De la boîte noire à la boîte à surprise, le programme Pêle-mél et les messageries électroniques

Touria Aït el Mekki
Bénédicte Grailles

MINISTÈRES SOCIAUX
Liberté
Égalité
Fraternité

1821 2021
PSL

LERIA

temos cnrs
temps · mondes · sociétés · UMR 9016

université angers

Soutenu par
MINISTÈRE DE LA CULTURE
Liberté
Égalité
Fraternité

Plan

1. Le contexte
   - État de l'art
   - Objectifs
   - Méthode
2. Le corpus
   - Un corpus hybride
   - Pré-traitement
3. Analyse fine du texte
   - Définitions
   - Extractions
   - Relations
4. Classification
   - Définitions
   - Thèmes
   - Plongement de documents
5. Quels enseignements ?
   - Réfléchir aux pratiques
   - Approcher les messageries
   - Au-delà des messageries

Équipe

Équipe

- Université d'Angers
  - Bénédicte Grailles (archivistique, Temos) benedicte.grailles@univ-angers.fr
  - Touria Aït El Mekki (informatique, Leria) touria.aitelmekki@univ-angers.fr
    - Tsanta Randriatsitohaina (informatique)
    - Chafik Akmouche (informatique)
    - Taimane Zerez (informatique)

Partenaires

- Mission des archives, Ministères sociaux
  - Anne Lambert (cheffe de mission)
  - Chloé Moser (cheffe de produit Archifiltre)
- École nationale des Chartes
  - Edouard Vasseur (archivistique, centre Jean-Mabillon)

Temos (Temps, mondes, sociétés) est une unité mixte de recherche du CNRS et un laboratoire en sciences sociales. Leria est un laboratoire d'informatique.

Programme soutenu par le ministère de la Culture dans le cadre de l'appel à projet "services numériques innovants"

Le contexte

La prise en charge des mails en France

- Approche Capstone validée par le programme VITAM (2013)

- ✓ Peu de collectes et d'expérimentations.
  Quelques exceptions : messageries des cabinets (depuis les années 2010, environ 3 % pour la Santé) ; messageries de fonds privés.

- SEDA

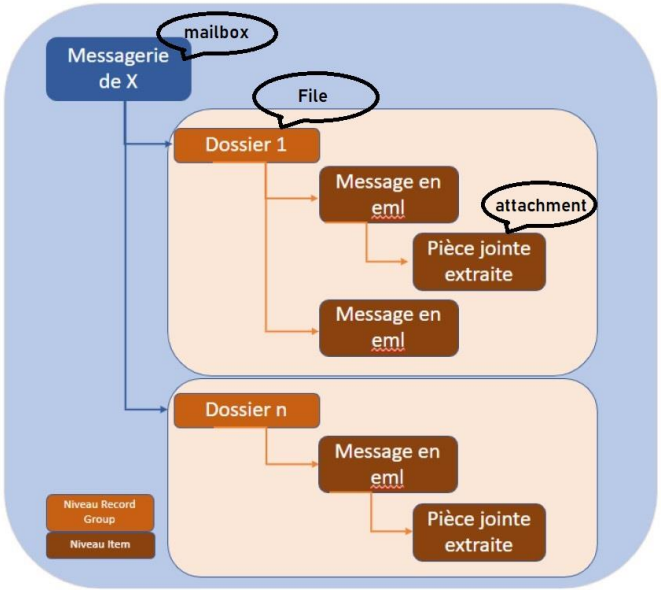
- Une bibliothèque java développée par Vitam (Mail Extract, intégrée à Sedalib) :
  - Extraction des messages en eml
  - Chaque niveau est documenté par un manifeste Seda
  - Préparation des submission information packages SIP

A#437-proiptc-2011	04/07/2022 16:57	Dossier de fichiers
_ArchiveUnitMetadata.xml	24/03/2021 15:43	XML
_BinaryMaster_1_-CB909FF5182D2C44B...	05/07/2022 21:48	Fichier EML


L'archivage des messageries électroniques

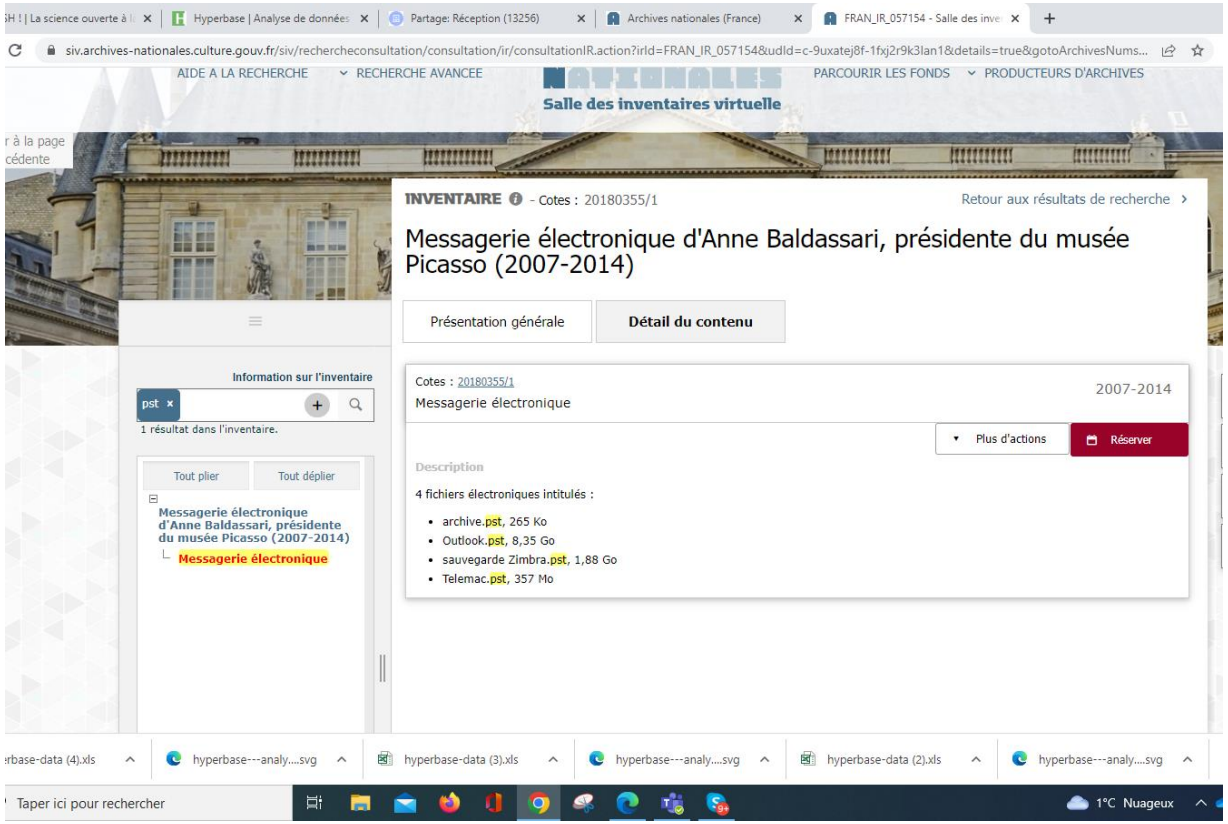
Preuve de concept VITAM

Version 1.2 – 31 octobre 2013



Un accès à l'information non garanti

Des boîtes noires ?



Accueil » Avis 20226133 - Séance du 15/12/2022

Avis 20226133 - Séance du 15/12/2022

Direction générale des patrimoines et de l'architecture

Monsieur X, X, a saisi la Commission d'accès aux documents administratifs, par courrier enregistré à son secrétariat le 2 septembre 2022, à la suite du refus opposé par le directeur général des patrimoines à sa demande de communication, par courriel, des documents suivants mentionnant le Health Data Hub ou son acronyme HDH :

1) les correspondances (courriers, courriels, ou autres) reçus ou envoyés par la ministre de la santé et des solidarités Madame X et son cabinet, au cours de la période de préfiguration du HDH, du 1er janvier 2018 au 31 décembre 2019 ;

2) les correspondances (courriers, courriels ou autres) échangés, entre le 1er janvier 2018 et le 30 mai 2022, entre la ministre de la santé et des solidarités Madame X et son cabinet, puis le ministre de la santé et des solidarités Monsieur X et son cabinet, d'une part, et d'autre part :

a) tout employé ou représentant de X ;

b) tout employé ou représentant d'X, de X ou de leur société commune X ;

Public-cible

Pour qui ?

Le réseau des archives publiques (Archives nationales, missions Archives des ministères, services archives des opérateurs de l'État, archives départementales, communales, intercommunales...). Mais aussi tout service d'archives privé confronté aux mêmes problèmes

Pourquoi ?

Les courriels sont la trace unique de processus décisionnels

C'est un enjeu patrimonial, de transparence de l'action publique mais aussi économique et environnemental

Les services d'archives doivent collecter et conserver des preuves et des témoignages pour le futur

Mais ils doivent aussi maîtriser cette collecte, c'est-à-dire conserver ce qui est important, pour le communiquer

Objectifs de Pêle-Mél
Plateforme d'évaluation, de livraison et d'exploration des méls

Tester différentes stratégies pour contextualiser les boîtes, les réseaux de correspondants et les contenus des messages

Mobiliser les techniques de traitement automatique de langue naturelle et les adapter à la langue française

Développer des critères pour améliorer l'évaluation archivistique et pour aider à la décision

Améliorer l'accès au contenu

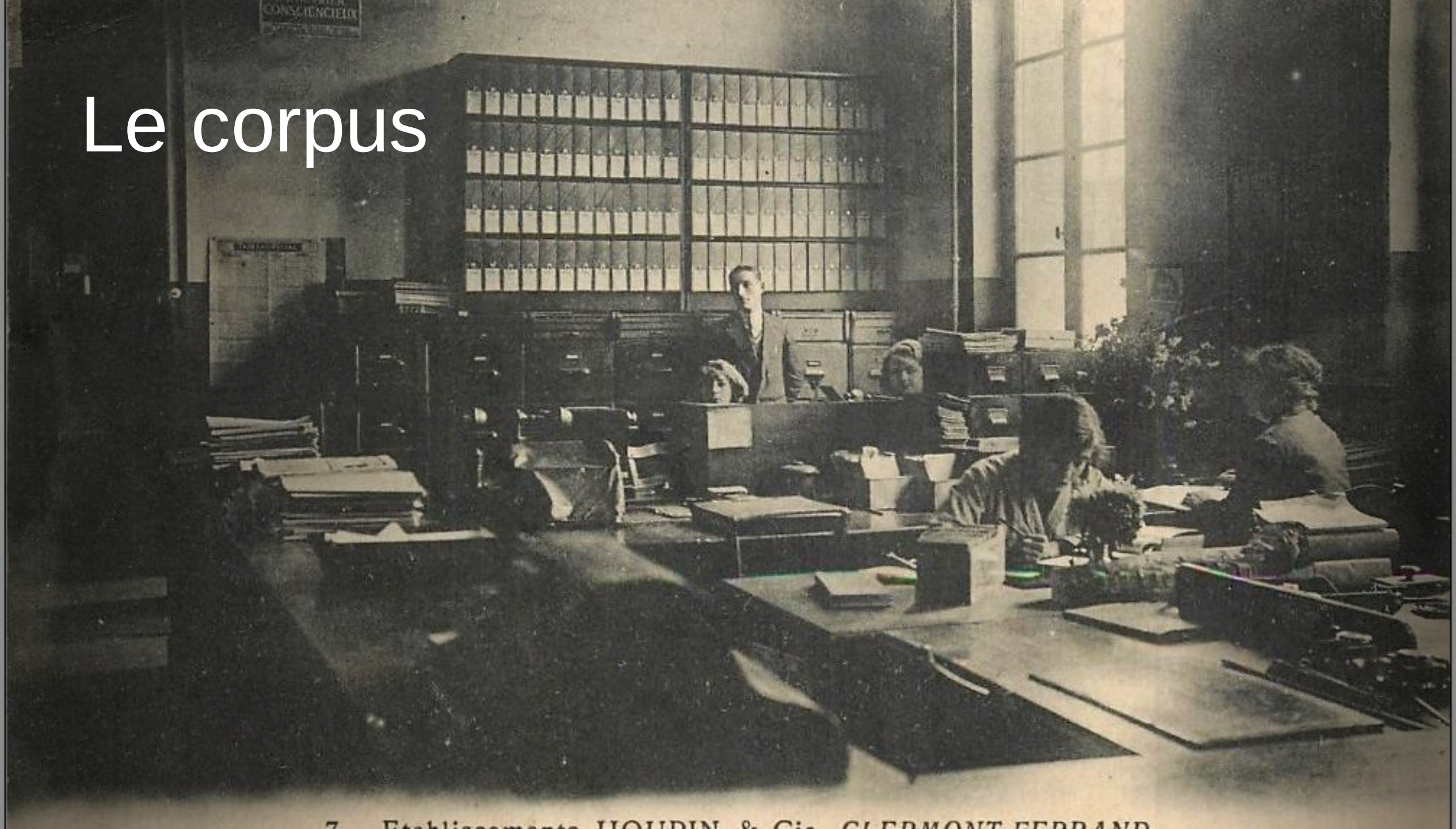
Développer des prototypes d'outils d'exploration et de visualisation des messages électroniques (version bêta)

Méthodes

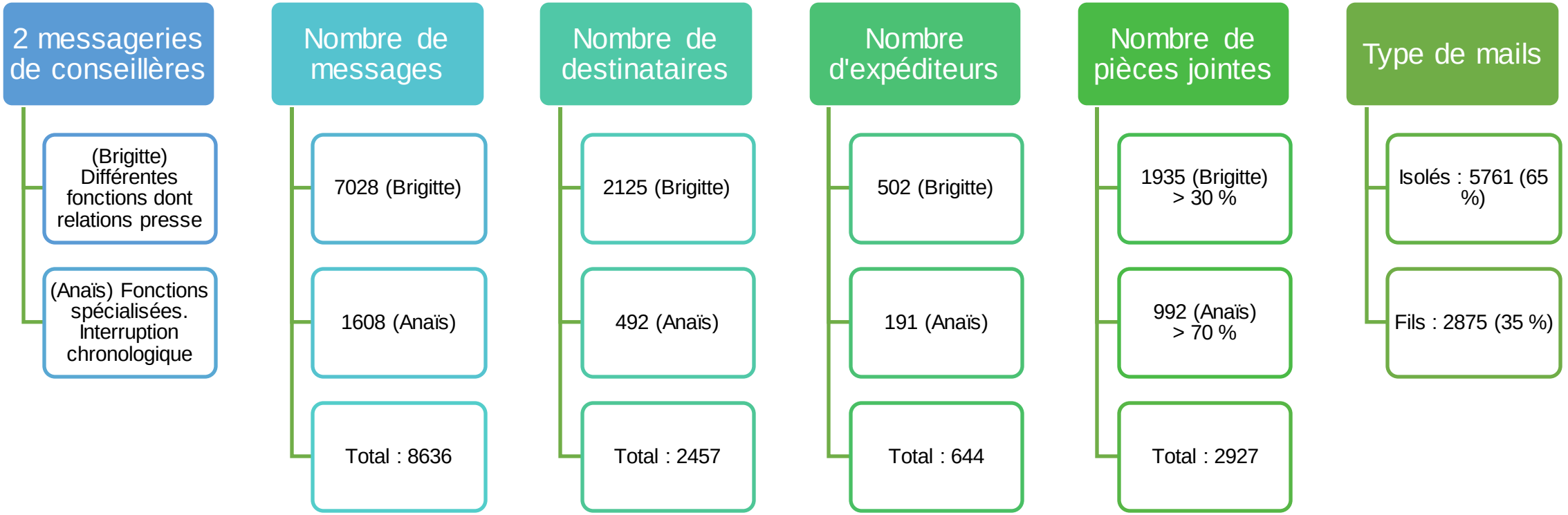
Intelligence artificielle

- Méthodes d'apprentissage automatique pour la classification et la catégorisation des contenus textuels
- Observation directe des messageries
- Sources externes

Le corpus



Zoom sur les messageries



Corpus

Deux messageries de conseillères du cabinet + pièces jointes

- Anaïs et Brigitte
  - Anaïs : juin 2011-nov. 2011
  - Brigitte : mai 2007-nov. 2010 ; juin-oct. 2011
- 8 636 messages + 2927 pièces jointes
- xml + pdf

Organigrammes et annuaires (cabinet, directions du ministère)

- Gouvernement, Cabinet, Direction générale de la santé, Direction générale de la cohésion sociale
- 14 documents (août 2010 - avr. 2012)
- jpeg ; pdf ; doc

Deux thésaurus

- 2014 & 2020
- 7000 descripteurs
- pdf

Discours (corpus communicable)

- 810 documents (nov. 2010 – mars 2012)
- doc / pdf

Pré-traitement

- Métadonnées (adresse exp/dest/CC, date)
- Signature (nom, rattachement)
- Corps (titre, contenu)
- Fichiers joints (nommage, contenu)
- = 1 unité

- Gestion de formats divers (pièces jointes)
- Passage en txt
- Caractères accentués
- Coupures de mots

Edition:(FRA)
Suppl.:
Rubrique:

économie

Médecins : devis obligatoire à partir de 70 euros

La ministre de la Santé, Roselyne Bachelot, annonce au « Figaro » plusieurs mesures pour favoriser la transparence des tarifs et sanctionner les praticiens qui dérapent. *Page 18*

SANTÉ Un entretien avec la ministre Roselyne Bachelot.

