Extraction de thèmes d'un corpus de demandes de support pour un logiciel de relation citoyen

Mokhtar Boumedyen Billami, Christophe Bortolaso, Mustapha Derras

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Extraction de thèmes d'un corpus de demandes de support pour un logiciel de relation citoyen

Mokhtar Boumedyen Billami Christophe Bortolaso Mustapha Derras
Berger-Levrault, 64 rue Jean Rostand, 31670 Labège
mb.billami@berger-levrault.com,
christophe.bortolaso@berger-levrault.com,
mustapha.derras@berger-levrault.com

RÉSUMÉ

Nous nous intéressons dans cet article à l'extraction de thèmes (*topics*) à partir de commentaires textuels provenant des demandes de support de l'éditeur de logiciel Berger-Levrault. Le corpus de demandes analysé est celui d'un outil de gestion de la relation citoyen. Ce corpus n'est pas formaté et est peu structuré avec plusieurs locuteurs qui interviennent (le citoyen et un ou plusieurs techniciens support). Nous décrivons une étude expérimentale qui repose sur l'utilisation de deux systèmes. Le premier système applique une LDA (Allocation Dirichlet Latente), tandis que le second combine l'application d'une LDA avec l'algorithme k-Moyennes (*k-Means*). Nous comparons nos résultats avec un échantillon de ce corpus, annoté par un expert du domaine. Nos résultats montrent que nous obtenons une classification de meilleure qualité comparable avec celle effectuée manuellement par un expert en utilisant une combinaison LDA/k-Moyennes.

ABSTRACT

Topic extraction from a corpus of support requests for citizen relations software.
In this paper, we are interested in topic modeling from textual comments provided by support requests from Berger-Levrault software editor. The selected corpus relates to a citizen relationship management platform. This corpus is not formatted and not well-structured involving multiple speakers (the citizen and one or more technicians). We describe an experimental study based on the use of two systems. The first system applies an LDA (Latent Dirichlet Allocation), while the second combines the application of an LDA with the k-Means algorithm. We compare our results with a datatest derived from this corpus, annotated by an expert in the field. Our results show that we obtain a better classification like the one carried out manually by an expert using a combination of LDA/k-Means.

MOTS-CLÉS : Modélisation de thèmes, Allocation Dirichlet Latente, k-Moyennes.
KEYWORDS: Topic Modeling, Latent Dirichlet Allocation, k-Means.

1 Introduction

La modélisation et l'identification du thème auquel appartiennent les documents d'une collection donnée de textes sont essentielles pour de nombreuses applications. Par exemple, la recherche d'information (Yi & Allan, 2009), les systèmes de recommandation (Al-Ghossein et al., 2018), la classification de textes (Guha Neogi et al., 2019), les sciences cognitives (Pirnay-Dummer & Walter, 2009) et l'analyse de sentiments dans les réseaux sociaux (Naskar et al., 2016), pour n'en

Volume 2 : Traitement Automatique des Langues Naturelles, pages 155–163.
HAL : hal-02784763.

citer que quelques-unes. Dans ce contexte, un topic est un groupe de mots clés qui est considéré intuitivement comme représentant un thème sémantique latent présent dans un document. L'extraction des thèmes consiste à effectuer une analyse distributionnelle sur un corpus de données pour en mesurer les probabilités de distribution des thèmes sur les termes du vocabulaire utilisé. La modélisation des thèmes (en anglais *topic modeling*) est un processus sémantique qui applique un regroupement de niveau supérieur sur les mots clés et repose sur des modèles statistiques qui capturent la probabilité d'apparition des mots sémantiquement liés dans un contexte défini. Par exemple, pour capturer l'idée que *demande* et *inscription* se rapportent au même thème, tandis que *demande* et *FAQ* (i.e. Foire Aux Questions) se rapportent à un thème différent.

Plusieurs travaux se sont concentrés sur la modélisation et l'identification du thème auquel appartiennent les documents d'un corpus de données (Lin, 1995 ; Medelyan et al., 2008 ; Varga et al., 2014 ; Venkatesaramani et al., 2019). Ces approches reposent sur l'analyse de termes (ou parfois d'entités nommées), extraits à partir des textes, sur lesquels un regroupement peut être effectué afin de les lier par thème (par exemple, dans notre cas d'étude, *inscription, gestion de doublons* et *renumérotation*). Par la suite, un étiquetage du thème (Sorodoc et al., 2017 ; Gourru et al., 2018) peut être (facultativement) appliqué pour attribuer à ce groupe de termes un identifiant approprié se référant au thème en question (par exemple, *doublons d'inscription*).

