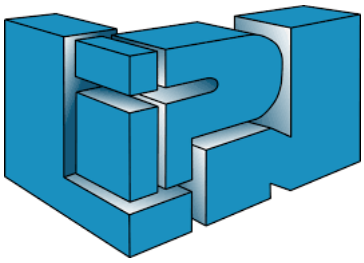


Institut National des Langues et Civilisations Orientales

Département Texte, Informatique et Multilinguisme

Comparaison entre la méthode symbolique et la méthode par apprentissage dans l'efficacité de la détection thématique d'articles de presse

Master
Traitement Automatique des Langues

Parcours :

Ingénierie Multilingue

par

Nicolas Scarcella

Directeur de mémoire :

Mathieu Valette

Encadrant :

Emmanuel Cartier

Année universitaire 2017/2018

TABLE DES MATIÈRES

Avant-propos

Le cadre de ce mémoire s'inscrit dans un projet visant à « doter les observateurs de la langue française d'un moteur de recherche s'appuyant sur la collection « Actualités » du dépôt légal du Web, conservée à la Bibliothèque Nationale de France (BnF) depuis 2010. »[E. Cartier, 2017].

Trois laboratoires dont le Laboratoire d'Informatique de Paris Nord (LIPN) où mon stage a été effectué contribuent à son élaboration.

Baptisé « Néonaute », cet outil a pour rôle de « suivre l'implantation des néologismes à partir des collections du dépôt légal du Web (bnf) ». [E. Cartier, 2017] ; « Un outil novateur d'analyse de données textuelles contemporaines, avec une fouille basée sur le plus gros corpus [journalistique] du Web français actuellement disponible. » [Langue et numérique, 2017].

La Bibliothèque Nationale de France archive depuis une vingtaine d'année des articles de presse qui paraissent sur le web ; une manne d'information colossale utile pour suivre le cycle de vie des lexies dont, bien-sûr, les néologismes.

Deux plateformes Logoscope[1] et Néoveille[2] se consacrent au repérage des néologismes par le croisement de diverses méta-informations et données textuelles.

Les articles [S. Ollinger, M. Valette, 2010] et [D. Bernhard, L. Bruneau, I. Falk, C. Gérard, A.L. Rosio, 2016] nous apprennent que le Logoscope use d'un processus de détection en plusieurs étapes avec tout d'abord la récupération quotidienne des articles provenant de plusieurs journaux, puis leur lecture et l'extraction de mot-candidats, aidée notamment d'une liste d'exclusion pour ensuite isoler les mauvais mots candidats en s'aidant du contexte du mot, du nombre d'occurrences ou de sa relative récence.

Enfin, un automate documente une base lorsque le néologisme est réemployé afin d'observer son évolution. Voici l'exemple de l'évolution d'emploi du mot « uberisation », tiré de [D. Bernhard, L. Bruneau, I. Falk, C. Gérard, A.L. Rosio, 2016].

1 http://logoscope.unistra.fr
2 https://lipn.univ-paris13.fr/neoveille/html/login.php?action=login

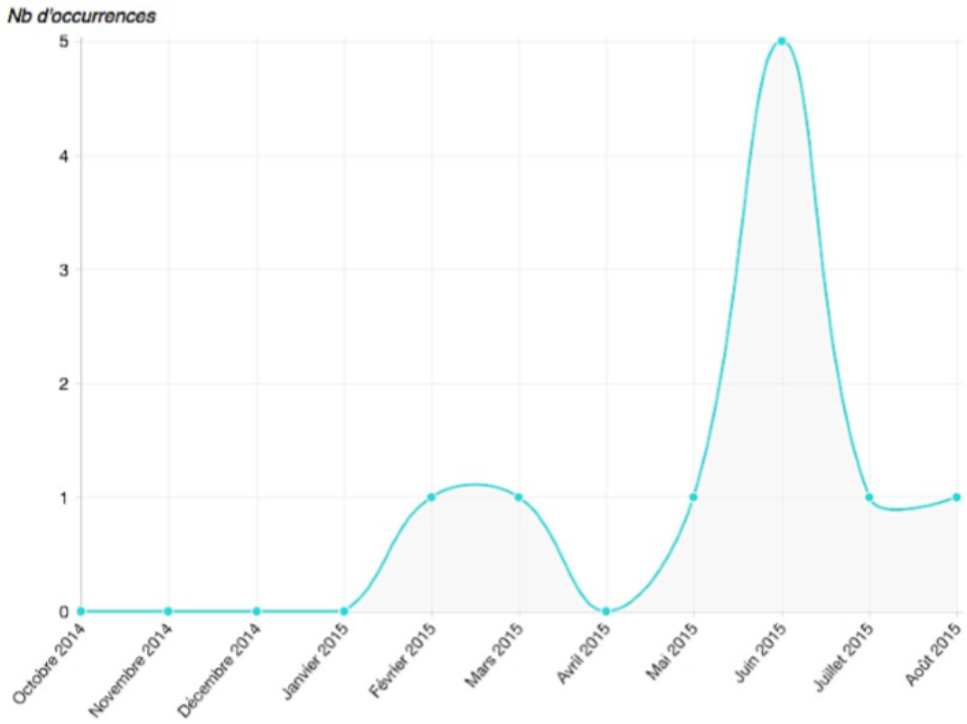


Figure 1 : Graphique de l'évolution du néologisme « ubérisation »

Grâce à cela, précise l'article [D. Bernhard, L. Bruneau, I. Falk, C. Gérard, A.L. Rosio, 2016] l'historien, le sociologue ou le journaliste peuvent interroger le Logoscope pour connaître les changements ou les événements qui se sont produits pour un sujet d'intérêt particulier, pendant une période donnée.

La plateforme Neoveille, quant à elle, vise à suivre, repérer et analyser automatiquement les néologismes dans 7 langues (français, grec, polonais, tchèque, portugais et portugais-brésilien, chinois et russe). Elle étudie également la notion d'innovation sémantique et propose de nouvelles procédures d'identification de ces nouveaux emplois.

Le projet Neonaute, travaillant sur la répertorisation des néologismes, s'inscrit donc dans un travail de veille lexicale. Cette objectif lexicographique, nous dit [M. Valette, 2009], est justifié par le fait que certaines grandes bases terminologiques telle que Frantext ne couvrent pas la période actuelle ce qui implique cette nécessité de recenser les phénomènes néologiques. La créativité lexicale au travers des néologismes, ajoute [S. Ollinger et M. Valette, 2010] a pour but d'enrichir les lexiques, de créer des métadonnées ou de sélectionner des contextes caractéristiques.

Ce projet associe nombre de tâches allant de la recherche d'entités nommées à l'indexation d'articles en passant par de la détection thématique de ces mêmes articles. C'est sur cette dernière tâche que le travail de ce mémoire a porté, à savoir une analyse des contenus textuels des articles de presse.

Remerciements

Je remercie l'ensemble du corps professoral du Master TAL de l'INALCO qui m'a transmis la formation nécessaire au bon déroulement de mon stage.

Je remercie, en particulier, Monsieur Mathieu Valette, enseignant à l’INACO et directeur du présent mémoire pour ses conseils et avis dans sa réalisation et Monsieur Emmanuel Cartier, mon tuteur de stage, maître de conférences en linguistique informatique à l’université Paris 13 et chercheur au LIPN pour son aide en informatique.

Également, je tiens à remercier Monsieur Christophe Gérard, maître de conférence à l’université de Strasbourg et chercheur au laboratoire Linguistique, Langues, Parole (LILPA) de Strasbourg pour nos échanges sur les thématiques et la néologie.

Résumé

Ce mémoire porte sur l'étude de la détection thématique d'articles de presse pour une recherche sur la néologie. L'objet du travail réside dans la confrontation de la méthode symbolique, appuyée sur des listes et un vocabulaire contrôlé, avec la méthode d'apprentissage automatique, basée sur des algorithmes et calculs mathématiques.

Ces deux procédures offrent l'analyse binaire visant à doter la machine d'une intelligence artificielle capable de détecter correctement les sujets abordés dans un texte dans le but de localiser les contextes qui aideront, dans un projet de détection de néologismes, à s'approcher de leur sens sémantique.

Dans un premier temps, le travail consiste en une réflexion sur le moteur de détection déjà existant puis, dans un second temps, en l'exploration de nouvelles techniques permettant d'optimiser la détection thématique.

Mots clés

Machine Learning, Clusterisation, Classification, Etiquetage, Ngrams, Matching

Introduction

Le langage donne lieu à des innovations qui sont hors des habitudes langagières habituelles. Une innovation qui s'établit sur deux niveaux historiques selon [D. Bernhard, L. Bruneau, I. Falk, C. Gérard, A.L. Rosio, 2016] :

- Celui des langues et dialectes, par la création d'un nouvel affixe, lexème ou nouvelle locution verbale,

- Celui des traditions discursives en créant et inventant un genre de discours.

Venons-en à la néologie proprement dite. Celle-ci, nous dit [S. Ollinger, 2011], peut être définie comme « l'ensemble des unités lexicales nouvelles dans un état de langue donné. »

Citons pour exemple, le terme « trumpisme », unité lexicale issue du nom propre « Trump » qui n'aurait sans doute jamais vu le jour si Monsieur Trump n'avait pas été élu président. Ces néologismes sont donc parfois la conséquence d'évènements temporels et factuels.

À présent, voyons le lien avec la thématique. L'article [C. Gérard, A. Grezka, L. Mercè, 2017] précise que « Tout néologisme apparaît dans un domaine particulier et se rattache nécessairement au thème développé au sein d'un texte ». C'est la raison pour laquelle, la détection de la thématique du texte et, plus localement, du contexte est une piste privilégiée pour trouver le sème d'un néologisme et obtenir une approche de sa signification lexicale. Trouver le contexte et vous aurez, au moins en partie, une définition du néologisme qui s'y trouve ; tel est le raisonnement. Et [D. Bernhard, I. Falk, C. Gérard, septembre 2014] de rajouter que « Tout néologisme s'inscrit nécessairement dans la progression thématique de son texte d'accueil ». Les mots ne sont pas employés par hasard dans un contexte et ils se rattachent de près ou de loin au thème local.

L'article [D. Bernhard, I. Falk, C. Gérard, septembre 2014] précédemment cité, donne l'exemple du terme « loto-impôt » qui conjoint les thèmes du « Jeu » et de la « Finance », en se basant sur la citation suivante :

> « Nous n'aimons pas les impôts, mais nous aimons les jeux de hasard. Alors pourquoi ne pas jouer au loto-impôts ? En payant 10 % de majoration on pourrait gagner le remboursement de l'impôt payé et une exonération d'impôt pendant 10 ans. (Événement du jeudi, 07/11/1985, courrier des lecteurs). »

Dans cet exemple, nous voyons deux termes, semblant appartenir à des domaines différents, réunis pour créer un néologisme dont le sens ressort grâce à l'identification des thèmes « Finance » et « Jeu ».

Plusieurs domaines pour une seule lexie, voilà qui n'est pas sans évoquer la polysémie. Cette dernière justifie d'autant plus le recours à la détection thématique puisque la langue française comporte pléthore d'homonymes et de termes polysémiques. Un mot comme « batterie » par exemple, sera utilisé en musique, en cuisine, en électronique… c'est à dire dans des contextes très variés.

Identifier la thématique, assure [D. Bernhard, I. Falk, C. Gérard, septembre 2014], mène à l'identification des innovations polysémiques (ex. flûte [musique] ; flûte [alimentation], flûte [marine]).

Cette polysémie est, selon [M. Valette, 2009], un « artefact résultant de l'isolement du mot, de sa décontextualisation » ; par conséquent, isoler le terme en question n'aiderait pas vraiment à dégager son sens. L'étude thématique ciblant l'analyse du contexte est donc indispensable.

L'article [F. Rastier, M. Valette, 2009] ajoute que, dès lors que la signification de toute unité n'est déterminable que par son contexte, elle varie avec lui ; en cela, le concept de néologie rejoint celui de néosémie, lequel, précise [F. Rastier, M. Valette, 2009] décrit la formation et l'évolution d'emplois nouveaux. Ainsi, néosémie et condition d'emploi vont de paire. Partant de là, [F. Rastier, M. Valette, 2009] précise qu'il est important de rattacher tout nouvel emploi avec le domaine, c'est-à-dire le groupe de classes sémantiques lié à une pratique sociale dans lequel il apparaît.

Le même article [F. Rastier, M. Valette, 2009] détaille ces changements d'emplois selon trois phénomènes :

- La domanialisation, cas où le lexème et son sémème sont intégrés dans un domaine auquel ils n'appartenaient pas. L'exemple du mot « grogne », autrefois réservé à l'animal puis désormais synonyme de mécontentement populaire, l'illustre bien.

— La grogne des urgentistes rebondit sur les généralistes. (Site liberation.fr, 31.03.2005)

- La dédomanialisation, évoquant le fait que l'emploi d'un sémème puisse carrément disparaître de son domaine initial pour être intégré dans un autre. Le mot « percuter » par exemple, qui signifiait « heurter », n'est plus utilisé aujourd'hui que dans le sens de « comprendre » ou « réagir rapidement ».

— Mais j'ai un problème avec l'oral. […] le temps que je percute ce que je dois dire et que je trouve les bons mots et le bon ordre il y a déjà 3 heures de passées !!! (Forum cultureco.com, 5.03.2005)

Par ailleurs, le changement est significatif puisque le terme « percussion », substantif de « percuter » s'est cantonné au domaine de la musique et il a fallu inventer le mot « percutage » pour retrouver un substantif ad hoc.

- La redomanialisation enfin, marque un changement complet de classe sémantique et de domaine.

L'exemple le plus typique est sans doute celui de « caviar », mot employé de manière égal pour le luxe ou/et l'aliment lui-même, qui se retrouve dans des contextes surprenants.

— Aujourd'hui la StarAc c'est du caviar comparé à ce qu'on a déjà vécu dans la musique industrielle (Site etnoka.fr, 29.09.2004)

— Un p'tit film d'animation hilarant, le début est un poil long mais restez jusqu'au bout : c'est du caviar (Site fumezlamoquette.com, 5.07.2004)

Tout ces points témoignent, d'une part de l'importance de la détection thématique d'un terme candidat à la néologie pour révéler son sens sémantique, voire son évolution de sens lorsqu'on effectue cette détection de façon chronologique, et d'autre part de sa potentialité à, nous dit [F. Rastier, M. Valette, 2009] « participer à l'émergence de nouveaux domaines ou sous-domaines ».

Outre les néologismes, la détection thématique, comme nous l'indique [M. Boumghar, C. Christophe, J. Velcin, 2018], offre un intérêt particulier pour le commerce et spécialement le domaine du retour client pour analyser les thématiques émergentes. En effet, la fouille d'opinions passe nécessairement par la détection des thématiques puisque l'on cherche à savoir de quoi parlent les gens dans leurs messages.

Plus ciblé, le cotexte ou l'environnement proche du mot qu'évoque [D. Bernhard, I. Falk, C. Gérard, septembre 2014] sert quand à lui de « preuve d'existence de la création lexicale d'exemple

d'emploi (à la manière d'une citation de dictionnaire) et surtout d'interprétant pour comprendre le sens du néologisme. »

Dans la continuité des travaux réalisés pour le projet Néonaute, nous tenterons de répondre à la problématique suivante :

Le choix de préférence entre méthode symbolique et méthode par apprentissage se justifie-t-il pour l'amélioration de l'outil de détection thématique du projet Neonaute ?

La réponse au sujet sera axée sur plusieurs réflexions. Dans une première partie, nous effectuerons un état de l'art sur la détection thématique, avant d'étudier dans une deuxième partie les déficiences de l'analyseur thématique du projet Neonaute. Enfin, dan une troisième partie, nous réaliserons plusieurs tests d'amélioration des techniques pour confronter la méthode symbolique et la méthode par apprentissage.

I. État de l’art

Dans l’analyse textuelle, la détection thématique a fait l’objet de plusieurs recherches qui suscitèrent de nombreuses réflexions sur les méthodes de détection et les moyens ainsi que la procédure à employer.

A) Généralités sur la détection thématique

Avant de nous intéresser dans la détection thématique et les travaux qui lui ont été accordés, citons l’article [D. Bernhard , I. Falk, C. Gérard, 2014] qui rappelle que le contexte d’un mot, et dans notre cas, un mot candidat à la néologie, est très lié à la création de mots. L’article affirme que la détection d’un contexte apporte une aide pour la détection des mots nouveaux, plus que pour celles des mots plausibles. En effet, il ne faut pas négliger le problème des faux candidats (variations idiosyncrasiques, entités nommées inconnues ou non étiquetées comme nom propre, mauvaise orthographe, mots étrangers en contexte hétérolingue…, exemple cités par les articles [S. Ollinger, M. Valette, 2010] et [D. Bernhard, L. Bruneau, I. Falk, C. Gérard, A.L. Rosio, 2016]), termes à écarter néanmoins de la liste des néologismes.

L’importance de la détection thématique ressort dans plusieurs articles comme [D. Bernard, I. Falk, C. Gérard, 2014] qui évoque trois types de traits permettant d’identifier les mots ou signes inconnus, c’est-à-dire les candidats à la néologie :

- Le trait de forme ou construction du token tel que son nombre de caractères, la présence ou non de majuscules ou de signes non alphabétiques, ou sa fréquence...)

- Le trait morpho-lexical, entendons la présence de préfixes ou suffixes particuliers ainsi que l’orthographe…

- Le trait lié à la thématique, à savoir le contexte d’apparition du token.

Selon les résultats de l’article, « Le groupe de traits thème (représentant les thématiques détectées pour le contexte textuel des mots/signes inconnus) permet d’identifier le plus grand nombre de vrais néologismes ».

L’article [M. Ballabriga, 2005] note que pour la recherche de contexte de néologisme, le terme de taxème serait préférable à celui de domaine. En effet, il est « le reflet de situations de choix dans des pratiques concrètes ou théoriques ».

Par exemple, illustre [M. Ballabriga, 2005], « les sémèmes {‘métro’, ‘train’, ‘autobus’, ‘autocar’} relèvent du domaine //transports// (moyens collectifs) articulé en deux taxèmes, le taxème //ferré// (comprenant les sémèmes ‘métro’ et ‘train’ opposés par les sèmes spécifiques /intra-urbain/ et /extra-urbain/) et le taxème //routier// (comprenant les sémèmes ‘autobus’ et ‘autocar’ différenciés, eux aussi, par les sèmes spécifiques /intra-urbain/ et /extra-urbain/). A noter que la présentation qui ferait de /intra-urbain/ et /extra-urbain/ des traits microgénériques et de /ferré/ et /routier/ des traits spécifiques correspondrait aux situations pragmatiques les plus courantes : en principe, on choisit un moyen de transport en fonction de sa destination, et non parce qu’il est ferré ou routier. ».

