Institut National des Langues et des Civilisations Orientales (INaLCO)

Département TIM (Textes, Informatique, Multilinguisme)

L'intégration du thésaurus dans le traitement de la catégorisation automatique

Domaine : Traitement automatique de langues naturelles

Grade : Master

Parcours : Ingénierie Multilingue

Auteur : Yingying MA

Directeur de mémoire : Cyril Grouin

Directeur de stage : Michel BERNARDINI

Présenté et soutenu le 21 novembre, 2014

REMERCIEMENT

En préambule à ce mémoire, je tiens à remercier, l'équipe pédagogique ainsi que les différents intervenants professionnels de la formation l'Ingénierie linguistique de l'INALCO d'avoir assuré mon apprentissage théorique.

Je tiens à exprimer ma reconnaissance à Monsieur Cyril qui en tant que directeur de mémoire a su être à l'écoute et le remercie pour l'aide et le temps qu'il a bien voulu consacrer à l'élaboration de ce mémoire.

Je souhaite adresser mes remerciements les plus sincères au Monsieur Michel Bernardini, mon tuteur, Responsable Informatique et Communication - BNP Paribas, CIB Etudes Economiques pour m'avoir accordé sa confiance, pour l'ensemble de ses conseils et sa disponibilité tout au long de ce stage. Il a largement contribué à l'élaboration de ce mémoire grâce aux informations qu'il a pu me transmettre.

Je tiens également à remercier toute l'équipe Leonard, notamment Paul Regnard, Chargé d'Information, Communication - Veille et les différents stagiaires pour leur générosité, leur partage d'expérience et leur coopération professionnelle tout au long de cette année d'apprentissage.

ENGAGEMENT DE NON PLAGIAT

Je soussignée MA Yingying certifie sur l'honneur que les travaux soumis en mon nom dans ce mémoire sont le fruit de mes propres efforts et réflexions personnelles et que toute idée ou tout document utilisé pour étayer ce travail et ne constituant pas une réflexion personnelle, est en conséquence, citée en référence.

Signature :

Yingying MA

Résumé

Ce mémoire étudie le projet « catégorisation automatique » sur la plateforme LEOnard du département Etudes économique de BNP Paribas. Ainsi il étudie le schéma du site LEOnard afin de comprendre les caractéristiques du corpus à catégoriser et afin de trouver la meilleure stratégie. Cependant, la richesse et la variété d'information rendent la catégorisation automatique plus complexe. Après avoir examiné l'approche de l'apprentissage automatique, nous nous rendons compte que cette approche n'est plus suffisante. Nous avons introduit le thésaurus dans la chaîne de traitement afin d'affiner l'extraction de termes dans le corpus. En calculant le score de termes dans le thésaurus, nous avons réussi à obtenir un résultat très satisfaisant.

Mot clé :

Apprentissage automatique, clustering, catégorisation automatique, thésaurus

Sommaire

Contenu

Liste de Figures

Liste de tableaux

1. Introduction

La technologie du web a complétement changé le mode de communication et la consommation de l'information à travers le monde. Cependant, elle génère également une multiplicité de l'information et rend plus complexe le moyen d'accéder à un contenu de qualité qui intéresse les utilisateurs. C'est une opportunité et à la fois un grand défi pour les individus comme pour les entreprises. D'une part, nous avons beaucoup d'informations qui restent à notre disposition et qui nous fournissent un grand panel de sources afin de prendre une décision, connaitre les dernières nouvelles etc., d'autre part, il n'est pas facile pour des internautes de trouver des informations de qualité et utiles avec une simple recherche de mots-clés. Pour optimiser les recherches d'information et la transformation digitale des entreprises, nous avons vu apparaitre ces dernières années la naissance du web sémantique et le Big Data.

Aujourd'hui, les éditeurs de logiciel proposent sur le marché les outils pour surveiller le web qui intègrent des crawleurs. Ces outils permettent aux entreprises d'obtenir de l'information de qualité. L'idée originelle étant de pouvoir présenter les contenus les plus récents et les plus pertinents dans certains domaines. Avec le grand volume d'informations, et même de données provenant du Big Data pour certaines d'entre elles, qui arrivent chaque jours sur leurs sites, classifier et catégoriser les documents afin d'en faciliter l'accès et la recherche devient un vrai enjeu pour les entreprises. Quelles que soient leur taille toutes les entreprises sont confrontées à ce problème. La solution actuelle en matière de classification automatique repose sur l'apprentissage automatique (apprentissage automatique).

Dans ce mémoire, nous réalisons une étude de catégorisation automatique d'articles de presse fondée sur un portail d'informations qui intègre des outils de Fouille de texte pour construire un corpus et récupérer des données dont nous avons besoin. Après une série de sélections, nous avons choisi un site de BNP Paribas qui s'appelle LEOnard. C'est un intranet au sein du groupe BNP Paribas. Nous avons choisi cette plateforme non seulement parce qu'il y a un grand volume de documents mais aussi parce que les thématiques sont très variées et touchent de nombreux domaines d'activité.

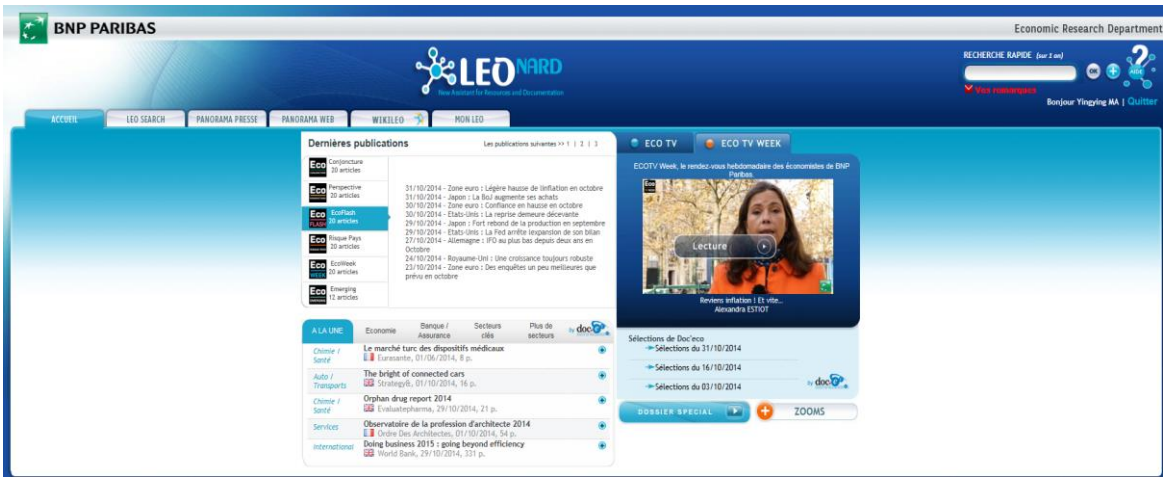


Figure 1 Page d'accueil de LEOnard

En tant que leader de la zone euro dans les services bancaires et financiers avec 4,8 Md€ de résultat net part du Groupe en 2013, BNP Paribas se préoccupe de l'importance de la qualité de ses données et de structurer les documents et leurs recherches. En 2004, le projet LEOnard a été lancé. L'idée originaire était de pouvoir présenter l'information de trois sources distinctes: l'information produite au sein de BNP Paribas, l'information issue de la presse et l'information qui vient du web (crawl). Cette plateforme s'appuyait sur deux technologies tierces :

- Un moteur de recherche statistique (PolySpot)
- Un outil de surveillance automatique de site web (KB Crawl).

La finalité de notre étude consiste à apporter une compétence en traitement automatique des langues pour répondre aux besoins de classifications des informations destinées aux analystes de BNP Paribas. Dans cette finalité, notre travail s'inscrit dans la plateforme LEOnard développée par BNP Paribas. L'idée de cette étude est de comprendre comment catégoriser automatiquement les informations de la presse pour les présenter d'une manière structurée et compréhensible pour les analystes de BNP Paribas. A partir de ce constat, nous analyserons la solution la plus pertinente et la plus facile à manipuler pour catégoriser les documents automatiquement.

2. Etat de l'art

2.1 Notre problématique

L'explosion de données et d'informations du web a demandé une catégorisation très robuste et très pertinente. Dans une étude [Augé et al., 2003], les chercheurs ont mis en place une méthode de apprentissage automatique (apprentissage automatique). Ils ont utilisé une machine SVM (Support Vector Machine) pour réaliser l'apprentissage automatique. Ce type d'algorithme a montré son efficacité aussi bien dans le cas de catégorisation plane (Joachims, 1998), que dans le cas de catégorisation hiérarchique [Dumais, 2000]. Dans le domaine de Fouille de texte, c'est une solution très courante pour réaliser la catégorisation automatique. L'idée principale est de segmenter les documents en mots. Ensuite, nous allons calculer le poids de chaque mot dans une catégorie. Nous collectons tous les poids des termes représentatifs dans une catégorie pour définir le poids moyen d'une catégorie.

Cependant, face à la richesse et la variété d'information, cette méthode ne nous semble pas très performante. En 2003, [Tikk et al., 2003] nous a présenté une méthode de catégorisation. Ici nous voulons affiner le travail et intégrer un thésaurus dans la chaîne de traitement du « apprentissage automatique ». Différente de la première méthode, celle-ci va calculer les poids des termes du thésaurus dans un grand corpus. Nous en servirons pour faire l'extraction dans un document à catégoriser. Le document appartient à la catégorie où se présentent les termes les plus pondérés.

A la fin, pour évaluer les deux approches, nous allons comparer les résultats obtenus de ces méthodes.

2.2 Travaux réalisés

2.2.1 Approche « apprentissage automatique »

- la constitution du plan de classement ;
- l'annotation manuelle du corpus d'apprentissage ;
- la définition de caractéristiques linguistiques utilisées par l'algorithme d'apprentissage.

Ces opérations peuvent être chronophages ; le résultat n'est généralement applicable qu'au domaine particulier concerné par les catégories prédéfinies, et aux types de documents représentatifs du corpus d'apprentissage.

Cette approche est très utile pour identifier des documents thématiques. Dans notre étude, nous suivons également le même schéma de traitement. Cependant, son travail se concentre sur le traitement du corpus multilingue. Dans notre étude, nous ne travaillons que sur la langue française.

Une question se pose à propos des plans de classement, généralement défini pour un domaine particulier. Quel jeu d'étiquettes prédéfini serait suffisamment couvrant pour catégoriser d'une façon raisonnablement générique un texte de la presse quotidienne ? [Yun et al., 2011] a présenté une démarche très proche pour catégoriser et étiqueter les documents provenant du web. L'idée principale est de trouver la relation et la hiérarchie entre des catégories différentes.

Dans notre recherche, nous avons utilisé des descripteurs définis par les documentalistes de BNP Paribas pour construire le plan de classement des catégories. Les descripteurs se trouvent dans l'entête des documents du corpus dans la zone métadonnée. Ces documents sont stockés dans la base de données interne de BNP Paribas.

Pour générer le modèle de l'apprentissage automatique, nous devons effectuer quelques prétraitements du corpus. Dans l'étude de [Foucault et al., 2013], il nous a proposé un schéma de traitement suivant :

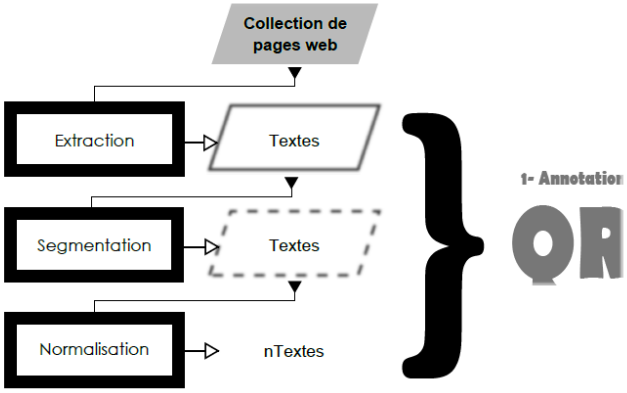


Figure 2 Prétraitement du corpus

A partir des pages web, il extrait et segmente des entités. Pour faciliter le traitement, il a également mis en place une phase de normalisation. Ce prétraitement nous permet de transformer les documents en mots afin que nous puissions calculer le poids de chaque terme.

Dans la thèse de [JALAM, 2003], pour répondre à la problématique d'apprentissage automatique et catégorisation de textes multilingues, il nous propose quelques différentes méthodes pour calculer le poids des termes dans le corpus :

1. Plus le terme t_k est fréquent dans un document d_j, plus il est en rapport avec le sujet de ce document.
2. Plus le terme t_k est fréquent dans une collection, moins il sera utilisé comme discriminant entre documents

Soient $\#(t_k, d_j)$ le nombre d'occurrences du terme tk dans le texte d_j, |T r| le nombre de documents du corpus d'apprentissage et #T r(t,k) le nombre de documents de cet ensemble dans lesquels apparaît au moins une fois le terme tk. Selon les deux observations précédentes, un terme tk se voit donc attribuer un poids d'autant plus fort qu'il apparaît souvent dans le document et rarement dans le corpus complet. La composante du vecteur est codée $f(\#(t_k, d_j))$, où la fonction f reste à déterminer. Deux approches triviales peuvent être utilisées. La première consiste à attribuer un poids égal à la fréquence du terme dans le document :

$W(k,j) = \#(t_k, d_j)$

et la deuxième approche consiste à associer une valeur booléenne :

$w(k,j) = 1$ Si $\#(t_k, d_j) > 1$ et $w(k,j) = 0$ Si $\#(t_k, d_j) < 1$.