Nous nous intéressons dans cet article à l'extraction de thèmes à partir de commentaires textuels évoqués par des clients et techniciens support pour le traitement des données du logiciel e.élections appartenant à l'éditeur Berger-Levrault. Il s'agit d'un logiciel de gestion des inscriptions électorales et de la préparation des scrutins. Plusieurs tâches sont nécessaires à la gestion des élections, par exemple, la modification des adresses des électeurs, l'optimisation d'édition des cartes électorales, la mise à jour du nom du maire, voire la gestion des transmissions dans un répertoire national induisant la gestion des doublons citée ci-dessus. Ce travail présente une étude expérimentale de méthodes d'extraction de thèmes à partir d'un corpus de demandes de support. Plus précisément, nous traitons la tâche d'identification des thèmes. Pour ce qui concerne l'étiquetage des thèmes, cette tâche n'est pas traitée dans ce travail mais constitue un complément essentiel.

Après avoir présenté dans la section 2 les travaux état-de-l'art d'extraction de thèmes, nous décrivons dans la section 3 le corpus de travail et d'évaluation de nos méthodes. Nous discutons du format de ces corpus et des conséquences sur le processus d'extraction de thèmes. Dans la section 4, nous présentons la méthodologie que nous avons suivie pour créer des modèles d'apprentissage permettant l'identification des thèmes. Enfin, les résultats d'expérimentation sont discutés dans la section 5 avant de terminer par une conclusion et quelques perspectives (section 6).

2 Travaux antérieurs d'extraction de thèmes

L'identification des thèmes peut être envisagée en appliquant directement les méthodes traditionnelles proposées dans la littérature telles que l'analyse sémantique latente–LSA (Deerwester et al., 1990), la LSA probabiliste–pLSA (Hofmann, 2001) ou l'allocation Dirichlet latente–LDA (Blei et al., 2003). Cependant, ces approches ont certaines limites. Premièrement, elles ne fonctionnent généralement que sur des mots individuels et non sur des expressions polylexicales, bien que des extensions aient été proposées pour tenir compte des termes multi-mots (Nokel & Loukachevitch, 2016; Blei & Lafferty, 2009). Deuxièmement, les thèmes sont considérés comme des variables latentes associées à une probabilité de générer des mots, et ne sont donc pas directement « étiquetés », ce qui les rend difficile à externaliser, même si des extensions pour étiqueter les thèmes sont disponibles (Gourru et al., 2018). Enfin, les mots sont difficiles à

interpréter sémantiquement et sont généralement considérés comme des références symboliques sur lesquelles une inférence statistique/probabiliste peut être appliquée. Toutefois, il existe une alternative qui permet de surmonter cette dernière limite lorsque le modèle LDA est appliqué. Il s'agit de l'outil de visualisation pyLDAvis (Sievert & Shirley, 2014). Cet outil permet de faciliter l'interprétation des thèmes à l'aide de la représentation graphique proposée. En effet, sélectionner un ou plusieurs termes sur le graphe permet non seulement de mieux expliquer le thème en question mais aussi de tirer des conclusions sur l'importance de certains termes pour certains thèmes. Par ailleurs, des approches ont émergé et proposent d'utiliser les informations structurées disponibles dans les bases de connaissances ontologiques pour améliorer l'identification des thèmes dans les textes (Jain & Pareek, 2010). Toutefois, ces ressources ne sont pas toujours disponibles principalement lorsque nous sommes confrontés au traitement d'un corpus de spécialité.

D'autres travaux se sont concentrés sur l'utilisation des méthodes traditionnelles en conjonction avec les informations extraites à partir de ressources standards, par exemple, la ressource DBpedia (Auer et al., 2007). Hulpuş et al. (2013) ont appliqué une LDA pour regrouper les mots en thèmes et ensuite, ils ont associé les mots groupés avec la ressource DBpedia pour étiqueter les thèmes. Todor et al. (2016), quant à eux, ont proposé un travail en sens inverse, c'est-à-dire, d'abord associer les termes avec la base DBpedia avant d'appliquer un algorithme d'identification de thèmes dans le but d'enrichir les textes avec des annotations en catégories de Wikipédia, hyperonymes, entités liées aux entités extraites, etc. Medelyan et al. (2008) ont proposé un travail qui consiste à indexer les thèmes sur tout le corpus Wikipédia par application d'un contrôleur de vocabulaire. Ce contrôleur repose sur l'utilisation des noms des articles de Wikipédia.

