorisation. Bien sûr, ils ont bénéficié d'un encadrement pour effectuer leur travail (formation et guide d'évaluation), mais il n'en demeure pas moins que le travail demandé en est un de catégorisation nécessitant certaines connaissances relatives aux thématiques, qu'il comprend une part de subjectivité et que, par conséquent, des différences s'observent dans leurs validations respectives. Lorsque des différences importantes apparaissent, elles sont relevées. Toutefois, nous tenterons de formuler nos conclusions à partir des points communs entre les deux évaluateurs.

[6] La performance de l'algorithme n'est pas dépendante de la langue, mais nous avons préféré limiter la création de nos échantillons aux langues française et anglaise pour en faciliter l'évaluation.

CHAPITRE 6 :
EXPÉRIMENTATIONS DU PROTOTYPE ET VALIDATION DES RÉSULTATS

Ce chapitre présente la mise en œuvre du cadre expérimental. La méthode de validation des résultats est d'abord explicitée pour chacun des trois niveaux de validation que nous avons retenus, puis nous présentons les résultats compilés pour l'ensemble des corpus.

Nous procédons ensuite à une analyse détaillée de deux corpus sur les thématiques de la politique canadienne et de la politique québécoise.

6.1 Validation de la création des champs lexicaux

6.1.1 Méthode de validation

La création de champs lexicaux s'effectue automatiquement par le regroupement des mots-clics ayant une forte occurrence. Nous avons repris l'exercice effectué sur le corpus de l'émission TopChef (p.55) sur nos 10 corpus du Tableau 5.2 en fournissant une description à chacun des mots-clics faisant partie du champ lexical pour ainsi permettre aux évaluateurs d'établir si ces derniers sont bel et bien en lien avec la thématique.

Les champs lexicaux évoluant selon l'actualité et leurs compositions étant recalculées aux 15 minutes, nous avons dû choisir un moment de façon aléatoire, mais devant répondre à deux critères : une période de la journée où l'actualité est fertile afin de s'assurer d'un maximum de données sur lesquelles faire nos calculs d'indicateurs sémantiques et une période de plus d'une heure après le début de la cueillette puisque certains calculs s'appuient sur les données recueillies dans la dernière heure.

Pour effectuer cette validation, nous avons constitué 10 tableaux (1 par thématique)

qui sont présentés dans l'Annexe B : champs lexicaux des 10 corpus d'évaluation (p.169).

1. Pour chacun des mots-clics présents dans le tableau, deux évaluateurs externes avaient pour tâche de déterminer si le mot-clic était en relation avec la thématique et, le cas échéant, de choisir une relation unissant le mot-clic candidat à la thématique. Leur sélection devait être faite à partir de l'un des quatre choix suivants : *None, owl:partOf, owl:equivalentTo, ou owl:hasPart.* Le résultat de cette validation permet de comparer les inférences de l'algorithme aux sélections des deux évaluateurs et de leur donner un score.
2. Un deuxième niveau d'analyse a permis de valider l'emploi de la loi de puissance de Zipf-Mandelbrot (p.47) en observant si la présence de relations partie-tout se maintient pour les mots-clics affichant une plus faible cooccurrence.

Les 10 tableaux constitués à partir des champs lexicaux créés par notre algorithme et présentés aux deux évaluateurs contenaient au total 259 mots-clics, donc une moyenne de 26 mots-clics par champ lexical, leur nombre variant en fonction du nombre de gazouillis recueillis dans l'heure précédente conformément à l'application de la loi de Zipf-Mandelbrot.

6.1.2 Résultats pour la création de champs lexicaux

Les champs lexicaux créés par l'algorithme se sont avérés riches en relations de type partie-tout. En effet, sur les 518 évaluations réalisées (259 mots-clics X 2 évaluateurs), 334 relations partie-tout, soit 64 % des mots-clics, ont été identifiées par les évaluateurs. Des différences existent entre les deux évaluateurs à ce chapitre. En effet, l'évaluateur 1 a identifié des relations partie-tout dans seulement 54 % des cas, alors qu'elles furent observables pour l'évaluateur 2 dans 74 % des cas. Ceci nous

renvoie à la part de subjectivité dans la validation dont nous discutions auparavant au sujet du processus de catégorisation propre à chacun.

La moyenne de cette validation effectuée sur 518 mots-clics indique tout de même que près de 2 mots-clics sur 3 des champs lexicaux déterminés par l'algorithme présentent des relations partie-tout avec le mot-clic thématique selon les évaluateurs. Soulignons que tous n'ont pas la même expertise des 10 thématiques étudiées. Malgré tout, il s'agit d'une forte proportion qui valide nos approches théorique et informatique permettant de passer d'une folksonomie à un champ lexical riche en relations partie-tout.

6.1.3 Résultats pour l'utilisation de la loi de Zipf-Mandelbrot

Rappelons que cette loi de puissance est utilisée afin de limiter le nombre de mots-clics sur lesquels les calculs d'indicateurs de cooccurrence sont effectués aux mots-clics présentant une forte occurrence. Nous avions fait l'hypothèse, suite à l'examen de données préliminaires, que plus la cooccurrence de mots-clics était forte, plus grande était la probabilité de relations partie-tout entre ceux-ci.

Pour tester cette hypothèse, nous avons repris les catégorisations des évaluateurs pour chacun des mots-clics (*None*, *owl:partOf, owl:equivalentTo, ou owl:hasPart)* et les avons isolées pour créer quatre sous-échantillons, un par relation. Pour chacun de ces échantillons, nous avons calculé le pourcentage moyen de cooccurrence. L'objectif demeurant de vérifier si les mots-clics ayant été identifiés par les évaluateurs comme n'ayant aucune relation partie-tout avec le mot-clic thématique présentaient réellement une plus faible cooccurrence.

Nous avons réuni ces moyennes dans ce tableau :

Tableau 6.1: Relations entre cooccurrence et probabilité de relations partie-tout

Type de relation	**Cooccurrence moyenne (évaluateur 1)**	**Cooccurrence moyenne (évaluateur 2)**	**Cooccurrence moyenne (évaluateur 1 et 2)**
None	4,0 %	2,9 %	3,6 %
partOf	6,5 %	6,0 %	6,2 %
EquivalentTo	6,4 %	7,2 %	6,8 %
hasPart	9,0 %	7,0 %	7,4 %

Note : La moyenne des 2 évaluateurs est influencée par le nombre, parfois différent, de relations identifiées. Par exemple, l'évaluateur 1 a identifié plus de relations de type « None », donc ses catégorisations ont une plus grande influence sur la moyenne.

À la lecture de ce tableau, nous observons une corrélation claire entre l'indicateur sémantique de cooccurrence de l'algorithme et la validation des évaluateurs. En effet, lorsque nous calculons la moyenne des cooccurrences des mots-clics identifiés comme n'ayant pas de relations partie-tout (*None*), nous observons qu'elle oscille entre 2,9 % (évaluateur 2) et 4,0 % (évaluateur 1). Par contre, lorsque nous refaisons le même calcul, mais cette fois-ci pour les mots-clics identifiés comme ayant des relations partie-tout (*partOf, EquivalentTo, hasPart)* la cooccurrence oscille plutôt entre 6,0 % et 9,0 %, soit un peu plus du double.

Ceci nous confirme deux choses. Premièrement, les résultats démontrent le lien entre une forte cooccurrence et la probabilité d'établir une relation partie-tout entre un mot-clic thématique et les mots-clics du champ lexical. De plus, ces résultats indiquent que l'emploi de la loi de puissance utilisée est efficace d'une part parce qu'elle produit un champ lexical riche en relations partie-tout en restreignant l'ajout de mots-clics ayant une trop faible cooccurrence et d'autre part parce qu'elle inclut suffisamment de mots-clics pour s'assurer que toutes les options soient épuisées. C'est-à-dire que l'emploi de cette loi de puissance inclut un peu plus de mots-clics que nécessaire, mais n’en omet aucun qui le sont. Si les évaluateurs avaient trouvé peu ou pas de mots-clics n'ayant aucune relation partie-tout (*None*), ceci aurait pu signifier que

l'algorithme n'en avait pas assez inclus dans le champ lexical. Le fait qu'il y ait 1 mot sur 3 qui ne présente aucune relation partie-tout et, surtout, que cette absence de relation soit reliée directement au faible indice sémantique de cooccurrence vient valider les aspects théoriques concernant le lien de causalité entre cooccurrence forte et relation partie-tout élaboré dans la section sur la formation des champs lexicaux (p.43).

Notons également que lorsque les évaluateurs ont estimé que le prédicat *None* était le plus approprié, l'algorithme est arrivé aux mêmes conclusions dans 96,7 % des cas. Dans les cas de « désaccords », 6 cas sur 184 (3,3 %), 5 cas exposaient également des désaccords entre les évaluateurs, ce qui est normal dû à une part de subjectivité dans la classification. Dans un seul cas sur 184, soit 0,05 %, les évaluateurs ont estimé qu'un mot-clic ne présentait aucune relation partie-tout alors que l'algorithme y a vu une relation de type *partieDe*. Il s'agit d'un pourcentage d'erreur avec lequel nous sommes confortable.

6.2 Validation des indicateurs sémantiques de cooccurrence inversée

6.2.1 Méthode de validation

L'attribution manuelle de prédicats RDF par les évaluateurs qui fut effectuée pour les 10 corpus ayant servi à la validation de la création de champs lexicaux nous a également permis de valider la qualité des inférences effectuées par l'algorithme lors de la sélection de candidats pour leur ajout à la requête en cours. Ici, la validation se fait en comparant les mots-clics candidats sélectionnés par l'algorithme aux prédicats RDF déterminés par les évaluateurs externes.

Nous vérifions pour chacun des mots-clics sélectionnés qu'il correspond bien à une relation pour laquelle un prédicat de type *owl :partOf ou owl:equivalentTo* a pu être déterminé. Autrement dit, nous isolons les mots-clics pour lesquels des calculs

d'indicateurs sémantiques de cooccurrence inversée sont supérieurs ou égaux à 0,5 et donc pour lesquels l'algorithme a attribué une relation *partieDe*, et nous les comparons avec les relations que les évaluateurs ont déterminé.

6.2.2 Résultats des indicateurs sémantiques de cooccurrence inversée « vs » classification des évaluateurs

Sur un total de 259 mots-clics répartis dans les 10 champs lexicaux, l'algorithme a attribué un prédicat *owl:partOf* à 69 mots-clics. Les deux évaluateurs n'avaient pas accès à cette information et ont attribué à ces mots-clics l'un des quatre prédicats RDF à leur disposition. L'intégral des classifications effectuées par chacun des évaluateurs se retrouve à l'Annexe C : Comparatif entre les décisions de l'algorithme et les classifications des évaluateurs sur les relations « partie de » (p.179). Nous proposons ici une synthèse en comparant les choix de l'algorithme à ceux des évaluateurs.

Sur les 138 classifications effectuées manuellement (69 mots-clics X 2 évaluateurs), un des deux évaluateurs, souvent les deux, a été en accord avec l'attribution du prédicat *owl:partOf* par l'algorithme à 89 %. Les deux évaluateurs ont quant à eux attribué le même prédicat RDF dans 86 % des cas, soit 59 choix identiques sur 69.

Nous avons cherché à savoir quel était le pourcentage d'efficacité de l'algorithme lorsque les deux évaluateurs s'entendaient sur le même prédicat et les résultats sont très positifs. Lorsque nous retirons les mots-clics pour lesquels les évaluateurs ont attribué des prédicats différents, le pourcentage d'efficacité de l'algorithme grimpe à 98 %. En fait, sur les 59 classifications communes, les évaluateurs n'ont été en désaccord avec l'algorithme qu'une seule fois.

6.3 Vérification de la cohérence du corpus

6.3.1 Méthode de validation

En tentant d'élargir la cueillette de gazouillis à partir d'une thématique donnée, notre

prototype augmente sensiblement le nombre de gazouillis dans un corpus. Cette augmentation quantitative ne doit pas être faite au détriment de la qualité du corpus. C'est donc dire qu'il faut s'assurer que le corpus reste cohérent par rapport à la thématique de départ. Il est donc primordial de mesurer cette cohérence dans une optique de contrôle de qualité afin de garantir que les inférences faites par notre algorithme résultent en une cueillette de gazouillis correspondant à la thématique de départ. Cette étape de validation est sans contredit la plus importante.

Nous avons mis en place une méthodologie de contrôle de qualité sur le contenu même des gazouillis afin de valider leur cohérence quant à la thématique choisie. La question à laquelle on tente de répondre se résume ainsi : « Pour un gazouillis X, peut-on affirmer qu'il est en lien avec la thématique? ».

Pour réaliser ce contrôle de qualité, nous avons créé deux sous-groupes d'échantillons de 100 gazouillis pour chacun des corpus. Le premier échantillon a été créé à partir de gazouillis faisant usage du mot-clic thématique et le deuxième a été créé à partir de gazouillis faisant seulement usage de mots-clics candidats sélectionnés par l'algorithme. Dans les deux cas, ces échantillons de 100 gazouillis ont été générés de façon aléatoire.

Nous ne connaissons pas pour l'instant d'outils permettant une validation automatisée de la thématique pour le contenu d'un gazouillis. Le texte d'un gazouillis étant très court et comportant parfois des erreurs grammaticales, des acronymes ou des abréviations, la seule méthode qui nous a semblé appropriée nécessitait une évaluation effectuée par un humain. Nous avons fait appel aux mêmes évaluateurs et ceux-ci ont évalué la pertinence de 4000 gazouillis (200 par thématique X 10 thématiques X 2 évaluateurs).

Afin d'éviter tout biais, les gazouillis des deux échantillons ont été présentés aux évaluateurs sans que ceux-ci ne puissent distinguer les gazouillis faisant usage du

mot-clic thématique de ceux faisant usage exclusivement de mots-clics candidats sélectionnés par l'algorithme. Pour chacun des gazouillis, les évaluateurs ont dû répondre si oui ou non ils considéraient que le contenu du gazouillis correspondait à la thématique. Il était également possible pour les évaluateurs de répondre « incertain ».

Le résultat de cette évaluation nous donne deux pourcentages de cohérence quant à la thématique, un pour les gazouillis utilisant le mot-clic thématique et un autre pour les gazouillis utilisant les mots-clics candidats sélectionnés. La différence entre les deux pourcentages de cohérence représentera notre indicateur de performance de l'algorithme de suivi de discussions et validera ou invalidera notre méthodologie et ses fondements théoriques.

6.3.2 Résultats de la validation de la cohérence du corpus

Sur les 10 corpus validés, un a dû être retiré suite à une erreur dans la requête SQL qui a généré les gazouillis à évaluer. Il s'agit du corpus à propos de la Conférence de Paris de 2015 sur le climat (#cop21).

Nous observons de très grandes variations quant au travail de validation des évaluateurs sur la question de la cohérence du corpus. Les résultats des évaluateurs sur certains corpus diffèrent de 42 %! Ces différences dans l'évaluation du contenu d'un gazouillis quant à son lien avec la thématique nous surprennent. Étant donné le peu d'options possibles dans cette portion de l'évaluation (oui, non ou incertain), nous nous attendions à une certaine uniformité dans les résultats, qu'ils soient positifs ou non.

Nous avons donc entrepris de faire un retour avec les évaluateurs sur cette question. Un point commun dans leurs commentaires concernait leurs connaissances des thématiques. En effet, les évaluateurs ont manifesté leurs difficultés à évaluer le

contenu à propos de thématiques leur étant peu familières. Suivant ces commentaires, nous avons soumis un questionnaire aux évaluateurs leur demandant d'indiquer leur niveau de connaissance pour chacune des 10 thématiques. Ils avaient le choix entre 5 options, soit « très faible », « faible », « moyenne », « bonne », « très bonne » auxquelles nous avons attribué une valeur de 1 à 5. Notre objectif est de vérifier si les forts écarts observés entre les évaluateurs peuvent être liés à leur familiarité avec la thématique. Les questionnaires remplis par les évaluateurs peuvent être consultés à l'Annexe D : Niveaux de connaissances des thématiques évaluées par les évaluateurs (p.184). Nous en présentons ici une synthèse.

Tableau 6.2: Comparatif des niveaux de connaissances des thématiques

Comment évaluez-vous votre connaissance des thématiques suivantes?	**Évaluateur 1**	**Évaluateur 2**	**Moyenne**
Les célébrations du 30e anniversaire du film Back To The Future	3	4	3,5
Conférence COP21 sur le Climat à Paris	3	2	2,5
Le BigData	3	1	2
La politique canadienne	4	4	4
Polémique autour d'une politique d'exclusion des homosexuels concernant leur don de sang	4	2	3
Le parti français Front National	4	2	3
La politique québécoise	5	4	4,5
La vague de support pour la mère de Taylor Swift suite à son cancer	2	3	2,5
L'équipe de football Paris Saint-Germain et le football européen	2	3	2,5
Le conflit en Syrie	4	2	3

Le premier constat qui ressort de l'analyse de ces formulaires est la très grande variété des niveaux de connaissance selon les thématiques. En compilant les résultats, on observe qu'une thématique telle que le « Big Data » était peu familière pour les deux évaluateurs avec un niveau de connaissance moyen se situant à « 2 » (faible), alors qu'il était de « 4,5 » (bonne-très bonne) pour la thématique de la politique québécoise.

Pour certaines thématiques, l'écart est grand entre les connaissances des deux

évaluateurs. Par exemple, pour la thématique de la « Polémique autour d'une politique d'exclusion des homosexuels concernant leur don de sang », l'évaluateur 1 a estimé sa connaissance « bonne » (4), alors que l'évaluateur 2 l'a estimée faible (2). Toutefois, pour certaines thématiques, le niveau de connaissance est comparable. Nous avons donc cherché à identifier des thématiques pour lesquelles le niveau de connaissance était à la fois semblable entre les évaluateurs et relativement élevé. Le but est ici de voir si les écarts entre les évaluations de chacun s'amincissaient lorsque les deux évaluateurs estimaient avoir une bonne connaissance de la thématique.

À la lecture de ce tableau, on observe deux thématiques pour lesquelles les deux évaluateurs ont indiqué des niveaux de connaissances « bonnes » ou « très bonnes ». D'abord, la thématique de la politique canadienne qui s'avère la seule pour laquelle les deux évaluateurs estiment avoir le même niveau de connaissances, soit « bonne » (4), puis la thématique de la politique québécoise pour laquelle le niveau moyen de connaissance s'est avéré le plus élevé soit « bonne-très bonne » (4,5).

Maintenant que nous avons une meilleure appréciation du niveau de connaissances des thématiques des évaluateurs, est-il possible de lier ce niveau de connaissances aux écarts observés entre leurs évaluations?

Le tableau qui suit présente l'écart entre les réponses des évaluateurs. Le calcul a été effectué à la fois pour le mot-clic thématique et les mots-clics sélectionnés et il représente le nombre total des différences d’évaluations par thématique.

Tableau 6.3: Écarts dans l'évaluation de la cohérence des gazouillis avec leur thématique

	Écart pour mot-clic thématique	**Écart pour mots-clics sélectionnés**	**Écart total**
Les célébrations du 30e anniversaire du film Back To The Future	40	42	82
Conférence COP21 sur le Climat à Paris	12	28	40
Le BigData	14	82	96
La politique canadienne	0	0	0
Polémique autour d'une politique d'exclusion des homosexuels concernant leur don de sang	16	76	92
Le parti français Front National	6	20	26
La politique québécoise	16	8	24
La vague de support pour la mère de Taylor Swift suite à son cancer	20	44	64
L'équipe de football Paris Saint-Germain et le football européen	48	84	132
Le conflit en Syrie	42	24	66

Si nous comparons ce tableau avec le précédent concernant le niveau de connaissances des thématiques par les évaluateurs, certaines corrélations se dégagent. Premièrement, les deux corpus sur lesquels nous observons les plus grandes variations en termes de jugement quant à la cohérence des gazouillis sont « L'équipe de football Paris Saint-Germain et le football européen », avec un écart total de 132, et le « Big Data », avec un écart total de 96. Il s'agit des thématique pour lesquelles les évaluateurs ont estimé leurs connaissances « faible-moyen» (2,5) et « faible » (2). À l'inverse, les deux corpus présentant les écarts les plus faibles soit « La politique québécoise », avec un écart de 24 et la «La politique canadienne », avec un écart de 0, sont effectivement des thématiques pour lesquelles les évaluateurs ont estimé avoir une connaissance « bonne-très bonne » (4,5) et « bonne » (4).

La corrélation entre le niveau de connaissances d'une thématique et une certaine cohérence dans l'évaluation de l'appartenance des gazouillis à cette dernière n'est

toutefois pas parfaite et nous observons pour des corpus où les évaluateurs ont estimé avoir des niveaux semblables des écarts qui eux sont relativement importants. Cette question mériterait d'être poussée, mais nous nous écarterions de notre problématique principale.

Néanmoins, il semblerait hasardeux de faire porter notre validation des résultats sur des évaluations de corpus sur lesquels les évaluateurs ne s'entendent pas. Nous avons donc choisi d'analyser les résultats des deux corpus présentant des variations qui nous ont semblé raisonnables et pour lesquels les évaluateurs ont estimé avoir un bon niveau de connaissance. Nous présenterons donc les résultats de l'évaluation de la cohérence des corpus de « La politique québécoise » et «La politique canadienne ».

Le tableau qui suit est une compilation des classifications des évaluateurs quant à l'appartenance de chacun des gazouillis. Rappelons que 200 gazouillis par thématique étaient soumis aux évaluateurs, 100 contenaient le mot-clic thématique et 100 autres contenaient seulement des mots-clics sélectionnés par l'algorithme.

Tableau 6.4: Validation de la cohérence des gazouillis

	Gazouillis avec mot-clic thématique			**Gazouillis sans mot-clic thématique**		
	Fait partie de la thématique?			Fait partie de la thématique?		
	oui	non	incertain	oui	non	incertain
Évaluateur 1						
La politique canadienne	90	7	3	93	1	6
La politique québécoise	88	10	2	85	15	0
Évaluateur 2						
La politique canadienne	90	7	3	93	1	6
La politique québécoise	86	4	10	85	11	4
moyenne	88,5	7	4,5	89	7	4

À la lecture de ce tableau, nous constatons l'importance d'avoir procédé à l'évaluation des gazouillis contenant le mot-clic thématique. En effet, il aurait été facile de

présumer que 100 % des gazouillis contenant le mot-clic thématique traitaient de cette dernière, alors que la réalité est tout autre. Selon les évaluateur, près de 1 gazouillis sur 10 contenant le mot-clic thématique ne faisant pas partie de cette dernière. Cette donnée pourrait être attribuée à la subjectivité des évaluateurs, mais à l'examen du contenu des gazouillis rejetés, une autre conclusion se dégage. Voici quelques gazouillis jugés hors thématique pour la politique canadienne, même s'ils contenaient le mot-clic #cdnpoli :

- Saudi Arabia no match for Iran: Aoun #uspoli #**cdnpoli** #ukpoli #gaza #plo #hamas http://t.co/qtmgnuytHH
- "RT @Bergg69: Crimea Is Now Putin's Problem Child http://t.co/EFCmzxqDcR #**cdnpoli** #yqr"
- RT @DAJHetherington: Our NATO ally #Turkey is letting a Kurdish city of 10,000 people burn to the ground. #**cdnpoli** #warcrime #Lice #Kurds
- The TVVH Urban Daily is out! http://t.co/NGrwiYozQ1 #**cdnpoli** #elxn2015 Stories via @TraderApple
- RT @ArtHealing44: #FreePalestine#Palestine#FreedomFlotilla #Gaza#Gaza1YearOn#Boycottisrael#ICC4israel#BDS#Europe#EU#**cdnpoli**

Ces gazouillis contenaient des liens vers des médias étrangers, des mentions d’usagers Twitter non canadiens, de l'autopromotion ou simplement une longue liste de mots-clics sans autre contenu. Il arrive parfois que des mots-clics populaires soient utilisés à l'intérieur de gazouillis dans le simple but d'attirer l'attention sur un autre sujet en manque de visibilité (Millette 2015) et #cdnpoli semble être un de ces mots.

