La méthode de modélisation thématique CFMf basée sur le clustering neuronal avec maximisation des traits : Comparaison avec LDA sur des études scientifiques

J.-C. Lamirel[1], F. Lareau[2], C. Malaterre[3]

[1] Université de Strasbourg, SYNALP-LORIA
[2] Université du Québec à Montréal, Computer Science Dept.
[3] Université du Québec à Montréal, Dept. of Philosophy & CIRST

lamirel@loria.fr, lareau.francis@courrier.uqam.ca, malaterre.christophe@uqam.ca

Résumé

L'amélioration des méthodes de modélisation thématique reste une préoccupation majeure pour l'analyse non supervisée des données textuelles. Nous proposons ici une approche de modélisation thématique basée sur le clustering neuronal et la maximisation des traits. Nous comparons ses performances à celles de LDA en appliquant les deux méthodes à un large corpus de référence d'articles de philosophie des sciences en texte intégral. Les résultats montrent des améliorations très significatives des mesures de performance quantitatives clés telles que la cohérence, ainsi que des résultats qualitatifs.

Mots-clés

Modélisation thématique, apprentissage non supervisé, LDA, clustering, maximisation des traits.

Abstract

The improvement of topic modeling methods remains a major concern for unsupervised analysis of textual data. We propose here a topic modeling approach based on neural clustering and feature maximization. We compare its performance to that of LDA by applying both methods to a large reference corpus of full-text philosophy of science articles. The results show very significant improvements in key quantitative performance measures such as coherence, as well as qualitative results.

Keywords

Topic modeling, unsupervised learning, LDA, clustering, feature maximization.

1 Introduction

En tant que résultats privilégiés de la recherche scientifique, les articles et leur contenu offrent des perspectives uniques pour comprendre la science. L'exploration par des méthodes informatiques du contenu textuel non structuré de ces articles peut être une solution efficace pour étudier des corpus de textes scientifiques trop volumineux pour une lecture manuelle. À cet égard, la modélisation thématique peut être utilisée pour inférer de manière fiable le contenu sémantique des publications, ce qui permet d'identifier les thèmes de recherche dominants dans des disciplines scientifiques spécifiques, y compris leur évolution dans le temps (par exemple, [1]; [2]; [3]). L'une de ces approches bien établies est le modèle Latent Dirichlet Allocation (LDA) [4] et ses variantes [5]. Des méthodes alternatives ont été récemment conçues, notamment certaines qui utilisent une combinaison de clustering neuronal et de maximisation des traits au moyen du contraste (« CFMc » pour neural Clustering and Feature Maximization with Contrast) [6]. Ces dernières ont montré des améliorations qualitatives prometteuses par rapport à LDA. Par contre, des tests approfondis sur un corpus de référence d'articles de philosophie des sciences en texte intégral (N=16917) qui avait déjà été analysé en détail au moyen d'un modèle thématique LDA [7] ont révélé des limites en termes d'interprétabilité des thèmes. Ces limites nous ont amenés à concevoir une nouvelle approche toujours basée sur le clustering neuronal avec maximisation des traits, mais qui s'appuie désormais sur la mesure F1 (« CFMf » pour neural Clustering and Feature Maximization with F1-measure). Dans la présente recherche en cours, nous décrivons l'approche CFMf. Pour évaluer sa performance par rapport à CFMc et LDA, nous appliquons ces méthodes au corpus de référence et comparons les modèles. D'abord quantitativement en termes de cohérence [8] pour plusieurs valeurs w du nombre de mots principaux et k du nombre de thèmes. Ensuite, nous évaluons également la performance qualitative en examinant l'interprétabilité des thèmes à $k = 25$ du point de vue de la connaissance experte. Les résultats montrent des améliorations très significatives apportées par CFMf, à la fois en termes de mesures de performance et d'évaluations qualitatives.