Nous considérons cette méthode pertinente pour notre problématique dans la mesure où elle prend ne compte à la fois le poids du terme dans un document et le poids du même terme dans le corpus entier. Il s'agit de la même technique de calcul que sur notre machine. Ce calcul prend en compte à la fois le poids des termes dans un document et le poids des termes dans un corpus. Cela nous semble plus logique et pertinent.

2.2.2 Approche « thésaurus »

Pour améliorer la performance de la catégorisation, nous voudrons intégrer un thésaurus dans la chaîne de traitement.

Au niveau de la construction du thésaurus, [Bertels et al., 2012] ont intriduit la méthode Stable Lexical Marker Analysis (SLMA) « marqueurs lexicaux stables » ou analyse des marqueurs lexicaux stables [Speelman et al., 2006]. Le but était d'identifier les variantes lexicales régionales typiques ou les « marqueurs lexicaux stables » des différences régionales entre le néerlandais utilisé aux Pays-bas et en Flandre (Belgique) [Speelman et al., 2008]. Cette méthode sert à compare des listes de fréquence d'un corpus d'analyse à des listes de fréquence d'un corpus de référence. Les « marqueurs lexicaux stables » sont des mots qui sont spécifiques dans la plupart de ces comparaisons.

Cette théorie nous permet d'identifier les termes du thésaurus. Nous nous concentrons sur des termes spécifiques, stables dans le corpus et homogènes.

Pour le calcul du poids du thésaurus, [Bertels et al., 2012] nous avons proposé le principe suivant : l'analyse des mots-clés se caractérise par une approche contrastive, qui consiste à comparer les fréquences relatives des mots dans un corpus spécialisé à celles dans un corpus de référence de langue générale. Un mot est « clé » ou spécifique dans le corpus spécialisé si sa fréquence relative dans ce corpus est plus élevée que sa fréquence relative dans le corpus de référence et si la différence de fréquence est statistiquement significative.

L'étude de [Kevers,2009] a déjà réalisé une indexation de textes avec thésaurus. C'est souvent le thésaurus qui est utilisé car il représente un bon compromis entre puissance descriptive et complexité acceptable du point de vue du développement et de la maintenance. Un thésaurus est un vocabulaire contrôlé qui regroupe un ensemble de concepts relatifs à un certain domaine. Il constitue un moyen de décrire ce domaine, d'en définir les concepts et de fixer la terminologie utilisée par un groupe de personnes.

Nous trouvons que dans son étude, le calcul de pondération des termes dans le thésaurus est très intéressant. Il a appliqué un calcul de TF-IDF(term frequency-inverse document frequency). Cette mesure est couramment utilisée pour évaluer le poids d'un terme par rapport à un corpus donné. Ce score de base sera éventuellement modifié pour déterminer le score final d'une expression. Les formules appliquées sont :

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \qquad idf_i = log\frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

où nij est la fréquence d'un terme i dans le document d_j, |D| étant le nombre de documents dans le corpus et $|\{d_j : t_i \in d_j\}|$ le nombre de documents dans lesquels le terme i est présent. La valeur finale du TF.IDF est obtenue par : $tf.idf_{(i,j)} = tf_{(i,j)} * idf_{(i)}$.

3. Matériel et Méthodes

Le corpus et les données que nous utilisons pour mener la recherche viennent du site LEOard et de la base interne de BNP. Toutes ces ressources contiennent des documents qui appartiennent à plusieurs domaines.

3.1 Présentation des données et du corpus

3.1.1 Intégration des données sur LEOnard

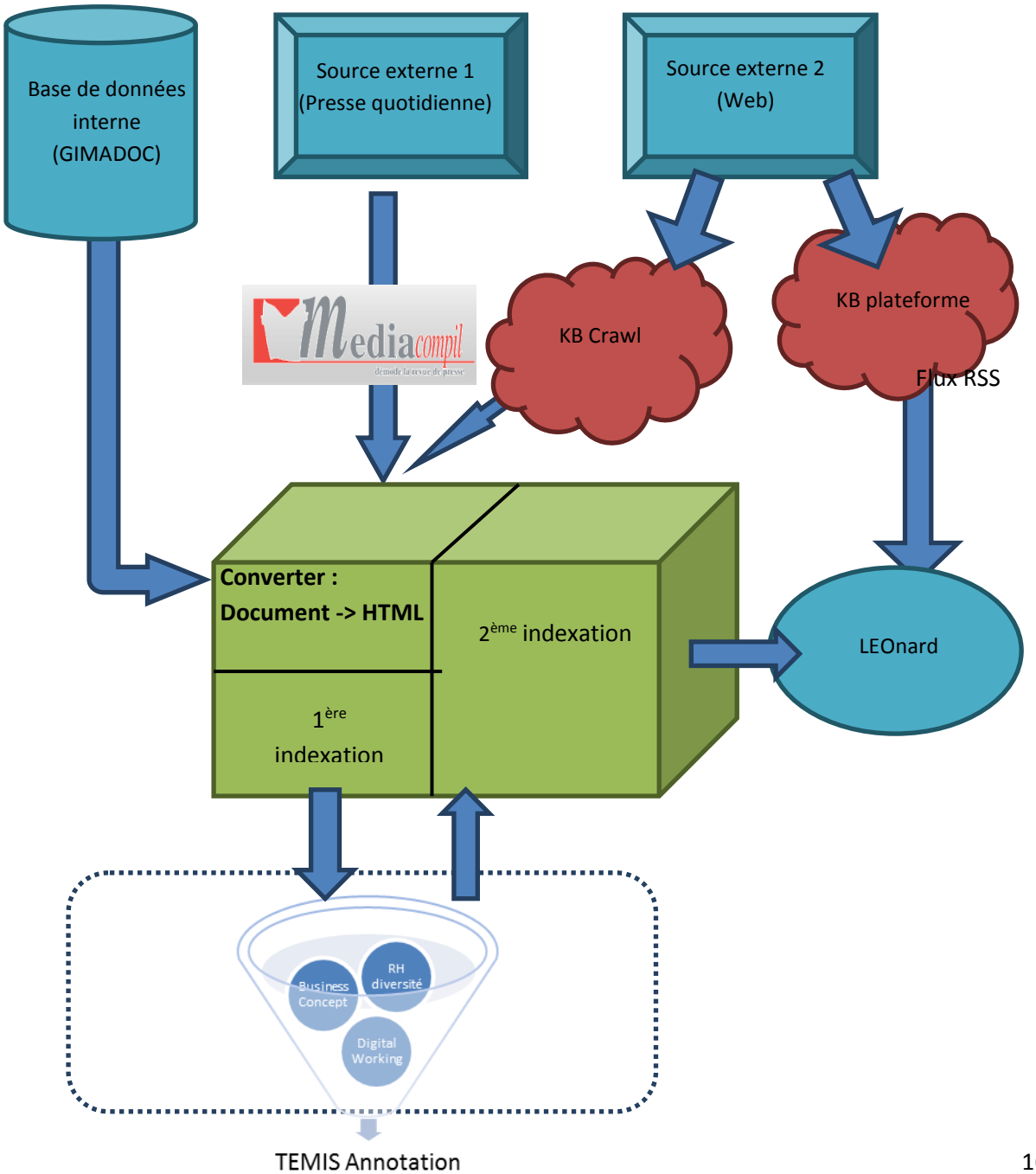


Figure 3 Architecture de LEOnard

La chaine de traitement de LEOnard se compose de cinq parties :

- **Les sources internes et externes**. Les sources internes sont la base de données GIMADOC alimentée par le centre de documentation du Groupe BNP Paribas. Ce sont des documents et des publications des Etudes Economiques et du Groupe Risque et Management. Les sources externes sont composées de deux parties : la page quotidienne numérisée et envoyée par le prestataire Mediacompil et les pages d'internet crawlées par l'outil de veille KB Crawl. KB Crawl génère également des flux RSS qui seront affichés directement sur LEOnard.

- **Les prestataires**. Il s'agit de deux prestataires : Mediacompil numérise la presse quotidienne et hebdomadaire et KB Crawl collecte les pages d'Internet sélectionnées par le webmaster.

- **Le moteur de recherche statistique Polyspot.** Les traitements réalisés par Polyspot est : convertir tous les formats en HTML et indexer les documents entrants.

- **Les cartouches de connaissances de Fouille de texte.** Cette couche de Fouille de texte comprend plusieurs cartouches de connaissances (Coupet et al. 205). Cette technique s'appuie sur l'extraction et l'analyse sémantique. Certaines cartouches extraient des entités nommées ; certaines cartouches extraient des documents pour un domaine spécifique[1]. Certaines cartouches effectuent la catégorisation automatique.

- **Le site LEOnard**. Les documents après l'indexation de PolySpot seront présentés sur LEOnard.

Le flux de traitement contient les étapes suivantes :

- Les sources internes Gimadoc arrivent sur PolySpot et sont indexées par PolySpot ;

- Les sources externe de la presse quotidienne numérisées par MediaCompil sont indexées par PolySpot ;

- Les pages crawlées par KB Crawl sont indexées par PolySpot ;

[1] L'extraction de documents thématiques est complémentaire de la catégorisation. Pour certains concepts (Digital Working par exemple) qui n'existe pas dans les catégories définies, on a créé des cartouches pour classifier les documents

- Toutes les sources indexées sont envoyés à la couche de l'annotation pour effectuer l'extraction ;

- Les résultats sont renvoyés à Polyspot pour ré-indexé par PolySpot. C'est ainsi que les documents paraissent sur la plateforme LEOnard.

Notre travail concerne la partie couche de connaissances de Fouille de texte (la partie dans le cadre de l'image ci-dessus). Le résultat de notre travail de catégorisation automatique sera présenté sous forme de cartouche de connaissances.

3.1.2 Caractéristiques des données et du corpus

La construction de corpus de référence a été effectuée sur les documents de GIMADOC. Dans GIMADOC, un document est composé de deux parties : une partie métadonnées qui contiennent les descripteurs attribués au document et le titre du document; Une autre partie contient le contenu textuel du document. (Voir l'image ci-dessous)

Figure 4 corpus de référence

Dans notre étude, avec un script Perl (voir annexe 7.2), nous avons récupéré 28 600 documents (soit 1.35Go) de la presse quotidienne en langue française dans la base de données GIMADOC. Tous les documents sont encodés en UTF-8. Ici, seulement les documents dont la taille est au-dessus de 4ko (environs 450 mots) sont conservés. Nous avons distribué les documents dans des catégories différentes en fonction de leurs descripteurs associés. Le

corpus construit servira à tester toutes les deux approches de catégorisation automatique. Pendant la récupération de données, nous avons remarqué quelques bruits qui pourraient avoir une influence importante sur le résultat de notre recherche.

- Il y a certains documents qui sont mal numérisés et qui contiennent des signes non français. Il en y a d'autres qui sont segmentés en lettre. Cette erreur vient principalement du processus de numérisation.

Figure 5 Document mal numérisé

- Il y a certains documents qui sont mal segmenté. Dans certains documents originaux, les textes sont présentés en plusieurs colonnes sur une page. Pendant la numérisation et la conversion en HTML, c'est difficile pour les outils de rassembler les mots qui avaient découpés en deux. C'est difficile pour les outils de gérer les césures.

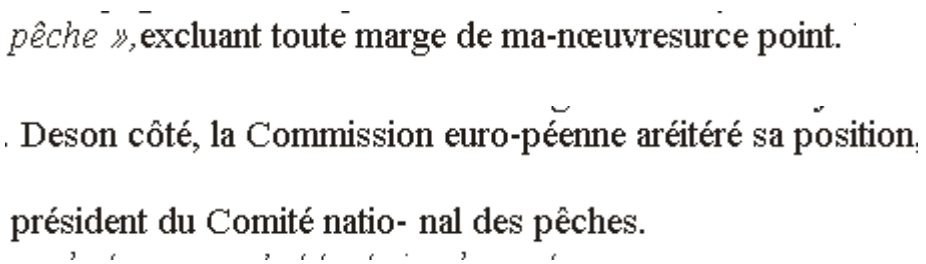


Figure 6 Phrase mal segmenté

3.2 Clustering[2] des données

Pour classifier les documents, nous nous servons des descripteurs qui ont été choisis par les documentalistes lors de la rentrée des documents dans la base de données GIMADOC.

Ensuite, notre travail est de définir les hiérarchies de toutes les catégories et leurs sous-catégories. L'objectif étant que ces hiérarchies soient assez claires et intuitives pour ceux qui ne possèdent pas de connaissances spécifiques des domaines. Nous avons effectué des traitements suivants :

- définir des champs lexicaux pour regrouper des descripteurs ;
- supprimer des catégories trop fines (voir l'image Figure 7)

La première étape est de définir ce qui nous intéresse. Pour faire cela, nous avons travaillé étroitement avec le centre de documentation de BNP Paribas. Pour étudier les descripteurs (mots-clés) associés aux documents, et les descripteurs les plus saisis par les documentalistes, nous avons regroupé des descripteurs afin de définir le plan de classement de tous les domaines d'activités utiles aux utilisateurs de LEOnard. [3]Dans l'image ci-dessous, nous vous présentons un petit aperçu de ce que nous avons trié. Les mots « Presse / Presse gratuite/ Presse professionnelle/Presse Quotidienne » vont tous dans la catégorie « MEDIA ». Pour certaines catégories, elles contiennent des sous-catégories. Les mots « Micro-informatique », « Micro-ordinateur » et « Portable » sont dans la sous-catégorie « Matériels » qui sont sous « Matériels-Logiciels-Equipement ». « Matériels-Logiciels-Equipement » elle-même n'est pourtant pas la catégorie racine. Elle est une catégorie du 2ème niveau dont la racine est « High-Tech ».

