Pêle-Mél Plate-forme d'exploration, de livraison et d'évaluation des méls

Touria Aït El Mekki, Bénédicte Grailles

Pêle-mél

Plate-forme d'exploration, de livraison et d'évaluation des méls

Rapport de recherche

Autrices : Bénédicte Grailles (Université d'Angers, Temos), Touria Aït el Mekki (Université d'Angers, Leria)

Contributeur.rices : Chafik Akmouche, Tsanta Randriatsitohaina, Taimane Zerez (Université d'Angers)

Partenaires : Anne Lambert, Chloé Moser (ministère de la Santé) ; Édouard Vasseur (École nationale des Chartes, Centre Mabillon)

Octobre 2022

Soutenu par

Abstract

This report presents the methodology and results of the Pèle-mél project aimed at improving access to electronic mailboxes held by archives. The search for information within large corpora of emails (one or more email boxes) can be improved by a better manipulation of the existing metadata but also by a fine analysis of the content, via the recognition and the extraction of named entities, of terms and classification. Classification is based on neural network methods, word embeddings and document embeddings. It is guided by themes suggested by the archivists and the establishment of relationships between terms and themes. The project has resulted in two prototypes, one for retrieval and classification, the other for exploration.

Keywords : electronic messaging, supervised automatic classification, natural language processing, archiving, access, word embeddings, document embeddings

Résumé

Ce rapport expose la méthodologie et les résultats du projet Pèle-mél visant à améliorer l'accès aux messageries électroniques conservées par les services d'archives. La recherche d'informations au sein de corpus volumineux de méls (une ou plusieurs boîtes méls) peut être améliorée par une meilleure manipulation des métadonnées existantes mais aussi par une analyse fine du contenu, via la reconnaissance et l'extraction d'entités nommées et de termes et de la classification. Cette dernière s'appuie sur des méthodes de réseaux de neurones, plongement de mots et plongement de documents. Elle est guidée par des thèmes suggérés par les archivistes et l'établissement de relations entre termes et thèmes. Le projet a abouti à deux prototypes, l'un permettant les extractions et la classification, l'autre l'exploration.

Mots-clés : messagerie électronique, classification automatique supervisée, traitement automatique de la langue naturelle, archivage, accès, méthode de plongement lexical, méthode de plongement de documents

Remerciements

Pêle-mél est un projet exploratoire qui propose d'importer des méthodologies nouvelles dans l'univers des archives. Il s'est nourri d'échanges préalables avec plusieurs archivistes. Qu'elles et ils soient toutes et tous remercié·es, avec une pensée particulière pour Aurélien Conraux, administrateur ministériel des données délégué au ministère de la Culture. Le programme a été accompagné par le Service interministériel des archives de France. Notre gratitude va à Françoise Banat-Berger, cheffe du Service interministériel des Archives de France, Violette Levy, cheffe du bureau de l'expertise numérique et de la conservation durable, Mélanie Rebours, cheffe du bureau du contrôle, de la collecte, des missions et de la coordination interministérielle, Dominique Naud, experte en archivage numérique. Lauréat de l'appel à projets Service numérique innovant du ministère de la Culture en 2020, il a bénéficié du suivi bienveillant d'Ariane Faraldi. Merci à elle. Merci aussi au laboratoire Temos, pour son soutien sans faille à l'archivistique et son aide dans la gestion et le montage du projet. Merci à Yves Denéchère, directeur, et Mireille Loirat, gestionnaire administratrice et aide au pilotage.

Ce projet n'aurait pas vu le jour sans notre partenaire culturel, la mission Archives des ministères sociaux, Anne Lambert, cheffe de la mission, et Chloé Moser, cheffe de produit Archifiltre. Qu'elles soient ici vivement remerciées. Nos remerciements vont également à Édouard Vasseur, École nationale des Chartes, pour son expertise jamais démentie, son implication et ses contributions tout au long de nos travaux.

Enfin, ce projet ne pouvait être imaginé sans corpus. Pour cela, il fallait obtenir de la ministre en exercice aux dates concernées l'autorisation d'accéder à des messageries et de les manipuler. Cette dérogation nous a été accordée sans aucune difficulté. Nous remercions Roselyne Bachelot-Narquin d'avoir accepté de nous ouvrir le fonds du cabinet.

Sommaire

Introduction

Le projet déposé avait pour objectifs de fournir de nouveaux outils d'exploration de corpus de messageries pour satisfaire les demandes d'information en sélectionnant les messages pertinents et de répondre aux enjeux d'accès et de mise en conformité avec les obligations de communication des documents administratifs et des archives (Code du patrimoine, Code des relations entre le public et les administrations). L'originalité du projet résidait dans la mobilisation de technologies de traitement automatique de la langue naturelle en français.

Les objectifs opérationnels du projet étaient : élaborer une base de connaissance recensant les métadonnées internes des messages, les contenus, les signatures, les pièces jointes et leurs métadonnées, les liens hypertexte via l'importation de messages au format eml (hors calendrier et agenda) ; enrichir la base de connaissance par l'injection d'ontologies, de thesaurus ou des listes d'autorité ; catégoriser et classifier les messages en se basant sur l'ontologie ; extraire les entités nommées ; proposer des fonctions de visualisation par graphe ; proposer des fonctions de recherche via des requêtes simples et expertes, des fonctions de filtre sur les métadonnées internes (permettant de trier les messages par auteur, destinataire, mots-clés etc), les classes de message, les réseaux ; implanter des fonctions d'export des résultats (listes de messages). Le projet mobilisait donc un certain nombre d'outils et de techniques de traitement automatique de la langue naturelle (Taln).

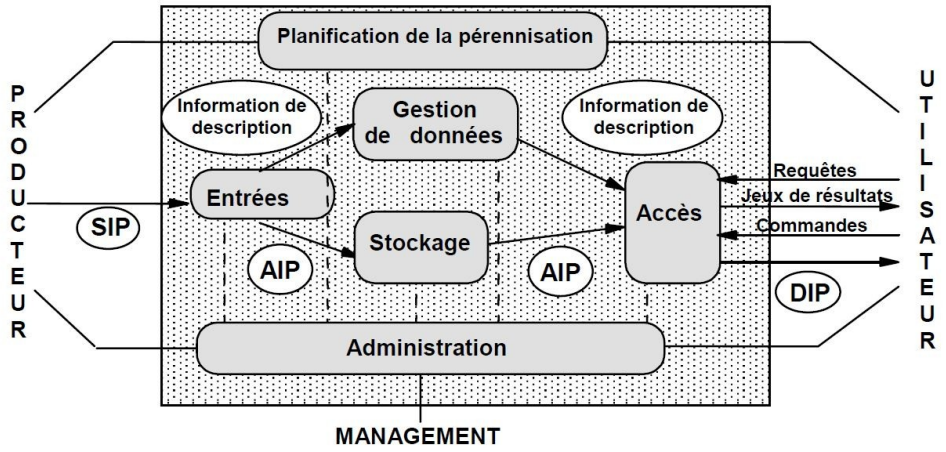
Les partenaires sont l'université d'Angers (Laboratoires Temos et Leria), le ministère de la Santé et des Solidarités (mission archives) et l'École nationale des chartes (centre Jean-Mabillon). Ont contribué à la réalisation de ce projet, sous la direction de Touria Aït el Mekki, maîtresse de conférences en informatique et spécialiste de Taln, et Bénédicte Grailles, maîtresse de conférences en archivistique : Tsanta Randriatsitohaina, ingénieure de recherche en informatique, Chafik Akmouche, stagiaire de deuxième année de master recherche en informatique de l'université d'Angers, Taimane Zerez, stagiaire de première année de master en informatique de l'université d'Angers. Pour la mission archives du ministère de la santé, le projet a bénéficié de l'appui, des conseils et des connaissances d'Anne Lambert, cheffe de la mission, et de Chloé Moser, cheffe de produit archifiltre. La connaissance du terrain, du fonctionnement du cabinet et plus généralement du ministère, était en effet indispensable à la réflexion méthodologique. Édouard Vasseur, professeur d'histoire des institutions, de diplomatique et d'archives contemporaines, a apporté son expertise en matière d'archivage électronique mais aussi ses connaissances sur les flux documentaires ministériels. Les différents partenaires se sont réunis une fois par mois, pour échanger sur l'avancement des travaux et valider les étapes intermédiaires.
Nous présenterons d'abord l'état de l'art et les objectifs (section 1), les corpus utilisés (section 2) pour ensuite traiter des questions méthodologiques (section 3) et des prototypes développés pour explorer et classifier (sections 4 et 5).
Un glossaire se trouve en fin de volume (p. 49) ainsi qu'une liste de technologies utilisées (p. 51).

1. Contexte et objectifs du projet

C'est à partir des années 1990 que les messageries électroniques se généralisent pour devenir un médium central de circulation de l'information dans les sphères personnelles et professionnelles. Dans le travail quotidien, elles sont devenues le support d'informations stratégiques et souvent les traces uniques de processus décisionnels (Bretesché S., Geffroy B., de Corbière F. 2018).
Depuis une dizaine d'années, les missions archives des ministères procèdent à la collecte systématique des boîtes mél des ministres et de leurs collaborateur.rices direct.es lors des remaniements gouvernementaux. Au sein des ministères sociaux (santé, solidarités, travail), les messages électroniques constituent en 2020 un part importante des documents collectés et représentent 45 % du volume des archives électroniques conservées. 200 comptes de messagerie ont déjà été collectés depuis 2012, la plus volumineuse représentant 70 Go de données.

1.1. Archivage des courriels

La pérennisation des courriels intéresse les archivistes depuis plusieurs années. Pour preuve, on trouvera plusieurs publications récentes (Prom C. 2019) et exemples de projets à l'étranger (ePadd [https://library.stanford.edu/projects/epadd], RATOM [https://ratom.web.unc.edu/]). En France, le programme interministériel d'archivage numérique Vitam [http://www.programmevitam.fr] a développé une réflexion (Programme Vitam 2013) et des outils notamment une librairie java, MailExtract, permettant d'extraire une arborescence de messages au format .eml des fichiers bruts exportés, et tenant compte des spécificités de la langue française (caractères accentués). La recherche s'est focalisée sur la préservation, ne s'intéressant que marginalement à la question de l'accès et de la restitution de l'information.
L'archivage des courriels s'inscrit dans le cadre de la norme ISO 14721:2012 ou norme OAIS (fig. 1) et du standard d'échange des données pour l'archivage (SEDA) promu par le service interministériel des Archives de France.



Légende : SIP : Paquet d'informations versé AIP : Paquet d'informations archivé
DIP : Paquet d'informations diffusé

Figure 1: Modèle conceptuel de données pour l'archivage et la préservation à long terme (source : norme Iso 14721 - Open archival Information System)

Chaque message constitue un submission information package (SIP) accompagné d'un manifeste xml conforme au SEDA.
Les missions archives des ministères ont en charge l'archivage intermédiaire. À ce titre, elles collectent des messageries qui ont ensuite vocation à être versées aux Archives nationales. La communication peut être sollicitée auprès des missions comme des Archives nationales. En août 2022, une recherche dans la salle de lecture virtuelle des Archives nationales permet d'identifier environ 130 messageries essentiellement issues du ministère de la Culture, mais de nombreuses messageries collectées par les missions n'ont pas encore été transférées. L'archivage des courriels est moins avancé au sein des services d'archives territoriaux ou des opérateurs. Ce sont principalement les services d'archives intermédiaires qui s'efforcent de les collecter mais de nombreux freins sont à l'œuvre, des freins techniques et organisationnels mais aussi méthodologiques.

Les organisations opèrent donc des choix en s'appuyant sur le niveau de décision du titulaire de la messagerie. Ainsi le ministère des Affaires étrangères ne collecte que 3 à 4 % des messageries existantes. La mission des ministères sociaux collecte quant à elle les messageries du cabinet, des directeur.rices et sous-directeur.rices d'administration centrale.

Le tri interne est en général laissé aux utilisateurs.

1.2. Traitement automatique de la langue naturelle (Taln) et courriels

L'explosion du volume de méls a rendu leur traitement automatique indispensable, notamment pour détecter les pourriels (Tang et al. 2013). L'exploration repose sur différentes méthodes. Elles peuvent utiliser des règles (Xia, 2020), de l'apprentissage (Nadjate et al., 2020) ou la combinaison des deux. L'efficacité des extracteurs automatiques des termes (Nazarenko et al. 2009) nécessite une évaluation. La catégorisation des courriels a été mobilisée dans le but de les organiser.
La question des contacts a aussi été explorée et touche à l'analyse du contenu : catégorisation de contacts ayant le même centre d'intérêt (Johansen, 2007) ; identification de contacts appartenant à une même communauté (Tyler et al., 2003).
Les interactions par méls pouvant être appréhendées comme un réseau, elles ont été explorées, en se fondant sur les statistiques des contributions de chaque contact au sein du réseau (Karagiannis, 2009) ou par l'apprentissage de l'objectif des méls envoyés, pour informer, enquêter et planifier (Lockerd, 2003). La détection d'évènements à partir des courriels a aussi été prospectée permettant l'identification de ceux (recrutement, discours, événement social) mentionnés dans les textes avec les détails comme la date, l'heure, le lieu. Des méthodes de reconnaissance d'entités nommées (Suárez et al. 2020) sont utilisées pour extraire ces informations qui sont ensuite soumis à validation (Nair et al., 2020).
Côté archives, le projet RATOM (université de Caroline du Nord) a travaillé à l'extraction d'entités nommées au moyen de librairies Taln dans un double but : identifier des informations potentiellement sensibles et préparer la diffusion. Notons que cette question

de l'accès aux contenus archivés – recherche au sein des corpus et communication à la demande – a été peu abordée, sauf aux États-Unis.

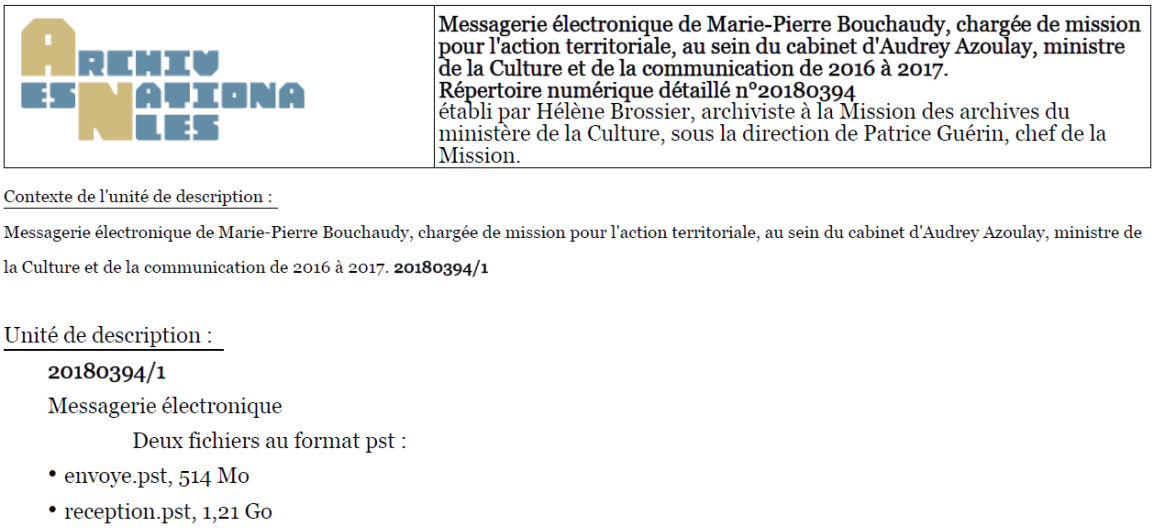
1.3. Analyse des besoins

L'intérêt de la conservation des courriels à des fins historiques et patrimoniales est reconnu (fig. 2).

Type de courriels	Exemples	Sort final
Courriels à valeur administrative, juridique, financière ou historique	- Politiques, procédures, directives, plans d'action	Conservés (attention, la DUA de certains de ces messages peut être très courte !)
	- Devis et soumissions	
	- Correspondance officielle	
	- Note de service à valeur stratégique	

Figure 2: Identification de la valeur des courriels (Source : Magnien A. 2012)

Les messageries des ministères et collectivités sont généralement produites via Microsoft Outlook et peuvent être exportées au format propriétaire pst. Si on souhaite ultérieurement consulter ces pst, il faut les réimporter dans Outlook, sous licence commerciale, ou utiliser une visionneuse pst, ce qui limite considérablement l'accès et l'utilisation de ces conteneurs. Les messageries sont de fait archivées dans des silos séparés. Un format conteneur type pst une fois intégré dans un SAE est une boîte noire : impossible de savoir ce qu'elle contient réellement (fig. 3).



Messagerie électronique de Marie-Pierre Bouchaudy, chargée de mission pour l'action territoriale, au sein du cabinet d'Audrey Azoulay, ministre de la Culture et de la communication de 2016 à 2017.
Répertoire numérique détaillé n°20180394
établi par Hélène Brossier, archiviste à la Mission des archives du ministère de la Culture, sous la direction de Patrice Guérin, chef de la Mission.

Contexte de l'unité de description :

Messagerie électronique de Marie-Pierre Bouchaudy, chargée de mission pour l'action territoriale, au sein du cabinet d'Audrey Azoulay, ministre de la Culture et de la communication de 2016 à 2017. **20180394/1**

Unité de description :

20180394/1

Messagerie électronique

Deux fichiers au format pst :

- envoye.pst, 514 Mo
- reception.pst, 1,21 Go

Figure 3: Exemple d'une description de messagerie [en ligne] sur le site des Archives nationales (consulté le 19 août 2022)

Or sans accès aux contenus des messages et des pièces jointes, il s'avère impossible :

- d'évaluer l'intérêt à long terme de chaque boîte mél comme des catégories de message qu'elle contient
- de jauger de l'originalité d'une boîte mail dans le réseau de mails de son organisation ou de son rôle pivot, autant d'éléments déterminants dans le processus de patrimonialisation des courriels
- d'effectuer un tri interne efficient
- de garantir l'exclusion des messages à caractère personnel inévitablement présents dans des messageries pourtant à caractère professionnel
- de proposer une description précise et une indexation pertinente
- d'analyser le flux informationnel d'une organisation
- de répondre aux besoins de recherche à caractère interne ou externe

De fait, les dispositifs socio-techniques déjà mis en place assurent la préservation mais ne prennent pas en compte l'accès au contenu des messages et pièces jointes, alors même que les deux tiers des demandes de communication portant sur les documents électroniques au sein de la mission archives des Ministères sociaux concernent des messageries (Dubois G. 2022). C'est une des raisons qui freinent l'archivage des messageries.