Un autre article, [C. Gérard, A. Grezka, M. Lorente, 2017] insiste sur la différence entre les termes thèmes et domaines. Définissant le premier comme type de contenu construit en différents endroits d'un texte et le second comme catégorie figée ayant un lexique bien circonscrit.

La notion de thème est principale car il constitue une information cruciale sur le contexte d'usage du néologisme. Il ne convient pas de se cantonner à un domaine bien délimité car les néologismes peuvent évoluer et changer de domaines comme l'a expliqué [F. Rastier, M. Valette, 2009].

Tout texte, ajoute [C. Gérard, A. Grezka, M. Lorente, 2017] est nécessairement pluri-thématique. C'est pourquoi, le néologisme est un réseau complexe de thèmes. Le domaine n'est utile que pour le classement des termes employés dans des textes spécialisés. Hors les textes journalistiques, contenus ciblés par le projet Neonaute, ne sont pas spécialisés et chaque journal offre des architectures de rubriques qui ne sont pas toujours similaires.

Bien qu'un article de presse soit bien souvent casé dans une seule rubrique d'un journal, il contiendra presque toujours des sous-thèmes. Ce problème des textes plurithématiques entraîne un besoin de détection plurithématique. Le souci reste d'établir la nomenclature des thèmes afin de catégoriser les contenus.

B) La construction d'une hiérarchie thématique

L'analyse thématique des textes, indique [B. Grau, N. Hernandez, 2003], économise du temps sur leur lecture et cible les propos, ce qui permet une pré-classification, préambule à une exploitation de la néologie. Mais un des problèmes réside dans le choix des thèmes, ces sortes de cases au sein desquelles devront être rangés les textes.

Divers objectifs, détaille [D. Bernhard, I. Falk, C. Gérard, septembre 2014], sont à parcourir lors d'une analyse thématique :

- « Détection de thèmes dans un texte afin de le segmenter en sous-unités thématiquement homogènes ». Le but étant de repérer les changements de thèmes. L'idée est que tout changement de vocabulaire indique un changement de thème ; d'où un découpage et une segmentation préalable en mots ou séquence de mots.

- Usage du topic modeling pour « identifier les thèmes 'latents' dans un corpus textuel » suivant un phénomène de clusterisation regroupant les termes selon diverses caractéristiques.

Notons que, bien qu'un texte contienne plusieurs thématiques, celles-ci n'ont pas toujours le même poids. Plusieurs moyens existent pour effectuer cette tâche et en particulier les systèmes LDA (Latent Dirichlet Allocation).

Plus récemment, la détection thématique dans le domaine commercial du retour client à l'aide du modèle probabiliste LDA a déjà été réalisé dans un autre article [C. Christophe, J. Velcin, M. Boumghar, 2018]. Ce modèle permet la création de thématiques à un temps donné. Ainsi, une thématique est décrite grâce à sa distribution de probabilité sur les termes et sur les documents.

La pertinence pour trier les néologismes ou plus largement les mots et signes inconnus semble assez bien revenir aux méthodes statistiques d'apprentissage automatique.

- « Annotation thématique à partir de thèmes prédéfinis » : réalisé à l'aide de l'outil ThemeEditor et son aspect de coloration thématique (affectation d'une couleur à chaque thème).

L'outil dénommé Termite indique, selon la thématique, les mots les plus saillants. En clair, l'article explique que « chaque mot inconnu est associé à deux textes : d'une part l'ensemble des phrases le comprenant, d'autre part les documents où il apparaît ».

L'article [D. Bernard, I. Falk, C. Gérard, 2014] précise que « la thématique se laisse concevoir de trois points de vue distincts : » :

- « Comme exploitation dérivée de la variable domaine lexical, elle permet d'obtenir un classement des néologismes ».

Un outil, dénommé Wortwarte et cité dans ce même article propose une classification thématique des néologismes. En clair, à partir d'un thème il est possible d'obtenir une liste des néologismes associés.

- « Comme variable textuelle (globale). Il s'agit de l'organisation par rubrique des journaux (politique, économie, culture, sport, etc.). Ce point de vue requiert l'observation de classification déjà existante.

- « Comme variable textuelle (locale), l'analyse thématique permet de caractériser tout texte individuel par une constellation de thèmes principaux et secondaires — tels que la représentent des outils comme ThemeEditor ou Termite. »

L'utilisation de topic modeling pour la création de clusters thématiques afin d'aider à cibler le contexte des mots candidats à la néologie a bien été testée. Comme l'explique [D. Bernhard, I. Falk, C. Gérard, juillet 2014], « un topic model est un modèle probabiliste permettant de déterminer des sujets ou thèmes abstraits dans une collection de documents ». L'article ajoute que le modèle le plus fréquemment utilisé est le LDA.

Dans l'article [D. Bernhard, I. Falk, C. Gérard, juillet 2014], les auteurs ont justement travaillé sur du « topic modeling » dans le but de détecter automatiquement l'apparition de nouveaux sens chez les mots connus. Cette observation d'un changement sémantique s'est faite sur la base de l'observation du nouveau sens de geste du mot « quenelle ». La première étape a été de constituer un corpus contenant le terme étudié. Ce corpus n'a pas été formé d'articles mais de paragraphes d'articles afin d'éviter la pollution lié à la polythématique et n'avoir que le contexte du terme. Dans le but de repérer le bon topic, des listes de mots pour analyser par comparaison le contexte du mot quenelle » ont été produits. Le mémoire de [L. Bruneau, 2014] ayant travaillé à la détection thématique du projet Néonaute procède par méthode symbolique à l'aide de la comparaison entre les articles et des listes de vocabulaire. Ces listes, pour la plupart sont liées à un travail de topic modeling, en particulier à l'aide de l'outil Millet.

Le topic modeling révèle un problème de dénomination des clusters. Quelle thématique attribuer à tel rassemblement de mots ? Il faut observer les mots regroupés et trouver un nom pour tous les référés.

Les entreprises elles-aussi se sont lancées dans la détection thématique de contenu. En 2015, un article[3] sur internet a mentionné que la société Proxem détenait un outil appelé « Ontology-Based Topic Detection » permettant de « dégager les différentes thématiques d'un document sur la base des connaissances organisées de Wikipedia ». Il indique les thématiques identifiées, sans passer par

3 Outil de détection thématique de l'entreprise Proxem : www.programmez.com

du topic modeling qui « renvoie des listes de mots caractéristiques du document dont il faut ensuite déduire un dénominateur commun » mentionne l'article.

C) L'interaction entre thèmes et listes de vocabulaire

Au début des années 2000, des outils, notamment ThemeEditor, pour l'acquisition de classes sémantiques offraient la possibilité d'un coloriage thématique des documents d'un corpus. L'objectif visait « l'étude des usages des mots dans les productions langagières (textes, dialogues...) » [P. Beust, 2002]. Cette étude de représentation de thème définit celui-ci comme le « sujet abordé dans un texte ou un corpus » [P. Beust, 2002].

Le repérage du thème se fait à l'aide d'un lexique, d'une liste de mots qui n'est autre que, précise [P. Beust, 2002], le moyen de représenter un thème. Puis, la liste d'exclusion « Morphalou » servit, par comparaison, à repérer les véritables néologismes.

L'article [P. Beust, 2002] indique aussi que le phénomène d'isotopie est particulièrement à relever car la récurrence de traits sémantiques est une indication sur le contexte et le domaine d'emploi. Leur analyse permet de saisir le sens d'un texte en dépit d'une lecture partielle. Par conséquent, le coloriage des thèmes sera d'affecter une couleur à chaque isotopie car repérer les isotopies, c'est repérer les thèmes. Un bon moyen utilisé est la recherche de n-grams. Cependant, relève [A. Guille, C. Favre, 2014], elle souffre de la critique de ne pas pouvoir capturer les relations entre des mots trop éloignés.

À présent, centrons notre propos sur l'équivocité des termes. Les homonymes entraînent un problème de polysémie impliquant qu'un terme apparaîtra dans plusieurs thèmes différents. Prenons comme exemple le mot « avocat » qui peut appartenir à un thème lié à la nourriture, aussi bien qu'un thème lié à la justice. Deux solutions sont exploitées pour coloriser correctement le terme.

Solution 1: prolonger le plus possible les isotopies du texte, c'est-à-dire favoriser la redondance. On attribue alors au mot la couleur la plus représentée dans le texte.

EX : En faisant mon marché, j'ai vu des poireaux , des concombres et des avocats.
En sortant du Tribunal, j'ai vu un avocat.

« Avocat » reçoit la bonne couleur car, dans le premier cas, la couleur verte domine et dans le deuxième cas le rouge domine.

Solution 2 : Analyse des voisinages des mots, on analyse la morphosyntaxe.

EX : En faisant mon marché, j'ai vu des poireaux , des concombres et des avocats.
En sortant du Tribunal, j'ai vu un avocat.

« Avocat » reçoit la bonne couleur car il est proche, dans le premier cas de concombres et poireaux et dans le deuxième cas de tribunal.

Les exemples paraissent identiques, c'est le mobile du choix de la couleur qui est différent.

L'on peut remarquer dès à présent qu'il ne faut pas que les autres mots soient aussi polysémiques mais en jouant sur le nombre de termes, on retombe théoriquement sur un même thème.

Il faut bien avoir en tête, rappelle [D. Bernhard, I. Falk, C. Gérard, juin 2014] que tout article est fortement susceptible d'être polythématique. Il peut y avoir plusieurs sujets et donc il convient de faire une estimation de proportion pour choisir les thèmes principaux.

Les travaux de Lauren Bruneau, dont le présent mémoire poursuit le travail, effectue la comparaison entre des listes de vocabulaire thématiques et les articles de presse. Cette comparaison se fait sur les lemmes. L'étiquetage des thèmes a été fait à l'aide de TreeTagger dans le but de différencier les catégories morphosyntaxiques pour limiter les ambiguïtés entre les thèmes. Par exemple « savoir » nom, n'est pas « savoir » verbe.

En guise d'étiqueteur, Treetagger a été choisi pour son efficacité mais notons que, pour la détection, spécifiquement de mots nouveaux, c'est, au dire de [I. Falk, D. Bernhard, C. Gérard , R. Potier-Ferry, 2014], StandfordTagger qui paraît être l'outil le plus adapté.

Analyser les listes permet de mettre en évidence leur structure lexicale indique [P. Beust, 2002].

Et, pour faciliter le remplissage des listes, le même article informe qu'il peut être nécessaire d'avoir recours à de l'extraction terminologique. On extrait automatiquement une liste de termes à partir d'un corpus spécialisé.

Si nous reprenons l'exemple de l'outil ThemeEditor, nous voyons dans [P. Beust, 2002] qu'il fournit des listes qu'il attribue à un thème, comme la liste (élus, élu, Lionel Jospin, assemblée, assemblées, premier ministre, ministre, ministres, président, présidents, député, députés, référendum local, ministère, commune, communes, vote, gouvernement) pour la politique.

Notons qu'une partie des listes des vocabulaires thématiques a été constituée au moyen du logiciel Antidote[4], un logiciel donnant des listes étiquetées de termes propres à un domaine (cf. Annexe 1 : Liste de vocabulaire thématique).

Au travers des réflexions précédentes, la deuxième partie du présent mémoire a pour objectif de commenter ces considérations sur la base des travaux réalisés par Lauren Bruneau dont le moteur de détection thématique sert pour le projet Néonaute, afin d'en analyser les faiblesses et pouvoir proposer des pistes d'amélioration.

4 https://www.antidote.info

II. L’analyseur thématique : corpus, thèmes et listes

Dans la suite des travaux antérieurs, l’objectif de cette partie va être de constater les améliorations possible pour la détection thématique conformément au projet Néonaute.

Mon travail s’inscrit dans la suite de celui de Loren Bruneau, personne ayant travaillé avec Christophe Gérard au laboratoire LILPA de Strasbourg. Son travail reprend un travail antérieur de détection thématique qu’elle a amélioré, notamment en terme de temps d’analyse qui était très long (amélioré par l’usage d’une double table de hashage), de coloration thématique jugée trop imprécise et de décompte des mots, peu correct auparavant.

Le système de coloration de [P. Beust, 2002] a été repris par [L. Bruneau, 2014] pour le rendue visuel de la détection thématique.

Au regard de mon domaine de compétences, j’ai souhaité poursuivre le travail rapporté par Lauren Bruneau mais en usant d’outils, n’ayant, semble-t-il, pas encore été utilisés pour la détection thématique.

A) Les Corpus

1) Le corpus journalistique : terrain de création lexicale

Le projet Néonaute s’est donné pour mission de repérer l’ensemble des néologismes français, ces innovations inhabituelles, et de suivre leur évolution thématique dans le temps. Pour ce faire, [S. Ollinger, 2011] exhorte à multiplier les corpus d’observation appartenant à la langue actuelle tout en prenant en compte les spécificités du corpus (typologie, date, auteur).

Le moteur de détection thématique doit travailler sur un corpus exclusivement journalistique. Ce corpus colossal est issu de l’archivage par la BNF, depuis plusieurs années, d’articles de presse et revues de tous types qui paraissent périodiquement sur le Net. L’intérêt pour la néologie est évoqué dans l’article [D. Bernhard, I. Falk, C. Gérard, juin 2014] qui affirme qu’un corpus journalistique est favorable à la création de mots. Néanmoins, on remarquera que les espaces d’expressions types réseaux sociaux sont aussi des viviers créateurs de nouveautés, mais la voix des journaux[5] aura autorité pour qu’un néologisme soit intégré au vocabulaire. Et même si l’heure est à internet et à l’auto rédaction via les blogs et autres tchats, la volonté de populariser un nouveau mot ne suffira pas sans une validation officielle.

Sur le même raisonnement, [M. Valette, 2009] précise : « Les néologismes non techniques (c’est-à-dire hors langue de spécialité) apparaissent le plus souvent dans des situations peu contraintes, mais pour qu’un mot intègre la langue, la tradition lexicographique impose qu’une autorité la valide. ».

Et [D. Bernhard, L. Bruneau, I. Falk, C. Gérard, A.L. Rosio, 2016] de compléter que les journaux sont une source privilégiée des veilleurs de néologie, les règles et normes expressives sont données par les genres de discours (écrit/oraux) et le style collectif de la publication.

5 Site de regroupement thématique des journaux : https://www.press-directory.com/presse-magazine/

L’on peut toutefois remarquer, selon [S. Ollinger et M. Valette, 2010] que, en fonction du genre d’écrit, l’apparition de néologismes est plus ou moins fréquente. C’est pourquoi, « parmi les genres argumentatifs du discours littéraire, le pamphlet est réputé créatif, l’essai est plus conservateur ». Les genres, à savoir les ensembles de normes de production des textes, présentent un potentiel néologique variable.

L’article mentionne aussi qu’il est possible que la production de néologisme, sur blog notamment, ne soit à exclure que provisoirement. Elle n’a pas de vocation lexicographique comme un article de presse mais il est possible que les journaux actent un nouveau terme apparu dans les réseaux sociaux. Le mot « hashtag » en est un bon exemple.

L’article [E. Cartier, 2018] mentionne également que : « il est bien connu que les néologismes sont en plus grand nombre dans les discours oraux que dans les discours écrits, d’une part et qu’une partie non négligeable des néologismes naissent dans des groupes sociaux ou socio-économiques particuliers : c’est ainsi que par exemple nombre d’anglicismes actuels proviennent du domaine informatique puis, de par la démocratisation de ce domaine, passent ensuite pour partie dans le langage courant... ».

Les considérations de productions langagières dans des corpus journalistiques m’ont mené à constituer, des corpus d’articles de journaux ou de revues pour mes expériences.

2) Les corpus réalisés pour les tests et expériences

Divers corpus on été formés pour la réalisation des tests et analyses contenus dans ce mémoire. En voici la liste :

- 35 articles sur le sport

  Journal : l’Équipe

- 50 articles sur la décoration et le bricolage

  Journaux : Prima, Madame le Figaro, Maxi-Mag, Marie France, Femme Actuelle, Avantage et Hello Coton.

- 20 articles sur la Météo

  Journal : Météo France

- 21 articles sur la géopolitique :

  Revue : Conflits

- 12 articles sur l’Education :

  Revue : Sciences Humaines

- 10 articles sur les Jeux vidéos :

  Journal : Le Temps

- 60 Articles sur l’Économie

  Journaux : La Croix, L’Express, La Tribune, Le Figaro

D’autres corpus on été créés pour des tests visant à améliorer le contenu des listes de vocabulaires thématiques mais n’ont pas été utilisés pour les réflexions exposées dans ce mémoire.

B) Amélioration de la hiérarchie thématique

La détection de contenu d’articles de presse doit leur permettre d’être rangés dans des catégories diverses reflétant, sous une appellation générique, leur contenu.

Deux méthodes, détaillées dans les sous-parties suivantes, ont été utilisées dans le cadre du projet Neonaute pour la construction d’une hiérarchie thématique.

Le schéma de cette hiérarchie thématique est consultable en annexe 2.

En somme, il y a plusieurs thèmes dits génériques (Sciences, Politique, Economie, Histoire, Culture, Philsophie, Defense...) contenant chacun un nombre plus où moins important de thèmes dits spécifiques (Botanique, ScienceTerre, EtatGouvernement, ConflistSociaux, Nucleaire, Metallurgique, AmeriqueNord, AmeriqueNordCentrale, Justice, CrimesTraffics, Musique, Danse, Medecine, Cancer, Mythologie, Religion...) dont les dénominations tentent d’être les plus explicites possibles.

1) Recherche dans des catégories existantes

Cette méthode peut être dite par déduction, c’est-à-dire, partir du général pour aller au particulier. En effet, elle vise tout simplement à observer les structures thématiques déjà établies pour s’en inspirer et créer sa propre thématique, adaptée à ses propres besoins.

Par exemple, les articles de Wikipédia sont structurés sous la forme de portails. Chaque portail générique contient une multitude de sous-thèmes plus spécifiques. Madame Bruneau s’est inspirée de cette structure pour développer son système de hiérarchie thématique. Une hiérarchie à deux niveaux : thème générique – thème spécifique. Le thème générique étant théoriquement général, dans lequel se trouvent plusieurs thèmes spécifiques s’y rattachant.