«Pas d'effort financier pour les patients»

Organisation du système de soins et budget 2009 de la Sécurité sociale : Roselyne Bachelot aborde l'automne avec deux gros textes «sur le feu». La ministre de la Santé promet au *Figaro* que les assurés ne seront pas davantage mis à contribution pour réduire les déficits. Elle annonce l'entrée en vigueur de mesures de transparence sur les prix des médecins et de sanctions pour ceux qui dérapent. Elle revient sur la philosophie de sa future loi et annonce 800 millions d'aides aux investissements hospitaliers.

LE FIGARO. – Transparence des tarifs, sanction des dépassements abusifs… Plusieurs mesures votées à l'automne dernier restent inappliquées. Pourquoi ce délai ?

Roselyne BACHELOT. – C'est long, je le reconnais, mais nous avons voulu laisser le temps de la concertation. Nous avons effectivement prévu un ensemble de mesures pour renforcer la transparence en faveur des patients. Dans les prochains jours paraîtra l'arrêté qui fixe à 70 euros le montant à partir duquel tout médecin et tout dentiste seront tenus de fournir au patient une information écrite préalable sur le tarif de leurs actes. Nous avions au départ proposé 80 euros, mais les partenaires sociaux préféraient 50. Par ailleurs, un texte paraîtra d'ici à la mi-septembre qui imposera un devis normalisé pour toutes les audio-prothèses, distinguant le prix de l'appareil lui-même et la prestation qui va avec. La variabilité des prix est trop forte dans ce domaine. Enfin, en octobre, paraîtra le décret permettant aux caisses d'assurance-maladie d'appli-

Page 1

Analyse fine du texte

Lexique

Terme

Un terme est un mot ou un groupe de mots qui désigne de manière univoque un objet ou un concept dans un domaine (par exemple, violence domestique, violence sexuelle, lutte contre la violence, etc.)

Entité nommée

peut renvoyer à un objet, une personne, un lieu, une date ou toute autre entité spécifique

Relation sémantique

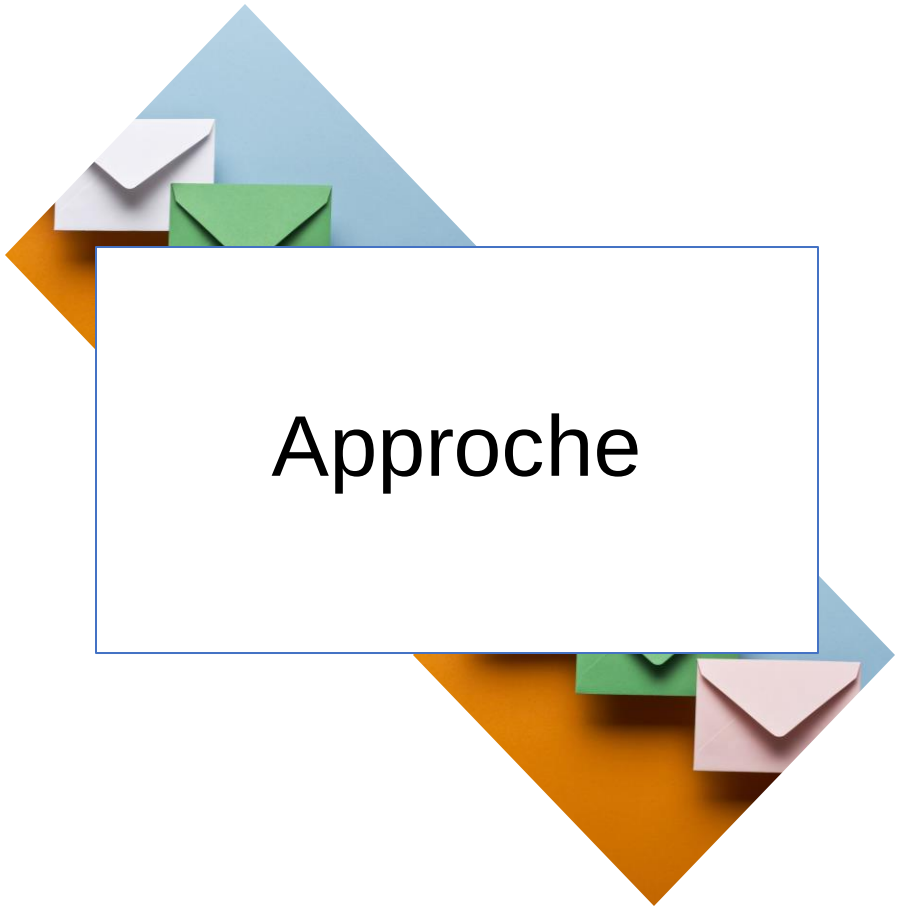
Lien entre deux entités lexicales (terme, entité nommée)

Patron lexical-syntaxique (pattern)

Ex. Terme 1 + être + dét + Terme 2
Ex. Le SIDA est une maladie incurable

-->

Hyperonymie
relation (Sida et maladie incurable)



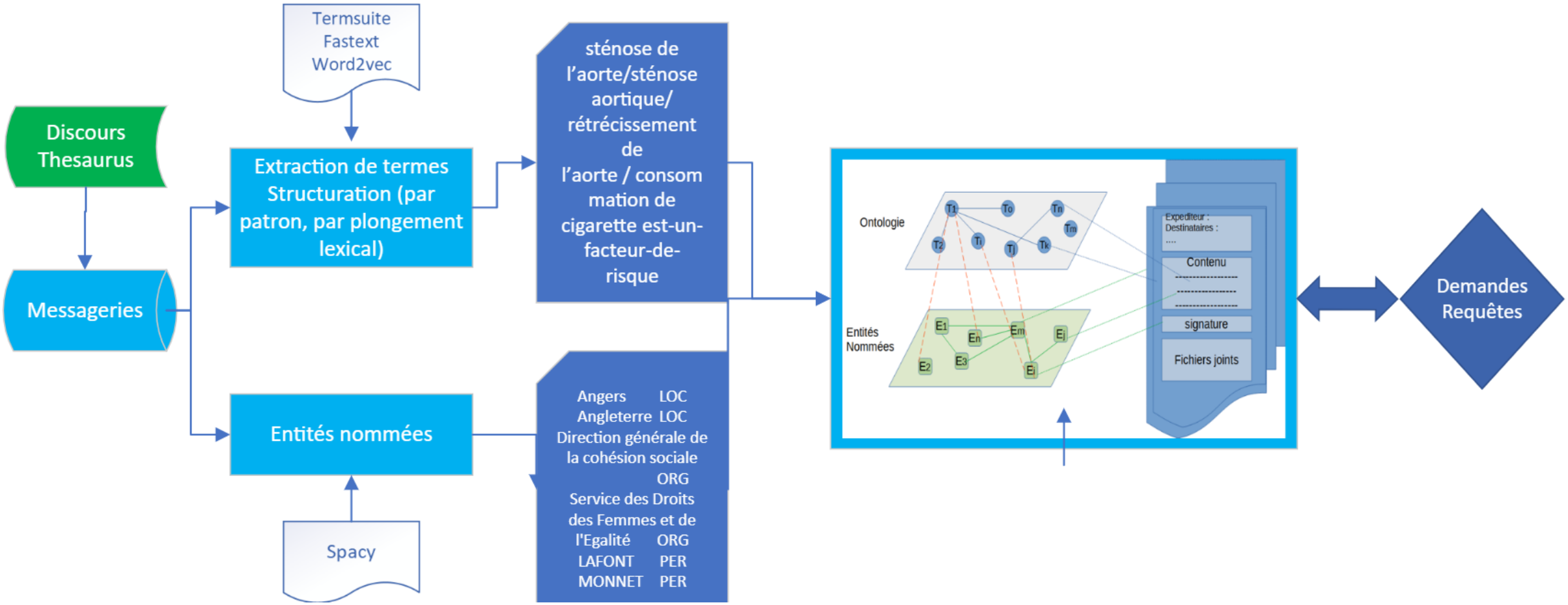
Approche

En entrée : un message = (contenu + objet + fichier joint + titre

Extraire la liste des termes et des entités nommées

Créer des nuages de termes et des entités nommées

Utiliser ces nuages de termes pour améliorer la classification thématique



Extraction et structuration de termes et d'entités nommées

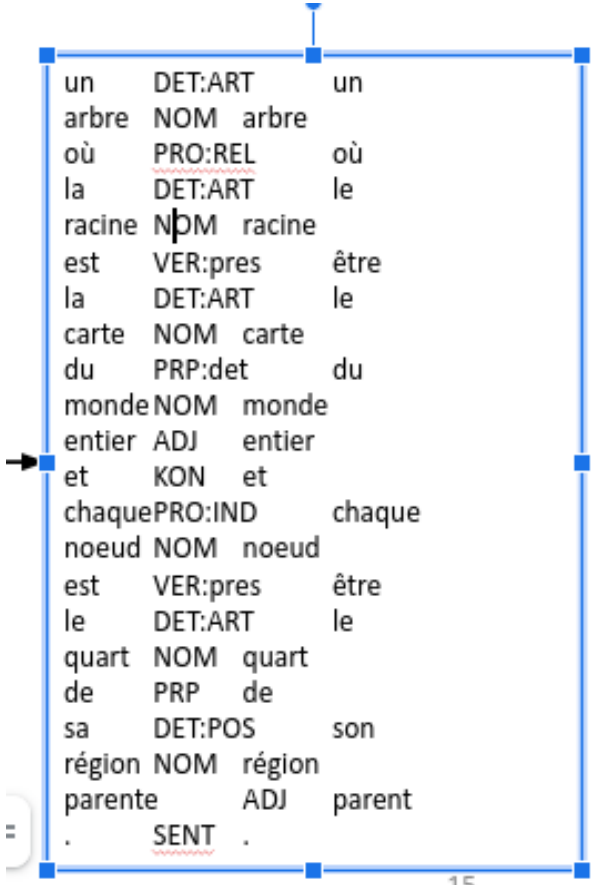
Text processing TreeTagger

- Un étiqueteur probabiliste
- Catégories grammaticales
- Informations morphosyntaxiques
- Informations de lemmatisation

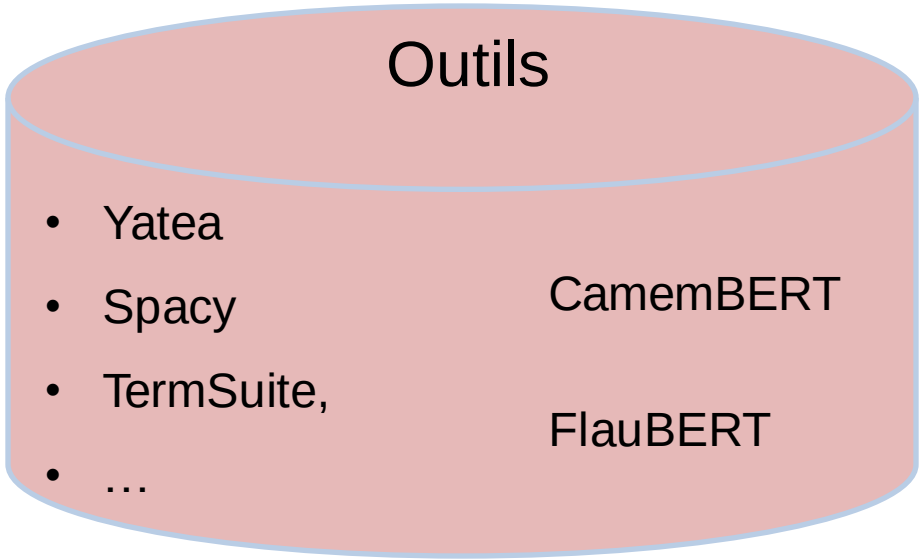
Python : TreeTaggerWrapper

Un arbre où la racine est la carte du monde entier et chaque nœud est le quart de sa région parente

TreeTagger



Extraction des termes



Différentes méthodes

des approches statistiques et ou à bases de règles

Apprentissage automatique

na: solution pratique
nra: plate-forme plus large
na: pression sanguin
npna: fondateur de aol [redacted]
npn: laboratoire de analyse
na: laurence [redacted]
na: outil informatique
npn: défense de vie
n: porteur
n: géant
a: social
n: hôpital
a: énorme
n: médecin
n: forme
npn: mise en place
n: retard
n: docteur
a: immense
n: concurrent

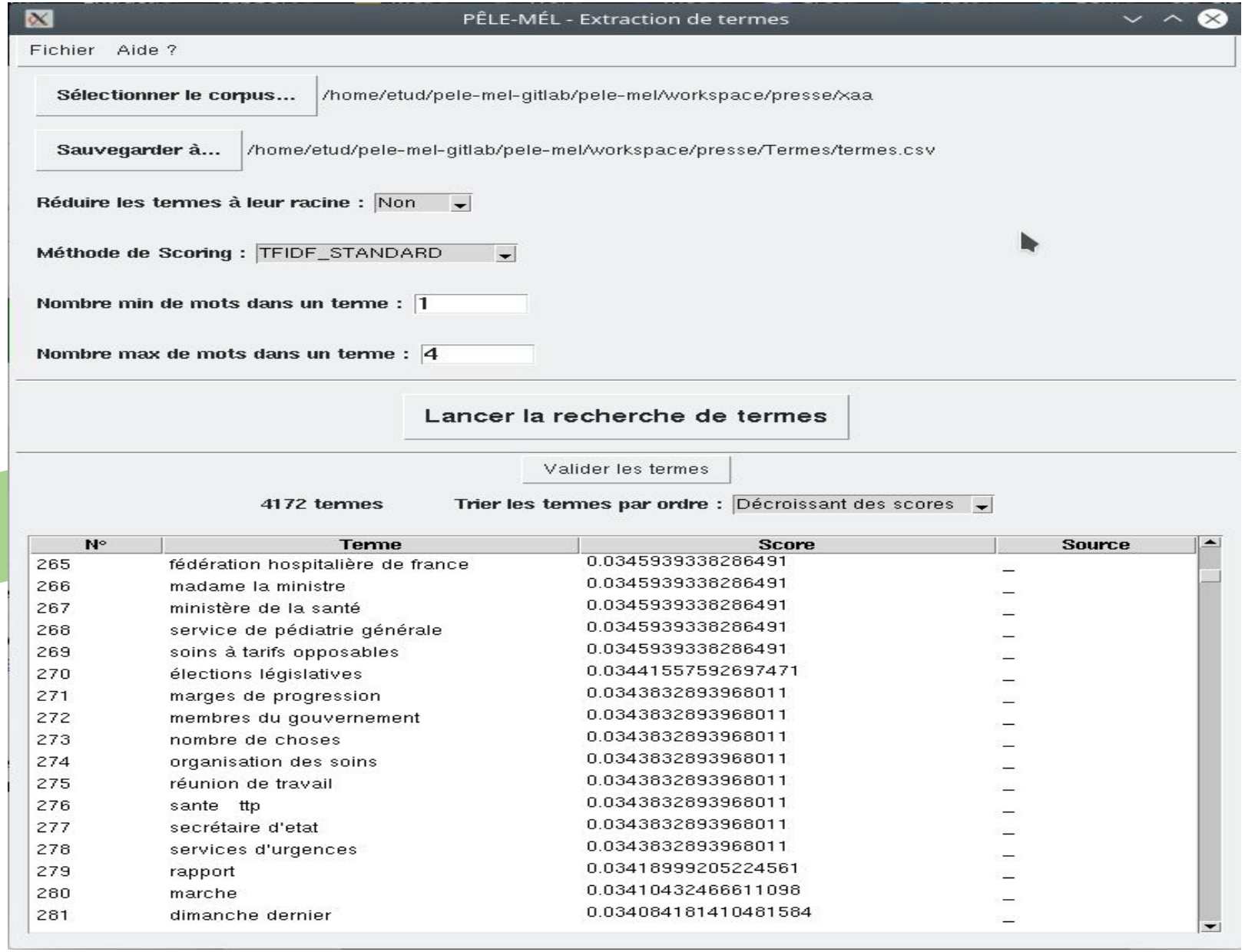
na: projet similaire
na: forme électronique
na: donnée médical
a: diamond
nnpn: privacy forum dans rapport
napn: support idéal de publicité
na: programmeur extérieur
n: revolutionhealth
nn: andrew [redacted]
npn: tau de glucose
na: interface central