Pour l'étiquetage des thèmes, des approches à base de graphes de connaissances ont été proposées en utilisant des ressources lexico-sémantiques ou ontologiques (Hulpuş et al., 2013 ; Allahyari & Kochut, 2015). Ce type d'approches a été aussi proposé pour l'amélioration de l'identification des thèmes en utilisant des plongements d'entités (*entity embeddings*), par exemple (Yao et al., 2017 ; Brambilla & Altinel, 2019). Yao et al. (2017) ont entraîné les représentations d'entités au moyen d'une utilisation des réseaux sémantiques comme WordNet (Miller, 1995) et des bases de connaissances en ligne comme FreeBase (Bollacker et al., 2008).

3 Données de travail

Nous utilisons un corpus français de demandes de support à propos du logiciel e.élections. Nous appellerons ce corpus par la suite le corpus support Élections. Ce dernier a plusieurs particularités (par exemple, mots non structurés, formatage, ponctuation faible, etc.). Cela s'explique par le fait que pour une demande donnée, plusieurs échanges ont eu lieu soit par écrit, soit par téléphone. Toutefois, le contenu textuel que nous avons récupéré pour chaque demande regroupe la question du citoyen et les différentes réponses ayant été rapportées à sa demande par suite des échanges effectués avec les techniciens support. Par ailleurs, le vocabulaire utilisé par les citoyens est très varié et parfois mal structuré. Il dépasse la plupart du temps les termes (mots) utiles pour apporter une réponse à une demande donnée. L'utilisation de l'ensemble des mots du vocabulaire du corpus peut facilement engendrer un biais. L'aspect de réduction du vocabulaire est traité dans ce travail et est discuté dans la section qui suit.

Le corpus de travail ayant servi à l'entraînement de nos modèles d'apprentissage contient 94 478 échanges représentant 31 036 demandes différentes. Ce corpus a été collecté durant le dernier trimestre de l'année 2019. Une première analyse de ce corpus avec Spacy (Honnibal & Johnson, 2015), chaîne du traitement automatique de la langue, nous a montré la diversité du vocabulaire

utilisé : 12 412 mots pleins différents, c'est-à-dire, mots portant du sens (noms, adjectifs, adverbes et verbes). Le corpus d'évaluation, quant à lui, représente un échantillon de données collecté durant le troisième trimestre de l'année 2019. Cet échantillon décrit 535 demandes différentes dont près de 74 % des demandes (394 cas) ont été annotées avec un seul thème par un expert du domaine.

4 Méthodologie

Cette section décrit l'architecture de notre approche et les modèles d'apprentissage que nous avons développés. Cette approche répond au problème d'identification des thèmes en trois parties différentes, à savoir : (1) quels sont les outils et les ressources nécessaires pour prétraiter les données et réduire le vocabulaire utilisé dans le corpus support Élections ? (2) quel est le modèle de représentation vectorielle à utiliser pour décrire les demandes ? et (3) quel est le modèle d'apprentissage non supervisé à utiliser pour apprendre les thèmes discutés dans le corpus ? La figure 1 présente l'architecture de notre approche.