Pour modifier et agrandir la hiérarchie déjà existante, il a fallu, en particulier, s’inspirer des rubriques de journaux tel que Le Monde, La Tribune, Le Point… mais également de presses plus spécialisées comme les journaux féminins pour certains thèmes comme la santé ou les loisirs.

Le choix relève parfois de la subjectivité, car tous les journaux n’ont pas la même façon d’organiser leur contenu informationnel et, dans notre cas, il doit servir dans une approche de détection de la néologie. L’objectif est en quelque sorte d’essayer de réaliser la thématique la plus englobante de la totalité des journaux tout en ajoutant des thèmes plus précis pouvant refléter des contextes langagiers propices à l’apparition de néologismes.

Comme l’explique [C. Gérard, A. Grezka, L. Mercè, 2017], en dehors des rubriques de journaux, d’autres arborescences thématiques avaient été étudiées :

- Eurovoc (UE) qui se limite à l’Union européenne.
- FranceTerme (experts) qui est circonscrit au secteur d’innovation et de découverte,

- Dictionnaires et encyclopédies, hélas trop spécialisés,
- RAMEAU (BNF, BU , …) dont la catégorie est inapplicable à la néologie car trop détaillée.

Malheureusement, les thématiques marquant le débat public sont absentes de ces grandes classifications, or beaucoup de néologismes peuvent apparaître plus spécialement dans ce cadre là. Notre recherche amena donc à créer de nouvelles thématiques, ciblant plus précisément certains contenus du débat public. Les thèmes spécifiques de « Immigration » ou « SoinsAlternatifs » en sont deux exemples.

Le problème de ce type de classification, basée sur l'observation d'arborescences thématiques, réside dans le fait qu'il est impossible de contenter l'ensemble des structures de rubriques journalistiques. On est alors porté à souvent réduire au plus petit dénominateur commun, à savoir les rubriques les plus communément admises telles que la Politique et l'Économie, les thèmes, ce qui ne peut suffire.

Aussi, il est parfois nécessaire de subdiviser un thème. Par exemple, nous avons détaché du thème générique « Economie », le thème spécifique « Industrie » pour en faire un thème générique à lui tout seul contenant les thèmes spécifiques « Agroalimentaire », « Aéronautique », « Nucléaire » ou « Automobile »…

Étant donné que notre analyseur fonctionne en comparant le texte d'articles à des listes de vocabulaires, nous avons réfléchi à la fusion de thèmes qui pourraient contenir les mêmes lexies ce qui générerait de l'ambiguïté sur le choix des thèmes du texte. Ainsi, nous avons, par exemple fusionné les thèmes « Éducation » et « Pédagogie » en « FormationPedagogie » et aussi Écologisme et Environnement en « EcologismeEnvironnementalisme ».

Il est recommandé de choisir des thèmes relativement disparates pour minimiser les chevauchements de vocabulaire. Dans l'optique de pouvoir comparer les vocabulaires de deux listes aux thèmes jugés trop proches, un script python a été réalisé, consultable en annexe 3.

À cela s'ajoute le problème du choix du thème générique car parfois, les thèmes spécifiques à fusionner n'appartiennent pas au même thème générique. Par exemple, les deux thèmes, « MouvementSocial » et « ConflitsSociaux » appartenaient, pour le premier au thème générique « Politique » et, pour le second, à celui intitulé « Société ». Le thème spécifique lié au conflit et au mouvement social devait-il être dans le thème générique politique ou bien société ? Nous avons choisi « Politique » sous l'appellation « ConflitsSociaux ».

Une autre solution permet de développer la réalisation d'une structure thématique : le topic modeling.

2) Usage du topic modeling

La seconde méthode part du particulier pour arriver au générale. Elle est donc inductive, en ne cherchant pas l'inspiration depuis une classification préétablie qui a ses propres rubriques et manières de scinder la réception de l'information mais au contraire, en trouvant les thèmes à partir d'un apprentissage non supervisé. L'expérience part des données d'un corpus que l'on traite selon

diverses méthodes statistiques choisies qui, au travers d'un algorithme, génèrent un nombre de clusters définis par l'utilisateur et auxquels il s'agit de trouver les définitions englobantes.

L'observation des différents clusters ou regroupements de mots formés permet aussi de découvrir des thèmes insoupçonnés ou d'enrichir la hiérarchie thématique déjà existante.

Ce travail de topic modeling a d'abord été effectué il y a quelques années à l'aide du logiciel Millet qui rapporte des paquets de mots aux contextes similaires. Le test a permis de constituer le contenu de plusieurs listes thématiques servant au moteur développé par Lauren Bruneau.

L'inconvénient de cette méthode est qu'elle regroupe des mots mécaniquement sans prendre en compte leur pertinence ou non. C'est pourquoi, de nombreuses listes se sont trouvées surchargées de termes ne ciblant pas le thème spécifique. Le topic modeling produit donc du bruit engendrant un déséquilibre lors de l'analyse. L'analyseur thématique calcule sur la fréquence des mots d'un texte et si un thème comporte trop de mots courants, le texte aura beaucoup plus de chance d'être matché pour ce thème là. Nous y reviendrons plus tard.

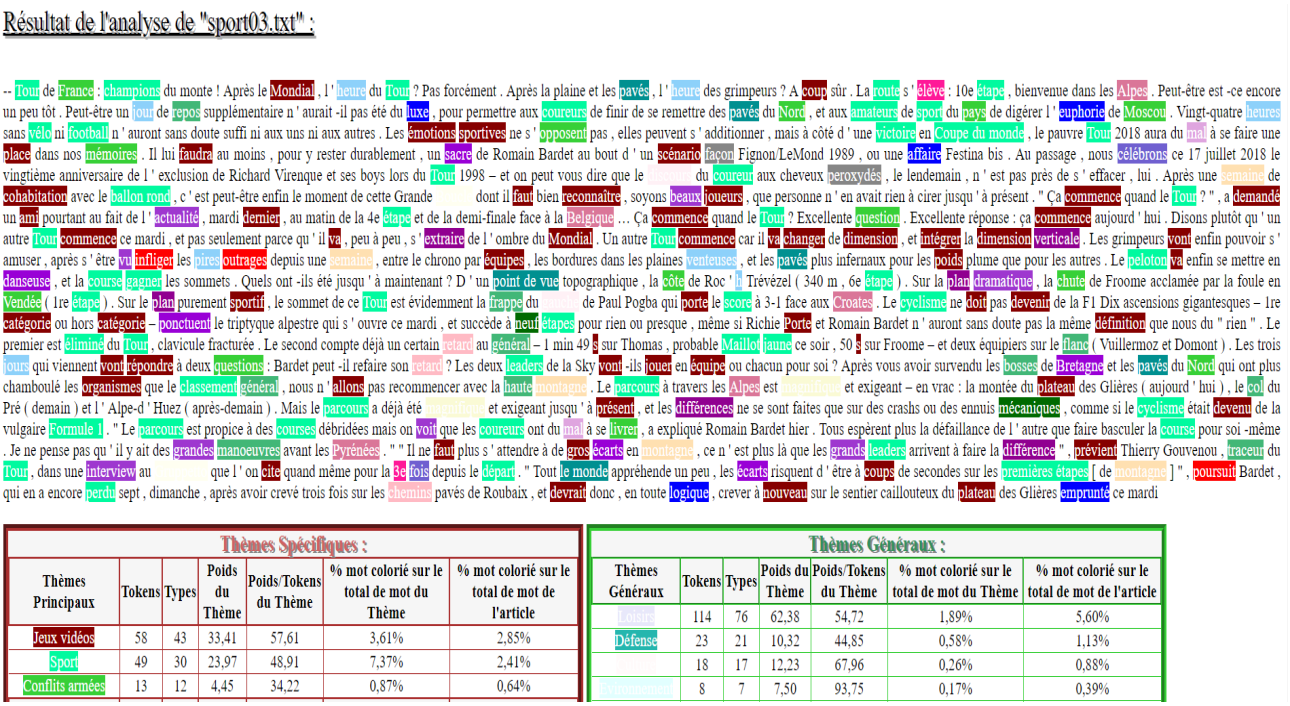
De plus, si la méthode ne sert qu'à découvrir des thèmes nouveaux, choisir un terme reflétant le contenu du cluster n'est pas chose aisée. D'autant que, bien souvent, les clusters regroupent des mots difficiles à concilier thématiquement. Il faut apprendre à donner du sens aux clusters. En troisième partie, nous verrons un test sur un corpus sportif au sujet de ce type de modélisation thématique.

Tout comme dans une recherche de catégories préétablies, il fallait, pour le choix des thèmes, faire attention au mélange des lexiques en prenant bien en compte que le vocabulaire d'un thème peut être similaire à un autre. Autrement dit, il est aussi nécessaire de parfois fusionner des clusters dont les vocabulaires, trop proches, auraient générés des ambiguïtés pour la détection.

La présence intempestive de lexies dites « hors sujet » entraîne une surcharge de contenu pouvant nuire à l'équilibre de calculs lors de la détection. L'exemple suivant (figure 2) évoque les effets d'une surcharge de liste, créée par du topic modeling. Nous pouvons y voir clairement que le vocabulaire contenu dans la liste portant sur le thème des jeux vidéos est trop éloigné du thème. Les mots tels que « commencer », « devenir », « devoir », « demander », « falloir »… n'ont pas à être attribués exclusivement à ce thème ni à aucun d'ailleurs, ils ont un sens trop large.

Néanmoins, il serait possible de justifier la présence d'un vocabulaire assez large au motif que certains types de vocabulaires de termes équivoques sont employés dans des contextes similaires.
L'inconvénient de cette approche est que le vocabulaire accentuera la coloration d'un thème précis, en général celui déterminé par les mots se regroupant dans un même thème. Rappelons que notre analyseur se base sur une sorte de taux de fréquence des mots. Plus la fréquence augmente, plus le poids du thème concerné augmente. Le principe est correct, mais si le vocabulaire de la liste est constitué de mots trop généraux, il y a un risque d'absorber les autres thèmes potentiels.

Nous pouvons voir (figure 3) que la suppression des termes équivoques de la liste jeux vidéos profite à un rééquilibrage du résultat de la détection.



Résultat de l'analyse de "sport03.txt" :

-- Tour de France : champions du monte ! Après le Mondial , l ' heure du Tour ? Pas forcément . Après la plaine et les pavés , l ' heure des grimpeurs ? A coup sûr . La route s ' élève : 10e étape , bienvenue dans les Alpes . Peut-être est -ce encore un peu tôt . Peut-être un jour de repos supplémentaire n ' aurait -il pas été du luxe , pour permettre aux coureurs de finir de se remettre des pavés du Nord , et aux amateurs de sport du pays de digérer l ' euphorie de Moscou . Vingt-quatre heures sans vélo ni football n ' auront sans doute suffi ni aux uns ni aux autres . Les émotions sportives ne s ' opposent pas , elles peuvent s ' additionner , mais à côté d ' une victoire en Coupe du monde , le pauvre Tour 2018 aura du mal à se faire une place dans nos mémoires . Il lui faudra au moins , pour y rester durablement , un sacre de Romain Bardet au bout d ' un scénario façon Fignon/LeMond 1989 , ou une affaire Festina bis . Au passage , nous célébrons ce 17 juillet 2018 le vingtième anniversaire de l ' exclusion de Richard Virenque et ses boys lors du Tour 1998 – et on peut vous dire que le du coureur aux cheveux peroxydés , le lendemain , n ' est pas près de s ' effacer , lui . Après une semaine de cohabitation avec le ballon rond , c ' est peut-être enfin le moment de cette Grande dont il faut bien reconnaître , soyons beaux joueurs , que personne n ' en avait rien à cirer jusqu ' à présent . " Ça commence quand le Tour ? " , a demandé un ami pourtant au fait de l ' actualité , mardi dernier , au matin de la 4e étape et de la demi-finale face à la Belgique … Ça commence quand le Tour ? Excellente question . Excellente réponse : ça commence aujourd ' hui . Disons plutôt qu ' un autre Tour commence ce mardi , et pas seulement parce qu ' il va , peu à peu , s ' extraire de l ' ombre du Mondial . Un autre Tour commence car il va changer de dimension , et intégrer la dimension verticale . Les grimpeurs vont enfin pouvoir s ' amuser , après s ' être vu infliger les pires outrages depuis une semaine , entre le chrono par équipes , les bordures dans les plaines venteuses , et les pavés plus infernaux pour les poids plume que pour les autres . Le peloton va enfin se mettre en danseuse , et la course gagner les sommets . Quels ont -ils été jusqu ' à maintenant ? D ' un point de vue topographique , la côte de Roc ' h Trévézel (340 m , 6e étape) . Sur la plan dramatique , la chute de Froome acclamée par la foule en Vendée (1re étape) . Sur le plan purement sportif , le sommet de ce Tour est évidemment la frappe du de Paul Pogba qui porte le score à 3-1 face aux Croates . Le cyclisme ne doit pas devenir de la F1 Dix ascensions gigantesques – 1re catégorie ou hors catégorie – ponctuent le triptyque alpestre qui s ' ouvre ce mardi , et succède à neuf étapes pour rien ou presque , même si Richie Porte et Romain Bardet n ' auront sans doute pas la même définition que nous du " rien " . Le premier est éliminé du Tour , clavicule fracturée . Le second compte déjà un certain retard au général – 1 min 49 s sur Thomas , probable Maillot jaune ce soir , 50 s sur Froome – et deux équipiers sur le flanc (Vuillermoz et Domont) . Les trois jours qui viennent vont répondre à deux questions : Bardet peut -il refaire son retard ? Les deux leaders de la Sky vont -ils jouer en équipe ou chacun pour soi ? Après vous avoir survendu les bosses de Bretagne et les pavés du Nord qui ont plus chamboulé les organismes que le classement général , nous n ' allons pas recommencer avec la haute montagne . Le parcours à travers les Alpes est et exigeant – en vrac : la montée du plateau des Glières (aujourd ' hui) , le col du Pré (demain) et l ' Alpe-d ' Huez (après-demain) . Mais le parcours a déjà été et exigeant jusqu ' à présent , et les différences ne se sont faites que sur des crashs ou des ennuis mécaniques , comme si le cyclisme était devenu de la vulgaire Formule 1 . " Le parcours est propice à des courses débridées mais on voit que les coureurs ont du mal à se livrer , a expliqué Romain Bardet hier . Tous espèrent plus la défaillance de l ' autre que faire basculer la course pour soi -même . Je ne pense pas qu ' il y ait des grandes manoeuvres avant les Pyrénées . " " Il ne faut plus s ' attendre à de gros écarts en montagne , ce n ' est plus là que les grands leaders arrivent à faire la différence " , prévient Thierry Gouvenou , traceur du Tour , dans une interview au Gruppetto que l ' on cite quand même pour la 3e fois depuis le départ . " Tout le monde appréhende un peu , les écarts risquent d ' être à coups de secondes sur les premières étapes [de montagne] " , poursuit Bardet , qui en a encore perdu sept , dimanche , après avoir crevé trois fois sur les chemins pavés de Roubaix , et devrait donc , en toute logique , crever à nouveau sur le sentier caillouteux du plateau des Glières emprunté ce mardi

Thèmes Spécifiques :

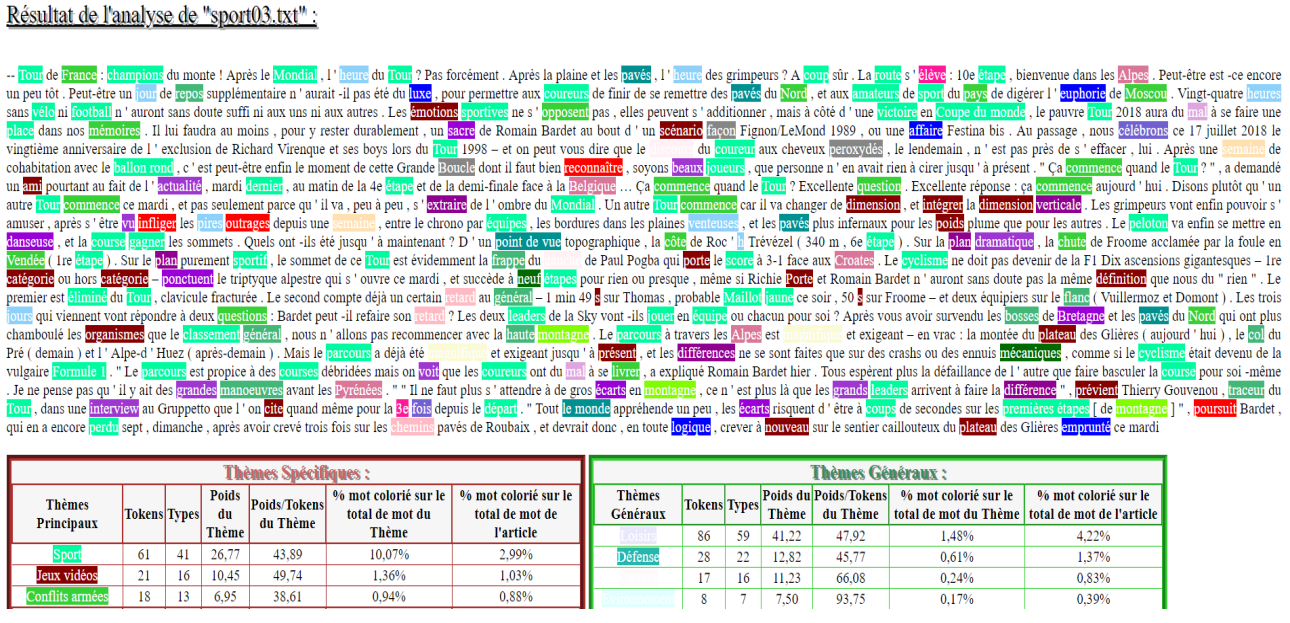
Thèmes Principaux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Jeux vidéos	58	43	33,41	57,61	3,61%	2,85%
Sport	49	30	23,97	48,91	7,37%	2,41%
Conflits armées	13	12	4,45	34,22	0,87%	0,64%

Thèmes Généraux :

Thèmes Généraux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Sport	114	76	62,38	54,72	1,89%	5,60%
Défense	23	21	10,32	44,85	0,58%	1,13%	
	18	17	12,23	67,96	0,26%	0,88%
Environnement	8	7	7,50	93,75	0,17%	0,39%

