Sélectionner, épurer, décrire. L'apport de l'IA dans le traitement archivistique des messages électroniques

Touria Ait El Mekki, Bénédicte Grailles

CULTURE ET RECHERCHE

MINISTÈRE DE LA CULTURE
Liberté
Égalité
Fraternité

N° 147 AUTOMNE-HIVER 2024

Recherche et intelligence artificielle

Recherche et intel

ligence artificielle

Dossier coordonné par

CATHERINE GRAINDORGE

Rédactrice en chef, Délégation générale à la transmission, aux territoires et à la démocratie culturelle, Sous-direction des formations et de la recherche, Bureau de la recherche

En couverture

La ville à l'heure du changement climatique – surréalisme migratoire

Sélectionner, épurer, décrire.

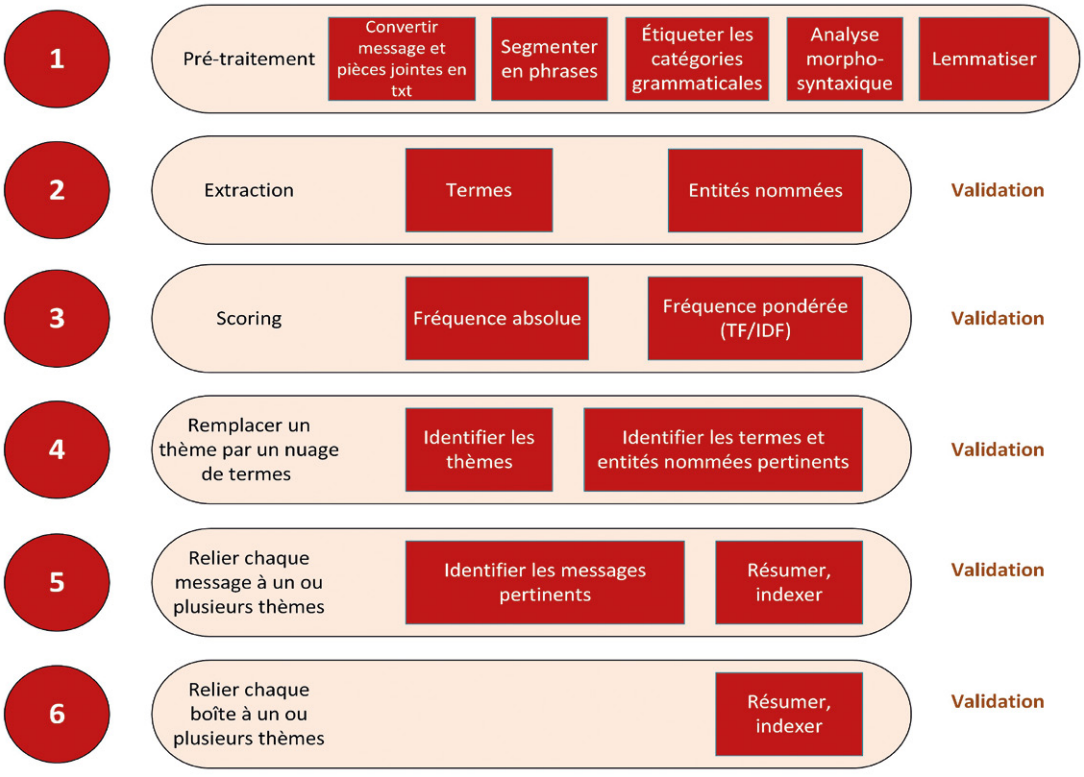
L'apport de l'IA dans le traitement archivistique des messageries électroniques

Le programme Pêle-mél[1], lauréat de l'appel à projets Services numériques innovants du ministère de la Culture en 2020, a permis de tester, sur un corpus de boîtes méls provenant du ministère de la Santé, des approches de traitement automatique de la langue naturelle reposant sur de l'extraction de termes, des relations sémantiques et des techniques d'apprentissage artificiel. L'objectif était d'explorer l'intérêt de la classification pour choisir les messageries à conserver définitivement et effectuer des tris internes, en associant, dans la démarche comme dans les résultats, l'expertise de l'archiviste. Il a donné lieu à l'élaboration de deux prototypes.

Une question de terminologie

Le projet et ses différentes étapes ont mobilisé de la terminologie computationnelle *via* de l'étiquetage grammatical, de l'analyse morpho-syntaxique, de l'extraction de termes et d'entités nommées et l'évaluation d'un score, et se sont appuyés sur de l'apprentissage artificiel pour la classification et la catégorisation des contenus textuels.

Les messages et leurs pièces jointes ont été dans un premier temps convertis en .txt, puis les catégories grammaticales (nom, verbe, adjectif, adverbe, déterminant, etc.), les fonctions et les phrases ont été



Phases du traitement des messageries proposé par le programme Pêle-mél (seuls les résumés automatiques n'ont pas été testés).