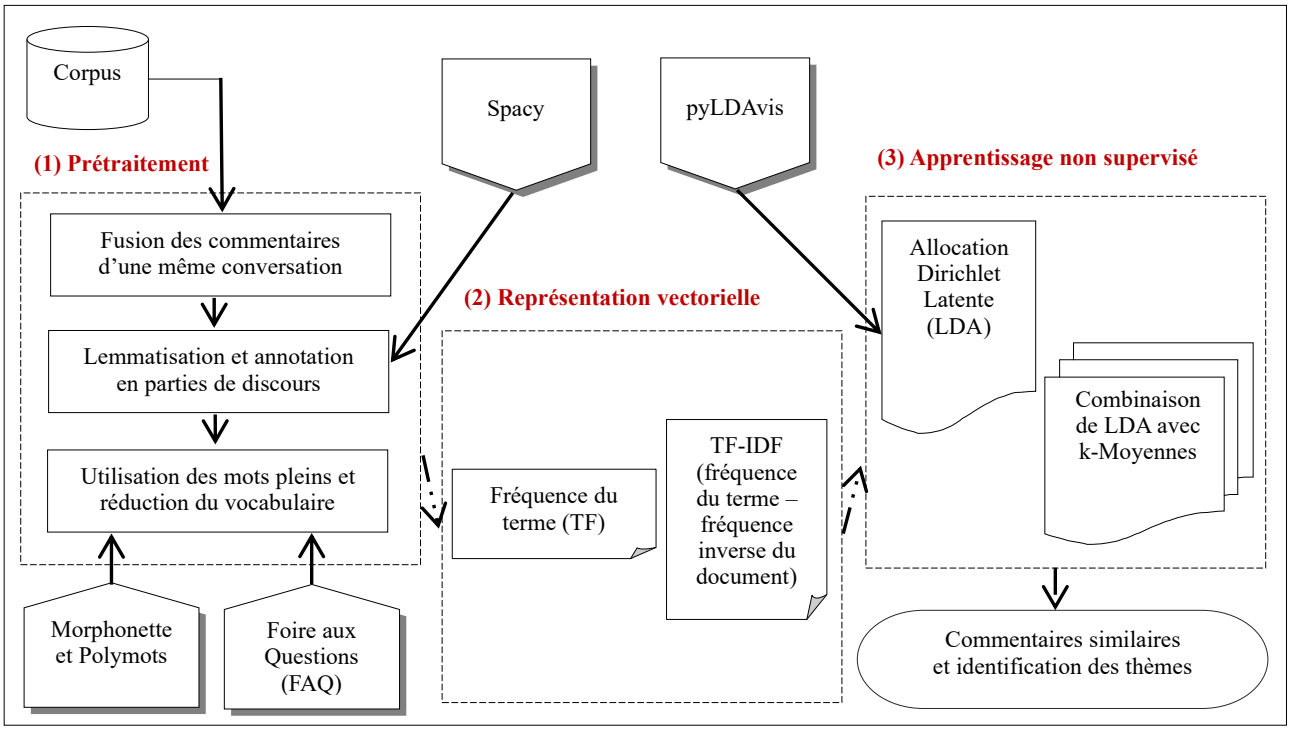

FIGURE 1: Approche d'identification des thèmes à partir des demandes de support

Pour la première partie, tout d'abord un travail de fusion des différents échanges liés à une même demande est effectué, en se référant au numéro de référence de la demande. Ensuite, nous utilisons Spacy (Honnibal & Johnson, 2015) pour extraire les formes lemmatisées des mots portant du sens.

L'étape suivante consiste à contrôler le vocabulaire exprimé dans les différents commentaires en utilisant un vocabulaire spécialisé et en effectuant des substitutions lexicales par l'utilisation de familles morphologiques. Plus précisément, nous avons utilisé un ensemble de documents FAQ liés aux corpus support Élections (22 documents avec une moyenne de 2 pages de contenu textuel, hors figures, par document). Ces documents FAQ permettent d'apporter des réponses aux questions des utilisateurs sur le produit e.élections. Un vocabulaire de 804 mots pleins a ainsi été obtenu.

Après avoir obtenu le vocabulaire lié aux documents FAQ, nous effectuons une substitution lexicale des mots du corpus par d'autres mots appartenant à une même famille morphologique. L'idée consiste à remplacer chaque mot plein du corpus par le mot le plus fréquent (dans le corpus). Pour cela, nous utilisons deux ressources, à savoir : Morphonette (Hathout, 2008) et Polymots (Gala & Rey, 2008). Morphonette est un réseau lexical morphologique construit à partir de la liste des mots du Trésor de la langue Française (Dendien & Pierrel, 2003), par utilisation des mesures de similarité morphologique. Ce réseau contient plus de 8 600 familles morphologiques. Polymots (Gala & Rey, 2008), quant à lui, représente une base de données lexicale permettant d'avoir des familles morpho-phonologiques. Cette base contient 8 000 mots communs en français et a été construite à partir de la continuité de sens et des formes phoniques comparables. Nous utilisons à la fois ces deux ressources puisqu'elles sont complémentaires. Par exemple, Polymots contient la famille morphologique *doublon*, *double*, *doublé*, *doubleur*, *doublage* et *doublement* alors que Morphonette ne propose pas de famille pour le mot *doublon*. D'un autre côté, Morphonette contient la famille *procuration*, *procurer*, *procureur* alors que Polymots ne propose pas de famille pour le mot *procuration*. La substitution que nous effectuons nous permet non seulement d'avoir une réduction supplémentaire du vocabulaire (nous passons de 804 à 513 mots) mais aussi d'ajouter une importance aux mots les plus fréquents. Par exemple, *doublon* est représenté avec 2 494 occurrences dans le corpus de travail contre 697 pour *double* et 3 pour *doublé*. La substitution lexicale donne ainsi un total d'occurrences de 3 194 pour *doublon*.