Figure 7 Regroupement des descripteurs

[2] Le clustering de données est une classification de données en

[3] Le fichier qui met en relation toutes les descripteurs avec les catégories se trouve dans **Annexe 5.1 (ajout de hyper lien)**

Nous manquons de connaissances pour regrouper les descripteurs. Le travail de réorganisation est effectué par les analystes de BNP Paribas. Dans l'image ci-dessous, nous voyons que le descripteur « Presse professionnelle » et le mot « Presse quotidienne » n'appartiennent pas au même niveau. « Presse professionnelle » s'agit plutôt de la presse de différents domaines et secteurs, alors que « Presse quotidienne » est au niveau de la fréquence de publication.

La seconde étape est d'affiner notre travail. Nous supprimons des catégories servant à annoter moins de 30 articles. Par exemple : Sport et Safety Defense (Voir la Figure 7). Nous considérons que ces catégories-là sont trop spécifiques et qu'elles ne sont pas assez présentes dans nos ressources.

Ainsi, nous obtenons un plan de classement de 4 niveaux, qui couvre presque tous les niveu. Il contient 8 grandes catégories :

- INDUSTRY
- BANKING-INSURANCE
- FOOD INDUSTRY-RETAILING-LUXURY-FASHION
- HIGH-TECH
- ENERGY-ENVIRONMENT
- MEDIA-ADVERTISING
- BUILDING CIVIL ENGINEERING-REAL ESTATE
- SERVICES-LEISURE ACTIVITIES[4]

[4] Pour être cohérent avec l'interface de LEOnard, nous avons traduit le plan en anglais.

Figure 8 Plan de classement sectoriel

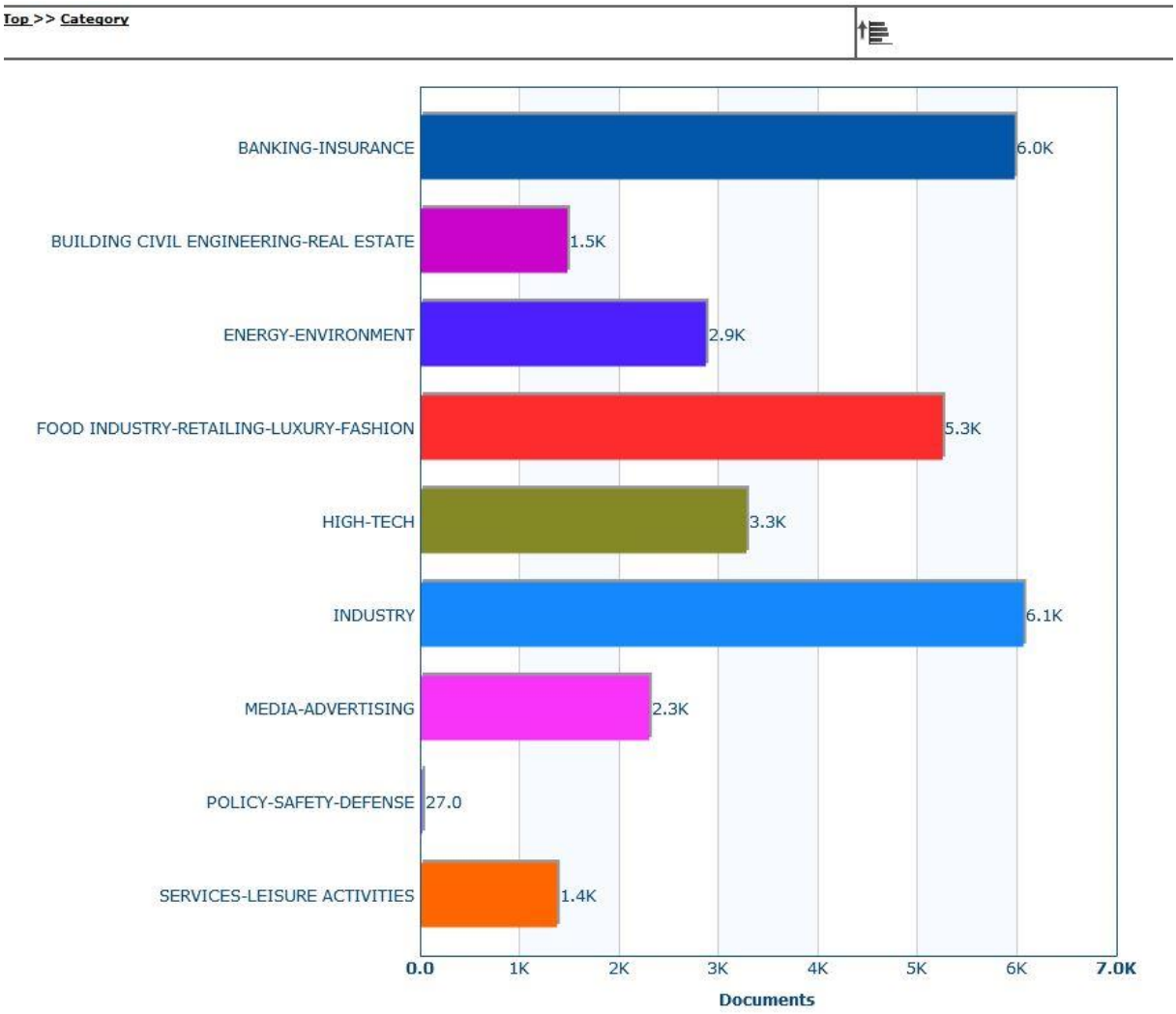


Figure 9 Le nombre d'articles associés à chaque catégorie

L'image au-dessus nous présente la répartition des documents dans le corpus français. Plus il y a des données dans cette catégorie, plus le résultat dans cette catégorie sera significatif. En conséquence, l'enrichissement du corpus nous permet de nous assurer d'avoir des résultats significatifs pour la suite de traitement. La catégorie Policy-Safety-Defense est supprimée car il y a trop peu de document collecté dans cette catégorie.

3.3 Méthode Apprentissage automatique statistique VS Méthode Thésaurus

Dans cette partie, nous voulons examiner en détail les deux approches de catégorisation automatique. L'intérêt est de savoir si l'introduction du thésaurus pourra augmenter la performance de la catégorisation.

3.3.1 Apprentissage automatique statistique

L'approche Apprentissage automatique se fait via l'outil Category Workbench, qui est une machine de « vecteur sémantique ». Le processus global est constitué des étapes suivantes :

- La machine extrait les termes dans un document catégorisé, les lemmatise et associe les étiquettes morphosyntaxiques
- Pour chaque lemme extrait, elle calcule sa fréquence[5]
- Chaque document est représenté comme un « vecteur sémantique » qui contient l'ensemble des couples terme/fréquence
- La machine prend en compte les types de descripteurs sélectionnés par l'utilisateur pour l'apprentissage
- La machine apprend la représentation de chaque catégorie en calculant une sorte de vecteur moyen de la catégorie
- Elle stocke tous les vecteurs moyens dans un modèle qui est prêt à assigner à d'autres documents inconnus.

3.3.1.1 Prétraitement du corpus

Pour commencer le processus, nous avons besoin d'effectuer un nettoyage et une vérification manuelle de la classification dans le corpus. L'apprentissage se fait sur un corpus de référence. C'est-à-dire que un corpus déjà préalablement catégorisé.

Dans le chapitre 3.2, nous avons classifié des descripteurs qui se trouvent dans la métadonnée à la tête des documents du corpus. D'après le plan de regroupement, nous avons élaboré un script [6]en Perl. La fonctionnalité de ce script est de consulter la base de données GIMADOC,

[5] La fréquence du mot est le nombre d'occurrence du terme dans le document
[6] Voir **Annexe(insérer hyperlien)**

récupérer des documents qui contiennent les descripteurs existant dans notre plan et les classer d'après le plan de classification.

En sortie, ce script nous a fourni un fichier TMX[7] qui contient les informations suivantes :

- L'identifiant du document avec la balise <doc id="XX">
- Le titre du document avec la balise <dc:title>
- La catégorie que nous avons associée au document de référence. Elle se trouve à l'intérieur de la balise <categories>
- Un groupe de descripteurs distribué au document au moment où il est rentré dans la base de données ces descripteurs sont entourés par <features>
- La dernière information concernant le document est le chemin du document. La machine peut aller chercher ce document avec ce chemin indiqué.

L'image ci-dessous nous montre un extrait du corpus d'apprentissage.

Ensuite, une fois nous avons récupéré ces documents de la base de données, nous allons effectuer le traitement de l'annotation.

```
<?xml version="1.0" encoding="UTF-8"?>
 <tm xmlns:dc="http://purl.org/dc/elements/1.1/">
<doc id="68">
<dc:title>68</dc:title>
<categories><c>ENERGY-ENVIRONMENT/ENERGY/OIL</c></categories>
<features>
<ft f="1">/Metadata/COMPAGNIE PETROLIERE</ft>
<ft f="1">/Metadata/CONTRAT</ft>
<ft f="1">/Metadata/CHINE</ft>
<ft f="1">/Metadata/IRAN</ft>
</features>
<text><file format="html" path="D:\apps\Catego\Sources\Gimadoc\Corpus-Gimadoc-Francais-Anglais\30197543446122.htm"/></text>
</doc>
```

Figure 10 Corpus d'apprentissage

[7] TMX est un format XML spécial pour la machine Category Workbench

Dans Category Workbench, nous appelons la cartouche (Coupet, Buschbeck, Six, Cardoso, & Huot, 2005) Analitics2. L'annotation du prétraitement par Analytics2 consiste principalement à segmenter le texte, désambigüiser, distribuer les étiquettes morphosyntaxiques et préparer pour être calculé par Category Workbench.

La première étape est la conversion de format. Le format de notre corpus est HTML. La cartouche le convertit dans TMX.

La seconde étape consiste à identifier la langue de document. Cette cartouche contient 16 modules linguistiques qui lui permettent d'analyser 16 langues[8]. Elle décide d'applique le module correspondant d'après la langue identifiée du document.

La troisième étape est l'analyse profonde par XeLDA[9].

XeLDA est un moteur unique d'analyse linguistique. Il utilise la technologie XFST (Technologie des automates à états finis développée par Xerox). Il contient plusieurs éléments indépendants qui peuvent s'intégrer dans les applications linguistiques en fonction des besoins des utilisateurs. Parmi les différents services proposés par XeLDA figurent les services suivants (ceux que nous avons utilisés sont suivis d'une étoile *) [10] :

- Identification de langue,
- Segmentation de phrases (en unité lexicale élémentaire) *
- Tokenisation,*
- Analyse morphologique,*
- Désambiguisation syntaxique,*
- Extraction de termes*
- Interrogation de dictionnaires*
- Reconnaissance d'expressions idiomatiques*
- Analyse morphologique relationnelle (prototype)*

[8] Les 16 langues sont : allemand, anglais, espagnol, français, grec, hongrois, italien, néerlandais, polonais, portugais, russe, tchèque, norvégien, finlandais, suédois et danois.
[9]XeLDA : Xerox Linguistic Development Architecture.
[10] http://www.atala.org/XeLDA

La dernière étape de prétraitement est de stocker tous les documents dans une base de données à laquelle se connecte Categorie Workbench.

En sortie, le document est sous forme :

- /Term/ COMMON-NOUN /AA
- /Term/NP[11]/BB
- /Term/ADJ/CC
- /Term/Verb/DD
- /Term/PROPERNAME/EE

Et une la catégorie de référence :

- /Category/Finance

3.3.1.2 Apprentissage automatique statistique utilisant Category Workbench

Pour effectuer les traitements suivants, nous avons divisé aléatoirement ce grand corpus en deux : la première qui occupe 85% du grand corpus (soit 24 310 documents) est un corpus d'apprentissage. Il va contribuer à construire le modèle de catégorisation automatique. La deuxième partie du grand corpus qui contient 4290 documents est un corpus d'évaluation. Nous utilisons ce corpus pour évaluer la qualité du modèle de catégorisation.

3.3.1.2.1 Phase d'apprentissage

Dans l'apprentissage, tous les traitements sont effectués sur le corpus d'apprentissage que nous venons de créer. Nous calculons les fréquences de tous les termes présents dans le document. Nous transformons le document en un « vecteur sémantique » qui contient l'ensemble des couples terme/fréquence (en anglais : Feature/Frequency).