Figure 2 : Résultat d’une analyse thématique d’un article sportif

Le thème principal détecté est celui des Jeux vidéos. Après avoir supprimé les mots généraux dans la liste de ce thème, nous obtenons le résultat ci-dessous (figure 3) :



Résultat de l'analyse de "sport03.txt" :

-- Tour de France : champions du monte ! Après le Mondial , l ' heure du Tour ? Pas forcément . Après la plaine et les pavés , l ' heure des grimpeurs ? A coup sûr . La route s ' élève : 10e étape , bienvenue dans les Alpes . Peut-être est -ce encore un peu tôt . Peut-être un jour de repos supplémentaire n ' aurait -il pas été du luxe , pour permettre aux coureurs de finir de se remettre des pavés du Nord , et aux amateurs de sport du pays de digérer l ' euphorie de Moscou . Vingt-quatre heures sans vélo ni football n ' auront sans doute suffi ni aux uns ni aux autres . Les émotions sportives ne s ' opposent pas , elles peuvent s ' additionner , mais à côté d ' une victoire en Coupe du monde , le pauvre Tour 2018 aura du mal à se faire une place dans nos mémoires . Il lui faudra au moins , pour y rester durablement , un sacre de Romain Bardet au bout d ' un scénario façon Fignon/LeMond 1989 , ou une affaire Festina bis . Au passage , nous célébrons ce 17 juillet 2018 le vingtième anniversaire de l ' exclusion de Richard Virenque et ses boys lors du Tour 1998 – et on peut vous dire que le du coureur aux cheveux peroxydés , le lendemain , n ' est pas près de s ' effacer , lui . Après une semaine de cohabitation avec le ballon rond , c ' est peut-être enfin le moment de cette Grande Boucle dont il faut bien reconnaître , soyons beaux joueurs , que personne n ' en avait rien à cirer jusqu ' à présent . " Ça commence quand le Tour ? " , a demandé un ami pourtant au fait de l ' actualité , mardi dernier , au matin de la 4e étape et de la demi-finale face à la Belgique … Ça commence quand le Tour ? Excellente question . Excellente réponse : ça commence aujourd ' hui . Disons plutôt qu ' un autre Tour commence ce mardi , et pas seulement parce qu ' il va , peu à peu , s ' extraire de l ' ombre du Mondial . Un autre Tour commence car il va changer de dimension , et intégrer la dimension verticale . Les grimpeurs vont enfin pouvoir s ' amuser , après s ' être vu infliger les pires outrages depuis une semaine , entre le chrono par équipes , les bordures dans les plaines venteuses , et les pavés plus infernaux pour les poids plume que pour les autres . Le peloton va enfin se mettre en danseuse , et la course gagner les sommets . Quels ont -ils été jusqu ' à maintenant ? D ' un point de vue topographique , la côte de Roc ' h Trévézel (340 m , 6e étape) . Sur la plan dramatique , la chute de Froome acclamée par la foule en Vendée (1re étape) . Sur le plan purement sportif , le sommet de ce Tour est évidemment la frappe du de Paul Pogba qui porte le score à 3-1 face aux Croates . Le cyclisme ne doit pas devenir de la F1 Dix ascensions gigantesques – 1re catégorie ou hors catégorie – ponctuent le triptyque alpestre qui s ' ouvre ce mardi , et succède à neuf étapes pour rien ou presque , même si Richie Porte et Romain Bardet n ' auront sans doute pas la même définition que nous du " rien " . Le premier est éliminé du Tour , clavicule fracturée . Le second compte déjà un certain retard au général – 1 min 49 s sur Thomas , probable Maillot jaune ce soir , 50 s sur Froome – et deux équipiers sur le flanc (Vuillermoz et Domont) . Les trois jours qui viennent vont répondre à deux questions : Bardet peut -il refaire son retard ? Les deux leaders de la Sky vont -ils jouer en équipe ou chacun pour soi ? Après vous avoir survendu les bosses de Bretagne et les pavés du Nord qui ont plus chamboulé les organismes que le classement général , nous n ' allons pas recommencer avec la haute montagne . Le parcours à travers les Alpes est et exigeant – en vrac : la montée du plateau des Glières (aujourd ' hui) , le col du Pré (demain) et l ' Alpe-d ' Huez (après-demain) . Mais le parcours a déjà été et exigeant jusqu ' à présent , et les différences ne se sont faites que sur des crashs ou des ennuis mécaniques , comme si le cyclisme était devenu de la vulgaire Formule 1 . " Le parcours est propice à des courses débridées mais on voit que les coureurs ont du mal à se livrer , a expliqué Romain Bardet hier . Tous espèrent plus la défaillance de l ' autre que faire basculer la course pour soi -même . Je ne pense pas qu ' il y ait des grandes manoeuvres avant les Pyrénées . " " Il ne faut plus s ' attendre à de gros écarts en montagne , ce n ' est plus là que les grands leaders arrivent à faire la différence " , prévient Thierry Gouvenou , traceur du Tour , dans une interview au Gruppetto que l ' on cite quand même pour la 3e fois depuis le départ . " Tout le monde appréhende un peu , les écarts risquent d ' être à coups de secondes sur les premières étapes [de montagne] " , poursuit Bardet , qui en a encore perdu sept , dimanche , après avoir crevé trois fois sur les chemins pavés de Roubaix , et devrait donc , en toute logique , crever à nouveau sur le sentier caillouteux du plateau des Glières emprunté ce mardi

Thèmes Spécifiques :

Thèmes Principaux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Sport	61	41	26,77	43,89	10,07%	2,99%
Jeux vidéos	21	16	10,45	49,74	1,36%	1,03%
Conflits armées	18	13	6,95	38,61	0,94%	0,88%

Thèmes Généraux :

Thèmes Généraux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Sport	86	59	41,22	47,92	1,48%	4,22%
Défense	28	22	12,82	45,77	0,61%	1,37%
	17	16	11,23	66,08	0,24%	0,83%
Environnement	8	7	7,50	93,75	0,17%	0,39%

Figure 3 : Résultat d’une analyse thématique d’un article sportif

On observe que la liste de vocabulaire sur le jeux vidéos, générée à l’aide du topic modeling, entraîne une erreur de détection.

D’autres tests plus conséquents seront présentés et commentés dans la troisième partie.

Un autre écueil soulevé à propos de la constitution de liste automatiquement est le fait que si l'on prend toute une masse d'articles d'un thème donné, on risque de se focaliser sur une période faisant ressortir un évènement en particulier. Ainsi, la liste de termes comportera des mots de cet évènement qui n'est pas foncièrement lié au thème.

Ex : « Paris » car attentat à « Paris » durant les épisodes de novembre 2013 alors que Paris n'est pas foncièrement lié a terrorisme.

C) Le contenu des listes thématiques

L'analyseur thématique développé par Lauren Bruneau fonctionne par comparaison entre l'article de presse et des listes de vocabulaires thématiques. Ainsi, le fond de la réflexion réside dans le choix des mots dans les listes. Les termes rares mais très spécifiques à leur thème ne seront pas nécessairement présents dans des articles de presse, sauf ceux spécialisés, censés s'adapter au lecteur néophyte. Les termes plus courant sont simples mais peuvent être équivoques. À côté de cela, le matching requiert un étiquetage précis des lexies.

1) Le problème de l'étiquetage

Les articles sont parsés et étiquetés par l'outil Treetagger, selon les choix de Lauren Bruneau. Il fallait donc étiqueter le contenu des listes selon les étiquettes de ce programme pour qu'il puisse correctement les reconnaître. Malheureusement, ceci pouvait être source de confusion. Prenons l'exemple du ngrams « haut de gamme » en comparant les deux étiquetages :

- Tree-tagger : NOM PREP NOM (soit haut_nc de_prp gamme_nc dans le fichier)

- Annotateur : ADJ PRP NOM (soit haut_adj de_prp gamme_nc dans le fichier)

Par conséquent, l'erreur d'étiquette du côté de l'annotateur avait empêché le matching de la portion.

Le problème est particulièrement présent avec les noms propres car, dans le script développé par Lauren Bruneau, les majuscules sont mises en minuscules et l'étiquette NAM de Treetagger servant à catégoriser ce type de mots, est fusionnée avec l'étiquette NOM, le tout regroupé sous l'appellation « nc » : nom commun.

L'erreur de l'annotateur engendre aussi une segmentation défectueuse. Si nous prenons l'exemple du verbe « s'établir » qui donne dans la liste s_pro &quote_pun établir_v, nous remarquons (photo ci-dessous) que pronom du verbe n'a pas été repéré. L'annotateur de nos listes qui étiquette ainsi : s_adj $quote_pun établir_v ne matche pas la portion.

Figure 4 : erreur d'étiquetage du verbe pronominal « s'établir »

Normalement le « s' » devient « se » et treetagger donne le résultat d'étiquetage suivant : se_pro établirent_v, qui produit (figure 5) le résultat suivant.

-- Les soldats s ' établirent sur la colline .

Figure 5 : bonne étiquetage du verbe pronominal « s'établir »

La segmentation avait dissocié l'apostrophe de la lettre « s ». Par conséquent, transformer l'apostrophe en « &quote » n'est pas forcément recommandée.

Par conséquent, la personne qui met les tags doit être vigilante et parfois même vérifier elle-même l'étiquetage de Treetagger qui, bien qu'étant un outil très performant, peut faire des erreurs de tagging. Plus loin, les exemples de « Afrique » et « Centrafrique » l'illustreront.

Mais regardons tout d'abord le problème de lemmatisation qu'il peut y avoir lors du traitement de Treetagger. Par exemple, selon l'outil, le mot « taux » a pour lemme « tau ». Il faut alors mettre dans la liste :
tau_nc
tau_nc d_adj &quote_pun intérêt_nc
tau_nc de_prp change_nc
tau_nc de_prp croissance_nc
...
Une solution très gênante qui impose d'écrire volontairement des termes mal orthographiés.

Ces problèmes et oublis d'étiquetage empêchent le repérage correcte et fausse la détection car des mots clés ne sont pas colorisés.

Prenons des exemples plus concrets et illustrés par notre analyseur thématique avec les termes « Afrique » et « Centrafrique ».

En faisant le test sur deux petites phrases à partir d'une seule liste (Theme_Continent-Afrique.txt), voici le résultat (figure 6) que nous obtenons :

Résultat de l'analyse de "test_article_rjdh.txt" :

-- L ' Afrique est un immense continent . La Centrafrique est au centre de l ' Afrique .

Thèmes Spécifiques :

Thèmes Principaux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article

Thèmes Généraux :

Thèmes Généraux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article

Figure 6 : Résultat de la détection thématique des mots Afrique et Centrafrique

Nous pouvons remarquer qu'aucun mot n'a été colorié et donc que les tableaux sont vides. Or, le fichier Theme_Continent-Afrique existe bien. Si l'on regarde à l'intérieur du fichier, on s'aperçoit qu'il manque le mot « Afrique »(figure 7) et que le mot « Centrafrique » (figure 8) a été étiqueté comme adjectif alors que c'est un nom.

```
afar_nc
africain_adj
africain_nc
afrikaan_nc
afrique_nc du_prp sud_nc
afrique_nc noir_adj
ahmed_nc abdallah_nc mohamed_nc sambi_nc
```

Figure 7 : « afrique » absent

```
centraficain_adj
centrafricain_adj
centrafricain_nc
Centrafrique_adj # ADJ => erreur
```

Figure 8 : « centrafrique » _adj

Le mot « centrafrique » est bien présent mais non étiqueté comme un nom. Il reçoit le tag réservé aux adjectifs alors que l'adjectif du mot est « centrafricain ».

Après correction (figure 9 et 10), les listes apparaissent ainsi :

```
afar_nc
africain_adj
africain_nc
afrikaan_nc
afrique_nc  # AJOUT
afrique_nc du_prp sud_nc
afrique_nc noir_adj
ahmed_nc abdallah_nc mohamed_nc sambi_nc
```

Figure 9 : « afrique » après correction

```
centraficain_adj
centrafricain_adj
centrafricain_nc
Centrafrique_nc # ADJ -> NC => :)
```

Figure 10 : « centrafrique » après correction

Avec les termes bien étiquetés, nous obtenons (figure 11) un matching correct des termes et le ciblage de bons thèmes. Rappelons à ce sujet que pour l'exemple précédent, la seule liste de vocabulaire du thème spécifique « Afrique » a été retenue lors de l'analyse. Nous verrons par la suite et au travers d'un autre exemple que, analysé sur plusieurs listes, le choix de coloration se fait dans l'ordre de lecture des listes par l'algorithme.

Résultat de l'analyse de "test_article_rjdh.txt" :

-- L ' Afrique est un immense continent . La Centrafrique est au centre de l ' Afrique .

Thèmes Spécifiques :

Thèmes Principaux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Afrique	4	3	4,00	100,00	0,74%	10,53%

Thèmes Généraux :

Thèmes Généraux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Continent	4	3	4,00	100,00	0,74%	10,53%

Figure 11 : Résultat de la détection thématique des mots Afrique et Centrafrique

D'autres exemples produisent des incohérences plus difficiles à résoudre. La nouvelle région « Grand Est » a été créée il y a peu. Regardons la manière dont elle est étiquetée.
Treetagger l'étiquette (figure 12) mal : « grand_adj est_nc ». Il faudrait, dans notre cas, qu'il soit étiqueté : « grand_nc est_nc.

Nous voyons ici que la majuscule, doublée d'une étiquette spéciale pour les noms propres permettrait d'éviter cet amalgame entre l'adjectif « grand » et le nom propre « Grand » et entre le nom « est » et le nom propre « Est ». De plus, étiqueter ainsi « grand_adj est_nc » afin qu'il soit bien repéré, paraît absurde. Néanmoins, c'est ce qu'il faut faire, à défaut de changer l'intérieur du script.

Cet exemple introduit de nouveaux problèmes exposés dans la sous-partie suivante. Mais avant, mentionnons, au travers d'un autre exemple que parfois, un terme sera indétectable correctement. Pour deux sens différents, le verbe « offrir » aura les mêmes compléments de tags dans la syntaxe de sa phrase :

- offrir (sens de montrer). Ex : Le jardin offre ses (dernières) floraisons => S V COD => V NC

- offrir (sens de faire un don). Ex : Il offre (un) cadeau => S V COD => V NC

Voici un dernier exemple concret illustrant les difficultés d'étiquetage. Dans le thème spécifique « BourseMarchéAssurance », il est question du token « AAA » ou « triple A ». Il faut, pour étiqueter cette deuxième expression, bien faire attention car Treetagger étiquette ainsi : « triple_ADJ a_NAM ». Hors, l'humain aurait tendance à mettre « triple_NC A_NAM » puisqu'on dit « Le triple A ».

Il faut donc, pour les expressions, bien vérifier comment a étiqueté Treetagger afin de mettre le bon tag dans la liste.

2) Le problème du repérage des entités nommées et des ngrams

Détecter les entités nommées, et donc de fait les noms propres ainsi que les ngrams, est particulièrement important dans un travail de détection thématique. En effet, il permet de cibler un thème de manière extrêmement précise. Par exemple, si l'on repère le nom d'un joueur de football, dans la mesure où son nom ne souffre d'aucune ambiguïté, il focalisera la détection sur un thème bien particulier.

Effet indésirable, le traitement effectué par le script de détection thématique réalise une suppression des majuscules ainsi que de l'étiquette NAM de Treetagger pour la regrouper avec l'étiquette NOM. Cette réduction peut générer des écueils. Ici (figure 12), l'exemple d'un nom de famille sans sa majuscule provoque une ambiguïté de détection (figure 13). Le nom « Porte » n'a rien à voir avec le mot « porte » du verbe « porter » présent dans la liste thématique « JeuxVidéos ».

même si Richie Porte et Romain Bardet

Jeux vidéos

Figure 12 : « Porte » Nom propre Figure 13 : « porte » thème principal

Dans l'exemple suivant, montrons l'intérêt des noms propres pour la détection de contenu.

Nous voyons dans le diagramme (figure 14) que le nom « Deschamp », faisant référence à l'entraîneur de foot Didier Deschamps, n'apparaît que dans les articles balisés « foot » et non dans ceux balisés « velo ».

Ainsi, sans même regarder une quelconque fréquence, le simple fait que l'entité nommée soit présente dans l'article permet de cibler le bon thème.

De même avec le nom du coureur « Junbels », nous pouvons constater (figure 15) que ce sont les articles sur le cyclisme qui ressortent.

Bien que ces deux noms ne semblent par générer d'ambiguïté, si l'on supprime leur majuscule, leur apparence reflète clairement le clivage entre le sous-corpus foot et le sous-corpus velo.

Le thème spécifique « Sport » sera sans doute diviser par la suite, afin de détecter séparément divers sports importants de l'actualité. C'est alors que la possibilité de cibler les noms propres aura tout son intérêt pour éviter les confusions ; ne pas mélanger ou confondre les noms des sportifs pour ne pas mélanger les sports. Nous reviendrons à cette problématique dans la troisième partie de ce mémoire.

