



Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Apports de la catégorisation automatique à la veille collaborative

MASTER

Traitement Automatique des Langues

Parcours :

Ingénierie Multilingue

par

Aurélie Jouannet

Directeur de mémoire :

Cyril Grouin

Encadrant :

Michel Bernardini

Remerciements

Je tiens à remercier toutes les personnes m'ayant permis de réaliser ce mémoire et d'effectuer ce stage dans les meilleures conditions, en particulier mon encadrant de stage Michel Bernardini ainsi que toute l'équipe Leonard, notamment Perrine Guy-Duché, pour toute l'aide qu'ils ont pu m'apporter dans la réalisation de ce projet.

Je remercie également mon directeur de mémoire, Cyril Grouin, pour ses précieux conseils et sa disponibilité tout au long de la rédaction de ce mémoire.

Enfin, je tiens à remercier l'équipe enseignante de la formation Ingénierie Multilingue de l'INALCO, pour les connaissances qu'ils m'ont transmises et que j'ai pu mettre en œuvre lors de la réalisation des différentes tâches qui m'ont été données durant ce stage.

Résumé

Ce travail aborde la question de l'utilisation de la fouille de texte dans le domaine de la veille, plus particulièrement la classification automatique de documents. Il s'agit de comparer différentes méthodes de catégorisation dépendant de problématiques linguistiques propres à chaque thématique traitée dans une plateforme de veille collaborative. Nous utiliserons pour cela les outils développés par Temis, qui nous permettent de développer différents outils linguistiques personnalisables selon les stratégies adoptées. Nous présenterons les résultats de trois méthodes de classification automatique, utilisant d'une part un modèle d'apprentissage et d'autre part des ressources linguistiques, à savoir l'utilisation d'une hiérarchie de concepts et d'un thésaurus.

Mots-clefs : fouille de texte, text mining, catégorisation automatique, thésaurus, hiérarchie de concepts, apprentissage automatique

Table des matières

Remerciements	
Résumé	
Introduction	1
I Etat de l'art	2
1.1 Les ressources linguistiques	2
1.2 L'apprentissage automatique	3
II Corpus et outils	5
2.1 Présentation de la plateforme	5
2.2 Les outils de fouille de texte	6
2.2.1 Les cartouches de connaissance	8
2.2.2 Les modèles d'apprentissage.....	9
III Classification par apprentissage supervisé.....	10
3.1 Présentation des catégories.....	10
3.2 Modification du secteur High Tech	13
3.3 Corpus et paramètres	14
IV Classification par hiérarchie de concepts	18
4.1 Description du projet Digital Working	18
4.2 Les concepts et leur hiérarchie	18
4.2.1 Français	18
4.2.2 Anglais.....	20
4.3 Constitution des corpus	21
V Classification basée sur un thésaurus	22
5.1 Historique du projet Macro-économie	22
5.1.1 Présentation	22
5.1.2 Pistes d'amélioration	23
5.2 Réorganisation du thésaurus.....	23
5.2.1 Structure du nouveau thésaurus.....	23
5.2.2 Réorganisation des contextes	25
5.3 Constitution des corpus	26
5.4 Paramétrages	26
5.4.1 Paramètres STF	27
5.4.2 Paramètres de gestion des contextes et de seuil des catégories.....	28
VI Evaluations.....	30
6.1 Modèle d'apprentissage automatique.....	30
6.1.1 Français	30

6.1.2 Anglais.....	32
6.2 Approche avec hiérarchie de concepts	33
6.3 Approche avec thésaurus	35
VII Discussion.....	37
Conclusion et perspectives	38
Bibliographie.....	39
Annexe	41

Introduction

Le domaine de la veille a vu la nécessité de s'adapter à la production croissante d'informations provenant du web, présentes dans des sources et des formats divers et variés. Cette évolution constante de la manière de produire de l'information a donc soulevé la question du traitement de toutes ces données hétérogènes. Les entreprises, dans leur quête de recherche de l'information pertinente pour répondre à diverses attentes, se retrouvent alors confrontées à des choix concernant la façon de filtrer ce volume de données.

Ainsi, après la notion de Big Data qui s'est popularisée ces dernières années, d'autres commencent à la remplacer, avec notamment l'apparition de la *smart data*. Depuis un peu plus de deux ans, cette nouvelle notion suscite de plus en plus l'intérêt des entreprises, qui s'interrogent sur la finalité de leur récolte de gros volumes de données. Il s'agit alors, comme le détaille une étude de Spark Digital¹, de donner la priorité à la qualité plutôt qu'à la quantité de l'information, cela impliquant de structurer l'information afin qu'elle soit facilement exploitable.

De fait, de plus en plus de sociétés se sont intéressées à cette problématique rencontrée dans des secteurs aussi bien publics que privés, formant ainsi un véritable marché de la fouille de données ou *data mining*. Ce domaine reposant sur des disciplines variées, allant des statistiques à l'intelligence artificielle, inclut un sous domaine spécialisé sur le traitement de l'information textuelle : la fouille de texte, ou *text mining*. Contrairement à la fouille de données s'exerçant sur des données structurées (bases de données), la fouille de texte se concentre sur l'exploitation des données textuelles qu'elles soient structurées (bases de données), semi-structurées (documents XML) ou non structurées (sites web). Cette particularité de la fouille de texte à pouvoir exploiter de l'information non structurée en fait un atout de choix pour les entreprises, les informations stratégiques ne se limitant plus aux documents internes ou produits par des sources officielles, mais se trouvant dorénavant sur des sites web, des blogs, sur les réseaux sociaux, etc.

Parmi les différents domaines applicables à la fouille de texte pour les entreprises, la classification automatique de documents en est un des plus utilisés. En effet, cette capacité à faciliter la recherche d'information en regroupant automatiquement des documents dans des thématiques particulières permet un gain de temps considérable quant à l'exploitation des informations, au vu de la diversité et de la quantité des données que les entreprises ont à traiter. Notre travail se situe ainsi dans une problématique d'indexation de documents liant la gestion de volume de données important (*big data*) et l'application de traitements en vue d'obtenir un ensemble plus restreint de données pertinentes (*smart data*).

Pour cela, nous nous appuyons sur le domaine de l'intelligence économique qui, de par ses activités en relation avec la collecte, le traitement et la diffusion de l'information, en fait un sujet de choix pour l'utilisation de la classification automatique. Afin de voir de quelle façon celle-ci permet de faciliter l'exploitation de documents provenant de sources diverses, nous nous appuyons sur une plateforme de veille collaborative de BNP Paribas.

¹ Spark Digital, *Turn your big data into smart data*, Août 2014 <<http://www.sparkdigital.co.nz/assets/fwd-live/5792-SD-FWD-Article-BigData-final.pdf>>.

Celle-ci devant gérer des documents en volume conséquent et sous différents formats, elle entre justement dans cette problématique de gestion de l'information nécessitant l'utilisation de la fouille de texte.

Nous commencerons par présenter un état de l'art des méthodes utilisées à ce jour pour la classification automatique de documents, avant de présenter le type de documents que nous devons prendre en compte pour le développement des outils de fouille de texte. Ensuite, nous détaillerons différentes méthodes d'indexation possibles grâce à ces outils et comparerons les avantages et les inconvénients de chacune, avant de conclure et de présenter des pistes d'améliorations.

I Etat de l'art

Il existe plusieurs méthodes pour catégoriser automatiquement des documents, notamment par l'utilisation de ressources linguistiques et de modèles d'apprentissage automatique. Les ressources linguistiques comprennent notamment les taxonomies, thésaurus et autres ontologies.

1.1 Les ressources linguistiques

Les taxonomies, dans leur organisation en termes présentés de manière hiérarchique, permettent de faciliter une recherche en l'élargissant ou au contraire, en la resserrant sur une thématique particulière. Une structure de ce type, nommée hiérarchie de concepts, est ainsi utilisée pour l'indexation automatique de textes par Englmeier *et al.* (2001), qui la définissent comme « une arborescence composée de concepts ou entrées, chaque entrée correspondant à un ensemble de termes ». Un document est ainsi catégorisé dans la classe représentée par la hiérarchie de concepts dès lors qu'il contient des termes présents dans celle-ci. Une méthode de vote a été appliquée pour déterminer l'importance d'un concept dans la hiérarchie par rapport à un texte, ce qui permet d'identifier les concepts les plus représentatifs de la hiérarchie. Ainsi, en comparant les termes les plus pertinents dans la hiérarchie de concepts aux termes les plus représentatifs dans le document, on obtient le taux de présence d'un concept de la hiérarchie dans un document. Cette approche permet d'obtenir une meilleure précision en supprimant les termes les plus ambigus, créant ainsi une sorte de vocabulaire contrôlé aidant à l'indexation et la recherche d'information pour des utilisateurs souhaitant s'informer sur un sujet dans un contexte bien défini (ex : des informations relatives à l'économie d'un pays en particulier).

Les thésaurus, quant à eux, sont aussi organisés de manière hiérarchique mais intègrent d'autres attributs permettant de décrire des thématiques : les concepts les plus hauts dans la hiérarchie sont appelés « broader terms » et contiennent des termes fils, appelés « narrower terms ». Une distinction est faite entre un terme et un concept : le premier désigne une unité lexicale simple, telle qu'elle apparaît dans un texte, tandis que le second regroupe plusieurs termes, notamment un terme préférentiel et des synonymes. Les concepts du thésaurus peuvent ainsi être liés par des relations de synonymie ou par d'autres relations regroupées sous le nom de « related terms ». Kervers (2009) a utilisé un thésaurus afin de faciliter la classification des documents aux catégories présentes dans celui-ci. Pour cela, il utilise une méthode se basant sur l'assignement de mots clefs, constituée de transducteurs capables de dériver des mots à partir des concepts présents dans le thésaurus. Ces transducteurs, sous forme d'expressions régulières, sont pondérés afin de faire remonter la catégorie du document d'après l'addition des poids des termes pour une catégorie donnée. Un seuil a été défini afin de ne conserver que les candidats

termes les plus pertinents, de même qu'un seuil prenant en compte les n premières catégories correspondant aux meilleurs scores, dans le but de restreindre le nombre de catégories possibles pour un même document. Parmi plusieurs expériences réalisées sur un corpus de documents législatifs en langue française, les plus performantes ont été celles qui utilisent la pondération des scores attribués aux concepts des différentes catégories. Les catégories présentant de moins bons résultats ont été pénalisées par la présence de termes trop génériques et l'absence de synonymes, donc à une hiérarchie moins travaillée que pour les catégories donnant de bons résultats.

Enfin, les ontologies ajoutent un degré supérieur de description par rapport aux ressources précédentes, car contrairement au nombre limité de relations que peuvent prendre en compte les thésaurus et taxonomies, les ontologies sont utilisées pour décrire les objets du monde à travers un nombre illimité de types, relations et autres propriétés nécessaires à la représentation d'un domaine. L'interaction entre les nombreuses propriétés et relations présentes dans une ontologie permet d'ajouter de la sémantique dans la tâche de recherche d'information. Un langage de description particulier (Resource Description Framework ou RDF) permet ainsi de formaliser ces propriétés et créer des processus de raisonnement ou inférences. Cette particularité se retrouve dans les travaux de Hernandez et Mothe (2004), où l'ontologie joue à la fois un rôle dans l'indexation des documents et dans l'exploration des informations contenues dans leur corpus grâce à une interface graphique. Une des difficultés rencontrées dans l'utilisation des ontologies pour la recherche d'information, soulevée par Slimani *et al.* (2007), repose sur les calculs utilisés pour mesurer la similarité sémantique entre les différents concepts de la hiérarchie. Or, les auteurs mentionnent que ces calculs reposant sur la notion de « voisinage » peuvent donner des poids inadéquats selon la méthode utilisée (calcul basé sur la distance entre les nœuds ou bien sur la hiérarchie), en considérant par exemple que les liens sémantiques entre deux nœuds possèdent le même poids, quelle que soit la distance qui les sépare dans la hiérarchie.

La complexité de la gestion des ontologies, de par leur formalisme très éloigné du langage naturel, ne nous semble pas adaptée pour les thématiques que nous aurons à traiter ici. En effet, celles-ci étant assez restreintes et prenant en compte un vocabulaire très précis et relativement limité, l'utilisation de taxonomies et de thésaurus nous semble plus pertinente.