TOURIA AÏT EL MEKKI

Maîtresse de conférences en informatique, Laboratoire d'étude et de recherche en informatique (LERIA), Université d'Angers

BÉNÉDICTE GRAILLES

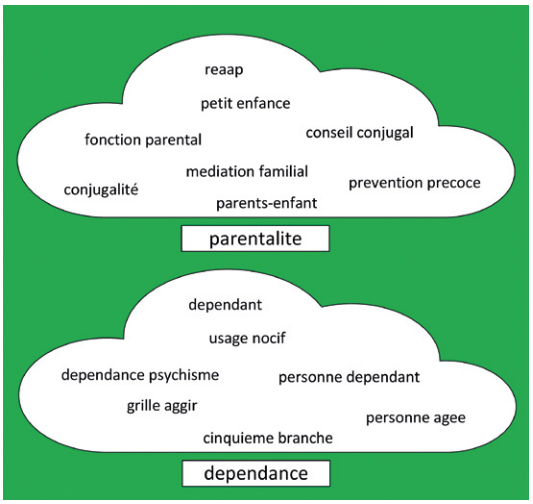
Maîtresse de conférences en archivistique, Université d'Angers, UMR CNRS 9016 Temps, mondes, société (TEMOS)

1. Porté par Bénédicte Grailles, Touria Aït el Mekki (Université d'Angers, TEMOS et LERIA), avec Chafik Akmouche, Tsanta Randriatsitohaina, Taïmane Zerez. Partenaires : mission Archives des ministères sociaux (Anne Lambert, Chloé Moser) ; École nationale des chartes (Édouard Vasseur).

identifiées et l'ensemble a été lemmatisé, c'est-à-dire qu'un verbe a été réduit à son infinitif, un substantif à son singulier, un adjectif à son masculin singulier. Le score a été établi soit par la fréquence, soit, plus sûrement, par une méthode de pondération dite TF-IDF (de l'anglais *Term frequency-Inverse document frequency*) qui permet de mesurer la quantité et la qualité d'un terme dans un contenu, une méthode très utilisée en fouille de textes[2] pour identifier les mots-clés pertinents.

Les termes – des unités lexicales d'un ou plusieurs mots représentant un concept (par exemple, « ministère de la Culture ») – et les entités nommées – une personne, un organisme, un lieu, un événement, un sigle – ont été extraits. Pour l'analyse fine du texte, nous avons utilisé différents outils (TreeTagger, Yatea, Termsuite, Spacy) qui contiennent des algorithmes « d'apprentissage profond[3] » capables d'extraire des informations en utilisant l'intelligence artificielle.

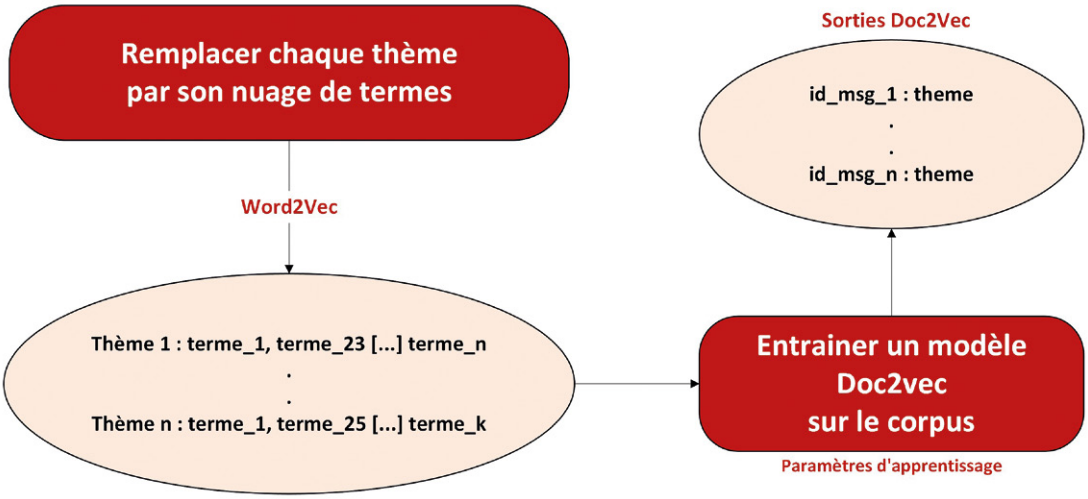
Les relations sémantiques, préalables à la classification, ont ensuite été établies par une approche symbolique non supervisée, en deux temps : grâce à Word2Vec, une méthode de plongement lexical utilisant un réseau de neurones artificiels à deux couches



Réseaux terminologiques représentant respectivement les thèmes *parentalité* et *dépendance* établis grâce à Word2Vec par comparaison des vecteurs et sélection des vecteurs les plus proches en distance avec une profondeur de recherche de 2. Les thèmes sont proposés par l'archiviste. Les termes ont été lemmatisés. Reaap désigne un dispositif particulier : les réseaux d'écoute, d'appui et d'accompagnement des parents. La grille Aggir (Autonomie gérontologique et groupe Iso ressources) est utilisée pour évaluer la perte d'autonomie.