Par exemple : pour un document dont le contenu serait : « Ceci est un document qui parle de finance. Vu que le document parle de finance, le mot finance est répété », le vecteur sémantique de ce document est présenté dans le tableau suivant :

Feature	Frequency
/Term/COMMON-NOUN/document	2
/Term/VERB/voir	1

[11] NP : Noun Phrases (groupes nominaux)

/Term/VERB/parler	1
/Term/COMMON-NOUN/finance	3
/Term/PRONOUN/Ceci	1
/Term/VERB/être	2
/Term/VERB/répéter	1

Tableau 1 Vecteur Sémantique

Dans Category Workbench, pour optimiser les résultats, nous devons définir les paramètres suivants :

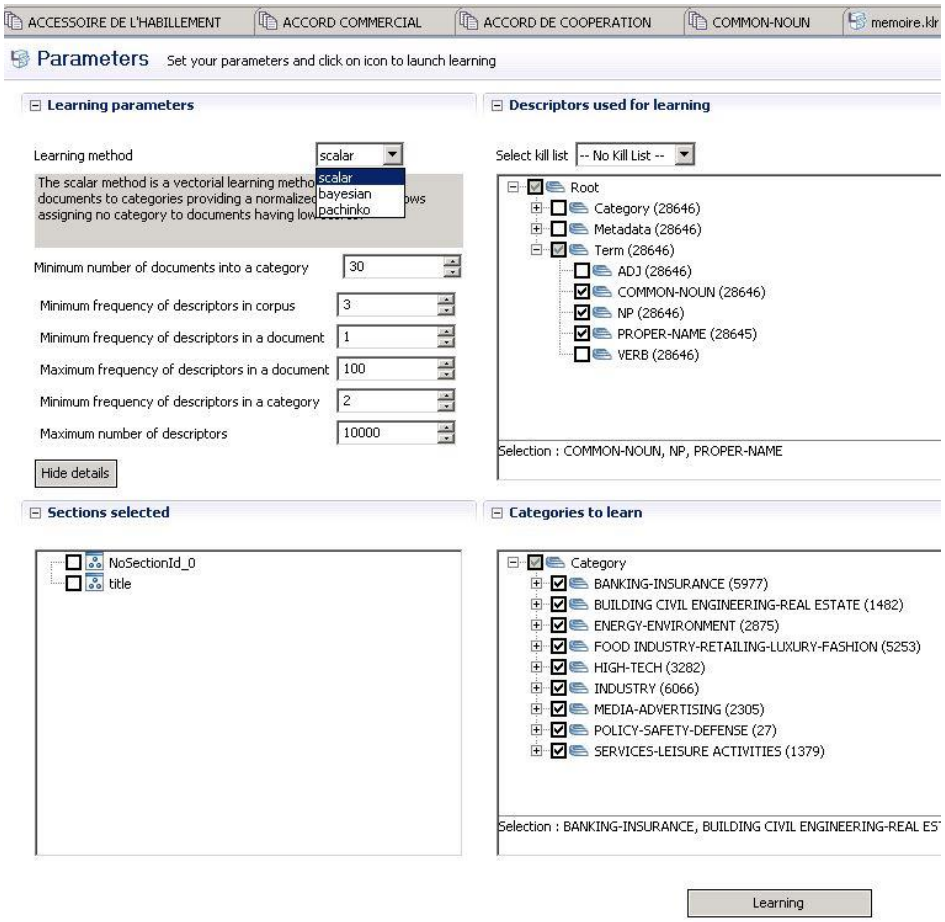


Figure 11 Paramètres d'apprentissage dans Category Workbench

- Nous choisissons la méthode d'apprentissage parmi les trois méthodes disponibles : Scalar, Bayesian et Pachinko.

  - Dans la méthode Scalar, toutes les catégories sont considérées comme plates (sans hiérarchie). C'est-à-dire qu'il y a un algorithme scalaire pour tous les niveaux de catégories.
  - Dans la méthode Pachinko, il y a un algorithme scalaire pour chaque nœud, y compris la racine. Cette méthode ne nous est pas très intéressante car elle

produit des bruits pendant l'assignement des catégories en croisant des nœuds de différentes branches.

- Pour la méthode Bayesian, nous ne le prenons pas en compte non plus car dans la machine, cette méthode n'utilise pas le score standard. Le résultat obtenu n'est pas très fiable.

Dans notre apprentissage, nous avons choisi Scalar comme l'algorithme d'apprentissage.

- Nous choisissons les types de descripteurs (dans l'outil, appelé « feature ») à prendre en compte pour l'apprentissage. Dans notre cas, nous ne conservons que les Nom Propre(NP) et nom commun (COMMON-NOUN) parce que les verbes et adjectifs sont souvent peu spécifiques aux thématiques.
- Le seuil minimum et maximum du nombre de fréquences de descripteurs dans un corpus. Nous prenons en compte seulement des termes dont les nombres de fréquences sont entre ces deux seuils. Cette méthode nous élimine les termes qui pourraient produire du bruit.

3.3.1.2.2 Calcul du vecteur moyen de la catégorie

La méthode d'assignement de catégorie à un document consiste à comparer le vecteur du document au vecteur de chaque catégorie. La catégorie avec le score inconnu sera attribuée à ce document. Dans notre cas, nous avons utilisé Category Workbench pour mettre en place un algorithme de scoring qui croise les scores des termes dans les documents avec ceux des documents dans un corpus de catégories. L'idée principale est de prendre en compte en même temps l'importance des termes spécifiques et celle des documents significatifs dans un corpus de catégorie.

Le processus de scoring contient trois étapes :

- Calculer les scores de termes ;
- Normaliser les scores de termes ;
- Combiner les scores de termes dans chaque document avec le score du document dans le corpus.

Le vecteur moyen d'une catégorie W(C) correspond à

$$W(C) = \begin{pmatrix} Term\ (i) \\ \dots\ \dots\ \dots \\ w(i,j) \\ \dots\ \dots\ \dots \end{pmatrix} Document(j)$$

Dont w (i,j) est le score de terme *i* dans le document *j* ; Document(j) est le score de document *j* dans le corpus *C* entier.

Pour calculer les scores des termes (soit w (i,j)), nous avons utilisé la méthode (Acherman, 2003), appelé « Smoothed-frequency scoring ». Considérons un document *j* et un terme *i*. S'il est présent dans *j*, son score égale à logarithme de sa fréquence, sinon ; son score égale à 0.

$$W(i,j) = \begin{cases} \log(1 + frequency(i,j)) & Si\ j\ existe\ dans\ i \\ 0 & Si\ j\ n'existe\ pas\ dans\ i \end{cases}$$

Cette méthode permet de normaliser les chiffres entre 0 et 1. Reprenons l'exemple du Tableau 1. D'après l'équation ci-dessus, le score de chaque mot dans le vecteur est :

Feature	Frequency	Weight
/Term/COMMON-NOUN/document	2	0.48
/Term/VERB/voir	1	0.30
/Term/VERB/parler	1	0.30
/Term/COMMON-NOUN/finance	3	0.60
/Term/PRONOUN/Ceci	1	0.30
/Term/VERB/être	2	0.48
/Term/VERB/répéter	1	0.30

Tableau 2 Vecteur avec le score des termes

Une étape de normalisation de score est réalisée pour permettre le fonctionnement des cartouches de connaissances, en utilisant l'équation suivante :

$$w_{(normalized)i}(j) = \frac{w_i(j)}{norm(i)} \text{ where } norm(i) = \sqrt{\sum_j w_i(j)^2}$$

Dont $w_{(normalized)\,i}(j)$ est le poids normalisé du terme *j* dans le document *i*; $w_i(j)$ *est* le poids du terme *j* dans le document *i* ; *norm(i)* est le facteur de normalisation pour un document *i* où le terme *j* est présent.

Le poids de document dans un corpus de catégorie dépend principalement de sa taille. Plus long un document, plus important il est dans le scoring. Le poids de document dans le corpus correspond à :

$w_i(C) = \log(1 + N(i))$

N(i) est la taille de document, qui correspond au nombre de terme dans le vecteur sémantique.

En conclusion, le calcul final de vecteur moyen de la catégorie contenant les scores des termes correspond à l'équation suivante :

$$w_{(smoothed-length)i}(j) = \frac{w_i(j) * \log(1 + N_i)}{norm(i)},$$

La machine va stocker l'ensemble des vecteurs sémantiques de toutes les catégories dans un fichier de modèle de catégorie.

L'image Figure 10 est un vecteur moyen représentatif pour la catégorie Industrie/*Pharmacie*. Les types de termes que nous avons choisis sont Nom commun et Nom propre.

L'image Figure 11 présente les termes représentatifs par catégorie.

/INDUSTRY/PHARMACY

Full name	Name	Score
/COMMON-NOUN/anticorps	anticorps	0,1025
/COMMON-NOUN/cament	cament	0,1025
/COMMON-NOUN/diabète	diabète	0,1025
/COMMON-NOUN/infection	infection	0,1025
/COMMON-NOUN/insuline	insuline	0,1025
/COMMON-NOUN/microbiologie	microbiologie	0,1025
/COMMON-NOUN/oncologie	oncologie	0,1025
/COMMON-NOUN/placebo	placebo	0,1025
/NP/anticorps monoclonal	anticorps monoclonal	0,1025
/NP/biologie moléculaire	biologie moléculaire	0,1025
/NP/laboratoire américain	laboratoire américain	0,1025
/NP/laboratoire suisse	laboratoire suisse	0,1025
/NP/laboratoire vétérinaire	laboratoire vétérinaire	0,1025
/NP/pharmacie médicament	pharmacie médicament	0,1025
/NP/phase II	phase II	0,1025
/NP/phase III	phase III	0,1025
/PROPER-NAME/Acomplia	Acomplia	0,1025
/PROPER-NAME/BMS	BMS	0,1025
/PROPER-NAME/Botox	Botox	0,1025
/PROPER-NAME/Jean-François De...	Jean-François Dehecq	0,1025
/PROPER-NAME/Jean-Pierre Garnier	Jean-Pierre Garnier	0,1025
/PROPER-NAME/Pfizer	Pfizer	0,1025
/PROPER-NAME/SANOFI AVENTIS	SANOFI AVENTIS	0,1025

Figure 12 Vecteur moyen pour catégorie Industrie/Pharmacie généré par Category Workbench

Trained documents | Filtered documents

Loading information ...

Title	BANKING-I...	BUILDING ...	ENERGY-E...	FOOD INDUST...	HIGH-TE...	INDUSTRY	M...	COMMON-NOUN	NP	PROPER-NAME
14846						PHARMACY		hausse; groupe;...	gestion agressif des ...	Bayer; Schering; Werner Wenning; r...
14864				FOOD INDUSTRY				marque; boisson...	pôle européen; marq...	Orangina Schweppes; Schweppes; O...
14868			GAS-ELEC...					énergie; courtier...	voie de le diversifica...	Michèle Assouline; Philippe Oddo; gro...
14898				FOOD INDUSTRY				tonne; productio...	côte d' ivoire; prévisi...	Londres; Nestlé; Laurent Pipitone; in...
14901				FOOD INDUSTRY				huile; soja; prix; ...	huile de soja; prix de...	Chicago; David Warne; Chine; USDA...
14932	PORTFOLI...							secteur; fonds; i...	private equity; ligne ...	Treasury Committee; méga-fond LBO...
14955					CONNEC...			télécom; entrepr...	Bouygues télécom; o...	Bouygues; Completel; Internet; Fran...
14967	SPECIALIZ...							crédit; foncier; c...	crédit foncier; caisse...	François Drouin; Nexity; Nicolas Méri...
14988						CAPITAL ...		entreprise; mach...	machine spécial; uni...	Inde; groupe Vallourec; coréen Hyun...
14998						AEROSPA...		acquisition; trés...	électronique de défe...	FINMECCANICA; Europe; PDG; Fran...
1501		CORPORA...						société; titre; ré...	Jacques de Chateau...	CBo Territoria; Jacques; Chateauvie...
15015						AEROSPA...		aéroport; opérat...	aéroport romain; év...	Macquarie Airports; Macquarie; MAC...
15026						CAR MAN...		euro; constructe...	euro fort; union eur...	Toyota; Honda; Europe; Japon; Var;...
15059						WOOD PA...		groupe; résultat...	carton ondulé; site d...	OPR; Otor; PSE; Otor Silesia; OTOR;...
15154	INSURANCE							groupe; santé; c...	chiffre d' affaire; tur...	Agrica; Hervé Bachellerie; groupe AG...
15174						CAR MAN...		modèle; États-U...	États uni; petit modè...	Nissan; Versa-Tiida; Japon; États; NI...
15207						CAPITAL ...		ascenseur; mod...	fédération des asce...	France; Europe; Processus; Marie Dc...

Figure 13 Termes représentatifs de catégories

3.3.1.2.3 Phase d'assignement

Pour évaluer la qualité du modèle généré dans le processus de apprentissage automatique par Category Workbench, nous appliquons notre modèle de catégorie sur un corpus d'évaluation pour évaluer le résultat de catégorisation. Chaque document dans ce corpus possède déjà une catégorie qui existe dans notre plan de clustering. Nous allons comparer la catégorie de référence qui est considérée comme la bonne catégorie avec la catégorie assignée par le modèle pour pouvoir avoir une idée de la qualité de cette méthode de catégorisation automatique.

Chaque document du corpus d'évaluation à annoter va, en sortie de l'extraction par Analytics2, pouvoir être représenté par son « vecteur sémantique ». Le prétraitement est le même que celui que nous avons présenté dans le chapitre « Phase d'apprentissage ».

Les scores de termes sont calculés de la même manière que les scores du corpus d'apprentissage. En sortie du calcul, chaque document du corpus d'évaluation est transformé également en vecteur sémantique avec un score associé à chaque terme.

Ce vecteur du document est ensuite comparé au vecteur moyen de chaque catégorie stocké dans le modèle de catégorie. Une distance est calculée entre 0 et 1 pour chaque catégorie.