1.2 L'apprentissage automatique

Concernant les modèles d'apprentissage, on distingue généralement les méthodes supervisées des méthodes non supervisées. Ces dernières impliquent que les classes auxquelles devront appartenir un ensemble de documents ne sont pas connues d'avance, contrairement aux méthodes supervisées, qui nous intéressent tout particulièrement dans cette étude. Parmi plusieurs méthodes ayant fait leur preuve dans la catégorisation de données textuelles, on peut citer les algorithmes bayésiens naïfs et les k plus proches voisins, ou encore les machines à vecteur de support ou SVM.

Les algorithmes bayésiens naïfs considèrent que tous les mots d'un document sont indépendants les uns des autres, s'appuyant ainsi sur une représentation simplifiée des documents, par exemple des sacs de mots, où la position des termes n'est pas prise en compte par le modèle. Le vocabulaire représentatif d'un document est alors formé par le nombre d'occurrences des mots qu'il contient. A partir de cette représentation, plusieurs approches peuvent être testées pour améliorer la catégorisation d'un document dans une classe prédéfinie, en limitant par exemple le vocabulaire à un nombre restreint de

descripteurs afin de diminuer le bruit causé par des termes trop généraux [Sureshkumar *et al.*, 2013].

L'algorithme des k plus proches voisins (ou KNN) consiste à déterminer la classe d'une variable aléatoire x par rapport à un ensemble des k plus proches voisins parmi un échantillon de données [Altman, 1992]. Karteeka Pavan *et al.* (2011) ont utilisé une version améliorée de cet algorithme dans leur tâche de classification automatique. Celle-ci permet de générer automatiquement un nombre de catégories optimal à partir d'un ensemble de documents donné en entrée. Une des difficultés de la classification automatique, mis à part le fait de classer correctement un ensemble de documents dans des catégories, est en effet de trouver le nombre idéal de catégories à classer. Leur algorithme KNN permet justement d'effectuer cette tâche avec plus ou moins de facilité selon les critères d'évaluation pris en compte.

L'utilisation d'espaces vectoriels pour représenter les documents est une autre méthode efficace pour la catégorisation automatique : elle consiste à représenter chaque document par un point dont les coordonnées, sur chaque axe, correspondent au poids d'un terme dans le document. Celui-ci est calculé d'après sa fréquence d'apparition. A partir de ces vecteurs, une fonction est appliquée afin de comparer le score des documents et ainsi savoir si le document appartient à la classe en question [Beney, 2008]. Cette méthode a notamment été utilisée par Roche *et al.* (2010), confrontés à une problématique de documents bruités à cause de leur océrisation. Ils ont ainsi comparé les algorithmes KNN, bayésien naïf et SVM pour déterminer l'algorithme le plus adapté pour ce type de documents. Ils utilisent le TF/IDF (*Term Frequency/Inverse Document Frequency*) comme méthode de pondération des termes, qui permet de donner un poids plus important aux termes les plus représentatifs d'un document en se basant sur leur fréquence d'apparition [Spärck Jones, 1972]. Au-delà des paramètres statistiques pris en compte dans leur méthode, les auteurs se sont aussi servis de certaines classes grammaticales pour ajouter de l'information sémantique dans leur modèle de classification. Les meilleurs résultats ont été obtenus avec l'algorithme SVM sur un corpus de textes journalistiques, puisque très riches concernant le nombre de mots. Les moins bons résultats ont été quant à eux dus à la présence des erreurs dans le corpus océrisé.

Nous nous inspirerons ainsi de cette méthode utilisant des vecteurs et prenant en compte les classes grammaticales, notre contexte de catégorisation d'articles de presse partageant la même problématique.

Tout au long de ce travail, nous étudierons les particularités linguistiques des documents que nous aurons à traiter afin d'en déduire les méthodes de classification les plus appropriées, au vu des différentes pistes évoquées par ces travaux antérieurs.

II Corpus et outils

2.1 Présentation de la plateforme

Les méthodes présentées ici s'appuieront sur la plateforme Leonard (*New Assistant for Resources and Documentation*) de BNP Paribas. Il s'agit d'un outil de veille collaborative initié aux Etudes Economiques en 2004 par Michel Bernardini, qui met à la disposition de ses utilisateurs des documents provenant de plusieurs sources et leur permet de partager des documents.

Tout d'abord, Leonard contient des articles de la presse quotidienne et hebdomadaire au format PDF, numérisés par un prestataire, Mediacompil. Ensuite, un outil de veille, KB Crawl, est utilisé pour sélectionner des sites web selon différents critères (présence de termes particuliers ou de noms d'entreprises, par exemple). Les sites ainsi crawlés sont présentés de plusieurs façons dans la plateforme : Unes, actualité internationale, flux RSS et répartition par secteurs (banque, industrie, automobile, etc.).

Ces différentes sources sont réparties en plusieurs onglets, dont un qui constitue l'élément central de la notion collaborative de Leonard : Wikiléo. Celui-ci permet à tous les collaborateurs inscrits d'ajouter des documents et ainsi de les partager avec les autres utilisateurs de la plateforme. Chaque collaborateur a aussi la possibilité de s'inscrire à des communautés (ex : Immobilier, Banque de détail, Economie Numérique, etc.) et choisir de recevoir des alertes par e-mail lorsqu'un nouveau document a été publié dans sa communauté.

Une cinquantaine de documents est ainsi ajoutée par jour dans cette section collaborative de la plateforme, constituée des articles de la presse quotidienne publiée dans des communautés ainsi que d'articles provenant du web en effectuant une veille manuelle sur des sujets intéressant les communautés (banque, assurance, économie numérique, big data, etc.).

Tous ces documents sont indexés par un moteur de recherche statistique, Polyspot, qui convertit au préalable les documents en HTML, et fait le lien avec les outils de fouille de texte que nous abordons ci-après. L'interaction de ces différents outils entre eux peut être schématisée de la manière suivante :

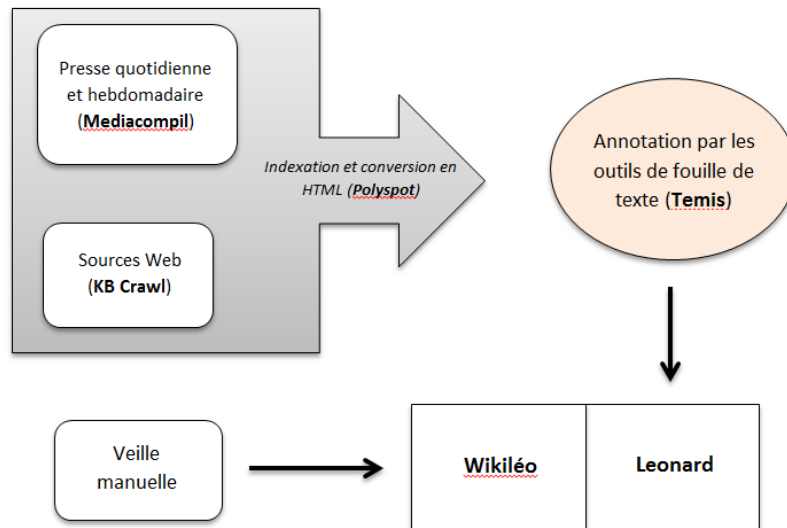


Fig.1 Processus d'intégration des documents dans la plateforme

Une base de données interne non liée à la plateforme, alimentée par le centre de documentation de BNP Paribas, sera aussi utilisée pour constituer nos corpus de documents.

Ce sont ces différentes sources qui alimenteront les corpus que nous utiliserons pour le développement des méthodes de fouille de texte, ces derniers étant présentés plus en détails par la suite car correspondent à des problématiques linguistiques propres à chaque thématique.

2.2 Les outils de fouille de texte

Tous ces documents intégrés dans la plateforme nécessitent d'être indexés afin d'en faciliter la recherche. Les utilisateurs peuvent ainsi filtrer leur recherche par source, langue, date mais aussi par catégorie, comme illustré ci-dessous.

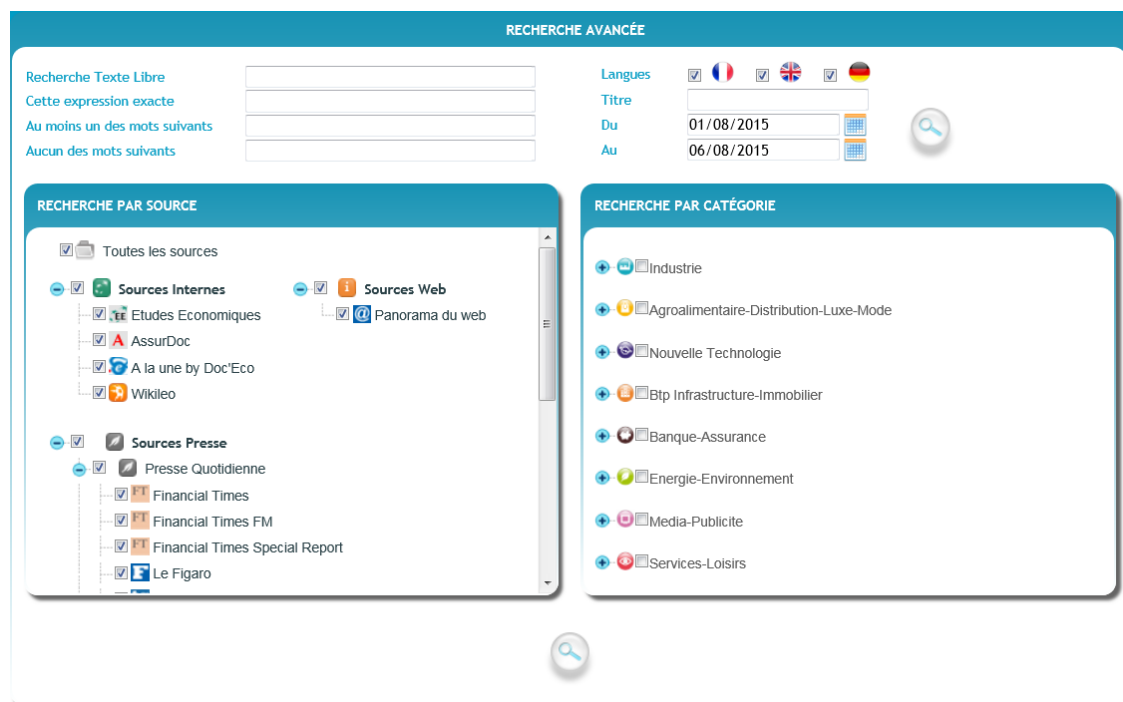


Fig.2 Onglet Leo Search

La recherche rapide, disponible à tout moment quel que soit l'onglet où l'utilisateur se trouve, permet aussi d'effectuer des recherches désambiguïsées grâce aux entités nommées. Cette tâche d'extraction d'information, consistant à reconnaître des objets textuels comme appartenant à des classes spécifiques (noms de personnes, de lieux, d'organisation, etc.), est en effet capitale pour faciliter la recherche d'information des collaborateurs lorsqu'ils souhaitent s'informer sur une société, une personne ou tout autre entité, surtout si elle est ambiguë sémantiquement. Ainsi, si un utilisateur recherche des articles parlant de la société Orange, une icône « société » lui sera proposée afin de filtrer les résultats uniquement sur les articles citant l'entreprise et ne proposera donc aucun autre homonyme.

Toutes ces fonctionnalités impliquant la fouille de texte sont effectuées grâce aux outils développés par Temis. Leur suite logicielle Luxid se base sur le principe des « cartouches de connaissances », permettant différentes stratégies pour mettre en place des applications selon les besoins des entreprises. Selon le logiciel utilisé, il est possible de gérer des tâches plus ou moins complexes, allant de la gestion de ressources linguistiques au développement de modèles d'apprentissage automatique.

Quelle que soit la méthode utilisée, le processus interne permettant de traiter les documents repose sur les étapes standards du traitement du langage naturel, à savoir :

- l'identification de la langue du document,
- la tokenisation,
- la segmentation en phrases,
- l'analyse morphologique,
- l'étiquetage morphosyntaxique,
- la désambiguïsation.

La plateforme Leonard utilise deux sortes d'indexation : l'une se fait via des cartouches de connaissance, et l'autre avec des modèles d'apprentissage.

2.2.1 Les cartouches de connaissance

Les cartouches ou « composants » de connaissance ont été créés par Temis pour permettre le développement d'outils d'extraction d'information fonctionnant sous forme de modules. Chaque module jouant un rôle particulier (conversion de document, calcul de fréquences, gestion de contextes, etc.), il est alors possible de combiner l'utilisation de ces différents modules pour créer des outils d'extractions d'information adaptables selon les besoins.