Pour la représentation des commentaires, nous utilisons Scikit-learn (Buitinck et al., 2013). Plus précisément, les modèles de représentation *CountVectorizer* et *TfidfVectorizer*. *CountVectorizer* permet de créer des représentations tenant compte seulement du nombre d'occurrences de mots dans un texte. *TfidfVectorizer*, quant à lui, permet de créer des représentations à base de TF-IDF, fréquence du terme-fréquence inverse du document (Jones, 1972). Cette technique donne des poids les plus élevés aux termes (i.e. mots) qui apparaissent plus fréquemment dans un texte par rapport aux autres textes du corpus. Pour l'expérimentation, nous avons utilisé ces deux techniques.

Pour le modèle d'apprentissage, nous utilisons l'allocation Dirichlet latente (Blei et al., 2003) et k-Moyennes (McQueen, 1967). L'allocation Dirichlet latente permet de représenter les textes comme des distributions de thèmes. En comparaison avec l'analyse sémantique latente probabiliste, la LDA suppose que les thèmes sont répartis entre les textes et les mots répartis entre les thèmes. Par ailleurs, k-Moyennes permet d'aboutir pour notre cas à une répartition des commentaires du corpus en *k* groupes (*clusters*). Les commentaires de chaque groupe partagent une certaine sémantique qui peut être identifiée en analysant la distribution des termes pour le groupe (i.e. fréquence des mots). Nous avons fait le choix de combiner LDA avec k-Moyennes pour voir le comportement de ce dernier sur des demandes évoquant au moins deux thèmes proches sémantiquement. Pour cette combinaison, l'algorithme LDA travaille sur la distribution des mots alors que l'algorithme k-Moyennes prend en entrée la distribution de probabilités des thèmes retournés par LDA.

5 Résultats d'expérimentation

Nous avons tout d'abord entrainé un modèle LDA sur les données du corpus de travail. Afin de choisir les meilleures valeurs pour les deux paramètres à fournir pour LDA, à savoir : le nombre de composants (thèmes) et la décomposition de l'apprentissage (*learning decay, ld*), nous avons appliqué la méthode *GridSearchCV* qui est implémentée dans Scikit-learn. Le nombre de thèmes a été varié entre 2 et 20 et nous avons pris une valeur pour la variable *ld* soit de 0,5, 0,7 ou 0,9. Les résultats obtenus ont montré que le paramétrage optimal pour notre modèle LDA serait de prendre 10 thèmes avec une valeur *ld* égale à 0,5, cela en utilisant les deux types de représentation des

commentaires, à savoir TF (fréquence du terme) et TF-IDF. Avec cette configuration, nous avons obtenu un rapport de vraisemblance de -1 204 185 et une perplexité de 123 par simple application du TF au type de représentation, ce qui est relativement bien pour une telle valeur de vraisemblance.

Nous avons comparé les résultats obtenus par l'utilisation des deux types de représentation des demandes de support. Cette comparaison nous a montré que le simple TF permet de mieux distinguer au moins quatre thèmes. La figure 2 montre les cartes retournées par pyLDAvis (Sievert & Shirley, 2014) sur les deux types de représentation. Nous remarquons que l'application de TF-IDF (voir la carte 'b)' de la figure 2) renvoie une projection très dense de la plupart des thèmes où la distance entre eux est très petite. Cela peut s'expliquer par le fait que la pondération TF-IDF accapare une très large part de la variance en raison de la présence des deux thèmes 9 et 10.