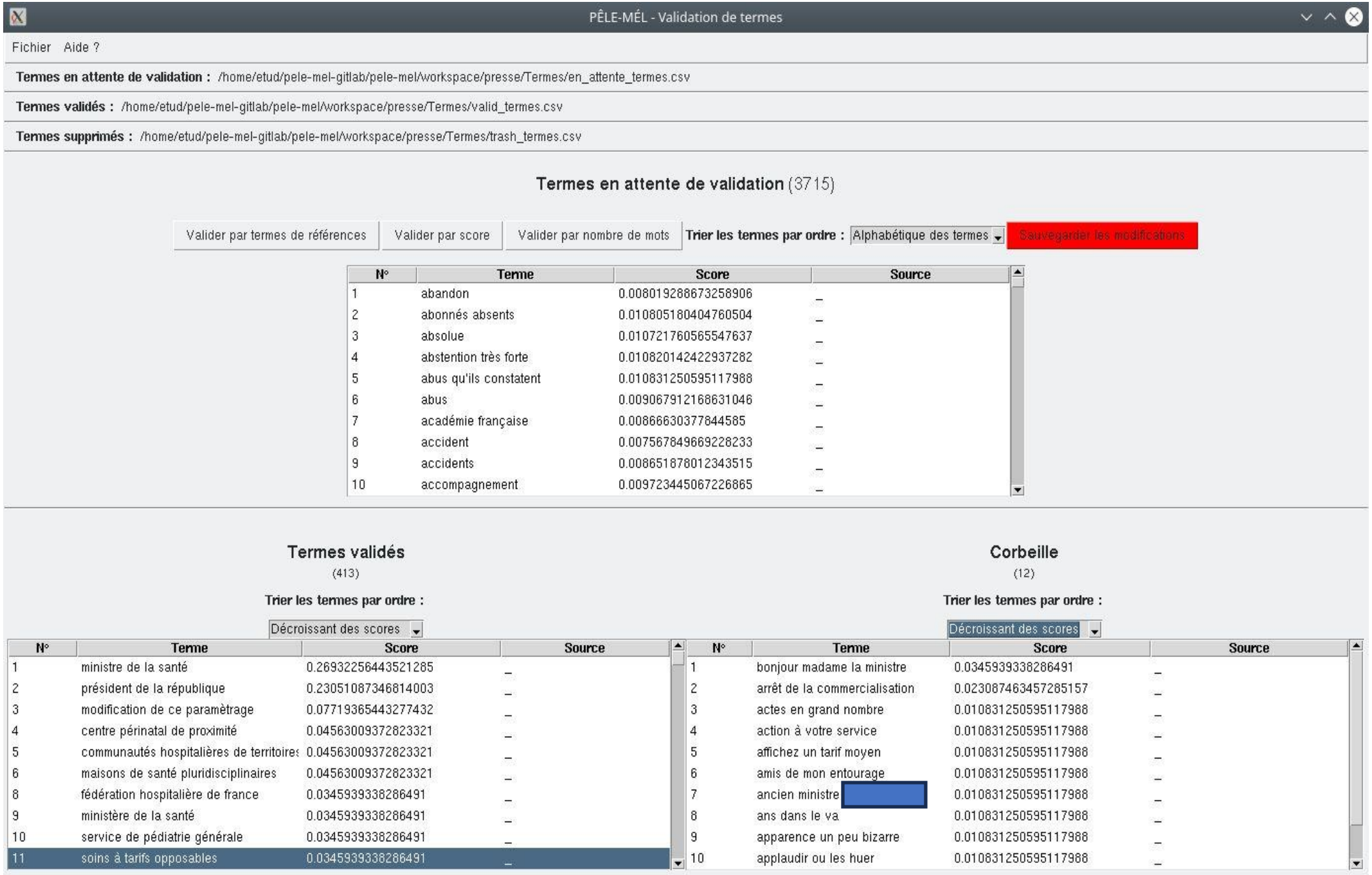
TF : plus le terme est fréquent plus son poids est élevé

IDF : mesure la rareté (poids plus élevé aux termes moins fréquents

TF_IDF : importance d'un terme dans un document par rapport à l'ensemble de corpus

Interface d'extraction de termes




PÊLE-MÊL - Validation de termes
Fichier Aide ?
Termes en attente de validation : /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Termes/en_attente_termes.csv
Termes validés : /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Termes/valid_termes.csv
Termes supprimés : /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Termes/trash_termes.csv
Termes en attente de validation (3715)
Valider par termes de références
Valider par score
Valider par nombre de mots
Trier les termes par ordre :
Alphabétique des termes
Sauvegarder les modifications
N° Terme Score Source
1 abandon 0.008019288673258906 _
2 abonnés absents 0.010805180404760504 _
3 absolue 0.010721760565547637 _
4 abstention très forte 0.010820142422937282 _
5 abus qu'ils constatent 0.010831250595117988 _
6 abus 0.009067912168631046 _
7 académie française 0.00866630377844585 _
8 accident 0.007567849669228233 _
9 accidents 0.008651878012343515 _
10 accompagnement 0.009723445067226865 _
Termes validés
(413)
Trier les termes par ordre :
Décroissant des scores
N° Terme Score Source
1 ministre de la santé 0.26932256443521285 _
2 président de la république 0.23051087346814003 _
3 modification de ce paramètrage 0.07719365443277432 _
4 centre périnatal de proximité 0.04563009372823321 _
5 communautés hospitalières de territoires 0.04563009372823321 _
6 maisons de santé pluridisciplinaires 0.04563009372823321 _
8 fédération hospitalière de france 0.0345939338286491 _
9 ministère de la santé 0.0345939338286491 _
10 service de pédiatrie générale 0.0345939338286491 _
11 soins à tarifs opposables 0.0345939338286491 _
Corbeille
(12)
Trier les termes par ordre :
Décroissant des scores
N° Terme Score Source
1 bonjour madame la ministre 0.0345939338286491 _
2 arrêt de la commercialisation 0.023087463457285157 _
3 actes en grand nombre 0.010831250595117988 _
4 action à votre service 0.010831250595117988 _
5 affichez un tarif moyen 0.010831250595117988 _
6 amis de mon entourage 0.010831250595117988 _
7 ancien ministre 0.010831250595117988 _
8 ans dans le va 0.010831250595117988 _
9 apparence un peu bizarre 0.010831250595117988 _
10 applaudir ou les huer 0.010831250595117988 _

Résultat de l'extraction des termes dans le corpus

Messages avec pièce(s) jointe(s) :
149 fichiers (pdf, doc, txt, xml)
14 563 phrases
316 496 mots

Nombre de termes proposés	**4493**
Nombre de termes corrects	4136
Nombre de termes incorrects	357

SPACY

Extraction des entités nommées

“

une bibliothèque Python gratuite et open source publiée sous la licence MIT
pour le traitement naturel du langage,

Elle peut être utilisé notamment pour développer des systèmes d’extraction d’information, de compréhension du langage naturel, ou encore pour pré-traiter des textes pour le Deep Learning.”

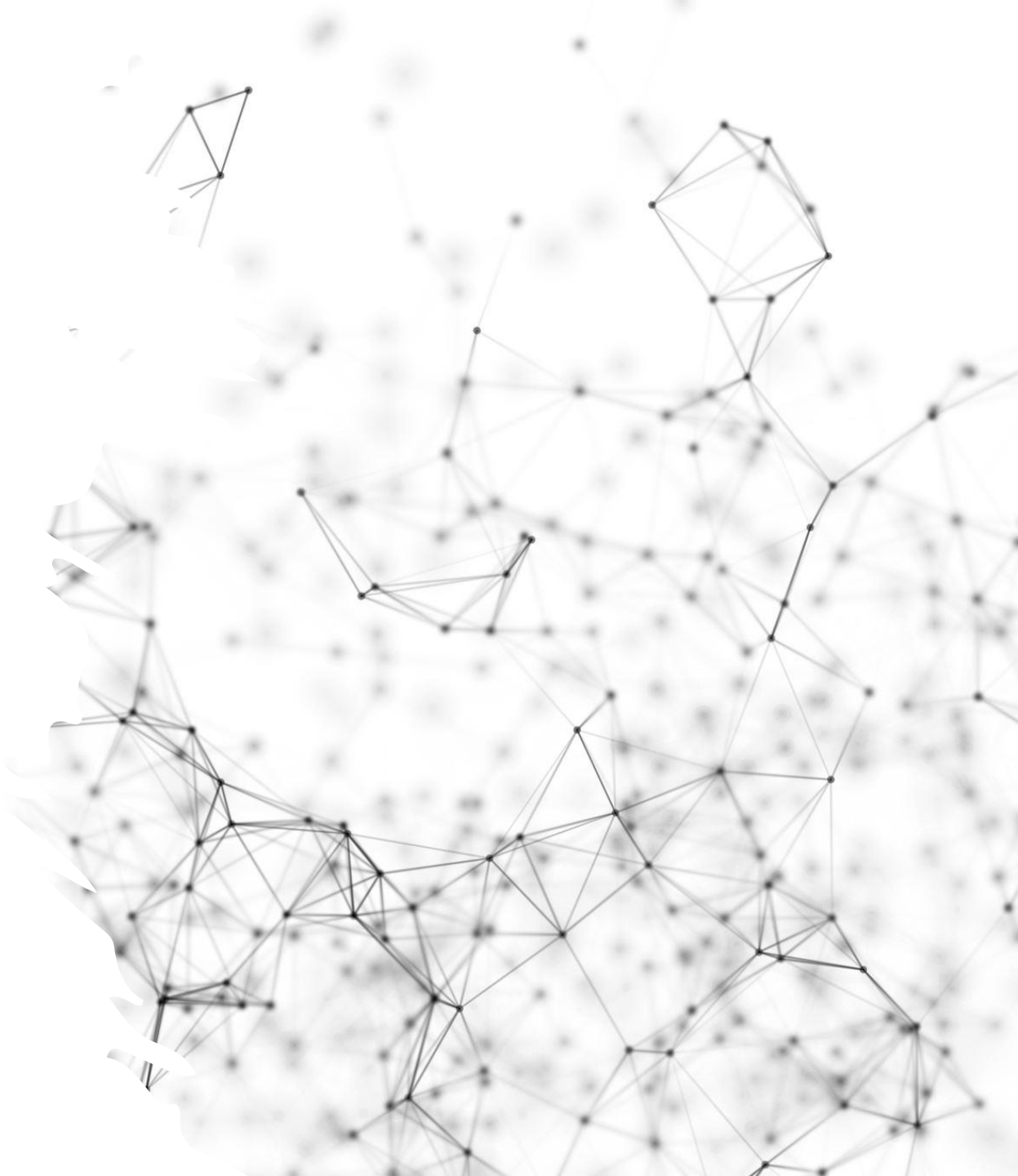
Spacy contient des algorithmes de Deep Learning qui permettent d’extraire des informations en utilisant l'intelligence artificielle.

Résultat de l'extraction des entités nommées dans le corpus

Messages avec pièce(s) jointe(s) : 149 fichiers (pdf, doc, txt, xml)

Entités nommées	Personnes	Organisations
Proposées	450	510
Correctes	424	366
Incorrectes	26	144

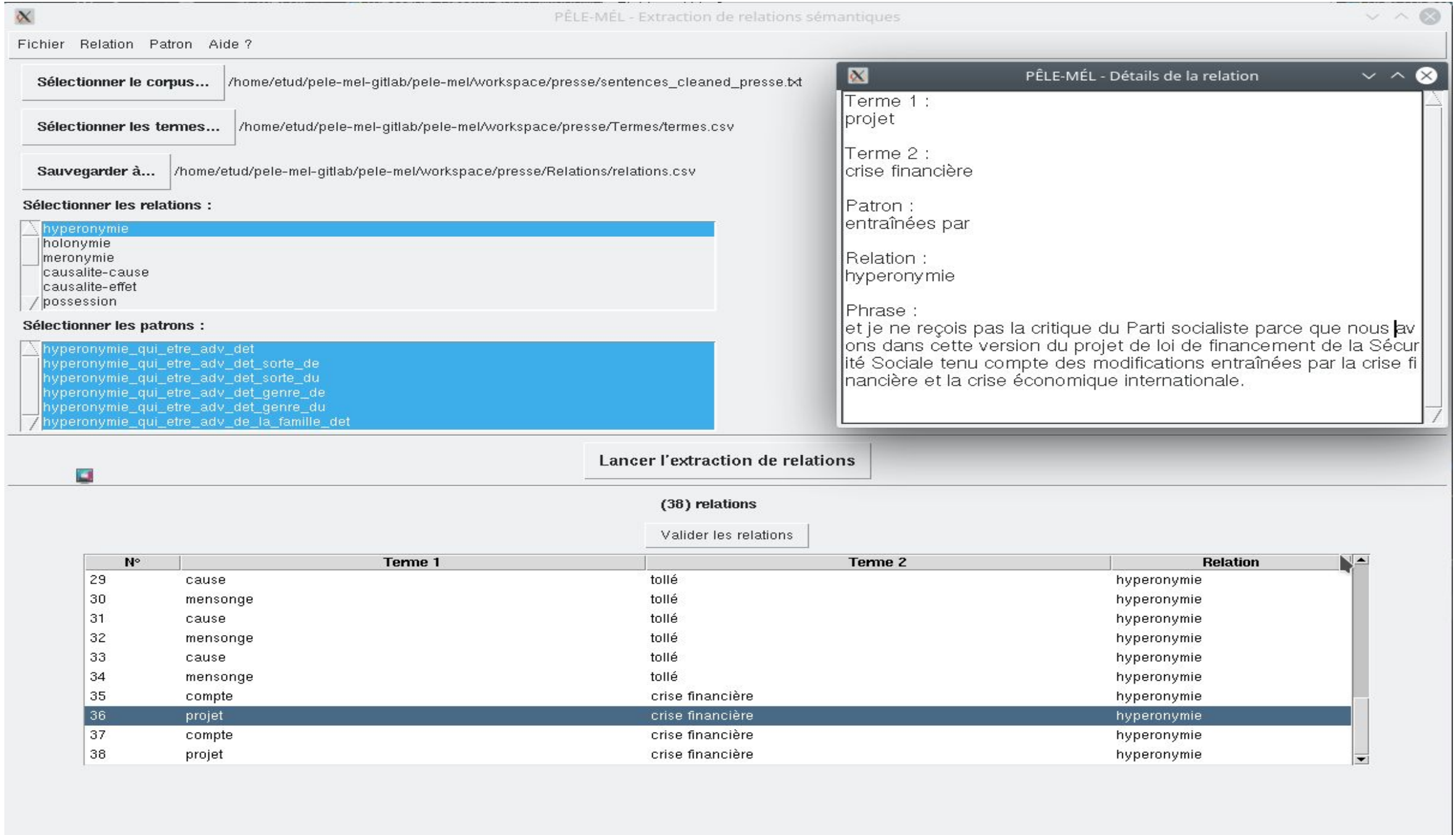
Extraction des relations entre les termes et aussi entre les termes et les entités nommées



Méthode par patrons

Recherche de régularités. Ex. :

Hyperonymie	Holonymie	Méronymie	Causalité (cause)	Causalité (effet)	Possession
• Être un • Être une sorte de • Être du genre • Être de la famille	• Composé de • Se compose • Constitué de • Se constitue	• Faire partie de • Compose le • Forme le • Constitue le • Être une partie de	• Causer • Provoquer • Engendrer • Produire • Déclencher • Générer	• Dû à • Causé par • Provoqué par • Engendré par • Produit par • Déclenché par	• Avoir • Posséder