Maintenant, sur la question qui nous intéressait de plus près, à savoir si les mots-clics sélectionnés par l'algorithme recueillent des gazouillis en lien avec la thématique, les résultats sont très satisfaisants. En effet, les évaluateurs ont jugé que les gazouillis contenant les mots-clics de candidats sélectionnés par l'algorithme s'inscrivaient tout

autant dans la thématique que le mot-clic thématique à une différence négligeable de 0,5 % en faveur de l'algorithme.

Cette dernière validation de données, la plus cruciale, démontre que le prototype et ses algorithmes sont en mesure d'augmenter le nombre de gazouillis relatifs à une thématique tout en maintenant sa cohérence. Par conséquent, l'utilisation de ce prototype à des fins de recherche représente une plus-value au plan de la quantité et de la qualité des corpus qui peuvent être recueillis. Par exemple, pour le corpus sur la politique canadienne recueilli du 9 au 27 juillet 2015, une cueillette n'utilisant que le mot-clic #cdnpoli aurait généré 323 681 gazouillis contre 479 317 avec l'utilisation de notre prototype alors que pour le corpus sur la politique québécoise, le mot-clic thématique aurait recueilli 151 954 gazouillis contre 266 116.

6.4 Analyse détaillée de deux corpus

Nous allons maintenant procéder à une analyse approfondie des corpus sur la politique canadienne et québécoise. D'une part, nous reprendrons certains des éléments de la validation en les détaillant davantage et d'autre part nous étendrons notre analyse à tout le corpus recueilli plutôt qu'à un moment précis comme ce fut le cas lors de la validation effectué par les évaluateurs.

6.4.1 Analyse du corpus sur la politique canadienne (#cdnpoli)

6.4.1.1 Création du champ lexical

Nous présentons ici un des champs lexicaux créé autour de la politique canadienne à partir du mot-clic thématique #cdnpoli. Le moment retenu a été le 26 juillet à 17:45:00. À ce moment, en plus du mot-clé thématique #cdnpoli, la requête en cours comprenait les mots-clés suivants :

- #CPCJesus
- #BCpoli
- #PolQC

- #cpc
- #canpoli
- #polCAN
- #elxn42
- #TM4PM
- #onpoli
- #StopHarper
- #lacmegantic
- #UCCB
- #c51
- #ndp
- #GPC
- #lpc

Cette requête en cours a recueilli un certain nombre de gazouillis contenant des mots-clics et ces derniers ont été triés par ordre décroissant automatiquement suivant le modèle de distribution de la loi de puissance de Zipf-Mandelbrot. Voici donc les 30 mots-clics sélectionnés pour faire partie du champ lexical.

Tableau 6.5: Champ lexical pour la thématique "politique canadienne" (#cdnpoli)

Mot-clic	**Description**[7]	**Prédicat RDF évaluateur 1**	**Prédicat RDF évaluateur 2**	**Cooc.**
lpc	« Liberal Party of Canada »	*owl :partOf*	*owl :partOf*	0.1555
ndp	« New Democratic Party »	*owl :partOf*	*owl :partOf*	0.1194
PolQC	politique québécoise	*owl :partOf*	*owl :partOf*	0.1150
cpc	« Conservative Party of Canada »	*owl :partOf*	*owl :partOf*	0.1150
BCpoli	« British-Columbia politics »	*owl :partOf*	*owl :partOf*	0.0800
TM4PM	« Thomas Mulcair for Prime Minister »	*owl :partOf*	*owl :partOf*	0.0624
onpoli	« Ontario politics »	*owl :partOf*	*owl :partOf*	0.0613
polCAN	politique canadienne	*owl:equivalentTo*	*owl:equivalentTo*	0.0570
CPCJesus	sarcasme sur les valeurs catholiques associées au parti conservateur	*owl :partOf*	*owl :partOf*	0.0493
pq	Parti Québécois	*owl :partOf*	*owl :partOf*	0.0416
assnat	Assemblée Nationale du Québec	*owl :partOf*	*owl :partOf*	0.0405
eglaw	circonscription de Eglinton-Lawrence ou la transfuge Eve Adams tentait de gagner une investiture pour le PLC	*owl :partOf*	*owl :partOf*	0.0383
BlocQC	Bloc Québécois	*owl :partOf*	*owl :partOf*	0.0383
c51	projet de loi fédérale	*owl :partOf*	*owl :partOf*	0.0372
elxn42	42e élection canadienne	*owl :partOf*	*owl :partOf*	0.0372
weekendofaction	attire l'action sur divers évènements	*none*	*none*	0.0350

[7] Les sources utilisées pour les descriptions sont Twitter, certains médias, Wikipédia, le site tagdef.com ou encore les connaissances générales du chercheur.

Mot-clic	**Description**	**Prédicat RDF évaluateur 1**	**Prédicat RDF évaluateur 2**	**Cooc.**
StopHarper	Arrêter Harper	*owl :partOf*	*owl :partOf*	0.0296
canpoli	« Canadian politics »	*owl:equivalentTo*	*owl:equivalentTo*	0.0263
UniRose	Circonscription fédérale à Toronto, University Rosedale	*owl :partOf*	*owl :partOf*	0.0252
Quebec	province	*owl :partOf*	*owl :partOf*	0.0241
Harper	Premier ministre canadien	*owl :partOf*	*owl :partOf*	0.0230
canada	pays	*owl :hasPart*	*owl :hasPart*	0.0230
PaysQc	mouvement indépendantiste du Québec	*owl :partOf*	*owl :partOf*	0.0208
BDS	Boycott, Divestment and Sanctions: movement to make Israel comply with international law and human rights	*none*	*none*	0.0197
Gaza	ville qui donne son nom au territoire de la bande de Gaza	*none*	*none*	0.0197
Europe	continent	*none*	*none*	0.0175
ICC4israel	Demanding that Israel gets referred to the International Criminal Court.	*none*	*none*	0.0175
FreePalestine	mouvement indépendant de la Palestine	*none*	*none*	0.0175
EU	European Union	*none*	*none*	0.0175
Gaza1YearOn	Opposition à la politique israélienne	*none*	*none*	0.0175

Le premier niveau de validation de ce champ lexical consiste à vérifier si des relations partie-tout ont pu être attribuées aux mots-clics sélectionnés par l'algorithme. Pour ce faire, des descriptions de chacun des mots-clics ont été récupérées à partir de différentes sources et à partir de ces descriptions, nous avons demandé aux deux évaluateurs externes de sélectionner l'un des quatre prédicats suivants: *owl:hasPart*, *owl:partOf*, *owl:equivalentTo* ou *None*.

Des 30 mots-clics identifiés par l'algorithme comme faisant partie du champ lexical, 8 ont été identifiés par les évaluateurs comme n'ayant aucune relation avec la thématique soit 27 % de ceux-ci. Ce pourcentage peut sembler élevé, mais à bien y regarder, on se rend compte que la plupart des mots-clics ayant été identifiés comme n'ayant aucune relation avec la politique canadienne sont reliés au conflit israélo-

palestinien. Or, le Canada, sous le gouvernement Harper, fut un proche allié d'Israël et prit position de façon régulière sur ce conflit. Il n'est donc pas complètement faux d'inclure des mots-clics en lien avec ce conflit dans le champ lexical de la politique canadienne, car il constitue bel et bien un enjeu important de la politique étrangère de ce pays. Toutefois, en termes de relation *partie-tout* la classification de ce conflit dans la politique canadienne est plus difficile. En effet, le conflit israélo-palestinien fait-il partie de la politique canadienne (*owl:partOf*) ou ou est-ce l'inverse? Probablement, qu'à certains moments de l'année on pourrait considérer que ce conflit fait effectivement partie de la politique canadienne, mais on pourrait aussi faire l'hypothèse qu'à ces mêmes moments il fait aussi partie de la politique américaine, française, britannique, allemande, etc. Ainsi on pourrait donner un sens à l'inclusion momentanée de ce conflit dans le champ lexical de la politique de chacun de ces pays.

Un deuxième niveau de validation et d'analyse consiste à observer si la présence de relations partie-tout se maintient pour les mots-clics affichant une plus faible cooccurrence. Dans le cas de ces 30 mots-clics sélectionnés par le modèle de distribution de la loi de puissance de Zipf-Mandelbrot, on observe que 7 des 8 mots-clics pour lesquels les évaluateurs n'ont pu trouver de relations partie-tout sont ceux ayant la plus faible cooccurrence. Cet état de fait est observable dans tous les corpus tel que nous l'avons mentionné dans la section Résultats pour l'utilisation de la loi de Zipf-Mandelbrot (p.98).

6.4.1.2 Indicateurs sémantiques de cooccurrence inversée

Il s'agit ici de comparer les résultats de l'algorithme aux classifications des évaluateurs quant aux types de prédicats RDF attribués aux mots-clics candidats, l'objectif du calcul de l'indicateur sémantique de cooccurrence inversée étant de déterminer à quels mots-clics on peut accoler un prédicat RDF *owl:parOf*. Pour ce

faire, nous allons reprendre le tableau précédent en conservant les relations partie-tout établies par les évaluateurs et en y ajoutant les résultats obtenus par l'algorithme.

Tableau 6.6: Comparatif de choix de prédicats: évaluateurs "vs" algorithme

Mots-clics	**Prédicat attribué par les évaluateurs 1 et 2**[8]	**Indice de cooccurrence inversée**	**Prédicat attribué par l'algorithme**	**Adéquation**
lpc	*owl :partOf*	0.6410	*owl :partOf*	oui
ndp	*owl :partOf*	0.6970	*owl :partOf*	oui
PolQC	*owl :partOf*	0.6296	*owl :partOf*	oui
cpc	*owl :partOf*	0.7381	*owl :partOf*	oui
BCpoli	*owl :partOf*	0.5476	*owl :partOf*	oui
TM4PM	*owl :partOf*	0.6031	*owl :partOf*	oui
onpoli	*owl :partOf*	0.6234	*owl :partOf*	oui
polCAN	*owl:equivalentTo*	0.6053	*owl :partOf*	oui
CPCJesus	*owl :partOf*	0.7108	*owl :partOf*	oui
pq	*owl :partOf*	0.0000	*none*	non
assnat	*owl :partOf*	0.0000	*none*	non
eglaw	*owl :partOf*	0.4706	*none*	non
BlocQC	*owl :partOf*	0.0000	*none*	non
c51	*owl :partOf*	0.8389	*owl :partOf*	oui
elxn42	*owl :partOf*	0.7471	*owl :partOf*	oui
weekendofaction	*none*	0.0737	*none*	oui
StopHarper	*owl :partOf*	0.8679	*owl :partOf*	oui
canpoli	*owl:equivalentTo*	0.5516	*owl :partOf*	oui
UniRose	*owl :partOf*	0.2121	*none*	non
Quebec	*owl :partOf*	0.0200	*none*	non
Harper	*owl :partOf*	0.5013	*owl :partOf*	oui
canada	*owl :hasPart*	0.0107	*none*	oui
PaysQc	*owl :partOf*	0.0000	*none*	non
BDS	*none*	0.0067	*none*	oui
Gaza	*none*	0.0000	*none*	oui
Europe	*none*	0.0000	*none*	oui
ICC4israel	*none*	0.0390	*none*	oui
FreePalestine	*none*	0.0000	*none*	oui

[8] Dans le cas du champ lexical du corpus de la politique canadienne, les prédicats choisis par les évaluateurs étaient identiques. Ils ont donc été jumelés dans une seule colonne.

Mots-clics	Prédicat attribué par les évaluateurs 1 et 2	Indice de cooccurrence inversée	Prédicat attribué par l'algorithme	Adéquation
EU	*none*	0.0000	*none*	oui
Gaza1YearOn	*none*	0.0064	*none*	oui

Dans la colonne « Adéquation », nous comparons le prédicat RDF attribué par les évaluateurs à la décision de l'algorithme en fonction de la valeur de l'indicateur sémantique de cooccurrence inversée. Rappelons que lorsque ce dernier est supérieur à 0,5, l'algorithme infère une relation *owl:partOf* et sélectionne le candidat pour l'inclure dans la requête en cours. Sur les 30 mots-clics, il y a 23 adéquations entre le prédicat RDF déterminé par les évaluateurs et la décision prise par l'algorithme. Regardons de plus près les mots-clics pour lesquels les choix des évaluateurs et de l'algorithme diffèrent.

Un premier groupe en lien avec la politique québécoise apparait clairement. Il est composé de 5 des 7 mots-clics pour lesquels les décisions des évaluateurs et de l'algorithme ne concordent pas. Ces mots-clics sont #pq, #assnat, #BlocQC, #Quebec et #PaysQC. Pour tenter de comprendre les inférences produites par l'algorithme, regardons de plus près les données ayant servi à les produire. Le tableau suivant montre les données obtenues lors de la cueillette d'informations auprès de la *Search API* ainsi que le calcul de l'indicateur sémantique de cooccurrence inversée en résultant.

Tableau 6.7: Différences dans les choix de prédicats (a)

Mots-clics	Prédicat RDF	Occurrence inversée	Nbre de gazouillis recueillis pour le candidat	Ind. de cooccurrence inversée
pq	*owl :partOf*	0	12	0.0000
assnat	*owl :partOf*	0	5	0.0000
BlocQC	*owl :partOf*	0	0	0.0000
Quebec	*owl :partOf*	9	451	0.0200
PaysQc	*owl :partOf*	0	2	0.0000

À l'exception du mot-clic #Quebec sur lequel les inférences ont pu être calculées sur un grand nombre de gazouillis (451), nous remarquons le faible nombre de gazouillis recueillis pour les 4 autres mots-clics : 12 pour #pq, 5 pour #assnat, 2 pour #PaysQC et 0 pour #BlocQC. Le mot-clic #Quebec a été exclu par l'algorithme à cause de sa forte entropie (voir fin de la section Formation de champs lexicaux p. 43). Il est normal que les évaluateurs lui aient attribué une relation *owl:partOf* dans un contexte où une majorité de mots-clics se référaient à la politique canadienne, mais il était impossible pour eux de tenir en compte de l'entropie de ce mot-clic.

Par contre, les quatre autres mots-clics ont une faible entropie et on pourrait arguer qu'ils font partie de la politique canadienne. Pourquoi n'ont-ils pas été sélectionnés par l'algorithme? Au moins deux pistes s'offrent à nous pour tenter de répondre à cette question. D'abord, comme on l'a mentionné, le nombre de gazouillis recueillis à partir de la *Search AP*I est très faible pour ces mots-clics. C'est une limite technique de Twitter qui est accrue par le fait de ne pouvoir interroger cette API que dans une seule langue à la fois. Le corpus étant majoritairement anglophone, les requêtes envoyées à la *Search API* ont tenté de récupérer des gazouillis en anglais. Une deuxième piste de réponses apparait lorsqu'on observe le nombre brut d'occurrences inversées. Aucun des 17 mots-clics de la requête en cours n'est en cooccurrence inversée avec les 4 mots-clics (#pq, #assnat, #PaysQC, #BlocQC). En effet, les indicateurs de cooccurrence inversée de chacun de ces mots-clics sont à zéro, un phénomène difficilement explicable. Il est possible ici que les différences culturelles ou politiques entrent en jeu et que, dans une certaine mesure, les discussions à propos de la politique québécoise fassent moins partie des discussions à propos de la politique canadienne que des discussions à propos de la politique d'autres provinces, mais encore une fois la quantité de données recueillies auprès de la *Search API* est trop faible pour pousser les investigations.

Les deux autres mots-clics pour lesquels on observe des différences entre les choix de

l'évaluateur et de l'algorithme sont #Unirose et #eglaw, deux circonscriptions fédérales de l'Ontario.

Tableau 6.8: Différences dans les choix de prédicats (b)

Mots-clics	Prédicat RDF	Occurrence inversée	Nbre de gazouillis recueillis pour le candidat	Ind. de cooccurrence inversée
eglaw	*owl :partOf*	24	51	0.4706
UniRose	*owl :partOf*	7	33	0.2121

La circonscription fédérale de Eglinton—Lawrence (#eglaw) a été sous les feux des projecteurs médiatiques lors de l'investiture libérale. Une des candidates à l'investiture, Eve Adams, fut fortement médiatisée d'une part parce qu'elle était une transfuge en provenance du Parti Conservateur et d'autre part parce qu'elle est fiancée à Dimitri Soudas qui avait dû démissionner de son poste de directeur général au Parti Conservateur suite à des allégations de favoritisme dans la nomination de sa fiancée lors d'une investiture dans une autre circonscription en 2014. L'investiture de Eglinton-Lawrence revêtait donc un caractère spécial au niveau national. L'analyse du mot-clic #eglaw permet de prendre toute la mesure de la dimension temporelle dans la construction d'un champ lexical. En effet, le mot-clic #eglaw a fait partie du champ lexical pendant un peu moins de 24 heures soit du 26 juillet, 14h36 au 27 juillet, 13h07, ce qui correspond au jour de l'investiture (26 juillet) et à la couverture médiatique qui en découla. La figure qui suit montre bien l'évolution de #eglaw dans le champ lexical.

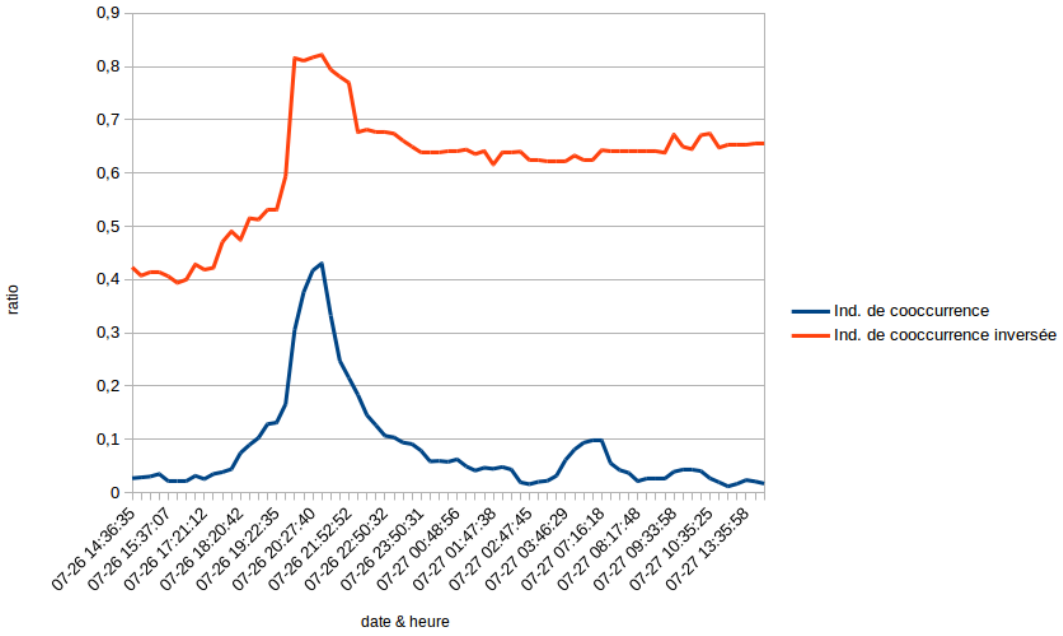


Figure 6.1: Évolution du mot-clic #eglaw dans le temps

On note d'abord l'inclusion de #eglaw dans le champ lexical en milieu d'après-midi (14h36) le dimanche 26 juillet puis l'augmentation de son indicateur sémantique de cooccurrence (bleu) de 18h00 à 21h00. Parallèlement, vers 18h30, l'indicateur sémantique de cooccurrence inversée (orange) passe la barre des 0,5 et #eglaw, désormais considéré par l'algorithme comme ayant une relation *owl:partOf* avec la thématique de la politique canadienne, est sélectionné comme candidat pour faire partie de la requête en cours à la *Streaming API*. Entre 20h30 et 21h00, l'indicateur sémantique de cooccurrence est à plus de 40 % et celui de cooccurrence inversée à plus de 80 %, ce qui signifie que l'investiture de Eve Adams est le sujet de l'heure pour la thématique de la politique canadienne en ce dimanche soir.

Au moment où nous avons choisi de prendre un instantané pour analyser et valider les données soit le 26 juillet à 17:45:00, #eglaw, dont l'indice de cooccurrence était un

peu en deçà de 0,5, était à quelques minutes d'être considéré par l'algorithme comme ayant une relation *owl:partOf* avec la thématique. Voilà ce qui explique la différence entre les décisions des évaluateurs et celle de l'algorithme à ce moment précis. L'indicateur sémantique de cooccurrence inversée s'est ensuite maintenu au-delà de 0,5 jusqu'à ce que #eglaw disparaisse du champ lexical le lendemain en début d'après-midi.

Le cas de la circonscription de University-Rosedale est très particulier au sens où il oppose deux candidates qui sont d'ex-journalistes populaires au Canada anglais qui sont très présentes sur le réseau Twitter. Jennifer Hollett, du NPD, fut VJ au réseau MuchMusic avant de faire partie des journalistes de l'émission d'affaires publiques « *Connect with Mark Kelley* » (« Jennifer Hollett », 2017). Elle donne également des conférences sur la politique et les nouveaux médias. La députée sortante, Chrystia Freeland, du PLC, fut journaliste au *Financial Times*, au *Globe and Mail* et au *Washington Post* avant d'occuper un poste dans la division numérique de *Thomson Reuters (« Chrystia Freeland », 2017).* En plus d'être des personnalités médiatiques, elles ont toutes deux des expertises en ce qui a trait aux médias et aux réseaux sociaux.

Leurs équipes médias respectives sont très actives sur Twitter et les citoyens de cette circonscription du centre-ville de Toronto font partie de la tranche de la population la plus impliquée sur Twitter. En effet, Twitter est plus populaire auprès des gens scolarisés et bien nantis vivant en milieu urbain (Duggan et al, 2014). Pour toutes ces raisons, il semblerait que le mot-clic #Unirose fasse irruption dans le champ lexical de la politique canadienne sans pour autant qu'il ne s'y déroule d'évènements spéciaux d'intérêt national.

La figure ci-dessous montre les trois périodes durant lesquelles #Unirose a été considéré comme faisant partie du champ lexical soit les 13, 15 et 26 juillet. On

remarque à la fois les faibles taux de cooccurrence et de cooccurrence inversée.

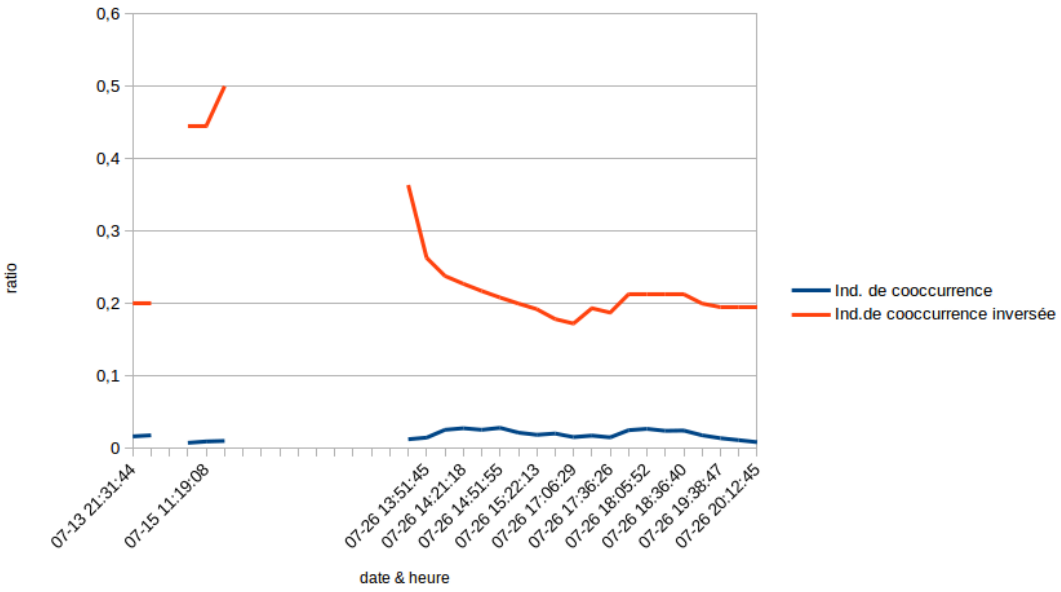


Figure 6.2: Évolution du mot-clic #Unirose dans le temps

Ces données ont été cumulées en début de campagne électorale durant la période estivale alors que le nombre de gazouillis sur cette thématique n'était pas à son plus fort. Quoi qu'il en soit, il est normal que les évaluateurs aient attribué un prédicat RDF *owl:partOf* à un mot-clic correspondant à une discussion à propos d'une circonscription fédérale lors d'une campagne électorale. Toutefois le faible taux de cooccurrence et de cooccurrence inversée calculé par nos algorithmes tend à montrer que la thématique entourant #Unirose en est une qui est plutôt locale que nationale.