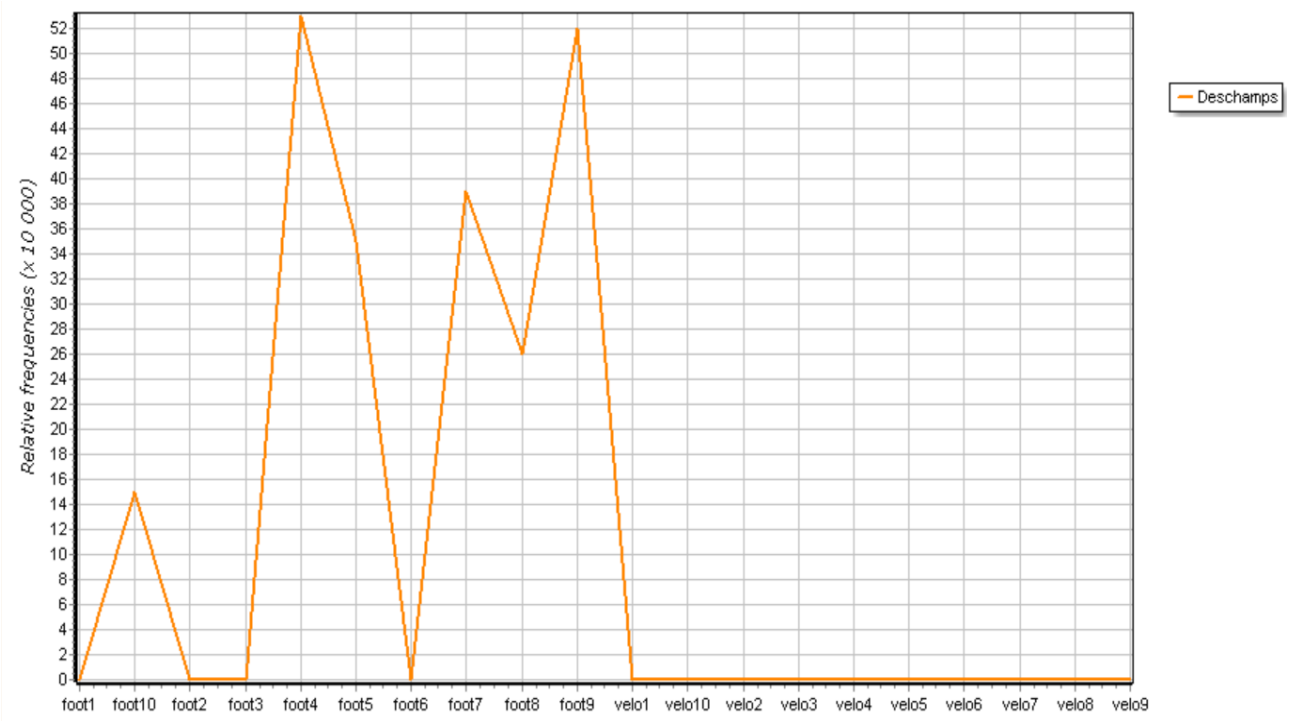


Figure 14 : Diagramme d'apparition du mot « Deschamps »

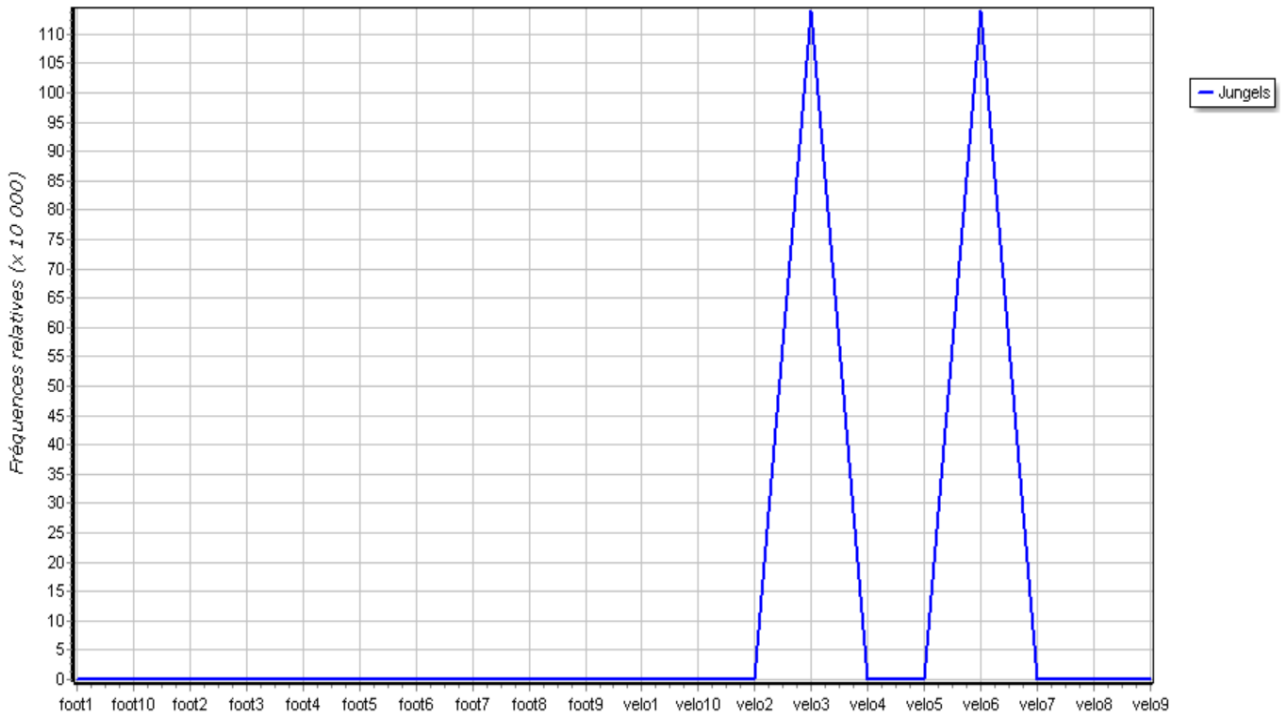


Figure 15 : Diagramme d'apparition du mot « Junbels »

Parallèlement aux noms propres, l'étude des « ngrams », où suites de mots, peut référer à des idées tout aussi importante. Il peut s'agir de mots qui se suivent pour ne former qu'une seule entité comme « laisser passer » ou bien d'entités nommées telle que « Organisation des Nations Unis ».

Les mots contenus dans ces ngrams peuvent être repérés isolément, ce qui pose un problème de détection selon l'ordre d'analyse des listes. Testons cela sur deux phrases simples :

- « Il s'agit d'un apprentissage machine. »
- « Il s'agit d'un apprentissage par renforcement. ».

Les deux expressions : « apprentissage machine » et « apprentissage par renforcement » ne sont que dans la liste du thème spécifique « RobotiqueIA ». Toutefois, les termes seuls (« apprentissage » et « renforcement ») sont présents dans la liste « FormationPedagogie ». Or, lors de l'analyse, c'est cette liste « FormationPedagogie » qui est filtrée et coloriée en premier. Par conséquent, les expressions de la liste « RobotiqueIA » ne sont pas reconnues. Voyons le résultat ci-dessous (figure 16) :

-- Il s ' agit de l ' apprentissage machine . Il s ' agit de l ' apprentissage par renforcement

Thèmes Spécifiques :

Thèmes Principaux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Formation pédagogique	3	2	1,00	33,33	0,53%	7,32%

Thèmes Généraux :

Thèmes Généraux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Education	3	2	1,00	33,33	0,25%	7,32%

Figure 16 : Résultat de la détection des expressions « apprentissage machine » et « apprentissage pas renforcement »

La liste FormationPedagogie a été retirée afin de vérifier si la liste RobotiqueIA allait prendre la relève. Effectivement, comme le montre l'exemple ci-dessous (figure 17), les expressions ont bien été repérées.

-- Il s ' agit de l ' apprentissage machine . Il s ' agit de l ' apprentissage par renforcement .

Thèmes Spécifiques :

Thèmes Principaux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Robotique et IA	2	2	2,00	100,00	2,27%	4,76%

Thèmes Généraux :

Thèmes Généraux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Téchnologie	2	2	2,00	100,00	0,08%	4,76%

Figure 17 : Résultat de la détection des expressions « apprentissage machine » et « apprentissage pas renforcement »

Il faudrait donc que le matching des ngrams l'emporte sur les autres, et ceci de manière croissante selon le schéma et l'exemple suivant :

apprentissage < apprentissage machine < apprentissage machine non supervisé.

Ce procédé permettrait de régler aussi le problème des termes équivoques qui, dès lors qu'ils seraient intégrés dans une suite de mots, verraient leur sens spécifié. Par exemple, le mot « position », inséré dans le bigram « position défensive » et « position familiale » révélerait le clivage et ciblerait immédiatement deux thèmes plus précis.

Dans la troisième partie, notre objectif sera de procéder à de nouveaux essais afin de varier les expériences sur la détection thématique.

III. Expériences, évaluations et résultats

Cette partie a pour objet de présenter certains tests permettant de proposer des améliorations pour la détection thématique d'articles de presse. Nous allons tenter de vérifier la pertinence de la modélisation de sujets par clustering, le travail sur la méthode symbolique par l'amélioration des listes et enfin l'apprentissage automatique qui semble être, à notre époque un moyen pour la compréhension des contenus textuels.

A) Éprouver le topic modeling

1) Clusterisation d'un corpus sur le sport

La réflexion a porté sur la transformation du thème spécifique "Sport" en thème générique pour le subdiviser en thèmes spécifiques "Football", "Rugby" … dans le but de cibler les articles plus spécifiques à tel ou tel dicipline.

Afin de vérifier l'efficacité d'une détection thématique sur plusieurs articles traitant de sports différents, un corpus de 35 articles issus du journal l'Équipe a été formé.

L'objectif était d'effectuer un clustering de ce corpus avec 7 clusters en sortie, correspondant au nombre de sous-corpus, chacun propre à un sport, afin d'observer dans un premier temps si les clusters formés regrouperaient bien des mots relatifs à chaque sport, et également, dans un second temps, voir les types de mots regroupés (noms propres, termes techniques...).

Ceci permettrait de sélectionner les termes les plus pertinents pour le repérage de tel ou tel sport et de tester l'efficacité de la clusterisation.

Les 7 sports choisis sont : football, rugby, voile, cyclisme, formule1, tennis et athlétisme.

Au sujet du choix du nombre de clusters pour une analyse, l' article [L. Derczynski, S. Chester, K.S. Bøgh, 2015] indique qu'à chaque taille de corpus correspond un nombre de classes optimales. Et inversement, à chaque nombre de classes correspond une taille de corpus. Et les auteurs ajoutent qu'un nombre de classes trop faible produit des classes d'inégales qualités puisqu'on aura, à la fois des classes rassemblant énormément de mots et des classes contenant très peu de termes. Et dans les classes rassemblant beaucoup de mots, il y aura une grande diversité de mots à l'intérieur et le contenu de ces classes ne sera pas homogène.

À l'inverse, en présence d'un trop grand nombre de classe (4000 termes → 4000 classes), on obtient de fait un terme par classe. Le clustering est alors inutile. Ainsi, un trop grand nombre de classes engendre des groupes composés de trop peu d'éléments pour être correctement interprétés.

En somme, les auteurs de l'article ont remarqué qu'à chaque taille de corpus, il est possible de trouver le nombre de classes optimales ».

Dans notre cas, nous connaissons le nombre de clusters à produire puisque nous avons créé notre corpus de sport en référençant les articles selon leur discipline. Notre hypothèse de travail est d'obtenir en sortie, des clusters regroupant les mots propres au football, au rugby, au cyclisme, à l'athlétisme, à la Formule 1, à la voile et au tennis. Les thèmes sont proches de la thématique du

sport, mais les vocabulaires et se rattachant à chaque sport, tant au niveau des noms propres (noms des sportifs…) que des noms communs (termes techniques, de terrain…) devraient différer.

La clusterisation a été réalisée à l'aide de l'algorithme de Brown. Ce dernier effectue ses calculs sur des bigrams. Cela pourrait donc présenter un intérêt pour les suites de mots type Prénom - Nom pour les appellations de sportifs.

Au préalable de ce test, un script bash (annexe 4) tiré des commandes de cours de fouilles de textes réalise le nettoyage du corpus (supprimer les mots vides et isoler la ponctuation).

2) Évaluation et résultats : les échecs de la clusterisation

Parmi les clusters (numérotés de 1 à 7) produits par les calculs de l'algorithme de Brown, les termes propres à chaque sport sont assez rares et pour le peu qu'il y a, leur regroupement est très décevant.

En détail et pour exemple, le cluster 1 a regroupé, du thème générique Formule1, les mots « Ferrari », « pilote » et « écurie » mais non « prix », « voiture » et « slicks », laissés respectivement aux clusters 2 et 4.

Les tokens relatifs au domaine de la voile se sont retrouvés dans le premier cluster (« équipage » et « nautique »), le deuxième (« marins et skipper »), le troisième (« sailing »), le quatrième (« bateau ») et le cinquième (« barreur », « rade », et « nacra »).

Concernant quelques entités nommées, ces dernières ont également été mal réparties. Les noms de « Mannshaft », « Virsliga » et « Ventspils » du cluster 1 propre au football n'ont pas été regroupés avec « Benfica », « Chelsea », « Gazélec » et « Villarreal », placés dans le cluster 5.

Remarquons également la répartition sur trois clusters de termes liés à la Formule1 avec « Heusden-Zolder » dans le premier cluster, « Silverstone » dans le cinquième et « Sauber » dans le sixième.

Les entités nommées et les termes techniques propres au même sport ne sont pas forcément regroupées. Les termes « Euskadi » du cyclisme dans le cluster 4 et « maillot » (pour « maillot jaune ») dans le deuxième cluster l'illustrent bien.

Précisons que les contextes des mots pouvant se chevaucher avec plusieurs sports ont été vérifiés.

À présent, pour compléter la vérification du test, focalisons notre attention sur les noms des sportifs.

Les listes ci-dessous, divisées en fonction des sports, comportent pour la plupart des dénominations de joueurs avec la distinction prénom et nom suivi du numéro du cluster entre parenthèses.

Tennis :
- andy (1) murray (6) : tennis
- kristina (2) mladenovic (1) : tennis
- steffi (3) graf (6) : tennis
- julien (4) benneteau (5) : tennis
- angelique (5) kerber (4) : tennis
- novak (5) djokovic (3) : tennis
- yannick (5) noah (6) : tennis

- caroline (6) garcia (7) : tennis
- michael (7) llodra (5) : tennis

Cyclisme :
- primoz (5) rodlic (4) : cyclisme
- patrick (2) lefevere (3) : cyclisme
- tanel (5) kangert (3) : cyclisme
- nairo (3) quintana (1) : cyclisme
- christopher (2) froome (3) : cyclisme
- bob (1) jungels (5) : cyclisme
- geraint (5) thomas (3) : cyclisme
- rafal (3) majka (5) : cyclisme
- alberto (4) contador (5) : cyclisme
- egan (5) bernal (6) : cyclisme
- guillaume (5) van (3) keirsbulck (4) : cyclisme
- tom (3) boonen (4) : cyclisme

Athlétisme :
- thiago (3) braz (0) : athlétisme
- timothy (3) cheruiyot (2) : athlétisme
- kevin (4) mayer (1) : athlétisme
- greg (5) rutherford (3) : athlétisme
- luvo (5) manyonga (3) : athlétisme
- emmanuel (5) korir (3) : athlétisme
- brianna (5) ncneal (3) : athlétisme
- renaud (5) lavillenie (3) : athlétisme
‘ serena (5) williams (3) : athlétisme
- dafne (4) schippers (5) : athlétisme
- garfield (5) darien (4) : athlétisme
- pierre-ambroise (5) bosse (3) : athlétisme
- frankie (1) fredericks (6) : athlétisme
- elena (6) vallortigara (7) : athlétisme
- sergey (6) shubenkov (7) : athlétisme
- shelly-ann (4) fraser-pryce (5) : athlétisme
- orphée (5) neola (4) : athlétisme
- ruth (5) jebet (4) : athlétisme
- maria (4) lasitskene (5) : athlétisme
- orlann (5) ombissa-dzangue (4): athlétisme

Formule 1 :
- lewis (1) hamilton (2) : formule 1
- valterri (3) bottas (1) : formule 1
- daniel (5) ricciardo (1) : formule 1
- carlos (4) sainz (5) : formule 1
- kimi (4) räikkönen (5) : formule 1
- romain (1) grosjean (3): formule 1
- toto (2) wolff (4) : formule 1

Rugby
- rieko (7) ioane (4) : rugby
- george (5) moala (3) : rugby
- simon (4) zebo (5) : rugby
- paul (4) couet-lannes (6) : rugby
- paddy (6) jackson (3) : rugby
- jerome (6) kaino (7) : rugby
- johan (5) goosen (4) : rugby

Voile :
- quentin (4) delapierre (5) : voile
- michel (7) desjoyeaux (5) : voile
- franck (4) cammas (5) : voile
- carolijn (6) brouwer (7) : voile
- mathieu (4) richard (5) : voile

Football :
- uli (3) hoeness (1) : football
- didier (3) deschamps (4) : football
- sabri (4) lamouchi (5) : football
- n'golo (5) kanté (6) : football
- alejandro (6) valverde (7) : football
- cristiano (2) ronaldo (6) : football

Les listes indiquent que la totalité des prénoms et noms des sportifs a été répartie séparément entre les clusters. Par exemple, « Cristiano », placé dans le cluster n°2 voit son nom de famille « ronaldo » dans le cluster n°6. Cet échec de calcul de bigrams peut être dû au manque de fréquence des termes. En effet, les noms des sportifs sont fréquents mais pas forcément récurrents et parfois, on a le nom de famille seul. Par conséquent, l'algorithme ne peut apprendre le lien existant entre le nom et le prénom. L'algorithme ne peut donc pas s'entraîner et comprendre l'association qu'il y a entre le prénom et le nom.

Il est même impossible de soupçonner que certains prénoms ou noms de sportifs d'une même discipline puissent avoir été rassemblés dans un seul cluster car les caprices de l'algorithme les ont vraiment totalement éparpillés.

Néanmoins, quatre relations se démarquent totalement :

- 12 avec prénom (5) nom (3)

- 12 avec prénom (4) nom (5)

- 8 avec prénom (5) nom (4)

- 6 avec prénom (6) nom (7)

Parmi ces relations, la majorité appartient aux adeptes de l'athlétisme avec 20/38 sportifs concernés et 15 des 20 athlètes ont soit leur prénom, soit leur nom rangés dans le cluster 5. Parmi eux, si l'on observe les fréquences, il y a deux noms apparaissant cinq fois et un nom quatre fois mais le reste est réduit à l'état d'hapax. Bien que plus élevées, ces fréquences ne permettent pas de comprendre cette classification.

B) L’enrichissement des listes

Basé sur le script Perl élaboré par Lauren Bruneau, un de mes objectifs de stage était de poursuivre cette méthode symbolique par la complétion des listes de vocabulaire des thèmes.

1) Entre surcharge et épuration lexicale

La deuxième partie de ce mémoire a montré l’inconvénient d’une liste fortement bruitée. Après divers tests de détection sur trois corpus concernant la météo, la géopolitique et l’éducation, la détection obtenait comme résultats les thèmes « JeuxVidéos » et, dans une moindre mesure « BourseMarcheAssurance » ; deux thèmes inappropriés en raison de leurs listes de vocabulaire surchargées et entrant de fait en concurrence avec des thèmes plus ciblés.

Revoyons cela, au travers de l’analyse de deux articles de presse, pris dans le domaine sportif, relatifs à la victoire de la France à la coupe du monde de football 2018. Nous voyons (figure 18) que le thème sport a bien été identifié en premier ; alors que dans l’autre (figure 19), c’est le thème ConflitsArmee qui a été retenu pour la première place. En effet, on peut voir que les mots « victoire », « libération », « drapeaux » ont porté à confusion. Ces termes pourraient aussi bien être présents dans un thème sur la guerre que sur une discipline sportive.