1.4 Solution cible

La démarche proposée allie des méthodes d'apprentissage automatique pour la classification et la catégorisation des contenus textuels et de l'intervention humaine dans un processus coopératif. La première étape voit l'élaboration de l'ontologie du domaine à partir du contenu des messageries et de corpus complémentaires. Outre l'identification des termes et la reconnaissance d'entités nommées, nous proposons une projection entre différents plans permettant de faire le lien entre l'ontologie, les entités nommées et les différentes thématiques (Alghamdi and Alfalqi 2015). Est pris en compte la totalité des informations des messageries (adresses, sujet, contenu, signature, pièces jointes). Les traitements automatiques permettent de construire un ensemble de connaissances à partir de textes et des ressources disponibles, de guider l'intervention humaine en assurant sa cohérence. Nous explorons comment différents types de ressources peuvent être exploités afin de valider des résultats ou pour spécialiser davantage des données préexistantes. Dans ce cadre, nous proposons une approche qui consiste à construire une liste des termes et des entités nommées que nous validons et que nous améliorons en injectant des connaissances issues des thesaurus spécialisés ou des métadonnées des messages (fig. 4).

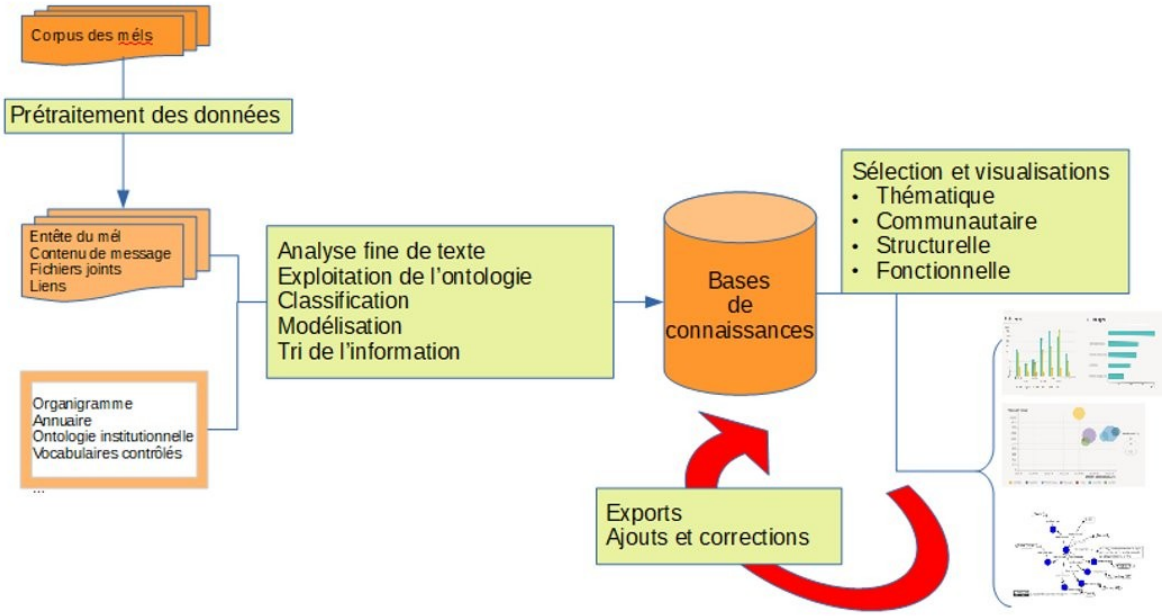


Figure 4: Solution-cible au démarrage du projet

2. Les corpus

Plusieurs corpus ont été mis à notre disposition pour réaliser ce projet que nous allons décrire (section 2.1) puis analyser (section 2.2).

2.1. Sélection des corpus

Quatre ensembles de corpus ont été mobilisés dans le cadre de cette expérimentation : des messageries, des organigrammes et annuaires, des discours de la ministre, ces trois catégories ayant été archivées par la mission, et des thesaurus génériques et de domaine.

2.2.1. Messageries

Les messageries confiées ont été choisies pour plusieurs raisons. Au nombre de deux que nous désignerons sous les noms de Brigitte et Anaïs, elles font partie des messageries les plus anciennes archivées par la mission et proviennent de conseillères en fonction au sein de cabinets successifs du ministère de la Santé, c'est-à-dire de collaboratrices personnelles choisies par la ministre, ayant pour mission de la conseiller et de l'assister dans la réalisation de l'ensemble de ses missions et jouant un rôle administratif et politique à ses côtés. Elles ont de fait un intérêt stratégique et patrimonial. Elles correspondent à une période sur laquelle il y a eu des demandes de recherche et de consultation en interne, notamment autour de la gestion de la crise sanitaire de la grippe A ou H1N1 ainsi que de celle des stocks de masque, mais aussi des sollicitations extérieures. Elles sont a priori relativement homogènes car issues des mêmes cabinets successifs d'une même ministre. Elles présentent un volume significatif (8636 messages). Elles semblent susceptibles de recevoir un avis favorable dans le cadre d'une demande de dérogation. En effet, les archives des cabinets répondent à un cadre légal spécifique : le

protocole de remise qui clarifie les conditions de traitement, de conservation et d'accès (Code du patrimoine, L213-4). Pour y accéder, il est nécessaire d'obtenir l'accord du / de la signataire du protocole, en l'occurrence, pour les messageries concernées, celui de Mme Roselyne Bachelot-Narquin[1].

Ces messageries sont issues d'outlook. Elles ont été exportées au format pst, la messagerie d'Anaïs constituant un seul pst et celle de Brigitte deux pst. Ces exports bruts en conteneur courriel ont été traités par la mission archives avec la bibliothèque de Vitam mailextract qui permet de générer une structure de répertoires et de fichiers, de normaliser l'extraction en eml et d'accompagner chaque niveau de cette hiérarchie d'un manifeste xml conforme au format Seda (ArchiveUnitMetadata), pour préparer les SIP avant transfert dans un système d'archivage électronique (fig. 5).

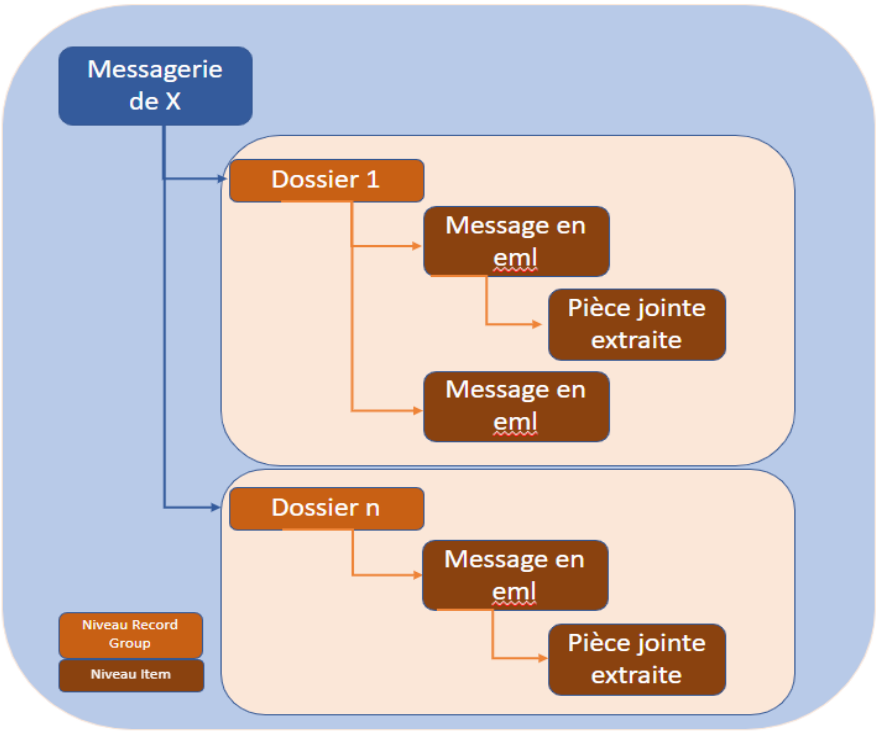


Figure 5: Messagerie après traitement du conteneur pst par l'outil Vitam Mail Extract

1 Nous remercions vivement Mme Bachelot-Narquin de nous avoir accordé l'accès à ces messageries. La question des dérogations a néanmoins été un point bloquant dans le déroulement du projet. En effet, la dérogation est individuelle et nominative et elle nécessite un temps de traitement. Il n'était pas possible d'enclencher la démarche pour les participants ponctuels qui devaient donc travailler sur d'autres corpus. La confidentialité des messages ne permet pas en outre de proposer des démonstrations complètes des interfaces développées.

Le projet se positionne clairement en aval de l'archivage des messageries. Il s'appuie sur des messages déjà documentés par un manifeste Seda et ramenés à un format eml. En revanche les formats des pièces jointes sont divers (pdf, doc, odt, jpeg…).

À l'intérieur de ce corpus, un sous-corpus « presse » a été identifié. Il regroupe des messages pour lesquels il n'y a pas de problèmes de confidentialité : messages adressés à la presse et comportant exclusivement des communiqués de presse ayant fait l'objet d'une publication. Ces messages ont été repérés, vérifiés et extraits manuellement. Ils proviennent tous de la messagerie de Brigitte.

2.2.2. Organigrammes et annuaires

À notre demande et afin d'améliorer l'identification des auteur.es et des destinataires des courriels comme des personnes nommées dans les textes, la mission a recherché des organigrammes et annuaires du ministère et des cabinets. 14 documents ont été mis à notre disposition aux formats jpeg, pdf et doc. Ils couvraient partiellement le cabinet et deux directions générales (santé et cohésion sociale) entre août 2010 et avril 2012. Ils ne permettent pas d'identifier et de tracer avec certitudes les personnes et les fonctions. La conseillère Anaïs par exemple ne figure sur aucun de ces organigrammes.

Par ailleurs, chaque document a une présentation qui lui est propre. En l'absence de régularités, une reprise manuelle a été inévitable.

On notera que les annuaires électroniques, type annuaires LDAP, et les répertoires de contact liés aux messageries et permettant de faire le lien entre une adresse et une personne et/ou entre une liste de diffusion et des adresses n'ont pas été archivés.

2.2.3. Discours

Pour augmenter le volume de données afin d'améliorer la classification, mais aussi pour disposer d'un ensemble non soumis à dérogation, d'autres données, cohérentes avec le corpus de messageries initiales dans les thématiques abordées comme par la période couverte (novembre 2010- mars 2012) ont été mises à disposition par la mission : il s'agit des projets de discours prononcés par la ministre dans ses fonctions aux formats doc et pdf. Des extraits plus ou moins complets, ainsi que des transcriptions d'entretiens radiophoniques, télévisuels ou des articles de presse écrite, sont également en ligne sur vie-publique.fr et ont été collectés.

2.2.4. Thesaurus

Une recherche sur l'existence de vocabulaires documentaires contrôlés a été menée. Différentes listes et thesaurus ont été identifiés mais qui ne couvraient que partiellement le domaine des messages. Le choix s'est finalement restreint aux thesaurus successivement utilisés au sein du centre de documentation du ministère de la santé (versions de 2014 et 2020). Seule une version pdf a pu nous être fournie. Au total, ces deux versions comprennent 7000 descripteurs cumulés de natures très différentes :

principalement des mots-matière mais également des noms de personnes morales et des acronymes.

2.2. Analyse

Une première phase a constitué en une analyse des corpus principaux. Elle a permis de vérifier la pertinence de la démarche.

2.2.1. Messageries

Au total les deux messageries (fig. 6) comprennent 8 636 messages reçus ou envoyés entre 2007 et 2011. Le contenu des messages en lui-même est court, avec une rédaction correcte, en moyenne 9 phrases par message et une médiane à 6 et 228 mots, mais le contenu reste répétitif.
Ces phrases comprennent des noms, des adjectifs et des verbes. Elles sont bien structurées. On est plus proche d'une rédaction écrite que d'une langue orale. Les pièces jointes sont présentes dans près de 30 % des messages pour une des boîtes et près de 70 % pour la seconde. 50% des mails contiennent moins de 3 phrases ce qui rend essentiel l'exploitation du contenu des fichiers joints.
Au total, c'est un réseau de plus de 2700 correspondants qui est concerné. On notera que les adresses fonctionnelles et listes de diffusion représentent 29 % de ce réseau. L'examen des adresses les plus fréquentes montre l'importance du trafic interne.

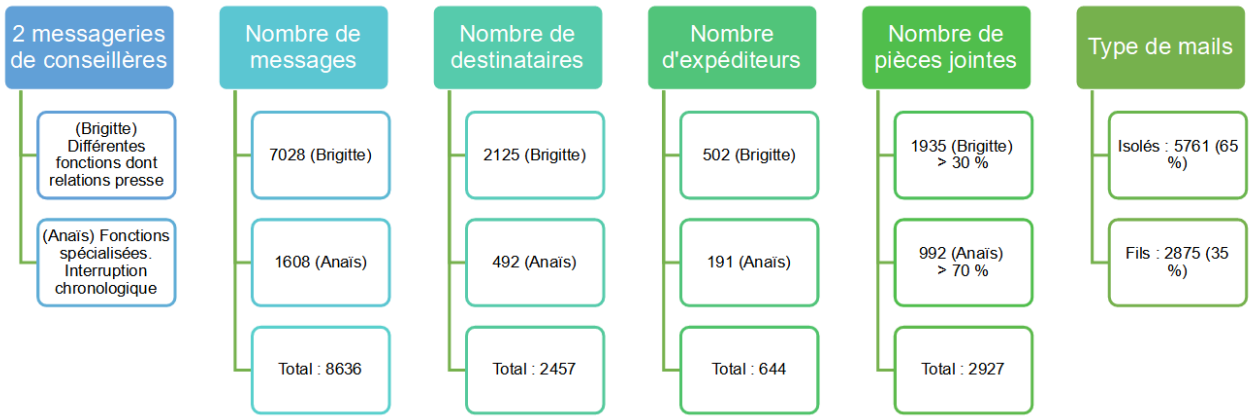


Figure 6: Données-clés des messageries de Brigitte et Anaïs

Nous considérerons toujours les messages comme un ensemble regroupant des métadonnées (adresses de l'expéditeur, du destinataire, des adresses en copie, date), un contenu (sujet et corps) et facultativement une signature et des fichiers joints (nommage et contenu) (fig. 7).

L'unité de réflexion est le message, son objet, le fichier et son titre.

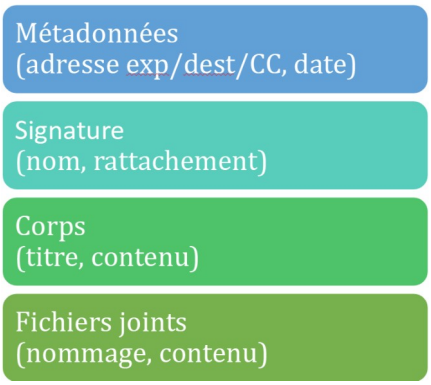


Figure 7: Structure des messages

2.2.2. Discours

Nous avions à disposition deux répertoires de discours : l'un rassemblant ceux collectés auprès du site vie publique et l'autre, ceux archivés par la mission et organisés en trois sous-répertoires correspondants à trois périodes : novembre 2010-décembre 2020, janvier 2011-décembre 2011 et janvier 2012-mars 2012. Les fichiers des discours sont généralement au format pdf ou doc. Le tableau ci-dessous résume les statistiques des répertoires de discours (fig. 8).

	Répertoires	Fichiers	Formats
Discours_en_ligne_vie_publique	1	533	pdf
Discours archivés (1re période)	40	74	Doc, pdf
Discours archivés (2e période)	376	672	Doc, pdf
Discours archivés (3e période)	44	51	Doc, pdf

Figure 8: Statistiques des discours

Pour pouvoir exploiter le contenu des discours, nous avons créé un module qui permet de convertir les fichiers pdf et doc en fichier texte. Le module prend en entrée le chemin des fichiers à convertir et produit en sortie les fichiers au format txt.

3. Méthodologie

Pour répondre aux besoins, nous proposons de mettre en œuvre une méthode composée des principaux modules suivants (fig. 9) :

- Un module de prétraitement : découpage de courriels, élaboration de réseaux de contacts, extraction des informations comme la fonction, le rattachement des personnes physiques ou morales à partir des adresses méls à relier avec les annuaires et les signatures. Une base de données qui contient l'ensemble de ces informations a été construite, ce qui permet de mettre en évidence les liens entre

les personnes, les fonctions et les services et facilite aussi différentes visualisations graphiques (section 3.1).

- Un module d'analyse fine de textes : extraction de termes, d'entités nommées (section 3.2).
- Un module de classification thématique afin de parcourir les messages selon les différents thèmes. Il permet d'avoir une vue plus globale de l'ensemble des messageries (section 3.3).

Figure 9: Méthodologie suivie

3.1. Pré-traitement

Il s'agit de constituer un corpus de travail permettant de manipuler les métadonnées et le contenu tout en uniformisant les pièces jointes en txt.

3.1.1 Découpage des courriels et appropriation des pièces jointes et corpus complémentaires

Dans le but d'exploiter les messages par les outils informatiques, nous avons créé des modules qui permettent d'extraire les différentes parties du message. Le module prend en entrée le dossier contenant les messages au format XML et crée en sortie des fichiers CSV :

- Messages.csv contenant l'id (OriginatingSystemId), l'objet (Title), le contenu (TextContent), la date d'envoi de chaque message (SentDate). Remarque : la date de réception (ReceivedDate) du message n'est pas toujours fournie.
- Annuaire.csv contenant l'annuaire général de la messagerie (la liste de toutes les adresses)
- Expediteur.csv contenant l'expéditeur (Writer) pour chaque message

- Destinaire.csv contenant les destinataires pour chaque message en précisant si l'adresse est le destinataire principal (Addressee) ou en copie (Recipient)
- Reponse.csv contenant les liens de réponses entre chaque message (mot-clé ReplyTo dans le fichier XML)
- PJ.csv contenant la liste des fichiers joints pour chaque message contenus dans les sous-dossiers dans les dossiers des messages

Les pièces jointes et des discours sont en général aux formats pdf, doc ou txt. Pour les exploiter, nous avons converti l'ensemble des contenus au format txt en utilisant l'outil Parser de la bibliothèque Tika. La forme initiale du texte est conservée : majuscule, minuscule, ponctuation par exemple. L'ensemble est segmenté en phrase pour faciliter la phase d'extraction.
Des difficultés ont été rencontrées. Lorsque le texte est présenté dans le format d'origine en colonnes, les outils n'arrivent pas à gérer les coupures de mots et à reconstituer les lignes. Il a fallu travailler au nettoyage des résultats produits par Tika, en utilisant des RegEx.
Les images, peu abondantes dans nos corpus, n'ont pas subi de traitement.

3.1.2. Les adresses

On trouve deux grandes catégories d'adresses :

- adresses individuelles : chaque adresse correspond à une personne physique avec un nom, un prénom, une fonction à un intervalle de date et un rattachement à un intervalle de date
- adresses fonctionnelles. Ce type d'adresses recouvre en fait deux cas de figure. L'adresse peut correspondre à un service (une fonction, un rattachement, un intervalle de dates) ou à une liste de diffusion. Remarque : la composition des listes de diffusion n'a pas été conservée. On ne peut donc pas relier une liste de diffusion aux personnes physiques qui y étaient rattachées.