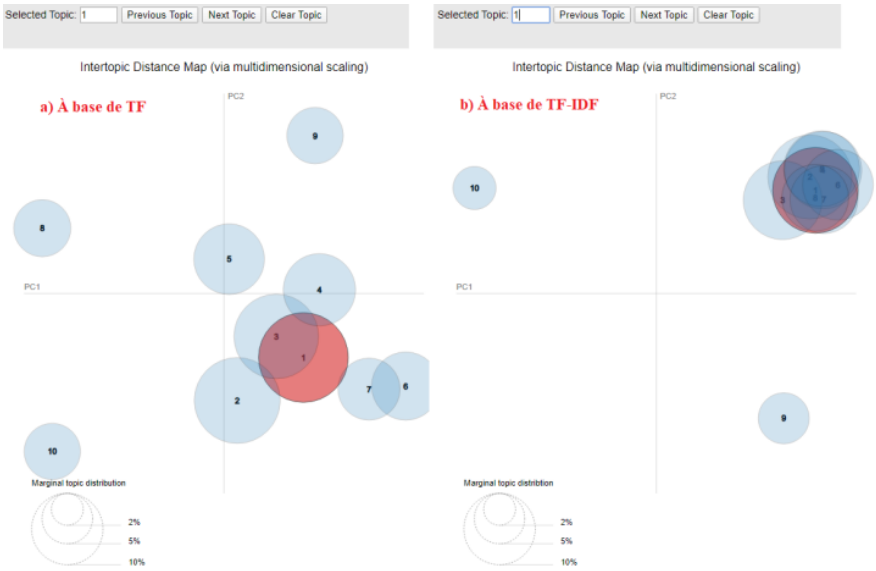

FIGURE 2: Visualisation des thèmes obtenus à l'aide de l'outil pyLDAvis

Pour ce qui suit, nous présentons nos résultats par application du type de représentation TF. Le tableau 1 décrit les mots les plus fréquents pour chaque thème du corpus de travail support Élections.

Thème	Liste de mots les plus fréquents
1	inscription ; notification ; traiter ; radiation ; pas ; date ; office ; avoir ; client ; compte.
2	logiciel ; mise ; jour ; rappeler ; nom ; synchronisation ; version ; avance ; usage ; pas.
3	demande ;insee ; transmettre ;pas; instruit ;message ; viser ;traiter ; notification ; erreur.
4	liste ; numéro ; avoir ; pas ; renumérotation ; apparaitre ; client ; fois ; carte ; électeur.
5	client ;connexion ; erreur ;pas ; message ;portail ; service ;période ;application ; besoin.
6	adresse ; doublon ; naissance ; changement ; ville ; commune ; gestion ; supprimer.
7	bureau ; vote ; synchronisation ; faire ; lancer ; pas ; base ; exister ; erreur ; sauvegarde.
8	faq ;client ;demande ;site ;lien ;recevoir ; cas ;disponible ;assistance ;réponse ; pouvoir.
9	carte ; retour ; procuration ; commission ; premium ; liste ; européen ; imprimer ;saisir.
10	tableau ; commander ; relatif ; espace ; solliciter ; récupérer ; transmettre ; condition.

TABLE 1 : Mots les plus fréquents de chaque thème par application du LDA

Nous remarquons que les thèmes numérotés en 5, 8, 9 et 10 se distinguent des autres puisque chacun est lié à une tâche de traitement différente, par exemple, 'problème de connexion et gestion des

messages d'erreurs' pour le thème 5 et 'demande des FAQ' pour le thème 8. Par ailleurs, certaines demandes n'ont pas pu être représentées (i.e. vecteurs nuls) en raison du texte très court utilisé par le citoyen (par exemple, *problème*) et l'absence de la réponse textuelle (fournie en général par téléphone pour ces cas). Il s'agit de 386 demandes (*Out-Of-Vocabulary, OOV*) sur 31 036.

Nous avons appliqué par la suite l'algorithme k-Moyennes en utilisant comme entrées les probabilités de distribution des thèmes renvoyées par LDA. Nous avons effectué une analyse de la silhouette en variant de la même façon le nombre de composants (thèmes dans notre cas) de 2 à 20. Le meilleur coefficient de la silhouette obtenu est 0,41 et revient à l'utilisation de 10 thèmes. Cela confirme d'une certaine manière le nombre de thèmes à prendre en compte pour classifier les demandes. La comparaison de nos deux systèmes sur la distribution des demandes du corpus de travail pour le thème le plus probable (majoritaire) nous montre des différences : 1 050 demandes évoquent un thème différent selon le système utilisé (3,38 % de cas). Le tableau 2 présente la distribution des demandes sur le thème majoritaire.