Ainsi, dans notre cas, l'utilisation des cartouches de connaissance est privilégiée pour la recherche par thème sur la plateforme Leonard (personnes, organisation, pays, etc.), comme illustré ci-dessous :



Fig.3 Illustration des thèmes avec mots clefs présents dans la presse quotidienne

En arrivant sur l'onglet « Panorama presse », les utilisateurs ont ainsi un aperçu des thématiques les plus abordées dans la presse du jour, et ont la possibilité d'accéder directement aux articles concernés en cliquant sur les mots clefs proposés par l'indexation. La taille des entités varie en fonction de leur nombre d'occurrences dans les articles, permettant une meilleure visibilité des informations les plus importantes du jour.

2.2.2 Les modèles d'apprentissage

Le second type d'indexation s'effectue via un modèle d'apprentissage automatique, utilisé pour indexer des secteurs (BTP infrastructure, banque-assurance, nouvelles technologies etc.). Celle-ci permet aux utilisateurs de Leonard de trouver des articles parmi huit secteurs contenant chacun plusieurs sous catégories, comme illustré ci-dessous :

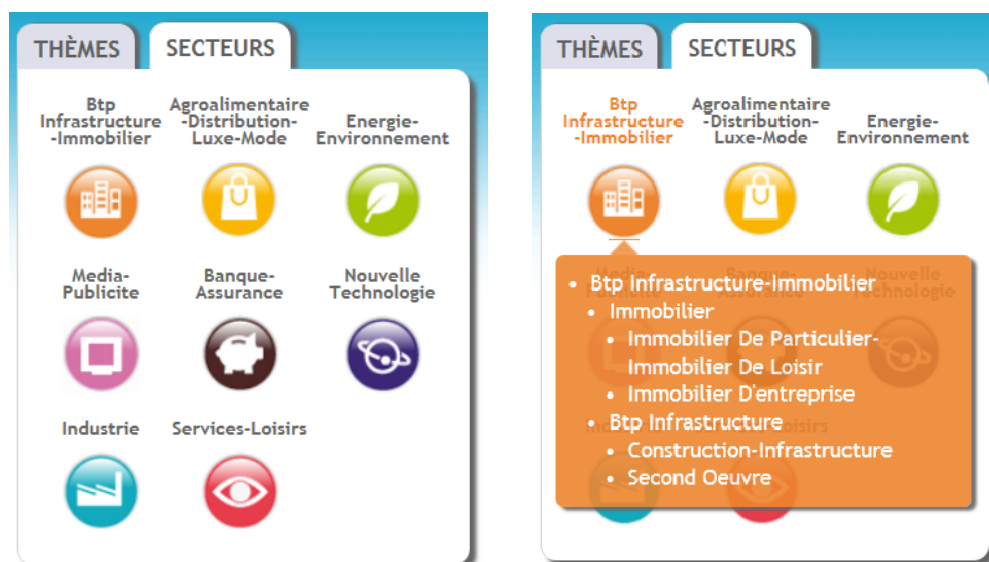


Fig.4 Illustration des secteurs avec survol des catégories

Cette indexation par secteurs est accessible via le panorama presse de même que par l'onglet de recherche, comme présenté plus haut (Fig.2).

Afin d'illustrer les différentes stratégies de développement possibles grâce à ces outils, nous comparerons les résultats de trois méthodes, utilisant un modèle d'apprentissage automatique et des ressources linguistiques.

III Classification par apprentissage supervisé

L'avantage de l'apprentissage automatique pour la classification de documents repose sur sa capacité à trouver les descripteurs les plus représentatifs d'un document et de s'appuyer sur ces derniers pour le classer en les comparant aux descripteurs d'une classe définie au préalable. Les méthodes de classification automatique supervisées reposent ainsi sur trois étapes : la définition de classes, l'apprentissage de celles-ci en utilisant un algorithme associé à un corpus d'apprentissage, puis l'évaluation des performances grâce à un corpus de test.

Les huit secteurs à prendre en compte dans la plateforme possédant un nombre de termes représentatifs non exhaustif, l'utilisation de ressources linguistiques telles qu'une taxonomie ou un thésaurus n'était pas adaptée. Or, les thématiques étant assez hétérogènes les unes des autres de par leur vocabulaire, l'utilisation d'un modèle d'apprentissage automatique s'est révélée être la solution la plus pertinente.

3.1 Présentation des catégories

Afin d'étudier le fonctionnement de cette méthode, nous nous basons sur huit catégories déjà mises en place dans la plateforme lors de précédents travaux [Ma, 2014], détaillées ci-dessous :

Banking-Insurance

- Banking
 - Banking Regulations
 - Investment banking
 - Islamic banking
 - Means of payment
 - Portfolio management
 - Retail banking
- Insurance

La catégorie Banque-Assurance, étant donné le secteur d'activité qu'elle prend en compte, est l'une des plus importantes pour les collaborateurs et utilisateurs de la plateforme. La sous-catégorie Banque est donc divisée en plusieurs thèmes afin d'offrir une recherche détaillée sur les sujets relatifs à la banque, incluant la régulation bancaire, les types de banques (banque d'investissement, banque de détail) et des services (gestion de portefeuille, moyens de paiement). Le thème de la Banque Islamique est aussi présent, car est devenu un sujet de plus en plus d'actualité ces dernières années.

Building civil engineering-Real estate

- Building civil engineering
 - Building materials
 - Building-Substructure
 - Construction terms
- Real estate
 - Corporate real estate
 - Individual real estate-Leisure real estate

La catégorie BTP-Infrastructure-Immobilier permet de différencier les articles en fonction du type d'immobilier abordé (immobilier d'entreprise ou de particulier), de même pour le bâtiment, séparant les infrastructures au sens général des infrastructures de seconde œuvre.

Energy-Environment

- Energy
 - Commodities
 - Gas-electricity-coal
 - Oil
 - Renewable energy
- Environment-Waste management
 - Environment
 - Waste management

La catégorie Energie-Environnement indexe les articles traitant des matières premières, avec des sous-catégories à part entière pour le gaz / l'électricité / le charbon, le pétrole et les énergies renouvelables, ces thématiques étant très présentes quotidiennement du fait de l'évolution de leur cours sur les marchés financiers.

La sous-catégorie Environnement-Gestion des déchets, quant à elle, ne nécessite pas d'être détaillée autant que pour la catégorie Energie.

Media-Advertising

- Advertising
- Media

La catégorie Média-Publicité a trait aux articles parlant des industries culturelles, à savoir la télévision, la presse, les médias numériques (plateformes de vidéos comme YouTube, Dailymotion), etc.

Food industry-Retailing-Luxury-Fashion

- Food industry
- Retailing-Luxury-Fashion and textile
 - Clothing-Accessory
 - Cosmetics
 - Luxury products
- Retailing
 - E-commerce
 - Retail industry
 - Specialty retail store

Ce secteur permet d'accéder aux documents traitant d'agroalimentaire, de la grande distribution et du e-commerce, des produits de luxe et de la mode.

Services-Leisure activities

- Leisure activities
 - Catering
 - Sport-Money-Games
 - Money games-Toys
 - Sport
 - Tourism
- Services
 - Consulting-Engineering-Outsourcing
 - Health-Welfare
 - Security-Cleanliness
 - Training-Temporary work

Dans ce secteur, la sous-catégorie Loisirs prend en compte la restauration, les sports, les jouets, les jeux d'argent et le tourisme, tandis que Services se concentre sur les métiers liés au consulting, à la sécurité et la propreté, à la santé et aussi plus généralement aux métiers temporaires, se rapprochant davantage d'une problématique de ressources humaines.

Industry

- Aerospace-Aeronautics-Airline industry
- Capital good
- Metal working industry-Steel industry
- Road and rail and goods transport
- Shipbuilding and sea transport
- Wood paper-Packaging-Furniture
- Automotive
 - Car manufacturer-Dealership
 - Equipment manufacturer and repairing
 - Financing-Car rental
- Chemistry-Pharmacy
 - Chemistry
 - Pharmacy

Le secteur Industrie contient le plus grand nombre de sous-catégories, à savoir tout ce qui concerne l'aéronautique, le transport ferroviaire et maritime, l'automobile, la chimie-pharmacie, etc.

High Tech

- Connectivity-Mobility
- Material-Software-Equipment
 - Material
 - Software
 - Equipment

Enfin, la catégorie Nouvelles Technologies traite des télécommunications et de l'informatique, cette dernière faisant la distinction entre le matériel informatique, l'équipement et le logiciel.

Un même document peut être classé dans deux catégories différentes au maximum s'il aborde une thématique recouvrant plusieurs secteurs. Ainsi, un article traitant d'une réglementation sur les moyens de paiements sera classé à la fois dans Régulation bancaire et Moyens de paiements. Dans le cas où un document traite de plus de deux secteurs, les deux catégories obtenant les meilleurs scores seront retenues. C'est ce que nous verrons par la suite lorsque nous chercherons à améliorer le secteur Nouvelles Technologies, qui devra répondre à de nouvelles attentes.

3.2 Modification du secteur High Tech

Au vu des documents indexés par le secteur High Tech, la modification de ce secteur s'est avérée nécessaire. En effet, il abordait des sujets trop techniques, créant un décalage avec les autres catégories. La distinction entre matériel, équipement et logiciel n'était pas non plus très claire. De plus, le nom de cette catégorie n'était pas adapté à ce qu'elle indexait, ses sous-catégories n'étant pas spécifiques aux nouvelles technologies. Nous avons donc décidé de la modifier pour qu'elle prenne en compte de nouvelles thématiques prenant de plus en plus d'importance dans l'actualité, à savoir :

- les PC, tablettes et smartphones,
- la cybersécurité,
- les réseaux sociaux,
- le big data,
- le Cloud,
- l'impression 3D.

Nous avons aussi gardé et renommé la sous-catégorie Connectivité-Mobilité en Télécoms, afin qu'elle soit plus explicite. Celle-ci indexe en effet des articles non pas techniques comme le laissait entendre son nom, mais des documents sur les opérateurs de télécommunication et leurs offres (Orange, Numéricable, Free, etc.).

D'après ces nouvelles thématiques, nous avons donc distingué trois sous-catégories : les thématiques relevant de l'informatique, celles traitant du numérique et enfin les télécommunications. Le nom du secteur a donc été modifié en conséquence, devenant IT-Digital-Telecoms, et dont les sous-catégories ont été classées de la manière suivante :

- IT
 - Cloud
 - PC-Tablet-Smartphone
 - 3D printing
- Digital
 - Big Data
 - Cybersecurity
 - Social Media
- Telecoms

Une fois cette nouvelle structure mise en place, il fallait réfléchir à la constitution de nouveaux corpus pour ces catégories, en étant vigilants quant à de possibles chevauchements entre certaines catégories risquant de partager des termes en commun.

3.3 Corpus et paramètres

La restructuration du plan a ainsi nécessité l'intégration de nouveaux documents dans le modèle, ainsi que la modification des corpus existants pour la catégorie Telecoms, contenant le corpus de l'ancienne catégorie Connectivity-Mobility. La constitution des corpus a été effectuée grâce à une veille en utilisant l'outil KB Crawl ainsi que par la collecte de documents présents sur la plateforme. Les documents ont ensuite été regroupés en un fichier TMX (format propriétaire XML de Temis) qui sera donné en entrée au logiciel Category Workbench, qui nous permet de développer les modèles d'apprentissage. Le fichier, créé via un programme Perl parcourant des répertoires et se basant sur leur nom pour donner aux documents leur catégorie (cf annexe), possède l'allure suivante :

```
<?xml version="1.0" encoding="UTF-8"?>
<tm xmlns:dc="http://purl.org/dc/elements/1.1/">
  <doc id="1">
    <dc:title>nom_du_document</dc:title>
    <categories><c>IT-DIGITAL-TELECOMS/IT/CLOUD</c></categories>
    <text><file format="html" path="chemin_du_fichier"/></text>
  </doc>
  [...]
</tm>
```

Fig.5 Aperçu de la structure d'un fichier TMX

Le fichier indique alors le chemin de chaque document et leur format ainsi que leur catégorie, balise qui sera utilisée par le logiciel pour créer la hiérarchie des différentes classes.