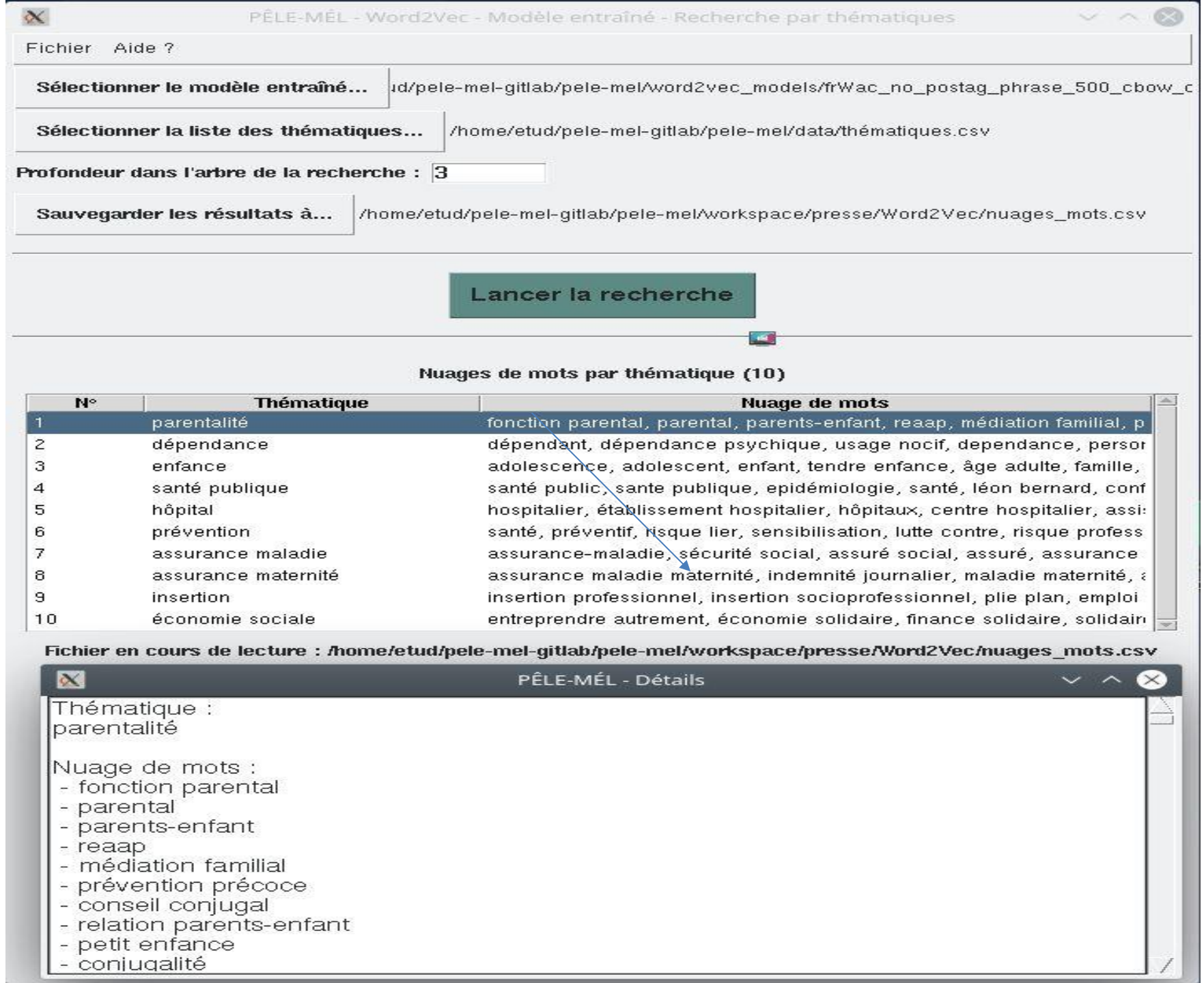

PÊLE-MÊL - Extraction de relations sémantiques
Fichier Relation Patron Aide ?
Sélectionner le corpus...
/home/etud/pele-mel-gitlab/pele-mel/workspace/presse/sentences_cleaned_presse.txt
Sélectionner les termes...
/home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Termes/termes.csv
Sauvegarder à...
/home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Relations/relations.csv
Sélectionner les relations :
hyperonymie
holonymie
meronymie
causalite-cause
causalite-effet
possession
Sélectionner les patrons :
hyperonymie_qui_etre_adv_det
hyperonymie_qui_etre_adv_det_sorte_de
hyperonymie_qui_etre_adv_det_sorte_du
hyperonymie_qui_etre_adv_det_genre_de
hyperonymie_qui_etre_adv_det_genre_du
hyperonymie_qui_etre_adv_de_la_famille_det
PÊLE-MÉL - Détails de la relation
Terme 1 :
projet
Terme 2 :
crise financière
Patron :
entraînées par
Relation :
hyperonymie
Phrase :
et je ne reçois pas la critique du Parti socialiste parce que nous avons dans cette version du projet de loi de financement de la Sécurité Sociale tenu compte des modifications entraînées par la crise financière et la crise économique internationale.
Lancer l'extraction de relations
(38) relations
Valider les relations
N° Terme 1 Terme 2 Relation
29 cause tollé hyperonymie
30 mensonge tollé hyperonymie
31 cause tollé hyperonymie
32 mensonge tollé hyperonymie
33 cause tollé hyperonymie
34 mensonge tollé hyperonymie
35 compte crise financière hyperonymie
36 projet crise financière hyperonymie
37 compte crise financière hyperonymie
38 projet crise financière hyperonymie

Plongement lexical

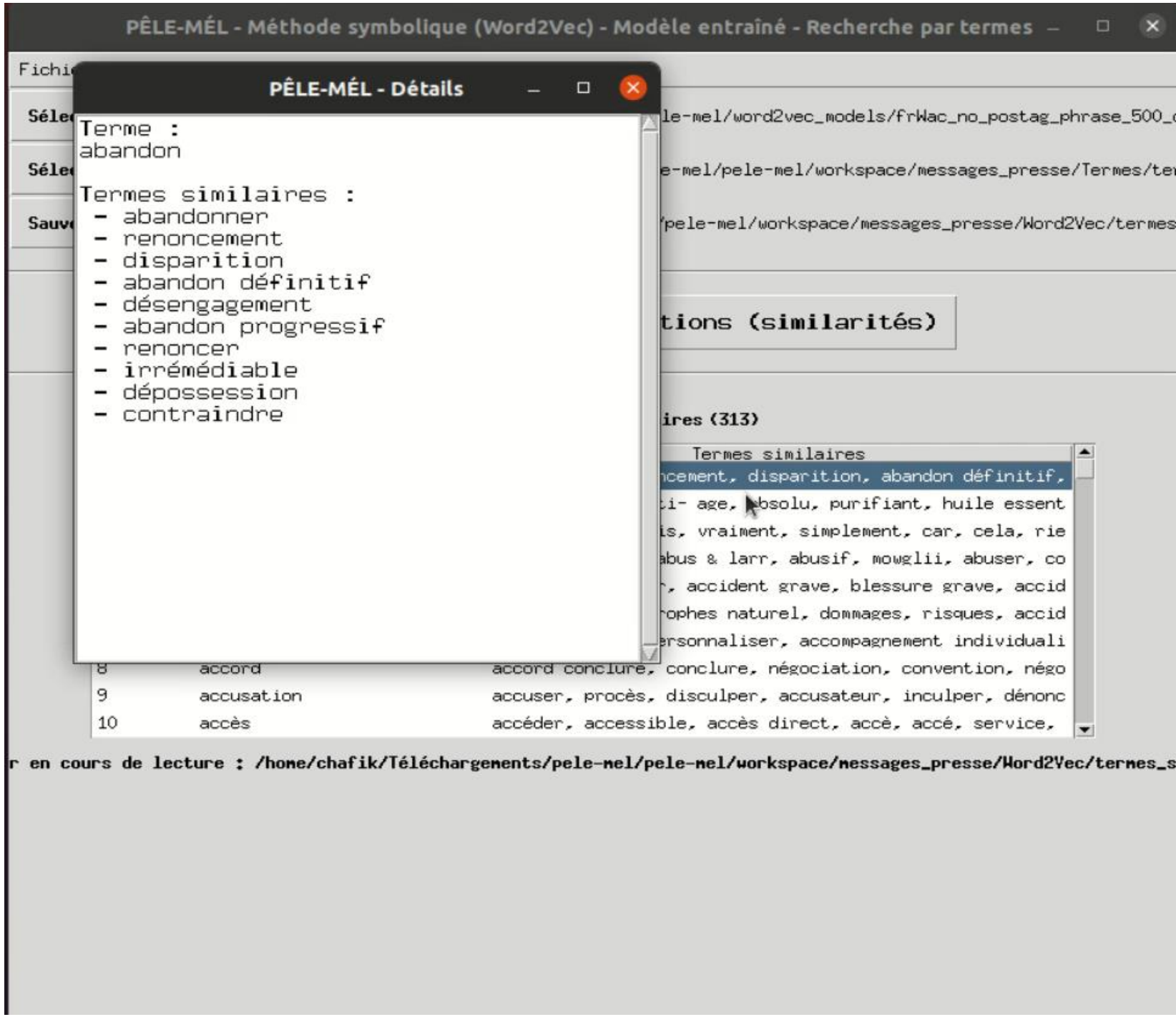
Word2Vec

Un réseau de neurones artificiels à deux couches entraînées pour le contexte linguistique des mots

- Objectif : représenter les mots en fonction de leur contexte en capturant les similarités sémantiques et syntaxiques > les termes et entités nommées sont représentés par des vecteurs de nombres réels. On récupère ensuite les vecteurs qu'on compare et on conserve les vecteurs (termes) les plus proches en distance.


PÊLE-MÊL - Word2Vec - Modèle entraîné - Recherche par thématiques
Fichier Aide ?
Sélectionner le modèle entraîné...
ud/pele-mel-gitlab/pele-mel/word2vec_models/frWac_no_postag_phrase_500_cbow_c
Sélectionner la liste des thématiques...
/home/etud/pele-mel-gitlab/pele-mel/data/thématiques.csv
Profondeur dans l'arbre de la recherche : 3
Sauvegarder les résultats à...
/home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Word2Vec/nuages_mots.csv
Lancer la recherche
Nuages de mots par thématique (10)
N° Thématique Nuage de mots
1 parentalité fonction parental, parental, parents-enfant, reaap, médiation familial, p
2 dépendance dépendant, dépendance psychique, usage nocif, dependance, perso
3 enfance adolescence, adolescent, enfant, tendre enfance, âge adulte, famille,
4 santé publique santé public, sante publique, epidémiologie, santé, léon bernard, conf
5 hôpital hospitalier, établissement hospitalier, hôpitaux, centre hospitalier, assi
6 prévention santé, préventif, risque lier, sensibilisation, lutte contre, risque profess
7 assurance maladie assurance-maladie, sécurité social, assuré social, assuré, assurance
8 assurance maternité assurance maladie maternité, indemnité journalier, maladie maternité,
9 insertion insertion professionnel, insertion socioprofessionnel, plie plan, emploi
10 économie sociale entreprendre autrement, économie solidaire, finance solidaire, solidair
Fichier en cours de lecture : /home/etud/pele-mel-gitlab/pele-mel/workspace/presse/Word2Vec/nuages_mots.csv
PÊLE-MÊL - Détails
Thématique :
parentalité
Nuage de mots :
- fonction parental
- parental
- parents-enfant
- reaap
- médiation familial
- prévention précoce
- conseil conjugal
- relation parents-enfant
- petit enfance
- conjugalité

Résultat (word2vec)



Exemple des liens entre termes et entités nommées

Num	Thématique	Nuage de mots
1	parentalité	fonction parental, parental, parents-enfant, reaap, médiation familial, prévention précoce, conseil conjugal, relation parents-enfant, petit enfance, conjugalité
2	dépendance	dépendant, dépendance psychique, usage nocif, dependance, personne âgé, cinquième branche, dépendance envers, personne dépendant, grille aggir, personne âgé dépendant
3	enfance	adolescence, adolescent, enfant, tendre enfance, âge adulte, famille, mère, enfance adolescence, jeunesse, parent divorcer
4	santé publique	santé public, sante publique, epidémiologie, santé, léon bernard, conférence régionale, épidémiologie, drassif, formation médicale continue, credes

- Liens entre termes

prévention : dépistage, contraception, transmissible, lutte contre la maltraitance,…
médicament : pharmaceutique, prescrit, fabrication, générique, circuit, remboursable

- Liens entre entités nommées et termes

personne 1 : ministère, prévention, jeunesse, sécurité, DGCS, vaccin, politique de la famille, solidarité, familiale, H1N1, etc.
personne 2 : parité, violence conjugale, victime, délégation européenne, vie associative, etc.

Nuage de termes

Messages avec pièce(s) jointe(s) : 149 fichiers (pdf, doc, txt, xml)

Utilisation d'une liste de 4136 termes.

10 termes similaires (en moyenne) pour 1681 termes.

Profondeur de recherche : 2.

La quasi-totalité des termes similaires extraits sont corrects.

Classification

Classification des messages

Apprentissage non supervisé (Clustering)

Pas de classes prédéfinies

Pas besoin d'annotation

Résultats : regroupement des données en fonction de leurs caractéristiques

Apprentissage supervisé (Machine learning)

Besoin d'exemples d'apprentissage annotés

Résultats : prédiction de la classe d'une donnée en fonction de ses caractéristiques

Règles

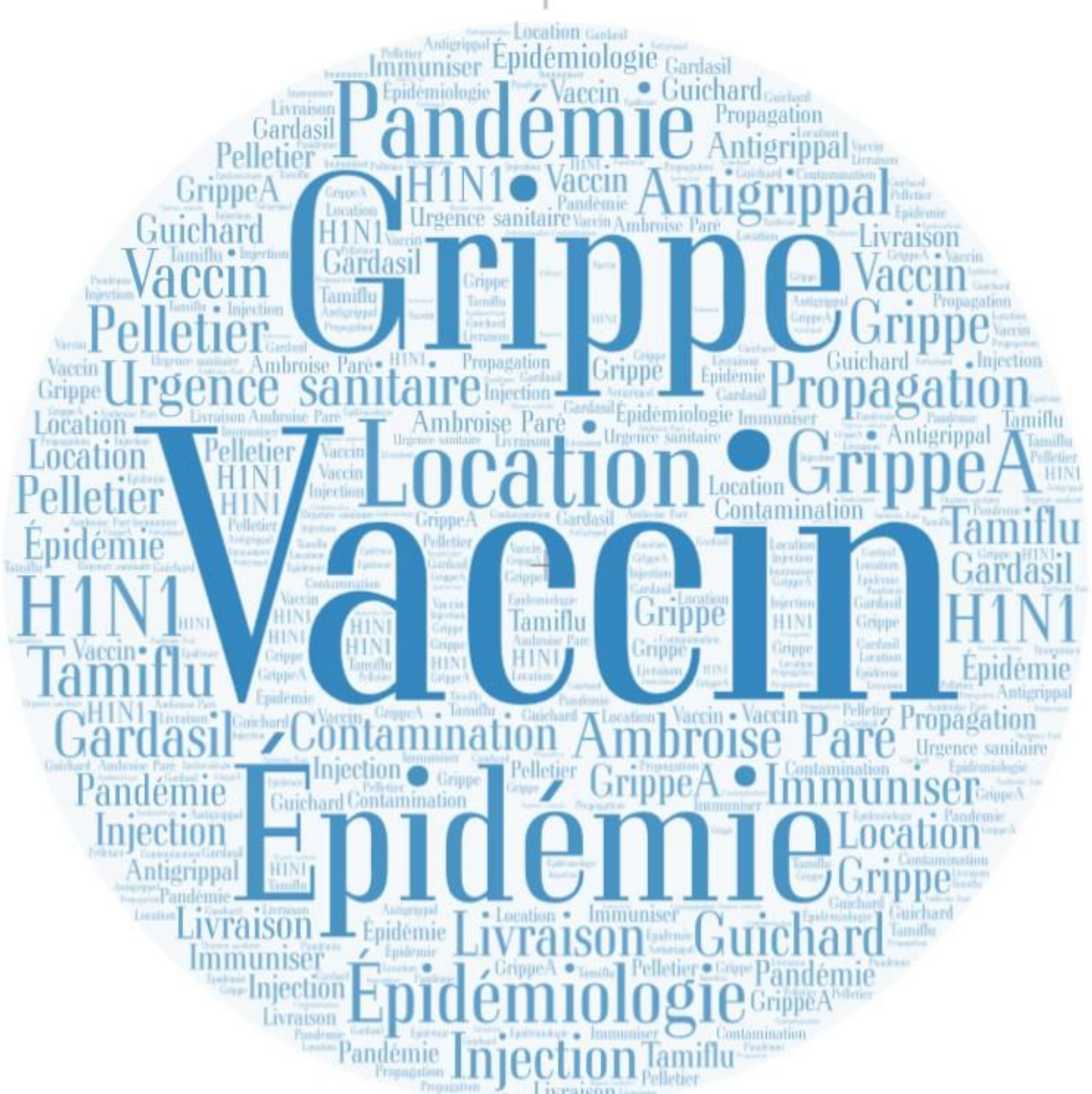
Recherche de régularité

Utilisation de patrons

Résultats : prédiction de relations

Vocabulaires

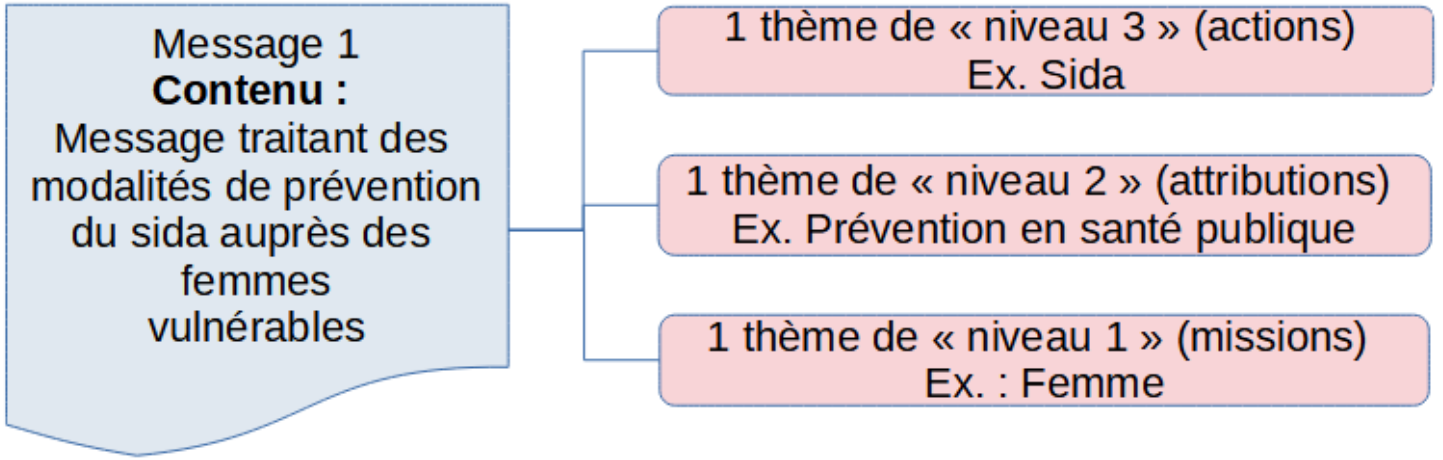
Thesaurus
Index



Guider la classification en associant des thèmes proposés par l'archiviste et le nuage de termes/entités nommées qui peut leur être associé

Étape suivante : associer des thèmes et des messages via le plongement de documents

- Représenter les messages par différentes thématiques en les reliant aux thèmes et à leur(s) nuage(s) en jouant sur les facettes / niveaux = associer un message à n thématiques
- Chaque unité est représentée par un vecteur de nombre réel qu'on compare aux vecteurs des thèmes
- Paramétrage : dimensionnalité des vecteurs, fréquence de mots à ignorer, nombre d'itérations



Pour guider la classification

Env. 70

Attributions

- Famille
- Handicap / dépendance
- Femme
- Enfant et adolescence
- Jeunesse
- Sport
- Vie associative
- Santé / santé publique
- Protection sociale

Missions

- Hôpital
- Prévention
- Médicament
- Santé publique
- Communication
- Protection sociale
- Professions médicales
- Etc.