Pour catégoriser les différentes adresses, nous avons eu recours à des expressions régulières correspondant au corpus concerné : FullName en majuscules ou point précédant l'arobase.

3.1.3. Fonctions et rattachement

Les fonctions et rattachements institutionnels peuvent partiellement être déduits des adresses.

En effet, dans le cas de nos messageries, on observe que les adresses fonctionnelles sont généralement de type fonction@rattachement. Pour les adresses individuelles, l'interprétation peut être guidée par la signature. Celles du ministère suivent le modèle Prénom Nom, fonction, rattachement. On peut aussi recourir à des sources externes (organigramme, annuaire) ou compléter manuellement.

3.2. Analyse fine de texte

Différents extracteurs appuyés sur des bibliothèques python ont été développés en environnement linux.

3.2.1 Extraction de termes

Un terme est un mot ou un groupe de mots représentant un concept spécifique d'un domaine. Par exemple, « aide sociale » ou « allocations familiales » sont des termes. Nous avons adapté un extracteur de termes existant (Cadiou A. extracteur PLDAC). À noter que cet extracteur ne traitant que de petits volumes, il est nécessaire de découper le corpus en plusieurs fichiers de traitement. L'extraction peut être paramétrée en jouant sur les variables suivants : lemmatisation, méthode de calcul du scoring (TF-IDF Standard, TF-IDF LOG, fréquence), nombre de mots (minimum, maximum).
Une phase de validation manuelle est nécessaire. Elle peut s'opérer avec les trois options suivantes ou en les combinant : validation par la fréquence, par nombre de mots ou à l'aide d'une liste de référence déjà validée par l'expert. On peut par exemple s'appuyer sur les termes figurant dans les thesaurus existants.

3.2.2 Extraction d'entités nommées

Les entités nommées sont des mots ou groupes de mots généralement assimilées aux noms propres (noms de personne physique ou morale), noms de lieu ou encore à des valeurs (date, heure...).
Leur extraction automatique peut être effectuée par des méthodes linguistiques reposant sur les catégories grammaticales, ou par des méthodes statistiques reposant sur les fréquence et distribution des mots dans le texte, ou encore par la combinaison des deux (TermSuite, Cram and Daille, 2016). Les sorties proposent des candidats-termes à soumettre à validation.
La reconnaissance d'entités nommées peut être effectuée à partir des bases de connaissances ou par méthode d'apprentissage. Parmi les outils qui prennent en charge le français et qui satisfont au besoin de confidentialité des messageries, nous avons testé plusieurs bibliothèques et avons utilisé les bibliothèques python SpaCy (Honnibal and Montani, 2017) et NLTK avant de filtrer via les regex (fig. 10).

	Messages	Fichiers joints	Discours
Personne	8 881 (28,9%)	18 208 (26,8%)	3 444 (23,8%)
Organisation	5 587 (18,2%)	13 267 (19,5%)	2 896 (20.0%)
Lieu	6 016 (19,6%)	11 573 (17,0%)	3 029 (20,9%)
Autres	10 240 (33,3%)	24 912 (36,7%)	5 098 (35,2%)
Total	30 724	67 960	41 157

Figure 10: Statistiques relatives à la reconnaissance d'entités nommées avec NLTK

La phase de validation s'organise en deux temps :

- par comparaison avec les descripteurs des thesaurus, puis avec la liste des abréviations validées (section 3.2.3), enfin en confrontant les noms de personne avec les métadonnées des messages et les annuaires
- en gardant les termes et les entités nommées associées, suivant la stratégie de Omrane et al., 2011 (section 4).

La projection des thesaurus métier qui recensent aussi des personnes morales comme la confrontation avec les annuaires et organigrammes s'avèrent peu utiles : peu de recoupements, recensement trop restreint eu égard au contenu des messages et pièces jointes.

3.2.3 Extraction des abréviations, acronymes, sigles

Un des enjeux de la contextualisation des messages est l'identification et la résolution des acronymes. Les étiqueteurs morphosyntaxiques proposent une catégorie « abréviations ». Celle-ci ne recense malheureusement qu'une petite partie des sigles que l'on trouve dans d'autres catégories comme les noms de personnes morales par exemple. Pour établir la liste la plus complète possible, nous avons extrait les abréviations du corpus, grâce à l'étiquetage morpho-syntaxique proposé par TreeTagger, via l'étiquette ABBR pour abréviation ainsi que la reconnaissance d'entités nommées via l'étiquette ORG (fig. 11).
Nous nous sommes appuyé.es sur les descripteurs des thesaurus que nous avons projeté sur les messages, sur les sujets et les pièces jointes, mais de nombreuses abréviations n'y figurent pas. Nous avons filtré et nettoyé via la mise en œuvre de règles (réduction du bruit sur l'étiquette ABBR de 70 %) et sur un dictionnaire des sigles et acronymes de l'administration (Maudry C. [https://www.data.gouv.fr/fr/datasets/dictionnaire-des-sigles-et-acronymes-de-ladministration/]) croisé avec l'identification d'entités nommées. Plus de 400 sigles ont été identifiés.

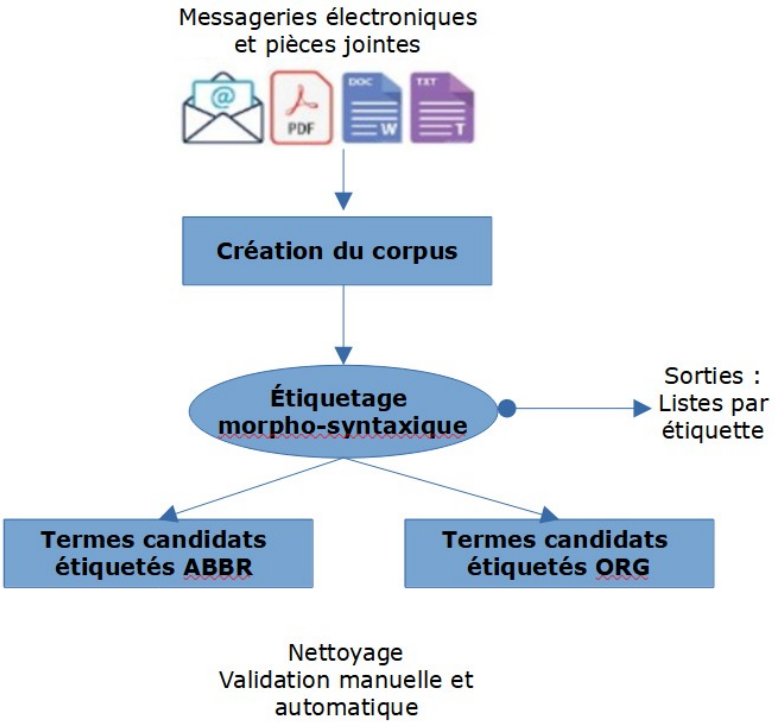


Figure 11: Schéma de la recherche des sigles et acronymes

3.3 Classification

Différentes méthodes de classification existent :

- À base de règles / de patrons
- Apprentissage non supervisé (Clustering). L'apprentissage non supervisé consiste à regrouper des données en se basant sur les caractéristiques communes à ces données. Il a l'avantage de ne nécessiter ni de classes prédéfinies ni de données déjà étiquetées. Toutefois, le résultat est imprévisible et ne permet pas d'obtenir des classes spécifiques.
- Apprentissage supervisé (Machine learning). Il consiste à fournir aux algorithmes des données étiquetées afin d'en déduire les caractéristiques propres à chaque classe et identifier la classe pour une nouvelle donnée. Une autre solution consiste à utiliser un modèle déjà pré-entraîné.

Les deux premières stratégies étaient privilégiées car facilement reproductibles dans d'autres environnements.

Les résultats obtenus avec la classification non supervisée nous ont amené.es à réfléchir à une stratégie de guidage pour les améliorer. Cette option est originale. Elle vise à améliorer les possibilités de recherche en s'appuyant sur un thème et sur les termes qui lui sont associés automatiquement ou manuellement par l'expert.

3.3.1. Premières tentatives : des résultats non concluants

Pour retrouver une information, on peut utiliser différentes méthodes. Une approche basique consiste à rechercher des mots-clef. Sur de gros volumes, elle produit autant de bruit que de silence. Il faut donc réussir à accéder au contenu des messages complétés par celui des pièces jointes. Nous avons donc testé plusieurs méthodes d'apprentissage et différents algorithmes pour aboutir à une méthodologie qui enchaîne différentes étapes.

3.3.1.1 Apprentissage non supervisé

En l'absence de données d'apprentissage, la classification peut se faire en mode non supervisé.

Dans un premier temps, nous avons développé le module de classification en utilisant les méthodes existantes pour regrouper les courriels avec leurs pièces jointes selon leurs similarités et différences en appliquant un modèle de clustering K-Means qui permet de séparer les données en k groupes. Le regroupement s'opère en fonction des similarités entre les données (distance TF-IDF).

Plusieurs essais ont été effectués : en donnant en entrée au modèle les courriels et pièces jointes représentés par tous les mots qu'ils contiennent, puis uniquement des noms, verbes, adjectifs, adverbes, ou par les termes identifiés afin qu'il les distingue par groupes. Les premiers résultats de cette classification n'ont pas été probants (fig. 12). Les centroïdes des clusters sont très proches. La similarité entre le contenu des messages ne permet pas au modèle d'identifier des points suffisamment discriminants pour orienter la classification. De plus, nous travaillons sur des textes très spécialisés.

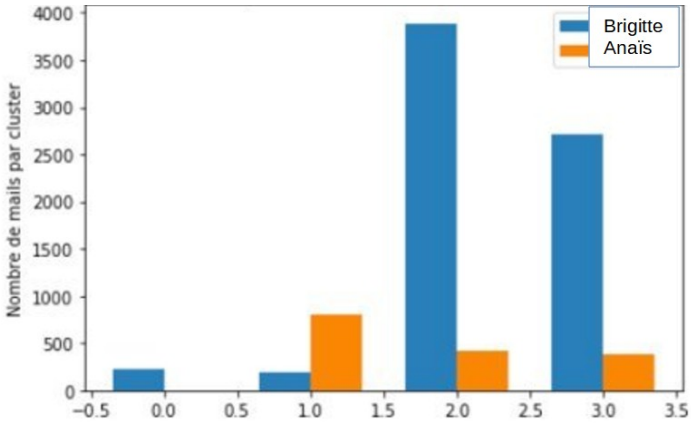


Figure 12: Répartition des mails et pièces jointes par cluster (algorithme K-Means)

Une tentative similaire a été menée avec le logiciel Iramuteq avec des résultats aussi peu concluants (fig. 13).

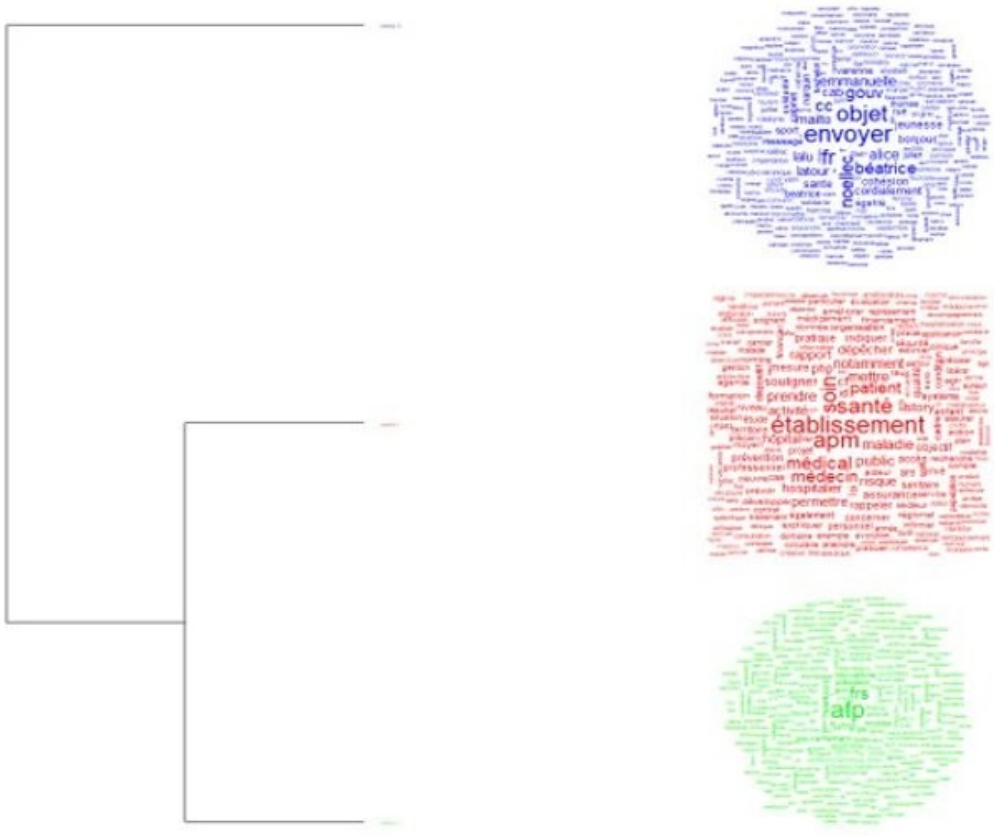


Figure 13: Clusters générés par Iramuteq

3.3.1.1 Projection des thesaurus

Pour améliorer la classification, nous avons exploré une voie qui consistait à s'appuyer sur les thesaurus et vocabulaires contrôlés existants au sein du ministère.
Les descripteurs du thesaurus ont été projetés sur les messages, leur objet et les pièces jointes. 60 % des 7000 descripteurs sont utilisés dans le corpus des mails. Malgré ce chiffre, le thesaurus n'est pas d'un grand secours car il ne suffit pas à discriminer des messages.

3.3.2. Améliorer la classification : trouver des relations entre les termes et les thèmes

Pour améliorer la classification, nous sommes parties de l'idée qu'un terme-clef ne se suffisait pas à lui-même mais que ce terme-clef était précisé et développé par tout un champ sémantique qui lui est propre, susceptible de d'orienter utilement la classification.
Nous avons donc procédé en deux temps : en identifiant des termes reliés à des thématiques (section 4.3.1) puis en classifiant les messages grâce aux termes afin de les relier aux thématiques (section 4.3.2).

3.3.2.1. Guider la classification en créant des nuages de termes

Nous avons amélioré le processus en réalisant une classification guidée par une liste de 70 thèmes structurés en trois niveaux par les spécialistes, donc en regroupant les classes autour de ces thèmes. La liste thématique a été fournie par la mission Archives qui s'est

appuyée sur les textes infra-réglementaires. De fait, on peut la lire comme un triptyque attributions/missions/actions et programmes, usuellement utilisé par les archivistiques pour analyser les flux documentaires et structurer les tableaux de gestion des archives.
Nous avons donc plusieurs « niveaux » de thèmes où un « niveau » représente une facette de classification. Un « niveau » donné contient plusieurs thèmes et chaque thème représente un cluster dans la classification. Chaque thème est lié avec plusieurs termes.
L'approche mise en œuvre ici se distingue des précédentes par la volonté de délimiter un thème par un nuage de termes, c'est-à-dire un ensemble de termes en relation directe ou indirecte avec un terme parent. L'objectif est de générer des nuages de mots présentant une similarité forte pour les utiliser ensuite dans la classification des messages.
Pour établir ces relations entre thèmes et termes, nous avons eu recours au plongement de mots ou plongement lexical. Les avancées récentes de l'apprentissage automatique ont permis de mettre en œuvre des méthodes permettant la représentation des mots en fonction de leur contexte. On parle de plongement lexical. Celui-ci capture les similarités sémantiques et syntaxiques des mots en fonction du contexte dans lequel ils apparaissent dans le texte. Pour notre projet, nous avons testé Fastext (Bojanowski et al., 2017), CamenBERT (modèle de BERT entraîné en français) et Word2Vec afin de comparer les termes ayant une similarité élevée, dans un premier temps, pour identifier les termes proches du thesaurus, dans un second temps, pour relier les thématiques des expertes avec des termes.
Les modèles de plongement lexical proposent une représentation mathématique sous forme de vecteur d'un mot entier (word2vec) ou d'une chaîne de caractères (Fastext). Les deux sont basés sur des fenêtres de contexte à dimensionner. Les relations sont ensuite établies par comparaison des vecteurs et recherche de similarité cosinus. BERT (Bidirectional Encoder Representations from Transformers) et ses dérivés (comme CamenBERT) s'appuient sur des transformateurs, des modèles de représentation textuelle où un même mot est représenté par des vecteurs différents selon le contexte, ce qui permet de prendre en compte des sens différents.
Plusieurs cas de figure ont été testés : messages avec ou sans pièce jointe, avec ou sans le corpus de discours, avec ou sans lemmatisation, avec ou sans racinisation, avec ou sans modèle pré-entraîné. La difficulté principale réside dans le volume de données, faible dans le cas de nos messageries. Or plus le volume est grand, plus le résultat est pertinent, d'où l'adjonction du corpus de discours pour entraîner les modèles.
Après recherches et tests, nous avons choisi Word2Vec. Facile à utiliser, paramétrable, adaptable et léger, on peut l'utiliser avec des machines de performance moyenne, contrairement à CamenBERT par exemple, qui possède l'avantage d'être entraîné sur des corpus en français très volumineux, mais qui est très lourd et nécessite des machines très performantes (https://camembert-model.fr/).
Word2Vec est un ensemble d'algorithmes de plongement lexical ou word embedding, non supervisé, qui permet de représenter les mots d'un texte par des vecteurs de nombres réels. Il utilise un réseau de neurones à 3 couches et a été conçu pour prédire les mots voisins d'un mot donné dans un texte ou prédire un mot à partir d'un ensemble de mots.
Le corpus doit subir une préparation spécifique : texte basculé en minuscule lemmatisation, suppression de la ponctuation, des mots vides, des mots ayant moins de 3 caractères.

Word2Vec peut être utilisé soit en utilisant un modèle pré-entraîné, soit en entraînant son propre modèle. Dans ce second cas de figure, la pertinence du résultat est corrélée au volume de données.
Cette option a été testée en jouant sur les paramètres essentiels, à savoir la dimensionnalité des vecteurs de mots (vector_size) et la taille de la fenêtre utilisée par Word2Vec quand il parcourt le texte (window). Le volume de données dont nous disposions n'étant pas suffisant, cette démarche n'a pas été concluante. Néanmoins, le prototype développé (section 5.3) permet de réaliser cette opération qui pourrait amener de meilleurs résultats dans une autre configuration de corpus.
Nous avons donc traité notre corpus en utilisant un modèle pré-entraîné en langue française, mis à disposition par Jean-Philippe Fauconnier (Fauconnier 2015), lequel propose plusieurs modèles entraînés sur Wikipédia, divers dictionnaires, etc. Le modèle utilisé est un modèle générique qui identifie des termes composés de plusieurs mots et a été entraîné sur des corpus différents : *frWac no_postag phrase_500 cbow_cut100*.
On recherche les termes qui se trouvent à la fois dans le modèle pré-entraîné et dans le corpus. Pour chaque terme du modèle présenté sous forme d'un vecteur, on récupère les termes (vecteurs) les plus proches en termes de distance (similarité cosinus).
En entrée, on a le modèle entraîné, le corpus à traiter, la liste des thématiques donnée par l'archiviste et la profondeur de l'arbre de recherche. À la fin de la recherche, pour chaque terme, on dispose d'une liste de termes similaires.
Par exemple, nous retrouvons le terme sida dans la même classe que ist, vih, maladie, vhc etc.; le terme médicament est associé à usage, produit, pharmaceutique, generique, prescrit, tamiflu etc.
Chaque thématique dispose, à l'issue du processus, de son nuage qu'on peut dès lors utiliser pour classifier les messages (section 3.3.3).