La constitution des corpus étant un procédé chronophage, nous avons commencé par tester la catégorisation de ce secteur avec une cinquantaine de documents pour chaque nouvelle classe. Les documents de l'ancienne catégorie Connectivity-Mobility ont été réutilisés pour alimenter les classes Telecoms et PC-Tablet-Smartphone, ce qui explique un nombre plus conséquent de documents pour celles-ci par rapport aux autres classes. En effet, le nombre de documents pour les secteurs existants est assez conséquent car ils ont été constitués automatiquement en extrayant les documents d'une base de données interne de manière automatique. Or, pour ces nouvelles catégories, les documents devaient être sélectionnés avec précision afin de minimiser les chevauchements de vocabulaire, à savoir les documents traitant à la fois de Big Data et de cloud, de réseaux sociaux et de Big Data, de cybersécurité et de cloud etc. Il a donc fallu procéder manuellement en laissant de côté les documents communs à plusieurs catégories, ce qui explique que le volume soit plus restreint pour les catégories autres que Telecoms et PC-Tablet-Smartphone.

Voici la répartition de nos corpus français et anglais :

	Français	Anglais
IT	292	169
3D PRINTING	47	52
CLOUD	53	53
PC-TABLET-SMARTPHONE	192	64
DIGITAL	196	147
BIG DATA	64	49
CYBERSECURITY	51	48
SOCIAL MEDIA	81	50
TELECOMS	1070	1870

Tableau 1 Répartition des corpus du secteur IT-Numérique-Télécoms

Ces nouveaux documents s'ajoutent ainsi au reste du corpus prenant en compte les autres secteurs, ce qui nous donne un total de 26748 documents pour le français et 10132 pour l'anglais. Cet écart s'explique par une plus forte présence d'articles en français dans la plateforme et la base de données par rapport à l'anglais.

Une fois les corpus constitués, il fallait lancer l'apprentissage du modèle. Pour une raison de cohérence, nous avons repris les paramètres utilisés par le modèle existant afin de ne pas altérer les résultats des autres secteurs. Le modèle est basé sur une méthode scalaire, ayant la particularité de comparer les catégories entre elles sans prendre en compte leur niveau hiérarchique. L'algorithme s'appuie alors uniquement sur la comparaison des différents descripteurs de chaque catégorie. Le paramétrage du modèle se fait ainsi en attribuant des seuils au niveau des termes, des documents et des catégories, comme nous le verrons ci-après.

- Nombre minimum de documents dans une catégorie

Un premier paramètre permet de choisir le nombre minimum de documents qu'une catégorie doit contenir pour être prise en compte lors de l'apprentissage. Cela permet d'ignorer certaines catégories si nous désirons tester le modèle sans les prendre en compte.

- Fréquence minimum des descripteurs dans le corpus

Ce paramètre permet de décider le nombre minimum d'apparition d'un descripteur pour qu'il soit pris en compte dans le modèle. Ainsi, si le terme « cloud » n'apparaît que deux fois dans tout le corpus et que nous avons fixé le seuil minimal à 3, il sera ignoré par le modèle.

- Fréquence maximum d'un descripteur dans un document

Nous pouvons définir le seuil maximal qu'un descripteur ne doit pas dépasser pour être pris en compte. Dans notre cas, comme nous voulons prendre en compte tous les descripteurs d'un document, nous avons fixé ce nombre à 10000 étant donné qu'il est peu probable qu'un terme puisse apparaître autant de fois dans un document.

- Fréquence minimum d'un descripteur dans un document

A l'inverse, nous pouvons aussi définir le seuil minimum qu'un descripteur doit avoir pour être pris en compte. Il est ici fixé à 2 afin de prendre en compte la longueur variable des documents du corpus, allant de 400 à plus de mille mots selon les articles.

- Fréquence minimale des descripteurs dans une catégorie

Avec ce seuil, si nous le fixons à 2 et qu'un terme n'apparaît qu'une seule fois dans chaque document d'une catégorie, il ne sera pas pris en compte dans le modèle.

- Nombre maximum de descripteurs

Ce dernier paramètre statistique permet de limiter le nombre global de descripteurs. De même que pour la fréquence maximale des descripteurs pour un document, nous avons augmenté ce nombre de façon à prendre en compte tous les descripteurs.

- Descripteurs utilisés pour l'apprentissage

Ce paramètre supplémentaire permet de prendre en compte uniquement certaines catégories morphosyntaxiques. Les verbes, adjectifs et autres adverbes n'étant pas pertinents pour identifier la thématique d'un document, nous avons donc choisi de sélectionner les noms communs, les noms propres et les groupes nominaux.

Voici un résumé des paramètres utilisés pour notre modèle :

Nombre minimal de documents dans une catégorie	30
Fréquence minimum des descripteurs dans le corpus	3
Fréquence maximum d'un descripteur dans un document	10000
Fréquence minimum d'un descripteur dans un document	2
Fréquence minimum des descripteurs dans une catégorie	5
Nombre maximum de descripteurs	1000000
Descripteurs utilisés	Noms communs, Groupes nominaux, Noms propres

Tableau 2 Paramètres du modèle d'apprentissage

Une fois l'apprentissage lancé sur 90% de notre corpus, nous pouvons vérifier si les descripteurs sont pertinents pour chaque catégorie. Par exemple, en observant les meilleurs descripteurs retenus pour la catégorie Cybersecurity, nous pouvons voir qu'ils sont bien restreints au vocabulaire de la cybersécurité, et qu'ils contiennent peu de termes pouvant prêter à confusion et qui seraient donc susceptibles de donner de moins bons résultats lors de la phase de test :

/IT-DIGITAL-TELECOMS/DIGITAL/CYBERSECURITY

Full name	Name
/COMMON-NOUN/cyber	cyber
/COMMON-NOUN/cyber-attaque	cyber-attaque
/COMMON-NOUN/cybercrime	cybercrime
/COMMON-NOUN/cybercriminel	cybercriminel
/COMMON-NOUN/cyberdéfense	cyberdéfense
/COMMON-NOUN/cyberespace	cyberespace
/COMMON-NOUN/cyberguerre	cyberguerre
/COMMON-NOUN/cybermenace	cybermenace
/COMMON-NOUN/malware	malware
/NP/Guillaume poupard	Guillaume poupard
/NP/pirate informatique	pirate informatique
/PROPER-NAME/ANSSI	ANSSI
/PROPER-NAME/Anonymous	Anonymous
/PROPER-NAME/Anssi	Anssi
/PROPER-NAME/CYBERSECU...	CYBERSECURITE
/PROPER-NAME/Cybersécurité	Cybersécurité
/PROPER-NAME/Dark Web	Dark Web
/COMMON-NOUN/cyberattaque	cyberattaque
/COMMON-NOUN/cybercrimin...	cybercriminalité
/PROPER-NAME/Symantec	Symantec
/NP/responsable de le sécurité	responsable de le sécurité
/NP/attaque informatique	attaque informatique
/COMMON-NOUN/pare-feu	pare-feu

Tableau 3 Termes les plus représentatifs de la catégorie Cybersecurity

On note par exemple comme terme le moins significatif la présence du nom propre *Guillaume poupard*, le modèle l'ayant considéré comme représentatif de la catégorie alors qu'il n'est pas aussi présent que les autres termes dans le corpus. En effet, les calculs implémentant le TF/IDF, certains descripteurs apparaissant dans peu de documents verront leur poids augmenter par rapport aux descripteurs très présents dans tous les documents. Dans ce cas précis, ce descripteur ne risque pas pour autant de provoquer du bruit étant donné qu'il s'agit d'un nom propre ayant peu de chance d'être significatif pour d'autres catégories, ce qui n'aurait pas été le cas de noms propres tels que des noms d'entreprise comme *Microsoft* et *Apple*.

Une fois les descripteurs de chaque classe jugés pertinents, le modèle est prêt à être testé sur les 10 % restants du corpus, ce que nous verrons en 6.1.

Nous avons vu, à travers la modification d'un modèle d'apprentissage préexistant, que l'approche par apprentissage automatique nécessite d'étudier les caractéristiques des documents du corpus d'apprentissage, afin de s'assurer que les descripteurs qui seront sélectionnés par le modèle seront les plus représentatifs possibles de chaque catégorie.

IV Classification par hiérarchie de concepts

4.1 Description du projet Digital Working

Un second projet de classification automatique concernait la thématique du Digital Working, dont l'objectif est d'identifier des documents relatifs à l'utilisation des outils du numérique dans le contexte professionnel. Cette thématique est au cœur de la transformation digitale des entreprises et intéresse donc de plus en plus de collaborateurs souhaitant se tenir informés quant aux changements induits par le digital dans le quotidien des salariés. Sont ainsi considérés comme faisant partie de la notion de *digital working* les concepts relatifs à la mobilité (ex : le télétravail), l'utilisation des réseaux sociaux en entreprise, le « BYOD » (*Bring Your Own Device*), le travail collaboratif et l'échange d'informations.

Pour cette thématique, l'utilisation de l'apprentissage automatique n'était pas pertinente : les termes représentatifs sont ici facilement identifiables et en nombre limité. Dans ce contexte, nous avons donc décidé de nous appuyer sur une approche entièrement linguistique via la création d'une hiérarchie de concepts, qui nous permettra de catégoriser un document dès lors qu'il contient des termes présents dans cette hiérarchie, ces termes étant associés à d'autres jouant le rôle de contextes. Cette approche reprend ainsi celle de Englmeier *et al.* (2001) mais sans utiliser de couche statistique pour sélectionner nos concepts.

4.2 Les concepts et leur hiérarchie

La stratégie choisie pour Digital Working a été de créer deux cartouches de connaissance distinctes, une pour chaque langue. En effet, au fur et à mesure de la constitution des concepts au regard de documents abordant cette thématique, il s'est avéré qu'une traduction d'un concept d'une langue à l'autre était impossible : il a donc fallu considérer ce thème comme deux projets distincts, chacun étant adapté à la langue en question.

Quelle que soit la langue, les deux cartouches suivent la même logique, possédant chacune une division entre outils et usages dans leur hiérarchie.

4.2.1 Français

La cartouche française est constituée de 205 concepts, soit 113 appartenant à la classe Outil et 92 à la classe Usage. Chaque classe est composée de plusieurs catégories, certaines se répétant d'une classe à l'autre :

outil	usage
appareil	BYOD
communication en temps réel	communication en temps réel
mobilité	mobilité
transfert d'informations	échange d'informations
messagerie	travail collaboratif
réseau social	

Tableau 4 Catégories de la thématique digital working (français)

Cette division en outils et usages digitaux s'explique par le fait que les concepts ont été créés au fur et à mesure de la constitution du corpus : dès qu'un article présent sur la plateforme a été identifié comme appartenant au domaine du *digital working*, il s'est avéré que les sujets traitaient soit d'usages du digital en entreprise, soit d'outils utilisés pour mettre en œuvre cette stratégie digitale. Cette logique a donc été conservée lors de la création des concepts.

La catégorie Mobilité située dans Outil décrit ainsi les outils mis en œuvre pour télétravailler, comme les *espaces de partage*, le *cloud*, les *bureaux virtuels* etc., tandis que les concepts de Mobilité dans Usages mettent l'accent sur le fait de télétravailler, avec des termes tels que *télétravail*, *demande de mobilité*, *dématérialisation*, etc.

Chaque concept de ces catégories peut être associé à un contexte Entreprise ou Général : le premier contient 64 termes appartenant aussi bien au domaine de l'entreprise que du digital comme *stratégie digitale*, *entreprise 2.0* etc., tandis que le second regroupe 79 termes du monde du travail plus généraux, tel que *collègue*, *cadre*, *collaborateur* etc.

e-management	multinationale
e-marketeur	mutation numérique
entreprise 2.0	nomadisme
entreprise digitale	numérique
entreprise du web	organisation
entreprise numérique	organisation collaborative
gouvernance 2.0	organisation digitale
industrie digitale	patrons du numérique
industrie du numérique	professionnel
industrie du web	professionnel du numérique
leadership	projet
leadership digital	ressources humaines
leadership numérique	RH 2.0
management stratégique	révolution numérique
	salarié

Fig.6 Extraits des contextes Entreprise et Général

Les contextes peuvent s'appliquer sur deux zones : le document ou la phrase. Par exemple, pour un concept dont le sens est très représentatif du domaine du *digital working*, comme *télétravail*, nous pouvons lui associer un contexte général sur le document. Cependant, pour des termes plus génériques comme ceux appartenant aux outils, telle que la catégorie Appareil (*smartphone*, *ordinateur portable*, *poste de travail*, etc.), nous leur assignons un contexte Entreprise et restreint à la phrase. Ainsi, si un document contient le mot *smartphone* mais qu'aucun terme du contexte Entreprise n'est présent dans la phrase dans laquelle il est utilisé, le concept ne sera pas extrait.