En conclusion de cette section sur les comparaisons entre les indicateurs sémantiques déterminés par notre algorithme et les décisions des évaluateurs à propos des prédicats RDF, notons deux éléments principaux. Premièrement, la variable multilingue d'un corpus représente un obstacle supplémentaire. Peut-on même aborder la notion de champ lexical construit à partir de folksonomies dans une

optique multilingue? Ensuite, il importe de souligner que les différences entre les évaluateurs et l'algorithme n'ont pas produit de faux positifs. C'est-à-dire que dans chacun des cas où il n'y avait pas adéquation entre les deux, la décision de l'algorithme fut de ne pas inclure le mot-clic en tant que sous-thématique. Par exemple, aucun mot-clic s'étant vu attribué un prédicat *owl:hasPart* ou *None* par les évaluateurs n'a été sélectionné par l'algorithme.

6.4.1.3 Cohérence du corpus

L'objectif ultime de notre recherche étant d'arriver à mieux suivre une thématique de discussion en y ajoutant des sous-thématiques pertinentes, il est attendu que le corpus de gazouillis augmente considérablement, car de nouveaux mots-clics sont ajoutés à la requête en cours à la *Streaming API* de Twitter. Par exemple, pour ce corpus recueilli du 9 au 27 juillet 2015, une cueillette de gazouillis ayant utilisé exclusivement le mot-clic thématique #cdnpoli aurait recueilli 323 681 gazouillis. Avec l'application de nos algorithmes, 155 636 gazouillis supplémentaires ont été ajoutés au corpus soit une augmentation de 48 %. Il s'agit d'une amélioration substantielle permettant de mieux capter les discussions au sujet de la politique canadienne dans l'espace public.

Nous avons déjà observé que selon les évaluateurs les gazouillis du corpus contenant le mot-clic thématique et ceux du corpus contenant des mots-clics sélectionnés par l'algorithme présentaient des niveaux de cohérence semblables par rapport à la thématique. Penchons-nous maintenant sur ces gazouillis jugés hors thématiques.

Dans le corpus de gazouillis contenant le mot-clic thématique #cdnpoli la majorité des gazouillis jugés hors thématiques concerne le conflit israélo-palestinien et le Moyen-Orient. Il est intéressant de noter que dans le corpus recueillis à partir des mots-clics ajoutés par l'algorithme, les gazouillis considérés hors thématique ou pour lesquels les évaluateurs avaient une incertitude ne sont jamais reliés au conflit israélo-

palestinien. Cela pourrait s'expliquer par le fait que des mots-clics relatifs à ce conflit sont en cooccurrence avec le mot-clic thématique #cdnpoli, mais pas en cooccurrence avec les mots-clics candidats sélectionnés représentant des sous-thématiques. Les gazouillis considérés hors thématique étaient plutôt reliés à des actualités provinciales possiblement jugées trop locales par les évaluateurs pour être considérées dans la grande thématique de la politique canadienne. La frontière entre les deux est en effet parfois très mince.

6.4.1.4 Autres éléments d'analyse : la découverte de synonymes

Les mots-clics synonymes d'un mot-clic thématique sont difficilement détectables par leur cooccurrence avec ce dernier. En effet, les utilisateurs emploient l'un ou l'autre, mais rarement les deux dans le même gazouillis. Dans le corpus du mot-clic #cndpoli, les évaluateurs ont attribué le prédicat RDF *owl:equivalentTo* à deux mots-clics. Ces deux mots-clics, #PolCAN et #canpoli, peuvent effectivement être considérés comme des synonymes du mot-clic thématique #cdnpoli.

Figure 6.3: Inclusion de synonymes au champ lexical

L'inclusion de mots-clics synonymes d'abord dans le champ lexical d'une thématique puis éventuellement dans la requête en cours ne se fait pas grâce à une cooccurrence directe avec le mot-clic thématique. Elle émerge plutôt d'une cooccurrence avec les mots-clics qui s'ajoutent progressivement à la requête en cours et nécessite par conséquent plusieurs itérations de l'algorithme. Par exemple, à la première itération de l'algorithme, le 9 juillet à 9h45, ni #PolCAN ou #canpoli ne faisaient partie du champ lexical. Toutefois, les mots-clics #NDP, #cpc, #LPC, #cdnecon et #qcpoli faisaient partie du champ lexical et furent parmi ceux présentant une cooccurrence inversée assez forte (> 0,5) pour être ajoutés à la requête en cours dès le début de la cueillette. Une fois ajoutés à la requête en cours, la cueillette de gazouillis avec ces mots-clics augmente le potentiel de découverte de synonymes puisque des mots-clics

comme #PolCAN ou #canpoli sont susceptibles de se retrouver en cooccurrence avec plusieurs des mots-clics de la requête en cours. L'ajout d'un mot-clic synonyme au champ lexical d'une thématique est donc le résultat de l'addition de sa cooccurrence avec les mots-clics de la requête en cours, alors que son inclusion à la requête en cours est l'addition de ses cooccurrences inversées avec ces derniers. Rappelons au passage que ces calculs sont des tests binaires de type VRAI ou FAUX. Par exemple, si un gazouillis contient un mot-clic de la requête en cours, ce gazouillis fait partie de la discussion à propos de la thématique principale et la valeur de la cooccurrence inversée sera augmentée de +1.

Modélisons maintenant la mécanique à l'œuvre dans l'inclusion des synonymes au champ lexical en utilisant un exemple tiré de ce corpus avec les mots-clics #PolCAN et #canpoli.

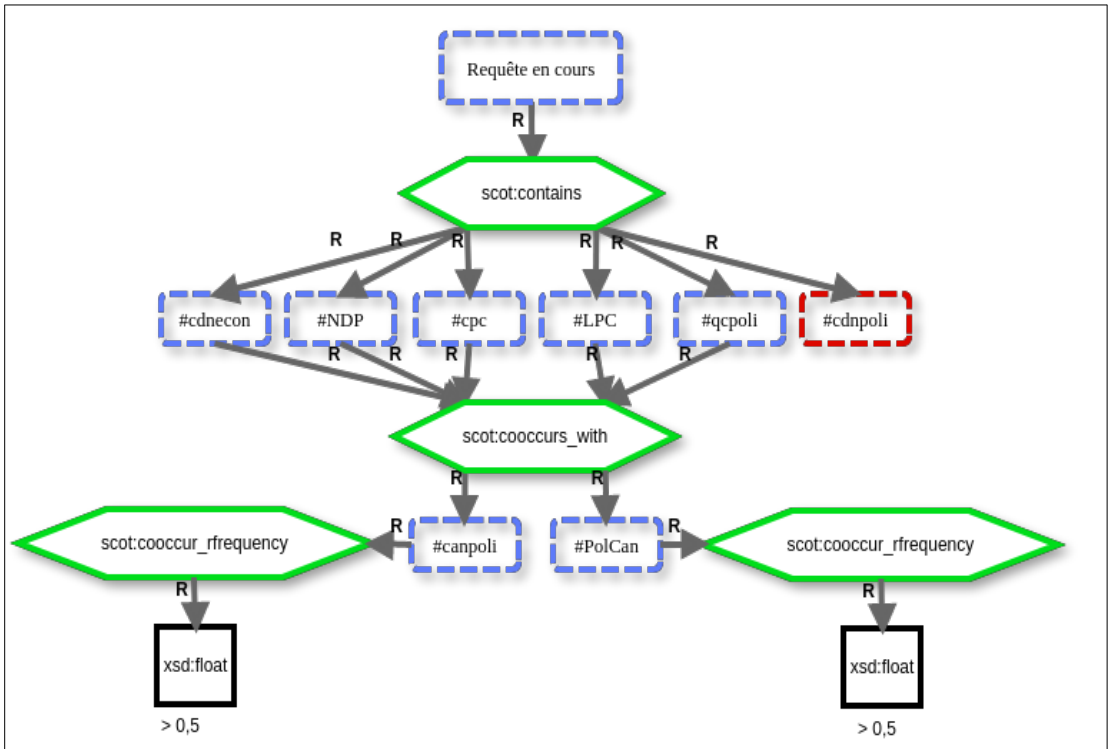


Figure 6.4: Inclusion de synonymes à la requête en cours

Dans cette figure, on observe d'abord que les mots-clics #NPD, #cpc, #LPC, #cdnecon et #qcpoli ont été inclus dans la requête en cours, car l'algorithme les a sélectionnés lors d'une inférence antérieure. Le mot-clic thématique, qui fait toujours partie de la requête en cours, est surligné en rouge. Les mots-clics #canpoli et #PolCAN, bien que n'étant pas en cooccurrence avec le mot-clic thématique #cdnpoli, sont utilisés fréquemment en remplacement de ce dernier par certains utilisateurs ayant publié des gazouillis contenant l'un des mots-clics de la requête en cours. Par conséquent, leur occurrence dans le corpus, illustré par le prédicat *scot:ooccurs_in*[9], augmente. C'est donc la somme de ces cooccurrences avec les mots-clics de la requête en cours qui permettra à #canpoli et #PolCAN d'éventuellement être inclus dans le champ lexical. Ce ratio, une valeur entre 0 et 1, exprimé par le prédicat *scot:cooccur_rfrequency* est évalué par l'algorithme appliquant la loi de Zipf-Mandelbrot et, s'il considère que le ratio est suffisamment élevé, il inclura le ou les mots-clics dans le champ lexical. À cette étape, notons qu'ils ne sont toujours pas ajoutés à la requête en cours puisque leur indice de cooccurrence inversée n'a pas encore été calculé, mais ils deviennent du moins des candidats.

Modélisons maintenant l'inclusion de ces deux mots-clics à la requête en cours en utilisant l'inférence qui permet à l'algorithme d'évaluer leur appartenance à une sous-thématique de la thématique principale, soit une relation *partieDe*.

Ce schéma illustrant le raisonnement inductif de l'algorithme ainsi que les inférences qu'il appliquera a plusieurs points communs avec le précédent schéma d'inclusion de synonymes au champ lexical. Par exemple, les inférences sont faites sur les mêmes mots-clics autant du côté des mots-clics présents dans l'ensemble de mots-clics de la requête en cours que des deux candidats #canpoli et #PolCan. Une première différence apparait par l'utilisation du prédicat *scot:cooccurs_with* plutôt que

[9] Le prédicat *scot:cooccurs_in* représente une relation où un mot-clic cooccurre avec un groupe de mots-clics

scot :cooccurs_in, ce dernier étant utilisé pour illustrer la relation d'un mot-clic qui cooccurre avec un ensemble de mots-clics, alors que *scot:cooccurs_with* est utilisé pour illustrer la relation d'un ensemble de mots-clics qui cooccurrent avec un mot-clic. Ce sont deux prédicats qui sont l'inverse de l'autre et c'est bien ce que nous cherchons à faire avec notre calcul d'indicateur sémantique de cooccurrence inversée.

Dans les deux schémas, nous utilisons le prédicat scot :cooccur_rfrequency, mais pour calculer deux valeurs différentes. En effet, dans le schéma d'inclusion au champ lexical, on calculait le ratio de cooccurrence de #canpoli et de #PolCAN avec les mots-clics de la requête en cours alors que dans ce schéma d'inclusion de synonymes à la requête en cours, le calcul est inversé.

Il est à noter que le prédicat *owl:equivalentTo*, qui aurait également pu être le prédicat *scot:synonym*, a été offert comme choix aux évaluateurs pour éviter de les obliger à choisir entre deux relations *partie-tout*, *owl:hasPart* et *owl:partOf*, qui ne correspondent ni l'une ni l'autre à la définition d'une relation entre deux objets presque identiques. Toutefois, les calculs de similarité sémantique de l'algorithme demeurent les mêmes, qu'il s'agisse de la découverte de sous-thématiques ou de synonymes. L'unique différence est que l'identification d'une sous-thématique ne requiert généralement qu'un fort ratio de cooccurrence inversée avec le mot-clic thématique alors que l'inclusion d'un synonyme nécessite l'addition de cooccurrences inversées avec plusieurs mots-clics de la requête en cours.

6.4.2 Analyse du corpus sur la politique québécoise

6.4.2.1 Création du champ lexical

Le tableau ci-dessous présente un des champs lexicaux créé autour de la politique québécoise à partir du mot-clic thématique #polQc. Le moment retenu a été le 20 aout 2015 à 17:30. À ce moment, en plus du mot-clic thématique #polQc, la requête en

cours comprenait les mots-clics suivants :

#PaysQc	#TournéePQ	#BlocQC
#PL44	#plq	#assnat
#caq	#pq	

Cette requête en cours a recueilli quelques centaines de gazouillis contenant des mots-clics et ils ont été triés par ordre décroissant automatiquement suivant le modèle de distribution de la loi de puissance de Zipf-Mandelbrot. Voici donc les 14 mots-clics sélectionnés pour faire partie de champ lexical.

Tableau 6.9: Champ lexical pour la thématique "politique québécoise"

Mot-clic	**Description**[10]	**Prédicat RDF** (évaluateur 1)	**Prédicat RDF** (évaluateur 2)	**Cooc.**
assnat	Assemblée Nationale	*owl :partOf*	*owl :partOf*	0,3578
elxn42	42e élection canadienne	*none*	*owl :hasPart*	0,2586
BlocQC	Bloc Québécois	*owl :partOf*	*owl :partOf*	0,2328
polcan	politique canadienne	*owl :hasPart*	*owl :hasPart*	0,1983
pq	Parti Québécois	*owl :partOf*	*owl :partOf*	0,1121
fed2015	Élection fédérale 2015	*none*	*owl :hasPart*	0,1034
PaysQc	mouvement indépendantiste du Québec	*owl :partOf*	*owl :partOf*	0,0690
plq	Parti libéral du Québec	*owl :partOf*	*owl :partOf*	0,0647
PL44	projet de loi 44 du Québec	*owl :partOf*	*owl :partOf*	0,0560
caq	Coalition Avenir Québec	*owl :partOf*	*owl :partOf*	0,0388
Pl59	projet de loi 59 du Québec	*owl :partOf*	*owl :partOf*	0,0388
Québec	province	*owl :hasPart*	*owl :hasPart*	0,0388
PKP	Pierre-Karl Péladeau	*owl :partOf*	*owl :partOf*	0,0345
cdnpoli	politique canadienne	*owl :hasPart*	*owl :hasPart*	0,0345

Tout comme pour le corpus à propos de la politique canadienne, notre première étape de validation consiste à vérifier la présence de relations partie-tout dans le champ lexical induit par l'algorithme. Dans le cas de cette thématique, encore plus que pour

[10] Les sources utilisées pour les descriptions sont Twitter, certains médias, Wikipédia, le site tagdef.com ou encore les connaissances générales du chercheur.

la politique canadienne, la proportion de relations partie-tout est très importante. L'évaluateur 2 a jugé que chacun des mots-clics du champ lexical pouvait être lié à la thématique par une relation *partie-tout*, alors que l'évaluateur 1 a considéré que cette relation était présente 12 fois sur 14. Les deux cas où l'évaluateur 1 a jugé qu'il n'existait pas de relation partie-tout sont ceux où les mots-clics faisaient référence à l'élection fédérale canadienne. Dans ce cas, l'évaluateur 2 a choisi une relation *owl:hasPart*. Dans le contexte sociopolitique québécois, ces différences de perceptions « sémantiques » étaient prévisibles… Pour tous les autres mots-clics, les évaluateurs sont arrivés aux mêmes conclusions quant à la relation partie-tout qui les unissait à la thématique.

Dans le corpus de la politique canadienne, nous avions observé que lorsque la cooccurrence diminuait, les relations de type partie-tout diminuaient également. Le champ lexical de la politique québécoise au moment où nous l'avons extrait ne permet pas d'observer le même phénomène. Il s'agit là d'une exception par rapport à la moyenne des 10 corpus analysés.

On remarquera l'inclusion au champ lexical des trois partis provinciaux les plus populaires au Québec et l'exclusion de plus petits partis comme Québec solidaire (7,63 % des voix aux élections de 2014) et Option Nationale (0,73 %). Notons toutefois que sur la période de cueillette du 20 aout au 1er octobre 2015, Québec Solidaire a été inclus au champ lexical plusieurs fois par jour alors qu'Option Nationale s'y est fait une place une à deux fois par semaine. Ces inclusions momentanées sont bien évidemment dues à leur faible ratio de cooccurrence et à l'application de la loi de Zipf-Mandelbrot. Leur inclusion ou leur exclusion n'a que très peu d'impact sur le corpus en terme quantitatif, mais une recherche qui porterait spécifiquement sur la politique québécoise pourrait opter pour une inclusion manuelle de ces deux mots-clics, malgré le faible volume de gazouillis en faisant usage.

6.4.2.2 Indicateurs sémantiques de cooccurrence inversée

Tout comme nous l'avons fait pour le corpus de la politique canadienne, nous allons maintenant comparer les prédicats déterminés par les évaluateurs aux inférences de l'algorithme. La colonne « Adéquation » sert à vérifier si les choix des évaluateurs et ceux de l'algorithme coïncident. Notons que l'algorithme n'effectue ses choix qu'entre une relation *owl:partOf* et aucune relation, donc l'adéquation sera considérée positive lorsque les évaluateurs auront choisi *owl:partOf* ou *owl:equivalentTo* et que l'algorithme aura choisi *owl:partOf* ou encore lorsque les évaluateurs auront choisi *owl:hasPart* ou *none* et que l'algorithme aura choisi *none*.

Tableau 6.10: Comparatif de choix de prédicats: évaluateur "vs" algorithme

Mot-clic	**Prédicat RDF** (évaluateur 1)	**Prédicat RDF** (évaluateur 2)	**Indice de cooccurrence inversée**	**Prédicat attribué par l'algorithme**	**Adéquation**
assnat	*owl :partOf*	*owl :partOf*	0,9512	*owl :partOf*	oui
elxn42	*none*	*owl :hasPart*	0,0469	*none*	oui
BlocQC	*owl :partOf*	*owl :partOf*	0,5766	*owl :partOf*	oui
polcan	*owl :hasPart*	*owl :hasPart*	0,4103	*none*	oui
pq	*owl :partOf*	*owl :partOf*	0,8023	*owl :partOf*	oui
fed2015	*none*	*owl :hasPart*	0,3744	*none*	oui
PaysQc	*owl :partOf*	*owl :partOf*	0,8679	*owl :partOf*	oui
plq	*owl :partOf*	*owl :partOf*	0,7746	*owl :partOf*	oui
PL44	*owl :partOf*	*owl :partOf*	0,5854	*owl :partOf*	oui
caq	*owl :partOf*	*owl :partOf*	0,8889	*owl :partOf*	oui
Pl59	*owl :partOf*	*owl :partOf*	0,2941	*none*	non
Québec	*owl :hasPart*	*owl :hasPart*	0,0352	*none*	oui
PKP	*owl :partOf*	*owl :partOf*	0,2333	*none*	non
cdnpoli	*owl :hasPart*	*owl :hasPart*	0,0049	*none*	oui

L'adéquation entre les choix des évaluateurs et de l'algorithme s'est avérée élevée. En effet, les choix convergent à 12 occasions sur 14 ce qui est exactement le même score que l'on obtient lorsqu'on compare les deux évaluateurs entre eux puisqu'eux aussi ont conclu à des prédicats différents à deux reprises.

Toutefois, même si la performance de l'algorithme est satisfaisante, nous avons tenté

de comprendre ce qui l'avait empêché de conclure que #PKP, possiblement le politicien le plus médiatisé au Québec à cette période, et #Pl59, un projet de loi de l'Assemblée nationale, étaient des *partieDe* la discussion à propos de la politique québécoise.

Pour tenter de comprendre cette décision de l'algorithme, nous avons porté notre analyse sur l'ensemble des inférences produites durant la période de cueillette qui dura un peu plus d'un mois. Notre objectif était d'observer si l'algorithme avait été constant dans ses décisions et si les calculs d'indicateurs sémantiques de cooccurrence inversée l'avaient toujours amené à conclure à l'absence de relation *owl:partOf* pour ces deux mots-clics.

Nous avons donc compilé l'ensemble des calculs d'indicateurs sémantiques de cooccurrence inversée et les avons placés dans deux diagrammes distincts pour chacun des mots-clics (#PKP et #Pl59).