Actions – Programmes

- Accueil familial
- Aide à domicile
- Alcool
- Autisme
- Dépendance
- Loi HPST
- Médiator
- SIDA
- etc

Thèmes proposés par l'archiviste conformes à ses méthodes d'analyse des flux documentaires
Différents niveaux ou facettes

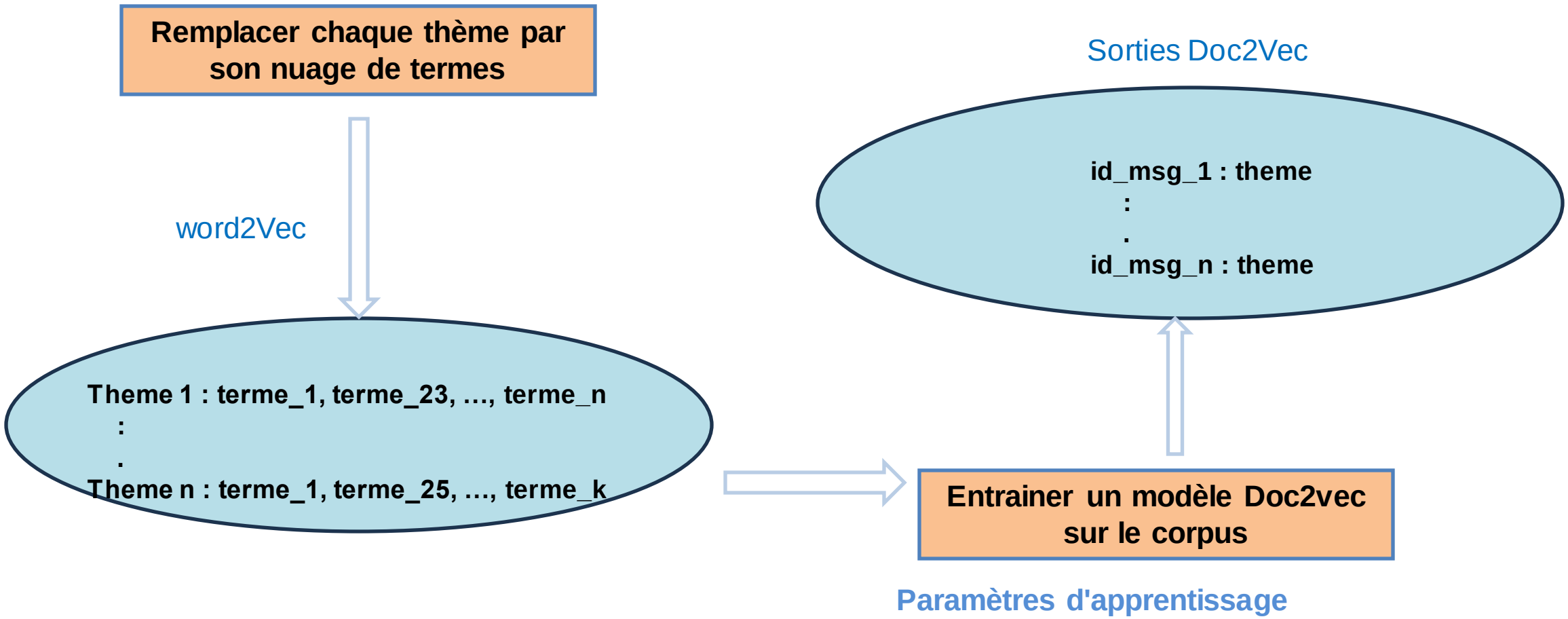
Doc2vec

- Similaire à word2vec

- Wor2vec : un vecteur pour chaque terme
- Doc2vec : un vecteur pour chaque document + un vecteur qui représente les documents
- le modèle Doc2vec associe chaque document à un vecteur numérique basé sur le contexte dans lequel il apparaît dans le corpus

Injecter des connaissances lexicales dans une méthode de plongement de mots qui utilise un réseau de neurones à deux couches

Classification des messages par thèmes



Classification des messages par thème

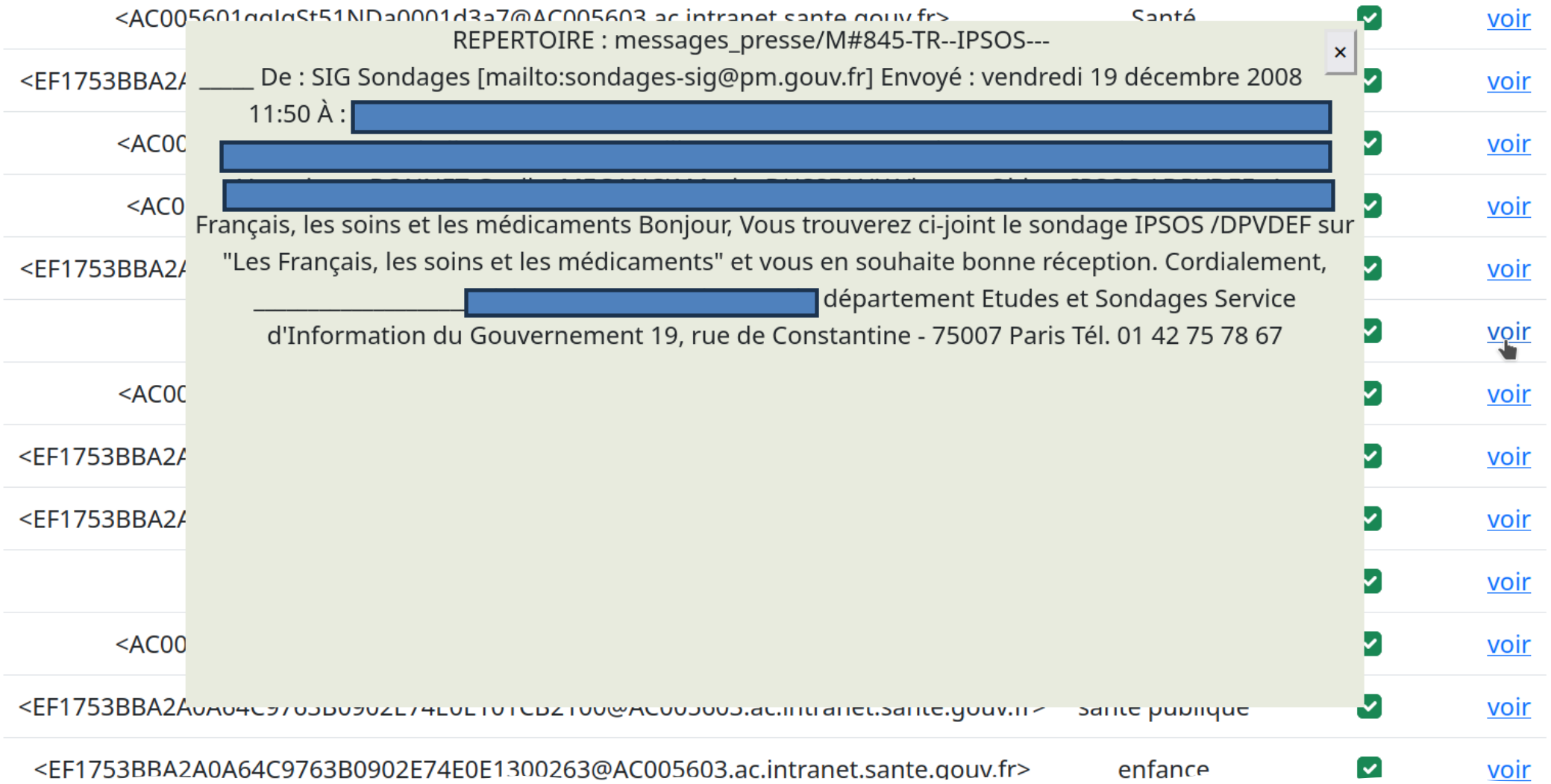
Id message	Extrait de message	thèmes
EF1753BBA2A0A64C9763B0902E74E0E1300263@AC005603.ac.intranet.sante.gouv.fr	**Sida, l'Etat dérape et fait silence…**	**Protection sociale**
EF1753BBA2A0A64C9763B0902E74E0E1442F02@AC005603.ac.intranet.sante.gouv.fr	**Projet de réforme du système de santé… Sur les 10 ministres les plus appréciés des Français, six sont des femmes… Ce nouveau virus contient des gènes de plusieurs virus d'origine porcine**	**Grippe A**
…	…	…

Résultat de la classification

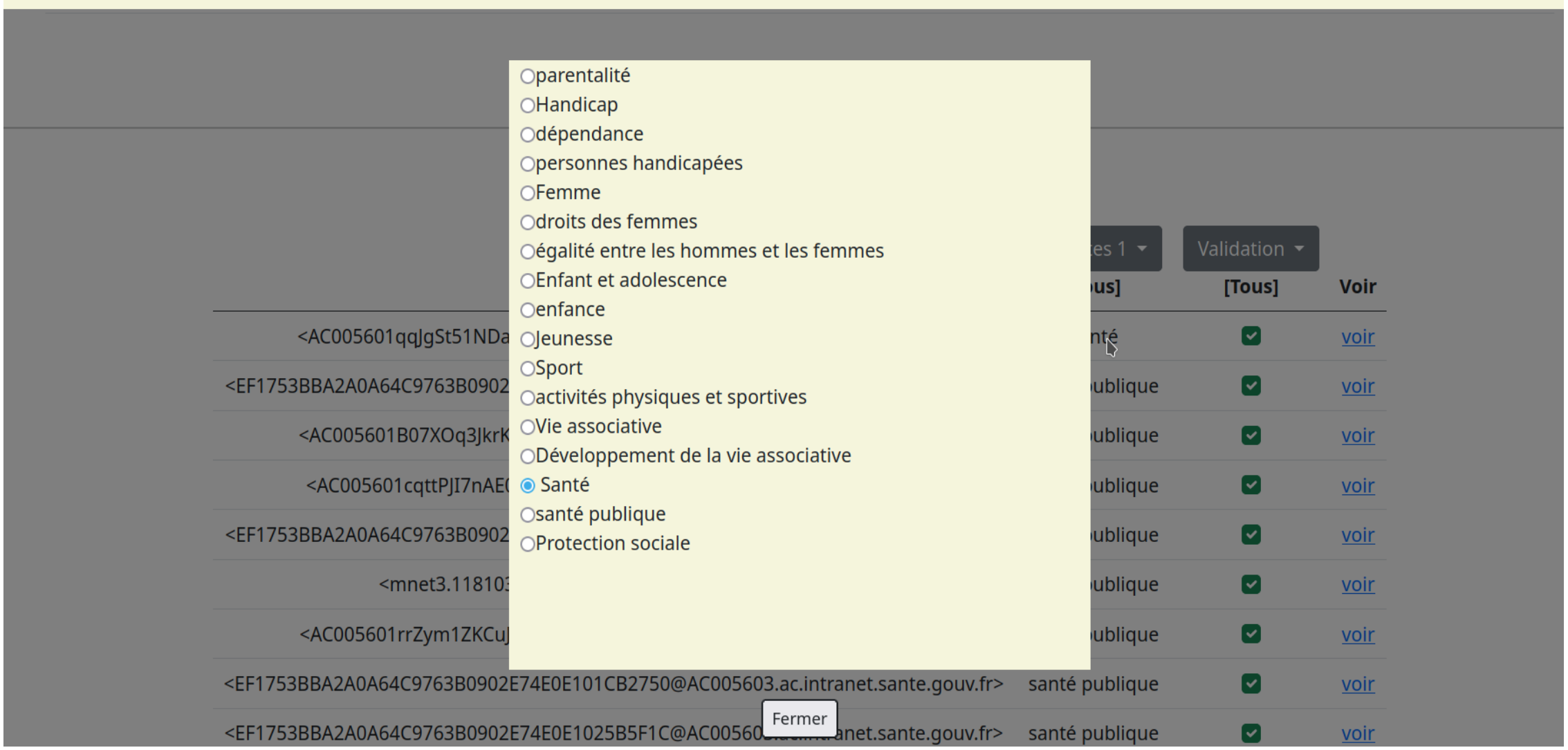
Télécharger

ID Message	Facettes 1 [Tous]	Validation [Tous]	Voir
<AC005601qqJgSt51NDa0001d3a7@AC005603.ac.intranet.sante.gouv.fr>	Santé	☑	voir
<EF1753BBA2A0A64C9763B0902E74E0E1025B5FE1@AC005603.ac.intranet.sante.gouv.fr>	santé publique	☑	voir
<AC005601B07XOq3JkrK00001b5c@AC005603.ac.intranet.sante.gouv.fr>	santé publique	☑	voir
<AC005601cqttPJI7nAE00014a2c@AC005603.ac.intranet.sante.gouv.fr>	santé publique	☑	voir
<EF1753BBA2A0A64C9763B0902E74E0E1025B5FE0@AC005603.ac.intranet.sante.gouv.fr>	santé publique	☑	voir
<mnet3.1181036948.29516.beatrice.noellec@noos.fr>	santé publique	☑	voir
<AC005601rrZym1ZKCuJ0000c371@AC005603.ac.intranet.sante.gouv.fr>	santé publique	☑	voir
<EF1753BBA2A0A64C9763B0902E74E0E101CB2750@AC005603.ac.intranet.sante.gouv.fr>	santé publique	☑	voir

Vérification du contenu



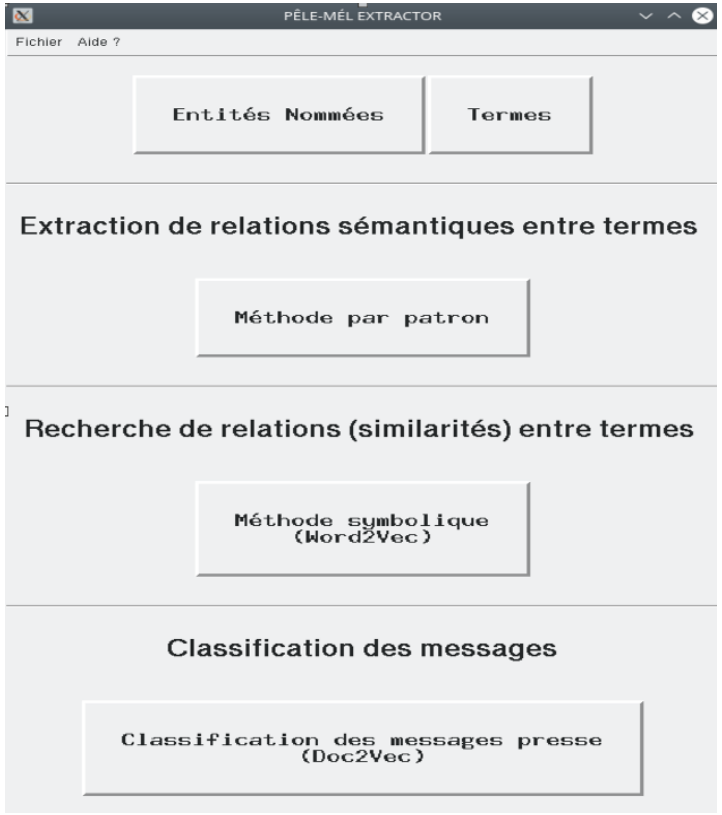
Modification de la classification



Exploitation

Présupposé : interface de classification

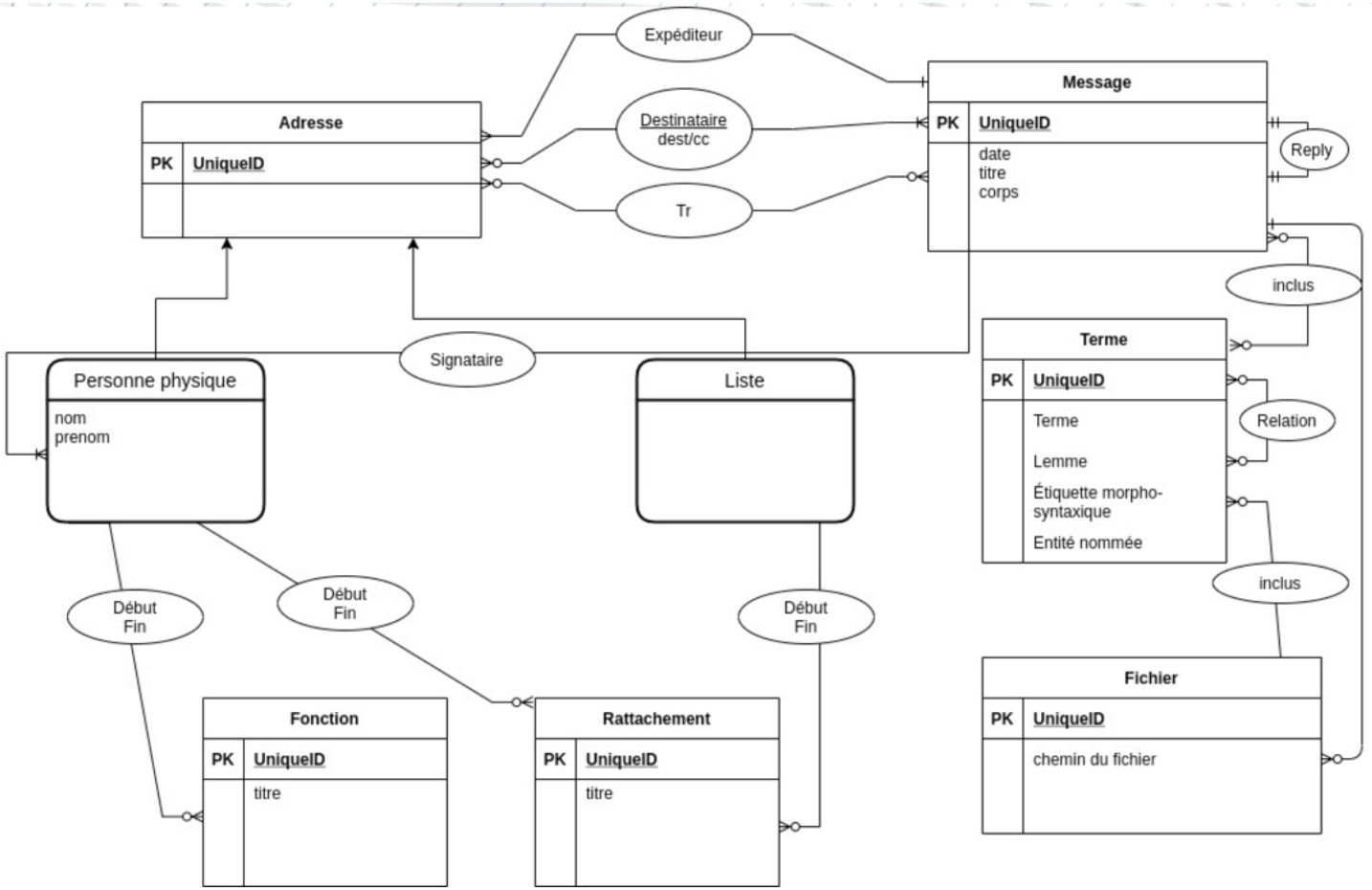
- Pré-traitement et segmentation en phrases
- Extractions d'entités nommées et de termes, validations automatiques et manuelles
- Création de relations entre thèmes et termes et classification avec un modèle pré-entraîné : création des nuages de termes
- Association des messages avec des thématiques