La catégorie qui a le meilleure score, donc celle dont le document est le plus proche en terme de vecteur moyen sera retenu est attribuée au document.

Après avoir assigné le modèle au corpus d'évaluation, la machine nous a retourné un résultat en classifiant les documents en trois sortes : correct, missed (en français : raté) et false (en français : incorrect).

Les images sont les résultats d'assignement du modèle de catégorie analysés par l'outil Category Workbench.

3.3.2 Méthode Thésaurus

Pour compléter la méthode de l'apprentissage automatique. Nous intégrons un thésaurus dans la machine. Nous voudrons savoir si l'introduction du thésaurus avec la méthode traditionnelle de l'apprentissage automatique va augmenter la qualité de la catégorisation automatique.

3.3.2.1 Construction des thésaurus

L'idée principale de la méthode du thésaurus est de catégoriser un document avec les termes dans le thésaurus. Chaque terme dans le thésaurus appartient à une catégorie. En paramétrant la manière d'extraction des termes dans un document, nous voulons que les mots extraits par le thésaurus dans un document correspondent à une ou deux catégories la (les) plus pertinente(s) pour le document.

Cette approche se fait via l'outil Annotation Workbench. C'est une plateforme de l'évaluation de qualité de l'annotation. Elle nous permet de modifier les termes du thésaurus, la hiérarchie du thésaurus et d'évaluer la qualité de la catégorisation automatique.

Cette méthode ne demande pas de prétraitement du corpus. En revanche, nous devons construire un thésaurus avant de commencer le traitement. Le processus global de catégorisation par thésaurus contient les étapes suivantes :

- Construire le thésaurus ;

- Importer le thésaurus dans Annotation Workbench ;

- Injecter le thésaurus dans la cartouche STF (Smart Taxonomy Facilitator) [12];

- Lancer le Learning du thésaurus sur un corpus général ;

- Annoter le corpus de référence avec le paramétrage défini de la cartouche STF

[12] http://www.temis.com/fr/luxid-skill-cartridge

Pour construire le thésaurus, nous avons collecté tous les descripteurs présents dans les documents en classifiant d'après le plan de clustering que nous avons défini dans le chapitre « Clustering des données ».

Nous essayons d'élaborer un thésaurus en format « SKOS ». SKOS est l'acronyme de Simple Knowledge Organization System est une recommandation du W3C depuis 2009. Il s'agit d'un vocabulaire RDF qui fournit un modèle commun pour partager et lier sur le web différents systèmes d'organisation des connaissances tels que thésaurus, taxinomies, système de classification, système d'index [Gandon et al., 2012]. Cette technologie permet de formaliser les connaissances et de les présenter d'une manière thématique et hiérarchique. Les vocabulaires principaux de SKOS sont présentés dans l'annexe 7.1.

```
<?xml version="1.0" encoding="UTF-8"?>

<rdf:RDF
    xmlns:skos ="http://www.w3.org/2004/02/skos/core#"
    xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:xsd ="http://www.w3.org/2001/XMLSchema#">

<skos:Concept rdf:about="#Banking-Insurance">
    <skos:prefLabel xml:lang="fr">Banking-Insurance</skos:prefLabel>
</skos:Concept>
<skos:Concept rdf:about="#INDUSTRY">
    <skos:prefLabel xml:lang="fr">INDUSTRY</skos:prefLabel>
</skos:Concept>
<skos:Concept rdf:about="#FOOD INDUSTRY-RETAILING-LUXURY-FASHION">
    <skos:prefLabel xml:lang="fr">FOOD INDUSTRY-RETAILING-LUXURY-FASHION</skos:prefLabel>
</skos:Concept>
<skos:Concept rdf:about="#Finance">
    <skos:prefLabel xml:lang="fr">HIGH-TECH</skos:prefLabel>
</skos:Concept>
<skos:Concept rdf:about="#Societe">
    <skos:prefLabel xml:lang="fr">ENERGY-ENVIRONMENT</skos:prefLabel>
</skos:Concept>
<skos:Concept rdf:about="#Entreprise">
    <skos:prefLabel xml:lang="fr">MEDIA-ADVERTISING</skos:prefLabel>
</skos:Concept>
<skos:Concept rdf:about="#BUILDING CIVIL ENGINEERING-REAL ESTATE">
    <skos:prefLabel xml:lang="fr">BUILDING CIVIL ENGINEERING-REAL ESTATE</skos:prefLabel>
</skos:Concept>
<skos:Concept rdf:about="#SERVICES-LEISURE ACTIVITIES">
    <skos:prefLabel xml:lang="fr">SERVICES-LEISURE ACTIVITIES</skos:prefLabel>
</skos:Concept>
```

Figure 14 1er niveau du Thésaurus

Nous avons effectué un traitement sur le corpus pour supprimer tous les termes qui sont trop ambigus ou qui sont trop peu présents dans le corpus de référence.

Après le nettoyage, le thésaurus contient 4041 termes. (Voir l'image Figure 14)

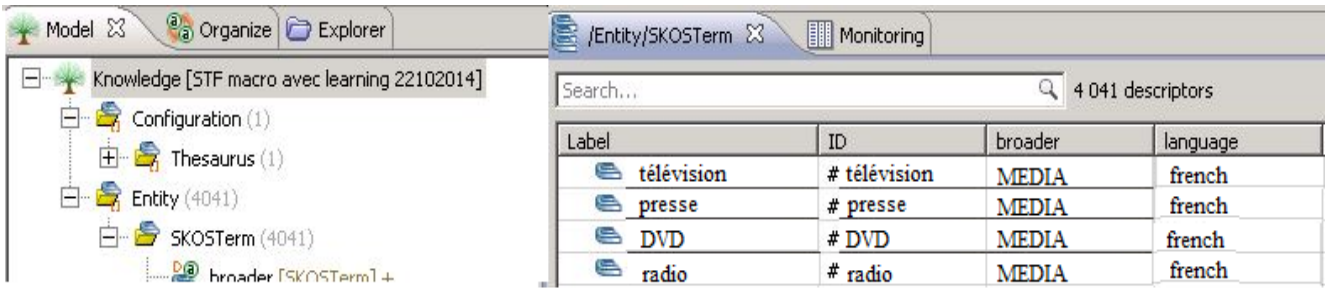


Figure 15 Thésaurus vu dans Annotation Workbench

3.3.2.2 Calcul d'IDF des termes du thésaurus

Dans le thésaurus, il y a des termes qui viennent de la presse quotidienne et qui ne sont pas assez spécifiques pour catégoriser un document. Il y a encore des termes qui sont rarement utilisés par les auteurs d'aujourd'hui. C'est-à-dire que dans le thésaurus, la pondération des termes est très variée. Nous voulons que les termes spécifiques possèdent un poids plus lourd que les termes courants et que ces termes aient une priorité dans l'extraction.

Pour faire cela, nous effectuons un calcul de « IDF » (Inverse Document Frequency). C'est une méthode de pondération classique. Avec cette mesure statistique, le poids d'un terme est inversement proportionnel à son nombre d'occurrences dans un « corpus général ». Un corpus général est différent qu'un corpus de référence ou qu'un corpus d'apprentissage. C'est un corpus qui ne contient pas de documents appartenant aux catégories que nous avons définies précédemment dans le chapitre 3.2 « Clustering des données ». Nous considérons que ces documents sont inintéressants. L'idée principale est : Si un terme est très fréquent dans le « corpus général », il ne nous est pas très intéressant car il n'est pas suffisamment spécifique pour définir une catégorie. Dans ce cas, nous allons baisser son poids dans notre corpus pour qu'il ne soit pas privilégié dans l'extraction. Dans notre calcul, les termes possédant des poids significatifs sont les termes qui sont très rares dans le corpus. En revanche, l'information qui est très dans les documents du corpus général aura un poids léger, même zéro s'elle est présente dans tous les documents du corpus.

Considérons un terme *j*. *N* est le nombre de document dans le corpus général, et *n* est l'occurrence du terme (le nombre de documents où *j* est présent) Le poids global de *j* égale au logarithme de *N* divisé par *n*. Nous avons :

$$W(j) = \log(N/n)$$

Dans l'image suivante, la courbe rouge est la relation «poids de termes/occurrence terme du corpus » dans un corpus de 100 documents. La courbe verte est la même relation mais dans un corpus plus petit : le corpus qui contient que 10 documents. La distance entre ces deux courbes nous intéresse. Plus le nombre de document est important dans un corpus, plus la différence de poids de termes sera significative. C'est-à-dire que dans un plus corpus, le poids d'un terme rare et spécifique est plus élevé que dans un petit corpus.

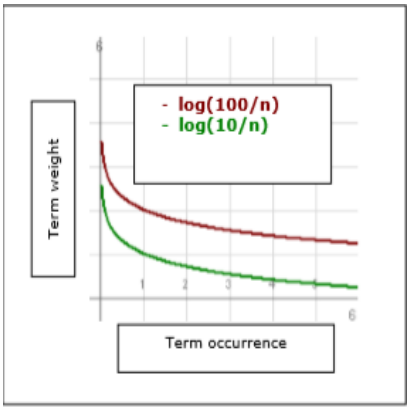


Figure 16 poids de termes dans des différents corpus

Pour que le résultat de ce learning soit intéressant, nous avons calculé les poids des termes du thésaurus sur un corpus des journaux qui parle de la politique, de la finance, de la macro économie et de la vie sociale. Ces catégories ne font pas partie de nos catégories en question. Les documents sont sous format *HTML*. Le nombre de termes par document est 500 en moyenne.

```
<html>
<head>
<title>Un projet de loi pour faire des prud_hommes de vrais juges</title>
</head>
<p>Il vous reste 74% de l'article à lire</p>
<p>Prud'hommes</p>
<p>Le projet de loi d'Emmanuel Macron contre les « trois maladies » françaises
La réforme des prud'hommes reportée</p>
<p>Un projet de loi pour faire des prud'hommes de vrais juges</p>
<p>LE MONDE
Le conseil des prud'hommes de Toulouse, en 2013 (photo d'illustration).
Le malaise est difficilement contestable : la justice prud'homale boitille depuis des dizaines d'années et la situation ne cesse de se dégrader. Un conflit entre un salarié et un employeur déposé aujourd'hui à Paris devant les prud'hommes ne sera pas examiné avant 2017, et la procédure pourra durer jusqu'à cinq ans…
Alain Lacabarats, le très respecté président honoraire de la chambre sociale de la Cour de cassation, a rendu en juillet un rapport sévère sur ces juridictions qui conclut sobrement qu'« a été franchi le seuil de tolérance ». Christiane Taubira, la garde des sceaux, et François Rebsamen, le ministre du travail, ont adopté les grandes lignes du rapport, et présenté le 6 novembre à Orléans un projet de loi qui devrait être adopté à la mi-décembre en conseil des ministres.
Lire le rapport d'Alain Lacabarats
Le constat est effectivement sombre. L'Etat a été condamné 51 fois en 2013 pour des dysfonctionnements prud'homaux et a payé 1,4 million d'euros de dédommagements au profit de salariés. « Quand la justice est condamnée pour déni de justice, je dors mal », a soupiré Christiane Taubira. La durée moyenne des affaires est de 11,9 mois (contre 5,4 en correctionnelle, ou 5,8 au tribunal d'instance), surtout le taux d'appel devant des magistrats professionnels atteint 62,1 % (contre 13 % en correctionnelle, 6,3 % en instance). Le taux de conciliation – mission première des prud'hommes – est en baisse constante et n'était plus, en 2013, que de 5,5 %.
« Délais déraisonnables »
L'alerte a été donnée le 18 janvier 2012, lorsque le tribunal de Paris a condamné l'Etat à verser des dommages et intérêts de 1 500 à 8 500 euros à 16 plaignants pour « délais déraisonnables » d'un à cinq ans, à l'initiative du Syndicat des avocats de France. Sur les 16 affaires, 13 concernaient les prud'hommes de Bobigny, dont le tribunal a dénoncé « l'encombrement récurrent », un problème qui ne peut en aucun cas « décharger l'Etat de sa responsabilité ».
```

Figure 17 Corpus général

Le résultat de learning sur le corpus général est enrigistré dans la cartouche STF. Il est prêt pour le traitement suivant.

3.3.2.3 Paramétrage de l'extraction avec STF

Nous utilisons la cartouche STF pour faire l'extraction des termes dans les documents à catégoriser. STF applique le thésaurus aux documents. Elle embarque des technologies qui aident à surmonter les deux faiblesses traditionnelles de l'indexation taxonomique. La première de ces technologies, le Fuzzy Term Matching, produit automatiquement des variantes des formes présentes dans le thésaurus, améliorant ainsi le rappel. La seconde, le Relevance Scoring, assigne heuristiquement un score de pertinence à chaque concept extrait pour écarter les moins pertinents, améliorant ainsi la précision de l'extraction.

Le calcul de score Relevance Scoring prend en compte des 4 scores principaux dans le thésaurus :

- Score de Fuzzy Matching
- Score de profondeur
- Score de nœud
- Score d'IDF

Pour le Fuzzy Matching, il s'agit de trois correspondances partielles : Permutation, Insertion Erreur orthographique. Par exemple, pour le terme « Acousto-optiques déflecteurs » qui est un terme dans notre thésaurus, il sera extrait avec un score of 1.0 car il existe tel qu'il est dans le thésaurus. Le terme « Optique et acoustique déflecteur » sera extrait avec un score de 0.8677 car il contient une permutation ; le terme « Optique et non-acoustique déflecteur » sera extrait avec un score de 0.8339 car il s'agit d'une insertion ; et le terme « Acousto-optiques déflectors » sera extrait avec un score de 0.925 car il contient une erreur orthographique.