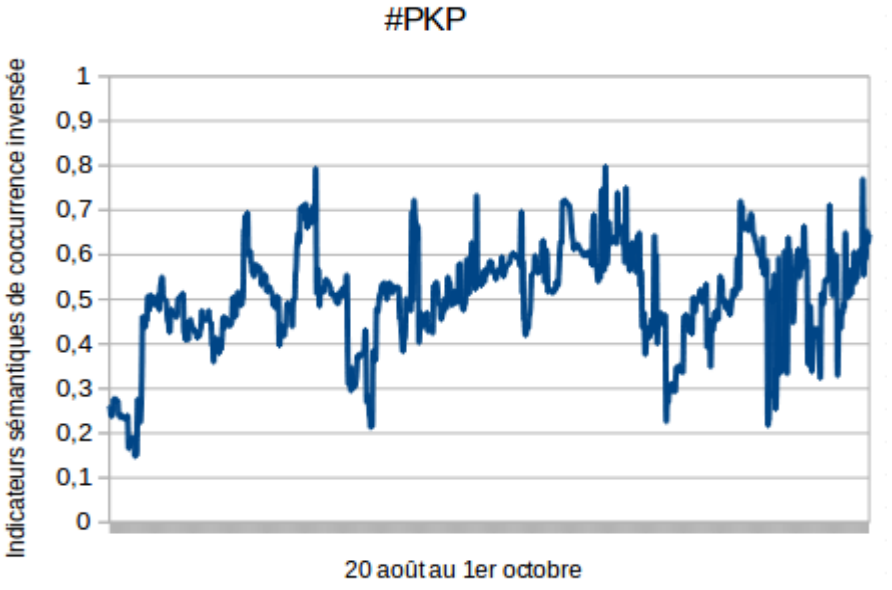


Figure 6.5: Évolution de l'indicateur de cooccurrence inversée (#PKP)

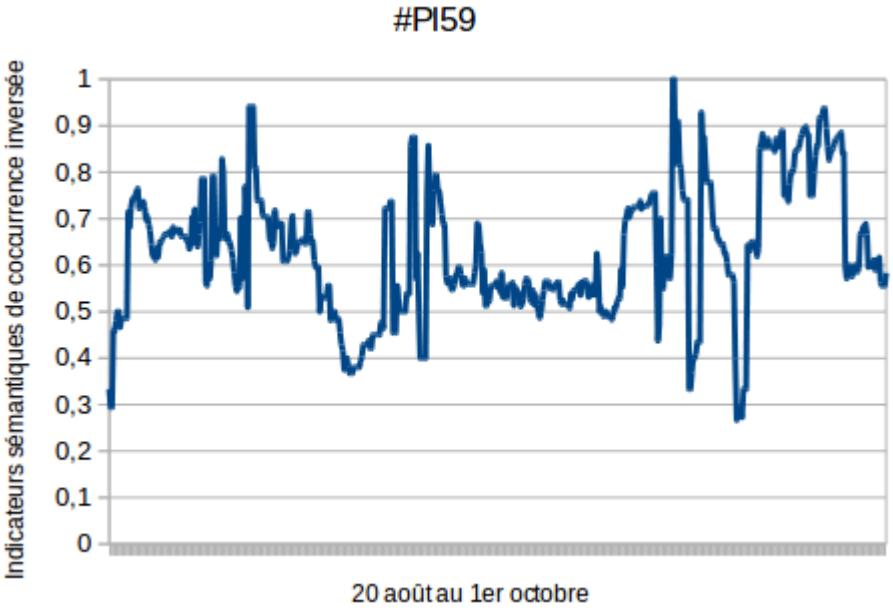


Figure 6.6: Évolution de l'indicateur de cooccurrence inversée (#Pl59)

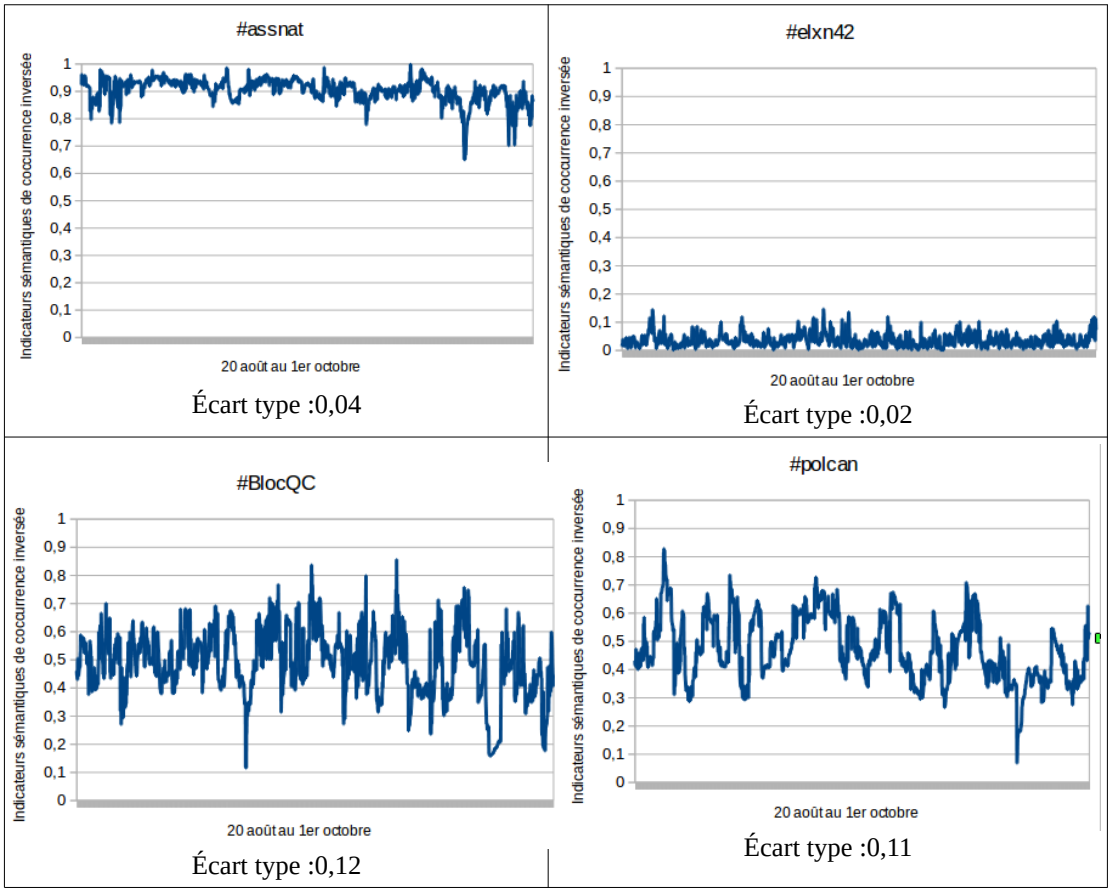
Ces deux diagrammes produisent un résultat quelque peu étonnant. En effet, bien qu'on s'attende à une fluctuation dans le temps de l'indicateur sémantique de cooccurrence inversée, les écarts que l'on peut observer semblent à priori d'une magnitude trop élevée ou du moins beaucoup plus importante que ce que l'on pouvait escompter. À la question qui était de savoir si les mots-clics #PKP et #Pl59 avaient été considérés comme faisant *partieDe* la discussion à propos de la politique québécoise, la réponse est positive. Ces deux mots-clics ont effectivement été inclus à la requête en cours plus souvent qu'ils n'ont été exclus, mais dans l'instantané que nous avons pris le 20 aout à 17h30, ils ont été exclus et ils l'ont été ensuite à plusieurs reprises, car leur indicateur sémantique de cooccurrence inversée a chuté en deçà de 0,5 et parfois de façon drastique. Pourquoi?

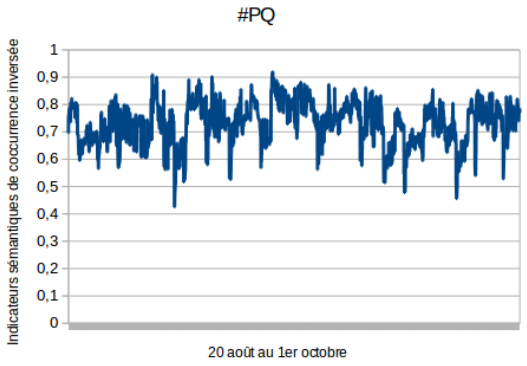
Cette question soulevée par ces deux inadéquations entre l'algorithme et les évaluateurs nous a amené à remettre en cause les 12 autres cas où il y eut adéquation.

Ces hauts niveaux de fluctuation dans les calculs de l'algorithme étaient-ils présents également pour les autres mots-clics? L'excellente adéquation entre les choix de l'algorithme et ceux des évaluateurs était-elle le fruit d'un heureux hasard?

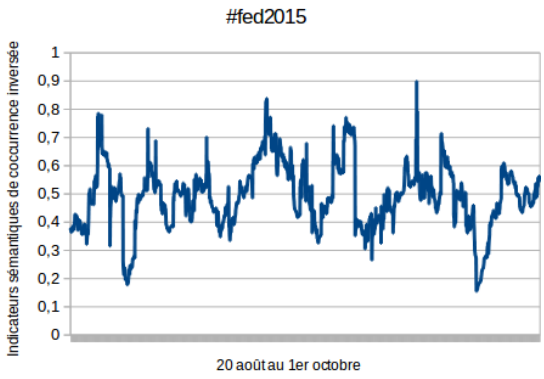
Pour répondre à ces interrogations, nous avons entrepris de compiler tous les calculs d'indicateurs sémantiques de cooccurrence inversée pour les 12 autres mots-clics du champ lexical et de refaire des diagrammes comme nous l'avions fait pour les mots-clics #PKP et Pl59. Pour mesurer le degré de fluctuation de ces calculs dans le temps, nous avons également calculé un écart type.

Tableau 6.11: Compilations d'indicateurs sémantiques de cooccurrence inversée pour le champ lexical de la politique québécoise

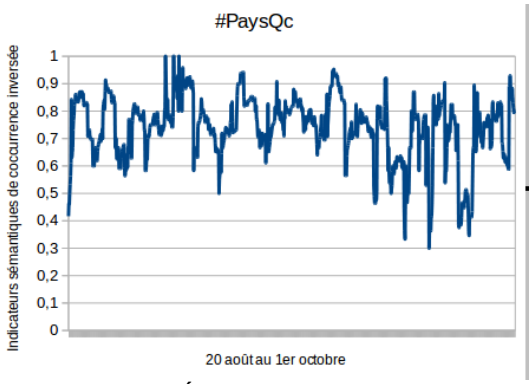




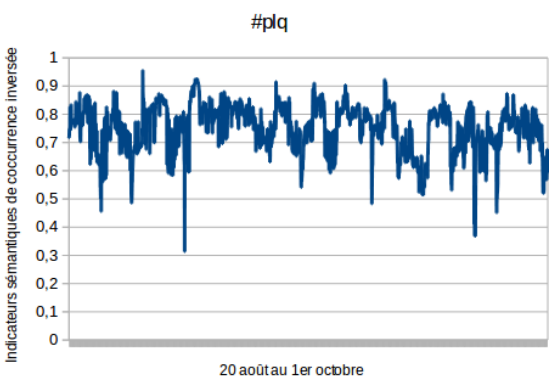
Écart type :0,08



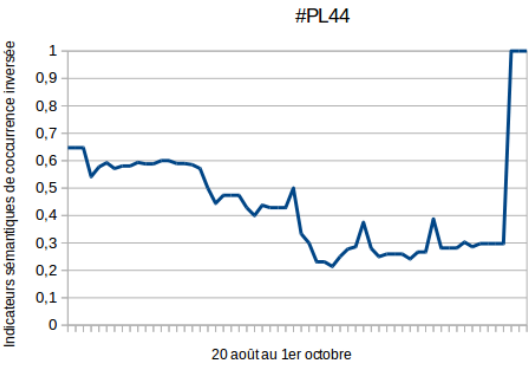
Écart type :0,11



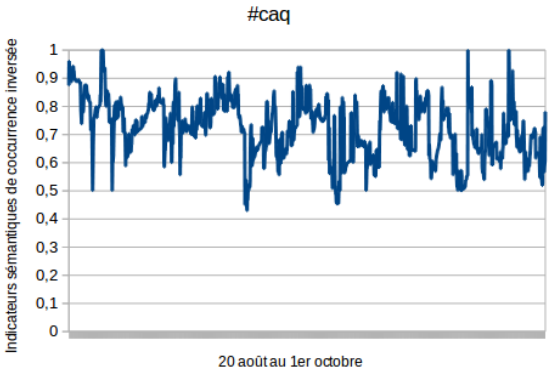
Écart type :0,12



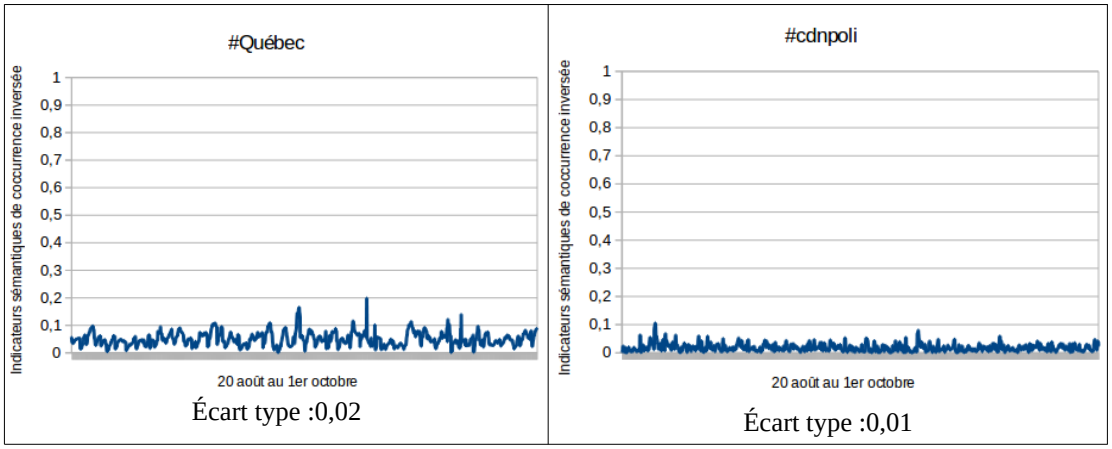
Écart type :0,08



Écart type :0,19



Écart type :0,10



Écart type :0,02

Écart type :0,01

Cette série de diagrammes démontre que d'autres mots-clics à l'intérieur du champ lexical affichent de fortes fluctuations quant à leurs calculs d'indicateurs sémantiques de cooccurrence inversée, alors que certains affichent des indicateurs très stables. Cette fluctuation se constate au premier coup d'œil en observant les courbes de nos diagrammes, mais nous l'avons également quantifiée par le calcul d'un écart type. Pour certains mots-clics, l'écart type est relativement faible, 0,02 pour le mot-clic #elxn42, alors que pour d'autres, il s'avère beaucoup plus élevé comme c'est le cas pour le mot-clic #PL44 qui a un écart type de 0,19.

On comprendra que certains de ceux-ci soient plus susceptibles de présenter des écarts substantiels lorsqu'ils représentent des thématiques fluctuantes dans l'actualité ou encore lorsque les usagers de Twitter les utilisent dans plusieurs contextes différents (entropie), mais, encore une fois, ces fluctuations nous semblent en décalage par rapport à la fluctuation attendue d'une thématique d'actualité sur Twitter.

Par ailleurs, nous savons que tout raisonnement inductif est susceptible de produire des inférences plus fiables et plus stables lorsque les données sur lesquelles sont appliquées ces inférences atteignent un certain volume. Nous avons cherché à comprendre sur quelles bases les calculs des indicateurs sémantiques de cooccurrence

inversée avaient été faits en calculant la moyenne de gazouillis utilisée pour effectuer ces calculs. L'hypothèse que nous souhaitons tester est le lien possible entre l'écart type observé dans les calculs des indicateurs et le nombre de gazouillis disponibles dans les échantillons pour effectuer ces inférences. Le tableau suivant met côte à côte ces deux variables.

Tableau 6.12: Gazouillis disponibles pour effectuer les calculs « vs » écarts types

Mot-clic	Moyenne du nombre de gazouillis disponibles pour effectuer les calculs d'indicateurs sémantiques de cooccurrence inversée	Écarts types observés dans les calculs d'indicateurs sémantiques de cooccurrence inversée
assnat	280	0,04
elxn42	389	0,02
BlocQC	91	0,12
polcan	331	0,11
pq	128	0,08
fed2015	185	0,11
PaysQc	38	0,12
plq	141	0,08
PL44	26	0,19
caq	68	0,10
Pl59	39	0,14
Québec	433	0,02
PKP	69	0,11
cdnpoli	378	0,01

En observant ce tableau, nous constatons l'effet que la limite de la *Search* API produit. L'écart type est inversement proportionnel au nombre de gazouillis disponibles pour effectuer les calculs d'indicateurs. Cette corrélation a été mesurée à -0,82, soit une corrélation très forte. Pour les deux mots-clics qui ne présentaient pas

d’adéquation avec les décisions des évaluateurs et de l'algorithme, #PKP et #Pl59, la moyenne de gazouillis disponibles pour appliquer nos calculs était respectivement de 69 et de 39. Ils présentent pour cette raison des écarts types qui reflètent ces faibles moyennes.

La cause principale de la fluctuation des inférences de l'algorithme pour des mots-clics comme #PKP et Pl59 est maintenant identifiée et elle est fortement liée à une limite technique imposée par le service même de Twitter. Elle ne remet en cause ni les fondements théoriques ni la méthodologie employée puisque lorsque le nombre de gazouillis disponibles pour faire des inférences est suffisamment élevé, l'algorithme produit des calculs d'indicateurs sémantiques de cooccurrence qui sont stables et en adéquation avec les décisions des évaluateurs. Néanmoins, sur des corpus de niche, comme l'est la politique québécoise par rapport à la twittosphère, il s'agit là d'une contrainte qui offre une opportunité d'amélioration.

Au moins trois avenues sont envisageables pour aider à réduire les écarts types observés sur les calculs de mots-clics à faible volume. Deux de celles-ci seront esquissées dans la section réservée aux limites (p.155) alors que la troisième a été mise en place. Elle ne consiste pas à augmenter le nombre de gazouillis sur lesquels les inférences sont faites, mais plutôt à limiter l'effet des fluctuations. En effet, comme nos calculs sont refaits toutes les 15 minutes, les fluctuations observées ont pour effet de faire basculer un mot-clic rapidement d'une relation *partieDe* à une relation qui ne fait pas *partieDe.* L'avenue que nous avons explorée et qui s'avère prometteuse est de constituer une moyenne des calculs d'indicateurs sémantiques de cooccurrence inversée. De cette façon, les fluctuations observées sont réduites et on obtient une plus grande stabilité dans les inférences. L'intervalle à l'essai pour le calcul des moyennes est sur une période de 24 heures et produit de bons résultats. Il aplanit efficacement les distorsions observables lorsque les écarts types sont trop grands. Néanmoins, il réduit quelque peu la réactivité de l'algorithme lorsqu'un mot-

clic ne mérite plus d'être inclus à la requête en cours.

6.4.2.3 Cohérence du corpus

Tout comme pour le corpus portant sur la politique canadienne, il s'agit maintenant de vérifier si les gazouillis recueillis avec les mots-clics ajoutés par l'algorithme à la requête en cours correspondent à la thématique. Pour ce corpus recueilli du 20 aout au 1er octobre 2015, une cueillette utilisant exclusivement le mot-clic thématique #polQc aurait recueilli 151 954 gazouillis contre 266 116 avec l'application de nos algorithmes. 114 162 gazouillis supplémentaires ont donc été ajoutés au corpus soit une augmentation de 75 %.

Les évaluateurs ont estimé que les gazouillis de l'échantillon contenant le mot-clic thématique et ceux de l'échantillon contenant des mots-clics sélectionnés par l'algorithme présentaient des niveaux de cohérence semblables par rapport à la thématique (voir Tableau 6.4: Validation de la cohérence des gazouillis p.107). Regardons de plus près ces gazouillis jugés hors thématiques.

Tout comme pour le corpus sur la politique canadienne, les gazouillis jugés hors thématique dans les deux échantillons sont de nature très différente. Dans l'échantillon contenant le mot-clic #polQc, la vaste majorité des gazouillis jugés hors thématiques étaient en relation avec la politique fédérale. En voici quelques exemples :

- RT @lepaisan: Jamais autant de députés fédéraux n'ont ignoré les intérêts de leurs propres électeurs... #polqc #fed2015 #elexn42
- RT @chevymo: #fed2015 #cdnpoli #polqc ... READ the NDP manual ... their 'plan' is detailed in it ... the consequences horrendous
- RT @ThDarkJedi_: SCOOP: Le fou furieux Conservateur change d'idée!!! #polqc #polcan #cdnpoli #npd #bloc #pcc #plc #pq #caq #qs #plq

- Imagine si Mulcair avait fondé Sun News et financé Harper? MDR @gobeillades #POLQC #PAYSQC #BLOCQC #ELXN42 http://t.co/dL1qESuVik

Dans ces quelques cas, on remarque que certains utilisateurs de Twitter marient à la fois des mots-clics relatifs à la politique québécoise et à la politique canadienne. La période de cueillette ayant eu lieu durant des élections fédérales, il est possible que cela ait eu une influence.

Les gazouillis jugés hors thématiques par les évaluateurs dans l'échantillon recueilli à partir de mots-clics sélectionnés par l'algorithme ont été très différents. La plupart des gazouillis rejetés l'ont été parce qu'ils traitaient de sujets d'intérêt pour les amateurs de foot d'Amérique latine! En effet, dans cette partie du monde le mot-clic #PQ est utilisé pour désigner une équipe de football alors qu'au Québec il est utilisé en référence au Parti Québécois. Il s'agit d'un cas intéressant où nous aurions pu filtrer les gazouillis en fonction de la langue, mais nous aurions exclu d'office les gazouillis de la communauté hispanophone du Québec. La meilleure alternative dans un cas comme celui-ci est de maintenir la cueillette telle quelle et de filtrer les gazouillis après coup dans la base de données.

6.5 Synthèse des résultats

En guise de conclusion à ce chapitre, soulignons d’abord la validation d’une forte proportion de relations de type *partie-tout* parmi les mots-clics présents dans les champs lexicaux formés par l’application de la loi de Zipf-Mandelbrot. En effet, sur l’ensemble des champs lexicaux analysés, près de 2 mots-clics sur 3 présentent des relations *partie-tout* avec le mot-clic thématique. Qui plus est, cette présence de relation de type *partie-tout* gagne en importance de façon proportionnelle au niveau de cooccurrence, ce qui confirme à la fois les notions avancées dans le cadre théorique et leurs implémentations dans le prototype à l’effet qu’une forte

cooccurrence entre mots-clics augmente la fréquence de relations de type *partie-tout* entre eux.

Les performances des calculs d'indicateurs sémantiques de cooccurrence inversée se sont également avérées satisfaisantes. Ces calculs arrivent efficacement à discriminer les mots-clics faisant partie d'une discussion de ceux qui ne le font pas. En effet, lorsqu'on tient en compte des prédicats identiques attribués par les évaluateurs et qu'on les compare à ceux déterminés par l'algorithme, on constate que l'algorithme arrive à prendre les mêmes décisions que des humains dans 98 % des cas; le 2 % de différence étant attribuable à une certaine « prudence » de l'algorithme. En effet, dans les rares cas où n'y a pas adéquation entre les décisions des évaluateurs et celles de l'algorithme, la décision de l'algorithme fut de ne pas inclure le mot-clic en tant que sous-thématique. Ainsi, aucun mot-clic s'étant vu attribué un prédicat *owl:hasPart* ou *None* par les évaluateurs n'a été sélectionné par l'algorithme. Autrement dit, notre prototype ne produit pas de faux positifs, ce qui favorise le maintien d'une cohérence des gazouillis recueillis par rapport à une thématique donnée.

À propos justement de cet aspect crucial qui est de s'assurer que le niveau de cohérence vis-à-vis de la thématique est maintenu même en ajoutant un grand nombre de nouveaux mots-clics à la requête en cours, les résultats obtenus sont concluants. En effet, dans les deux corpus analysés plus en profondeur à propos de la politique canadienne et québécoise, les mots-clics ajoutés automatiquement par notre prototype à la requête en cours ont permis de recueillir des gazouillis tout aussi cohérents avec la thématique que ceux recueillis par le mot-clic thématique lui-même. De plus, cette requête en cours, dont l'objectif est de rendre plus exhaustif le suivi d'une discussion à propos d'une thématique, a permis d'augmenter le nombre de gazouillis de 48 % dans le cas de la thématique sur la politique canadienne et de 75% pour la thématique à propos de la politique québécoise.

L’analyse de ces résultats produits par l’implémentation computationnelle de notre cadre théorique répondent à notre avis de façon probante à notre questionnement de départ à savoir comment réaliser une automatisation de la découverte de mots-clics nécessaires au suivi d'une discussion ayant cours dans l'espace public à propos d'une thématique donnée.

CHAPITRE 7 :
EXEMPLES D’APPLICATIONS

Nous présenterons d’abord dans ce chapitre diverses applications de notre prototype à des recherches dans le champ des études s’intéressant à la circulation de l’information et aux études des pratiques des usagers sur Twitter. Nous poursuivrons ensuite avec un exemple de recherche comparant les méthodes couramment utilisées dans les humanités numériques avec celles que notre prototype permet désormais. Finalement, nous démontrerons comment l’utilisation de notre prototype peut être étendue à d’autres dispositifs sociotechniques.

7.1 Application de notre prototype pour la recherche

Cette thèse s'est intéressée à la problématique du suivi de discussions sur Twitter en amont, c'est-à-dire à l'étape primordiale de la cueillette. Toute étude de discours sur Twitter ou sur d'autres dispositifs sociotechniques dépend à la base du corpus sur lequel seront effectuées lesdites analyses. Le prototype issu de notre recherche a démontré qu'il pouvait améliorer sensiblement qualitativement et quantitativement la collecte de données.

Au fil des ans, la popularité de Twitter a généré un grand intérêt dans la communauté scientifique qui a produit une littérature abondante sur ce dispositif socionumérique. Un de ces courants de recherche s'est intéressé à la structure des réseaux interpersonnels. En observant les réseaux d'abonnés, Java et al. (2007) ont pu observer une grande variété dans les usages des utilisateurs. En effet, le même utilisateur peut jouer de multiples rôles à l'intérieur de différentes communautés montrant une certaine analogie avec ses comportements hors-ligne. De leur côté, Huberman et al. (2009) ont analysé les interactions sociales et découvert que le plus grand motivateur d'utilisation de la plateforme était en fait un petit réseau d'abonnés fait d'amis dans la « vraie vie » ou de collaborateurs en ligne et que les autres relations avec les nombreux abonnés liés à un compte étaient en fait la plupart du temps sans intérêt.

Un autre courant de recherche s'est intéressé à « l'influence » de certains utilisateurs et à leur importance dans la circulation de l'information. Pour ce faire, Cha et al. (2010) ont comparé trois mesures d'influence distinctes soit le nombre d'abonnés, le nombre de retweets et les mentions des utilisateurs. Ils ont ainsi pu observer que le nombre d'abonnés ne se traduisait pas nécessairement en terme « d'influence » et les utilisateurs ayant un grand nombre d'abonnés ne voyaient pas nécessairement leurs gazouillis plus souvent relayés ou leur nom plus souvent mentionné dans les gazouillis des autres utilisateurs. Dans le même sens, Romero et al. (2010) ont montré que la corrélation entre la popularité et l'influence d'un utilisateur était plutôt faible, car la plupart des utilisateurs sont passifs et retweetent rarement l'information qu'ils lisent. C'est d'ailleurs là une limite de ces études, car elles ne peuvent que mesurer les retweets et non le nombre de personnes ayant lu un message puisque cette donnée ne peut être capturée par Twitter. Toujours dans le domaine de la circulation de l'information, Yang et al. (2010) ont créé un modèle d'analyse permettant de mesurer la vitesse et l'ampleur de la diffusion d'une information. Ils ont pu établir que la

propagation d'un message est plutôt liée aux propriétés du gazouillis lui-même qu'à l’auteur de ce dernier.