Prototype d'exploration : deux applications en une

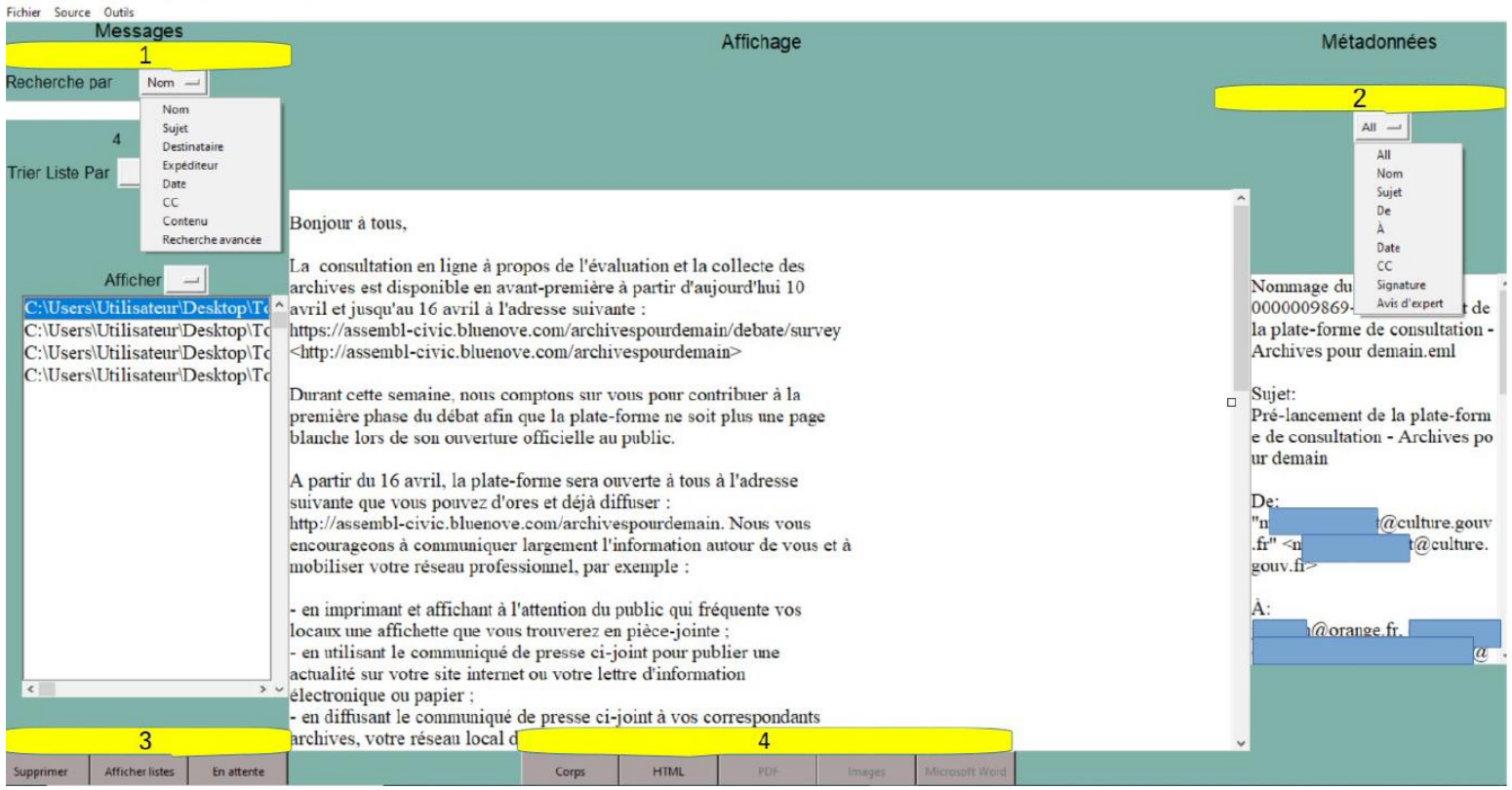
Une application très légère, installable par copier / coller directement sur un bureau

Une base de données relationnelle qui peut être connectée ouvrant d'autres possibilités de recherche et intégrant les fichiers de sortie de la classification

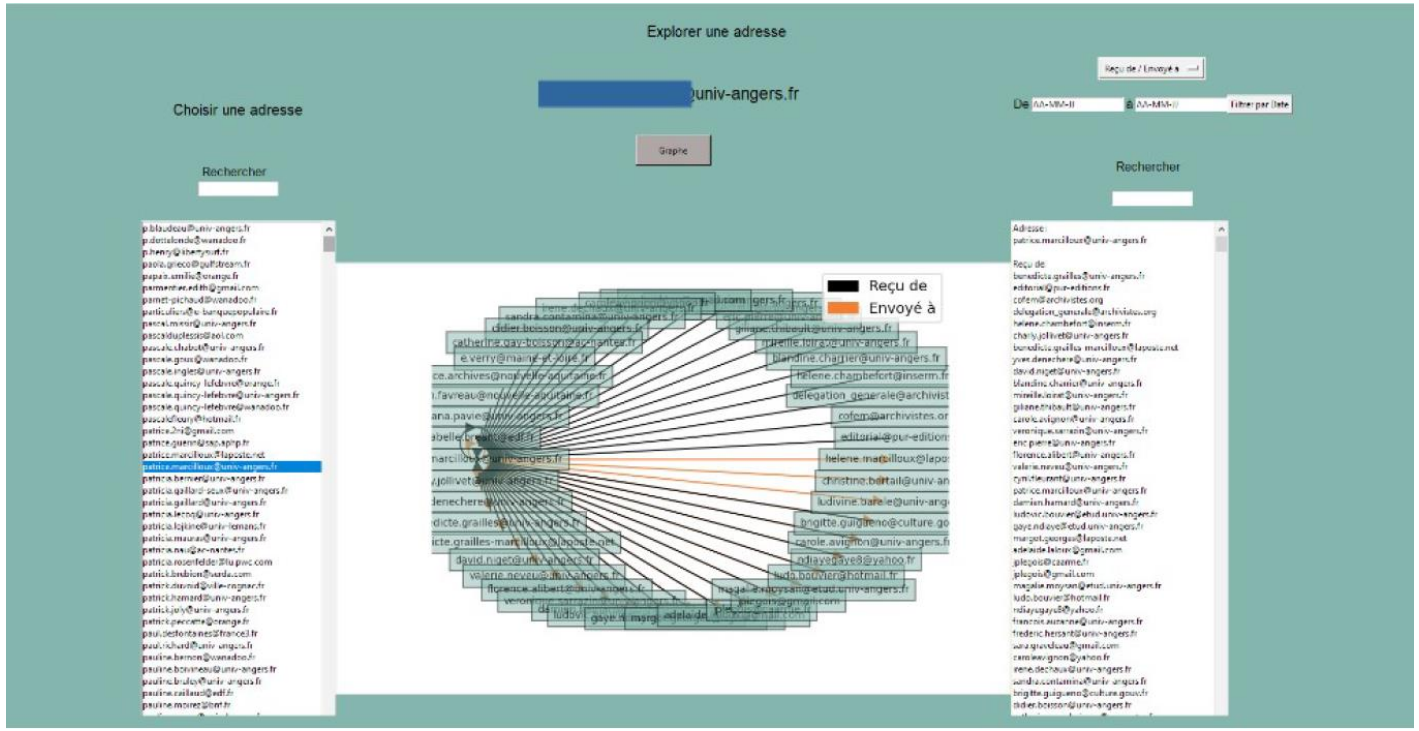


Fonctionnalités de base (1)

- Interrogation des métadonnées et du contenu des messages et filtrage les résultats [1, 2] : nom, objet, destinataire, expéditeur, date, CC, contenu (message et pièces jointes). Une recherche avancée est également possible.
- Liste des messages [3] affichable et modifiable dans une fenêtre.
- Contenu des messages [4] et pièces jointes affichables dans leur format d'archivage (pdf, format image, word).



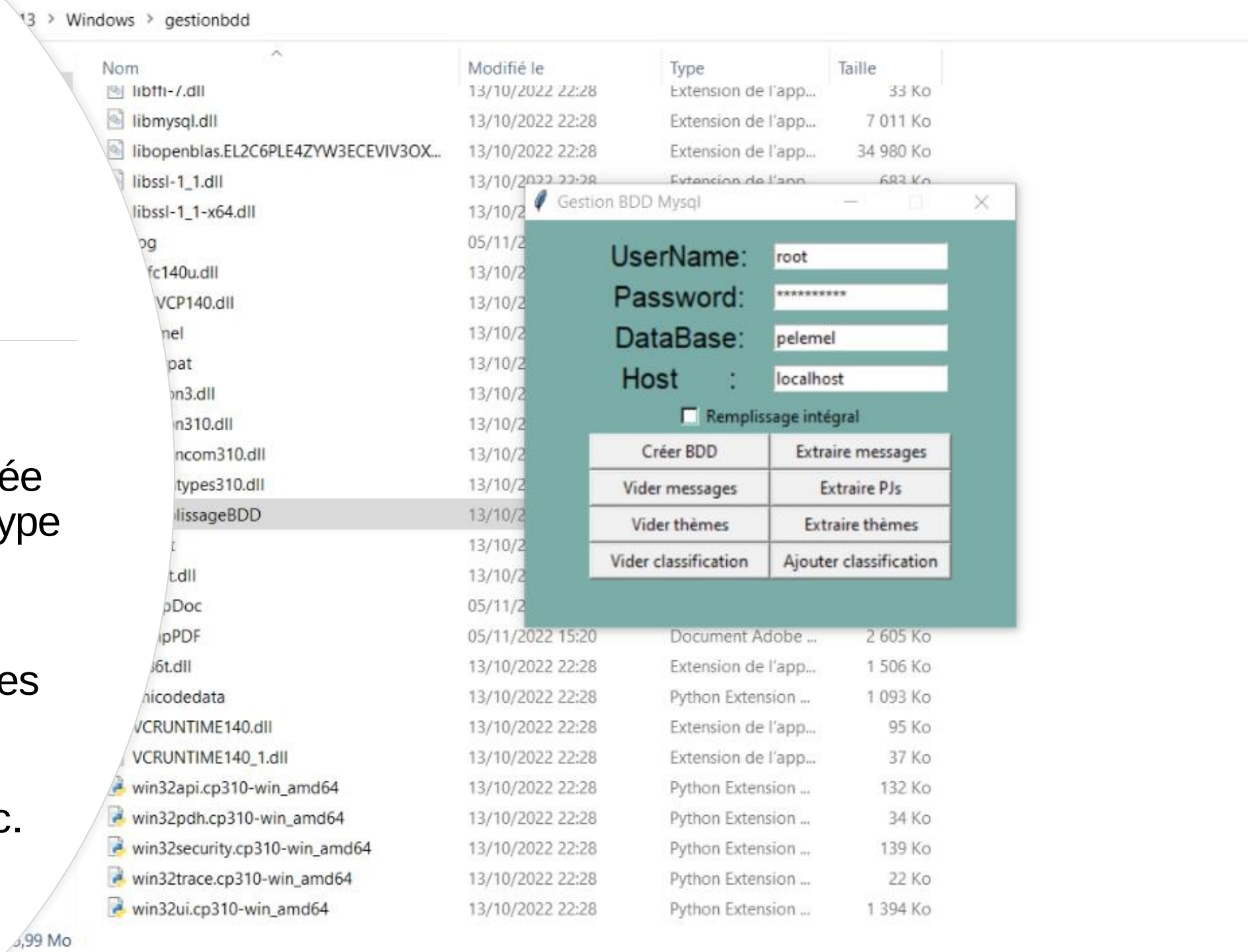
Fonctionnalités de base (2)



- Relations entre une adresse et ses correspondants (messages reçus et envoyés) sous la forme d'un graphique
- Génération d'un graphique dynamique et paramétrable possible
- Fonctions de filtrage et de recherche d'adresses et de dates pour affiner le graphique, dont l'image peut être exportée.

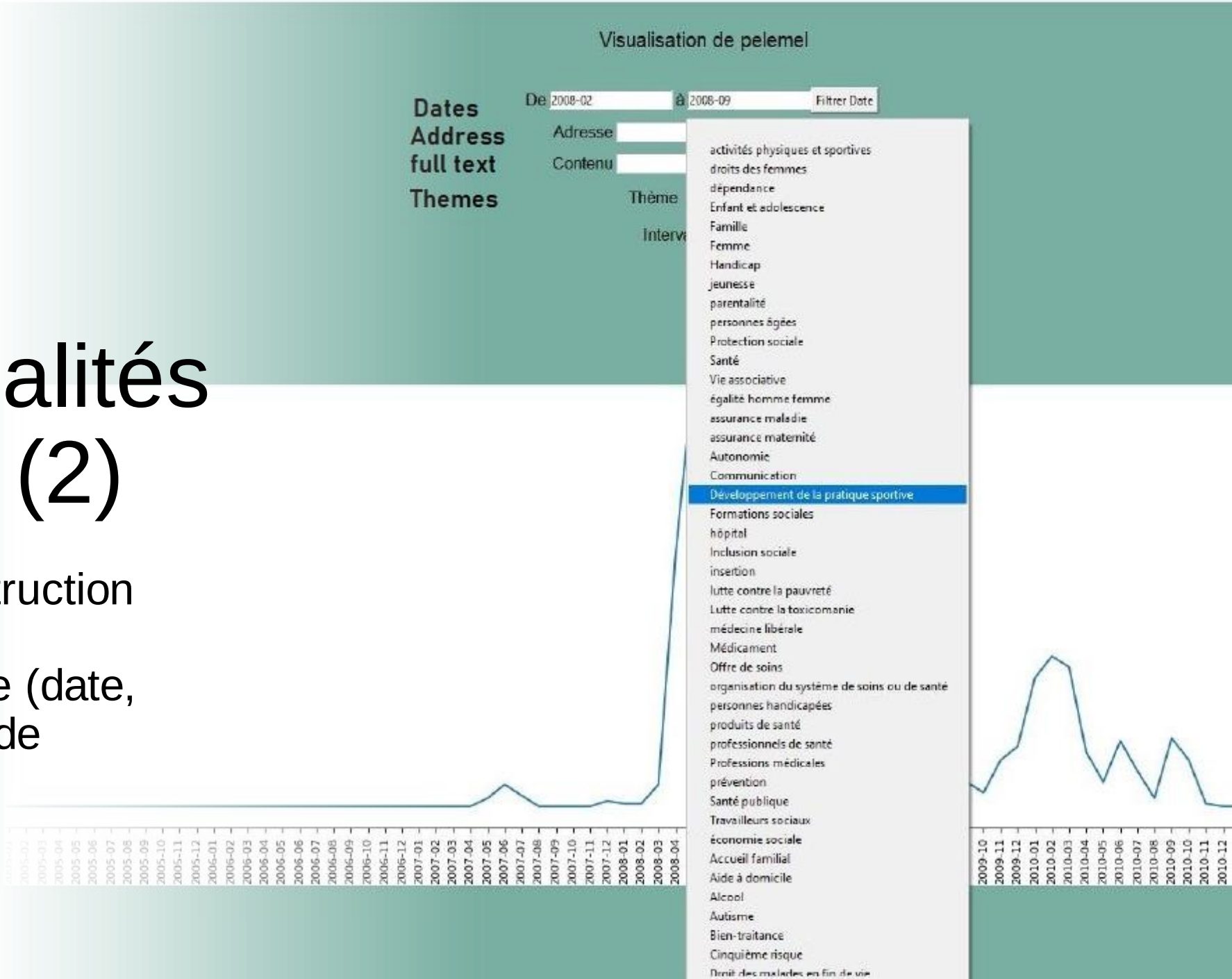
Fonctionnalités avancées (1)

- Exploitation de la classification effectuée sur le premier prototype via une base de données mysql
- Chargement des listes de thèmes et des fichiers de sortie de word2vec et doc2vec.