Le score de profondeur correspond à la distance entre le terme extrait et la racine. Dans l'image ci-dessous (Figure 18), le score de ; profondeur du terme *ratio* égale à 3. Le score de profondeur du terme *prévoyance* égale à 2. Nous considérons que le terme *ratio* est plus spécifique que le terme *prévoyance* car il est plus profond dans hiérarchie. En conséquence, le terme *ratio* sera plus pondéré dans le thésaurus.

Le score de nœud est basé sur le positionnement du terme extrait dans le thésaurus entier. Considérons que terme *j* du thésaurus est extrait dans un document. Le score de nœud du terme *j* est le nombre des « Sibling terms »[13] du *j* dans le thésaurus. Par exemple, si le score de nœud du *j* égale à 4, c'est-à-dire que dans le thésaurus, il y a 4 autres termes qui sont liés directement au même terme que *j*. Cette mesure augmente le poids d'un terme s'il n'est pas tout seul dans la hiérarchie.

Le score d'IDF est le résultat du learning que nous avons lancé dans le chapitre précédant et que nous avons stocké dans la cartouche STF.

Nous projetons la cartouche sur notre grand corpus de référence. Nous avons obtenu un résultat évalué par Annotation Workbench.

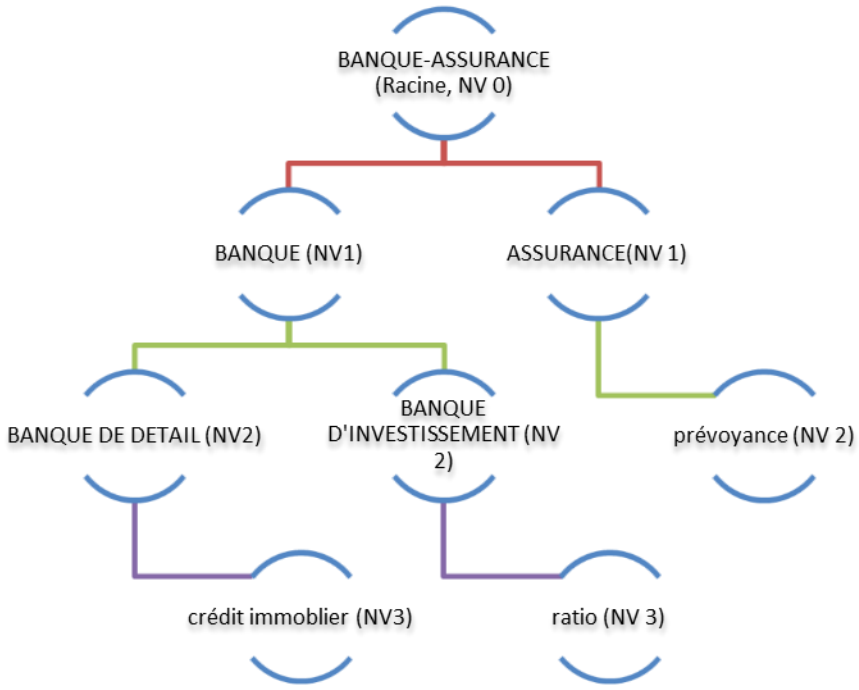


Figure 18 Hiérarchie du thésaurus

[13] « Sibling term », appelé également « terme voisin », est le terme qui appartient au même terme que le terme *j* et qui se trouve au même niveau que *j* dans la hiérarchie.

4. Evaluation et discussion

4.1 Mesure d'évaluation

Pour évaluer la qualité des résultats et les comparer, nous avons utilisé les mesures classiques d'évaluation de la qualité qui est : la précision et le rappel.

- Précision = Correct / (Correct + Bruit)
- Rappel = Correct / (Correct + Silence)

Correct signifie qu'un document *j* qui appartient à la catégorie *A* est annoté en catégorie *A* ; Bruit signifie qu'un document *j* qui appartient à la catégorie *B* est annoté en catégorie *A*, et vice versa ; Silence signifie qu'un document *j* qui appartient à la catégorie A n'est pas catégorisé en A. Dans le cas précédent du Bruit, *j* est considéré comme un bruit pour la catégorie *A*, en même temps, il est également considéré comment un silence pour la catégorie B.

Précision indique le nombre de documents correctement catégorisés par l'outil par rapport au nombre total de documents catégorisés ;

Rappel mesure le nombre de documents correctement catégorisés par l'outil par rapport au nombre total de documents de catégories de référence.

4.2 Résultats de catégorisation automatique

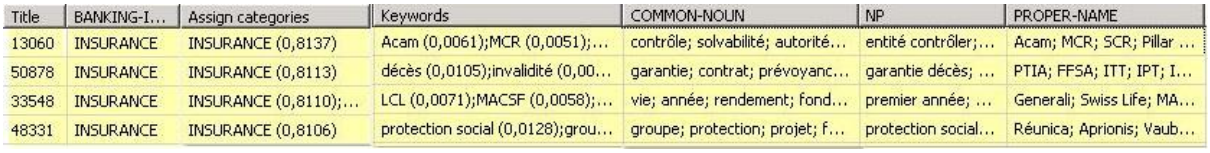
4.2.1 Résultat de l'apprentissage automatique statistique

Le bloc en violet est les documents « missed ».D'après les termes dans ces documents, nous n'avons pas pu trouver une catégorie qui correspond aux vecteurs moyens de ces documents. Par conséquent, la colonne « Assign categories » est vide.

Title	BANKING-INSUR…	Assign categories	Keywords	COMMON-NOUN	NP	PROPER-NAME
15868	INSURANCE			assurance; maison; réseau; …	GAN assurance; homme de le …	GAN; Benoît Maes…
11616	INSURANCE			saga; association; fonds; en…	automobile association; million…	Permira; STÉPHAN…
38705	INSURANCE			bénéfice; croissance; milliard…	bénéfice net; croissance ann…	Swiss Life; wiss Lif…
24567	INSURANCE			hausse; marge; group; béné…	bénéfice net; marge opératio…	April; APRIL; résul…

Figure 19 Résultat d'assignement-Silence généré par Category Workbench

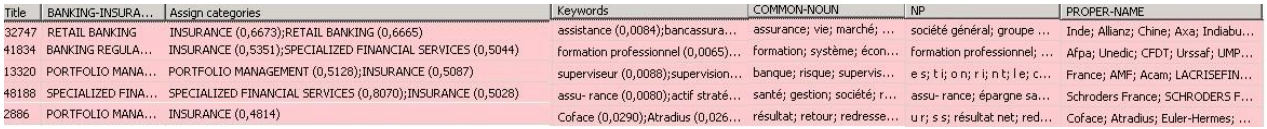
Le bloc en jaune présente les documents qui sont bien assignés. Ses catégories de référence correspondent aux catégories assignées. C'est-à-dire que basant sur les scores de termes, la machine a réussi à associer la bonne catégorie aux documents.



Title	BANKING-I...	Assign categories	Keywords	COMMON-NOUN	NP	PROPER-NAME
13060	INSURANCE	INSURANCE (0,8137)	Acam (0,0061);MCR (0,0051);...	contrôle; solvabilité; autorité...	entité contrôler;...	Acam; MCR; SCR; Pillar ...
50878	INSURANCE	INSURANCE (0,8113)	décès (0,0105);invalidité (0,00...	garantie; contrat; prévoyanc...	garantie décès; ...	PTIA; FFSA; ITT; IPT; I...
33548	INSURANCE	INSURANCE (0,8110);...	LCL (0,0071);MACSF (0,0058);...	vie; année; rendement; fond...	premier année; ...	Generali; Swiss Life; MA...
48331	INSURANCE	INSURANCE (0,8106)	protection social (0,0128);grou...	groupe; protection; projet; f...	protection social...	Réunica; Aprionis; Vaub...

Figure 20 Résultat d'assignement-Correct généré par Category Workbench

Le bloc rouge contient les documents qui sont mal catégorisés. En comparant le vecteur sémantique du document avec celui de la catégorie, nous sommes tombés sur des mauvaises catégories.



Title	BANKING-INSURA...	Assign categories	Keywords	COMMON-NOUN	NP	PROPER-NAME
32747	RETAIL BANKING	INSURANCE (0,6673);RETAIL BANKING (0,6665)	assistance (0,0084);bancassura...	assurance; vie; marché; ...	société général; groupe ...	Inde; Allianz; Chine; Axa; Indiabu...
41834	BANKING REGULA...	INSURANCE (0,5351);SPECIALIZED FINANCIAL SERVICES (0,5044)	formation professionnel (0,0065)...	formation; système; écon...	formation professionnel; ...	Afpa; Unedic; CFDT; Urssaf; UMP...
13320	PORTFOLIO MANA...	PORTFOLIO MANAGEMENT (0,5128);INSURANCE (0,5087)	superviseur (0,0088);supervision...	banque; risque; supervis...	e s; t i; o n; r i; n t; l e; c...	France; AMF; Acam; LACRISEFIN...
48188	SPECIALIZED FINA...	SPECIALIZED FINANCIAL SERVICES (0,8070);INSURANCE (0,5028)	assu- rance (0,0080);actif straté...	santé; gestion; société; r...	assu- rance; épargne sa...	Schroders France; SCHRODERS F...
2886	PORTFOLIO MANA...	INSURANCE (0,4814)	Coface (0,0290);Atradius (0,026...	résultat; retour; redresse...	u r; s s; résultat net; red...	Coface; Atradius; Euler-Hermes; ...

Figure 21 Résultat d'assignement-Incorrect généré par Category Workbench

Après la phase d'assignement, nous avons obtenu le résultat du modèle de catégorisation évalué par Category Workbench.

Qualité	74,35%
Précision	75,9%
Rappel	72,8%

Tableau 3 Qualité généré par Category Workbench

L'image Figure 13 présente la qualité générale de ce modèle. Le résultat n'est pas très satisfaisant. L'objectif de ce modèle est de catégoriser les documents sur LEOnard. Cependant, une plateforme d'information telle que LEOnard attache de l'importance à la précision de la catégorisation. Après avoir observé et analysé les données, nous en conclurions les points suivants :

4.2.2 Résultat de l'approche thésaurus

Pour l'approche Thésaurus, nous avons obtenu un résultat très satisfaisant. Il y a très peu de bruit ainsi que de silence.

Catégorie	Précision	Rappel	F-Mesure
Ensemble	97,4%	97,45%	98,7%
INDUSTRY	98,6%	99,6%	99,7%
BANKING-INSURANCE	99,8%	99,1%	99,4%
AGRO-ALIMENTAIRE	97%	96,1%	96,8%
HIGH-TECH	95,2%	95,2%	95,2%
ENERGY-ENVIRONMENT	98,6%	97,4%	98%
MEDIA-ADVERTISING	98,6%	97,4%	97,6%
BUILDING CIVIL - ENGINEERING-REAL ESTATE	97%	97,4%	97,2%
SERVICES-LEISURE ACTIVITIES	95,2%	97,4%	96,7%

Tableau 4 Qualité évalué par Annotation Workbench

4.3 Discussion

4.3.1 Observation des résultats

A partir des résultats récoltés dans les deux approches, nous avons observé des points intéressants.

Pour l'approche apprentissage automatique statistique :

1) Comme présenté dans le chapitre 3.1.2 Caractéristiques des données et du corpus, les documents sont mal numérisés et mal segmentés à l'origine. Parce que le processus d'apprentissage dépend largement de la qualité du corpus, un corpus bruité ne permet pas de générer un modèle efficace
2) Dans certaines catégories qui obtient les niveaux de précision très bas, telles que Industrie/Chimique, il y a quasiment un quart de documents qui contiennent du bruit. Dans ce cas, les vecteurs moyens de ces catégories ne sont pas corrects. Ils génèrent du faux score pendant le calcul.

3) Les images ci-dessous nous présentent la qualité de chaque catégorie. Nous trouvons qu'il y a beaucoup de catégories qui sont N/A. C'est-à-dire que la machine ne les a pas trouvées dans le corpus d'évaluation. En conséquence, elle n'a pas réussi à évaluer la qualité de ces catégories. Dans les 52 catégories que nous avons définies, il y a 29 catégories qui n'existent pas dans le corpus d'évaluation, équivalent à 55,77% de catégories qui n'existent pas dans le corpus d'évaluation. Ce problème vient du fait que la division du grand corpus en corpus d'apprentissage et en corpus d'évaluation est aléatoire. Les seuils de division sont : 90% pour corpus d'apprentissage et 10% pour corpus d'évaluation. Apparemment, 10% pour corpus d'évaluation n'est pas suffisant. Il est probable que ce 10% du grand corpus ne contient pas certaines catégories. Il est même possible que tous ces documents de 10% sont dans la même catégorie. Cette division aléatoire rend le résultat d'assignement et d'évaluation moins fiable.