D'autres études comme celle de Asur et Huberman (2010) se sont penchées sur le potentiel prédictif de l'analyse de gazouillis publiés par l'ensemble des utilisateurs afin de prédire le succès d'un film au boxoffice. Leur modèle, très simple, basé sur l'adage « parlez-en en bien, parlez-en en mal, mais parlez-en! » a montré une corrélation très forte entre le nombre de gazouillis à propos d'un film et l'achat de billets de cinéma. Plusieurs autres recherches (Tumasjan et al., 2010; Lampos et Cristianin, 2010; Antweiler et Frank, 2004; Gilbert et Karahalios, 2010) ont également réalisé des analyses prédictives liant des données et des métadonnées de corpus de gazouillis à des phénomènes boursiers, des épidémies, des nominations aux Oscars. Le spectre d'utilisation des données recueillies à partir des réseaux sociotechniques est très large et continuera de s'élargir, car il devient un reflet de plus en plus riche de l'activité humaine.

Dans la plupart des recherches que nous venons de mentionner, notre prototype de suivi de discussion aurait pu être employé lors de la cueillette. Nous allons maintenant illustrer deux exemples d’utilisation qui pourraient s'appliquer dans des contextes de recherche. Le premier exemple est inspiré d'une recherche à laquelle j'ai participé et qui s'est intéressé à l'usage de Twitter par les minorités francophones au Canada anglais (Rocheleau et Millette, 2015), le second explorera les possibilités de modélisation des données recueillies afin de créer un modèle ontologique mettant en lien des connaissances acquises par notre algorithme dans Twitter avec d'autres réseaux sociotechniques.

7.2 Méthodologie de recherche sur les réseaux socio-techniques

Cet exemple d'utilisation est celui dans lequel la problématique de cette thèse nous est d'abord apparue. Il s'agit d'une recherche menée par Mélanie Millette (2015) pour sa

thèse en communication et durant laquelle j'ai élaboré une partie du cadre théorique de ma thèse. De ce fait, le développement de notre prototype ayant été effectué après le forage de données qu'a mené Millette en 2012-2013, plusieurs des outils n'ont pu être utilisés dans le cadre de sa recherche. Dans un premier temps, nous mettrons en parallèle les méthodes utilisées par Millette, donc avant le développement de notre prototype, et celles qu'il serait désormais possible d'envisager pour de nouvelles recherches. Dans un deuxième temps, nous résumerons certaines des analyses issues de notre collaboration illustrant comment un corpus de gazouillis peut être approché dans un contexte d'étude sur l'espace public.

7.2.1 Comparatifs méthodologiques

La recherche de Millette débute comme plusieurs recherches en sciences sociales portant sur les réseaux sociotechniques, c'est-à-dire par une phase exploratoire qui fait suite à l'identification d'une problématique. Cette phase exploratoire est souvent constituée de plusieurs éléments, mais, dans le cas de travaux sur les réseaux sociotechniques, une veille en ligne est l'un des éléments que l'on retrouve presque toujours. Cette recherche n'en diffère pas :

> « Après deux mois de veille quasi quotidienne, nous avions repéré plusieurs mots-clics, dont cinq très populaires : #FrCan (Francophonie canadienne), #OnFr (Ontario francophone), #AbFr (Alberta francophone), #languesofficielles, #FANE (Fédération Acadienne de la Nouvelle-Écosse). »
> (Ibid. p. 116)

Observons cette citation et faisons maintenant un lien avec nos concepts théoriques mis de l'avant jusqu'à maintenant. On peut y lire « nous avions repéré plusieurs mots-clics, dont cinq très populaires ». Au cours de ses deux mois de veille, la chercheure a lu et pris en note des mots-clics qui apparaissaient dans les gazouillis de la communauté. Elle a, informellement, créé une forme de folksonomie faite de mots-clics appartenant à sa thématique de recherche. Parmi les mots-clics repérés, cinq ont

été sélectionnés pour leur popularité. Il s'agit du même processus auquel nous avons déjà fait référence, c'est-à-dire que la chercheure a procédé à une sélection des mots-clics de sa folksonomie informelle pour conserver ceux le plus souvent utilisés. Notre algorithme, en procédant au regroupement des mots-clics puis en utilisant la loi de Zipf-Mandelbrot pour extraire les mots-clics présentant les plus fortes cooccurrences, réalise une tâche analogue, mais automatisée, par la création d'un champ lexical à partir d'une folksonomie. En effet les cinq mots-clics retenus portent tous les caractéristiques d'un champ lexical autour de la thématique de la francophonie canadienne.

Portons maintenant une attention aux cinq mots-clics retenus. Pourrait-on en identifier un comme étant un mot-clic thématique? La thèse de Millette porte sur les minorités francophones au Canada. Or, parmi les mots-clics sélectionnés, on retrouve #frcan (Francophonie canadienne) qui pourrait correspondre au mot-clic thématique. Millette en parle d'ailleurs en ces termes : « Nous avons identifié ce mot-clic comme étant un « méta-mot-clic », c'est-à-dire le mot-clic thématique représentant le niveau le plus englobant, la hiérarchie la plus haute, dans un ensemble de messages ayant une parenté thématique. » (Ibid., p.144). Si on regarde les quatre mots-clics restants, il serait facile d'y voir des relations *partieDe* pour au moins trois de ceux-ci soit #onfr (Ontario francophone), #abfr (Alberta francophone), #fane (Fédération Acadienne de la Nouvelle-Écosse) alors que le cas de #languesofficielles nécessiterait une analyse plus approfondie.

En empruntant une méthode et un vocable différents, on observe que tous les concepts clés de notre approche peuvent trouver leur équivalent dans une méthodologie de recherche en sciences sociales portant sur l'étude des réseaux sociotechniques : folksonomie, champ lexical, mot-clic thématique et relations *partieDe.* Toutefois, une méthode recourant à une veille menée par un chercheur nécessite beaucoup de son temps et ne peut, pour des raisons humaines (il faut bien

dormir!) être exhaustive. De plus, comme on l'a déjà mentionné dans la section Relations sémantiques et temporalité (p.67), sa veille court le risque de devenir désuète rapidement.

En employant notre prototype, comment pourrait-on faire différemment pour libérer le chercheur de la longue phase exploratoire et de sa veille afin que plus de temps puisse être consacré à l'analyse?

Notre prototype de cueillette et de suivi de discussion a pour point de départ un mot-clic thématique. Ce mot-clic peut être établi de différentes manières. Parfois, il est connu du chercheur avant qu'il entame sa cueillette, mais dans le cas présent le mot-clic thématique #frcan a été découvert à postériori, donc nous ne le tiendrons pas pour acquis et emploierons la méthode décrite dans notre description d'un mot-clic thématique (p.29). Cette méthode consiste à identifier un ou plusieurs leadeurs dans une communauté afin d'analyser les mots-clics qu'ils utilisent le plus souvent. Un tel leadeur existe dans la communauté franco-canadienne et a d'ailleurs été identifié par l'auteure :

> « La FCFA est un organisme particulièrement central et englobant dans la constellation des organisations franco-canadiennes, ne serait-ce que par son mandat pancanadien et son statut de fédération. La FCFA occupe ainsi une posture privilégiée pour quiconque s'intéresse aux francophonies canadiennes et acadiennes... »
>
> (*Ibid.* p. 161)

En cherchant sur Twitter, on découvre que la FCFA détient un compte Twitter et qu'on pourra procéder à l'extraction du mot-clic thématique à partir de ce dernier. Une fois le mot-clic thématique inséré dans le prototype, le module de cueillette de gazouillis puis le module de sélection de mots-clics se mettent en marche (p.89). La découverte de ce qui a été décrit comme des mots-clics populaires (champ lexical) et la découverte de mots-clics jugés pertinents d'être ajoutés à la cueillette (requête en

cours) sont complètement automatisées. De plus, grâce à l'interface graphique, le chercheur est à même de consulter les inférences des algorithmes ainsi que leur historique.

Nul doute que les deux mois de la phase exploratoire ont permis de faire plus qu'une identification de mots-clics pertinents pour la recherche, mais en automatisant cette tâche, on libère du temps dont le chercheur peut faire usage pour accomplir d'autres tâches.

7.3 Extension à d'autres outils sociotechniques

Nous nous sommes concentrés sur le réseau sociotechnique de Twitter afin de mieux suivre des discussions d'intérêt public, mais nous croyons que notre modèle théorique et les méthodologies mises de l'avant pourraient être adaptés à tout autre réseau sociotechnique reposant sur l'étiquetage social. De plus, nous estimons possible d'étendre à d'autres plateformes les d'inférences effectuées à partir de données recueillies sur Twitter par notre prototype.

Nous allons maintenant présenter un autre exemple d'application dans lequel, en plus de la cueillette de gazouillis effectuée par notre prototype, nous explorerons une nouvelle avenue pour des chercheurs intéressés à effectuer un forage de données multi-plateforme à propos d'une thématique d'intérêt public. Nous ne nous mettrons pas dans une posture de recherche où un accent particulier serait mis sur une plateforme en particulier (Facebook, Youtube, Twitter, etc.) ou sur un type de contenu (blogue, sites web de médias, images), mais nous tenterons plutôt d'esquisser une solution pour un type de recherche qui s'intéresserait à l'analyse globale d'une thématique d'intérêt public sur le web.

Nous ajouterons donc aux messages de Twitter, ceux de Facebook, des images de Flickr, des vidéos de Youtube ainsi que le contenu de plusieurs pages web pour la

plupart issues de sites web de médias.

Notre prototype sera utilisé à la fois pour la cueillette de gazouillis et comme moteur pour réaliser le forage de textes à partir des API d'autres sources. Pour simplifier notre démonstration, nous utiliserons comme exemple un corpus que l'on a bien décrit jusqu'à maintenant soit celui de la politique canadienne.

Les raisonnements effectués par nos algorithmes nous permettent de passer d'une folksonomie à un champ lexical puis à la découverte de mots-clics représentant des sous-thématiques. L'hypothèse que nous testerons dans cette mise en situation consiste à vérifier si les mots-clics utilisés sur Twitter se retrouvent sous forme de mots-clés dans d'autres dispositifs sociotechniques. Prenons par exemple le mot-clic #StopHarper, identifié comme sous-thématique de #cdnpoli et vérifions s'il est utilisé sur Flickr, Youtube ou Facebook et, surtout, s'il est utilisé pour étiqueter des contenus correspondant au mot-clic et à sa thématique. Tous ces dispositifs sociotechniques disposent d'API, mais il sera plus facile d'illustrer les résultats de notre vérification par l'utilisation de captures d'écran réalisées à même leurs sites web respectifs.

Commençons d'abord par lancer une recherche avec le mot-clé « StopHarper » dans l'outil de recherche de Filckr.

Figure 7.1: Images récupérées de Flickr pour le mot-clé "StopHarper"

Source : Flickr (https://www.flickr.com/)

Les vignettes ci-dessus sont une copie d'écran des premières images retournées par Flickr pour le mot-clé « StopHarper ». Comme on peut le constater, le mot-clic #StopHarper peut être utilisé sur cette plateforme et donner des images intéressantes pour la thématique.

Voyons maintenant ce qu'il en est si on lance une recherche sur Youtube toujours avec le mot-clé « StopHarper ».

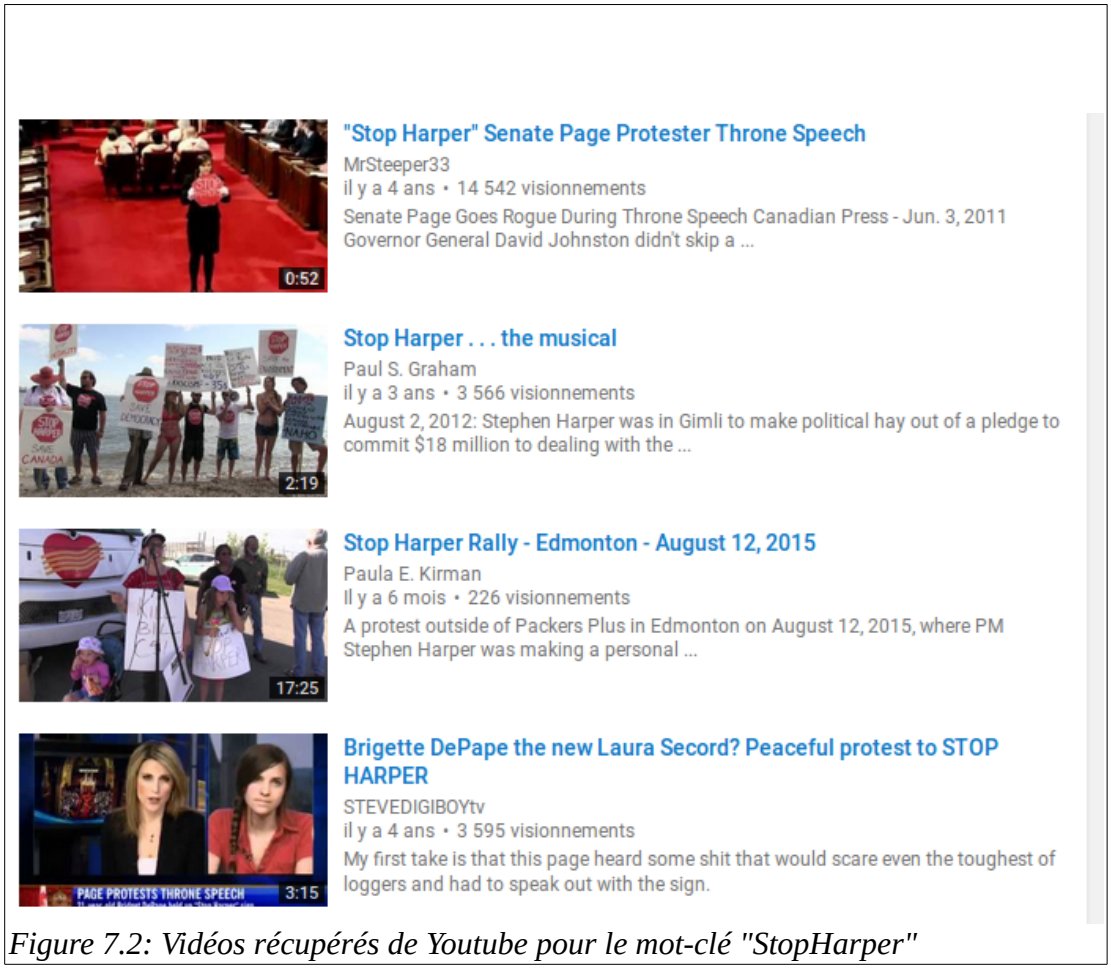


Figure 7.2: Vidéos récupérés de Youtube pour le mot-clé "StopHarper"
Source : Youtube (https://www.youtube.com/)

Encore une fois, les contenus associés au mots-clics #StopHarper identifié comme sous-thématique dans Twitter sont similaires à ceux que l'on retrouve sur Youtube en utilisant le mot-clé StopHarper. La situation s'avère identique lorsqu'on répète l'opération sur Facebook :

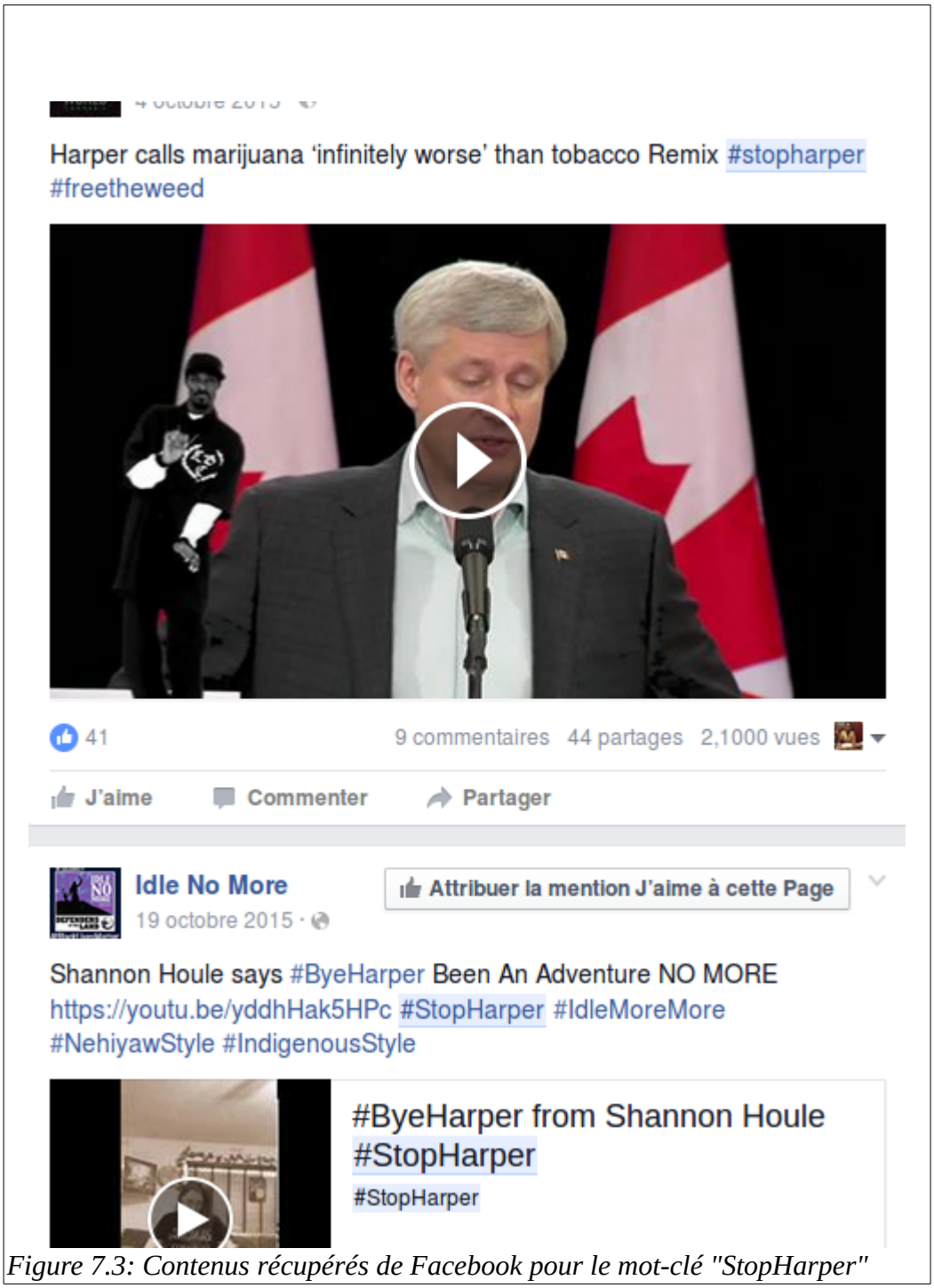


Figure 7.3: Contenus récupérés de Facebook pour le mot-clé "StopHarper"
Source : Facebook (https://www.facebook.com/)

À la lumière de ces premiers résultats qui semblent attester l'hypothèse selon laquelle des mots-clics identifiés par notre algorithme comme faisant partie d'une thématique peuvent ensuite être utilisés pour lancer des requêtes vers d'autres dispositifs sociotechniques, il est donc possible d'envisager une automatisation de la cueillette de contenus à propos d'une même thématique sur de multiples plateformes de façon simultanée.

Le premier avantage d'une telle approche est bien sûr d'obtenir un corpus aux contenus plus variés qui peut s'adapter à divers objets d'étude comme l'analyse d'images, l'analyse de discours publics, la prise en compte de contenus générés par les utilisateurs, l'analyse de la circulation de contenus d'une plateforme à l'autre, l'étude de la redondance de l'information, l'observation de pratiques participatives sur ces plateformes, etc.

Toutefois, le plus grand bénéfice d'une telle approche en termes de richesse sémantique est qu'elle nous permet d'imaginer des scénarios pour contourner les limitations de Twitter dont nous avons déjà fait mention soit la limite de 140 caractères, l'emploi d'abréviations et de néologismes ainsi qu'un nombre important de fautes d'orthographe. Ces limitations nous empêchaient d'utiliser une foule de ressources linguistiques ou d'outils d'informatique cognitive, notamment ceux que l'on retrouve du côté du traitement automatique du langage naturel (p.12).

Avant de traiter de l'utilisation de ces outils, notons au passage une fonction qu'on retrouve sur la plupart de ces plateformes soit celle de pouvoir inclure des liens vers des pages web. Ces pages web sont souvent celles de médias et proposent des contenus plus substantiels et ayant une meilleure qualité orthographique. Prenons par exemple le gazouillis suivant contenant le mot-clic #canpoli :

- RT @georgiastraight: Stephen Harper stealthily lays his imprint over the Canadian judiciary **http://t.co/lr7yzNqfQl** #canpoli

On remarque qu'il fait référence à une URL raccourcie. Des outils de moissonnage web avec lesquels nous travaillons pour d'autres projets permettent de suivre ces URL et d'en extraire le contenu. En suivant l'URL raccourcie « http://t.co/lr7yzNqfQl » on découvre qu'il s'agit en fait d'un article publié par « The Georgia Straight », le plus important hebdomadaire de Colombie-Britannique. Comme nous venons de le mentionner, un article d'un média ouvre une panoplie d'options pour l'analyse automatisée de contenu. En voici quelques-unes qui pourraient être envisagées.

Un article média permet en premier lieu l'identification d'entités sémantiques telles que des noms de personnes, des lieux, des organisations. Plusieurs outils existent pour accomplir ce type de tâches, nous utilisons ici celui de *Alchemy*[11]. Après analyse, cette API retourne un fichier JSON avec une liste de personnes dont il est question dans l'article soit Stephen Harper, Rona Ambrose, Charlie Smith, Sean Fine, Jason Kenney, Peter MacKay, James Moore et Joe Oliver que l'outil a identifié comme étant pour la plupart des politiciens. Il réussit à faire ces inférences d'abord de façon simple par l'identification de deux mots consécutifs avec des majuscules, mais également de façon plus « intelligente » par une fouille dans le web de données liées (linked data) notamment en recherchant des correspondances pour ces noms de personnes dans Dbpedia, Freebase et Yago. La possibilité de créer des liens entre un gazouillis, un article et ensuite des entités sémantiques documentées dans le web des données représente un potentiel énorme du point de vue de la construction d'une base de connaissances à propos d'une thématique.

[11] Les données ont été obtenues à partir de l'outil en démonstration à https://alchemy-language-demo.mybluemix.net/

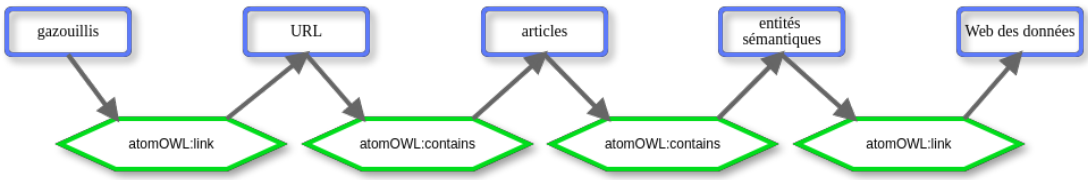


Figure 7.4: Degrés de séparation entre un gazouillis et le web des données

En plus de l'extraction de noms de personnes, l'analyse automatisée de texte permet aussi l'extraction de la thématique de l'article, de concepts liés en plus des métadonnées comme le titre, la date ou l'auteur. Ces informations à propos de contenus partagés via Twitter ou Facebook, une fois bien structurées dans un *triplestore*[12], rendraient possibles des recherches très pointues ainsi que la constitution de sous-échantillons dans un cadre de recherche se focalisant sur un objet bien précis d'une thématique.