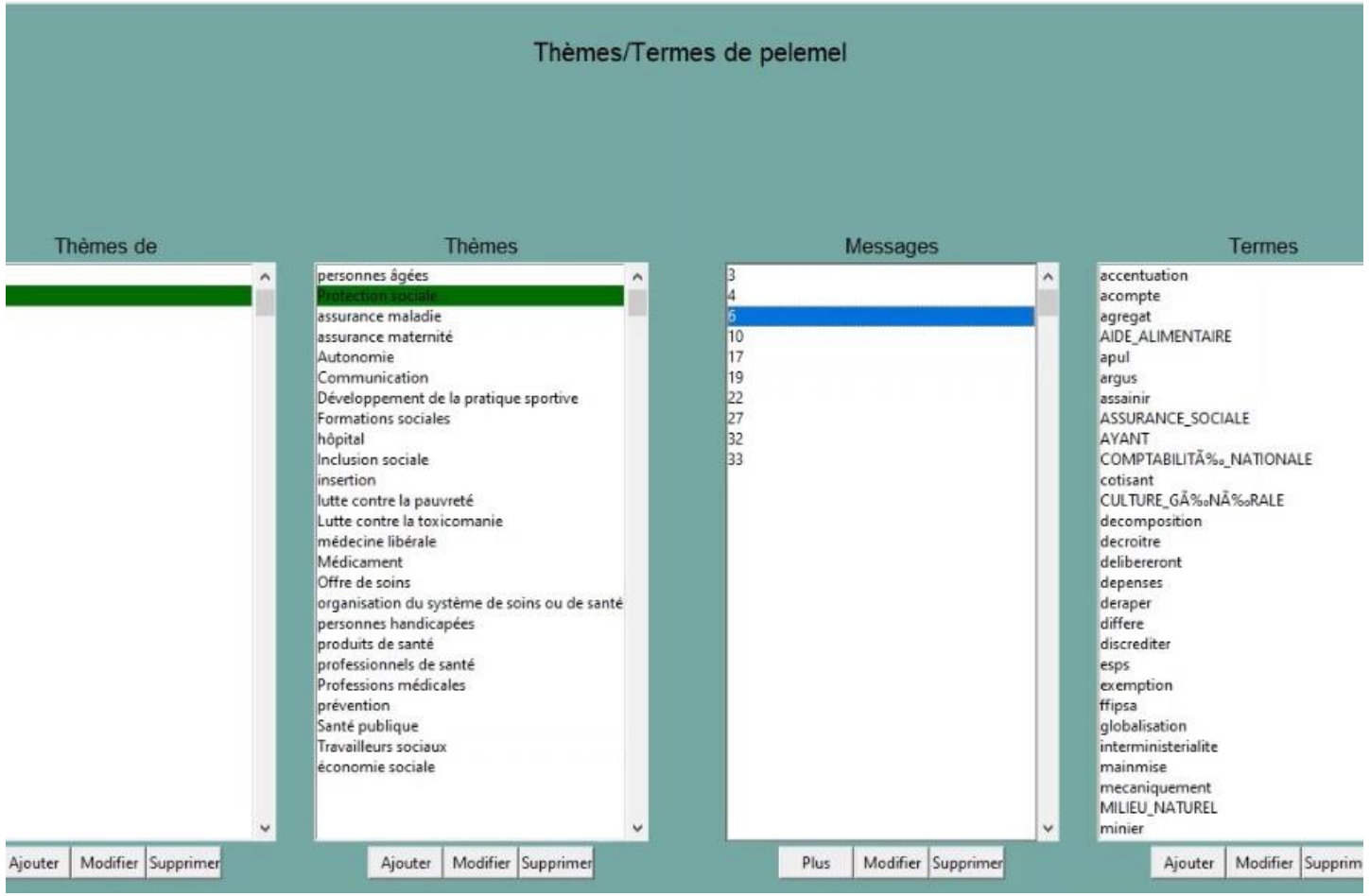
Fonctionnalités avancées (2)

Visualisation : construction de graphiques avec possibilité de filtrage (date, adresse, thème) et de recherche en texte intégral (résultat exportable)



Fonctionnalités avancées (3)

L'archiviste peut modifier les tableaux, annoter et corriger les relations thèmes-termes-messages via la base de données,



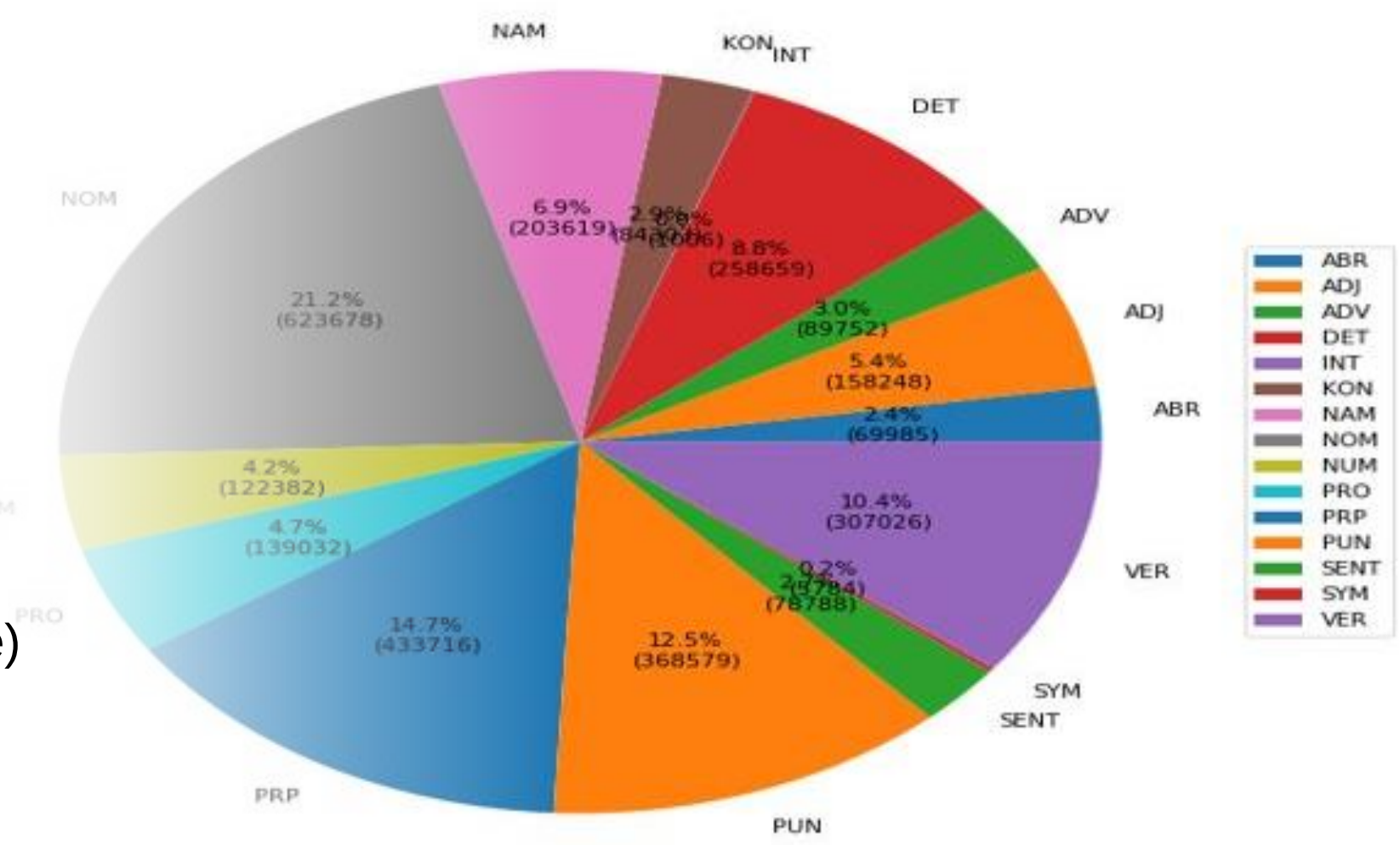
Quels enseignements ?

Identifier les manques / modifier sa politique

- Sigles
- Annuaire Ldap
- Liste de diffusion / adresse fonctionnelle
- Liens vers serveur
- Ne plus travailler "à la pièce" (passer du "client" au "serveur")

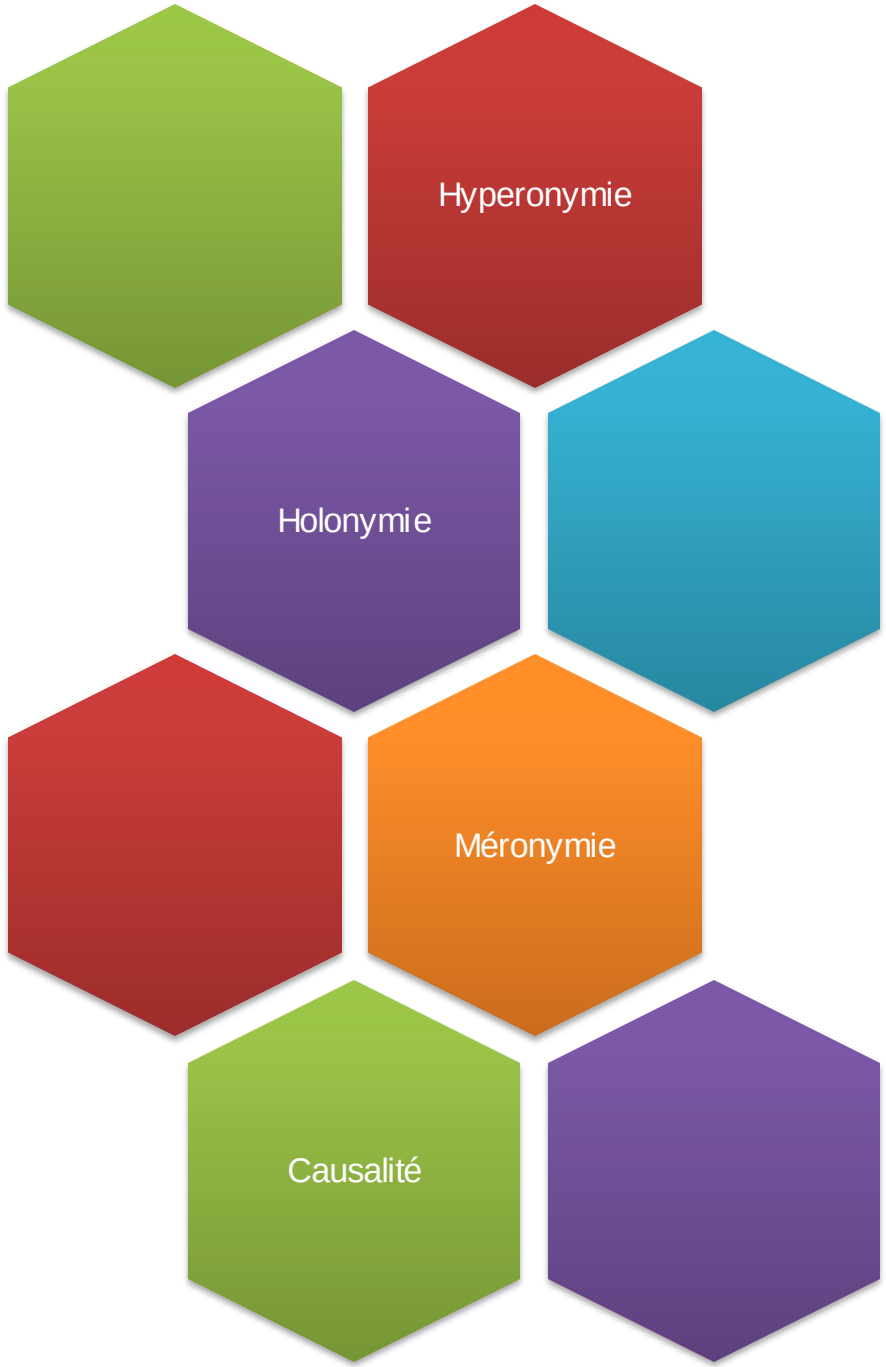
Évaluation archivistique
Choisir les boîtes

- Qualité et complexité de l'écriture (étiquetage morphosyntaxique, longueur des phrases)
- Réseaux (visualisation)
- Variété des correspondants
- Type de flux (interne, externe)
- Proportion des adresses fonctionnelles
- Proportion des copies à
- Fréquence



Évaluation archivistique
Faire du tri interne

- Messages "privés" non signalés, messages dit "engageants" (méthode des patrons)
- Conversation ping pong
- Long fil de discussion



Améliorer l'accès aux archives

- Intégration des éléments dans les instruments de recherche pour mieux décrire et caractériser les messageries (registre des correspondants, facettes, évolution chronologique ...)
- Recherche à la demande

Cada | Commission d'accès aux documents administratifs

Contacter la CADA | Rechercher

Accueil » Avis 20226133 - Séance du 15/12/2022

Avis 20226133 - Séance du 15/12/2022

Direction générale des patrimoines et de l'architecture

Monsieur X, X, a saisi la Commission d'accès aux documents administratifs, par courrier enregistré à son secrétariat le 2 septembre 2022, à la suite du refus opposé par le directeur général des patrimoines à sa demande de communication, par courriel, des documents suivants mentionnant le Health Data Hub ou son acronyme HDH :

1) les correspondances (courriers, courriels, ou autres) reçus ou envoyés par la ministre de la santé et des solidarités Madame X et son cabinet, au cours de la période de préfiguration du HDH, du 1er janvier 2018 au 31 décembre 2019 ;

2) les correspondances (courriers, courriels ou autres) échangés, entre le 1er janvier 2018 et le 30 mai 2022, entre la ministre de la santé et des solidarités Madame X et son cabinet, puis le ministre de la santé et des solidarités Monsieur X et son cabinet, d'une part, et d'autre part :

a) tout employé ou représentant de X ;

b) tout employé ou représentant d'X, de X ou de leur société commune X ;

c) Madame X.

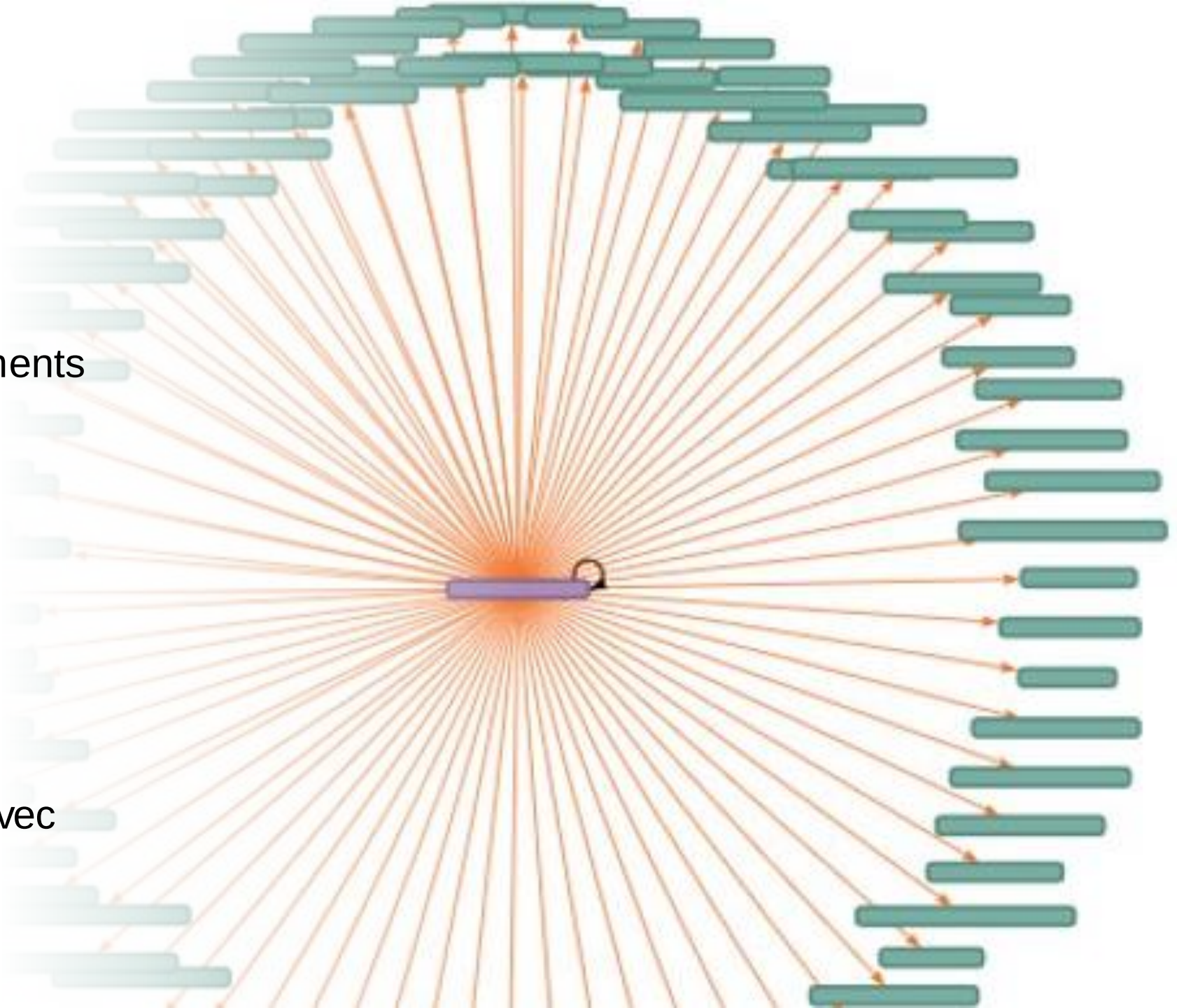
IA mais pas que...