« L'objectif final était d'obtenir des agrégats (*clusters*) de messages incluant leur nommage et leur(s) pièce(s) jointe(s), afin de comprendre et de suivre des fils de conversation. »

entraînées pour reconstruire le contexte linguistique des mots – les mots sont représentés en fonction de leur contexte par capture des similarités sémantiques et syntaxiques –, puis avec Doc2Vec[4], une méthode prédictive de plongement de documents qui permet également de prendre en compte le contexte dans lequel un mot a été trouvé. Chaque mot *via* Word2Vec, puis chaque message *via* Doc2Vec est représenté par un vecteur de nombres réels. Les mots puis les messages utilisés dans des contextes similaires, supposés avoir des significations proches, sont représentés dans l'espace vectoriel par des vecteurs proches. L'objectif final était d'obtenir des agrégats (*clusters*) de messages incluant leur nommage et leur(s) pièce(s) jointe(s), afin de comprendre et de suivre des fils de conversation. En l'absence de corpus pré-étiquetés, un modèle générique déjà pré-entraîné sur de larges corpus en français a été utilisé. Nous avons pu ainsi relier des thèmes à leur nuage de termes et d'entités nommées puis classifier les messages en les reliant à un ou plusieurs thèmes grâce à leur nuage. La structuration a été guidée par l'archiviste qui a proposé une liste de thèmes. La même méthode pourrait être élargie à la classification de boîtes complètes au sein d'un système complexe et à l'échelle d'une organisation.



Méthode de classification des messages. Les messages peuvent être associés à 0, 1 ou *n* thèmes.

2. Voir l'article : Didier Thébault, « Faites parler vos données ! Le *text mining* (la fouille de textes) en documentation », *Culture et Recherche*, n° 144, printemps-été 2023, p. 85-87.

3. « L'apprentissage profond » (*deep learning*) est une méthode d'intelligence artificielle qui apprend aux ordinateurs à traiter les données d'une manière inspirée du cerveau humain avec ses réseaux de neurones. Les modèles « d'apprentissage profond » peuvent ainsi reconnaître des modèles complexes dans des images, du texte, des sons et d'autres données pour produire des informations et des prévisions précises.

4. Nous avons également testé le *transformer* BERT, un modèle « d'apprentissage profond », dans sa version adaptée au français camenBERT. Pour le volume de données dont nous disposions, Doc2Vec s'est révélé plus efficace et plus rapide.

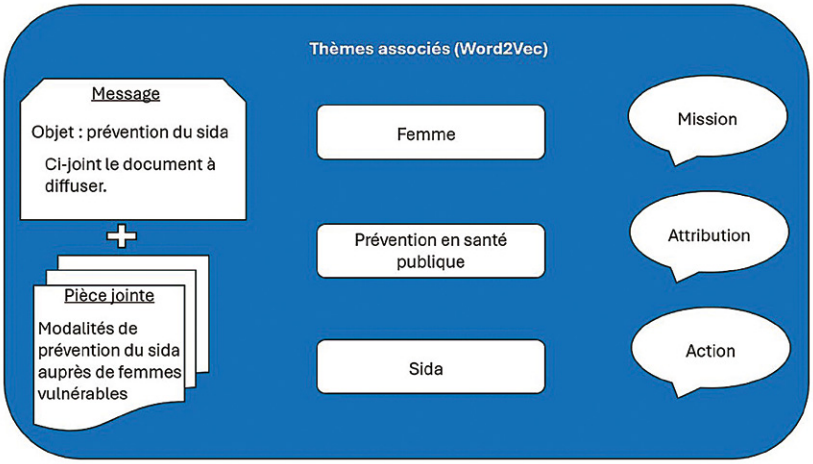
Une compréhension fine et augmentée

Ces différentes techniques permettent d'améliorer la compréhension des messageries à l'échelle unitaire mais surtout à celle d'une organisation. Elles peuvent concourir à mieux les sélectionner, en fonction d'autres critères que la simple fonction occupée par son ou sa titulaire (approche dite *Capstone* validée par le programme Vitam[5]), comme par exemple la diversité des sujets de conversation, la place au sein du réseau interne, la fréquence des relations externes, son originalité ou au contraire sa représentativité. Elles peuvent aider à effectuer des tris à l'intérieur d'une boîte, entre documents privés et institutionnels, ou en repérant les messages jugés importants par le ou la titulaire de la boîte. Elles peuvent enfin contribuer à améliorer la description comme l'accès, grâce à la construction d'une ontologie et à l'identification des entités nommées contextualisant les événements et les correspondants. La génération de résumés automatiques sur un thème donné, non testée dans le projet, serait un complément intéressant.