Revenons maintenant aux contenus extraits des deux autres plateformes (Youtube et Flickr) tout en gardant en tête ce potentiel de liaison entre les données de notre corpus et le web des données à l'intérieur d'un triplestore. En plus de proposer des types de contenus complémentaires (images et vidéo), ces deux plateformes n'imposent pas de limites quant au nombre de caractères dans la définition d'étiquettes. Ce faisant les utilisateurs emploient, en plus des mots-clics populaires sur Twitter, des mots-clés qui relèvent du vocabulaire usuel, lesquels peuvent, dans certains cas, être reliés à des entités du web des données. Sélectionnons une image et une vidéo qu'une requête avec « StopHarper » aurait recueillie.

[12] Type de base de données conçue pour le stockage de données RDF

Figure 7.5: Vidéo Youtube avec le mot-clé "StopHarper"

Dans cette vidéo recueillie de Youtube avec le mot-clé « StopHarper » dans lequel on peut visionner une manifestation, plusieurs informations intéressantes peuvent être extraites de l'étiquetage social qui fut effectué pour identifier ce contenu. Les mots-clés qu'on y découvre sont les suivants : « Edmonton (City/Town/Village), stephen harper, conservatives, canada, politics, rally, protest ». Aucune abréviation, aucune contraction, aucun mot-valise, seulement des termes définis dans la langue anglaise et qui rendent possible l'utilisation de ressources linguistiques. Dans une optique de récolte de données dans un *triplestore*, il serait dès lors possible d'associer au mot-clic « #StopHarper » - qui rappelons-le est incompréhensible pour une machine puisque comme la plupart des mots-clics il n'est défini dans aucune base de connaissances ou dictionnaire - des noms de personnes (stephen harper), des lieux (Edmonton, canada), une organisation (conservatives) et des concepts (rally, politics, protest) qui eux sont définis dans le web des données.

Du côté de Flickr aussi, les contenus associés à « StopHarper » sont étiquetés à l'aide

Figure 7.6: Image Flickr avec le mot-clé "StopHarper"

de mots-clés pour lesquels des ressources linguistiques existent. En effet, les mots-clés suivants ont été ajoutés à l'image ci-contre : des objets (sign et stop) et un lieu en Colombie-Britannique (mount pleasant). De plus, Flickr offre une fonctionnalité appelée « autotag » qui ajoute des étiquettes suite à une analyse automatisée des images. Suite à cette analyse, les mots-clés suivants ont été ajoutés à l'image : « text », « signboard », « street sign » et « outdoor ». Tout comme

dans le cas de Youtube, ces mots-clés présentent le net avantage sur la plupart des mots-clics de Twitter d'être définis dans des ressources linguistiques et sur le web des données.

L'objectif de cet exemple d'utilisation était de démontrer que les mots-clics représentant des sous-thématiques, fruits des calculs d'indicateurs sémantiques de cooccurrence inversée, peuvent être réutilisés pour recueillir d'autres types de contenus sur d'autres plateformes. On peut ainsi imaginer une cueillette de contenu beaucoup plus large sur une thématique donnée. La plus-value en terme sémantique est également appréciable et pourrait mener à la création d'une base de connaissances où les contenus et les métadonnées des différentes plateformes viendraient enrichir notre compréhension des différentes sous-thématiques. Une autre avenue intéressante de recherche serait de mettre en lien toutes ces informations avec le web de données dans le but d'apporter des définitions aux mots-clics qui enrichiraient des services de définitions de mots-clics comme tagdef.com.

CHAPITRE 8 :
LIMITES ET APPORTS

Dans ce chapitre, nous aborderons les limites de cette recherche et nous relèverons les apports pour les sciences de la communication, les sciences cognitives et les sciences informatiques puis, avant de conclure, nous présenterons quelques pistes de développements futurs.

8.1 Limites

8.1.1 Restrictions de l’accès aux archives de Twitter

Twitter restreint l'accès à ses archives de plusieurs façons. D'une part, il n'est pas possible d'obtenir des gazouillis datant de plus d'une semaine et, d'autre part, seul un faible pourcentage des gazouillis publiés peut être récupéré jusqu'à un maximum de 500 par requête. La limite de retour dans le temps n'a pas d'impact sur notre prototype, car nos requêtes ne s'intéressent guère aux gazouillis publiés il y a plus de 24 heures. Le maximum de 500 gazouillis n'est pas non plus problématique, car, jusqu'à un certain point, les algorithmes réussissent à faire des inférences fiables même avec de plus petits nombres de gazouillis.

Le faible pourcentage de gazouillis retournés par la *Search* API a certainement un impact négatif dans certains cas sur la qualité des calculs d'indicateurs sémantiques de cooccurrence inversée. Cette limite n'a que peu d'impact sur les thématiques qui sont populaires sur Twitter, mais lorsque la thématique est plus marginale et que les mots-clics candidats le sont encore plus, cette limite pose problème, car elle oblige parfois notre algorithme à faire des inférences sur un faible nombre de gazouillis. Nous en avons donné un exemple lors de l'analyse des données du corpus sur la politique québécoise (p.128). Une première solution consistant à calculer une moyenne du calcul des inférences pour éviter une trop grande volatilité des décisions de

l'algorithme a été mise en place pour contourner cette limite, mais deux autres solutions pourraient être envisagées dans le futur.

La première consisterait à recueillir systématiquement tous les gazouillis présents dans le champ lexical dans une deuxième base de données afin de constituer une archive sur laquelle les calculs d'indicateurs de cooccurrence inversée pourraient être effectués. On se retrouverait ainsi à recréer l'archive de Twitter pour les mots-clics qui nous intéressent et nos inférences pourraient être exécutées sur des échantillons complets plutôt que sur les échantillons tronqués par les limites de la *Search API*. Ce serait une solution qui nous rendrait autonomes par rapport à la *Search API* de Twitter, mais qui augmenterait nos besoins en ressources informatiques dûe à la création d'une deuxième base de données qui recueilleraient de forts volumes de gazouillis et à l'envoi de requêtes SQL chargées de receuillir des échantillons précis de gazouillis à intervalles réguliers de 15 minutes. C'est d'ailleurs par souci d'économie de ressources informatiques que la *Search API* ne retourne que des échantillons partiels. Par contre, comme cette limite se fait surtout sentir pour des thématiques qui sont moins populaires au sein de Twitter, un seuil de déclenchement de cette méthode paliative pourrait être mis en place nous évitant ainsi une surcharge de données lors de la cueillette de gazouillis sur des thématiques plus populaires.

Un deuxième moyen d'éliminer cette limite serait le recours aux services tarifés, tel que Gnip.com, qui offrent un accès complet aux archives de Twitter. Toutefois, compte tenu d'autres contraintes propres au financement en recherche universitaire, cette avenue n'a pas été explorée.

8.1.2 Absence des gazouillis sans mots-clics

Une autre limite qu'a pu observer le lecteur est le fait que nous nous sommes concentré uniquement sur les gazouillis contenant des mots-clics alors qu'une très large proportion n'en contiennent pas! Il s'agit d'une balise que nous nous sommes

imposé volontairement, car la thèse repose sur un type particulier de raisonnement inductif appliqué à des folksonomies, donc des ensembles de mots-clics. Il serait possible et tout à fait avisé dans un contexte de recherche de recueillir à la fois des gazouillis contenant un mot-clic comme #Montreal et le mot-clé Montréal lorsqu'il est présent dans des gazouillis sans être précédé du dièse. Toutefois, comme l'ont noté Cecaj et Mamei (2016), il existe parfois des différences sémantiques entre un mot-clic et un mot ordinaire dans Twitter. En effet, dans le corpus qu'ils ont étudié, ces deux chercheurs ont pu observer que dans les gazouillis contenants le mot-clic #Milan, la thématique principale était cette ville d'Italie, alors que lorsque le nom de la ville est employé sans le dièse, il y est plus souvent question d'une des équipes de foot de cette ville, l'AC Milan. L'ajout de mots ordinaires à la requête en cours est donc envisageable, en autant que la possibilité de changement de sens entre un mot-clic et un mot ordinaire soit prise en compte.

8.1.3 Corpus multilingues

Certains des mots-clics thématiques que nous avons employés sont utilisés dans plus d'une langue. Ce fut le cas des thématiques sur la politique canadienne et québécoise pour lesquels des corpus en anglais et en français ont été recueillis ou encore pour des thématiques comme celle du cancer de la mère de la vedette country-pop Taylor Swift (#PrayForMamaSwift) qui embrasa la planète pendant quelques jours.

Il est tout à fait possible de spécifier les langues dans lesquelles on souhaite interroger les API de Twitter, mais nous avons choisi de conduire nos expérimentations sans les spécifier. Nous avons pu faire deux observations suite à ces expérimentations.

La première nous a permis de distinguer deux pratiques d'usagers différentes. Certains usagers gazouillent dans une langue autre que la langue dans laquelle a émergé le mot-clic thématique, par exemple des usagers italiens utilisant #PrayForMamaSwift, mais n'ajoutent à leur gazouillis que des mots-clics issus de la

langue d'origine, dans ce cas l'anglais. Dans un tel cas, même si la langue du gazouillis est italienne, les mots-clics en cooccurrence demeurent dans la langue anglaise et participent à la formation d'une folksonomie unilingue. Par contre, une autre pratique, plus répandue, consiste à utiliser un mot-clic tel que #PrayForMamaSwift puis de mettre en cooccurrence des mots-clics permettant d'inscrire la discussion dans un territoire, une culture ou un groupe particulier. Cette pratique résulte en la formation de folksonomies et éventuellement de champs lexicaux multilingues.

La deuxième observation concerne la formation de champs lexicaux et la découverte de sous-thématiques. Lorsque'une discussion sur une thématique se déroule majoritairement dans une langue en particulier, comme ce fut le cas pour la discussion autour de #PrayForMamaSwift qui se déroula principalement en anglais, les mots-clics issus d'autres langues auront des indices sémantiques de cooccurrence faibles les excluant la plupart du temps du champ lexical. Aucun indice sémantique de cooccurrence inversée ne sera alors calculé pour ces mots-clics et bien sûr aucune sous-thématique ne sera découverte. Notre prototype est donc en mesure de capter des corpus de gazouillis multilingues, mais la façon dont il crée ses champs lexicaux donne un avantage à la langue principalement utilisée dans une discussion. Par contre, dans des corpus plus équilibrés en termes d'utilisation de la langue, comme pour la thématique sur la politique canadienne, l'algorithme est en mesure de découvrir à la fois des sous-thématiques en anglais et en français. Selon les objectifs d'une recherche, ce comportement des algorithmes peut présenter un avantage et, dans d'autres cas, une limite.

Une piste pour dépasser cette limite serait de modifier le fonctionnement des algorithmes pour qu'ils créent autant de folksonomies que de langues présentes dans un corpus pour ensuite créer un champ lexical par langue et effectuer la découverte de sous-thématiques pour chacune de celles-ci. Cette voie comporte toutefois plusieurs

défis techniques, dont ceux liés aux restrictions des API et à la fiabilité de l'identification de la langue, que nous n'avons pu explorer dans le cadre de cette thèse.

8.2 Apports pour les sciences de la communication

En quelques années, Twitter est devenu un incontournable en ce qui a trait à la diffusion d'actualités, supplantant même régulièrement les chaînes radiophoniques et télévisuelles spécialisées dans l'information lors de la diffusion de primeurs (Hahn, 2013). Ce faisant, cette plateforme et les discussions qui y ont cours ont pris une place importante dans l'espace public, devenant un objet de recherche fort important pour les sciences de la communication. Cette thèse contribue à un effort de réflexion visant l'adaptation et le développement d'outils d'informatique cognitive qui apporteront un éclairage nouveau sur certaines problématiques issues du domaine des communications et plus particulièrement sur celles de l'espace public. Plusieurs phénomènes auxquels s'intéresse la recherche en communication, notamment la circulation de l'information, la redondance de l'information et l'influence des relations publiques sur le journalisme nécessitent de plus en plus l'analyse de vastes corpus de contenus médiatiques sur de longues périodes et nous croyons que cette recherche et l'inclusion de notre prototype qui en émane constituent un apport utile.

8.3 Apports pour les sciences cognitives

Par ailleurs, bien que cette recherche se concentre sur Twitter, nous croyons que notre approche et les méthodologies qui en sont issues sont applicables à d'autres plateformes faisant appel aux usagers pour catégoriser des contenus à partir de mots-clés. On pense ici notamment à Facebook qui a introduit les mots-clics (Wired, 2013), mais aussi à des sites web spécialisés proposant à leurs usagers de partager des contenus et d'en découvrir de nouveaux tels que Digg, Pinterest, Buzznet, Tagged, Youtube, Flickr, ou StumbleUpon. Tous ces services reposent sur la création de folksonomies comme moyen d'organiser des masses énormes de contenus. Par

conséquent, des thématiques émergent de ces processus de catégorisation et l'automatisation de la découverte de sous-thématiques permettraient, tout comme dans le cas de Twitter, de regrouper des contenus étant liés sémantiquement et d'enrichir le contenu proposé à l'usager.

La découverte de sous-thématiques à l'intérieur de folksonomies peut également améliorer la qualité des fonctions de recherche d'informations, notamment en utilisant les sous-thématiques comme mots-clés pour l'expansion de recherches. En effet, un prototype s'inspirant de notre approche, pourrait de façon dynamique déterminer des liens sémantiques émergeant d'une folksonomie et proposer, par exemple, des résultats de recherches incluant des contenus à la fois à propos de la thématique recherchée, mais aussi à propos des sous-thématiques découvertes par le prototype.

Nous voyons aussi des applications dans l'automatisation de publication de contenus par des robots. Sur des sites comme Twitter, plusieurs organisations utilisent des robots (*Twitterbots*) pour publier des contenus sur divers sujets. Le choix de mots-clics est primordial afin que leurs messages puissent s'insérer dans les discussions de leurs publics cibles. Notre prototype étant en mesure d’identifier des sous-thématiques à partir d'une thématique principale, un robot pourrait facilement être alimenté pour inclure ces sous-thématiques dans ces gazouillis et maximiser ainsi sa présence sur cette plateforme.

8.4 Apports pour les sciences informatiques

Du côté des sciences informatiques, les apports les plus importants se situent au plan des algorithmes qui, dans un premier temps, permettent de circonscrire un champ lexical de mots-clics autour d'une thématique pour ensuite opérer des calculs statistiques cherchant à déterminer des relations sémantiques de type *owl:partOf* et *owl:equivalent_to* et ce, même en l'absence d'une analyse linguistique qui s'avère impossible dans notre cas.

En ce qui concerne les champs lexicaux, nous sommes confiants que la méthodologie, les algorithmes et le prototype mis en place traduisent bien les éléments cognitifs décrits dans le cadre théorique et constituent une approche efficace pour extraire des champs lexicaux à partir de folksonomies. Les analyses de corpus nous ayant permis de déterminer que ces derniers suivaient un modèle de distribution correspondant à la loi de Zipf-Mandelbrot représentent également un atout majeur en terme d'optimisation des ressources.

Quant aux calculs d'indicateurs sémantiques, nous croyons que ces derniers constituent une innovation permettant de pallier le manque de ressources linguistiques pour traiter des folksonomies contenant des mots-clés inventés par les utilisateurs et pour lesquels aucune ressource linguistique ne fournit de définitions. L'utilisation de cet algorithme est un élément clé de notre prototype, car il permet l'identification de sous-thématiques à l'intérieur d'un corpus de mots-clics.

8.5 Développements futurs

Cette recherche et le prototype qui en est issu s'inscrivent dans un projet plus large initié par le Centre de recherche interuniversitaire sur la communication, l’information et la société (CRICIS, UQAM) et le Centre de recherche en Informatique Cognitive et Environnements de Formation (LICEF, Téluq) nommé l'Observatoire de la circulation de l'information (OCI). Cet observatoire a pour mission de développer les outils et les méthodes d'analyse nécessaires à la construction d'un vaste corpus d'informations circulant dans les médias canadiens, certains réseaux sociaux, les agences de presse, l'industrie des relations publiques, ainsi que les gouvernements provinciaux et fédéral afin de le rendre accessible aux chercheurs universitaires et aux mouvements sociaux.

Cette thèse alimentera les développements de l'axe 1 de l'OCI, soit l'extraction d'informations et la fouille de textes et a pour objectif de devenir le moteur de

cueillettes dynamiques de sujets d'actualité sur Twitter ainsi que sur d’autres plateformes.

En étendant notre cueillette d’information à d’autres réseaux sociotechniques comme nous avons tenté de le faire pour le cas du mot-clic « #stopHarper », ou encore à d’autres médias d’informations, il pourrait également devenir envisageable de mettre en évidence des contenus informationnels circulant dans diverses sphères de l’écosystème web. On pourrait ainsi éventuellement mettre en lumière des liens, des discours qui se font échos où encore des discours en vase clos au sein de plusieurs espaces publics partiels. Les défis techniques seraient de taille, la quantité de données nous rapprocherait rapidement du Big Data et l’analyse d’une telle masse d’information requerrait la formation d’une équipe de chercheurs, mais de telles recherches auraient la possibilité de nous donner un portrait inédit de la circulation des flux informationnels entre différents espaces publics. L’obstacle le plus sérieux à une telle démarche serait toutefois le mouvement de relative fermeture de l’accès aux API des principales plateformes ainsi que les changements de leurs conditions d’utilisation.

CHAPITRE 9 :
CONCLUSION

Rappelons notre problématique qui cherchait à automatiser la découverte de mots-clics nécessaires au suivi d'une discussion ayant cours dans l'espace public à propos d'une thématique donnée. Cette problématique nous a d'abord mené sur la piste des folksonomies créées par regroupement de mots-clics que nous avons envisagées en tant qu'intersection des processus cognitifs de catégorisation d'un ensemble d'utilisateurs. Cette posture aura fourni un socle théorique à ce projet de recherche, car elle permet de mieux saisir la valeur ontologique des choix de catégorisation des utilisateurs. Comme nous l'avons démontré, de cette base émergent des champs lexicaux autour d'un mot-clic thématique et parmi les mots-clics formant ces champs lexicaux, des relations de type *partie-tout* sont observables.

L'émergence de champs lexicaux entre mots-clics affichant une forte cooccurrence entre eux nous a conduits à poser la question des relations sémantiques entre ces derniers. Ce faisant, nous avons noté dans les corpus analysés que des relations partie-tout sont fréquemment observables dans les champs lexicaux émergeant des folksonomies créées par notre prototype. Par la recherche de définitions de mots-clics, on a identifié des types de relations revenant le plus souvent et auxquels les prédicats RDF *owl :partOf* et *owl :hasPart* ainsi que *owl:equivalent_to* peuvent être attribués. La difficulté d'automatiser la recherche de définitions et l'attribution de relations sémantiques entre mots-clics ont nécessité le développement d'une méthode faisant usage de statistiques sur la cooccurrence entre mots-clics.

Par l'attribution de relations sémantiques à partir de calculs statistiques ou, comme nous les avons appelé, des indicateurs sémantiques, nous avons été en mesure de mieux circonscrire les champs lexicaux et d'identifier les mots-clics unis par la relation *owl :partOf.*

Notre cadre théorique a été mis en œuvre dans un prototype capable de façon autonome de réaliser une cueillette et d'automatiser la découverte de mots-clics représentant des sous-thématiques. Notre objectif principal demeure l'observation de la circulation de l'information dans l'espace public, donc sur des thématiques d'intérêt public, mais nos expérimentations ont démontré que le prototype pouvait très bien être utilisé pour des thématiques aussi diverses que le sport, la fiction télévisuelle ou les faits divers.

L'hypothèse que nous avions formulée à savoir que « le suivi d'une discussion sur Twitter à propos d'un mot-clic représentant une thématique discutée dans l'espace public peut être automatisé par une approche de raisonnement inductif appliquée à l'analyse de mots-clics cooccurrents ainsi qu'à l'extraction de relations sémantiques entre ces mots-clics », s'est avérée. Chacun des trois concepts sur lesquels elle s'appuyait soit [1] la création de folksonomies par le regroupement de mots-clics, [2] l'émergence de champs lexicaux entre mots-clics affichant une forte cooccurrence entre eux et [3] l'attribution de relations sémantiques à des mots-clics à l'aide de calculs statistiques a fait l'objet d'une validation rigoureuse sur une dizaine de corpus. Les résultats positifs de la validation des inférences algorithmiques réalisée pour chacun de ces trois concepts par deux évaluateurs externes attestent à la fois de la pertinence du cadre théorique et de la qualité de son implémentation computationnelle.

En conclusion, nous croyons que ce prototype faisant usage d'une logique d'induction permettra de mieux recueillir l'ensemble d'une thématique de discussion et de ses sous-thématiques sur Twitter, et ce, de façon dynamique et non supervisée. De plus, comme nous l'avons abordé dans le chapitre traitant des exemples d'utilisation, ces fondements théoriques et la méthodologie qui en est issue offrent des potentiels d'application pour plusieurs autres dispositifs sociotechniques reposant sur les folksonomies comme base d'organisation du contenu.

ANNEXE A :
DIRECTIVES AUX ÉVALUATEURS

Guide de l'évaluateur externe

1. Ouvrir le fichier LibreOffice Calc nommé « Évaluation#(nom du mot-clic).ods
2. Aller à l'onglet « Thématique » pour prendre connaissance de la thématique traitée dans ce corpus de gazouillis.

Évaluation du champ lexical et attribution de prédicat RDF

1. Aller à l'onglet « ChampLexical »

2. La colonne à remplir se nomme « Prédicat RDF »

3. Chacune des cellules de cette colonne contient un menu déroulant. Vous devez sélectionner un des 4 choix suivants:

    1. **« *owl:partOf* » :** Si à la lecture de la description du mot-clic, vous pouvez le considérer comme faisant partie du mot-clic thématique.
    Ex : le mot-clic #lpc (Liberal Party of Canada) est considéré comme faisant partie (*owl:partOf*) du mot-clic thématique #cdnpoli (politique canadienne).
    2. **« *owl :hasPart* »** : Si à la lecture de la description du mot-clic, vous pouvez considérer que le mot-clic thématique en fait partie.
    Ex : le mot-clic #canada a comme partie (*owl:hasPart*) le mot-clic thématique #cdnpoli (politique canadienne).
    3. **« *owl:equivalentTo* »** : Si à la lecture de la description du mot-clic, vous pouvez considérer qu'il est équivalent ou synonyme du mot-clic thématique.
    Ex : le mot-clic #polCAN (politique canadienne) est équivalent au mot-clic thématique #cdnpoli (politique canadienne).
    4. **« *None* »** : Si à la lecture de la description du mot-clic, vous pouvez considérer qu'il n'a juste pas rapport avec la thématique.
    Ex : le mot-clic #Gaza1YearOn (opposition à la politique israélienne) n'a pas rapport avec le mot-clic thématique #cdnpoli.

Cohérence du corpus

Les onglets « GazouillisAleatoiresT » et les onglets « GazouillisAleatoiresC » contiennent 100 gazouillis extraits de façon aléatoire d'un corpus. Les directives sont les mêmes concernant l'évaluation des contenus de ces deux onglets.

1. Aller à l'onglet « GazouillisAleatoiresT » et ensuite « GazouillisAleatoiresC »
2. La colonne à remplir se nomme « Fait partie de la thématique? »
3. Elle est divisée en trois sous-colonnes permettant de répondre « oui », « non » ou « incertain »
4. Parfois le contenu du gazouillis suffit pour répondre à la question, parfois il est nécessaire de suivre un hyperlien s'il est présent.