3.3.2.2. Classification à base de règles ou de patrons

Les patrons permettent d'identifier les relations sémantiques qui relient des entités lexicales : hyperonymie, holonymie, méronymie, causalité, possession… Un patron caractérise une relation et une seule mais une relation peut être détectée par plusieurs patrons. 12 relations et 111 patrons (en annexe) ont été créés avec la volonté de rester assez générique et en ajoutant la possibilité d'un mot facultatif.

Exemple d'une relation d'hyperonymie :

L'hépatite A est une maladie
(terme) (être) (det) (terme)

L'hépatite A est aussi une maladie
(terme) (être) (adv) (det) (terme)

Pour traduire des patrons, les « matchers » proposés par SpaCy ont été utilisés. Ils fonctionnent sur la base d'expressions régulières, en utilisant l'analyse morphosyntaxique de la phrase (verbe, adverbe, déterminant etc.). Notons que certaines fonctions sont sensibles à la casse typographique. Le corpus doit subir un pré-traitement : segmentation en phrase, extraction des termes et des entités nommées, nettoyage (fig. 14).

Patron correspondant à l'exemple précédent :

$$[\{"LEMMA": "être"\}, \{"POS": "ADV", "OP": "?"\}, \{"POS": "DET"\}]$$

Ce patron identifie les phrases qui commencent par un terme suivi du verbe être, suivi d'un adverbe, suivi d'un déterminant et d'un autre terme et relie le premier terme au deuxième.

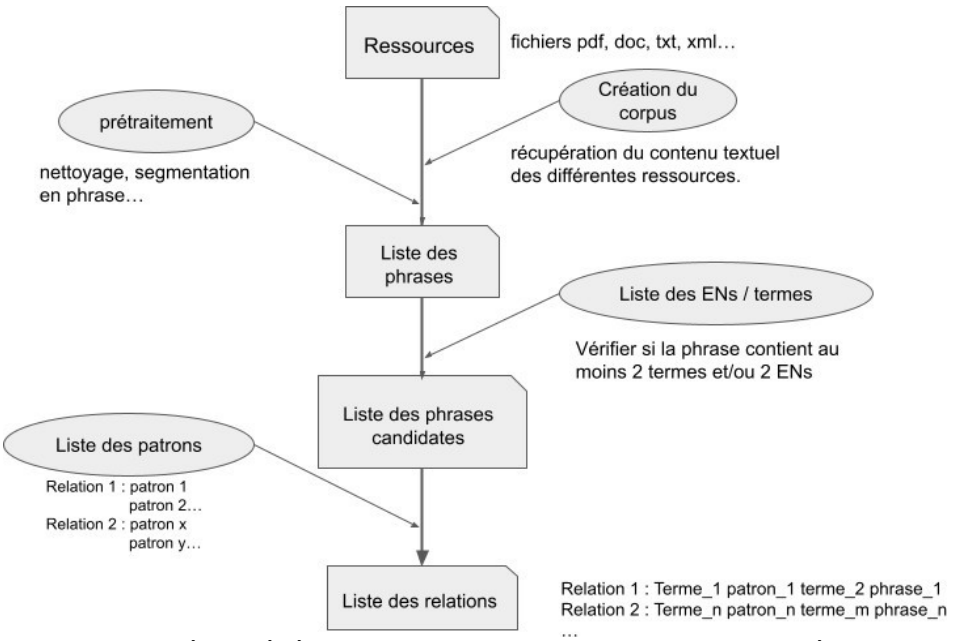


Figure 14: Schéma général de la recherche des relations par la méthode des patrons sémantiques (source : Akmouche C. 2022)

ENs = entités nommées

Cette méthode peut paraître peu efficace dans notre cas. Les messages comprennent en effet un nombre limité de phrases peu complexes, ce qui réduit les chances de trouver des relations par ce biais. Par ailleurs, elle est assez chronophage, puisqu'il faut créer les patrons manuellement. Néanmoins, elle ne nécessite pas de données étiquetées pour l'apprentissage ni de corpus volumineux : une seule occurrence suffit à identifier une relation.

3.3.3. Classer les messages grâce aux nuages de termes

L'objectif poursuivi est d'associer un message n à des thématiques :

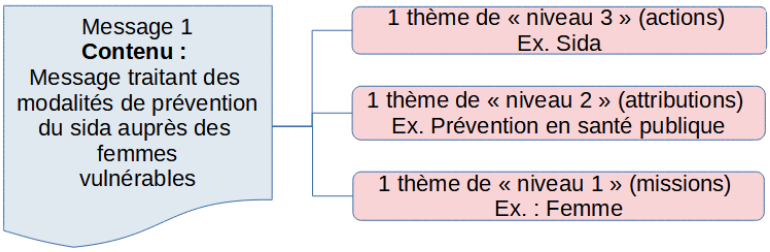


Figure 15: Identification des thèmes d'un message

Pour ce faire, nous avons opté pour du clustering, de la classification non supervisée. En entrée, nous disposons d'un corpus constitué de messages et de pièces joints et des nuages de termes des thématiques auquel nous avons appliqué une méthode de plongement de documents, Doc2Vec, qui est une généralisation de Word2Vec.

Doc2Vec permet de représenter chaque unité du corpus (message + pièces jointes) par un vecteur de nombre réel. On lance ensuite un calcul de similarité (fonction most_similar) entre les vecteurs des messages et les vecteurs des thématiques. On peut paramétrer différents paramètres dont la dimensionnalité des vecteurs (vector_size), le nombre d'itérations (epochs), la faculté d'ignorer les mots dont la fréquence totale est inférieure à une valeur choisie (min_count).

Un fichier de sortie CSV est créé qui peut ensuite être utilisé et visualisé.

4. Interfaces de visualisation

Les interfaces de visualisation permettent de s'orienter les messages provenant d'une ou plusieurs messageries, d'explorer leur contenu et celui des pièces jointes. Conçues en python, elles peuvent être déployées sur Linux ou Windows avec les mêmes bibliothèques. Elles sont complémentaires. L'interface de base est une application desktop très légère à installer en local. Elle peut être utilisée seule. L'interface avancée s'appuie sur une base de données relationnelle.

4.1. Interface principale

Elle permet au choix depuis le menu de la page d'accueil (fig. 16) :

- d'importer un répertoire de fichiers, par exemple après traitement par MailExtract. On peut importer une ou plusieurs messageries. Il faut donc préalablement préparer le dossier en fonction des messageries que l'on souhaite explorer

- d'interroger la base de données relationnelle. Plus de fonctionnalités sont alors offertes en termes d'exploration et de visualisation et une plateforme d'annotation est mise à disposition (fig. 17).

Figure 16: Menu principal

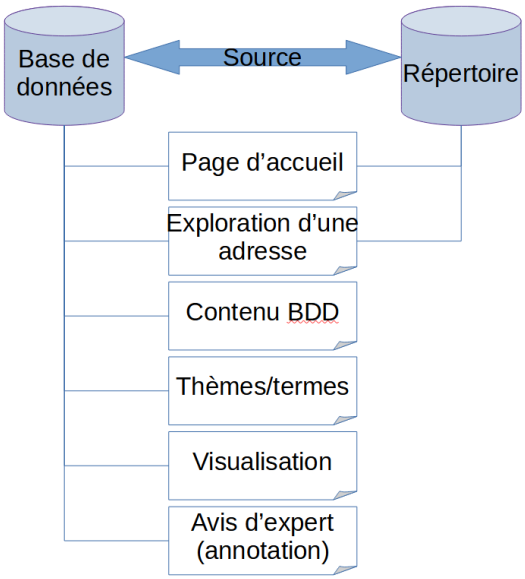


Figure 17: Schéma d'accessibilité

4.1.1. Page d'accueil

À partir de cette page (fig. 18), on peut interroger les métadonnées des messages, leur contenu et filtrer les résultats à partir des critères suivants [1, 2] : nom, sujet, destinataire, expéditeur, date, CC, contenu (message et pièces jointes). Une recherche avancée est aussi possible.

> Exemple de syntaxe : date=2010||2011 ; sujet=mot1 ; contenu=mot2||mot3
> (affiche dans la liste tous les messages des années 2010 ou 2011, présentant le mot 1 dans le libellé du sujet et les mots 2 ou 3 dans le corps des messages)

La liste des messages peut être travaillée manuellement (boutons en bas à gauche de la page principale [3]) via une fenêtre : affichage de la liste, fonction de suppression de la liste, fonction de mise en attente, fonction de rétablissement d'un message supprimé de la liste.
Dans la partie centrale de la page d'accueil [4], des boutons permettent d'afficher le contenu des messages si ils sont en html et les pièces jointes dans leur format archivé (pdf, format image, word). Il est possible de naviguer entre les pièces jointes et de les sauvegarder sur son bureau si besoin.

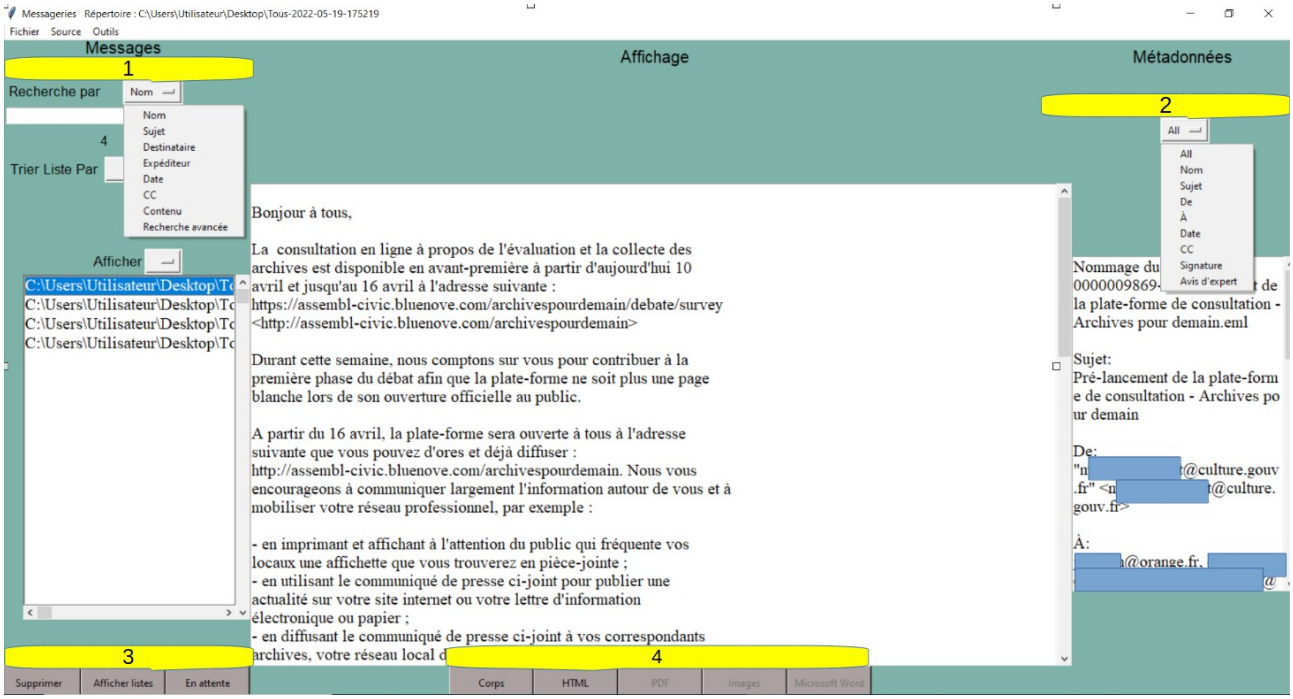


Figure 18: Page principale des interfaces de visualisation

4.1.2. Explorer une adresse

Cette fonction est accessible quelle que soit la source choisie. Elle permet de visualiser les relations entre une adresse et ses correspondants (messages reçus et envoyés) sous forme de graphe (fig. 19). Un graphe dynamique et paramétrable peut être généré grâce à Pyvis (fig. 20). Des fonctions de filtrage par adresses et dates et de recherche permettent d'affiner le graphe dont l'image peut être exportée.

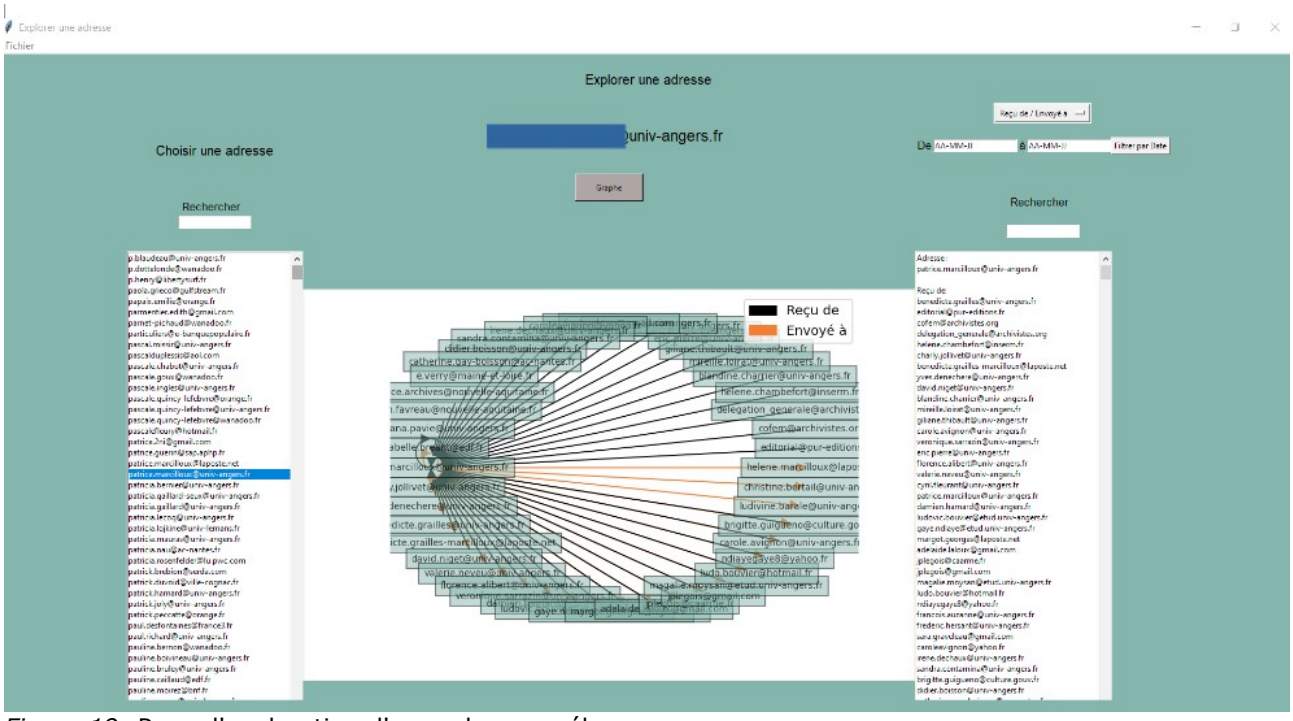


Figure 19: Page d'exploration d'une adresse mél

Figure 20: Exemple de graphes dynamiques

4.2. Interface avancée

Si sur la page d'accueil, on choisit comme source la base de données, une fenêtre s'ouvre pour se connecter à celle-ci. Une fois la connexion effectuée, on a accès à des fonctionnalités supplémentaires :

- Contenu BDD : permet de modifier le contenu des tables de la base (section 4.2.1)
- Thèmes/termes : la page permet de modifier les thèmes et les termes et leurs relations (section 4.2.2)
- Visualisation BDD (section 4.2.3)
- Avis d'expert : permet à l'expert de classifier des messages manuellement et de les enregistrer dans la base (section 4.2.4)

4.2.1. Fonction de modification de la base de données

Depuis cette page (fig.21), on peut modifier les tables message, personne physique, pj, thème, terme, thème-terme à l'aide de deux menus d'option, l'un pour choisir la table ([1]), l'autre la colonne dans la table ([2]). Une barre de recherche est à disposition.

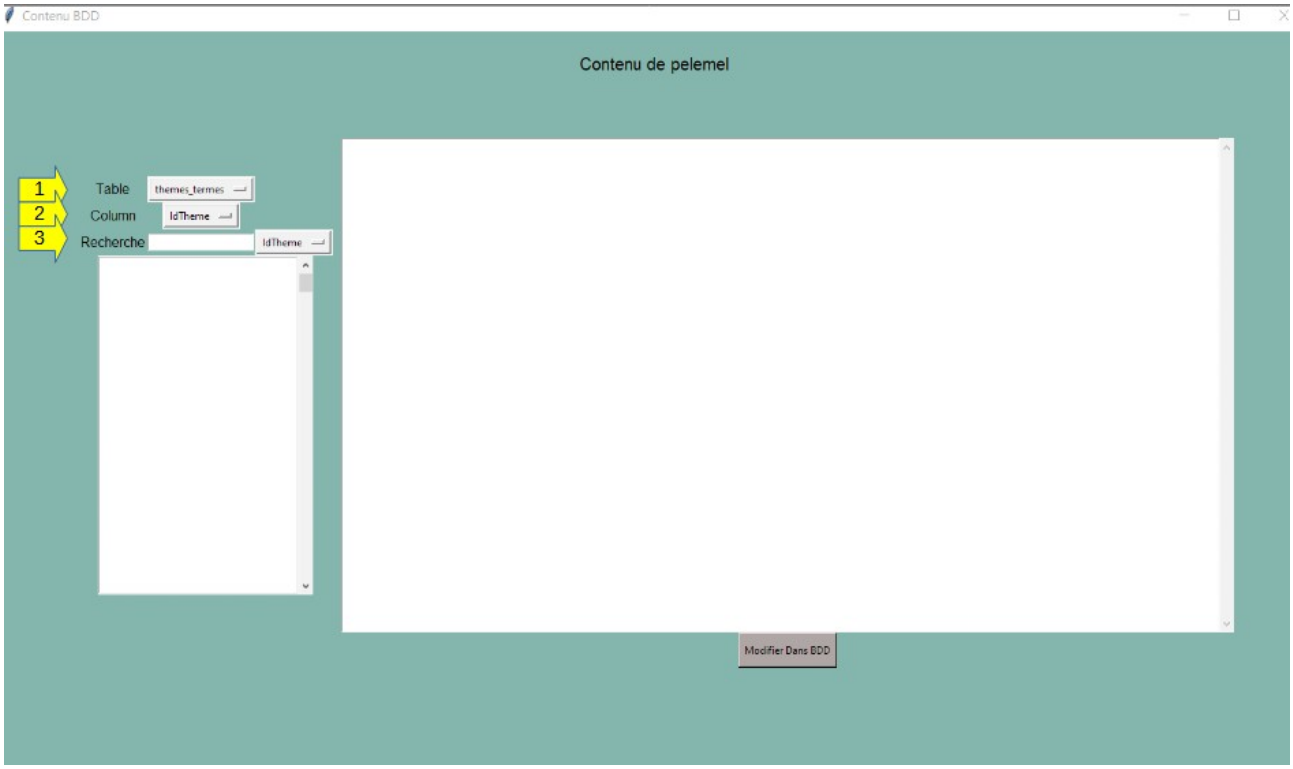


Figure 21: Page de modification du contenu de la base

4.2.2. Fonction thèmes termes

La page thèmes-termes (fig. 22) permet d'afficher et de modifier la classification guidée. Elle permet d'afficher les thèmes, pour chaque thème les termes associés et les messages classifiés. L'expert.e peut décider d'intervenir sur chaque élément des listes : ajout, modification, suppression.