Remarquons que le thème « JeuxVidéos » est commun aux deux alors qu’il ne s’agit pas de jeux-vidéos. Et rappelons que la liste « JeuxVidéos » a été partiellement épurée dans la partie précédente. Ceci témoigne encore d’une surcharge de cette liste. Nous le constatons clairement par son vocabulaire ; les termes « compose », « choisi » ou « porté » ne devraient pas être matchés, surtout pas au profit de la liste sur les jeux vidéos. En revanche, ces termes pourraient servir à être intégrés dans des ngrams comme par exemple : « composer une musique », « porter le drapeau »…

Résultat de l'analyse de "sport01.txt" :

-- L ' équipe de France a remporté la deuxième Coupe du monde de son histoire . On dresse le bilan de la compétition de chacun des 23 joueurs qui la compose . Ils ont décroché leur étoile . Un mois et demi après l ' annonce de la liste des 23 par Didier Deschamps , les joueurs qu ' il a choisi pour composer l ' équipe de France ont remporté , dimanche 15 juillet , la deuxième Coupe du monde de l ' histoire des Bleus . Portés par Paul Pogba , Antoine Griezmann , Hugo Lloris et le benjamin de la bande , Kylian Mbappé , les Français ont battu les Croates au terme d ' une finale échevelée (4-2) et rejoignent , vingt ans après , l ' équipe de leur entraîneur , sacrée en 1998 . Portée par une ossature solide , l ' équipe de France a réalisé un Mondial de haut vol . Mais quel bilan pour chacun des vingt-trois joueurs qui la compose ?

Thèmes Spécifiques :

Thèmes Principaux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Sport	21	13	8,94	42,56	3,19%	6,12%
Jeux vidéos	7	6	3,67	52,38	0,50%	2,04%
Bourse marché et assurance	3	3	0,87	28,89	0,23%	0,87%
Thèmes Secondaires	**Tokens**	**Types**	**Poids du Thème**	**Poids/Tokens du Thème**	**% mot colorié sur le total de mot du Thème**	**% mot colorié sur le total de mot de l'article**
Cinéma	3	2	0,26	8,68	0,33%	0,87%
Europe	2	2	2,00	100,00	0,43%	0,58%
Conflits armées	3	1	0,60	20,00	0,07%	0,87%
Télévision	2	1	2,00	100,00	0,27%	0,58%

Thèmes Généraux :

Thèmes Généraux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Loisirs	28	19	12,60	45,02	0,47%	8,16%
Economie	3	3	0,87	28,89	0,13%	0,87%
Culture	3	2	0,26	8,68	0,03%	0,87%
Continent	2	2	2,00	100,00	0,07%	0,58%
Défense	3	1	0,60	20,00	0,03%	0,87%
Communication médiatique	2	1	2,00	100,00	0,06%	0,58%

Figure 18 : Résultat de l'analyse thématique d'un article sur le sport

Résultat de l'analyse de "sport02.txt" :

-- Coupe du monde : " Le d ' une victoire -nous vivants " La France entière a fêté la victoire des Bleus , dimanche . Pour beaucoup de Français , l ' occasion de danser autour d ' un drapeau qu ' ils n ' avaient sorti que lors d ' occasions funestes . La France qui danse . Il n ' est pas tout à fait 18 h 57 , dimanche 15 juillet , et le pays entier ressemble à une immense fan-zone . Le coup de sifflet de la finale du Mondial 2018 n ' a pas encore retenti , le compte à rebours est à peine entamé , et déjà le pays entier envahit les avenues , les places , les ronds-points et les . Cinq , quatre , trois , deux , un ... A la 90e minute de jeu , vuvuzelas , les trottoirs de Paris , de Marseille et de Lyon se transforment en dance floor . La France a gagné ? Alors on danse , dirait Stromae . " C ' est comme un bouchon de . La pression , la libération , le débordement " , sourit Nicolas , un trentenaire de Gradignan , en Gironde , venu regarder le match " à Chaban " , ce stade du cœur de Bordeaux . Ici , et dans toutes les rues de France , les drapeaux tricolores s ' exhibent sous toutes les formes : en cape , en robe , en bandeau ou en chouchou dans les cheveux pour les , en bandana pour les . Le bleu-blanc-rouge se décline sur les joues et les lèvres , mais aussi sur les casquettes , les perruques ou les hauts-de-forme , sans oublier les colliers en papier crépon . Au milieu de cette marée tricolore , quelques drapeaux tunisiens , marocains ou algériens s ' agitent avec la même folie et la même énergie . Sous le soleil de la capitale ou entre deux averses dans les du sud de la France , on fait la fête . Quelques packs de , un lecteur MP3 branché sur une enceinte , trois fois rien suffisent quand on a la joie au cœur . " On avait ce rôle -là , aussi , de rendre les ", explique Kylian Mbappé sur TF1 , avant d ' oser quelques pas de gwara gwara en direct depuis Moscou . " La la la la , lalalalala " , chante la fan-zone de Bondy , cette ville de Seine-Saint-Denis où le surdoué a fait ses débuts . Un peu plus tôt , sur la pelouse du stade Loujniki , le défenseur Samuel Umtiti avait lui aussi esquissé un pas de danse . Ça bouge ...

Thèmes Spécifiques :

Thèmes Principaux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Conflits armées	22	14	6,68	30,37	1,02%	2,32%
Spectacle vivant	13	11	6,83	52,56	1,06%	1,37%
Jeux vidéos	11	11	3,85	35,02	0,92%	1,16%
Thèmes Secondaires	**Tokens**	**Types**	**Poids du Thème**	**Poids/Tokens du Thème**	**% mot colorié sur le total de mot du Thème**	**% mot colorié sur le total de mot de l'article**
	9	9	6,75	75,00	1,32%	0,95%
Sport	7	6	5,03	71,90	1,47%	0,74%
Europe	6	6	5,50	91,67	1,29%	0,63%
Danse	6	5	3,17	52,78	0,73%	0,63%

Thèmes Généraux :

Thèmes Généraux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Loisirs	23	22	12,89	56,02	0,59%	2,42%
Culture	24	21	13,75	57,29	0,32%	2,53%
Défense	22	14	6,68	30,37	0,39%	2,32%
Continents	9	9	8,50	94,44	0,33%	0,95%
Societe	9	9	6,75	75,00	0,48%	0,95%
Communication médiatique	4	4	2,50	62,50	0,25%	0,42%
Santé	4	3	4,00	100,00	0,35%	0,42%
Environnement	3	3	1,83	61,11	0,07%	0,32%
Politique	2	2	1,33	66,67	0,06%	0,21%

Figure 19 : Résultat de l'analyse thématique d'un article sur le sport

Ainsi, nous observons (figure 18) que la suite de mots « Coupe du monde » a bien été matchée. Séparer « coupe » de « monde » n'aurait pas eu d'intérêt. Malheureusement, les suites de mots n'ont pas toujours été repérées de cette manière. Ainsi, nous pouvons remarquer (figure 19) qu'aucun ngram du sport n'a été trouvé dans le deuxième article. Par exemple, les deux substantifs « victoire » et « Bleus » de la portion « victoire des Bleus » ont été distingués à tort ; le terme « Victoire » attribué au thème « ConflitsArmées » et « Bleus » attribué au sport.

Cependant, bien que fréquent, un ngram n'est pas forcément pertinent. Par exemple, le trigram « conférence de presse », assez fréquent dans le corpus, ne concerne par le sport à proprement parlé. Cette suite de mots est dans la liste du thème spécifique « Presse ».

La confusion entre le thème « ConflitsArmées » et « Sport » n'étonne pas puisque de nombreux mots de vocabulaire sont communs tels que « victoire » et « défaite ». L'on peut également citer des ngrams tels que « la victoire de » et « face à la » qui, suivis du nom d'un pays permettent d'aider à la détection thématique dans le corpus. Sur les 50 articles du monde du sport, le segment « face à la » et « la victoire de » apparaissent 11 fois chacun. Mais ces suites de mots, si elles ne sont pas propres à un thème, ne seront pas forcément pertinentes pour localiser un thème précis. En effet,

ces expressions peuvent se retrouver dans un tout autre contexte. Ajouter des ngrams trop communs générerait pareillement une surcharge de liste qui handicaperait l'analyseur tout autant qu'une surcharge de mots simples. L'ajout de ngrams n'est donc pas toujours indispensable.

Toutefois, cet ajout peut-être particulièrement indispensable. Dans le domaine du sport « Coupe du monde », « Tour de France », « Roland Garros » ou encore « équipe de France » et « tirs au but »… en sont de bons exemples. L'ajout de ces nouvelles suites, ne pourraient qu'améliorer la détection pour un article traitant du sport.

Ajoutons que le calcul par proximité ne serait pas forcément utile puisque ces suites de mots apparaîtront toujours dans cet ordre. Il suffit de repérer les suites de mots à l'aide d'un logiciel comme Lexico et de les relever afin de compléter nos listes thématiques.

Néanmoins, repérer les ngrams ne règle pas le problème de relation entre les mots, comme le dit [A. Guille, C. Favre, 2014]. Les ngrams sont de expressions figées ne consédérant pas les liens de positions entre les termes.

L'idéal serait d'attribuer une sorte de hiérarchie entre les mots. Les adjectifs prendraient la couleur du substantif qu'il qualifie. Par exemple, dans « belle victoire », l'adjectif « belle », qui n'appartiendrait à aucune liste, recevrait la couleur du substantif « victoire » auquel il est lié. Les adverbes auraient la couleur du verbe duquel il précise l'action. Ainsi, dans « Combattre fortement », l'adverbe obtiendrait la couleur du verbe.

Une autre solution pourrait être envisagée, en particulier en appliquant un poids plus important au terme vraiment propre au sujet ou au ngrams. Un poids qui pourrait faire pencher la balance et éviter d'avoir à supprimer tous les mots équivoques.

Notons qu'à l'inverse d'une liste surchargée, un token présent dans une seule liste peut aussi faire pencher la balance vers son thème. Le test a été confirmé avec un article ayant obtenu comme thème principal : « BourseMarchéAssurance », relatif au thème générique de l'économie, simplement parce qu'il contenait une importante fréquence du symbole « € ». Ce symbole, présent dès qu'un prix est mentionné génère une confusion. Face à ce problème, plusieurs solutions sont alors à envisager :

- Supprimer le token de la liste. Mais s'il est propre au thème, c'est illogique,

- Ajouter le token dans les autres listes pour les thèmes où il serait susceptible d'apparaître. Dans ce cas, le « € » pourrait être ajouté au thème spécifique « Commerce ».

- Intégrer le token dans une expression qui ciblerait plus le thème en question, tout en s'assurant que cette expression serait bien usitée dans les articles. Par exemple, si nous reprenons le symbole « € », l'expression « devise en € » serait plus proche du thème « BourseMarchéAssurance » qu'une expression telle que « une robe à 40€ ».

Ajoutons qu'il est probable que, hors du contexte de prix, le symbole « € » apparaîtrait surtout sous la forme « euros ». On parle de « devise en euros » et non « en € ». Par contre, on écrit 30,99 € et non 30,99 euros.

Il est impossible d'obtenir des listes ne contenant que des termes propres à un seul thème spécifique.

2) Le choix de purger les listes

Après ces observations, procédons à quelques réflexions sur la surcharge de listes. Il apparaîtrait deux solutions possibles :

- Soit surcharger toutes les listes afin qu'elles soient toutes à peu près au même niveau ;

Remarque : L'inconvénient avec cette technique est que le poids (dépendant de la fréquence) fera pencher vers le thème qui a le plus de terme équivoque. Or, il faudrait qu'il penche envers les thèmes qui ciblent bien le sujet.

- Soit épurer toutes les listes pour ne garder que le vocabulaire minimum (le plus ciblant) ;

Remarque : L'inconvénient est que des thèmes qui ne comportent que peu voire pas de vocabulaires pertinents ne seraient pas du tout correctement évalués.

Il faut tout de même admettre que ce n'est pas tant la taille des listes qui pose problème mais ce qu'elles contiennent. En effet, s'il s'agit d'une longue liste contenant des termes attribués exclusivement à son thème ou presque, cela ne pose pas de problème, bien au contraire. Par exemple, dans notre base de données, la liste sur la météorologie et le climat contient 3619 lexies et elle n'est pourtant jamais détectée de manière impromptue car son vocabulaire cible parfaitement bien le domaine.

En revanche, si la taille démesurée d'une liste résulte d'un vocabulaire équivoque, il y a fort à parier que ce thème fera pencher la balance de la détection de son côté au détriment d'une autre au vocabulaire plus précis mais plus mince.

Il convient donc d'épurer les listes et de les nettoyer d'un vocabulaire parasitant la détection.

Cependant, bien que la presse s'efforce d'employer des termes assez spécifiques et lorsqu'elle traite d'un sujet afin d'aiguiller les lecteurs, que ce soit par des termes techniques ou bien par des noms propres, elle ne peut pas pour autant rédiger dans un langage totalement hermétique au risque de perdre l'attention du lecteur. Des termes légèrement équivoques doivent aussi être conservés pour des contextes syntaxiques bien particuliers.

Toutefois, ajoutons que les préposés de la presse travaillant dans des revues spécialisées ont certainement affaire à un lexique précis. À titre d'exemple, un mot comme « mogotte » appartenant au domaine de la géomorphologie ne serait pas forcément compris par le commun des mortels mais plus certainement par les puristes d'un journal spécialisé. Mais alors, rien ne coûte d'ajouter ses mots à nos listes puisqu'ils ne seront que rarement et spécifiquement pris en compte.

Illustrons à présent par plusieurs tests le phénomènes d'allègement de listes.

Voici un tableau résumant les résultats de l'analyse thématique, avec les thèmes les plus fréquents, sur les 50 articles d'un thème portant sur la décoration et le bricolage :

Thèmes repérés	Thème principaux	Thème secondaire (les 3 premiers)
Jeux vidéos	44/50	3/50
Cuisine et Gastronomie	31/50	16/50
Arts plastiques	34/50	9/50

Tableau 1 : Score de présence des trois thèmes gagnants de l'analyse thématique du corpus sur le Bricolage avant nettoyage de la liste « JeuxVidéos »

Le thème spécifique des jeux vidéos est majoritaire alors qu'aucun article n'aborde le sujet.

Lorsqu'on regarde les termes colorisés pour ce thème intrus, on obtient de nombreux mots portant un sens assez général mais certainement pas propre au domaine des jeux vidéos. De nombreux tokens ont été supprimés de la liste. Il est donc certain que ce vocabulaire a fait pencher la balance en sa faveur.

Cependant, il est parfois difficile de supprimer des mots. Le verbe « jouer » pose ce problème parce qu'il est aussi présent dans notre thème spécifique des jeux de sociétés ou du sport mais là est la justification de garder, modestement, un vocabulaire équivoque. Par exemple, dans un article, il y a la phrase « on adore jouer avec le papier » et le terme jouer a été classé comme jeu vidéo du fait du poids important qu'avait les autres termes. Pourtant, ce verbe est présent dans les thèmes spécifiques « Cinéma », « Musique », « JeuxSociété » et « Sport ». L'important était donc de le lier à son cooccurrent « papier ».

Le mot « soigner », peut être utilisé dans un article parlant des jeux vidéos si l'on évoque, par exemple, l'action d'un personnage ; mais alors, il vaut mieux qu'il soit classé dans santé et non jeux vidéo puisque la sémantique de mot ne se rapproche pas du thème générique « Loisirs » mais plutôt « Santé ». Nous retrouvons ici la pertinence du choix exclusif des termes précis dans la composition des listes de vocabulaires thématiques.

Ceci nous amène à comprendre qu'à défaut de pouvoir matcher tous types de suites de mots, l'analyse des cooccurrents pourrait s'avérer être une méthode de choix pour rapprocher les mots entre eux et désambiguïser la sémantique. Voici des exemples, tirés de l'observation de l'épuration de la liste, qui en témoignent :

- Un mot comme « règle » (non supprimé de la liste) peut être utilisé dans de nombreux contextes. C'est pourquoi, dans le cas de jeux vidéo, il faut que les cooccurrents soit « jeu », « partie » ou « gameplay »...

- Le mot « touche » spécifiant l'élément d'une manette de jeu ou d'un clavier apparaît dans l'article de bricolage dans la suite de mots : « une touche de confort » ce qui est différent.

Pour cibler les jeux vidéos, ce terme devrait sans doute être rapproché de mots tels que « pouce » et de verbes tels que « appuyer ».

- Le verbe agir a été repéré comme Jeux vidéos ; or, dans un article, il est dit : « laisser agir ». Il faut donc garder ces deux verbes car le sens renvoi à un produit qu'on applique sur quelque chose et non à une quelconque action. Il s'agit de repérer la construction du verbe « laisser » suivi d'un autre verbe.

- Dans la suite de tokens « Soulever délicatement la dentelle », on pourrait trouver une pertinence à localiser le bigram « V + adv » souvent présent lorsqu'il est question d'une directive puisque nos articles de bricolage ressemblent à des notices ou des petits modes d'emploi.

Retournons aux résultats et aux deux autres thèmes « CuisineGastronomie » et « Artsplastiques ». Leur présence en tant que thèmes principaux surprend moins car ils comportent un vocabulaire plus proche du bricolage et de la décoration.

Concernant le thème « ArtsPlastiques », aucun des termes ciblant ces thèmes n'a été colorisé de manière erroné. Des mots tels que « bois », « papier », « peinture », « spatule »… se retrouvent aussi bien dans les articles parlant de bricolage que dans ceux traitants d'Arts plastiques. Ce léger chevauchement de vocabulaire entraîne une concurrence entre les deux thèmes spécifiques : « ArtsPlastiques » et « DecoBricolage ».

Pour le thème « CuisineGastronomie », il semble que la raison soit différente. Son score, non négligeable, est probablement dû au contexte d'une partie des articles du corpus, faisant parfois référence au repas et à la manière de l'embellir. De ce fait, de nombreux termes de la cuisine y apparaissent et induisent en erreur l'analyseur thématique.