Comprendre les comportements des utilisateurs

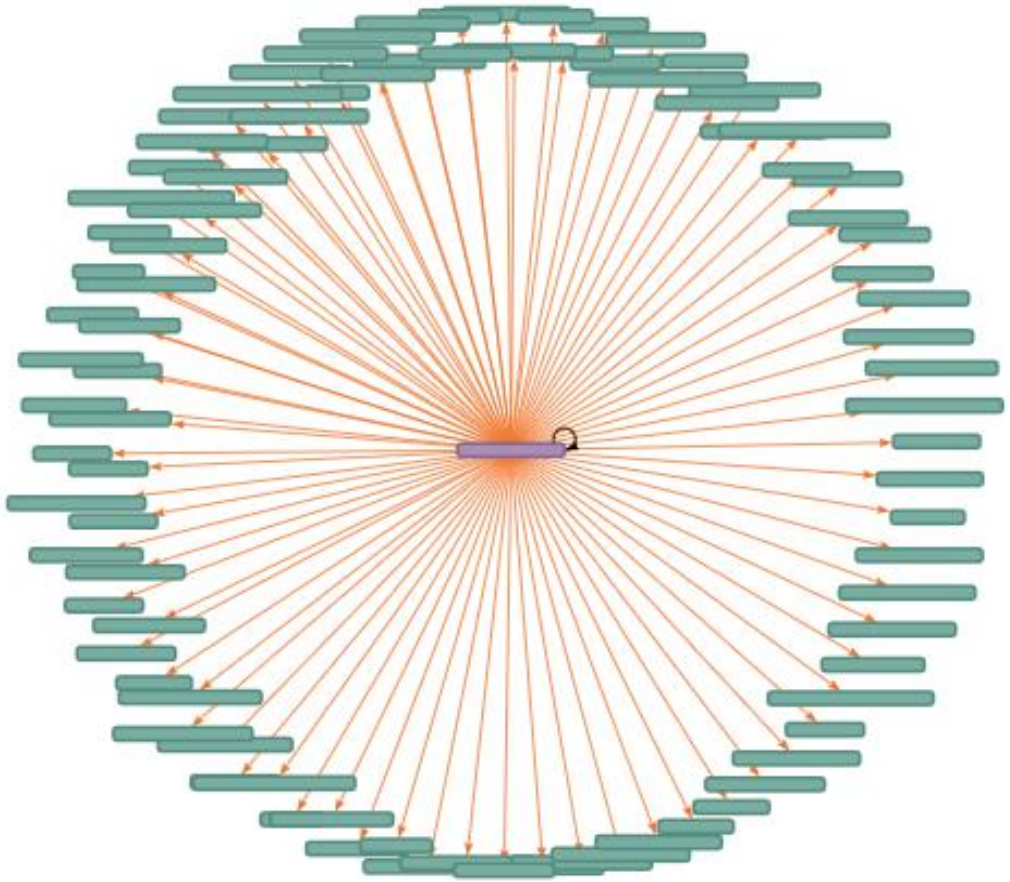
- Projet de recherche Comprendre les administrateurs et leurs relations aux méls (CARAméls)

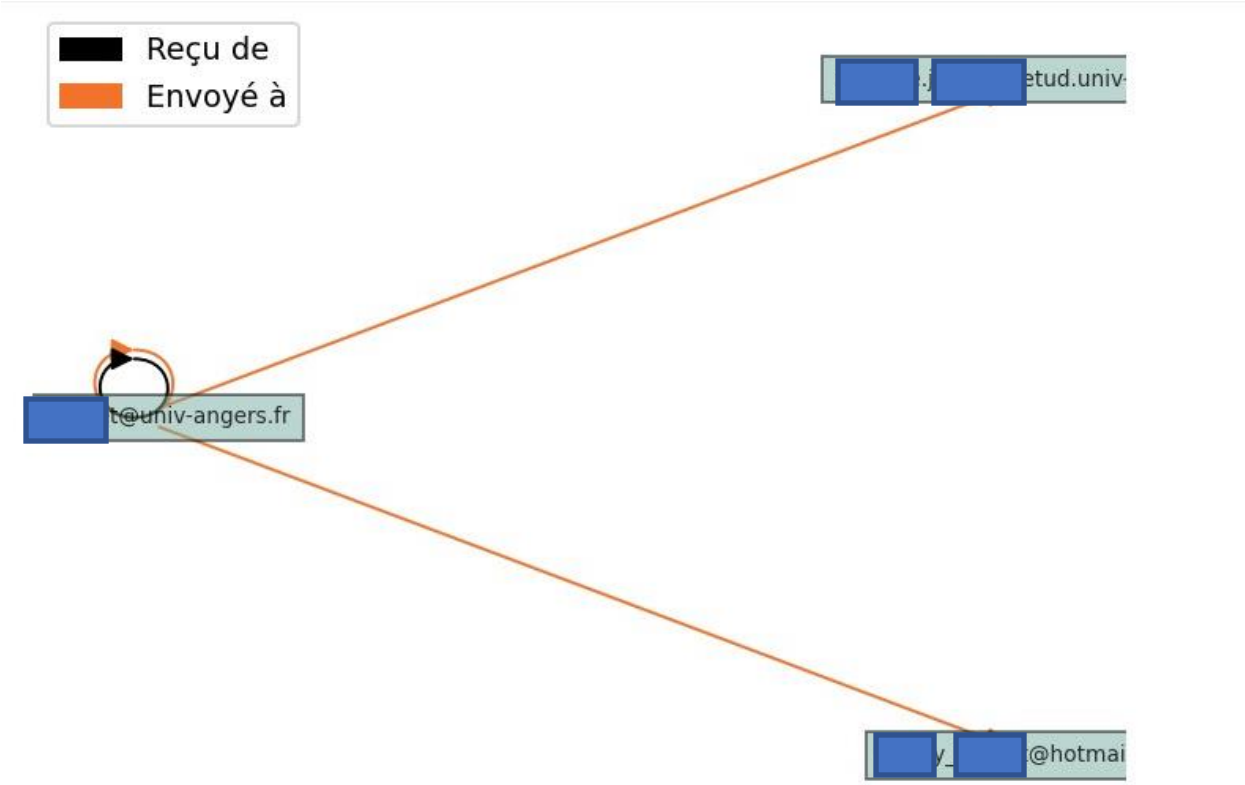
université angers

Relier les comportements avec des indicateurs et des visualisations



Exemple : autoarchivage et tranfert entre SI



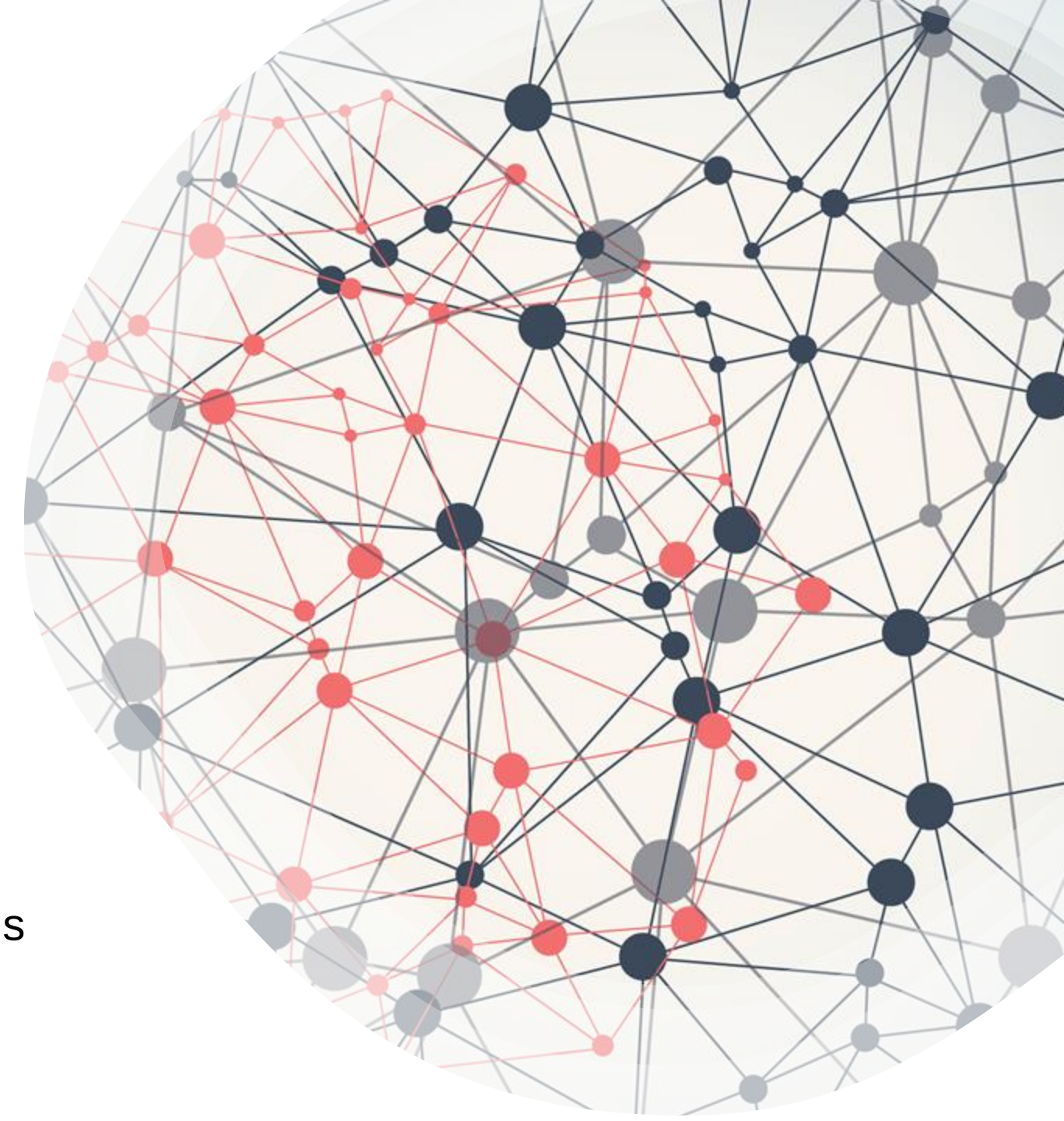


IA mais pas que...

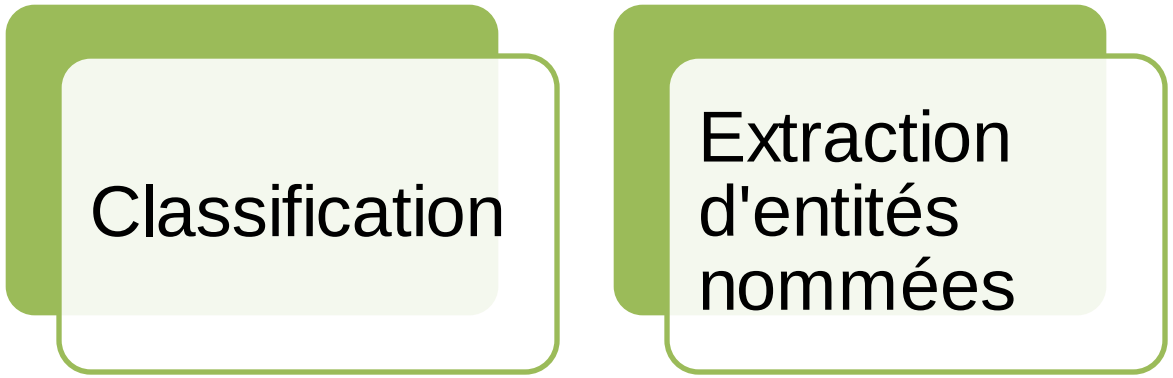
Améliorer l'évaluation globale

- Sortir de l'approche Capstone (sélection les courriels en fonction du travail et/ou de la fonction du propriétaire du compte de messagerie)

- Tenir compte de la sociologie des organisations et des transformations des rapports sociaux et des organisations bureaucratiques



Au-delà des messageries : alimenter des salles de lecture en ligne à la demande

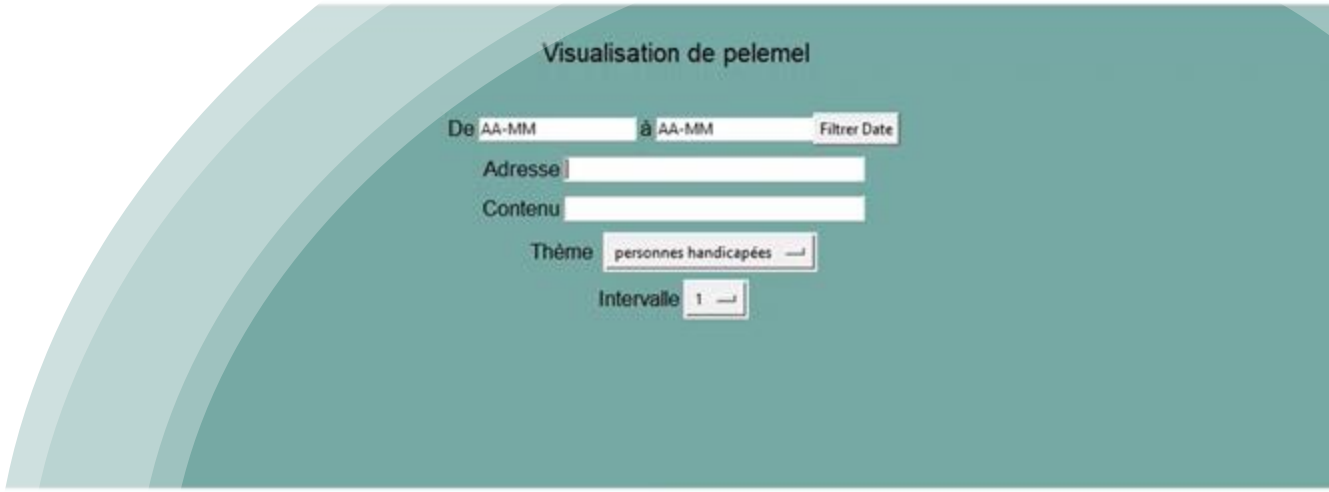


Créer des points d'entrée transversaux en exploitant mieux les instruments de recherche existants et en les complétant par les contenus informationnels

Délivrer des DIP pertinents à partir d'une pré-sélection à la demande ou pour des portails thématiques

Sortir de l'illusion de la mise à disposition immédiate de tout, tout le temps, en proposant des points d'entrée de pré-sélection

Proposer de nouvelles manières d'exploiter les archives



- Fournir des éléments de data-visualisation ou utilisables pour produire des data-visualisation

Conclusions

Limites et obstacles

- ✓ La taille du corpus : à tester sur des volumes plus importants. Option BERT et sa variante française, CamenBERT pourraient être appropriées ?
- ✓ Le contexte spécialisé : l'expérience devrait également être transposée aux autres types de contexte, spécialisé et non spécialisé.
- ✓ Prototype : l'ergonomie doit être améliorée.
- ✓ Un investissement à long terme dont les bénéfices ne sont pas immédiatement perceptibles. Les premiers lots de messages nécessitent beaucoup d'attention et de temps pour corriger, valider et améliorer la classification.

Innovations et bénéfices

TALN en français

Nous pensons avoir validé la preuve de concept : méthodes symboliques, patrons, messages + pièces jointes

De nombreux éléments peuvent être reproduits dans un autre environnement (résultats généralisables)

Les outils peuvent être utilisés pour enrichir la réflexion des archivistes sur l'évaluation des boîtes et leur tri interne

L'expertise archivistique reste au centre : outils d'aide à la décision

Au-delà des messageries, ce projet donne à apercevoir des voies complémentaires d'accès aux archives

Solutions ?

- ✓ Développement d'expertises croisées intermédiaires
- ✓ Une acculturation nécessaire, un portage par le réseau.
- ✓ Développement de projets de recherche : corpus inaccessibles, dérogations inadaptées, financements

benedicte.grailles@univ-angers.fr

touria.aitelmekki@univ-angers.fr