4.2.2 Anglais

La cartouche anglaise, quant à elle, contient moins de concepts que pour le français avec 161 concepts, 86 pour la classe Outil et 75 pour la classe Usage.

outil	usage
device	BYOD
realtime communication	realtime communication
mobility	mobility
transfert of informations	exchange of informations
messaging	collaborative work
social media	

Tableau 5 Catégories de la thématique digital working (anglais)

Les concepts inclus dans ces catégories sont semblables à ceux de la cartouche française. Le nombre de contextes est cependant doublé par rapport au français, avec 4 classes : Entreprise, General, Digital et Collaborative Work. L'ajout de ces deux contextes supplémentaires a été nécessaire du fait de l'utilisation plus générale des concepts dans les documents en anglais par rapport au français.

Name	expression.en...	variant.english
corporate data		
entreprise 2.0		
entreprise mobility		
Human resources		director of HR;director of human resources;HR;HR director;human resources director
IT administrator	:IT / administrator	
IT businesses	:IT / businesses?	
IT department	:IT / department	information technology department
IT leader	:IT / leader	
IT organization	:IT / organization	
multinational company		multinational corporation;multinational enterprise

Tableau 6 Extrait des contextes Entreprise en anglais

Le contexte Entreprise, comme pour le français, contient des termes spécifiques au monde de l'entreprise et admet en plus de nombreux synonymes et des expressions régulières. Par exemple, la notation « :IT / administrator » signifie que le contexte doit être composé des initiales IT en respectant la casse, suivi du nom *administrator*.

Le contexte Digital prend en compte 30 termes relatifs au digital tels que *digital agency*, *digital company*, *digital innovation*, *workplace transformation* etc.

De même, le contexte Collaborative Work est composé de 12 termes comme *collaborative groupwork*, *collaborative workspace*, *virtual team* afin de cibler uniquement les concepts dans un contexte de travail collaboratif, qui est un des sujets principaux pris en compte dans la notion de « Digital Working ».

Enfin, le contexte General regroupe les termes de tous les contextes présentés ci-dessus et est utilisé pour les concepts spécifiques au *digital working* qui ne nécessitent pas d'être associés à un contexte trop restreint.

4.3 Constitution des corpus

Le corpus français est constitué de 99 documents provenant aussi bien de la plateforme Leonard que de sites web, notamment L'Atelier² qui contient une thématique Digital Working à part entière dans laquelle des articles sont publiés quotidiennement. Contrairement au projet High Tech utilisant un fichier TMX pour associer chaque document à une catégorie, les documents du corpus n'ont pas été répartis selon les classes Appareil, Mobilité etc. étant donné qu'elles contiennent toutes plusieurs termes appartenant aux différentes catégories : un document abordant la question de la mobilité en entreprise ne va pas contenir uniquement les termes de la catégorie Mobilité, mais aussi ceux des catégories Appareil et Echange d'informations, par exemple. Les documents, en PDF ou HTML, sont donc intégrés directement dans le logiciel Annotation Workbench où ils sont prétraités par des cartouches créées à cet effet. Ce logiciel permet de tester les extractions effectuées par les cartouches de connaissance sur des corpus, et ainsi d'évaluer leur qualité grâce aux mesures de Précision, de Rappel et de F-Mesure, comme nous présenterons par la suite.

Pour l'anglais, 83 documents ont été sélectionnés de la même manière que pour le français, à l'exception qu'une veille a été effectuée sur d'autres sites que L'Atelier car celui-ci ne publie que des articles en français.

Un contre corpus a aussi été constitué afin de vérifier le bon fonctionnement de la cartouche. Ont ainsi été sélectionnés des documents abordant les réseaux sociaux et autres outils du numérique dans un contexte autre que celui de l'entreprise. La question du *digital working* est en effet particulièrement délicate : le contre corpus permet de vérifier que la cartouche n'identifie pas les concepts du digital lorsque ceux-ci ne sont pas liés à l'entreprise.

Après avoir présenté une première méthode statistique via un modèle d'apprentissage automatique, nous venons de voir une approche opposée, basée uniquement sur la linguistique par la modélisation manuelle d'une hiérarchie de concepts. Voyons à présent une dernière méthode mêlant ces deux aspects, via l'utilisation d'un thésaurus.

² <http://www.atelier.net/>

V Classification basée sur un thésaurus

Une autre thématique très présente dans la plateforme de veille ne disposant pas d'indexation automatique concerne les documents traitant de macro-économie, ayant la particularité de regrouper des sujets à la fois politiques, économiques, financiers, sociaux et relevant aussi des relations internationales. Ces différents thèmes étant voués à se croiser dans un même document, l'apprentissage automatique n'apparaît pas comme une bonne solution, le vocabulaire étant trop hétérogène. À l'inverse, l'utilisation d'une simple liste de termes n'est pas suffisante car il serait impossible de décrire de façon exhaustive tous les termes liés à ces différents aspects économiques, sociaux et politiques. Nous nous sommes donc tournés vers l'utilisation d'un thésaurus, qui apparaît comme un bon compromis entre méthode linguistique et statistique. En effet, à l'instar de la méthode de Kervers (2009), nous pouvons gérer l'extraction des concepts présents dans le thésaurus en y ajoutant une couche statistique nous permettant de fixer des seuils et autres paramètres précis qui nous permettraient de construire une stratégie adéquate pour cette thématique.

5.1 Historique du projet Macro-économie

5.1.1 Présentation

Afin de débiter sur une base existante en matière de concepts économiques, il a d'abord été décidé d'utiliser Eurovoc, un thésaurus multilingue de l'Union Européenne. Celui-ci a ensuite été restructuré une première fois afin d'obtenir les plus grandes thématiques qui serviront à classer les documents selon 5 thèmes principaux, correspondant aux catégories mères du thésaurus :

- finances
- vie économique
- vie politique
- relations internationales
- questions sociales

Chacune de ces catégories était constituée d'un certain nombre de concepts répartis en plusieurs sous-catégories, la profondeur du thésaurus allant jusqu'à trois niveaux.

D'autres concepts ont aussi été créés pour servir de contextes à ceux du thésaurus. De cette manière, les concepts associés à des contextes n'étaient extraits que si les concepts présents dans ces contextes se trouvaient aussi soit dans le document, soit dans la phrase. Trois contextes ont alors été créés, le premier contenant des termes relatifs à l'économie (*économie, économique, économiste, ...*), le second à la finance (*finance, bourse, ...*) et le troisième à l'emploi (*travail, chômage, emploi, ...*).

Les premières observations effectuées ont soulevé plusieurs problématiques qui furent prises en compte dès le début du projet. En effet, comme le domaine de la macro-économie fait appel à des concepts pouvant appartenir à plusieurs catégories selon le point de vue de chacun, la façon de hiérarchiser le thésaurus est sujette à la subjectivité de la personne chargée de le construire. Ce fait s'est ainsi démontré pour les catégories Finance et Vie économique, où des contradictions ont été mises en évidence après observation des extractions effectuées par la cartouche :

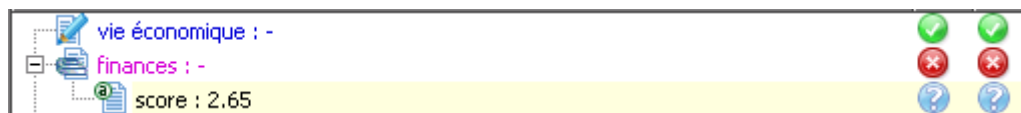


Fig.7 Observation des concepts extraits dans Annotation Workbench

En effet, nous retrouvions pour de nombreux cas la situation illustrée ici, à savoir qu'un document annoté comme Vie économique dans le corpus de référence se retrouvait indexé dans la catégorie Finances. Cela s'explique par le fait que le corpus de référence provenait d'une base de données documentaire utilisant un autre thésaurus, dont la structure différait alors de celle de notre thésaurus. Les notions de finance et de vie économique ne coïncidaient donc pas d'un thésaurus à l'autre, ce qui explique cette contradiction que l'on retrouvait pour la majorité des cas entre ces deux catégories.

Afin de pallier ces problèmes, nous nous sommes concentrés sur plusieurs pistes d'améliorations similaires à celles évoquées par Kervel, à savoir structurer plus finement les concepts du thésaurus d'une part, puis mettre en œuvre un système de seuils limitant l'extraction des concepts et des catégories.

5.1.2 Pistes d'amélioration

La première piste d'amélioration a été le choix de restructurer le thésaurus. En effet, il était constitué d'un grand nombre de termes trop généraux, qui étaient la cause de bruit. Ce manque de précision devait alors être le premier point à régler avant d'envisager d'autres solutions. De même, la constitution d'un nouveau corpus de référence devait être nécessaire afin de résoudre le problème décrit précédemment, à savoir la contradiction entre le thésaurus actuel et le corpus de référence basé sur le thésaurus de la documentation.

Une seconde piste consistait à modifier les paramètres pris en compte par la cartouche, qui permettent d'agir sur la façon dont les termes du thésaurus doivent être extraits. Plusieurs stratégies sont ainsi possibles en combinant plusieurs paramètres, ce qui fera l'objet d'une partie à part entière à la suite de ce mémoire (cf 5.4).

5.2 Réorganisation du thésaurus

5.2.1 Structure du nouveau thésaurus

Cette première étape a nécessité la participation de plusieurs collaborateurs afin de rendre la hiérarchie cohérente pour chaque point de vue. Nous avons donc décidé de nous appuyer dans un premier temps sur une seule personne afin de réorganiser les concepts, avant de mettre en commun les points de vue des différents collaborateurs associés au projet pour confirmer la pertinence de cette nouvelle hiérarchie. Voici un aperçu de la structure finale du thésaurus, d'après ses deux premiers niveaux hiérarchiques :

Catégories mères	Catégories filles
<i>Vie économique</i>	analyse économique comptabilité nationale conjoncture économique finances politique monétaire politique économique prix structure économique
<i>Vie politique</i>	cadre politique parlement parti politique pouvoir exécutif procédure électorale sécurité publique
<i>Economie sociale</i>	cadre social démographie et population emploi niveau de vie politique sociale protection sociale
<i>Relations internationales</i>	commerce international organisations internationales politique de coopération politique internationale

Tableau 7 Extrait de la structure du thésaurus (deux premiers niveaux)








Le thésaurus a été réduit à quatre catégories mères, les concepts de Finance et de Vie économique ayant été fusionnés, réglant ainsi le problème initial de similarité entre ces deux catégories. De nouveaux concepts furent créés et d'autres supprimés de sorte que cette nouvelle hiérarchie se détache dorénavant de celle du thésaurus de départ. En effet, l'objectif d'Eurovoc étant de constituer une sorte de dictionnaire des concepts économiques, beaucoup d'entre eux n'étaient plus d'actualité ou ne correspondaient pas à la vision macro-économique telle qu'elle est présentée dans les documents que nous devons catégoriser.

Le niveau de profondeur du thésaurus a aussi été augmenté, passant de 3 à 4 niveaux maximum. Ainsi, les termes les plus profonds dans la hiérarchie gagnent en précision et ont plus de poids par rapport à ceux des niveaux supérieurs.










Une fois le thésaurus restructuré, il fallait aussi faire de même pour les contextes.