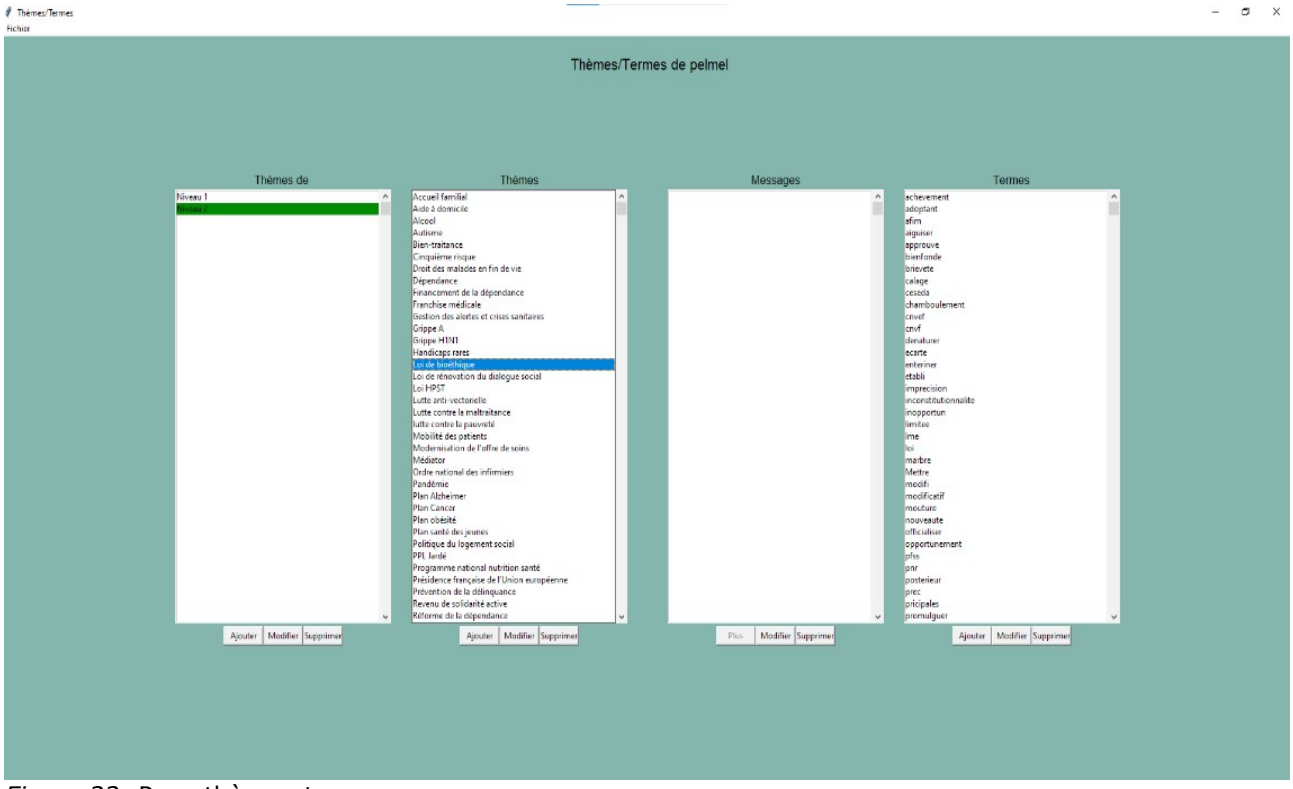


Figure 22: Page thèmes-termes

4.2.3. Fonction d'exploration

La page (fig. 23) permet d'explorer et visualiser les messages de la base de données pour une période donnée avec des possibilités de filtrage par adresse(s), par recherche dans le contenu (corps du message et des pièces jointes), par thème (avec un menu d'option pour naviguer dans les thèmes). On peut choisir un ou plusieurs filtres.

Par défaut, les dates de début et de fin sont celles du message le plus ancien et du message le plus récent et l'échelle horizontale est d'un mois. Cette échelle peut être modifiée. On peut notamment générer des graphiques qui permettent de voir la fréquence et l'évolution chronologique d'une thématique (le thème et son nuage de mots) dans les sujets de conversation.

Le graphique obtenu peut être sauvegardé et exporté.

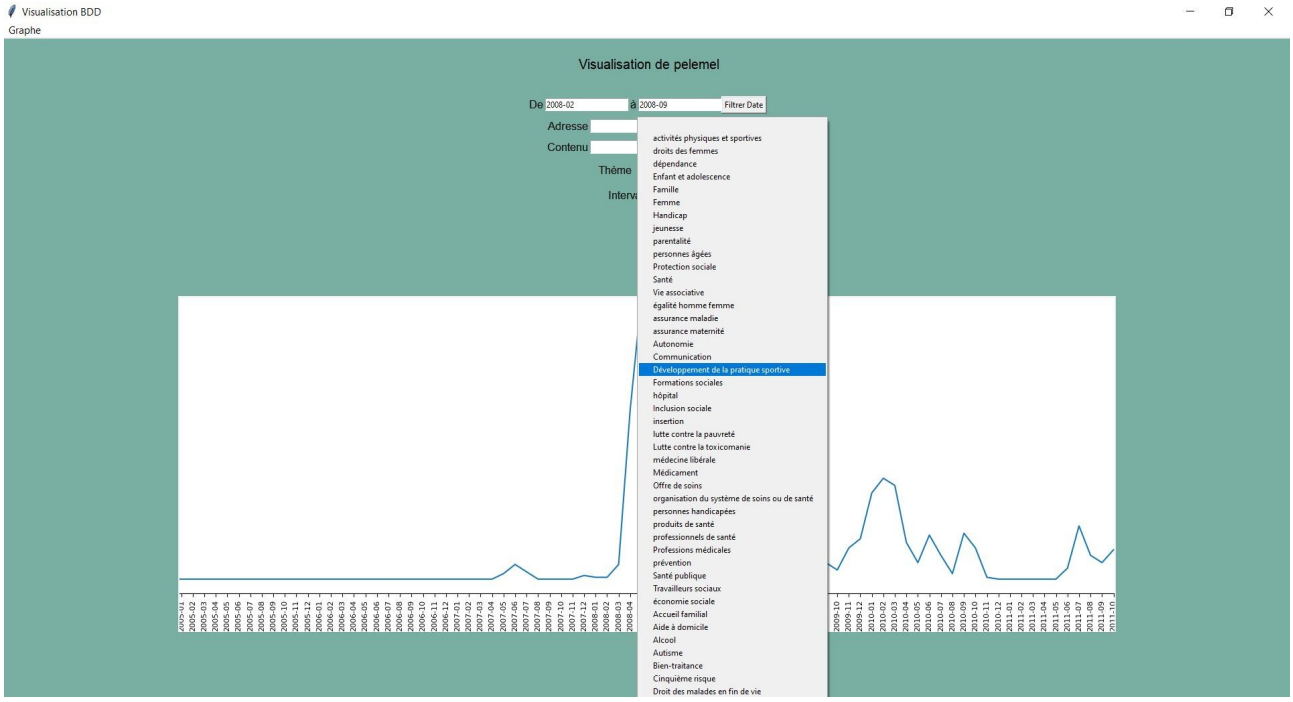


Figure 23: Page de visualisation

4.2.4. Fonction avis d'expert.e

Il s'agit d'une plateforme d'annotation manuelle pour la classification supervisée (fig. 24). L'expert.e choisit un nombre de messages qui lui sont fournis par l'application de manière aléatoire. Il peut travailler sur l'ensemble des messageries téléchargées ou sur une messagerie spécifique.

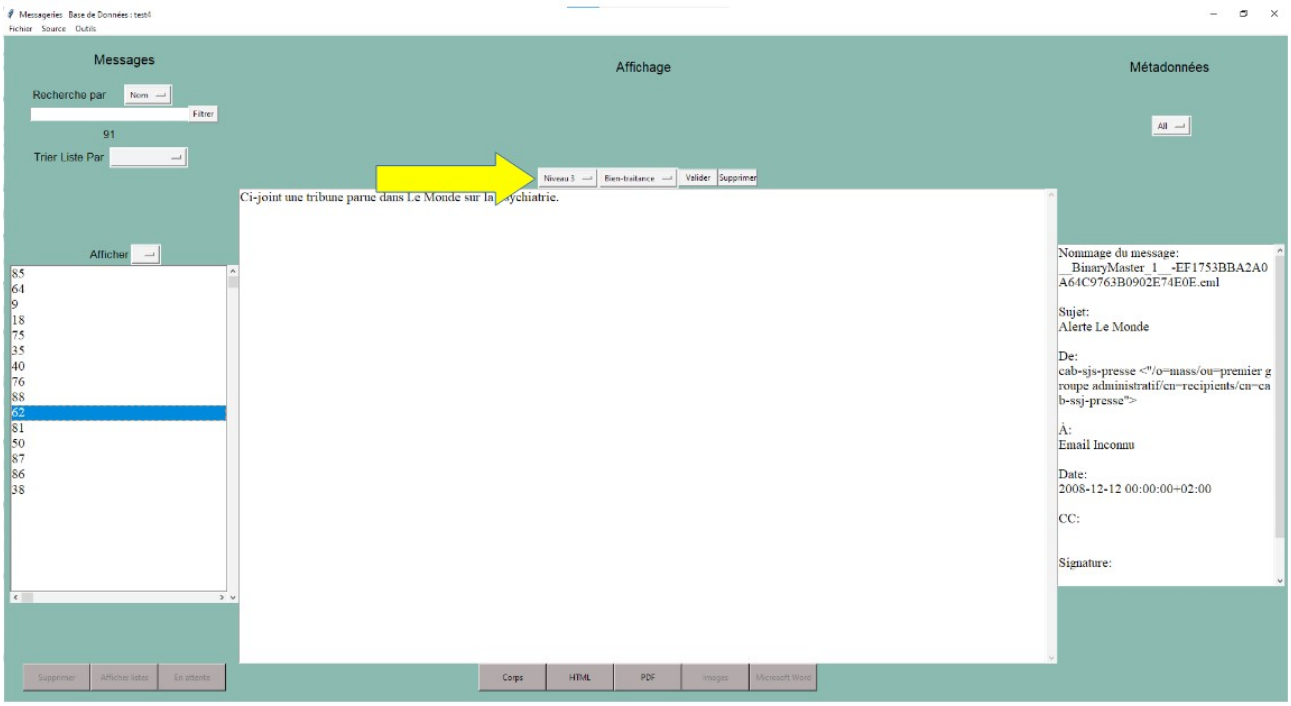


Figure 24: Page avis d'expert

Pour chaque message fourni, l'expert.e peut associer un ou plusieurs thèmes et pour chaque thème les termes voulus ou les supprimer si nécessaire.

4.3. Interface de gestion de la base de données

La base de données (mysql et python connector) est gérée via une interface séparée. Le modèle conceptuel est représenté en fig. 25.

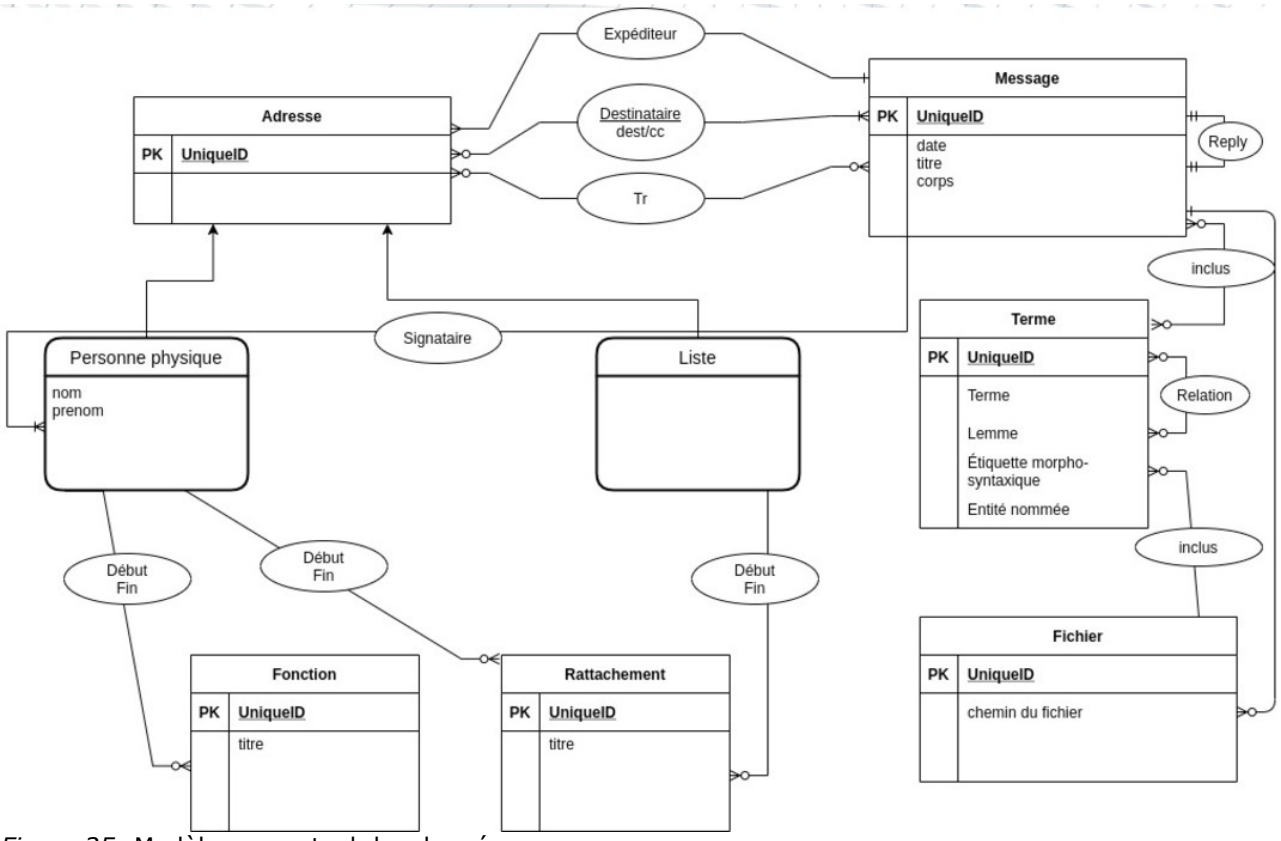


Figure 25: Modèle conceptuel des données

Le script permet de créer une base de données et de remplir les tables en extrayant les messages avec les pièces jointes, en extrayant les thèmes et en intégrant la classification qui a été produite grâce à l'interface de classification sous linux via des fichiers CSV (section 5). Il est également possible de supprimer les messages, les pièces jointes, les thèmes ou la classification si nécessaire (fig. 26).

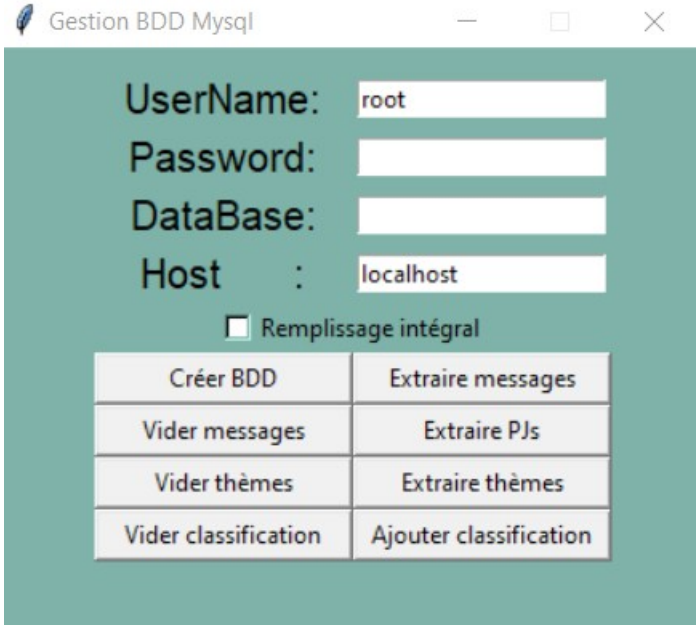


Figure 26 : Boîte de dialogue pour gérer la base de données MySql

5. Interface de classification[2]

L'outil réalisé est une application desktop à installer en local appuyée sur des bibliothèques python, sur Anaconda (gestion des paquets et déploiement), PyCharm pour ses avantages en analyse de code en temps réel, de refactoring avancé, de debugging, etc.
Elle permet de réaliser les opérations de création du corpus, de prétraitement, d'extraction des entités nommées et de validation, d'extraction de termes et de validation, de détection des relations sémantiques par les différentes méthodes existantes et classification des messages.

5.1. Création de corpus et prétraitement

Le menu fichier permet de sélectionner les répertoires contenant sous-répertoires et fichiers (txt, pdf, doc, docx, xml) à inclure dans la classification, à en récupérer le contenu textuel et à créer un fichier de sortie, le tout essentiellement grâce à la bibliothèque tika. Le pré-traitement (SpaCy et NLTK) effectue la segmentation en phrases. Une fois le corpus pré-traité, les différentes opérations sont possibles et offertes via l'écran principal (fig. 28).

2 Les pages qui suivent portant sur l'interface de classification sont largement inspirées de Akmouche C. 2022, p. 15-23.