Modèle	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**	**OOV**
LDA	5010	4223	4077	4322	4060	2407	2013	1105	2071	1362	386
LDA/k-Means	4870	4429	3962	4380	3924	2639	2035	1009	2078	1324	386

TABLE 2 : Nombre de demandes associées au thème majoritaire

Nous avons évalué la qualité de nos deux systèmes sur 394 nouvelles demandes de support ayant été annotées par un expert du domaine. Sur l'ensemble de ces cas, 8 thèmes ont été évoqués, à savoir : (a) changement d'adresse, (b) demande d'inscription/radiation, (c) doublon d'inscription, (d) inscrits d'Office, (e) erreur du numéro d'émargement, (f) synchronisation, (g) suppression de la radiation et (h) divers. Nous avons remarqué plusieurs équivalences de thèmes entre ceux retournés par nos systèmes et ceux annotés par l'expert, voir même de nouveaux thèmes que l'expert a pris comme divers. Par exemple, l'édition des cartes électorales (thème 9). Il est à noter que l'expert était libre de choisir les thèmes jugés pertinents et n'avait pas l'obligation de spécifier exactement 8 thèmes. Par ailleurs et afin de mesurer la performance de nos systèmes, nous avons mis en correspondance les thèmes générés automatiquement et ceux proposés par l'expert. L'application du système LDA/k-Moyennes s'est vu meilleure que l'utilisation seule de LDA lorsque les demandes évoquent au moins deux thèmes importants. Nous avons obtenu un taux d'exactitude de 72 % pour LDA contre 77,69 % pour LDA/k-Moyennes.

6 Conclusion et perspectives

Dans cet article, nous avons présenté deux méthodes d'identification des thèmes. La première repose sur l'utilisation de LDA, tandis que la seconde est en deux temps : elle applique d'abord une LDA sur la distribution des mots pour ensuite appliquer l'algorithme k-Moyennes sur la distribution des thèmes retournés par LDA. Nous avons montré que l'utilisation de la seconde méthode à deux niveaux est plus performante lors du traitement des commentaires ayant au moins deux thèmes proches sémantiquement. Par ailleurs, en raison de la variété du vocabulaire exprimé par les locuteurs (principalement les citoyens), une réduction de la liste des mots du corpus était nécessaire en utilisant non seulement un vocabulaire spécialisé (documents FAQ liés au corpus support Élections) mais aussi certaines ressources de familles morphologiques. Deux perspectives s'ouvrent à ce travail : (1) une intégration des expressions polylexicales par l'utilisation des N-grammes (Nokel & Loukachevitch, 2016) est possible afin d'améliorer nos modèles; et (2) un étiquetage des thèmes (Gourru et al., 2018) est envisageable afin de faciliter leur intégration dans l'amélioration de plusieurs applications du traitement automatique du langage naturel.

Références

ALLAHYARI M. & KOCHUT K. (2015). Automatic Topic Labeling Using Ontology-Based Topic Models. *International Conference on Machine Learning and Applications (ICMLA)*, p. 259–264.

AL-GHOSSEIN M., MURENA P. A., ABDESSALEM T., BARRE A. & CORNUEJOLS A. (2018). Adaptive collaborative topic modeling for online recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, p. 338–346. DOI : 10.1145/3240323.3240363.

AUER S., BIZER C., KOBILAROV G., LEHMANN J., CYGANIAK R. & IVES Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *International Semantic Web Conference (ISWC-ASWC)*, p. 722–735.

BLEI D. M. & LAFFERTY J. D. (2009). Visualizing Topics with Multi-Word Expressions. *arXiv preprint* arXiv:0907.1013.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, **3**(Jan), p. 993–1022.

BOLLACKER K., EVANS C., PARITOSH P., STURGE T. & TAYLOR J. (2008). Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. In *Proceedings of the ACM SIGMOD international conference on Management of data*, p. 1247–1250.

BRAMBILLA M. & ALTINEL B. (2019). Improving Topic Modeling for Textual Content with Knowledge Graph Embeddings. *AAAI-MAKE Spring Symposium*.

BUITINCK L., LOUPPE G., BLONDEL M., PEDREGOSA F., MUELLER A., GRISEL O., NICULAE V., PRETTENHOFER P., GRAMFORT A., GROBLER J., LAYTON R., VANDERPLAS J., JOLY A., HOLT B. & VAROQUAUX G. (2013). API design for machine learning software: experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, p. 108–122.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**, p. 391–407.