5.2.2 Réorganisation des contextes










Les termes composant les contextes Economie, Finance et Emploi étaient d'une part constitués de peu de termes et, d'autre part, ne correspondaient plus à la nouvelle structure du thésaurus mêlant désormais vie économique et finance. Nous avons donc fusionné les concepts des contextes Economie et Finance et remplacé celui d'Emploi par un contexte Social. Un contexte Politique a aussi été créé, regroupant des termes trop généraux pour être intégrés au thésaurus mais restant tout de même représentatifs de cette catégorie. C'est aussi le cas pour les contextes Social et Economie, comme nous pouvons le voir dans le détail des différents contextes présenté ci-dessous :

ID	prefLabel.english	expression.engl...	prefLabel.french	expression.french
 actions	equities	:equities :equity	actions	
 bourse	stock exchange		bourse	boursi.*
 debt	debt		dette	
 economy	economy	econom.*	économie	économ.*
 emprunt	bond		emprunt	
 finance	finance	financ.*	finance	financ.*
 macroeconomy	macroeconomy	macro-?econom.*	macroéconomie	macro-?économ.*

Contexte Economie

ID	prefLabel.english	expression.engl...	prefLabel.french	expression.french
 aide_etat			aide de l'Etat	aides? / de / l'Etat
 conseil	Council	\pCouncil	Conseil	\pConseil
 depute			député	
 election	election	elections?	élection	élections?
 geopolitics	geopolitics		géopolitique	
 law	law project	legal.*	projet de loi	légal.* législ.*
 prestations_sociales			prestations_sociales	
 reform	reform		réforme	
 senateur	senator		sénateur	

Contexte Politique

ID	prefLabel.english	expression.engl...	prefLabel.french	expression.french
 emploi	employment		emploi	
 household	household	households?	ménage	ménages?
 human_resource	human resources		ressources humaines	
 pension	pension		pension	
 population	population		population	
 salarié	worker	workers?	salarié	salariés?
 social	social		social	sociales?
 travail			travail	travail.*
 unemployment	unemployment		chômage	chôm.*

Contexte Social

Tableaux 8,9,10 Contextes économique, politique et social

Les contextes peuvent contenir des expressions régulières, qui nous permettent d'éviter de créer des listes non exhaustives de termes et de préciser les formes que nous voulons prendre en compte. L'expression « économ.* » nous permet ainsi de capturer toutes les variantes du contexte : « économie », « économique », « économiste », etc., tandis que nous pouvons aussi choisir de préciser les formes plurielles d'un terme ou même sa casse, via l'expression « \p » précédant le terme en question.

5.3 Constitution des corpus

Afin d'évaluer les modifications effectuées précédemment, il a fallu constituer un nouveau corpus reflétant les nouvelles notions macro-économiques ajoutées au thésaurus.

Pour cela, nous avons sélectionné des articles correspondant aux thèmes Politique, Social, Economie et Relations internationales en interrogeant la base de données interne des Etudes Economiques et en récoltant des articles sur la plateforme Leonard. Voici la répartition des corpus pour chaque catégorie :

	Français	Anglais
Economie sociale	63	40
Relations internationales	34	24
Vie économique	112	101
Vie politique	78	87
Total	287	252

Tableau 11 Répartition des corpus pour Macro-Economie

Les corpus pour Economie sociale et Relations internationales sont moins conséquents que les autres catégories, car les articles ayant pour thème principal le commerce international et les questions sociales sont moins nombreux. Les deux catégories fortes de la macro-économie sont ici représentées par l'économie et la politique, qui disposent d'un plus grand nombre d'articles.

Tous ces documents, au format PDF, sont intégrés dans un fichier TMX de la même manière que pour le modèle d'apprentissage automatique abordé précédemment.

Un contre-corpus de 140 documents a aussi été créé, constitué à partir de documents présents dans Leonard à partir des secteurs ne traitant pas de macro-économie, à savoir le Big Data, l'éducation, le climat, mais aussi la banque. Cette dernière thématique nous permet ainsi de vérifier si les concepts de notre thésaurus relatifs à la finance ne créeront pas de bruit lorsque les documents traiteront du secteur bancaire, et dans le cas contraire, nous permet d'ajouter des contextes en observant les extractions posant problème.

5.4 Paramétrages

Une fois le corpus constitué, nous pouvions passer à l'étape la plus cruciale et la plus chronophage du projet : le paramétrage des extractions des concepts. En effet, chaque modification de paramètre nécessite d'en tester les conséquences sur le processus d'extraction.

Cette méthodologie permet ainsi de savoir quels paramètres sont les plus pertinents au fil des changements effectués.

Les différents paramètres pris en compte pour l'extraction des concepts du thésaurus sont répartis en plusieurs cartouches : on parle alors de *plan d'annotation*. Les cartouches composant un plan d'annotation sont connectées entre elles selon un ordre précis, de sorte que certaines cartouches ont un impact direct sur l'extraction tandis que d'autres jouent un rôle relevant davantage de filtre supplémentaire permettant l'affinement des paramètres des cartouches principales.

5.4.1 Paramètres STF

La cartouche STF (*Smart Taxonomy Facilitator*) est la cartouche principale du plan d'annotation : c'est elle qui nous permet de gérer les paramètres dédiés à l'extraction des concepts du thesaurus. Elle permet de gérer des paramètres allant de l'identification des concepts jusqu'à leur synchronisation et la façon dont ils doivent être évalués.

- Synchronisation

Tout d'abord, des paramètres de synchronisation sont nécessaires afin de spécifier les langues utilisées (ici le français et l'anglais) et la méthode de normalisation à appliquer sur le lexique (le lemme ou la forme). Nous avons choisi la lemmatisation car celle-ci permet d'appliquer de futurs traitements sur les concepts grâce à d'autres paramètres relatifs à la syntaxe des phrases.

- Matching

Ensuite, les paramètres relatifs à l'extraction, ou « matching », offrent plusieurs angles d'approche. Nous pouvons par exemple choisir de désambiguer les concepts pouvant être présents plusieurs fois dans le thesaurus mais à différents niveaux. Dans ce cas, seul le terme possédant le plus haut score sera extrait. Ce paramètre n'est cependant pas pertinent dans notre cas étant donné qu'aucun concept n'est présent plusieurs fois dans la hiérarchie. Par contre, notre stratégie prend en compte la méthode de recherche à utiliser dans le cas où une même portion de phrase contient plusieurs concepts qui se chevauchent. Par exemple, dans la phrase : « il faut diversifier les sources de financement des entreprises », trois concepts du thesaurus peuvent être extraits, à savoir « sources de financement », « financement des entreprises » et « source de financement des entreprises ». Trois méthodes sont alors possibles :

- l'extraction de chaque concept individuellement (*all matches*),
- l'extraction du concept le plus long (*longest match*),
- l'extraction du concept le plus long en autorisant l'insertion d'un mot (*relaxed longest match*).

Nous avons choisi d'extraire les concepts les plus longs car ils sont susceptibles d'être plus spécifiques à une catégorie, et donc plus pertinents lors de la catégorisation.

En reprenant la phrase précédente, si nous sélectionnons « all matches », les trois concepts vont être extraits. Avec l'option « longest match », seul « source de financement des entreprises » sera extrait. Cela va donc créer deux concepts avec un score moins élevé que s'il était extrait avec le concept le plus long « sources de financement des entreprises ».

Nous pouvons aussi choisir la distance d'édition maximum pour qu'un terme soit extrait (ici 1, soit la forme du concept tel qu'il est écrit dans le thesaurus).

- Evaluation

Viennent ensuite les paramètres relatifs à la façon d'évaluer les extractions, composés de plusieurs procédures qui joueront sur la qualité de nos résultats.

Tout d'abord, nous définissons la formule qui permettra de calculer les scores des termes, ici le *Weighted TF/IDF*. En plus de la formule de TF/IDF de base prenant en compte la fréquence des termes dans un document pour calculer leur score, cette variante y ajoute le

score de correspondance (*matching*) et prend en compte la longueur des termes ainsi que leur niveau de profondeur dans la hiérarchie du thésaurus.

Un second paramètre nous permet de préciser la procédure à utiliser pour sélectionner les termes en se basant sur leur score. Nous pouvons par exemple choisir de conserver uniquement les n termes les plus significatifs en matière d'occurrences ou même en pourcentage, ou bien décider de conserver tous les termes ou uniquement ceux obtenant les meilleurs scores. C'est cette dernière approche que nous avons choisie, car elle nous permet de voir quels termes sont les plus représentatifs de nos documents en situation réelle, c'est-à-dire ceux qui seront extraits d'après les paramètres de *matching* définis précédemment.

Une dernière procédure nous donne la possibilité de donner plus de poids aux termes selon leur classe grammaticale. Nous avons ainsi donné plus de poids aux noms car, de la même manière que pour la définition de notre modèle d'apprentissage présenté plus haut, les noms sont les plus représentatifs d'une thématique comparé aux autres classes tels que les verbes et autres adjectifs.

Une seconde cartouche rattachée à la STF permet d'augmenter sa précision, en indiquant par exemple l'unité de base sur laquelle repose l'extraction, dans notre cas la phrase.

5.4.2 Paramètres de gestion des contextes et de seuil des catégories

Une troisième cartouche permet de gérer les contextes liés aux concepts du thésaurus. Il suffit alors d'indiquer le nom de la classe contenant les contextes ainsi que celle contenant les concepts du thésaurus. Des paramètres de synchronisation permettent aussi de choisir d'ignorer ou non la casse lors de l'identification des concepts, et si celle-ci doit s'effectuer sur les lemmes ou la forme.

Dans notre cas, les concepts sont extraits d'après leur lemme et la casse est ignorée. Cependant, pour des cas particuliers comme les acronymes et autres termes que nous souhaitons extraire en respectant la casse et même la forme, nous pouvons le préciser directement en sélectionnant le concept en question dans le thésaurus, comme représenté ci-dessous.



Fig.8 Illustration du concept « crédit agricole » dans la vue Webstudio

En prenant l'exemple du concept « crédit agricole », nous devons donc restreindre son extraction à la forme et la casse exactes si nous voulons éviter qu'il soit confondu avec le nom propre Crédit Agricole, ce qui aurait été le cas en laissant le paramètre de *fuzzy matching* par défaut.

En effet, la méthode d'extraction appliquée par défaut sur tous les termes du thésaurus est basée sur le *fuzzy matching*. Cette technique de correspondance partielle de texte permet d'identifier les variantes orthographiques d'un concept tel que le pluriel, mais admet aussi l'insertion, la permutation de mots et les erreurs orthographiques. Un concept identifié tel qu'il est écrit dans le thésaurus aura ainsi un score plus élevé (soit le maximum : 1.0) par rapport à un concept identifié par correspondance partielle. Plus précisément, un concept identifié avec une permutation ou une insertion de mots aura un score d'environ 0.8, tandis qu'une variante orthographique donnera un score de 0.9. Nous avons choisi par ailleurs de ne pas prendre en compte les erreurs orthographiques, puisque cette option a tendance à provoquer beaucoup de bruit.

Enfin, un plug-in rattaché à la cartouche gérant les contextes permet d'attribuer un seuil minimal sous lequel une catégorie ne sera pas extraite. Il nous permet aussi de définir le nombre maximum de catégories possibles pour un même document. Dans notre cas, nous autorisons deux catégories maximum pour l'indexation d'un document, avec un seuil minimum de 1.0. En effet, après observation des extractions sur les corpus français et anglais ainsi que sur le contre-corpus, nous avons remarqué que ce seuil était un bon compromis pour empêcher l'indexation de documents dans le contre corpus sans créer trop de silence dans le corpus d'apprentissage. L'évolution des résultats en fonction du seuil appliqué est présentée ci-dessous :

Seuil	0.6	0.8	1.0	1.2	1.4	2.0
Documents manqués	10	13	18	33	51	95
Documents mal catégorisés	42	26	13	10	6	1
Nb de documents indexés dans le contre-corpus	51	45	36	31	28	22

Tableau 12 - Evolution par rapport au seuil minimum des catégories

D'après ce test, effectué sur le corpus français et sur la totalité du contre-corpus, on constate que l'augmentation du seuil minimum est proportionnelle à celle du nombre de documents manqués, tandis qu'à l'inverse, les erreurs de catégorisations diminuent. De même, il est logique que le nombre de documents catégorisés dans le contre-corpus baisse au fur et à mesure que l'on augmente le seuil. On remarque ainsi que dès que le seuil dépasse 1.0, l'écart entre le nombre de documents manqués et mal catégorisés (correspondant aux mesures de précision et de rappel que nous présentons ci-après) devient de plus en plus important. Ce seuil de 1.0 nous est donc apparu comme le plus adapté, offrant des résultats plus équilibrés.

Nous venons de voir une méthode mêlant linguistique et couche statistique, constituant une sorte de méthode hybride combinant les approches des projets Nouvelles Technologies et Digital Working. Nous comparerons à présent les résultats obtenus pour chacune de ces approches.

VI Evaluations

L'évaluation de la performance de chacune de ces méthodes est effectuée de deux manières dans notre cas. D'une part, l'évaluation quantitative est observable via l'application des formules de rappel, précision et F-mesure intégrées dans les logiciels utilisés lorsque nous leur soumettons les corpus de référence présentés précédemment.