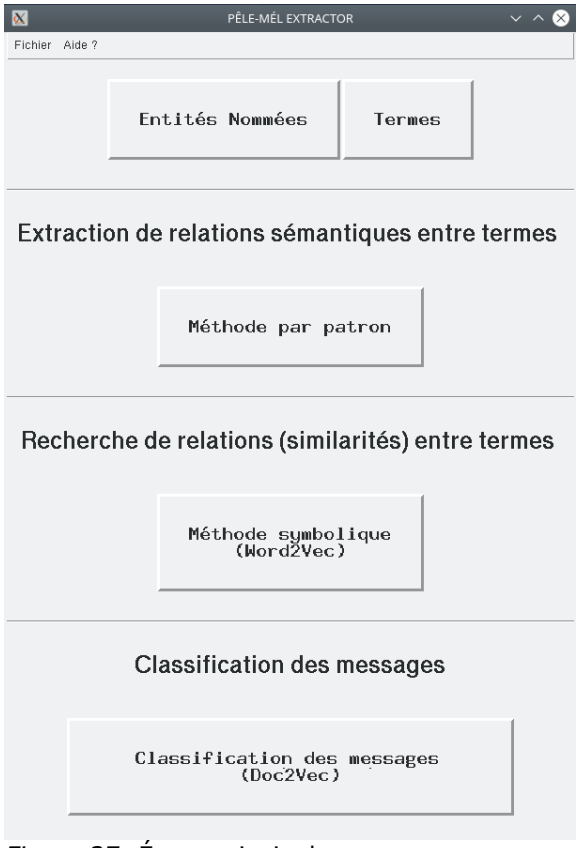


Figure 27: Écran principal

5.2. Extraction des entités nommées et des termes

Des interfaces permettent de lancer les extractions, d'effectuer des validations et de récupérer en sortie des fichiers CSV.

5.2.1. Entités nommées (personnes physiques et morales)

Les entités nommées de type personne et organisation (section 3.2.2) sont extraites du corpus sélectionné par l'utilisateur.rice et sauvegardées dans deux listes différentes à un emplacement préalablement choisi. Les deux listes sont également affichées (fig. 28). Si le corpus est très volumineux, l'outil le découpe automatiquement en plusieurs corpus de taille inférieure.

Comme on a affaire à du langage naturel, après extraction, il y a toujours du bruit (fausses entités nommées). L'interface de la figure 29 permet à l'utilisateur de valider ou de nettoyer manuellement les entités nommées extraites (valider une entité nommée, supprimer ou laisser en attente de validation). La validation peut combiner des opérations manuelles et automatiques (via une liste pré-existante).

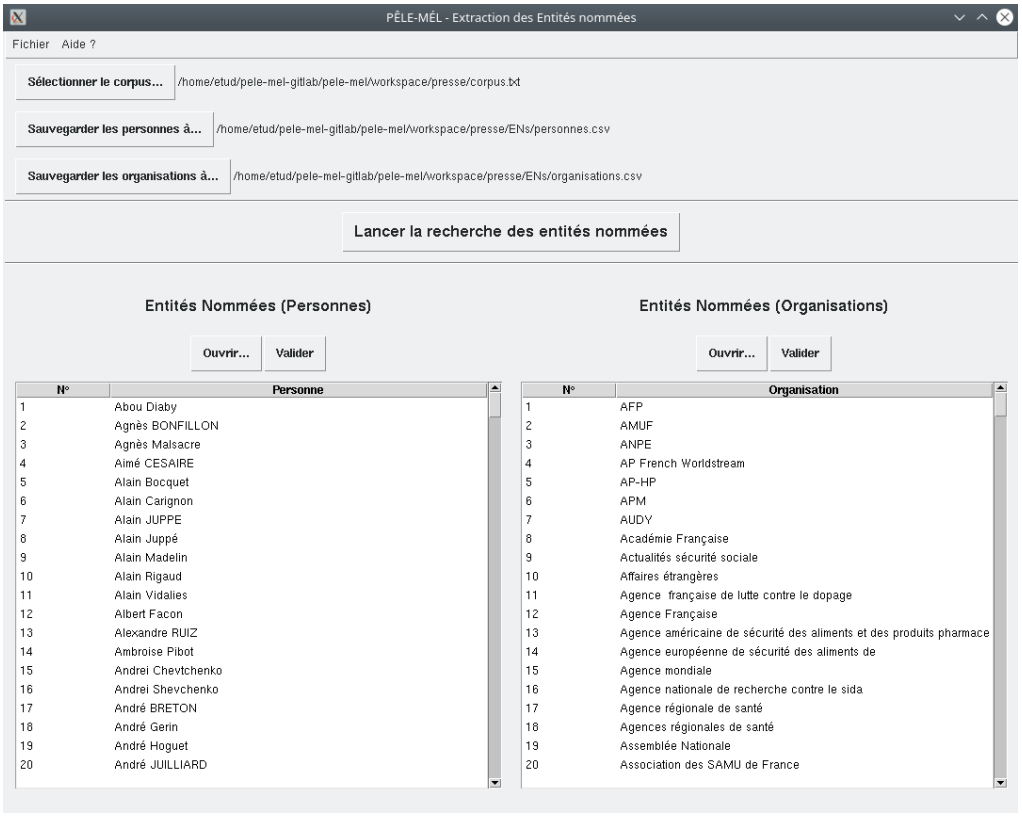


Figure 28: Extraction et affichage des entités nommées

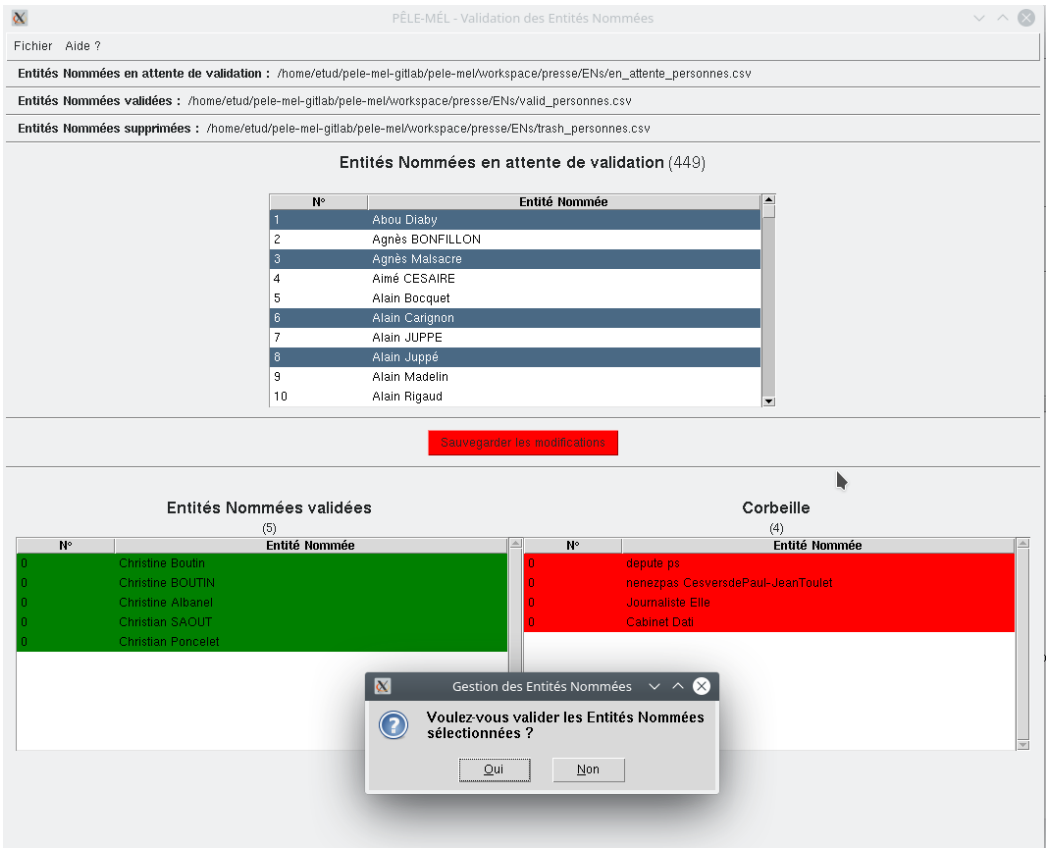


Figure 29: Validation des entités nommées

5.2.2. Termes

L'interface (fig. 30) active l'extracteur adapté à nos besoins (section 3.2.1) et permet de choisir les paramètres d'extraction. Un fichier de sortie en csv est sauvegardé à un emplacement préalablement choisi et la liste s'affiche.

L’utilisateur.rice (fig. 31) peut ensuite valider, supprimer ou laisser en attente de validation les termes extraits manuellement, en fonction de la fréquence ou d’un nombre de mots en choisissant une valeur. Il peut également confronter la liste à une ou plusieurs listes de référence. Ces listes peuvent être externes (par exemple les descripteurs des thesaurus) ou issus d’un traitement précédent du même corpus ou de corpus de même nature.

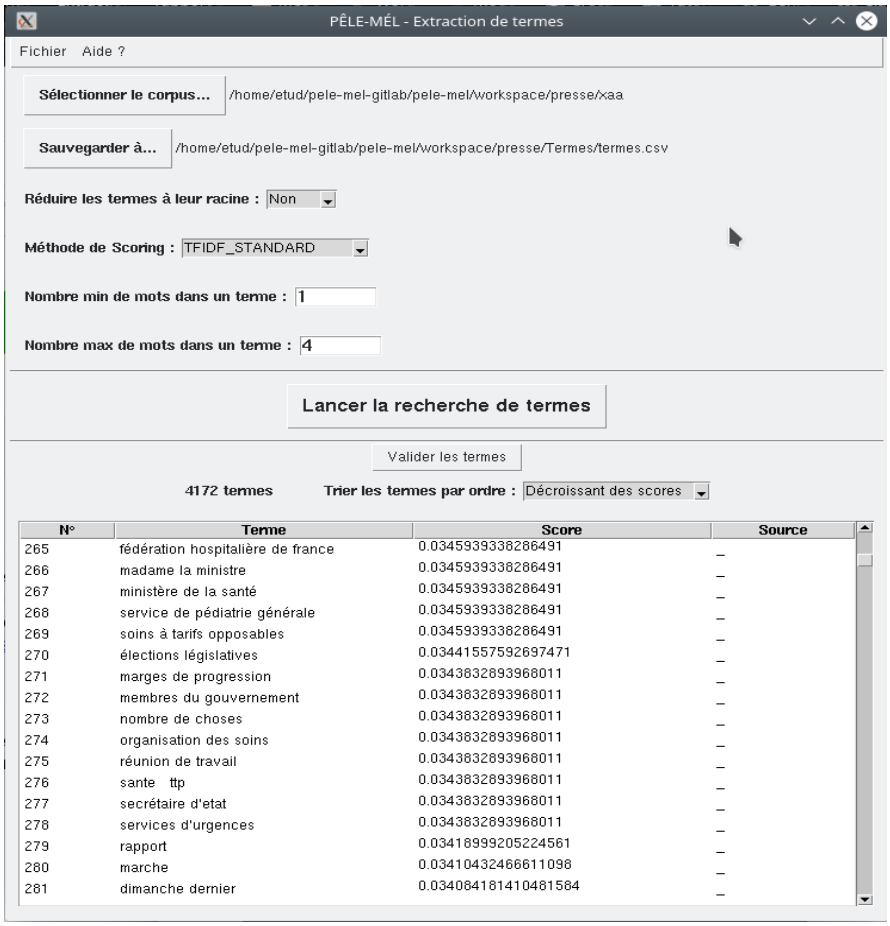


Figure 30: Extraction et affichage des termes

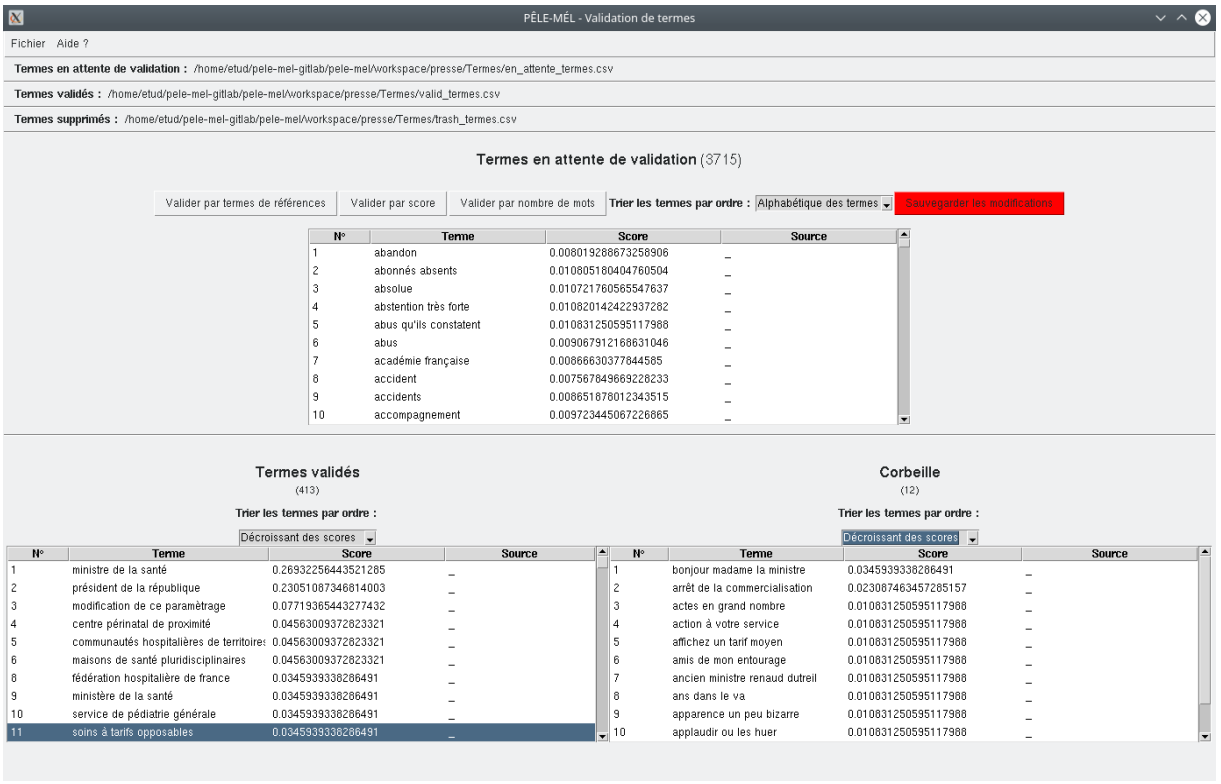


Figure 31: Validation des termes extraits

5.3. Gestion et extraction des relations

5.3.1. Méthode des patrons

Cette interface (fig. 32) permet à l'utilisateur.rice de paramétrer sa recherche de relations (liste de termes à utiliser, type de relation recherchée, les patrons recherchés…). Une fois la recherche terminée, en double-cliquant sur la ligne de résultat souhaitée, des informations détaillées sont disponibles et la phrase où la relation a été identifiée s'affiche dans une fenêtre.

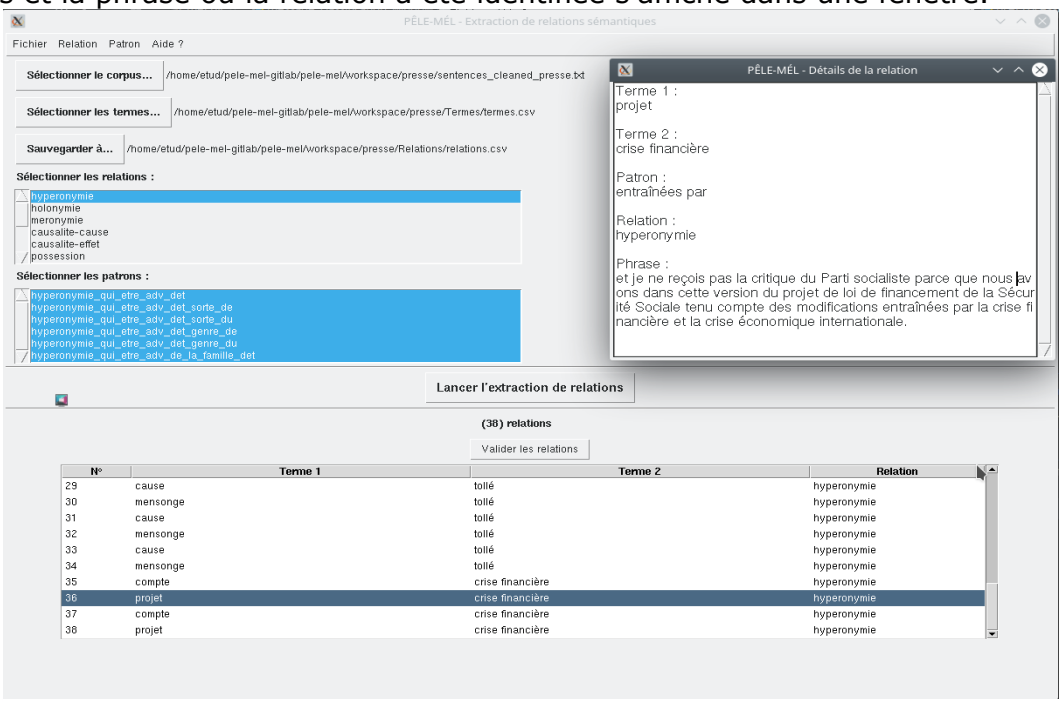


Figure 32: Extraction et affichage des relations sémantiques

Comme pour les entités nommées et les termes, on peut valider, supprimer ou garder une relation en attente de validation (fig. 33).

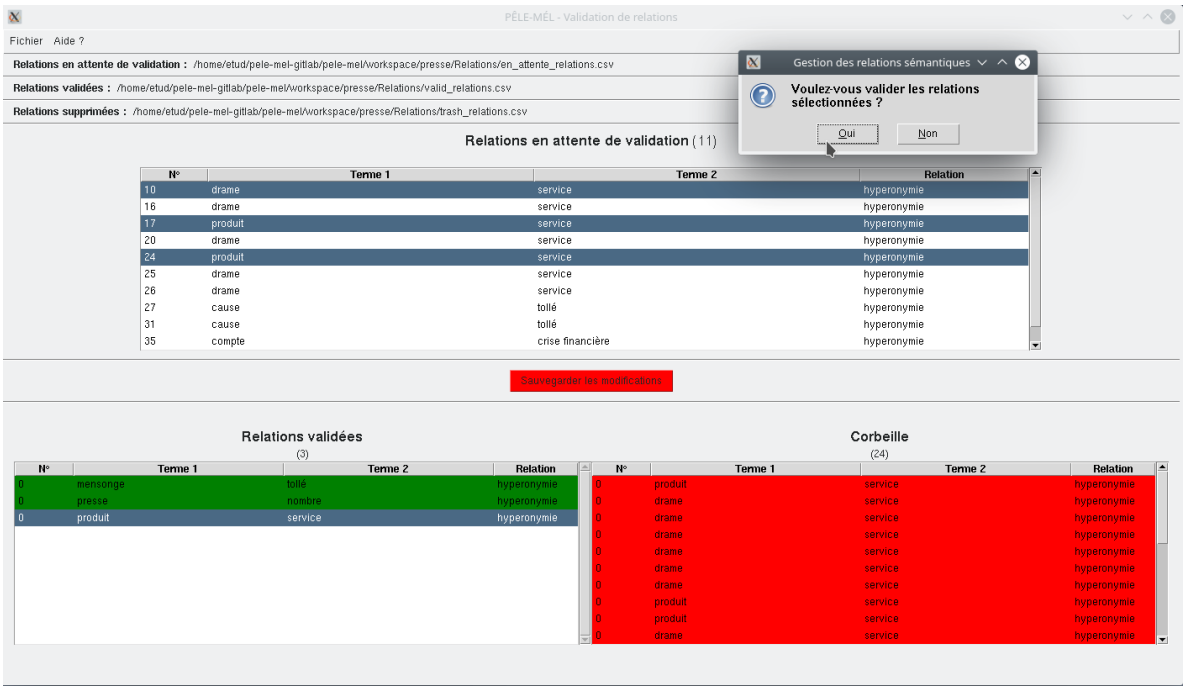


Figure 33: Validation des relations sémantiques extraites

L'utilisateur.rice peut créer une nouvelle relation et/ou un nouveau patron (fig. 34) en se référant au site de SpaCy pour connaître la syntaxe à adopter (https://spacy.io/).