Category	Precision	Recall	Correct	Missed	False	Test
/BANKING-INSURANCE/BANKING REGULATIONS	N/A	0.0	0	8	0	8
/BANKING-INSURANCE/INSURANCE	96.7	91.1	206	20	7	226
/BANKING-INSURANCE/INVESTMENT BANKING	N/A	0.0	0	29	0	29
/BANKING-INSURANCE/ISLAMIC BANK	N/A	0.0	0	5	0	5
/BANKING-INSURANCE/MEANS OF PAYMENT	N/A	0.0	0	16	0	16
/BANKING-INSURANCE/PORTFOLIO MANAGEMENT	82.6	90.0	190	21	40	211
/BANKING-INSURANCE/RETAIL BANKING	58.4	100.0	38	0	27	38
/BANKING-INSURANCE/SPECIALIZED FINANCIAL SERVICES	66.3	98.3	61	1	31	62
/BUILDING CIVIL ENGINEERING-REAL ESTATE/BUILDING MATERIALS	N/A	0.0	0	21	0	21
/BUILDING CIVIL ENGINEERING-REAL ESTATE/BUILDING-SUBSTRUCTURE	71.9	97.6	41	1	16	42
/BUILDING CIVIL ENGINEERING-REAL ESTATE/CONSTRUCTION TERMS	N/A	0.0	0	8	0	8
/BUILDING CIVIL ENGINEERING-REAL ESTATE/CORPORATE REAL ESTATE	68.0	96.0	49	2	23	51
/BUILDING CIVIL ENGINEERING-REAL ESTATE/INDIVIDUAL REAL ESTATE-LEISURE REAL ...	N/A	0.0	0	24	0	24
/Category/COMMODITIES	0.0	N/A	0	0	81	0
/Category/GAS-ELECTRICITY-COAL	0.0	N/A	0	0	118	0
/Category/OIL	0.0	N/A	0	0	81	0
/ENERGY-ENVIRONMENT/COMMODITIES	N/A	0.0	0	65	0	65
/ENERGY-ENVIRONMENT/ENVIRONMENT	N/A	0.0	0	20	0	20
/ENERGY-ENVIRONMENT/GAS-ELECTRICITY-COAL	N/A	0.0	0	96	0	96
/ENERGY-ENVIRONMENT/OIL	N/A	0.0	0	71	0	71
/ENERGY-ENVIRONMENT/RENEWABLE ENERGY	N/A	0.0	0	23	0	23
/ENERGY-ENVIRONMENT/WASTE MANAGEMENT	N/A	0.0	0	10	0	10
/FOOD INDUSTRY-RETAILING-LUXURY-FASHION/CLOTHING-ACCESSORY	76.0	94.7	108	6	34	114
/FOOD INDUSTRY-RETAILING-LUXURY-FASHION/COSMETICS	66.6	100.0	34	0	17	34
/FOOD INDUSTRY-RETAILING-LUXURY-FASHION/E-COMMERCE	N/A	0.0	0	10	0	10
/FOOD INDUSTRY-RETAILING-LUXURY-FASHION/FOOD INDUSTRY	92.5	92.1	222	19	18	241
/FOOD INDUSTRY-RETAILING-LUXURY-FASHION/LUXURY PRODUCTS	N/A	0.0	0	25	0	25
/FOOD INDUSTRY-RETAILING-LUXURY-FASHION/RETAIL INDUSTRY	84.2	98.4	64	1	12	65
/FOOD INDUSTRY-RETAILING-LUXURY-FASHION/SPECIALTY RETAIL STORE	57.1	100.0	32	0	24	32
/HIGH-TECH/COMPANY	82.9	94.4	34	2	7	36
/HIGH-TECH/CONNECTIVITY-MOBILITY	91.8	95.3	124	6	11	130
/HIGH-TECH/EQUIPMENT	79.6	96.2	51	2	13	53
/HIGH-TECH/MATERIAL	80.7	91.3	63	6	15	69
/HIGH-TECH/SOFTWARE	70.2	89.1	33	4	14	37
/INDUSTRY/AEROSPACE-AERONAUTICS-AIRLINE INDUSTRY	96.0	93.8	122	8	5	130
/INDUSTRY/CAPITAL GOOD	69.5	91.4	32	3	14	35
/INDUSTRY/CAR MANUFACTURER-DEALERSHIP	91.0	93.8	92	6	9	98
/INDUSTRY/CHEMISTRY	73.7	91.8	45	4	16	49
/INDUSTRY/EQUIPMENT MANUFACTURER AND REPAIRING	N/A	0.0	0	29	0	29
/INDUSTRY/FINANCING-CAR RENTAL	N/A	0.0	0	6	0	6
/INDUSTRY/METALWORKING INDUSTRY-STEEL INDUSTRY	N/A	0.0	0	24	0	24
/INDUSTRY/PHARMACY	89.8	97.5	80	2	9	82
/INDUSTRY/ROAD AND RAIL AND GOODS TRANSPORT	73.6	98.1	53	1	19	54
/INDUSTRY/SHIPBUILDING AND SEA TRANSPORT	N/A	0.0	0	26	0	26
/INDUSTRY/WOOD PAPER-PACKAGING-FURNITURE	81.8	92.6	63	5	14	68
/MEDIA-ADVERTISING/ADVERTISING	86.0	82.2	37	8	6	45
/MEDIA-ADVERTISING/MEDIA	92.3	91.8	170	15	14	185
/POLICY-SAFETY-DEFENSE/SAFETY-DEFENSE	N/A	0.0	0	2	0	2
/SERVICES-LEISURE ACTIVITIES/CATERING	N/A	0.0	0	9	0	9
/MEDIA-ADVERTISING/MEDIA	92.3	91.8	170	15	14	185
/POLICY-SAFETY-DEFENSE/SAFETY-DEFENSE	N/A	0.0	0	2	0	2
/SERVICES-LEISURE ACTIVITIES/CATERING	N/A	0.0	0	9	0	9
/SERVICES-LEISURE ACTIVITIES/CONSULTING-ENGINEERING-OUTSOURCING	N/A	0.0	0	18	0	18
/SERVICES-LEISURE ACTIVITIES/HEALTH-WELFARE	N/A	0.0	0	9	0	9
/SERVICES-LEISURE ACTIVITIES/MONEY GAMES-TOYS	N/A	0.0	0	20	0	20
/SERVICES-LEISURE ACTIVITIES/SECURITY-CLEANLINESS	N/A	0.0	0	7	0	7
/SERVICES-LEISURE ACTIVITIES/SPORT	N/A	0.0	0	9	0	9
/SERVICES-LEISURE ACTIVITIES/TOURISM	79.6	92.1	47	4	12	51
/SERVICES-LEISURE ACTIVITIES/TRAINING-TEMPORARY WORK	N/A	0.0	0	12	0	12

Figure 22 Qualité par catégorie

4) La raison pour laquelle nous avons eu un rappel très élevé est que tout le corpus d'apprentissage vient de la même source : la base interne de BNP Paribas. Les

documents de la même catégorie sont plus ou moins homogènes. Malgré tous les calculs de scores que nous avons utilisés pendant l'apprentissage, ces documents ne peuvent pas nous définir un champ lexical suffisamment large pour couvrir tous les termes spécifiques dans une catégorie. En conséquence, quand un nouveau document contenant un champ lexical différent de celui du corpus d'apprentissage est analysé par le modèle de catégorie, ce dernier n'arrive pas à reconnaître la bonne catégorie.

5) La machine Category Workbench elle-même a mal fonctionné pendant le processus d'apprentissage. Le fait qu'elle est connectée à une base de données produit du cache à chaque fois qu'on lance le processus. Dans ce cas, le cache du dernier exercice pollue le résultat. C'est un problème déjà connu par le développeur. Par exemple, dans la Figure 19, nous voyons la catégorie *OIL* qui est sous *Category*. En revanche, dans notre classification des données, la catégorie *OIL* est bien dans Energie-Environnement.

Pour l'approche thésaurus :

Nous avons obtenu un résultat très satisfaisant grâce aux raisons suivantes :

1) Le thésaurus est suffisamment complet pour définir les catégories dans notre recherche. Nous avons 4041 termes dans 8 grandes catégories. C'est-à-dire que 500 par catégorie. D'ailleurs, les termes sont très riches au niveau du champ lexical.
2) Le calcul de pondération de termes ne dépend pas de la qualité de corpus de référence. La mesure la plus importante dans cette méthode est l'IDF qui ne concerne que le corpus général et le terme lui-même. Vu que le corpus général sur lequel nous avons effectué le learning pour apprendre l'IDF de termes est très propre et complet, le résultat du thésaurus est très satisfaisant.

Nous trouvons que cette approche est plus facile à manipuler. Elle est plus performante que la première approche « apprentissage automatique statistique ». Le résultat produit par cette approche nous satisfait. D'ailleurs, nous pouvons définir les paramètres pour personnaliser la catégorisation pour qu'elle convienne à un besoin spécifique. Cela nous permet de personnaliser le thésaurus en fonction de nos besoins.

Néanmoins, le résultat de l'extraction dépend à 100% des termes dans le thésaurus même si nous avons la possibilité de faire le « fuzzy matching ». Dans ce cas-là, nous devons faire

attention à ce que les termes appartenant à une catégorie dans le thésaurus soient complets et couvre tous les domaines de cette catégorie. Dans ce cas, la construction de thésaurus demande beaucoup de temps et d'énergie. Le thésaurus devrait contenir beaucoup de termes spécifiques dans une catégorie pour pouvoir catégoriser un document inconnu. Tout cela demande beaucoup de travail de Fouille de texte.

Nous en concluons que la méthode de thésaurus nous permet d'avoir un résultat beaucoup plus précis et performant. Néanmoins, elle est plus pratique si les données sont simples. C'est-à-dire que nos données sont plus ou moins homogènes et la classification de données n'est pas très compliquée.

4.3.2 Travaux futurs

D'après l'observation tout au long de notre étude et les résultats obtenus, il reste plusieurs traitements potentiels à réaliser pour améliorer la performance des outils et obtenir des meilleurs résultats.

Pour l'approche apprentissage automatique :

Pour améliorer l'apprentissage automatique, nous souhaitons un corpus qui est plus propre et bien formalisé. Nous pouvons collecter des pages du format texte brut et sans faute orthographique. Cela nous permet d'améliorer le résultat d'apprentissage. Nous pouvons affiner le prétraitement pour enfin éviter les signes inconnus, les caractères bizarres ainsi que la mauvaise segmentation de mot ou de phrase.

Pour les domaines Sport et Safety Defense et d'autres catégories qui ne contiennent pas beaucoup de documents, nous pourrons alimenter le corpus de référence de ces catégories pour perfectionner le travail de catégorisation automatique.

Pour diminuer le bruit, nous pouvons élaborer une liste de termes qui sont interdits. Cette liste contient des termes qui sont très ambiguës et qui ne sont pas intéressants pour définir une catégorie.

Pour l'approche thésaurus :

Nous estimons qu'un thesaurus plus grand et mieux compréhensible permettrait de diminuer le taux de silence. L'augmentation de la taille du thesaurus est possible grâce aux corpus gratuits disponibles, tel que EUROVOC[14]. Cependant, il faut noter que la plupart des thésauri sont disponibles en format pdf pour faciliter la lecture des utilisateurs. Pour intégrer ce type de thésaurus dans notre projet, il faudrait effectuer une conversion de format pdf en format skos. Cela demande beaucoup de travail sachant que la taille du thésaurus est importante.

Ensuite, vu que la classification de ce thésaurus est très vague, il y a beaucoup de termes qui ne sont pas pertinents pour notre classification de catégories. Nous constatons qu'il y a

[14] EuroVoc est un thésaurus multilingue de l'Union européenne qui couvre la terminologie des domaines d'activité de l'UE. Il est généré par l'Office des publications.

beaucoup de mots qui ne sont jamais extraits dans notre corpus. Pour augmenter l'efficacité du projet, il faudrait effectuer un nettoyage dans le thésaurus.

5. Conclusion

Notre étude a pour but de réaliser une catégorisation automatique des données sur le site LEOnard. Les documents à catégoriser viennent de la presse quotidienne et des sites web crawlés. Ces informations possèdent deux caractéristiques remarquables : la variété du contenu des données et la quantité importante d'informations. La solution traditionnelle « apprentissage automatique » n'est plus suffisante pour traiter ces informations aussi riches. Nous avons besoin d'une nouvelle approche de catégorisation automatique. Dans ce cas, nous avons introduit un thésaurus dans le processus apprentissage automatique pour affiner la catégorisation et nous appuyer sur des termes vraiment spécifiques.

Nous avons construit un corpus de référence et un corpus de test qui possèdent les mêmes caractéristiques que les documents à traiter sur LEOnard. Ces deux corpus sont déjà associés aux bonnes catégories. Après l'apprentissage automatique effectué par l'outil Category Workbench, nous avons construit un modèle de catégorisation. Pour évaluer la qualité et la performance de ce modèle, nous l'avons projeté sur un corpus.

Pour l'approche thésaurus, c'est différent. Nous avons regroupé tous les descripteurs de chaque catégorie pour construire d'abord un thésaurus. Nous l'avons projeté sur un corpus général. Ce corpus contient des documents qui ne sont pas dans les catégories définies dans notre projet. Suite à un calcul de poids de termes, les termes qui sont très fréquents dans le corpus général sont considérés comme des termes inintéressants pour notre catégorisation automatique. Nous avons annoté le corpus de référence avec le thésaurus.

La comparaison de ces deux approches se fait sur deux critères : la qualité de la catégorisation automatique et la faisabilité.

Après une phase d'évaluation, nous avons obtenu des résultats très différents. L'évaluation de la qualité du modèle généré par Category Workbench avec l'approche « apprentissage automatique statistique » génère une précision de 75,9% et un rappel de 72,8%. Cela ne nous semble pas très satisfaisant.