DENDIEN J. & PIERREL J.-M. (2003). Le Trésor de la Langue Française Informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues*. Sous la direction de M. ZOCK & J. CAROLL, **44**(2), p. 11–37.

GALA N. & REY V. (2008). POLYMOTS: une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. *Actes de TALN 08: Traitement Automatique des Langues Naturelles*.

GOURRU A., VELCIN J., ROCHE M., GRAVIER C. & PONCELET P. (2018). United we stand: Using multiple strategies for topic labeling. *NLDB: Natural Language Processing and Information Systems*, p. 352–363. DOI : 10.1007/978-3-319-91947-8_37, HAL : lirmm-01910614.

HATHOUT N. (2008) Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the COLING Workshop Textgraphs-3*, p. 1–8.

HOFMANN T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, **42**(1), p. 177–196.

HONNIBAL M. & JOHNSON M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1373–1378. DOI : 10.18653/v1/D15-1162.

HULPUŞ I., HAYES C., KARNSTEDT M. & GREENE D. (2013). Unsupervised graph-based topic labelling using DBpedia. In *Web Search and Web Data Mining (WSDM)*, p. 465–474. ACM.

JAIN S. & PAREEK J. (2010). Automatic Topic(s) Identification from Learning material: An Ontological Approach. In *Computer Engineering and Applications (ICCEA)*, volume 2, p. 358–362. IEEE.

JONES K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation,* **28**, p. 11–21.

LIN C.-Y. (1995). Knowledge-based automatic topic identification. In *Annual meeting of the Association for Computational Linguistics (ACL)*, p. 308–310.

MACQUEEN J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume **1**, p. 281–297.

MEDELYAN O., WITTEN I. H. & MILNE D. (2008). Topic indexing with Wikipedia. In *2008 AAAI workshop "Wikipedia and Artificial Intelligence: An Evolving Synergy"*.

MILLER G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, **38**(11), p. 39–41.

NASKAR D., MOKADDEM S., REBOLLO M. & ONAINDIA E. (2016). Sentiment Analysis in Social Networks through Topic modeling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 46–53.

NOKEL M. & LOUKACHEVITCH N. (2016). Accounting ngrams and multi-word terms can improve topic models. In *Proceedings of the 12th Workshop on Multiword Expressions*, p. 44–49. DOI : 10.18653/v1/W16-1806.

GUHA NEOGI P. P., DAS A. K., GOSWAMI S. & MUSTAFI J. (2019). Topic Modeling for Text Classification. *Emerging Technology in Modelling and Graphics*, p. 395–407, Springer. DOI : 10.1007/978-981-13-7403-6_36.

PIRNAY-DUMMER P. & WALTER S. (2009). Bridging the world's knowledge to individual knowledge using latent semantic analysis and web ontologies to complement classical and new knowledge assessment technologies. *Technology, Instruction, Cognition & Learning*, **7**(1).

SIEVERT C. & SHIRLEY K. E. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, p. 63–70. DOI : 10.3115/v1/W14-3110.

SORODOC L., LAU J. H., ALETRAS N. & BALDWIN T. (2017). Multimodal Topic Labelling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, p. 701–706.

TODOR A., LUKASIEWICZ W., ATHAN T. & PASCHKE A. (2016). Enriching topic models with DBpedia. *On the Move to Meaningful Internet Systems*, p. 735–751, Springer.

VARGA A., CANO BASAVE A. E., ROWE M., CIRAVEGNA F. & HE Y. (2014). Linked knowledge sources for topic classification of microposts: A semantic graph-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, **26**, p. 36–57. DOI : 10.1016/j.websem.2014.04.001.

VENKATESARAMANI R., DOWNEY, D., MALIN, B. & VOROBEYCHIK, Y. (2019). A Semantic Cover Approach for Topic Modeling. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, p. 92–102. DOI : 10.18653/v1/S19-1011.

YAO L., ZHANG Y., WEI B., JIN Z., ZHANG R., ZHANG Y. & CHEN Q. (2017). Incorporating Knowledge Graph Embeddings into Topic Modeling. In *Thirty-First AAAI Conference on Artificial Intelligence*, p. 3119–3126.

YI X. & ALLAN J. (2009). A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Boughanem M., Berrut C., Mothe J. & Soule-Dupuy C. (eds) Advances in Information Retrieval. ECIR 2009*. Lecture Notes in Computer Science, vol 5478, Springer.