Pour être parfaitement concluante, l'approche informatique et linguistique doit cependant être combinée avec une analyse relevant de la gestion personnelle de l'information, car les comportements restent très individualisés, et mobilisant la théorie des réseaux. Cette stratégie a été validée dans le cadre du projet CARAméls[6] qui a permis de détecter des profils de messagerie différents, présentant des valeurs archivistiques différenciées.

Chacun son rôle

Si l'IA ouvre des horizons, elle ne fait que proposer des outils de compréhension et de visualisation nouveaux : l'impulsion comme la décision restent du côté du service d'archives. Dans notre démarche, nous avons sollicité l'expertise archivistique pour guider le choix des thèmes qui ont servi à la constitution des agrégats ou nuages de termes puis de messages. C'est donc bien l'archiviste qui a orienté la création de l'ontologie en s'appuyant sur ses méthodes habituelles d'analyse des flux documentaires, à savoir le triptyque missions, attributions et actions. De même, le prototype lui permet d'intervenir à presque toutes les étapes en choisissant des validations automatisées et/ou manuelles. Il n'en reste pas moins que, grâce à l'IA, il est possible d'appréhender et de catégoriser de gros volumes avec un degré de précision inconnu jusqu'à présent. L'archiviste dispose de moyens d'exploration pour lesquels il n'est pas, pour le moment, outillé théoriquement. Par exemple, la théorie des valeurs – primaires et secondaires – se révèle insuffisante pour confirmer des choix. Concernant les réseaux de messageries, détecter des communautés différentes au sein d'une même organisation est désormais réalisable. La question devient alors de choisir des boîtes méls pertinentes pour chacune de ces communautés, soit parce qu'elles sont centrales, soit parce qu'elles articulent la circulation entre deux communautés, soit parce qu'elles sont des nœuds de transmission internes. Cette sélection ne pourra se faire qu'en fonction d'une politique d'archivage validée, en croisant les informations avec des méthodes issues des sciences sociales, et nécessitera plus de temps et d'énergie que la simple identification par fonction.



Exemple de résultat de la classification d'un message à partir des réseaux terminologiques précédemment structurés. Le message ici proposé est relié à trois thèmes différents qui correspondent à des missions, attributions et/ou actions du producteur d'archives.

Les potentialités ouvertes par l'intelligence artificielle et plus spécifiquement par le traitement automatique de la langue sont réelles. Elles promettent un gain qualitatif évident. Mais elles ne sont rien sans une réflexion archivistique approfondie et sans une analyse des comportements des utilisateurs et utilisatrices. Si l'IA autorise des actions jusque-là inenvisageables étant donné les volumes et la complexité des échanges, elle replace l'archiviste et ses compétences métier au centre du dispositif et, paradoxalement, complexifie ses décisions en lui donnant accès à de nouvelles variables. ■

5. Pour « Valeurs immatérielles transmises aux archives pour mémoire ». Ce programme interministériel français d'archivage électronique est porté par trois ministères, Europe et Affaires étrangères, Culture, Armées, tous trois responsables de la conservation des archives de l'État.

6. « Comprendre les administrateurs et leurs relations à leurs méls », financement par la communauté urbaine Angers Loire Métropole. Projet porté par Patrice Marcilloux et Bénédicte Grailles (Université d'Angers, TEMOS), avec Edgar Lejeune.

Bibliographie

Touria Aït El Mekki, Bénédicte Grailles et Tsanta Randriatsitohaina, « Création d'une base de connaissances à partir de messageries spécialisées pour améliorer l'exploitation et l'archivage des méls », *JADT 2022 Proceedings of the 16th International Conference on statistical analysis of textual data*, Vadistat press, Editzioni Erranti, 2022, p. 52-59.

Bénédicte Grailles, Touria Aït El Mekki et Édouard Vasseur, « Improving the archiving and contextualization of electronic messaging in French », *International Conference on Digital Preservation – iPress 2022 Proceedings iPres 2022 Glasgow 12–16 September 2022*, 2022, p. 374.

Bénédicte Grailles et Touria Aït El Mekki, *Pêle-Mél. Plateforme d'exploration, de livraison et d'évaluation des méls. Rapport de recherche*, Université d'Angers – TEMOS, 2022, 64 p. (hal-04647186).

Bénédicte Grailles, *Pêle-mél. Plateforme d'exploration, de livraison et d'évaluation des méls. Rapport d'évaluation des usages*, Université d'Angers – TEMOS, 2023, 29 p. (hal-04647179).