Exemples de « oui » pour le mot-clic thématique #cdnpoli (politique canadienne)

RT @stephenlautens: .@happyfamily837 Then there's this... #cdnpoli http://t.co/L97LfWcUFM
The TVVH Urban Daily is out! http://t.co/NGrwiYozQ1 #cdnpoli #elxn2015 Stories via @TraderApple
RT @Bergg69: Harper's pretend Senate Position!http://t.co/tnlhTcQTW4#cdnpoli @PMHarper http://t.co/4NPRVUXFbW
This is going to be interesting. #cdnpoli https://t.co/P5YiCZ7C2e
RT @Can_ada: Fire Crew Given Stop Work Order to Accommodate Photo-op - Harper explains how his pants are always on fire. #cdnpoli http://t…
RT @CDHill9: This walking blunder shut 9 #Veterans Affairs off &refused 3600 for PTSD treatment &ppl STILL defend him! #Cdnpoli http://t.co…
RT @Can_ada: Fire Crew Given Stop Work Order to Accommodate Photo-op - Harper explains how his pants are always on fire. #cdnpoli http://t…

Exemples de «non» pour le mot-clic thématique #cdnpoli (politique canadienne)

RT @DAJHetherington: Our NATO ally #Turkey is letting a Kurdish city of 10,000 people burn to the ground. #cdnpoli #warcrime #Lice #Kurds h…
RT @ArtHealing44: #FreePalestine#Palestine#FreedomFlotilla#Gaza#Gaza1YearOn#Boycottisrael#ICC4israel#BDS#Europe#EU#cdnpoli
Ex-Hasid Wrote Email Excoriating Ultra-Orthodox Jewish Community A Week Before Suicide http://t.co/AN3MNRucDH #uspoli #cdnpoli #zionis

RT @Bergg69: Crimea Is Now Putin's Problem Child http://t.co/EFCmzxqDcR #cdnpoli #yqr
#FreePalestine#Palestine#FreedomFlotilla#Gaza#Gaza1YearOn#Boycottisrael#ICC4israel#BDS#Europe#EU#cdnpoli
Saudi Arabia no match for Iran: Aoun #uspoli #cdnpoli #ukpoli #gaza #plo #hamas http://t.co/qtmgnuytHH

Exemples de «incertain» pour le mot-clic thématique #cdnpoli (politique canadienne)

RT @DianeShears: #UnavoidablyUnsafe #abpoli #bcpoli #onpoli https://t.co/8iPHP7pcua
RT @timescolonist: Ex-employee says it’s routine for B.C. government to destroy emails #yyj #bcpoli http://t.co/tPJf7ohW4N
RT @sunlorrie: Here's our Sunday Toronto Sun editorial: Patrick Brown’s job is to win http://t.co/ftbE0GTD4b #onpoli #topoli
RT @waitinginBC: Provincial elections in BC can't come soon enough. #timetotakeoutthetrash #bcpoli
RT @Norm_Farrell: I'm hearing threats & nasty warnings also tools to ensure silence of gov't staff who know about #LNG deals. #bcpoli

ANNEXE B :

CHAMPS LEXICAUX DES 10 CORPUS D'ÉVALUATION

Champ lexical du mot-clic #backtothefuture (2015-10-23 01:18)

Mot-clic	Description	Cooccur-rence	Cooccur-rence inversée
lootcrate	boite à surprise pour geeks	0.0118	0.3356
BTTFDAY	Back to the future Day	0.0150	0.3941
KimmelinBrooklyn	venue de l'animateur de talk-show Jimmy Kimmel dans Brooklyn	0.0171	0.3600
CONFIDENT	CONFIANT	0.0407	0.0000
BackToFutureDay	célébration entourant le 30e anniversaire de Back to The Future	0.0396	0.4052
BTTF	Back to the future, le film	0.0889	0.5000
BTTF2015	Back to the future, le film	0.0567	0.5410
MartyMcFly	personnage principal de BTTF	0.0450	0.7522
Delorean	voiture emblématique de BTTF	0.0225	0.5822
usatoday	journal américain	0.0171	0.2143
BackToTheFutureDay	célébration entourant le 30e anniversaire de Back to The Future	0.1156	0.3623
tbt	ThrowBack Thursday(s)	0.0203	0.0086
DocBrown	personnage de BTTF	0.0139	0.5115
Kimmel	animateur de talk-show Jimmy Kimmel	0.0096	0.3088
MichaelJFox	acteur principal dans BTTF	0.0150	0.5542
ChristopherLloyd	acteur dans BTTF	0.1638	0.7062
mcfly	personnage principal de BTTF	0.0128	0.5058
21OCT15	jour où le personnage principal de BTTF retourne dans le futur	0.0182	0.9385
nikemag	paire de chaussures créée par l'entreprise popularisée grâce à BTTF	0.0075	0.4549
FueledByEverything	particularité de la voiture de BTTF de pouvoir être propulsée par n'importe quoi	0.0086	0.5556
NIKE	entreprise d'équipements sportifs	0.0096	0.0440
hoverboard	évolution fictive du skateboard inventé dans le film BTTF	0.0128	0.4476
Future	le futur	0.0086	0.0170
BenghaziCommittee	comité américain chargé d'enquêter sur une attaque de leur embassade à Benghazi	0.0332	0.0150
PredictiveProgramming	référence à un montage de scènes de BTTF qui aurait prédit les attentats du 11 sept.	0.0075	0.9714
Pluto	planète naine	0.0075	0.0959
viralchat	chat très populaire	0.0139	0.0968
LeaThompson	actrice dans BTTF	0.0107	0.3500
FutureDay	Le jour du futur	0.0075	0.4486
F1	Grand Prix de course automobile	0.0075	0.0057
SHAYTARDS	youtubers populaires	0.0075	0.2000
business	monde des affaires	0.0064	0.0000
EasyThrowback	en référence à Throwback Thursdays	0.0064	0.0667
ToyotaMirai	promotion d'une voiture à l'hydrogen	0.0278	0.3538

WSHH	World Star Hip Hop	0.0075	0.0000
Cubs	club de baseball	0.0086	0.1034
SciFi	science-fiction	0.0128	0.0162
TheWalk	titre d'un film réalisé par le même réalisateur que BTTF	0.0086	0.3962
nyc	New-York City	0.0118	0.0000
9ll	référence aux attentats du 11 sept.	0.0139	0.9796
Backtothefuture30thanniversary	30e anniversaire du film BTTF	0.0910	0.5479

Champ lexical du mot-clic #cop21 (2015-11-13 17:17)

Mot-clic	Description	Cooccurrence	Cooccurrence inversée
climat	climat	0,0261	0,5212
ClimateChange	changement climatique	0,1229	0,1067
climate	climat	0,0536	0,0781
FossilFuel	énergie fossile	0,0118	0,0719
24HoursofReality	événement spécial incluant conférences et spectacle en préparation de Cop 21	0,0575	0,1362
WhyImWatching	événement spécial incluant conférences et spectacle en préparation de Cop 21	0,051	0,093
Paris	capitale de la France, lieu où se déroule Cop 21	0,1412	0,0045
energy	énergie fossile	0,0144	0,0113
CambioClimático	changement climatique	0,0261	0,1719
ActOnClimate	leitmotiv « Agissons pour le climat maintenant »	0,0196	0,1063
energía	énergie	0,0484	0,0504
climateaction	leitmotiv « Agissons pour le climat »	0,0144	0,3433
Euro2016	Championnat d'Europe de football 2016	0,0275	0,0138
Attaques	relatif aux attentats de Paris le 13 nov.	0,0366	0,0149
TPP	Trans-Pacific Partnership	0,0196	0,0152
VANPOLI	politique de la ville de Vancouver	0,0144	0,0283
YVR	Aéroport international de Vancouver	0,0144	0,0079
attentat	relatif aux attentats de Paris du 13 nov.	0,0261	0
TSX	indice de la bourse de Toronto	0,0118	0,0323
csoj	« Ce soir ou jamais »	0,0183	0,4688
iOT	Internet of Things	0,0144	0,0054
BigData	mégadonnées	0,0157	0,0052
FINTECH	Financial technology	0,0118	0,0052
Action2015	Initiative onusienne de lutte aux inégalités socio-économiques	0,0601	0,15
Bataclan	salle de spectacle parisienne lieu d'attentats terroristes	0,0392	0,0024
cdnpoli	politique canadienne	0,0248	0,0262
Explosion	relatif aux attentats de Paris le 13 nov.	0,0314	0,0041
hollande	président de la France	0,0379	0,0242
auspol	Politique australienne	0,017	0,0812

Fusillade	relatif aux attentats de Paris le 13 nov.	0,1621	0
Valls	Premier ministre de la France	0,0118	0,013
Tecnologia	technologie	0,0118	0,0203

Champ lexical du mot-clic #BigData (2015-10-11 16:01)

Mot-clic	**Description**	**Cooccur-rence**	**Cooccur-rence inversée**
IoT	Internet Of Things	0,5	0,7051
DataScience	science des données	0,1091	0,543
Security	sécurité informatique	0,3758	0,6688
InfoSec	sécurité informatique	0,3272	0,7218
analytics	Analyse de données	0,0894	0,3355
KDN	site web à propos de Business Analytics, Big Data, Data Mining, and Data Science	0,0099	0,1212
SaaS	Software as a Service (technologie)	0,0145	0,1227
Statistics	statistiques	0,0112	0,0226
Cloud	infonuagique	0,0329	0,068
DataViz	Visualisation de données	0,0118	0,2292
Rstats	application du langage R à l'analyse statistique	0,0112	0,3103
WebRTC	technologie de communication en temps réel	0,0158	0,1034
Data	données	0,0125	0,0726
SEO	Search Engine Optimization	0,2562	0,2905
Hadoop	système de fichiers souvent utilisé pour le Big Data	0,0177	0,4222
cybersecurity	sécurité informatique	0,0145	0
digital	numérique	0,0112	0,0063
tech	technologie	0,2635	0,1698
machinelearning	apprentissage par des logiciels	0,0348	0,2933
Opines	exprime une opinion	0,0217	0,1373
Hiring	embauche	0,0099	0
AI	Intelligence artificielle	0,0112	0,0926
Wearables	technologie portable (vêtements ou accessoires)	0,0427	0,0796
API	Application Programming Interface	0,0315	0,1996
deeplearning	Deep Learning (méthodes d'apprentissage automatique)	0,0158	0,2059
RT	demande de retweet	0,0164	0
M2M	communication machine à machine	0,0355	0,4191
Jobs	Emplois	0,0151	0
Java	langage de programmation	0,0151	0,0203
job	Emplois	0,0164	0
Hacking	Bidouillage informatique	0,0118	0
Tecnologia	technologie	0,0118	0,0203

Champ lexical du mot-clic #cdnpoli (2015-07-26 17:51)

Mot-clic	Description	Cooccur-rence	Cooccur-rence inversée
polCAN	politique canadienne	0.0570	0.6053
canpoli	*Canadian politics*	0.0263	0.5516
lpc	*Liberal Party of Canada*	0.1555	0.6410
ndp	*New Democratic Party*	0.1194	0.6970
PolQC	politique québécoise	0.1150	0.6296
cpc	*Conservative Party of Canada*	0.1150	0.7381
BCpoli	*British-Columbia politics*	0.0800	0.5476
TM4PM	*Thomas Mulcair for Prime Minister*	0.0624	0.6031
onpoli	*Ontario politics*	0.0613	0.6234
CPCJesus	sarcasme sur les valeurs catholiques associées au parti conservateur	0.0493	0.7108
c51	projet de loi fédéral	0.0372	0.8389
elxn42	42e élection canadienne	0.0372	0.747p
StopHarper	Arrêter Harper	0.0296	0.8679
Harper	premier ministre canadien	0.0230	0.5013
pq	Parti Québécois	0.0416	0.0000
assnat	Assemblée nationale du Québec	0.0405	0.0000
eglaw	circonscription de Eglinton-Lawrence ou la transfuge Eve Adams tentait de gagner une investiture pour le PLC	0.0383	0.4706
BlocQC	Bloc Québécois	0.0383	0.0000
UniRose	Circonscription fédérale à Toronto, University Rosedale	0.0252	0.2121
Quebec	province	0.0241	0.0200
PaysQc	mouvement indépendantiste du Québec	0.0208	0.0000
canada	pays	0.0230	0.0107
weekendofaction	attire l'action sur divers événements	0.0350	0.0737
BDS	Boycott, Divestment and Sanctions: a global movement to apply nonviolent, economic pressure to Israel to comply with international law and human rights	0.0197	0.0067
Gaza	ville qui donne son nom au territoire de la bande de Gaza	0.0197	0.0000
Europe	continent	0.0175	0.0000
ICC4israel	This hashtag was created to hold Israel accountable for its war crimes by demanding that Israel gets referred to the International Criminal Court (ICC4ISRAEL).	0.0175	0.0390
FreePalestine	mouvement indépendantiste de la Palestine	0.0175	0.0000
EU	« European Union »	0.0175	0.0000
Gaza1YearOn	Opposition à la politique israélienne	0.0175	0.0064

Champ lexical du mot-clic #DonDuSang (2015-11-05 07:30)

Mot-clic	Description	Cooccur-rence	Cooccur-rence inversée
DonDuSangPourTous	Slogan contre les restrictions imposées par la ministre Touraine aux dons de sang par les gays	0,0875	0,5238
MarisolTouraine	ministre des Affaires sociales et de la Santé de France	0,0875	0,6667
freeprepnow	collectif revendiquant l'accès à la prophylaxie pré-exposition (PrEP)	0,025	0
homosexuels	Personnes ayant des pratiques sexuelles avec des individus du même sexe	0,0375	0,3889
gay	homme ayant une attirance pour les individus du même sexe	0,2	0,0781
homophobie	hostilité, explicite ou implicite, envers des homosexuels	0,1	0,3636
homosexuel	Personne ayant des pratiques sexuelles avec des individus du même sexe	0,15	0,25
LGBT	Lesbiennes, gays, bisexuels et transgenres	0,15	0,1957
LePen	politicien français lié au parti Front National	0,2125	0,037
Mediapart	journal en ligne d'information généraliste	0,2125	0,0149
Sarkozy	politicien français lié au parti « Les Républicains »	0,2125	0,007
AirCocaine	deux pilotes français arrêtés pour trafic de drogue	0,2125	0,0385
Benzema	footballeur ayant une implication présumée dans une histoire de chantage à la «sextape»	0,2125	0,005
NeverGiveUp	Lâche pas	0,05	0,3333
LGJ	« Le Grand Journal », téléjournal en France	0,0625	0

Champ lexical du mot-clic #fn (2015-10-30 13:00)

Mot-clic	Description	Cooccur-rence	Cooccur-rence inversée
Bettati	politicien s'étant nouvellement rallié au FN	0,0526	1
GUD	organisation étudiante française d'extrême droite réputée pour ses actions violentes	0,0226	1
Ravier	sénateur et maire FN ayant défrayé les manchettes suite à l'embauche de son fils à la mairie	0,0376	0,6364
Desport	Adrien Desport, ex-FN, condamné à 3 ans de prison pour avoir incendié en groupe treize voitures	0,0376	0,5
Ménard	maire FN De Béziers, populaire pour sa tentative d'interdiction du kebab dans sa ville	0,0602	0,0479
Regionales2015	élections régionales françaises de 2015	0,2707	0,1158
UMP	parti politique français de droite, actif de 2002 à 2015	0,0301	0,35
Culture	ensemble des traits distinctifs, spirituels, matériels, intellectuels et affectifs, qui caractérisent une société	0,0376	0,0215
ps	Parti socialiste de France	0,0752	0,125
Marseille	ville de France	0,1654	0,0858
Gard	région ou se déroulent des élections régionales	0,0602	0,0172
Beziers	commune de France faisant régulièrement l'actualité à cause de son maire « coloré » et très à droite	0,0526	0,0758
PACA	désigne la région Provence-Alpes-Côte d'Azur	0,0526	0,092
kebab	désigne différents plats à base de viande grillée	0,0677	0,0467

Champ lexical du mot-clic #polQc (2015-08-20 17:30)

Mot-clic	Description	Cooccur-rence	Cooccur-rence inversée
assnat	Assemblée nationale	0,3578	0,9512
pq	Parti Québécois	0,1121	0,8023
plq	Parti libéral du Québec	0,0647	0,7746
PL44	projet de loi 44 du Québec	0,056	0,5854
PaysQc	mouvement indépendantiste du Québec	0,069	0,8679
BlocQC	Bloc Québécois	0,2328	0,5766
caq	Coalition Avenir Québec	0,0388	0,8889
Pl59	projet de loi 59 du Québec	0,0388	0,2941
PKP	Pierre-Karl Péladeau	0,0345	0,2333
polcan	politique canadienne	0,1983	0,4103
Québec	province	0,0388	0,0352
cdnpoli	politique canadienne	0,0345	0,0049
elxn42	42e élection canadienne	0,2586	0,0469
fed2015	élection fédérale 2015	0,1034	0,3744

Champ lexical du mot-clic #PrayForMamaSwift (2015-04-10 16:45)

Mot-clic	Description	Cooccur-rence	Cooccur-rence inversée
StayStrongMamaSwift	support pour la mère de Taylor Swift	0,0498	0,9081
PrayersForMamaSwift	prières pour la mère de Taylor Swift (Andrea)	0,0272	0,5323
DirectionersAreHereForSwifties	support des fans de One Direction à la famille Swift	0,427	0,9136
staystrongtaylor	support à Taylor Swift	0,0464	0,6856
KatyCatsAreHereForSwifties	support des fans de Katy Perry à la famille Swift	0,0306	0,7558
SwiftiesAreHereForYouTaylor	support à Taylor Swift de la part de ses fans	0,0294	0,9145
SwiftStrong	support à Taylor Swift	0,0159	0,6742
5SOSFamarehereforTaylor	support des fans de 5 Seconds of Summer à la famille Swift	0,0159	0,7387
WeLoveYouMamaSwift	nous vous aimons maman Swift	0,0147	0,767
SelenatorsAreHereForSwifities	support des fans de Selena Gomez aux fans de Taylor Swift	0,0136	0,7674
BeliebersAreHereForTaylor	support des fans de Justin Bieber à Taylor Swift	0,0113	0,7797
BeliebersAreHereForSwifties	support des fans de Justin Bieber à Taylor Swift	0,0102	0,86
PrayForAndrea	prier pour la mère de Taylor Swift (Andrea)	0,0113	0,9053
Tribunale	tribunaux italiens	0,0408	1
Swifties	groupe de fans de Taylor Swift	0,0227	0,6
TaylorSwift	Taylor Swift	0,0227	0,0501
1DisFinallyFreeFromModest	fin du contrat de One Direction avec la firme Modest	0,0204	0,037
CelebrityAwards2015	gala des célébrités	0,0136	0,0238
CatchGG	mot-clic promouvant le groupe Girl's Generation	0,0113	0,0742
kindle	Liseuse électronique	0,0113	0
libro	Livre (en espagnol ou en italien)	0,0113	0
CallMeBaby4thWin	prix remporté par le groupe EXO	0,0113	0,2948
libros	livre (en espagnol ou en Italien)	0,0102	0
amazon	service de vente en ligne	0,0102	0

Champ lexical du mot-clic #PSG (2014-03-31 13:53)

Mot-clic	Description	Cooccur-rence	Cooccur-rence inversée
CHELSEA	équipe de foot	0,1088	0,5179
Ibrahimovic	joueur du PSG	0,039	0,5843
ParisChampionsDream	support à l'équipe du PSG	0,0227	0,6226
TeamPSG	équipe de foot	0,4118	0,6286
Livel1	ligue 1 en direct	0,033	0,9814
Pariskop	site d'actualité sur le Paris Saint-Germain	0,0115	1
PSGCHE	Paris Saint-Germain « vs » Chelsea	0,1264	0,2981
PSGCHELSEA	Paris Saint-Germain « vs » Chelsea	0,0944	0,4362
PSGCFC	Paris Saint-Germain « vs » Chelsea	0,0496	0,453
CFC	Chelsea Football Club	0,0262	0,1333
ICICESTPARIS	support à l'équipe du PSG	0,0214	0,4444
Zlatan	prénom d'un joueur du PSG	0,0189	0,4233
cavani	joueur du PSG	0,017	0,4656
Mourinho	entraineur de l'équipe Chelsea	0,0128	0,4619
LDC	Ligue Des Champions	0,1219	0,3005
ChampionsLeague	Ligue des Champions	0,0352	0,2051
PARIS	la ville	0,0285	0,0723
foot	sport	0,0256	0,1096
UCL	UEFA Champions League	0,0186	0,3824
liguedeschampions	Ligue des Champions	0,0115	0,3953
LDCLiveCamp	diffusion des match de la LDC	0,0618	0,3902
DreamBigger	Elvis Gratton déguisé en fan de foot	0,0374	0,4645
Opta	fournisseur de données de performances sportives	0,0246	0,1429
L1	Ligue 1	0,0243	0,2449
Ligue1	Ligue 1	0,0214	0,0577
OM	équipe Olympique de Marseille	0,0154	0,0574
RT	demande de retweeter	0,0138	0
Monaco	la ville	0,0122	0,1136

Champ lexical du mot-clic #Syria (2015-07-29 11:31)

Mot-clic	Description	Cooccur-rence	Cooccur-rence inversée
Damascus	ville de Damas	0,0287	0,6557
Assad	président de la Syrie	0,0533	0,6545
Aleppo	Alep, ville de Syrie	0,0287	0,6087
Zabadani	ville de Syrie, lieu de conflits armés	0,0342	0,75
Zahle	ville libanaise près de la frontière syrienne où des tirs de rocket ont été observés	0,0273	0,75
ISIS	Islamic State of Iraq and Syria	0,138	0,1368
IS	Islamic State	0,0342	0,0969
Kurds	Kurdes	0,0383	0,2404
Daesh	Acronyme arabe pour « État islamique »	0,0191	0,0774
YPG	bras armé des Kurdes de Rojava (Kurdistan syrien)	0,0205	0,2482
AlQaeda	mouvement salafiste djihadiste	0,0219	0,2162
ISIL	Islamic State of Iraq and the Levant	0,0287	0,0212
middleeast	Moyen-Orient	0,0178	0,0638
iraq	Iraq, le pays	0,0779	0,1921
Iran	Iran, le pays	0,1134	0,0287
euronews	média européens	0,0574	0,0305
Israel	Israël, le pays	0,0451	0,0352
interfax	agence de presse russe	0,0191	0,0894
Turkey	Turquie, le pays	0,123	0,1782
news	actualités	0,0355	0
IranDeal	négociations entre les É-U et Iran	0,0178	0,003
UN	United Nations	0,0437	0,0495
UNSC	United Nations Security Council	0,1093	0,1514
LeMonde	Média français	0,0369	0,0939
US	United-States	0,0355	0,0211
Lebanon	Liban, le pays	0,0464	0,0816
yemen	Yémen, le pays	0,026	0,1455
Irantalks	négociations entre les É-U et Iran	0,0191	0,0081
google	entreprise	0,0574	0,0071
vocDT	radio musulmane d'Afrique du sud	0,0273	0,5

ANNEXE C :
COMPARATIF ENTRE LES DÉCISIONS DE L'ALGORITHME ET LES CLASSIFICATIONS DES ÉVALUATEURS SUR LES RELATIONS « PARTIE DE »

Mot-clic	Prédicat RDF	Indicateur sémantique de cooccurrence inversée	Mot-clic thématique	Évaluateur
21OCT15	owl:partOf	0,9385	BacktoTheFuture	1
21OCT15	owl:partOf	0,9385	BacktoTheFuture	2
9ll	none	0,9796	BacktoTheFuture	1
9ll	owl:hasPart	0,9796	BacktoTheFuture	2
backtothefuture30thanniversary	owl:equivalentTo	0,5479	BacktoTheFuture	1
backtothefuture30thanniversary	owl:equivalentTo	0,5479	BacktoTheFuture	2
BTTF2015	owl:equivalentTo	0,541	BacktoTheFuture	1
BTTF2015	owl:partOf	0,541	BacktoTheFuture	2
ChristopherLloyd	owl:partOf	0,7062	BacktoTheFuture	1
ChristopherLloyd	owl:partOf	0,7062	BacktoTheFuture	2
Delorean	owl:partOf	0,5822	BacktoTheFuture	1
Delorean	owl:partOf	0,5822	BacktoTheFuture	2
DocBrown	owl:partOf	0,5115	BacktoTheFuture	1
DocBrown	owl:partOf	0,5115	BacktoTheFuture	2
FueledByEverything	owl:partOf	0,5556	BacktoTheFuture	1
FueledByEverything	owl:partOf	0,5556	BacktoTheFuture	2
MartyMcFly	owl:partOf	0,7522	BacktoTheFuture	1
MartyMcFly	owl:partOf	0,7522	BacktoTheFuture	2
mcfly	owl:partOf	0,5058	BacktoTheFuture	1
mcfly	owl:partOf	0,5058	BacktoTheFuture	2
MichaelJFox	owl:partOf	0,5542	BacktoTheFuture	1
MichaelJFox	owl:partOf	0,5542	BacktoTheFuture	2
PredictiveProgramming	owl:partOf	0,9714	BacktoTheFuture	1
PredictiveProgramming	owl:partOf	0,9714	BacktoTheFuture	2
5SOSFamarehereforTaylor	owl:partOf	0,7387	PrayForMamaSwift	1
5SOSFamarehereforTaylor	owl:partOf	0,7387	PrayForMamaSwift	2
Aleppo	owl:partOf	0,6087	Syria	1
Aleppo	owl:partOf	0,6087	Syria	2
Assad	owl:partOf	0,6545	Syria	1
Assad	owl:partOf	0,6545	Syria	2
assnat	owl:partOf	0,9512	polQc	1
assnat	owl:partOf	0,9512	polQc	2
BCpoli	owl:partOf	0,5476	cdnpoli	1
BCpoli	owl:partOf	0,5476	cdnpoli	2
BeliebersAreHereForSwifties	owl:partOf	0,86	PrayForMamaSwift	1
BeliebersAreHereForSwifties	owl:partOf	0,86	PrayForMamaSwift	2
BeliebersAreHereForTaylor	owl:partOf	0,7797	PrayForMamaSwift	1
BeliebersAreHereForTaylor	owl:partOf	0,7797	PrayForMamaSwift	2