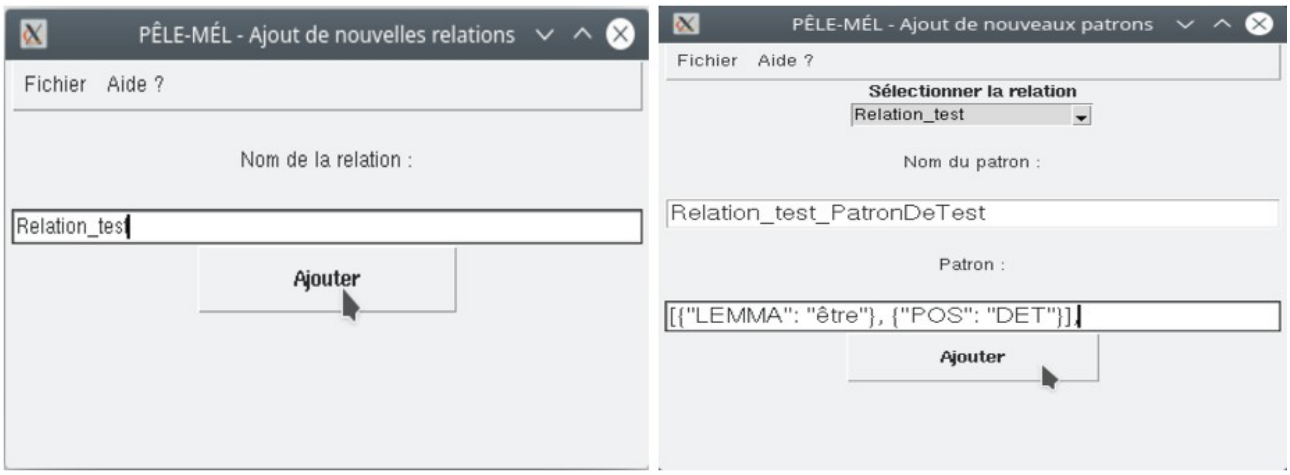


Figure 34: Fonctions d'ajout d'une relation ou d'un patron

5.3.2. Méthode symbolique (réseau de neurones)

L'interface permet de paramétrer le pré-traitement (fig. 35). Celui-ci achevé, deux possibilités s'offrent à l'utilisateur.rice : utiliser un modèle pré-entraîné ou entraîner un nouveau modèle Word2Vec.

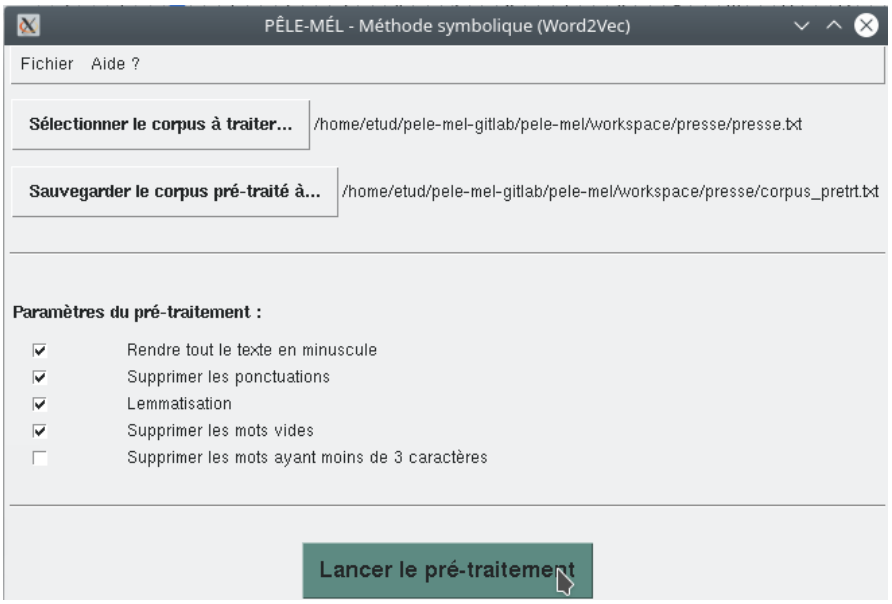


Figure 35: Prétraitement d'un corpus

Si on choisit d'entraîner un nouveau modèle Word2Vec, il faut sélectionner les paramètres d'apprentissage (dimensionnalité des vecteurs, et la taille des fenêtres). L'outil commence par extraire les termes puis peut afficher dans une nouvelle fenêtre les termes similaires (nuage de mots) (fig. 36).

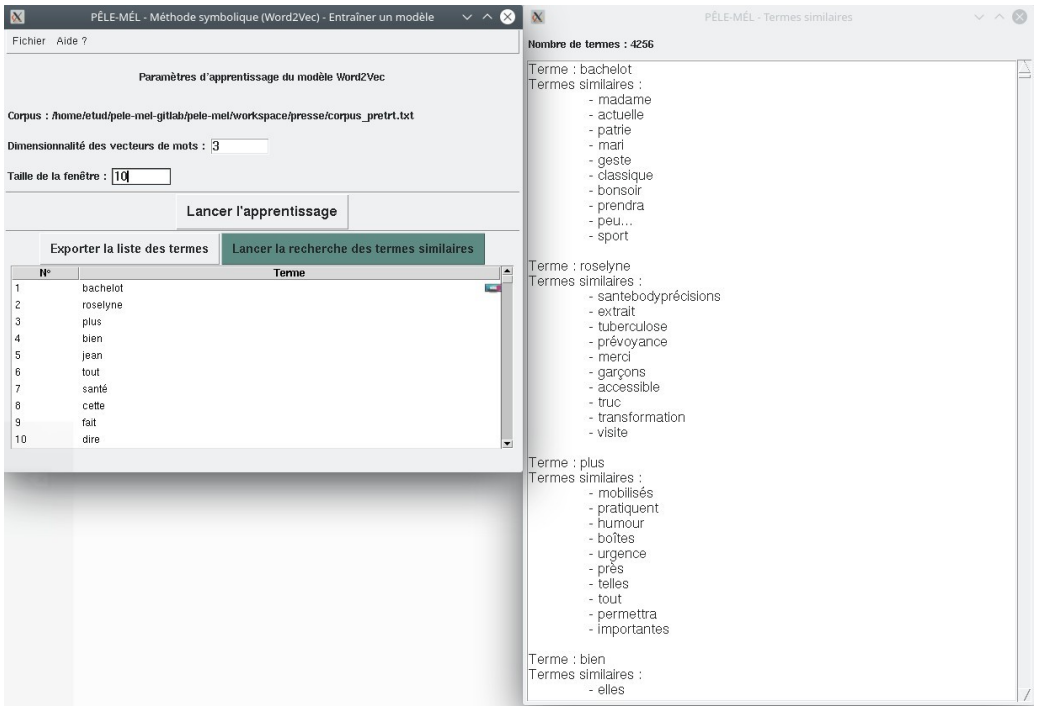


Figure 36: Entraînement, extraction et affichage des termes similaires avec Word2Vec

On peut s'appuyer sur un modèle Word2Vec pré-entraîné qui permet de faire une recherche de termes similaires (fig. 37) ou de nuage de mots par thématique (fig. 38).

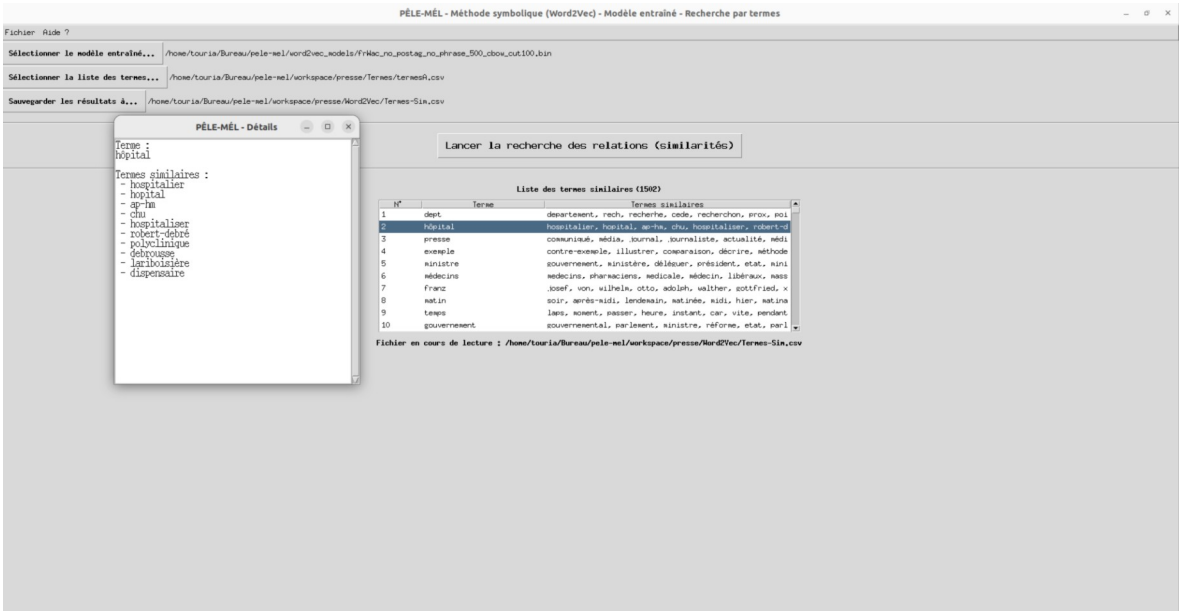


Figure 37: Extraction et affichage des termes similaires en utilisant un modèle entraîné

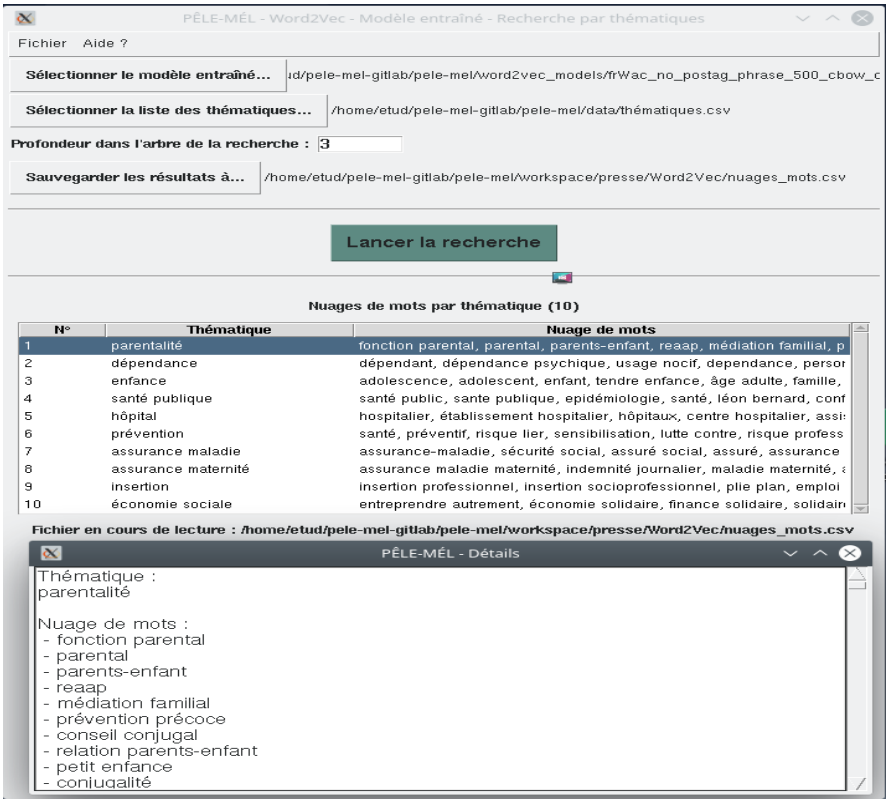


Figure 38: Extraction et affichage des nuages de mots similaires avec Word2Vec

5.4. Classification des messages

Comme expliqué dans la section 3.3, cette méthode crée des nuages de termes par thématique (fig. 39). Dans un second temps, on utilise Doc2Vec pour associer chaque message à une thématique (fig. 40). L'objectif est d'avoir un fichier CSV en sortie contenant pour chaque message son identifiant unique et sa thématique. Ce fichier peut ensuite être téléchargé dans l'interface de visualisation (section 4.3).

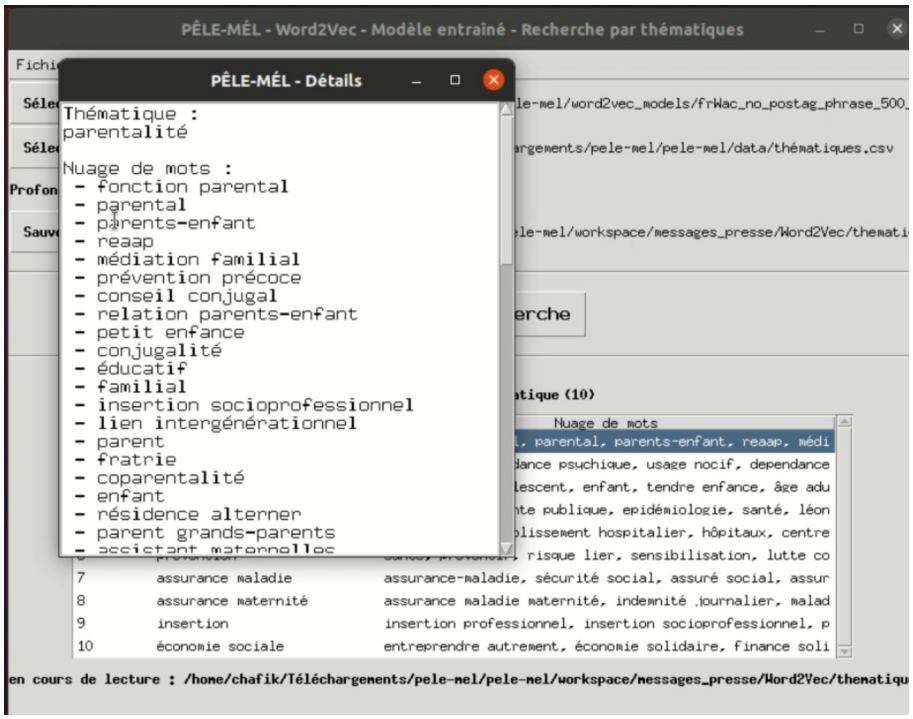


Figure 39: Exemple d'un nuage de termes

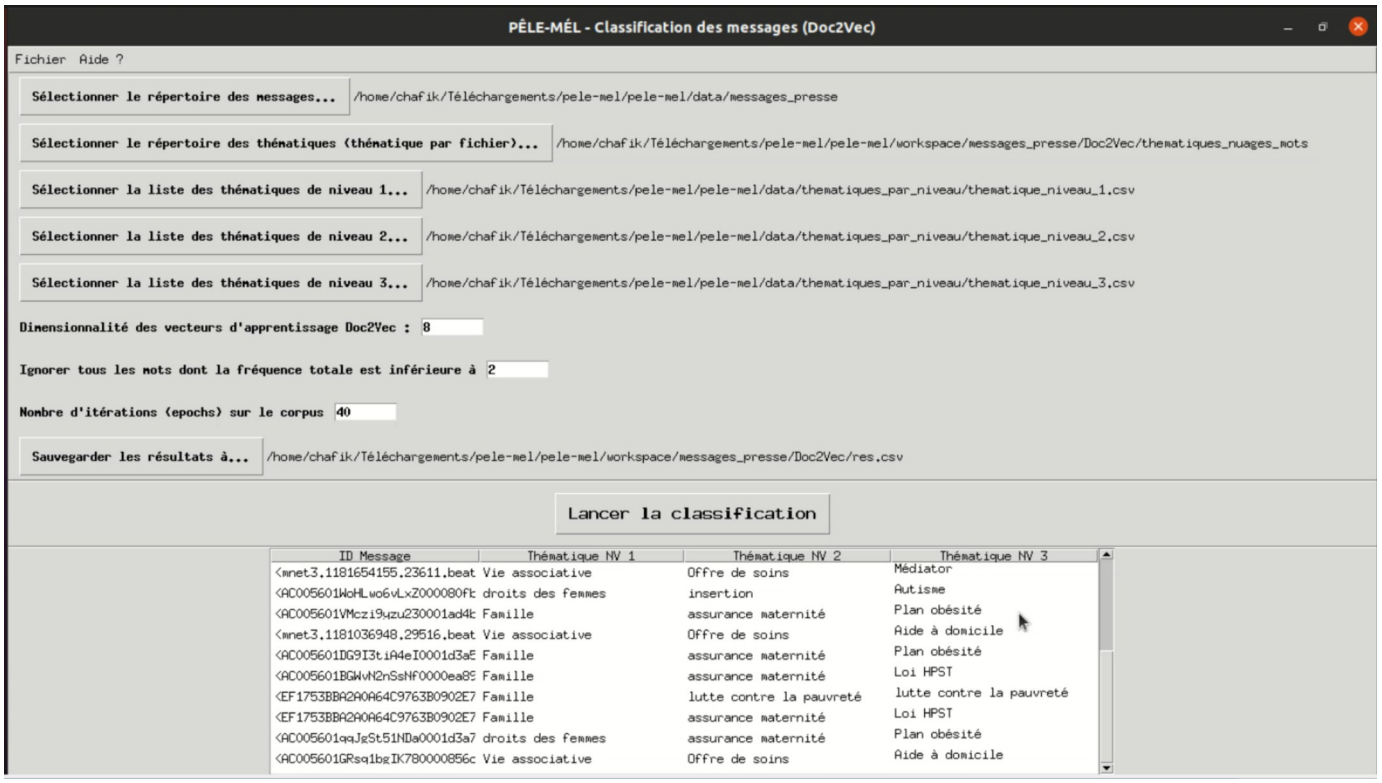


Figure 40: Classification des messages avec Doc2Vec

Conclusion

Le programme Pêle-mél a innové de différentes manières.