Cependant, l'approche thésaurus nous a fourni une bonne performance avec une précision de 97,4% et un rappel de 97,45%. C'est ce que nous nous attentions au début.

Au niveau de la faisabilité, nous trouvons que l'approche de « apprentissage automatique statistique » est plus facile à mettre en place et à manipuler. Néanmoins, l'approche thésaurus demande beaucoup plus de temps et d'énergie afin de construire et nettoyer le thésaurus. Pour optimiser la performance du thésaurus, nous pouvons profiter des thésaurus déjà disponibles en ligne.

L'introduction du thésaurus dans le travail de catégorisation automatique est innovante. L'amélioration de la performance de cette méthode nécessitera de nouveaux travaux.

6. Bibliographie

[Augé et al., 2003] Jérôme Augé, Kurt Englmeier, Gilles Hubert, Josiane Mothe (2007) Catégorisation automatique de textes basée sur des hiérarchies de concepts, ftp://irit.fr/IRIT/SIG/2003_BDA_AEHM.pdf

[Bertels et al., 2012], Ann Bertels, Dirk De Hertog, Kris Heylen, Etude sémantique des mots-clés et des marqueur lexicaux stables dans un corpus technique, Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2 : TALN, page 239-252

[Chaumartin, 2013], François-Régis Chaumartin Apprentissage d'une classification thématique générique et cross-langue à partir des categories de la Wikipédia, TALN-RECTAL 2013, 17-21 juin, les Sables d'Olonne

[Coupet et al.]Coupet, P., Buschbeck, B., Six, A., Cardoso, F., & Huot, C. Le Text Mining multilingue: application au monde de l'intelligence économique1.

[Gandon et al., 2012], Gandon, F., Corby, O., & Faron-Zucker, C. (2012). Le Web sémantique:Comment lire les données et les schémas sur le web? Dunod.

[Dumais,2000] S. Dumais, H. Chen, « Hierarchical classification of Web documents », 23nd Annual International ACM Conference on Research and Development in Information Retrieval SIGIR'2000, Athenes, 2000.

[Foucault et al., 2013], Nicolas Foucault, Sophie Rosset, Gilles Adda Pré-segmentation de pages web et sélection de documents pertinents en Questions-Réponses, TALN-RECTAL 2013, 17-21 juin, les Sables d'Olonne

[JALAM, 2003], Radwan JALAM Apprentissage automatique et catégorisation de textes multilingues, UNIVERSITÉ LUMIÈRE LYON2, PhD soutenu le 4 juin 2003

[Joachims,1998]T. Joachims, « Text categorization with Support Vector Machines: Learning with many relevant features », 10th European Conference on Apprentissage automatique ECML'98, p. 137-142, 1998

[Kevers,2009], Laurent Kevers, Indexation semi-automatique de textes : thésaurus et transducteurs, CORIA 2009 - Conférence en Recherche d'Information et Applications

[Speelman et al., 2006], Speelman, D., Grondelaers, S., & Geeraerts, D. (2006). A profile-based calculation of region and register variation: the synchronic and diachronic status of the two main national varieties of Dutch. Language and Computers, 56(1), 181-194.

[Speelman et al., 2008] Speelman, D., Grondelaers, S., & Geeraerts, D. (2008). Variation in the choice of adjectives in the two main national varieties of Dutch. Cognitive sociolinguistics: language variation, cultural models, social systems.

[Tikk et al., 2003], Domonkos Tikk , Jae Dong Yang , Sun Lee Bang, Hierarchical text categorization using fuzzy relational thesaurus, Kybernetika, Vol. 39 (2003), No. 5, [583]—600 Persistent URL: http://dml.cz/dmlcz/135557

[Yun et al., 2011], YUN, J. JING, L., YU, J., HUANG, H. ZHANG, Y. (2011). Document Topic Extraction Based on Wikipedia Category. Actes de Computational Sciences and Optimization (CSO).

7. Annexe

7.1 Vocaulaires du format skos dans notre thésaurus

Les vocabulaires que nous avons utilisés dans notre thésaurus sont les suivants :

- <skos:Concept rdf:about="#XXX"> Cette balise désigne l'ID du terme dans le thésaurus. Cette ID commence par un #. Elle est une clé unique dans le thésaurus. Le thésaurus n'accepte pas le doublon.

- <skos:prefLabel xml:lang="fr"> Cela contient la forme du terme dans le thésaurus ainsi que la langue du terme. Au contraire que l'ID du terme, cette valeur peut être répétitive à condition que les IDs soient différentes.

- <skos:broader> cette balise indique le parent du terme qui est directement lié au terme présent mais qui se trouve au niveau précédent du terme présent. Et <skos:narrower> pointe au terme fils qui est directement lié au terme présent mais qui se trouve au niveau plus bas dans la hiérarchie. Cela permet de lier les termes ensemble pour décrire le réseau sémantique.

- <note> Note indique le contexte dans lequel le terme se présente. Un terme ne va pas être pris en compte sauf qu'il est présent avec le mot de contexte.

7.2 Script Perl pour le prétraitement du corpus

```
#!/usr/bin/perl

use utf8;
use DBI;
use DBD::ODBC;
use Term::UI;
use Term::ReadLine;
use Params::Check qw[check allow last_error];

###################paramètres à saisir######

$repertoire="D:\\APPS\\Catego\\Sources\\Gimadoc\\Corpus-Gimadoc-Francais-Anglais"; # mettre le chemin absolu correspondant au répertoire contenant les fichiers Gimadoc
$fichier_sortie= "D:\\APPS\\Catego\\CorpusApprentissage\\Gimadoc\\Gimadoc-Francais\\corpusGimadocFR-adapteHIGHTECH.tmx"; #mettre le chemin absolu du fichier de sortie TMX. ex: "C:\\Categorisation\\toto.xml"
$fichier_categories= "D:\\APPS\\Catego\\plan_de_classement\\Gimadoc\\EN\\planClassementTM_ENavcHighTech.txt"; #mettre le chemin absolu de la table de correspondance métadonnées->catégories
$fichier_temporaire="D:\\APPS\\Catego\\temp1.xml"; # chemin absolu du fichier temporaire. Ne pas modifier
$langue="fr"; # filtrage du corpus au niveau de la langue: mettre "fr" ou "en" selon la langue désirée
$taille_fichiers=50000; #filtrage du corpus au niveau de la taille des fichiers: mettre un chiffre correspondant au nombre d'octet maximal des fichiers à sélectionner
@selection_cat=();   # Mettre le nom des catégories à sélectionner sur le modèle: "BANQUE", "ENERGIE", "ENTREPRISE DE SERVICE"

#Paramètres de connexion à SQL server

my $dsn = 'CategoLeo'; #source de données ODBC
my $dbuser = 'CategoLeo'; # utilisateur
my $dbpass = 'CategoLeo'; #mot de passe

###################Fin paramètres##########

my $dbh = DBI->connect ("dbi:ODBC:$dsn", $dbuser, $dbpass) or die "$DBI::errstr\n";

$requete = $dbh->prepare("if exists (select * from INFORMATION_SCHEMA.TABLES where TABLE_NAME='categories') drop table categories");
$requete->execute() or die "Echec de la suppression de la table categories\n";
$requete->finish();

$requete = $dbh->prepare("if exists (select * from INFORMATION_SCHEMA.TABLES where TABLE_NAME='documents') drop table documents");
$requete->execute() or die "Echec de la suppression de la table documents\n";
$requete->finish();


```

```
37
38
39  $requete = $dbh->prepare("create table categories (descripteur VARCHAR(150), categorie
    VARCHAR(150));");
40  $requete->execute() or die "Echec de la creation de la table des categories\n";
41  $requete->finish();
42
43
44  $requete = $dbh->prepare("create table documents (document Integer, descripteur
    VARCHAR(150));");
45  $requete->execute() or die "Echec de la creation de la table des documents\n";
46  $requete->finish();
47
48  $compteur=1;
49
50  # Lecture des fichiers et ouverture du fichier de sortie en écriture
51  unlink ("$fichier_temporaire");
52  open(H, ">>$fichier_temporaire");
53  print H "<?xml version=\"1.0\" encoding=\"UTF-8\"?>\n <tm
    xmlns:dc=\"http:\/\/purl.org\/dc\/elements\/1.1\/\">";
54
55
56  opendir (REP, "$repertoire") or die "impossible d'ouvrir le repertoire contenant le corpus";
57  @fichiers = readdir REP;
58  foreach $fichier (@fichiers){
59
60  #Test sur la taille
61
62      if ($fichier=~/[A-Za-z]+/){
63          $taille=(-s "$repertoire\\$fichier");
64          if ($taille < $taille_fichiers){
65              print "$fichier\n";
66              open (F,"$repertoire\\$fichier") or die "impossible d'ouvrir le fichier du
                corpus";
67              @tab = <F>;
68              $texte = join ("",@tab);
69
70  #Test sur la langue
71
72              if ($langue eq "fr" ) {
73                  $expReg1=qr/ et /i; $expReg2=qr/ un /i;
74              }else {
75                  $expReg1=qr/ the /i; $expReg2=qr/ to /i;
76                  }
```

```
77
78
79              if (($texte=~$expReg1) and ($texte=~$expReg2)) {
80                  print "OK\n";
81  #Récupération des descripteurs
82                  $texte=~m/<b>descripteur:<\/b><\/font> <font color="black">([^<]+)<\/font>/;
83                  @desc=$1=~/(.+)+/ig;
84
85  #Ecriture dans le fichier TMX
86
87                  print H "\n<doc id=\"$compteur\">\n";
88                  print H "\<dc:title>$compteur</dc:title>\n";
89                  print H "<categories><c></c></categories>\n";
90                  print H "<features>\n";
91
92
93                  foreach $desc(@desc){
94                      print H "<ft f=\"1\">/Metadata/$desc</ft>\n";
95                  }
96
97                  print H "</features>\n";
98                  print H "<text><file format=\"html\"
                    path=\"$repertoire\\$fichier\"/></text>\n";
99                  print H "</doc>\n";
100
101
102 #Ecriture dans la BD
103
104                 foreach $desc(@desc){
105                     $desc=~s/'/ /g;
106                     $requete = $dbh->prepare("insert into documents(document, descripteur)
                        VALUES ($compteur,'$desc')");
107                     $requete->execute() or die "Echec requete insertion document\n $desc";
108                     $requete->finish();
109                 }
110
111
112                 $compteur++;
113             }
114             close F;
115         }
116     }
117 }
118
119 print H "</tm>\n";
120 close H;
121 closedir (REP);
122
```

```
122
123   #Alimentation de la BD Catégories
124
125   open(J, "$fichier_categories");
126   @tab=<J>;
127   foreach $ligne (@tab){
128       $ligne =~/([^\t]+)\t([A-Za-z\-' \/]+)/;
129       $mot1=$1;
130       $mot2=$2;
131
132       $mot1=~s/'/ /g;
133       $mot2=~s/'/ /g;
134       $requete = $dbh->prepare("insert into categories(descripteur, categorie) VALUES
          ('$mot1','$mot2')");
135       $requete->execute() or die "Echec requete insertion categorie\n";
136       $requete->finish();
137
138   }
139
140   #$requete = $dbh->prepare("select distinct documents.document, categorie from documents,
      categories where categories.descripteur=documents.descripteur group by document, categorie
      having count(document)=1;");
141   $requete = $dbh->prepare("select T.document, D.categorie  FROM (select distinct
      documents.document FROM documents, categories where
      categories.descripteur=documents.descripteur group by document having count (distinct
      categorie)=1) As T, (select distinct documents.document, categorie FROM documents,
      categories where categories.descripteur=documents.descripteur) As D WHERE
      T.document=D.document  order by T.document;");
142   $requete->execute() or die "Echec requete selection de documents\n";
143
144
145   close (J);
146
147
148   ##########################################
149   ##Ecriture du  TMX####
150   ##########################################
151
152
153   unlink ("$fichier_sortie");
154   open(K,"$fichier_temporaire");
155   open (L,">>$fichier_sortie");
156   print L "<?xml version=\"1.0\" encoding=\"UTF-8\"?>\n <tm
      xmlns:dc=\"http:\/\/purl.org\/dc\/elements\/1.1\/\">";
157
158   $ref = $requete -> fetchall_arrayref;
159   $compteur=0;
160   foreach $r (@$ref) {
161
```

```
161
162         $id=$r->[0];
163         $cat=$r->[1];
164     #print "---------\n";
165         $bool= fait_partie ($cat, @selection_cat);
166     #print "$cat \t $bool\n";
167     #print "---------"  ;
168         print "\n$id\t$cat";
169         if ($bool==1){
170
171             print " OK\n";
172             $/="</doc>";
173             while($doc=<K>){
174
175                 if ($doc=~/id="$id"/){
176                     $doc=~s/<c><\/c>/<c>$cat<\/c>/;
177                     print L "$doc";
178                     last;
179                 }
180             }
181         }
182     }
183     print L"<\/tm>";
184     $requete->finish;
185     close L;
186
187     ##################
188
189     sub fait_partie{
190         @args=@_;
191         $cat= shift(@args);
192         $bool=0;
193
194         if (scalar(@args)>0) {
195             foreach $el(@args){
196
197                 if ($el =~/^$cat$/i) {
198                     $bool=1;
199                 }
200             }
201         }else{
202             $bool=1;
203         }
204
205         return $bool;
206     }
```