Bettati	owl:partOf	1	fn	1
Bettati	owl:partOf	1	fn	2
BlocQC	owl:partOf	0,5766	polQc	1
BlocQC	owl:partOf	0,5766	polQc	2
c51	owl:partOf	0,8389	cdnpoli	1
c51	owl:partOf	0,8389	cdnpoli	2
canpoli	owl:equivalentTo	0,5516	cdnpoli	1
canpoli	owl:equivalentTo	0,5516	cdnpoli	2
caq	owl:partOf	0,8889	polQc	1
caq	owl:partOf	0,8889	polQc	2
CHELSEA	owl:partOf	0,5179	PSG	1
CHELSEA	owl:partOf	0,5179	PSG	2
climat	owl:partOf	0,5212	cop21	1
climat	owl:hasPart	0,5212	cop21	2
cpc	owl:partOf	0,7381	cdnpoli	1
cpc	owl:partOf	0,7381	cdnpoli	2
CPCJesus	owl:partOf	0,7108	cdnpoli	1
CPCJesus	owl:partOf	0,7108	cdnpoli	2
Damascus	owl:partOf	0,6557	Syria	1
Damascus	owl:partOf	0,6557	Syria	2
DataScience	owl:partOf	0,543	BigData	1
DataScience	owl:hasPart	0,543	BigData	2
DirectionersAreHereForSwifties	owl:partOf	0,9136	PrayForMamaSwift	1
DirectionersAreHereForSwifties	owl:partOf	0,9136	PrayForMamaSwift	2
DonDuSangPourTous	owl:partOf	0,5238	DonDuSang	1
DonDuSangPourTous	owl:equivalentTo	0,5238	DonDuSang	2
elxn42	owl:partOf	0,7471	cdnpoli	1
elxn42	owl:partOf	0,7471	cdnpoli	2
GUD	owl:partOf	1	fn	1
GUD	owl:partOf	1	fn	2
Harper	owl:partOf	0,5013	cdnpoli	1
Harper	owl:partOf	0,5013	cdnpoli	2
Ibrahimovic	owl:partOf	0,5843	PSG	1
Ibrahimovic	owl:partOf	0,5843	PSG	2
InfoSec	owl:partOf	0,7218	BigData	1
InfoSec	owl:partOf	0,7218	BigData	2
IoT	owl:partOf	0,7051	BigData	1
IoT	owl:partOf	0,7051	BigData	2
KatyCatsAreHereForSwifties	owl:partOf	0,7558	PrayForMamaSwift	1
KatyCatsAreHereForSwifties	owl:partOf	0,7558	PrayForMamaSwift	2
Livel1	none	0,9814	PSG	1
Livel1	owl:hasPart	0,9814	PSG	2
lpc	owl:partOf	0,641	cdnpoli	1
lpc	owl:partOf	0,641	cdnpoli	2

MarisolTouraine	owl:partOf	0,6667	DonDuSang	1
MarisolTouraine	owl:hasPart	0,6667	DonDuSang	2
ndp	owl:partOf	0,697	cdnpoli	1
ndp	owl:partOf	0,697	cdnpoli	2
onpoli	owl:partOf	0,6234	cdnpoli	1
onpoli	owl:partOf	0,6234	cdnpoli	2
ParisChampionsDream	owl:partOf	0,6226	PSG	1
ParisChampionsDream	owl:partOf	0,6226	PSG	2
Pariskop	none	1,0000	PSG	1
Pariskop	owl:partOf	1	PSG	2
PaysQc	owl:partOf	0,8679	polQc	1
PaysQc	owl:partOf	0,8679	polQc	2
PL44	owl:partOf	0,5854	polQc	1
PL44	owl:partOf	0,5854	polQc	2
plq	owl:partOf	0,7746	polQc	1
plq	owl:partOf	0,7746	polQc	2
polCAN	owl:equivalentTo	0,6053	cdnpoli	1
polCAN	owl:equivalentTo	0,6053	cdnpoli	2
PolQC	owl:partOf	0,6296	cdnpoli	1
PolQC	owl:partOf	0,6296	cdnpoli	2
pq	owl:partOf	0,8023	polQc	1
pq	owl:partOf	0,8023	polQc	2
PrayersForMamaSwift	owl:equivalentTo	0,5323	PrayForMamaSwift	1
PrayersForMamaSwift	owl:equivalentTo	0,5323	PrayForMamaSwift	2
PrayForAndrea	owl:hasPart	0,9053	PrayForMamaSwift	1
PrayForAndrea	owl:equivalentTo	0,9053	PrayForMamaSwift	2
Ravier	owl:partOf	0,6364	fn	1
Ravier	owl:partOf	0,6364	fn	2
Security	owl:partOf	0,6688	BigData	1
Security	owl:hasPart	0,6688	BigData	2
SelenatorsAreHereForSwifities	owl:partOf	0,7674	PrayForMamaSwift	1
SelenatorsAreHereForSwifities	owl:partOf	0,7674	PrayForMamaSwift	2
StayStrongMamaSwift	owl:equivalentTo	0,9081	PrayForMamaSwift	1
StayStrongMamaSwift	owl:equivalentTo	0,9081	PrayForMamaSwift	2
staystrongtaylor	owl:partOf	0,6856	PrayForMamaSwift	1
staystrongtaylor	owl:equivalentTo	0,6856	PrayForMamaSwift	2
StopHarper	owl:partOf	0,8679	cdnpoli	1
StopHarper	owl:partOf	0,8679	cdnpoli	2
Swifties	none	0,6000	PrayForMamaSwift	1
Swifties	owl:hasPart	0,6	PrayForMamaSwift	2
SwiftiesAreHereForYouTaylor	owl:partOf	0,9145	PrayForMamaSwift	1
SwiftiesAreHereForYouTaylor	owl:partOf	0,9145	PrayForMamaSwift	2
SwiftStrong	owl:partOf	0,6742	PrayForMamaSwift	1
SwiftStrong	owl:equivalentTo	0,6742	PrayForMamaSwift	2

TeamPSG	owl:hasPart	0,6286	PSG	1
TeamPSG	owl:equivalentTo	0,6286	PSG	2
TM4PM	owl:partOf	0,6031	cdnpoli	1
TM4PM	owl:partOf	0,6031	cdnpoli	2
Tribunale	none	1,0000	PrayForMamaSwift	1
Tribunale	none	1,0000	PrayForMamaSwift	2
WeLoveYouMamaSwift	owl:partOf	0,767	PrayForMamaSwift	1
WeLoveYouMamaSwift	owl:equivalentTo	0,767	PrayForMamaSwift	2
Zabadani	owl:partOf	0,75	Syria	1
Zabadani	owl:partOf	0,75	Syria	2
Zahle	owl:partOf	0,75	Syria	1
Zahle	owl:partOf	0,75	Syria	2

ANNEXE D :
NIVEAUX DE CONNAISSANCES DES THÉMATIQUES ÉVALUÉES PAR LES ÉVALUATEURS

	Évaluateur #1				
	Très faible	Faible	Moyenne	Bonne	Très bonne
Comment évaluez-vous votre connaissance des thématiques suivantes (entrer « 1 » pour votre choix)					
Les célébrations du 30e anniversaire du film Back To The Future et de la journée fétiche du 21 octobre 2015, date à laquelle le personnage principal se retrouvait dans le futur.			1		
Conférence COP21 sur le Climat à Paris			1		
Le BigData			1		
La politique canadienne				1	
Polémique autour d'une politique d'exclusion des homosexuels concernant leur don de sang				1	
Le parti français Front National				1	
La politique québécoise					1
La vague de support pour la mère de Taylor Swift suite à son cancer		1			
L'équipe de football Paris Saint-Germain et le football européen		1			
Le conflit en Syrie				1	
	Évaluateur #2				
	Très faible	Faible	Moyenne	Bonne	Très bonne
Les célébrations du 30e anniversaire du film Back To The Future et de la journée fétiche du 21 octobre 2015, date à laquelle le personnage principal se retrouvait dans le futur.				1	
Conférence COP21 sur le Climat à Paris		1			
Le BigData	1				
La politique canadienne				1	
Polémique autour d'une politique d'exclusion des homosexuels concernant leur don de sang		1			
Le parti français Front National		1			
La politique québécoise				1	
La vague de support pour la mère de Taylor Swift suite à son cancer			1		
L'équipe de football Paris Saint-Germain et le football européen			1		
Le conflit en Syrie		1			

ANNEXE E :
REQUÊTES SQL

Requête SQL pour créer une folksonomie

`SELECT tweet_tags.tag, COUNT( * ) FROM tweet_tags`	Sélection de l'ensemble des mots-clics et leur décompte à partir de la table où sont stockés les mots-clics (tweet_tags)
`LEFT JOIN tweets ON tweet_tags.tweet_id = tweets.tweet_id`	Création d'un lien avec le gazouillis d'origine afin d'y extraire la date et l'heure de publication
`WHERE DATE_ADD( NOW( ) , INTERVAL -#{time_frame} SECOND ) < tweets.created_at`	Condition filtrant les mots-clics des gazouillis publiés dans la dernière heure. La variable *time_frame* prend généralement la valeur d'une heure, mais fait partie des paramètres que nous pouvons modifier.
`GROUP BY tweet_tags.tag`	Regroupe les mots-clics (agrégation)
`ORDER BY COUNT( * ) DESC`	Ordonne les mots-clics en ordre décroissant.

Requête SQL calculant le nombre d'occurrence du mot-clic thématique

`SELECT hashtag COUNT( * ) FROM main_hashtags`	Sélection du mot-clic thématique (main_hashtags) et de son décompte à partir de la table où sont stockés les mots-clics (tweet_tags)
`LEFT JOIN tweet_tags ON main_hashtags.hashtag = tweet_tags.tag`	Création d'un lien entre le mot-clic thématique et son occurrence dans la table contenant tous les mots-clics (tweet_tags)
`LEFT JOIN tweets ON tweet_tags.tweet_id = tweets.tweet_id`	Création d'un lien avec le gazouillis d'origine afin d'y extraire la date et l'heure de publication
`WHERE DATE_ADD( NOW( ) , INTERVAL -#{time_frame} SECOND ) < tweets.created_at`	Condition filtrant les mots-clics des gazouillis publiés dans la dernière heure. La variable *time_frame* prend généralement la valeur d'une heure, mais fait partie des paramètres que nous pouvons modifier.
`GROUP BY tag`	Regroupe toutes les occurrences du mot-clic thématique pour en donner le compte

BIBLIOGRAPHIE

Abu Abbas, Osama. 2008. « Comparison Between data Clustering Algorithms ». *The International Arab Journal of Information Technology*, vol. 5, no 3, p. 320-325.

Anyanwu, Kemafor, Angela Maduko et Amit Sheth. 2007. « Sparq2l: towards support for subgraph extraction queries in rdf databases ». Dans *Proceedings of the 16th international conference on World Wide Web*, p. 797-806. ACM.

Bourion, Évelyne. (2001). *L'aide à l'interprétation des textes électroniques*. Université Paris 10, Paris.

Caron, Yves. 2004. « Contribution de la loi de Zipf à l'analyse d'images ». Tours.

Cattuto, Ciro, Vittorio Loreto et Luciano Pietronero. 2007. « Semiotic dynamics and collaborative tagging ». *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no 5, p. 1461-1464.

Clauset, Aaron, Cosma Rohilla Shalizi et M.E.J. Newman. 2009. « Power-law distributions in empirical data ». *SIAM Review*, no 4, p. 661-703.

Déglise, Fabien. 2010. « Les comportements humains cartographiés ». *Le Devoir*, 14 novembre 2010.

Dhiraj, Murthy. 2012. « Towards a Sociological Understanding of Social Media: Theorizing Twitter ». *Sociology*, vol. 46, no 6, p. 1059-1073.

Ding, Ying, Elin K. Jacob, Zhixiong Zhang, Schubert Foo, Erjia Yan, Nicolas L. George et Lijiang Guo. 2009. « Perspectives on social tagging ». *Journal of the American Society for Information Science and Technology*, vol. 60, no 12, p. 2388-2401.

Dye, Jessica. 2006. « Folksonomy ». *EContent*, vol. 29, no 3, p. 38-43.

Eschenbach, Carola et Michael Gruninger. 2008. *Formal ontology in information*

systems: proceedings of the Fifth International Conference (FOIS 2008). Coll. « Frontiers in artificial intelligence and applications v. 183 ». Amsterdam, Netherlands ; Fairfax, VA : IOS Press.

Evandro, L.T. et Cunha Paradela. 2011. « Analyzing the dynamic evolution of hashtags on twitter: a language-based approach ». Dans *Workshop on Language in Social Media - LSM2011*. Portland : aclweb.org. En ligne. <http://www.academia.edu/2541577/Analyzing_the_dynamic_evolution_of_hashtags_on_twitter_a_language-based_approach>. Consulté le 16 avril 2013.

Fraser, Nancy. 2003. « Repenser l'espace public : une contribution à la critique de la démocratie réellement existante ». Dans *Où en est la théorie critique?*, p. 103-134. La Découverte. Paris.

Frath, Pierre, Rochdi Oueslati et François Rousselot. 2000. « Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques ». *Ingénierie des connaissances. Évolutions récentes et nouveaux défis*, p. 291-304.

Fu, Wai-Tat. 2008. « The microstructures of social tagging: a rational model ». Dans *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, p. 229-238. ACM.

Guy, Marieke et Emma Tonkin. 2006. « Folksonomies: Tidying up Tags? ». *D-Lib Magazine*, vol. 12, no 1. En ligne. <http://www.dlib.org/dlib/january06/guy/01guy.html#21>. Consulté le 13 avril 2015.

Hahn, Nadja. 2013. *What good is Twitter? The value of social media to public service journalism*. Coll. « Eurovision Media Strategy Publication ». London, UK : London School of Economics and Political Science. En ligne.

<http://www.lse.ac.uk/media@lse/Polis/home.aspx>. Consulté le 9 juillet 2015.

Hintikka, Merrill B. et Jaakko Hintikka. 1991. *Investigations sur Wittgenstein*. Editions Mardaga.

Hong, Lichan, Gregorio Convertino et Ed H. Chi. 2011. « Language Matters In Twitter: A Large Scale Study. » Dans *ICWSM*.

Hotho, Andreas, Robert Jäschke, Christoph Schmitz, Gerd Stumme et Klaus-Dieter Althoff. 2006. « Folkrank: A ranking algorithm for folksonomies ». Dans *LWA*, vol. 1, p. 111–114.

Huberman, B. et S. Golder. 2005. *The Structure of Collaborative Tagging Systems*. cs/0508082.

Hummel, John E. et Keith J. Holyoak. 1997. « Distributed representations of structure: a theory of analogical access and mapping. » *Psychological Review*, vol. 104, no 3, p. 427.

Jiang, Jay J. et David W. Conrath. 1997. « Semantic similarity based on corpus statistics and lexical taxonomy ». *arXiv preprint cmp-lg/9709008*. En ligne. <http://arxiv.org/abs/cmp-lg/9709008>. Consulté le 19 mars 2015.

Kadlec, Jan. 2010. « Measures of semantic similarity in folksonomies ». En ligne. <http://projekter.aau.dk/projekter/files/61077941/1275544766.pdf>. Consulté le 9 avril 2015.

Kim, Hak Lae, Sung-Kwon Yang, Seung-Jae Song, John G. Breslin et Hong-Gee Kim. 2007. « Tag Mediated Society with SCOT Ontology. » Dans *Semantic Web Challenge*.

Kochut, Krys J. et Maciej Janik. 2007. « SPARQLeR: Extended SPARQL for semantic association discovery ». Dans *The Semantic Web: Research and Applications*, p. 145-159. Springer.

Laniado, David et Peter Mika. 2010. « Making sense of twitter ». Dans *The Semantic Web–ISWC 2010*, p. 470–485. Springer.

Lemaire, Patrick. 1999. *Psychologie cognitive*. Belgique : De Boeck Supérieur.

Ley, Tobias et Paul Seitlinger. 2010. « A cognitive perspective on emergent semantics in collaborative tagging: the basic level effect ». Dans *Proceedings of the Workshop on Adaptation in Social and Semantic Web (SASWeb 2010). CEUR Workshop Proceedings*, vol. 590, p. 13–18. Citeseer.

Liu, Chunnian, Dehui Yang et Yonglong Wang. 2011. « Domain ontology and semantic web applications for study of web competitive intelligence analysis system ». *Int. J. Web Science*, vol. 1, no 1/2, p. 99-113.

Liu, Tao, Shengping Liu, Zheng Chen et Wei-Ying Ma. 2003. « An evaluation on feature selection for text clustering ». Dans *ICML*, vol. 3, p. 488–495.

Loureiro-Koechlin, Cecilia et Tim Butcher. 2013. « The Emergence of Converging Communities via Twitter ». *The Journal of Community Informatics*, vol. 9, no 3.

Magnuson, Lauren. 2013. « Folksonomies: Meaning, Discourse, and Information Retrieval ». Dans *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*.

Martinez, William. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*. Université de la Sorbonne nouvelle – Paris 3, Paris.

Mathes, Adam. 2004. « Folksonomies - cooperative classification and communication through shared metadata ».

McConchie, Alan. 2015. « The Pop vs. Soda Page ». En ligne. <http://popvssoda.com/>. Consulté le 8 mai 2015.

Mika, Peter. 2005. « Ontologies are us: A unified model of social networks and semantics ». Dans *The Semantic Web–ISWC 2005*, p. 522-536. Springer.

Millette, Mélanie. 2015. « L'usage des médias sociaux dans les luttes pour la visibilité : le cas des minorités francophones au Canada anglais ». Montréal : UQAM.

Millette, Mélanie et Sylvain Rocheleau. 2014. « Tactiques de mise en visibilité : Usage de Twitter par des acteurs des minorités franco-canadiennes ». Dans *Actes du Colloque international Communication électronique, Cultures et Identités (CECI)*, sous la dir. de S. Zlitni, F. Liénard, D. Dula, et C. Crumière, p. 487-504. Le Havre, France : Éditions Klog.

Nikam, Mr KK et A. N. Mulla. 2014. « Pattern Deploying and Pattern Evolving Approaches for Text Mining ». *International Journal of Research Studies in Computer Science and Engineering*, vol. 1, no 4, p. 53-62.

Oldenburg, Ray. 1999. *The Great Good Place: Cafes, Coffee Shops, Bookstores, Bars, Hair Salons, and Other Hangouts at the Heart of a Community*. Third Edition edition. New York : Berkeley, Calif. : Da Capo Press.

Oussalah, M., F. Bhat, K. Challis et T. Schnier. 2013. « A software architecture for Twitter collection, search and geolocation services ». *Knowledge-Based Systems*, vol. 37, p. 105-120.

Paolillo, John C. et Elijah Wright. 2006. « Social network analysis on the semantic web: Techniques and challenges for visualizing FOAF ». Dans *Visualizing the semantic web*, p. 229-241. Springer.

Passant, Alexandre et Philippe Laublet. 2008. « Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. » Dans *LDOW*.

Peterson, Elaine. 2006. « Beneath the Metadata: Some Philosophical Problems with

Folksonomy ». *D-Lib Magazine*, vol. 12, no 11.

Posch, Lisa, Claudia Wagner, Philipp Singer et Markus Strohmaier. 2013. « Meaning as collective use: predicting semantic hashtag categories on twitter ». Dans *Proceedings of the 22nd international conference on World Wide Web companion*, p. 621–628. International World Wide Web Conferences Steering Committee.

Rocheleau, Sylvain et Mélanie Millette. 2014. « Political Uses of Twitter Hashtag Co-Occurrences: Efficiently Exploring Organizational Tagging With Small Data ». Dans *Hashtag Publics*, sous la dir. de Nathan Rambukkana. Coll. « Digital Formations Series 3 ». Toronto : Peter Lang Press.

Rosch, Eleanor. 1975. « Cognitive reference points ». *Cognitive psychology*, vol. 7, no 4, p. 532-547.

Rosch, Eleanor H. 1973. « Natural categories ». *Cognitive psychology*, vol. 4, no 3, p. 328-350.

Rosch, Eleanor, Carolyn B Mervis, Wayne D Gray, David M Johnson et Penny Boyes-Braem. 1976. « Basic objects in natural categories ». *Cognitive Psychology*, vol. 8, no 3, p. 382-439.

Russell, Stuart J., Peter. Norvig et Ernest. Davis. 2010. *Artificial intelligence: a modern approach*. Upper Saddle River, NJ : Prentice Hall.

Sen, Shilad, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper et John Riedl. 2006. « Tagging, communities, vocabulary, evolution ». Dans *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, p. 181–190. ACM.

Shannon, Claude. 1948. « A Mathematical Theory of Communication ». *Bell System*

Technical Journal, vol. 27, p. 379-423 et 623-656.

Sinha, Rashmi. S.d. « A cognitive analysis of tagging ». *Rashmi's blog*. En ligne. <http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/>. Consulté le 14 avril 2015.

Soulé, B. 2007. « Observation participante ou participation observante ». *Usages et justifications de la*, vol. 27, no 1, p. 127–140.

Specia, Lucia et Enrico Motta. 2007. « Integrating folksonomies with the semantic web ». Dans *The semantic web: research and applications*, p. 624–639. Springer.

Tate, Ryan. 2013. « Facebook introduces hashtags (Wired UK) ». *Wired UK*, juin 2013. En ligne. <http://www.wired.co.uk/news/archive/2013-06/13/facebook-hashtags>. Consulté le 11 juin 2015.

Trant, Jennifer. 2009. « Studying social tagging and folksonomy: A review and framework ». *Journal of Digital Information*, vol. 10, no 1.

Twitter Inc. 2015. « About Twitter ». *Twitter*. En ligne. <https://about.twitter.com/company>. Consulté le 7 mai 2015.

Vidhya, K.A. et G. Aghila. 2010. « Text Mining Process, Techniques and Tools : an Overview ». *International Journal of Information Technology and Knowledge Management*, vol. 2, no 2, p. 613-622.

Wakamiya, Shoko, Ryong Lee et Kazutoshi Sumiya. 2011. « Crowd-Powered TV Viewing Rates: Measuring Relevancy between Tweets and TV Programs ». *Computer Science*, vol. 6637, p. 390-401.

Weigend, Andreas S., Erik D. Wiener et Jan O. Pedersen. 1999. « Exploiting Hierarchy in Text Categorization ». *Information Retrieval*, vol. 1, no 3,

p. 193-216.

Weinberger, David. 2005. « Taxonomies to Tags: From Trees to Piles of Leaves ». *Release 1.0*, vol. 23, no 2.

Winston, Morton E., Roger Chaffin et Douglas Herrmann. 1987. « A Taxonomy of Part-Whole Relations ». *Cognitive Science*, vol. 11, no 4, p. 417-444.

Zhang, Dell, Jun Wang, Deng Cai et jinsong Lu. 2010. « Self-taught hashing for fast similarity search ». Dans *33rd international ACM SIGIR conference on Research and development in information retrieval*, p. 18-25. New York : ACM.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. Addison-Wesley Press.