La précision désigne le nombre de documents correctement catégorisés sur le nombre total de documents catégorisés, représentée par la formule suivante :

$$\text{Précision}_n = \frac{\text{documents correctement catégorisés dans la catégorie } n}{\text{nombre total de documents appartenant à la catégorie } n}$$

Le rappel est calculé d'après le nombre de documents correctement catégorisés par rapport au nombre total de documents des catégories de référence :

$$\text{Rappel}_n = \frac{\text{documents correctement catégorisés dans la catégorie } n}{\text{nombre de documents dans la catégorie de référence } n}$$

Quant à la F-mesure, il s'agit d'une moyenne calculée à partir de la précision et du rappel, représentée par la formule suivante :

$$\text{F-mesure} = \frac{2 \cdot (\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})}$$

D'autre part, l'évaluation qualitative des outils se fait en les testant directement sur la plateforme Leonard après une vérification des documents catégorisés sur une période donnée.

6.1 Modèle d'apprentissage automatique

6.1.1 Français

L'évaluation du modèle d'apprentissage s'effectue via le logiciel Category Workbench une fois la phase d'apprentissage terminée et le modèle exporté lorsque celui-ci nous convient. Le logiciel nous donne alors les statistiques globales ainsi que les résultats pour chaque catégorie. Nous obtenons ainsi un résultat global tournant autour de 87% de F-Mesure, avec une précision de 82% et un rappel de 92%. Les résultats de nos nouvelles catégories pour le français sont les suivants :

/IT-DIGITAL-TELECOMS/DIGITAL/BIG DATA	54.5	100.0
/IT-DIGITAL-TELECOMS/DIGITAL/CYBERSECURITY	100.0	100.0
/IT-DIGITAL-TELECOMS/DIGITAL/SOCIAL MEDIA	72.7	100.0
/IT-DIGITAL-TELECOMS/IT/3D PRINTING	62.5	100.0
/IT-DIGITAL-TELECOMS/IT/CLOUD	100.0	100.0
/IT-DIGITAL-TELECOMS/IT/PC-TABLET-SMARTPHONE	76.0	100.0
/IT-DIGITAL-TELECOMS/TELECOMS	85.4	93.4

Tableau 13 Précision et Rappel des catégories IT-Digital-Telecoms (français)

On remarque que le rappel est satisfaisant, mais que la précision reste à améliorer. En effet, en comparant ces résultats avec la précision des autres catégories, on remarque que celle des nouvelles catégories créées est inférieure aux autres. Cela s'explique par le fait que les corpus étant beaucoup moins volumineux que pour les anciennes catégories, le nombre des descripteurs pris en compte par le modèle s'en retrouve lui aussi réduit. En effet, la fréquence des descripteurs ne dépasse pas 80 pour nos nouvelles catégories (excepté PC-Tablet-Smartphone et Telecoms, ayant des corpus plus conséquents) tandis que les descripteurs des autres catégories sont présents plus de 200 fois en moyenne.

Il ne faut cependant pas s'appuyer uniquement sur ces résultats quantitatifs, car même si le modèle admet qu'un même document puisse être classé dans deux catégories différentes, cela n'est pas pris en compte dans l'évaluation. Ainsi, en observant l'assignement des documents pour la catégorie Réseau social, on remarque que les documents considérés comme mal catégorisés parlent bien de réseaux sociaux même s'ils appartiennent à une autre catégorie, par exemple E-commerce ou encore Médias.

Les performances réelles sont ainsi observables en testant les modèles sur la plateforme de test de Leonard. Après une semaine d'indexation sur les articles de presse quotidienne, on obtient les résultats suivants pour le français :

Catégories	Nb de docs catégorisés	Nb de docs corrects	Nb de docs incorrects	% correct
IT	14	13	0	92%
PC-Tablette-Smartphone	6	6	0	100%
3D printing	2	1	0	50%
Cloud	6	6	0	100%
Numérique	14	13	1	92%
Cybersécurité	3	3	0	100%
Big Data	6	6	0	100%
Réseaux sociaux	5	4	1	80%
Télécoms	10	10	0	100%
Total	38	36	1	94%

Tableau 14 Résultats qualitatifs du modèle français

On remarque que les résultats correspondent globalement à ce qui était attendu au regard de l'évaluation quantitative, à savoir que les catégories offrant le plus de précision sont Cybersécurité, Télécoms, Cloud et PC-Tablette-Smartphone. La catégorie Big Data, quant à elle, a donné de très bons résultats malgré le manque de précision observée lors de l'apprentissage. Il s'avère que les documents indexés pendant cette semaine de tests étaient très significatifs de cette thématique et ne provoquait alors pas de bruit, ce qui explique cette absence d'erreurs. La catégorie 3D Printing souffre des mêmes lacunes que lors de l'apprentissage puisqu'elle indexe des documents traitant d'impression mais pas d'impression 3D. Concernant Réseaux Sociaux, une seule erreur était due au fait qu'il était précisé qu'un politique était « suivi par plusieurs millions d'utilisateurs sur son compte Twitter » mais le réseau social n'était pas le thème principal de l'article.

Nous comparerons à présent ces résultats avec ceux du modèle anglais.

6.1.2 Anglais

Pour l'anglais, nous obtenons de la même manière les résultats suivants :

/IT-DIGITAL-TELECOMS/DIGITAL/BIG DATA	57.1	100.0
/IT-DIGITAL-TELECOMS/DIGITAL/CYBERSECURITY	100.0	100.0
/IT-DIGITAL-TELECOMS/DIGITAL/SOCIAL MEDIA	71.4	100.0
/IT-DIGITAL-TELECOMS/IT/3D PRINTING	100.0	100.0
/IT-DIGITAL-TELECOMS/IT/CLOUD	62.5	100.0
/IT-DIGITAL-TELECOMS/IT/PC-TABLET-SMARTPHONE	60.0	100.0
/IT-DIGITAL-TELECOMS/TELECOMS	96.0	92.3

Tableau 15 Précision et Rappel des catégories IT-Digital-Telecoms (anglais)

On remarque que certaines catégories obtiennent de meilleurs résultats en anglais qu'en français, notamment 3D Printing, ce qui s'explique par une meilleure prise en compte des descripteurs lors de l'apprentissage. En effet, les descripteurs anglais sont plus pertinents car prennent en compte le groupe nominal « 3D + nom » contrairement au français, comme nous pouvons le voir ci-dessous :

/IT-DIGITAL-TELECOMS/IT/3D PRINTING		/IT-DIGITAL-TELECOMS/IT/3D PRINTING	
Full name	Name	Full name	Name
/COMMON-NOUN/3D-printed	3D-printed	/COMMON-NOUN/stéréolit...	stéréolithographie
/COMMON-NOUN/cad	cad	/COMMON-NOUN/trachée	trachée
/COMMON-NOUN/prosthetics	prosthetics	/NP/Big delta	Big delta
/COMMON-NOUN/prototyping	prototyping	/NP/défense militaire	défense militaire
/COMMON-NOUN/stereolitho...	stereolithography	/NP/fabrication additif	fabrication additif
/NP/3D print	3D print	/NP/main fabriquer	main fabriquer
/NP/3D print technology	3D print technology	/NP/marché de le impression	marché de le impression
/NP/3D printer	3D printer	/NP/objet du quotidien	objet du quotidien
/NP/3D printer market	3D printer market	/NP/partie de crâne	partie de crâne
/NP/3D printer technology	3D printer technology	/NP/service d' impression	service d' impression
/NP/3D system	3D system	/PROPER-NAME/Autodesk	Autodesk

Tableaux 16, 17 Les onze principaux descripteurs de la catégorie 3D Printing pour l'anglais et le français

On obtient ainsi une F-Mesure de 100% pour cette catégorie en anglais, tandis que la précision pour le français est dégradée car les documents sont confondus avec d'autres catégories telles que Wood Paper-Packaging-Furniture qui contient des sujets relatifs à l'imprimerie.

Quant à la catégorie Big data, il s'agit de la moins performante que ce soit en anglais ou en français, étant donné qu'elle possède le vocabulaire le moins homogène car les documents du corpus traitent toujours du Big Data en l'associant à des thématiques différentes, comme la gestion de relation client, le stockage de données et le cloud, etc. Cette dernière catégorie se retrouve justement bruitée par le Big Data, les documents tests étant souvent associés à ces deux catégories en même temps.

Comme pour le modèle français, nous avons effectué une évaluation qualitative sur la plateforme de tests sur une semaine, qui nous a donné les résultats suivants :

Catégories	Nb de docs catégorisés	Nb de docs corrects	Nb de docs incorrects	% correct
IT	21	20	1	95%
PC-Tablet-Smartphone	11	11	0	100%
Impression 3D	1	0	1	0%
Cloud	9	9	0	100%
Digital	21	19	2	90%
Cybersecurity	12	12	0	100%
Big Data	1	1	0	100%
Social media	8	6	2	75%
Telecoms	1	1	0	100%
Total	43	40	3	93%

Tableau 18 Résultats qualitatifs du modèle anglais

Les performances sont équivalentes au français, avec les mêmes lacunes concernant les catégories 3D Printing et Social Media. En effet, malgré l'avantage de l'anglais qui prend en compte le groupe nominal « 3D + N », il prend tout de même en compte d'autres descripteurs plus généraux et notamment *printer*, ce qui provoque les mêmes erreurs que pour le français.

6.2 Approche avec hiérarchie de concepts

L'évaluation quantitative des cartouches française et anglaise s'est effectuée en procédant à une annotation manuelle après l'import des corpus dans le logiciel Annotation Workbench. En effet, l'utilisation d'une annotation de référence via un fichier TMX n'était pas adaptée pour ce projet car les documents sont difficilement catégorisables dans une seule des catégories composant la thématique du *digital working*. Chaque document abordant ce thème n'est pas uniquement centré sur la communication en temps réel, la mobilité, etc. mais contiennent plusieurs de ces aspects en même temps, ce qui explique la nécessité d'annoter manuellement les extractions pour chaque document du corpus de test afin d'obtenir les résultats pour chacune des catégories.

	Précision	Rappel	F-Mesure
Outil	94,3	91,2	92,7
Appareil	94,3	91,7	93
Communication en temps réel	100	100	100
Messagerie	85,7	85,7	85,7
Mobilité	100	100	100
Réseau social	86,2	86,5	86,4
Transfert d'informations	100	83,3	90,9
Usage	98,2	97,4	97,8
BYOD	99,4	95,3	97,3
Communication en temps réel	100	100	100
Echange d'informations	100	100	100
Mobilité	100	100	100
Travail Collaboratif	91,7	91,7	91,7
Total	96,2	94,3	95,2

Tableau 19 Résultats pour le français

Pour le français, nous remarquons des résultats très satisfaisants dans l'ensemble, avec une précision légèrement plus élevée que le rappel. Les catégories Communication en temps réel, Mobilité, Echange et Transfert d'informations sont celles obtenant les meilleurs résultats avec 100% de F-Mesure. Concernant les catégories donnant de moins bons résultats, on peut noter que Messagerie et Réseau social sont celles qui posent le plus problème. En effet, les termes qu'elles contiennent sont les plus susceptibles de causer du bruit, en particulier *email* et les noms de réseaux sociaux *Twitter*, *Facebook* etc. qui, malgré leur contexte Entreprise au niveau de la phrase, peuvent encore être extraits dès lors que le contexte est trop générique. C'est par exemple le cas pour la phrase : « le ministre de la Mobilité et des Travaux Publics a répondu sur Twitter (...) », où le contexte Entreprise associé au concept *Twitter* comprend le terme *mobilité*.

	Precision	Recall	F-Mesure
Outil	98,4	90,4	94,4
Device	90,9	87,8	89,3
Messaging	100	82,6	91,3
Mobility	100	92,4	96
RealtimeCommunication	100	100	100
SocialMedia	100	87,3	93,6
TransfertOfInformation	100	92,8	96,4
Usage	98,7	97,7	98,2
BYOD	100	100	100
CollaborativeWork	95,2	93,6	94,4
ExchangeOfInformation	100	100	100
Mobility	98,6	94,9	96,7
RealtimeCommunication	100	100	100
Total	98,5	94	96,3

Tableau 20 Résultats pour l'anglais

Les résultats pour l'anglais sont semblables au français, avec une précision encore meilleure. En effet, on remarque que huit catégories sur onze ont une précision maximale, ce qui s'explique par l'utilisation des contextes supplémentaires, plus précis. Le rappel est quant à lui équivalent à celui du français, avec la même problématique que précédemment pour les catégories Messagerie et Réseau social.