Tout d'abord, dans les méthodes proposées. Il s'agit de la première adaptation de méthodes symboliques utilisant les réseaux de neurones au cas de corpus de méls en français. Il s'agit aussi d'une expérimentation concrète aboutissant à une stratégie précise de mise en œuvre.

Ensuite, par la manière d'appréhender la question des méls et celle de la classification thématique. L'unité sur laquelle il a été choisi de réfléchir est constitué du message et des pièces jointes et pas du message seul. Autre innovation, le thème est délimité non par un mot-clef ou des mots-clef mais bien par un nuage de termes préalablement extraits et reliés par une recherche de similarités via une méthode de plongement lexical.

L'expérimentation s'est développée sur un terrain particulier mais les résultats sont a priori généralisables à d'autres contextes et d'autres environnements.

Il n'en reste pas moins que subsistent plusieurs obstacles.

Le premier concerne les principaux intéressé.es, à savoir les archivistes et de leur nécessaire acculturation aux concepts et contraintes du Taln. Un temps de formation est nécessaire tant sur les questions de linguistique que sur les méthodes et leur fonctionnement. On notera que les DSI ne possèdent pas généralement non plus cette compétence. Un accompagnement est donc indispensable. L'investissement initial à produire dans la classification elle-même est important. Les validations des listes (entités nommées, abréviations, termes) sont indispensables et en grande partie manuelles au départ, mais capitalisables sur le long terme.

Le second touche au périmètre de ce projet, à commencer par les volumes des données manipulées. Celles-ci n'ont pas été suffisantes pour tester toutes les hypothèses et pour entraîner des modèles spécifiques. Concernant les réalisations finales, deux prototypes sont proposés. Ils permettent d'expérimenter, mais il y a encore du chemin entre ces prototypes et un logiciel intégré et convivial.

Néanmoins, on notera que ce projet a suscité beaucoup d'intérêt au plan international.

Deux papiers dans des congrès internationaux ont été acceptés et présentés :

- *Création d'une base de connaissance à partir des messages spécialisées pour améliorer l'exploration et l'archivage des méls*, papier présenté le 7 juillet 2022 lors de la 16e conférence internationale Statistical Analysis of textual Data à Naples (Aït el Mekki et al. 2022) ;
- *Improving recordkeeping and contextualization of electronic messaging in French* présenté le 16 septembre 2022 lors de l'International Conference on Digital Preservation, Ipres, à Glasgow (Grailles et al. 2022).

Glossaire

Apprentissage	Non supervisé (clustering), supervisé (machine learning)
Apprentissage non supervisé	Situation d'apprentissage automatique sans données étiquetées. L'objectif est d'extraire des classes d'individus partageant des caractéristiques communes.
Apprentissage supervisé	Situation d'apprentissage automatique à partir d'exemples annotés et de données étiquetées. Capacité d'apprendre une fonction de prédiction.
Candidats-termes	Mots ou suites de mots susceptibles d'être des unités terminologiques
Clustering	Apprentissage non supervisé
Entité nommée	Mot ou groupe de mots de noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.
Étiqueteur morpho-syntaxique	Associe aux mots d'un texte les informations grammaticales correspondantes comme la partie du discours, le genre, le nombre, etc.
Extracteur de termes	Outil établissant une liste de candidats-termes dans une masse au préalable indifférenciée d'unités lexicales
Fréquence « brute »	Nombre d'occurrences d'un terme dans un corpus. Voir TF-IDF
Machine learning	Apprentissage supervisé
Ontologie	Ensemble structuré des termes et concepts représentant le sens d'un champ d'informations. Modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts
Plongement de document	Permet de créer une représentation numérique d'un document par un vecteur de nombres réels. Des documents utilisés dans un contexte similaire seront représentés par des vecteurs proches. L'approche est prédictive : un document prédit son contexte et un contexte prédit un document.
Plongement lexical	Permet de créer une représentation numérique de chaque mot d'un corpus par un vecteur de nombres réels. Des mots utilisés dans un contexte similaire seront représentés par des vecteurs proches. L'approche est prédictive : un mot prédit son contexte et un contexte prédit un mot. On parle également de plongement de mots.
Regex	Regular expressions ou expressions régulières ou motifs, permettent de décrire selon une syntaxe précise ou règles un ensemble souhaité de chaînes de caractères

Réseau de neurones	Modèle informatique dont la structure en couches est similaire à la structure en réseau des neurones du cerveau, avec des couches de nœuds connectés. C'est une structure capable d'apprendre par l'expérience, qui peut être entraînée à reconnaître et à classer
Terme	Tout mot ou groupe de mots représentant un concept spécifique d'un domaine
TF-IDF	*term frequency-inverse document* frequency. Schéma de pondération qui mesure l'importance d'un terme dans l'ensemble du corpus. Des termes moins fréquents sont considérés comme plus discriminants et améliore la pertinence.
Thesaurus	Liste organisée de termes contrôlés et normalisés

Récapitulatif des technologies mobilisées

Base64	Module python pour le décodage des base64
CamenBERT	Modèle de langue contextuel pré-entraîné en français (2020) de BERT (bidirectional encoder representations from transformers, 2018)
Dateutil	Module python pour la manipulation des dates
Doc2Vec	Algorithme d'apprentissage automatique non supervisé (2014) qui est utilisé pour convertir un document en un vecteur (fonctionnement similaire à Word2Vec)
Email	Module python pour l'extraction des métadonnées d'un message
Fastext	Modèle permettant de créer un algorithme d'apprentissage supervisé ou non supervisé pour obtenir des représentations vectorielles des mots. Il existe de nombreux modèles pré-entraînés
Fr WaC	Corpus de 1,6 milliard de mots construit à partir du domaine .fr et en utilisant des mots de moyenne fréquence du corpus du *Monde diplomatique* et de dictionnaires français de base
Iramuteq	Logiciel libre proposant des analyses lexicométriques (classification hiérarchique descendante et analyse de similitude)
KMeans	Algorithme de classification non supervisée
Matplotlib	Module python pour l'affichage des graphes de réseau
Netwrokx	Module python pour la création du réseau pour visualiser les relations entre emails
NLTK	Suite de bibliothèques de traitement de texte python pour la classification, la tokénisation, l'étymologie, le balisage, l'analyse syntaxique etc.
PLDAC	Extracteur de termes
Pyvis	Module python pour l'affichage avancé des graphes
Quopri	Module python pour le décodage des chaînes de caractères
SpaCy	Bibliothèque logicielle python pour le Taln
Tika	Module python pour la manipulation des documents Microsoft Word
Tkinter	Interface et widgets
Treetagger	Étiqueteur morpho-syntaxique
Word2Vec	Algorithme publié en 2013 utilisant un modèle de réseau de neurones pour apprendre les associations de mots à partir d'un large corpus de textes. Il représente chaque mot distinct par une liste particulière de chiffres appelée vecteur. Les vecteurs permettent de capturer les qualités sémantiques et syntaxiques des mots. Le niveau de similarité sémantique est donné par une fonction mathématique simple (similarité en cosinus)

Bibliographie et références

Aït el Mekki T., Grailles B., Randriatsitohaina T. (2022), Création d'une base de connaissance à partir des messages spécialisées pour améliorer l'exploration et l'archivage des méls. 16e conférence internationale Statistical Analysis of textual Data (association Vadistat).

Akmouche C. (2022). PELE-MEL : Plate-forme d'exploration, de livraison et d'évaluation des méls. Extraction et classification de connaissances à partir d'une base de connaissances et un corpus de messageries. Rapport de stage de master 2 Informatique – Intelligence décisionnelle, Université d'Angers, 40 p.

Alghamdi R. and Alfalqi K. (2015). A survey of topic modeling in text mining. International Journal of Advanced Computer Science and Applications, vol. 6., pp. 147-153.

Bissonnette N. (2012-2013). Gestion des courriels : stratégies, technologies et bonnes pratiques, Revue Archives (Association des archivistes du Québec), vol. 44, n° 1, pp. 77-113.

Bojanowski P., Grave E., Joulin A., and Mikolov T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, vol. 5, pp.135-146.

Bretesché S., de Geffroy B., de Corbière F. (2018). E-bureaucratie: le travail emmailé des cadres, Paris: Presses des Mines.

Cadiou A. Extracteur de termes PLDAC [https://github.com/antocad/PLDAC] et [https://antocad.github.io/PLDAC/]

CamenBERT. [https://camembert-model.fr/]

Cram D. and Daille B. (2016). Terminology Extraction with Term Variant Detection. In Proceedings of ACL-2016 System Demonstrations, pp. 13-18.

ePADD Project Website Homepage, University of Standford. [https://library.standford.edu/projects/epadd]

Fauconnier J.-P. (2015), French Word Embeddings, [http://fauconnier.github.io].

Grailles B., Aït el Mekki T., Vasseur É. (2022). Improving recordkeeping and contextualization of electronic messaging in French, International Conference on Digital Preservation, Ipres.

Honnibal M., and Montani I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Johansen L., Rowell M., Butler K. R., and McDaniel P. D. (2007). Email Communities of Interest. In CEAS The Fourth Conference on Email and Anti-Spam, 2-3 August 2007, USA.

Karagiannis T., and Vojnovic M. (2009). Behavioral profiles for advanced email features. In Proceedings of the 18th international conference on World Wide Web, pp. 711-720.

Lockerd A., and Selker T. (2003). DriftCatcher: The Implicit Social Context of Email. In INTERACT'03, pp. 813-816.

Magnien A. dir. (2012). La gestion et l'archivage des courriels. Manuel pratique. Version 2. Archives nationales.
[https://www.archives-nationales.culture.gouv.fr/documents/10157/11411/2013_12_vademecum_courriel.pdf/d9df3809-0cc3-44b6-8859-320cba1987fa]

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., et al. (2019). CamemBERT: a Tasty French Language Model. [https://hal.inria.fr/hal-02445946]

Maudry C. Dictionnaire des sigles et acronymes de l'administration [https://www.data.gouv.fr/fr/datasets/dictionnaire-des-sigles-et-acronymes-de-ladministration/]

Maynard D., Li Y. and Wim P. (2008). NLP Techniques for Term Extraction and Ontology Population. In Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, NLD, pp. 107-127.

Maynard D., Li Y., Wim P. (2008)). "NLP Techniques for Term Extraction and Ontology Population." In Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, NLD, pp. 107–127.

Nadjate S., Adi K. and Allili M. (2020). Semantic Representation Based on Deep Learning for Spam Detection. Foundations and Practice of Security, pp. 72-81.

Nair A. M., Justus A. A., Ramesh A., and Rajan B. (2020). Event Extraction from Emails. International Journal of Computer Applications, vol. 176, n°41, pp. 1-8.

Nazarenko A., Zargayouna H., Hamon O. and Puymbrouck J. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. Revue TAL, ATALA, vol. 50, pp. 257-281.

Omrane N., Nazarenko A. and Szulman S. (2011). Le poids des entités nommées dans le filtrage des termes d'un domaine. In Proceedings of Conférence internationale de Terminologie et Intelligence Artificielle, Paris, pp. 80-86.

Programme VITAM (2013). L'archivage des messageries électroniques. Preuve de concept VITAM, Paris: Ministère des Affaires étrangères/Ministère de la Culture/Ministère de la Défense.

Programme VITAM (2013). L'archivage des messageries électroniques. Preuve de concept, 103 p.

Prom C. (2019). Preserving email., 2nd ed., London: DPC. [https://www.dpconline.org/docs/technology-watch-reports/2159-twr19-01/file]

RATOM Project Website Homepage, University of North Carolina. [https://ratom.web.unc.edu/]

Suárez P., Dupont Y., Muller B., Romary L., Sagot B. (2020). "Establishing a New State-of-the-Art for French Named Entity Recognition" in LREC 2020 - 12th Language Resources and Evaluation Conference, May 2020, Marseille, France. [hal-02617950v2]

SpaCy. [https://spacy.io/]

Tang G., Pei J., & Luk W.S. (2014). "Email mining: tasks, common techniques, and tools", Knowledge and Information Systems, vol. 41, n°1, pp. 1-31.

Tang G., Pei J., and Luk W. S. (2014). Email mining: tasks, common techniques, and tools. Knowledge and Information Systems, vol. 41, pp. 1-31.

Texier B. (2020). Quand le Quai d'Orsay archive ses emails. Archimag [https://www.archimag.com/archives-patrimoine/2020/08/21/archivage-electronique-quai-orsay-archive-emails]

Tyler J.R., Wilkinson D.M., Huberman B.A. (2003). Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. In Huysman M., Wenger E., Wulf V. (eds) Communities and Technologies. Dordrecht, pp. 81-96.

Xia T. (2020). A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule-Based Filtering Systems. IEEE Access, vol. 8, pp. 82653- 82661.

Zerez T. (2022). Stage Apprentissage, traitement et visualisation des messageries. Rapport de stage de M1 Informatique, Université d'Angers, 20 p.

Dubois G. (2022). Enjeux de l'archivage des courriels et des messageries aux Archives nationales : le cas de l'ADEME. Mémoire de licence professionnelle Métiers de l'information, Le Cnam, 40 p.

Annexes

Liste des relations et des patrons

12 relations ont été testées : hyperonymie, holonymie, méronymie, causalité-cause, causalité-effet, possession, caractérisation, opposition, localisation, accompagnement, traitement, signe.

Le tableau ci-dessous regroupe quelques types de relations les plus utilisées dans les différents tests avec leurs patrons (source : Akmouche C. 2022).

Relation	Patrons
hyperonymie	[{"ORTH": "qui", "OP": "?"}, {"LEMMA": "être"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}],
	[{"ORTH": "qui", "OP": "?"}, {"LEMMA": "être"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}, {"ORTH": "sorte"}, {"ORTH": "de"}],
	[{"ORTH": "qui", "OP": "?"}, {"LEMMA": "être"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}, {"ORTH": "sorte"}, {"ORTH": "du"}],
	[{"ORTH": "qui", "OP": "?"}, {"LEMMA": "être"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}, {"ORTH": "genre"}, {"ORTH": "de"}],
	[{"ORTH": "qui", "OP": "?"}, {"LEMMA": "être"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}, {"ORTH": "genre"}, {"ORTH": "du"}],
	[{"ORTH": "qui", "OP": "?"}, {"LEMMA": "être"}, {"POS": "ADV", "OP": "?"}, {"ORTH": "de"}, {"ORTH": "la"}, {"ORTH": "famille"}, {"POS": "DET"}],
holonymie	[{"ORTH": "qui", "OP": "?"}, {"LEMMA": "être"}, {"POS": "ADV", "OP": "?"}, {"LEMMA": "composer"}, {"POS": "ADP"}],
	[{"POS": "PRON"}, {"LEMMA": "composer"}, {"POS": "ADV", "OP": "?"}, {"POS": "ADP", "OP": "?"}],
	[{"ORTH": "qui", "OP": "?"}, {"LEMMA": "être"}, {"POS": "ADV", "OP": "?"}, {"LEMMA": "constituer"}, {"POS": "ADP"}],
	[{"POS": "PRON"}, {"LEMMA": "constituer"}, {"POS": "ADV", "OP": "?"}, {"POS": "ADP", "OP": "?"}],
méronymie	[{"LEMMA": "faire"}, {"ORTH": "partie"}, {"ORTH": "de"}, {"POS": "DET"}],
	[{"LEMMA": "faire"}, {"ORTH": "partie"}, {"ORTH": "du"}],
	[{"LEMMA": "faire"}, {"ORTH": "partie"}, {"ORTH": "des"}],
	[{"LEMMA": "composer"}, {"ORTH": "le"}],
	[{"LEMMA": "composer"}, {"ORTH": "la"}],

	[{"LEMMA": "composer"}, {"ORTH": "les"}],
	[{"LEMMA": "former"}, {"ORTH": "le"}],
	[{"LEMMA": "former"}, {"ORTH": "la"}],
	[{"LEMMA": "former"}, {"ORTH": "les"}],
	[{"LEMMA": "constituer"}, {"ORTH": "le"}],
	[{"LEMMA": "constituer"}, {"ORTH": "la"}],
	[{"LEMMA": "constituer"}, {"ORTH": "les"}],
	[{"LEMMA": "être"}, {"ORTH": "une"}, {"ORTH": "partie"}, {"ORTH": "de"}, {"POS": "DET"}],
	[{"LEMMA": "être"}, {"ORTH": "une"}, {"ORTH": "partie"}, {"ORTH": "du"}],
	[{"LEMMA": "être"}, {"ORTH": "une"}, {"ORTH": "partie"}, {"ORTH": "des"}],
	[{"LEMMA": "être"}, {"ORTH": "une"}, {"ORTH": "partie"}, {"ORTH": "du"}],
	[{"LEMMA": "être"}, {"ORTH": "une"}, {"ORTH": "partie"}, {"ORTH": "des"}],
causalité-cause	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "causer"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "causer"}, {"POS": "ADV", "OP": "?"}, {"POS": "ADP"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "provoquer"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "provoquer"}, {"POS": "ADV", "OP": "?"}, {"POS": "ADP"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "engendrer"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "engendrer"}, {"POS": "ADV", "OP": "?"}, {"POS": "ADP"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "produire"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "produire"}, {"POS": "ADV", "OP": "?"}, {"POS": "ADP"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "déclencher"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "déclencher"}, {"POS": "ADV", "OP": "?"}, {"POS": "ADP"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "générer"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "générer"}, {"POS": "ADV", "OP": "?"}, {"POS": "ADP"}],

	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "entraîner"}, {"POS": "ADV", "OP": "?"}, {"POS": "DET"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "entraîner"}, {"POS": "ADV", "OP": "?"}, {"POS": "ADP"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "affecter"}, {"ORTH": "le"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "affecter"}, {"ORTH": "la"}],
	[{"ORTH": "qui", "OP": "?"}, {"POS": "VERB", "OP": "?"}, {"LEMMA": "affecter"}, {"ORTH": "les"}],

Liste des tutoriels joints à ce rapport

	Nom du fichier	Durée
Module classement		
Lancement du module, prétraitement des fichiers avant toute manipulation	01_pretraitement	02:40:00
Extraction et validation des entités nommées	02_entites_nommees	03:21:00
Extraction et validation des termes	03_termes	04:45:00
Identification des relations entre des termes par la méthode des patrons (expressions régulières)	04_relations_patrons	09:19:00
Identification des relations entre des termes par la méthode symbolique (plongement lexical) et classification des termes pour constituer les nuages de termes (Word2Vec)	05_relations_methode_symbolique	07:08:00
Classification des messages par plongement de documents (Doc2Vec)	06_classification_messages	04:54:00
Module exploration		
Lancement du module et recherche par métadonnées ou contenu	07_affichage_recherche	05:36:00
Exploration d'une adresse et génération de graphe	08_explorer_adresse	00:58:00
Interface avec la base de données Mysql et intégration des données provenant du module classement	09_remplissage_bdd	02:10:00
Autres manipulations : corrections, annotations, visualisation par graphique	10_manipulation_bdd_explication	04:50:00

Table des figures

Table des matières