Face à ce problème, il semble qu'il n'y ait guère de solution hormis en élargissant l'analyse du texte ; peut-être en repérant de quel type de journal il provient, à savoir si c'est un journal de cuisine ou de décoration. Le principal problème de détection, réside bien dans le poids du vocabulaire. Un article de décoration florale, présent dans notre corpus, contient énormément de termes sur la botanique et n'a malheureusement pas catégorisé Decobricolage du moins en première position. L'astuce pourrait être de détecter le vocabulaire du titre de l'article qui peut influencer. À ce propos, [D. Bernhard, I. Falk, C. Gérard, septembre 2014] évoque le rôle crucial des conditions textuelles en parlant de la position du néologisme dans le texte (titre de l'article, corps du texte ou note de bas de page ; début ou fin de paragraphe) qui n'est jamais neutre.

Après avoir nettoyé la liste « jeux vidéos » des termes gênants pour nos 50 articles, le souhait a été de relancer le traitement afin de voir si le thème qui passerait devant serait Arts plastiques et Cuisine-Gastronomie ou bien un autre.

Thèmes repérés	Thème principaux	Thème secondaire (les 3 premiers)
Jeux vidéos	3/50	7/50
Cuisine et Gastronomie	40/50	4/50
Arts plastiques	37/50	6/50
Bourse marché et assurances	18/50	10/50

Tableau 2 : Score de présence des trois thèmes gagnants de l'analyse thématique du corpus sur le Bricolage après nettoyage de la liste « JeuxVidéos »

Au regard des résultats (tableau 2), nous remarquons, sans surprise, l'importante baisse pour le thème « JeuxVidéos ». Une perte qui a profité aux deux autres thèmes concurrents. Mais, les « points » semblent aussi s'être reportés sur un thème spécifique du thème générique « Économie », celui intitulé « BourseMarcheAssurance » qui n'a pas beaucoup de rapport avec le sujet de notre corpus. Lorsqu'on analyse les mots colorisés comme appartenant à ce nouveau thème, force est de constater qu'ils ne peuvent être retirés. En voici des exemples : « *produit », « fabriquer », « bénéficier », « clientèle », « boutique », « somme », « bon », « montant », « aide », « plafond »*…

Néanmoins, la liste du thème Bourse – Marché – Assurances est énorme (plus de 1000 termes) dont beaucoup sont très génériques. Après relance du traitement, sans cette liste, il a été constaté que le thème des arts plastiques restait en tête, ce qui rassure en quelque sorte.

Plus intéressant, un autre thème dénommé « FormationPédagogie » n'a été repéré qu'une seule fois alors qu'il conviendrait puisque la plupart de ces articles de Bricolage sont des « modes d'emploi » pour fabriquer des décorations. Mais le contenu de cette liste reste à améliorer. Autre observation, le thème « ForcesArmees » a pris plus de poids en raison de la présence du substantif « décoration », très usité dans le domaine militaire.

Ces exemples démontrent l'importance de commencer le parsing de l'article par l'analyse des cooccurrents des substantifs afin de procéder à la sélection des listes les plus conformes aux relations sémantiques du texte.

Pour finir, une liste a été constituée pour le thème « DecoBricolage », selon les observations du vocabulaire utilisé dans le corpus. Sur la cinquantaine des articles du corpus, 43 ont obtenu ce thème comme principal, ce qui est une bonne nouvelle mais il est difficilement possible d'affirmer si ce repérage correct fonctionnera avec un autre corpus sur le bricolage. Même si le nouveau thème est globalement bien détecté, il est en forte concurrence avec le thème « ArtPlastiques » dont le vocabulaire est proche.

Autre difficulté, cette liste concurrente est parsée par le programme avant notre liste nouvellement créée. Elle obtient donc la prédominance lorsque son vocabulaire y est présent. Nous nous heurtons au même problème que lorsque le mot « apprentissage » était colorisé à la place du bigram « apprentissage machine ».

Les observations ont révélé plusieurs constructions syntaxiques pertinentes pour la détection de ce thème DecoBricolage. Par exemple, « Laisser + V » comme dans « laisser sécher » ; ou « V + -y » comme dans « placez-y » ou « insérez-y ». Également, les adverbes sont assez fréquents pour préciser les intensités des actions : « fermement », « précisément », « facilement », « soigneusement », délicatement »…

Aussi, il y aurait les verbes à l'impératif (« Pliez le papier », « Collez la feuille ») mais la lemmatisation transforme tout à l'infinitif et empêche de préciser cela.

D'autres locution verbales ont été importantes pour ce thème et notamment « donner vie » et « donner du caractère ».

En fait, il y a un point sur lequel on peut être satisfait. Certains articles, rangés dans la rubrique bricolage dans le journal reçoivent un tout autre thème en résultat de détection. Ces quelques articles traitaient par exemple de l'ouverture de magasin de bricolage ont reçus comme thème principale de détection « BourseMarcheAssurance » ce qui est tout à fait correct.

Ainsi, nous pouvons voir que sous une conduite supervisée, nous aimerions classifier tel et tel article dans la catégorie Bricolage, comme le fait le journal mais en réalité, certains articles ne traitent pas à proprement parlé de décoration ou de bricolage. Il est donc normal que ce thème ne soit pas principal. c'est pourquoi, sur les 7 articles non reconnus comme « DecoBricolage », on parlait souvent d'entreprise de Bricolage avec un sujet commercial et leur thème principal était bien « BourseMarcheAssurance ».

Pour conclure, du fait de la prédominance du thème « JeuxVidéos » et l'influence de l'épuration d'une liste, un test a été réalisé sur un corpus d'articles traitant de ce sujet. Sans surprise le thème était correctement reconnu. Sur la totalité des 10 articles, le thème « JeuxVidéos » apparaissait en première place de la détection.

Ensuite, la réduction de la liste « JeuxVidéos » permet le rendu comparatif suivant (tableau 3) :

Articles se Jeux vidéos	Poids du thème avant réduction	Poids du thème après réduction	Position en tant que thème principal
01	50,08	29,97	Toujours1er
02	28,59	21,03	Toujours 1er
03	11,70	4,61	Spectacle vivant et Medecine
04	10,09	50,26	Sport
05	30,74	17,06	Toujours 1er
06	18,44	15,14	Toujours 1er
07	74,66	66,56	Bourse-Marché-Assurance
08	19,62	13,28	Bourse-Marché-Assurance
09	50,71	36,14	Toujours 1er
10	12,25 (2nd position derr Bourse)	7,64	Toujours 2nd
11	52,45	39,54	Toujours 1er

Tableau 3 : Comparaison de l’analyse d’un corpus d’articles sur le JeuxVidéos avant et après l’épuration de sa liste

La réduction de la liste a eu pour effet de baisser le score du thème « JeuxVidéos ». Ceci présente un risque car, bien que le thème reste dominant dans la majorité des articles, il a perdu sa première place dans d’autres et cela aurait peut-être été plus grave si la liste avait encore été épurée. Cette dernière expérience nous confronte au dilemme de savoir jusqu’à quel point purger une liste est conseillé.

Néanmoins, il est fort possible que les thèmes qui arrivent à prendre le dessus soient également des thèmes dont les listes contiennent des termes trop équivoques et elles aussi se retrouvent à faire pencher la balance en leur faveur. C’est clairement le cas de la liste du thème « BourseMarcheAssurance » qui, au cours des divers tests était constamment présente parmi les thèmes principaux.

Le mieux reste d’épurer les listes car le temps de calcul n’en sera que plus réduit. Autre avantage, nous travaillons dans une optique de recherche néologique. Or, réduire les vocabulaires des listes rendra la coloration plus précise sur les contextes entourant un mot candidat à la néologie. En effet, l’objectif n’est pas de détecter les thèmes principaux mais tous les thèmes contextuels qui se cacheraient au sein du texte. Tout changement de vocabulaire est potentiellement un changement de thème ainsi que l’évoque [D. Bernhard, I. Falk, C. Gérard, septembre 2014], cité dans l’état de l’art.

Précisons un peu cette pensée. On est en droit de se demander si le fait que l'analyseur ait mal choisi un thème soit forcément une erreur. En effet, nous pratiquons nos tests sous un mode supervisé, en souhaitant bien que, pour tel article, tel thème soit choisi par l'analyseur dans telle catégorie. Pourtant, songeons qu'un texte peut être de manière générale d'un thème spécifique particulier et pour autant comporter de nombreux contextes focalisés sur d'autres thèmes spécifiques. L'analyseur aura alors toutes les bonnes raisons de relever les thème « non désirés » puisque les probables néologismes qui y apparaîtraient seraient en lien, non pas avec le thème général du document, mais bien plutôt avec les contextes présents. Par exemple, dans des articles de jeux vidéos du corpus étudié plus haut, le terme « pédagogique » était matché comme appartenant à ce thème. Cela ne devrait pas, quand bien même l'article traiterait principalement de ce thème. Ce mot appartient à la liste « FormationPedagogie » et c'est bien dans un contexte de ce genre qu'il est employé dans l'article.

Par conséquent, il vaut mieux plusieurs thèmes révélant ainsi les différents thèmes évoqués en contexte, si petits soient-il, que la détection d'un seul thème général prédominant qui obscurcirait les résultats de thèmes tout aussi pertinents.

Le principal problème d'une liste très épurée réside dans le fait qu'un article ne comportant presque pas de mots clés ou d'entités nommées ne sera pas correctement analysé. D'où l'importance d'appliquer un poids plus important, dès la moindre entités nommées ou le moindre ngrams détecté.

C) Apprentissage supervisé/non supervisé

Le travail de cette sous-partie vise à tester l'efficacité de certains algorithmes d'apprentissage automatique pour la prédiction.

1) L'apprentissage automatique pour distinguer les thèmes

Le machine learning connaît un succès grandissant depuis quelques années dans le domaine de l'intelligence artificielle. Basé sur des méthodes de calculs d'observation, il offre un terrain d'expertise intéressant pour notre détection thématique. Le test d'apprentissage automatique que nous avons effectué a été fait à l'aide du logiciel Weka qui implémente plusieurs algorithmes dans ce domaine.

Précisons tout d'abord que cette méthode est supervisée, par conséquent, nous attribuons à chaque articles un seul thème particulier, correspondant au nom de la rubrique dans laquelle il était classé dans le journal. Tout article est plurithématique, mais l'objectif de ce test n'est pas de trouver à tout pris les thématiques les plus proches du contenu mais, dans un premier temps, de vérifier si par apprentissage automatique supervisé, un algorithme est capable de catégoriser correctement des articles en fonctions des bonnes étiquettes.

Pour effectuer ce travail d'apprentissage automatique, un corpus lié au thème générique « Economie » a été choisi avec trois thèmes spécifiques : « Agriculture », « BourseMarcheAssurances » et « BudjetFiscaliteComptabilite ».

Un corpus d'entraînement de 30 articles et un autre corpus de test de 30 articles ont été créés. Chaque corpus ayant 10 articles par thème spécifique.

La logiciel Weka, utilisé au travers du langage python (cf. annexe 5), a permis de lancer les algorithmes sur les corpus. Trois algorithmes ont été sélectionnés :

- KNN : K plus proche voisin qui, selon [L. Ralaivola, 2007] est « l'algorithme le plus simple d'apprentissage automatique supervisé qui (...) permet d'obtenir de très bons résultats de classification. »

- J48 : Arbre de décision où, selon [M. Ballabriga, 2005] « on cherche à discriminer les exemples selon leur classe et en fonction d'attributs considérés comme les meilleurs parmi tous les autres au sens d'un critère de données ».

- Multilayer Perceptron : Perceptron multicouche qui est un petit réseau de neurones, méthode très utilisée dans le machine learning.

Concernant les caractéristiques permettant de discriminer les trois thèmes spécifiques, le logiciel Lexico a permis de faire ressortir les fréquences pertinentes ainsi que les segments répétés et les différences entre les articles.

Ainsi, pas moins de 120 features ont été choisis dans le corpus train. Elles peuvent être subdivisées en trois types :

- Les features propres au thème. Par exemple, les mots « bétail », « PAC », « hectare », « sècheresse »… ne figurent que dans les articles traitant de l'agriculture. De même pour le thème de la fiscalité avec les termes « TVA », « ISF », « taxe d'habitation », « prélèvement » ...

- Les features en concurrence avec un des deux autres thèmes : des tokens tels que « aides » ou « cotisations » se retrouvent dans des articles à la fois sur l'agriculture et sur la fiscalité. Également, les mots « investissement » et « exportations » sont à la fois présents pour l'agriculture et pour le thème « BourseMarcheAssurance ».

- Les features où les trois thèmes sont en concurrence : des mots tels que « consommation » et « pays » en font partis.

Ces concurrences entendent simuler la confrontation que l'on rencontre entre plusieurs listes comportant des termes équivoques lors d'une analyse par méthode symbolique.

Aussi et de la même manière qu'avec l'analyseur thématique et les listes de vocabulaires, les features ne jouent que sur la fréquence des mots.

Précisons que, d'après les observations de caractéristiques, les thèmes « Agriculture » et « Fiscalité » sont les deux thèmes qui entrent le moins en concurrence.

Les trois algorithmes d'apprentissage automatique ont été lancés à deux reprises selon deux étapes :

- Lors de la première étape, nous conservons l'ensemble des features en concurrence.

- Lors de la seconde, nous ne travaillons qu'avec les termes propres au thème.

L'idée est de voir si un vocabulaire ciblant parfaitement les articles catégorise convenablement le texte ou bien s'il est nécessaire d'utiliser des termes périphériques bien qu'équivoques.

2) L’apprentissage automatique : bilan correct

L’ensemble des caractéristiques inspectées pour le corpus train sont a présent expérimentées sur le corpus test.

Si nous prenons en compte les 120 features pour notre première étape, nous obtenons les résultats suivants (tableau 4, 5 et 6) :

Algorithme : J48

	Agriculture	BourseMarcheAssurance	BudjetFiscaliteComptabilite
Précision	1,0	0,75	0,88
Rappel	0,9	0,9	0,8
Précision globale	0.87		
Rappel globale	0.86		
F-mesure	0.87		

Tableau 4 : Résultat des prédictions de l’algorithme J48 pour le corpus sur l’Économie à l’aide des features concurrentes

Algorithme: KNN

	Agriculture	BourseMarcheAssurance	BudjetFiscaliteComptabilite
Précision	1,0	1,0	1,0
Rappel	1,0	1,0	1,0
Précision globale	1,0		
Rappel globale	1,0		
F-mesure	1,0		

Tableau 5 : Résultat des prédictions de l’algorithme KNN pour le corpus sur l’Économie à l’aide des features concurrentes

Algorithme: Perceptron

	Agriculture	BourseMarcheAssurance	BudjetFiscaliteComptabilite
Précision	1,0	1,0	0.90
Rappel	1.0	0,9	1,0
Précision globale	0.96		
Rappel globale	0.96		
F-mesure	0.96		

Tableau 6 : Résultat des prédictions de l’algorithme Perceptron pour le corpus sur l’Économie à l’aide des features concurrentes

De par ces trois premiers tableaux, la performance de l'algorithme des K plus proche voisin qui a réussi à systématiquement attribuer la bonne étiquette à l'article, ressort indubitablement.

Pour les deux autres algorithmes quelques observations des articles non correctement prédits s'impose. Concernant J48, un article du thème « BourseMarcheAssurance » a été classé comme « BudjetFiscaliteComptabilite » alors qu'il traitait pourtant de Wall Street et comportait des sujets sur la FED et les taux d'intérêts. Inversement, deux articles du thème « BudjetFiscaliteComptabilite » ont été prédits comme thème « BourseMarcheAssurance », les deux ont un titre éloquent comportant le mot « fiscalité » : « Le Crédit agricole freiné par la fiscalité » et « Fiscalité immobilière, ce que vous réserve 2013 » ce qui montre que la prédiction semble vraiment s'être trompée. Confirmons-le par l'observation du contenu qui révèle un de features importantes pour caractériser le thème « BudjetFiscaliteComptabilite » telles que les termes « surtaxe », « épargne », « contribuables », « plafonnement », « revenus fonciers », « taxe sur les dividendes », « impôt sur le revenu », « prélèvement à la source » et « niches fiscales ».

Enfin, rappelons que de nombreux termes sont en concurrences entre le domaine de la fiscalité et celui de la bourse. Ce chevauchement perturbe la prédiction, tout comme avec la méthode symbolique. Après avoir rajouté les features manquantes, l'article a été correctement prédit.

Un seul article du thème spécifique « Agriculture » a reçu l'étiquette « BourseMarcheAssurance » mais l'explication est simple, le texte évoquait surtout le marché de l'emploi dans le secteur du bio. Par ailleurs, si l'on analyse cet article avec notre analyseur à méthode symbolique, basé sur l'ensemble des listes de vocabulaire thématique, nous obtenons le même résultat : le thème BourseMarcheAssurance en pôle position des thèmes principaux. La méthode symbolique se recoupe avec le machine learning alors que c'est plutôt le thème spécifique de l'emploi et du travail qui aurait peut-être dû être matché. Le contenu de cette liste semble insuffisant, ou bien la prédominance du vocabulaire d'une liste par rapport à l'autre a à nouveau eu lieu.

Avec le Perceptron, cet article sur l'emploi dans le bio a correctement été prédit comme faisant partie du domaine de l'agriculture. En revanche, pour les autres articles mal classés de J48, ce sont d'autres articles qui ont été mal attribués. Et il est possible d'affirmer qu'ici, c'est le Perceptron qui a eu raison. Par exemple, il a prédit la classe « BudgetFiscaliteComptabilite » au lieu de « BourseMarcheAssurance », l'étiquette initiale, pour un article ne traitant effectivement que de fiscalité. L'erreur vient donc de l'annotateur humain. Le reste des articles a été convenablement répertorié.

Après avoir supprimé les features en concurrence, il n'en restait que 65, les résultats ont été les suivants (tableau 7, 8 et 9) :

Algorithme: J48

	Agriculture	BourseMarcheAssurance	BudjetFiscaliteComptabilite
Précision	1.0	1.0	0.77
Rappel	1.0	0.7	1.0
Précision globale	0.92		
Rappel globale	0.9		
F-mesure	0.91		

Tableau 7 : Résultat des prédictions de l'algorithme J48 pour le corpus sur l'Économie sans features concurrentes

Algorithme: KNN

	Agriculture	BourseMarcheAssurance	BudjetFiscaliteComptabilite
Précision	1.0	1.0	1.0
Rappel	1.0	1.0	1.0
Précision globale	1.0		
Rappel globale	1.0		
F-mesure	1.0		

Tableau 8 : Résultat des prédictions de l'algorithme KNN pour le corpus sur l'Économie sans features concurrentes

Algorithme: Perceptron

	Agriculture	BourseMarcheAssurance	BudjetFiscaliteComptabilite
Précision	1.0	1.0	0.91
Rappel	1.0	0.9	1.0
Précision globale	0.97		
Rappel globale	0.97		
F-mesure	0.97		

Tableau 9 : Résultat des prédictions de l'algorithme Perceptron pour le corpus sur l'Économie sans features concurrentes

KNN conservant les meilleurs résultats, analysons toujours les deux algorithmes contenant des erreurs de prédiction, à savoir, J48 et le Perceptron.