Une fois ces résultats jugés satisfaisants, une seconde évaluation est effectuée pour tester la catégorisation sur le corpus de la plateforme Leonard. Après une semaine d'évaluation des documents catégorisés par la cartouche, on obtient les résultats suivants :

Catégorie	Anglais	Français	% correct par catégorie
appareil	8/8	10/10	100%
messagerie	1/1	1/2	66%
mobilité	11/13	23/24	92%
comm. temps réel	1/1	10/11	90%
réseau social	8/8	5/7	86%
transfert d'informations	3/3	5/5	100%
BYOD	7/7	9/10	94%
échange d'informations	4/4	5/5	100%
travail collaboratif	2/2	6/6	100%
% correct par langue	95%	92%	

Tableau 21 Résultats de l'évaluation qualitative

Les écarts de résultats entre l'anglais et le français s'expliquent par le fait que les articles en langue anglaise traitant du *digital working* sont moins nombreux que ceux en français dans la presse quotidienne. On peut notamment constater cette différence avec la catégorie Mobilité, très présente en français mais presque moitié moins fréquente en anglais. Un seul document parlant de messagerie et de communication en temps réel ayant été catégorisé sur la période d'évaluation, celle-ci n'est pas la plus complète possible mais nous permet néanmoins de constater quelles thématiques sont les plus représentatives du domaine dans la presse, à savoir la mobilité, les appareils et les réseaux sociaux.

Voyons à présent les résultats obtenus pour la dernière méthode de notre étude, à savoir la classification basée sur un thésaurus.

6.3 Approche avec thésaurus

L'évaluation de la cartouche Macro-Economie s'est effectuée en deux temps : les mesures de Précision, Rappel et F-Mesure ont d'abord été calculées en fonction de la catégorie de référence associée à chaque document, puis une annotation manuelle a été ajoutée afin de prendre en compte les cas où un document appartenait à deux catégories et que celles-ci étaient correctes.

Catégorie	Précision	Rappel	F-Mesure
relations internationales	97,2	87,5	92,1
vie politique	94,6	95,6	95,1
vie économique	100	98,1	99,1
économie sociale	90,5	91,8	91,2
total	96,4	95,1	95,7

Tableau 22 Résultats pour le français

Catégorie	Précision	Rappel	F-Mesure
relations internationales	90	94,7	92,3
vie politique	92,3	89,4	90,8
vie économique	100	100	100
économie sociale	89,1	95,3	92,1
total	95	95,6	95,3

Tableau 23 Résultats pour l'anglais

On remarque que les résultats sont équivalents d'une langue à l'autre et plutôt homogènes, le français obtenant une précision légèrement supérieure par rapport à l'anglais, tandis que leurs rappels sont presque similaires. La précision maximale pour vie économique s'explique par le fait que tous les documents, qu'ils soient classés en premier dans les catégories vie politique, économie sociale ou relations internationales, abordent tous des concepts économiques. Il est donc juste que la seconde catégorie attribuée à certains documents soit Vie économique. Le rappel à 98% pour le français est dû à un manque de concepts de Vie économique dans trois documents manqués, le score des concepts donnant un résultat inférieur au seuil fixé à 1.0.

Concernant le contre-corpus, 36 documents sur 140 ont été classés alors qu'ils ne font pas partie du thème macro-économique. Ces documents posant problème sont en majorité ceux traitant du domaine bancaire, étant donné qu'ils partagent certains concepts avec Vie économique (*action financière, banque centrale, réglementation financières, etc.*). Ces termes ont été extraits car les contextes qui leurs sont associés sont aussi présents dans les documents. La plus grande difficulté de cette thématique est que des termes très généraux utilisés comme contextes (*économie, finance, social* et leurs variantes) se retrouvent en effet dans des documents n'abordant pas forcément des questions macro-économiques.

Cependant, ces contextes ne peuvent pas être supprimés car ils sont les plus représentatifs des thématiques économiques, sociales et politiques pour la grande majorité des documents de nos corpus.

VII Discussion

D'après les différents tests effectués pour chacune des trois approches présentées précédemment, nous pouvons en dégager plusieurs avantages, de même que des inconvénients.

Premièrement, l'inconvénient de la méthode utilisant la hiérarchie de concepts par rapport au thésaurus est que même si l'on peut réduire le bruit en associant un terme à son contexte, il suffit que ces derniers ne soient présents qu'une fois dans un document pour qu'il soit catégorisé. Par exemple, dans le cas du projet Digital Working, un document abondant la politique a été identifié comme parlant de réseau social étant donné qu'il contenait une occurrence de *Twitter* associé à son contexte *entreprise*. Or, les statistiques utilisées dans la méthode par thésaurus permettent justement d'éviter cela puisqu'une catégorie ne sera validée que si le score cumulé des termes qui la représente dépasse un certain seuil.

En revanche, l'avantage de la méthode par hiérarchie de concepts est qu'elle est plus facile à manipuler et plus rapide à mettre en œuvre, tandis que l'utilisation d'un thésaurus demande un développement plus fastidieux car nécessite d'étudier les interactions entre les différents paramètres statistiques et leurs conséquences sur l'extraction des termes. Aussi, plusieurs cartouches peuvent être développées pour un même projet en étant adaptées à des langues différentes, ce qui n'est pas possible pour la méthode thésaurus : il faut donc pour cette dernière être sûr qu'il n'y ait pas d'ambiguïtés entre les termes d'une langue à l'autre.

Malgré ses avantages, l'inconvénient de la méthode par thésaurus est que la couche statistique nous empêche de contrôler parfaitement l'extraction et nous confronte à des questionnements sur certaines d'entre elles : on peut se demander pourquoi certains termes ne sont pas extraits alors qu'ils devraient l'être, ce qui est du à la combinaison des différents paramètres.

Le même problème se retrouve pour la méthode par apprentissage automatique, dont la principale difficulté consiste à trouver le bon paramétrage qui permettra d'obtenir des résultats satisfaisants pour toutes les catégories. Or, plus le nombre de catégories à prendre en compte augmente, plus il devient difficile de conserver des résultats optimaux étant donné que le vocabulaire des différentes catégories risque de se croiser. Cependant, si elles possèdent un vocabulaire assez spécifique, cette méthode a l'avantage d'être plus rapide à mettre en place à partir du moment où les classes sont définies et les corpus rapides à constituer.

Conclusion et perspectives

Nous avons abordé plusieurs méthodes permettant de catégoriser des documents en vue d'une meilleure accessibilité de ces derniers pour les utilisateurs d'une plateforme de veille collaborative. Chaque méthode dépend ainsi de la thématique à catégoriser, chacune nécessitant une stratégie particulière : les modèles d'apprentissage conviennent aux sujets plus généraux mais possédant des vocabulaires assez distincts pour ne pas se chevaucher, les hiérarchies de concepts simples sont bien appropriées aux thématiques possédant un vocabulaire très précis et restreint, tandis que l'utilisation d'un thésaurus opère un compromis entre ces deux méthodes, mêlant à la fois ressource linguistique et couche statistique.

Au vu de leurs points faibles évoqués plus haut, chacune de ces méthodes pourrait être améliorée. Le modèle d'apprentissage pourrait ainsi prendre en compte une liste de termes que nous ne souhaitons pas inclure parmi les descripteurs (*stop words*). Ceci est en effet applicable pour toutes les catégories, mais il aurait été utile de préciser une liste de termes pour une seule catégorie, de sorte que cette suppression n'affecte pas les descripteurs des autres catégories pour lesquelles elle serait pertinente. Par exemple, le terme *client* crée du bruit dans la catégorie Big Data, mais peut être pertinent pour une autre et ne peut donc pas être supprimé pour toutes les catégories. Concernant la hiérarchie de concepts, il serait utile d'ajouter un seuil minimum d'apparition des concepts pour classer un document comme traitant du Digital Working. Cela permettrait d'éviter qu'un document soit catégorisé alors qu'il ne contient qu'un concept et son contexte, car il s'agit de la cause principale d'erreurs de classification pour cette cartouche.

Enfin, l'utilisation d'une méthode de construction automatique de thésaurus pourrait permettre de faciliter sa création, en appliquant un algorithme capable de sélectionner des concepts par rapport à leur contexte proche et de les organiser hiérarchiquement en conséquence [Claveau *et al.*, 2014]. Aussi, une piste pour réduire le bruit causé par les concepts apparaissant dans le contre-corpus consisterait à restreindre l'indexation effectuée par la cartouche à une zone précise du document, par exemple le titre et le chapeau des articles de presse. En effet, comme ils permettent de situer le sujet abordé dans les articles, les termes les plus significatifs d'une catégorie auraient tendance à apparaître plus souvent dans cette zone, tandis que les termes causant le bruit ont plus de chance d'apparaître au cœur de l'article. Il faudrait cependant prendre en compte les cas où les articles ne possèdent pas de chapeau, par exemple en les indexant avec les *n* premiers mots de l'article.

Bibliographie

N.S. Altman, An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, *The American Statistician*, Vol. 46, No. 3, pp. 175-185, Août 1992.

K. Englmeier, G. Hubert, J. Mothe, Classification automatique de textes basée sur des hiérarchies de concepts, *Veille stratégique, scientifique et technologique*, Barcelone, 2001. < <http://www.irit.fr/~Josiane.Mothe/pub/VSSST01.pdf>>.

J. Beney, Classification supervisée de documents : théorie et pratique, *Hermes Science*, février 2008, 184 p.

B. Buschbeck, L. Grivel, S. Guillemin-Lanne, C. Lautier, Une Application Industrielle d'Extraction de l'Information pour l'Intelligence Economique, *EGC 2002 Extraction et Gestion des Connaissances*, 15 p. 2002.

V. Claveau, E. Kijak, O. Ferret, Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. *21^{ème} conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Marseille, France, 12 p, 2014.

T. Finley, Supervised Clustering with Support Vector Machines, *Department of Computer Science*, Cornell University, Ithaca, NY., 8 p, 2005.

N. Hernandez., J. Mothe, Ontologies pour l'aide à l'exploration d'une collection de documents. *Veille Stratégique Scientifique & Technologique Systèmes d'information élaborée*, Bibliométrie, Toulouse, 12 p, 2004.

M. Ikonomakis, S. Kotsiantis, V. Tampakas, Text Classification Using Machine Learning Techniques, *WSEAS Transactions on Computers*, Issue 8, Volume 4, pp 966-974, 2005.

K. Karteeka Pavan, A. Appa Rao, A.V. Dattatreya Rao, An automatic clustering technique for optimal clusters, *International Journal of Computer Science, Engineering and Applications*, Vol.4, No.4, 2011.

L. Kervers, Indexation semi-automatique de textes : thésaurus et transducteurs. In *Actes de CORIA09 (Sixième Conférence Francophone en Recherche d'Information et Applications)*, Presqu'Île de Giens, France, pp 151-167, 2009.

S. Laroum, N. Bechet, H. Hamza, M. Roche, Classification automatique de documents bruités à faible contenu textuel, *Revue Nouvelles Technologies de l'Information*, Hermann, E-18 (Numéro spécial : Fouille de Données Complexes), 25 p, 2010, <lirmm-00394668>.

Y. MA, L'intégration du thésaurus dans le traitement de la catégorisation automatique, *Institut National des Langues et des Civilisations Orientales (INALCO)*, mémoire de Master 2, 59 p, 2014.

X. Polanco, Text Mining et intelligence économique : aujourd'hui et demain, *Colloque Veille Technologique, Intelligence Economique et Bibliométrie*, Université Catholique de Louvain-la-Neuve, Belgique, 9 p, 23-24 janvier 2001.

T. Slimani, B. Ben Yaghlane, K. Mellouli, Une extension de mesure de similarité entre les concepts d'une ontologie, *4th International Conference : Sciences of Electronic, Technologies of Information and Telecommunications*, Tunisie, 10 p, 25-29 Mai 2007.

K. Spärck Jones, A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, 28, 1972, pp 11-21. (Réimprimé dans le *Journal of Documentation*, 60, pp 493-502, 2004).

K. Sureshkumar, M. Umadevi, N.M. Elango, Divisive Clustering method using Naive Bayes Algorithm for Text Categorization, *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 2, Issue 4, 2013.