Au sujet des résultats du J48, l'article précédemment cité dont l'erreur provenait de l'annotateur, a correctement été rangé comme le Perceptron l'avait fait. Ceci démontre qu'avoir supprimé les vocabulaires en concurrence a aidé le J48 dans ses prédictions. Deux autres articles de « BourseMarcheAssurance » ont été classés en tant que « BudgetFiscaliteAssurance ». Le résultat est étonnant pour le premier article qui parle de l'histoire d'une entreprise et de son commerce. Le second article avait des lacunes lexicales : ces quelques mots clés appartenaient au domaine de la fiscalité.

Le Perceptron refait la même heureuse erreur de classer l'article traitant de l'emploi dans l'agriculture bio selon l'étiquette « BourseMarcheAssurance ».

En conclusion de cette dernière sous-partie, nous pouvons remarquer la différence de prédiction à la fois entre les articles et les algorithmes. Ces observations confirment encore, au travers de l'apprentissage automatique l'importance des suites de mots qui permettent de circonscrire la sémantique d'un terme. Elles précisent considérablement les thèmes à privilégier pour la sélection d'un thème. Par exemple, le mot « revenu » peut être rapprocher du verbe « revenir » mais « impôt sur le revenu » est univoque et ne sera que dans la liste du vocabulaire de la fiscalité.

CONCLUSION

La détection thématique de texte nécessite de se pencher sur la quasi totalité des problèmes de compréhension du langage naturel par un ordinateur. Elle nous place au cœur de l'intelligence artificielle puisque, si une machine catégorise bien les documents reçus en entrés, c'est qu'elle les a « compris ». Quand un ordinateur ne détecte pas bien un thème, cela s'apparente au quiproquo entre êtres humains. On utilise des mots que chacun comprend mais le sujet de conversation est différent. Pareillement, l'ordinateur repère bien les mots dans les listes mais choisit les mauvaises.

Pour cette détection, les méthodes de machine learning demeurent trop imperméables à la manipulation et nuisent au dialogue de compréhension entre l'homme et la machine.

Les méthodes symboliques, au contraire, sont parfois à privilégier pour permettre une correction consciente des données. En effet, cette méthode offre un avantage certain, celui de connaître l'état d'avancement de la détection et de savoir où effectuer les modifications pour en améliorer les performances. Seul bémol, l'approche paraît très descendante et l'utilisation de vocabulaires contrôlés, s'ils ne sont pas utilisés dans la vraie presse journalistique rend le travail chimérique.

C'est pourquoi, l'appui de corpus divers constitue la base sereine d'une progression maîtrisée.

ANNEXES

Annexe 1 : Liste de vocabulaire thématique (logiciel Antidote)

Annexe 2 : Schéma de la hiérarchie thématique

Annexe 3 : Script Python de comparaison de contenu du vocabulaire entre deux listes

Annexe 4 : Script bash de nettoyage du corpus sur le sport pour la clusterisation

Annexe 5 : Parti du script python interagissant avec le logiciel Weka pour les calculs algorithmiques

Champ lexical de urbanisme, n. m. **Force**

Noms (118)

- adjoint
- agglomération
- aménagement
- aménageur
- architecte
- architecture
- artificialisation
- autorisation
- banlieue
- bâti
- bâtiment
- centre-ville
- cité
- cité-jardin
- code
- cohérence
- collectivité
- commune
- compétence
- concertation
- conseil
- constructibilité
- construction
- copropriété
- décentralisation
- délivrance
- démolition
- densification
- densité
- dérogation
- développement
- disposition
- document
- écologie
- écoquartier
- élaboration
- élu
- embellissement
- environnement
- équipement
- espace
- étalement
- expropriation
- foncier
- friche
- géographie
- gratte-ciel
- Grenelle
- H.L.M.
- habitants
- habitat
- habitation
- hectares
- îlot

Annexe 1 : Liste de vocabulaire thématique (logiciel Antidote)

Champ lexical de urbanisme, n. m.	**Force**
immeuble	
immobilier	
implantation	
infrastructure	
institut	
intercommunalité	
littoral	
logement	
loi	
lotissement	
lotisseur	
maire	
mairie	
métropole	
mitage	
mixité	
modification	
municipalité	
occupation	
opération	
parc	
parcelle	
patrimoine	
paysage	
paysagisme	
paysagiste	
périmètre	
périphérie	
permis	
plan	
planification	
préemption	
préfet	
préservation	
projet	
promoteur	
quartier	
réalisation	
réaménagement	
reconstruction	
règle	
règlement	
réglementation	
réhabilitation	
rénovation	
révision	
revitalisation	
schéma	
secteur	
servitude	
sol	
stationnement	
terrain	
territoire	
tramway	

Annexe 1 : Liste de vocabulaire thématique (logiciel Antidote)

Champ lexical de urbanisme, n. m. — **Force**

- transport
- urbain
- urbanisation
- urbaniste
- urbanité
- ville
- voirie
- zonage
- zone

Noms propres (3)

- Chandigarh
- Haussmann
- Le Corbusier

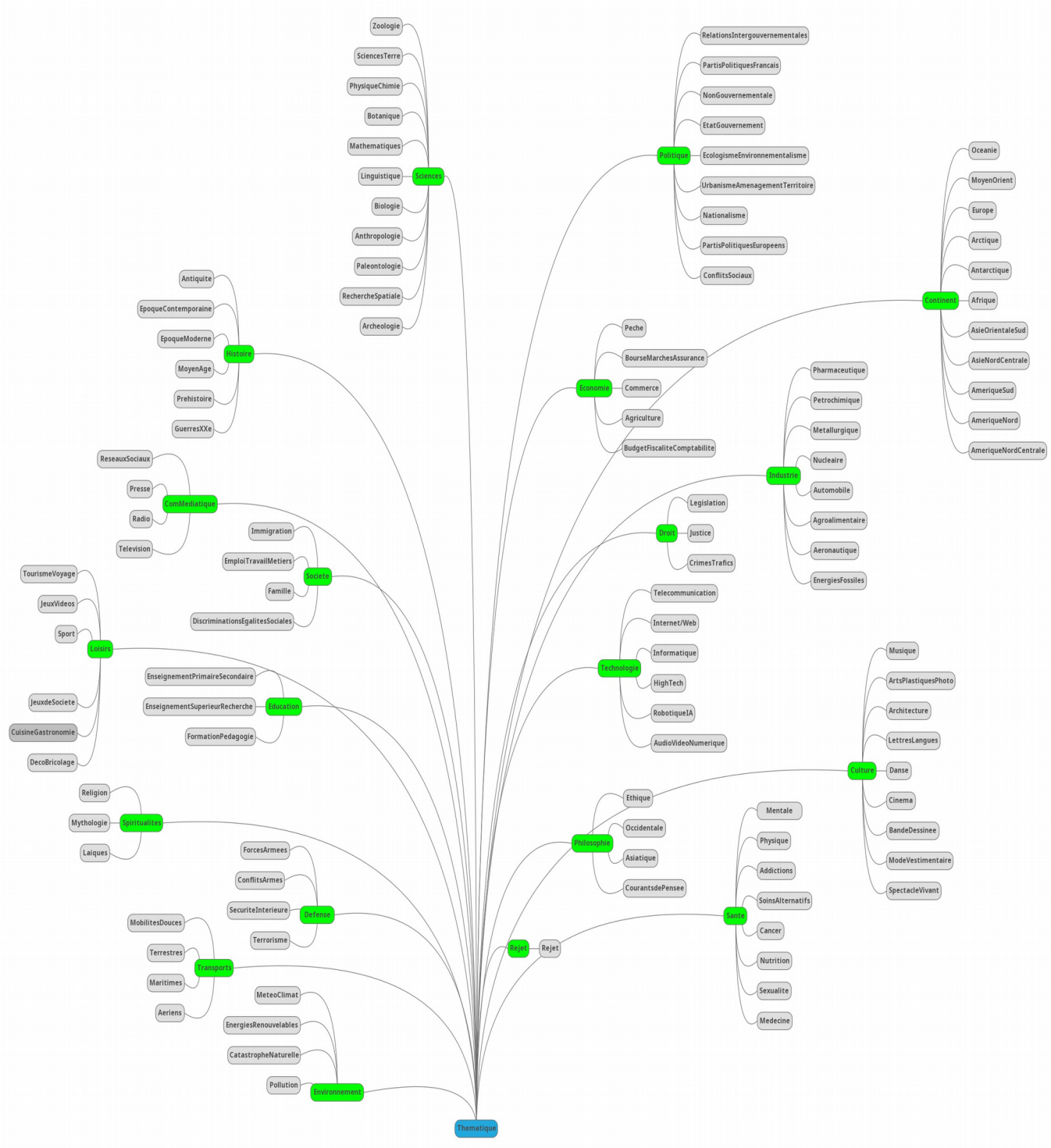
Adjectifs (42)

- administrative
- anarchique
- architecturale
- bâti
- carré
- chargé
- commercial
- communale
- concerté
- constructible
- consultatif
- départementale
- directeur
- durable
- écologique
- environnementale
- foncière
- haussmannien
- immobilier
- inconstructible
- inondable
- insalubre
- intercommunale
- littoral
- local
- locatif
- municipal
- opposable
- pavillonnaire
- paysagère
- périurbain
- piétonne
- public
- rénové
- résidentiel
- territoriale
- urbain
- urbanisable
- urbanisé
- urbaniste
- urbanistique

Annexe 1 : Liste de vocabulaire thématique (logiciel Antidote)

Champ lexical de urbanisme, n. m.	Force
vert	
Verbes (11)	
aménager	
architecturer	
bâtir	
construire	
délimiter	
démolir	
densifier	
élaborer	
prévoir	
pucer	
urbaniser	

Annexe 1 : Liste de vocabulaire thématique (logiciel Antidote)



Annexe 2 : Schéma de la hiérarchie thématique

```python
'''
Script de comparaison entre deux listes:
    - Entrée: 2 listes thématiques
    - Sortie:   Le nombre de mots de chaque liste
                Le nombre de mot communs
'''
chemin="../../../themes/Theme_"
extension=".txt"
liste1 = input("Quelle liste voulez-vous comparer ? (ThèmeGénérique-ThèmeSpécifique)\n")
liste2 = input("Quelle autre liste voulez-vous comparer ? (ThèmeGénérique-ThèmeSpécifique)\n")

with open(chemin+liste1+extension, "r") as f:
    contenu = f.readlines()
    ens1={}
    ens1 = set(ens1)
    for ligne in contenu:
        mot = ligne.split("_")[0]
        ens1.add(mot)

with open(chemin+liste2+extension, "r") as f:
    contenu = f.readlines()
    ens2={}
    ens2 = set(ens2)
    for ligne in contenu:
        mot = ligne.split("_")[0]
        ens2.add(mot)

print("Le fichier ", liste1 ," comporte ",len(ens1)," mots.")
print("Le fichier ", liste2 ," comporte ",len(ens2)," mots.")
mots_communs = ens1.intersection(ens2)
print("Les fichiers ", liste1 ," et ", liste2 ," ont ", len(mots_communs)," mots en communs.")
```

Annexe 3 : Script python de comparaison de contenu du vocabulaire entre deux listes

```
cat sports_corpus.txt | awk -F'\t' '{print $5}' |
perl -ne "\$_=lc(\$_); s/<[>]+>//g; s/\&\#\d+\;/
/g; s/[\'\|]/ /g; s/[\;\:\.\,\[\]\(\)\d]*//g; s/
(le|la|les|ce|cette|cet|du|de|au|qui|que|et|se|à|en
|par|pour|sur|sous|on|un|une|des|dès) / /g; s/ +/
/g; print $_" >sports_corpus.normalise
```

Annexe 4 : Script bash de nettoyage du corpus sur le sport pour la clusterisation

```
""" WEKA """

print('Démarrage de la jvm et chargement des données dans weka')
jvm.start()
loader = Loader(classname="weka.core.converters.ArffLoader")
data = loader.load_file("train_fr.arff")
data.class_is_last()

print('Apprentissage du modèle') # Le choix des algo de weka s'effectue ici
classifier = Classifier(classname="weka.classifiers.trees.J48") # J48
# classifier = Classifier(classname="weka.classifiers.lazy.IBk") # KNN
# classifier = Classifier(classname="weka.classifiers.functions.MultilayerPerceptron") # Perceptron


classifier.build_classifier(data)

### Ici enregistrer le modèle dans le fichier dont le chemin est fourni par args.model
serialization.write(args.model, classifier)

print('Extinction de la jvm')
jvm.stop()
```

Annexe 5 : Parti du script python interagissant avec le logiciel Weka pour les calculs algorithmiques

RÉFÉRENCES BIBLIOGRAPHIQUES

[P. Beust, 2002] Un outil de coloriage de corpus pour la représentation de thèmes, JADT 2002 : 6es Journées internationales d'Analyse statistique des Données Textuelles, GREYC CNRS UMR 6072 & ModeSCoS – Université de Caen – 14032 Caen Cedex, par Pierre Beust – 2002

[B. Grau, N. Hernandez, 2003] Analyse thématique de textes pour permettre une lecture rapide ; Groupe LIR LIMSI – 2003

[M. Ballabriga, 2005] Comparaison de méthodes de classifications, par Adrien Haccoun – 2005

[L. Ralaivola, 2007] Algorithme des K-plus-proches-voisins, par L. Ralaivola – 06 février 2007

[F. Rastier, M. Valette, 2009] De la polysémie à la néosémie ; par François Rastier de CNRS et de l'INALCO et Mathieu Valette de l'Atilf – janvier 2009

[S. Ollinger, M. Valette, 2010] La créativité lexicale : des pratiques sociales aux textes ; par Sandrine Ollinger, Mathieu Valette – 9 novembre 2010

[M. Valette, 2009] Méthode pour la veille lexicale ; par Mathieu Valette – 4 décembre 2009

[S. Ollinger, 2011] –POMPAMO : Détection automatique de candidats à la néologie ; par Sandrine Ollinger de l'Atilf – 2011

[I. Falk, D. Bernhard, C. Gérard , R. Potier-Ferry, 2014] –Étiquetage morpho-syntaxique pour des mots nouveaux ; par Delphine Bernhard, Ingrid Falk, Christophe Gérard, Romain Potier-Ferry – 2014

[D. Bernhard, I. Falk, C. Gérard, juillet 2014] De la quenelle culinaire à la quenelle politique : identification de changements sémantiques à l'aide des Topic Models ; par Delphine Bernhard, Ingrid Falk et Christophe Gérard – juillet 2014

[D. Bernhard, I. Falk, C. Gérard, juin 2014] From Non Word to New Word : Automatically Identifying Neologisms in French Newspapers ; par Delphine Bernhard, Ingrid Falk et Christophe Gérard – 2 juin 2014

[D. Bernhard, I. Falk, C. Gérard, septembre 2014] Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique ; par Delphine Bernhard, Ingrid Falk et Christophe Gérard – 10 septembre 2014

[A. Guille, C. Favre, 2014] Une méthode pour la détection de thématiques populaires sur Twitter ; par Adrien Guille, Cécile Favre du Laboratoire ERIC de l'Université Lumière Lyon 2 – 2014
+ Guide de Coloriage Thématique ; par Lauren Bruneau de l'Université de Strasbourg du Laboratoire LiLPa – 2014

[L. Bruneau, 2014] Rapport de stage BRUNEAU Lauren 2e année Master Linguistique Informatique et Traduction LiLPa – Université de Strasbourg Tuteurs : Delphine BERNHARD , Ingrid FALK, Christophe GERARD Octobre 2014 – Décembre 2014

[L. Derczynski, S. Chester, K.S. Bøgh, 2015] Tune Your Brown Clustering, Please ; par Leon Derczynski de University de Sheffield, Sean Chester et KennethS.Bøgh de l'University d'Aarhus – janvier 2015

[www.programmez.com, Proxem, 2015] Proxem : une API de détection de la thématique d'un document ; par fredericmazue – 30 juin 2015

[K. Deturck, 2015] Détection de contenu utile depuis des sites d'actualité ; MASTER TRAITEMENT AUTOMATIQUE DES LANGUES ; Parcours : Traitement Automatique des Langues / Ingénierie Multilingue, par Kévin DETURCK – 2015

[D. Bernhard, L. Bruneau, I. Falk, C. Gérard, A.L. Rosio, 2016] Le Logoscope: observatoire des innovations lexicales en français contemporain ; par Christophe Gérard, Lauren Bruneau, Ingrid Falk, Delphine Bernhard, Anne-Lise Rosio – 2016

[C. Gérard, A. Grezka, M. Lorente, 2017] –

[C. Gérard, A. Grezka, L. Mercè, 2017] Pour une une documentation pluri-thématique des néologismes ; par Christophe du l'Université de Strasbourg, Aude Grezka du laboratoire LIPN-CNRS de l'Université Paris 13 et Lorente Mercè de l'Université de Pompeu Fabra

[Langue et numérique, 2017]

[E. Cartier, 2017] Néonaute : un moteur de recherche pour suivre l'implantation des néologismes à partir des collections du dépôt légal du web (BnF). Étude ciblées sur la néologie générale et terminologique, et la féminisation des noms de métiers, titres, grades et fonctions ; par Emmanuel Cartier de l'Université Paris 13 Sorbonne Paris Cité, Labex EFL et LIPN-RCLN CNRS UMR 7030 – 3 mai 2017

[E. Cartier, 2018] – Néoveille, système de repérage et de suivi des néologismes en sept langues ; par Emmanuel Cartier du laboratoire LIPN CNRS UMR 7030 de l'Université Paris 13 Sorbonne Paris Cité – 21 février 2018

[M. Boumghar, C. Christophe, J. Velcin, 2018] Utilisation de techniques de modélisation thématiques pour la détection de nouveauté dans des fluxs de données textuelles ; par Manel Boumghar et Clément Christophe de l'Entreprise EDF R&D et Julien Velcin du Laboratoire ERIC de l'Université Lumière Lyon2 – 2018