2 Méthodes

Alors que les approches CFMc avaient conduit à des résultats prometteurs lorsque testées sur un corpus d'articles de recherche chinois dans le domaine des « sciences de la science », nos expériences préliminaires sur le jeu de don-

nées plus complexe d'articles de philosophie des sciences - précédemment analysé avec la LDA [7] - ont mis en évidence trois limites. Premièrement, la représentation binaire des mots dans les documents, telle que mise en œuvre dans CFMc, semblait trop restrictive, encourageant ainsi l'utilisation des fréquences de mots. Deuxièmement, le contraste semblait ne pas être applicable dans les corpus dont les documents ne contenaient pas de thèmes discriminants clairement définis, d'où la nécessité de modéliser les documents comme intégrants plusieurs thèmes. Troisièmement, aucune comparaison quantitative avec la LDA n'avait été réalisée. La nouvelle approche que nous proposons ci-après (CFMf) répond à ces limitations. Elle s'appuie toujours sur le clustering neuronal et la maximisation des traits FMax, mais utilise désormais la mesure F1 au lieu du contraste. L'approche s'appuie également sur la représentation des documents en sac de mots (« BoW » pour Bag of Words), mais en exploitant l'information fréquentielle des mots. Les documents sont partitionnés en utilisant le clustering neuronal GNG [10]. Il s'agit d'une approche de type « winner-take-most » basée sur l'apprentissage Hebbien, moins sujette aux problèmes connus du clustering concernant la sensibilité aux valeurs aberrantes et à l'initialisation (comme c'est notamment le cas pour des méthodes classiques telles que k-means). Dans une étape ultérieure, les mots-clés représentatifs des clusters qui représenteront les thèmes sont extraits des documents associés à chaque cluster à l'aide de l'approche FMax (associée à la mesure F1), qui est un schéma générique de comparaison de données pouvant être utilisé comme une alternative aux métriques usuelles telles que le chi2, la métrique euclidienne ou la similarité cosinus lorsqu'il s'agit de traiter des données éparses et fortement multidimensionnelles, comme c'est le cas des données textuelles quand elles sont représentées en mode BoW. L'approche FMax offre des capacités de sélection et de pondération de variables sans nécessiter d'utiliser de paramètre [9] et s'est avérée très utile dans de nombreuses tâches d'exploration de données, y compris pour le plongement de mots et le plongement de graphes [11]. Dans le cas que nous traitons, les variables (c.àd. les traits) sont des mots et les données sont les documents associés à chaque cluster. FMax est basée ici sur l'estimation de la mesure F1 qui représente la moyenne harmonique (1) du rappel de trait, qui estime le pouvoir de discrimination d'un mot vis-à-vis d'un cluster ; et (2) de la prédominance de trait, qui évalue la capacité de généralisation du mot vis-à-vis de ce même cluster.

Considérons une partition C qui résulte d'une méthode de clustering appliquée à un ensemble de documents D représenté par un ensemble de mots F. Les mesures de rappel de traits, de prédominance de traits et la mesure F1 sont respectivement définies comme suit :

$$FR_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{c \in C} \Sigma_{d \in c} W_d^f} \tag{1}$$

$$FP_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{f' \in F_c, d \in c} W_d^{f'}} \tag{2}$$

avec

$$F1_c(f) = 2\left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)}\right) \tag{3}$$

où W_d^f représente le poids du mot f pour le document d et F_c représente l'ensemble des mots présents dans l'ensemble des documents associés au cluster c.

Pour réduire le bruit, nous éliminons au sein d'un cluster donné les mots qui répondent à au moins une des conditions suivantes : a) une mesure F1 inférieure à la moyenne des mesures F1 de ce même mot pour tous les clusters dans lesquels le mot est présent ou b) une mesure F1 inférieure à la moyenne des mesures F1 de tous les mots de tous les documents.

Ainsi, l'ensemble S_c des mots qui sont caractéristiques d'un cluster c issu d'une partition C est défini par :

$$S_c = \left\{ f \in F_c \mid F1_c(f) > \overline{F1}(f) \text{ et } F1_c(f) > \overline{F1}_D \right\} \tag{4}$$

avec

$$\overline{F1}(f) = \Sigma_{c' \in C} \frac{F1_{c'}(f)}{|C_{/f}|} \text{ et } \overline{F1}_D = \Sigma_{f \in F} \frac{\overline{F1}(f)}{|F|} \tag{5}$$

où $C_{/f}$ représente le sous-ensemble des clusters de C dans lequel le mot f est présent.

Il en résulte un profil de mesure F1 (sur un lexique réduit) pour chaque cluster, ce dernier étant alors considéré comme un thème. D'où la possibilité d'extraire les mots les plus importants sur la base de leur classement selon la mesure F1.
De plus amples détails sur les mesures ci-dessus mentionnées sont également donnés dans [9].

3 Protocole expérimental

Le corpus est constitué de tous les articles de recherche en texte intégral provenant de huit revues majeures de philosophie des sciences qui ont été rassemblées dans le cadre de l'étude de la philosophie des sciences [13] : le *British Journal for the Philosophy of Science*, le *European Journal for Philosophy of Science*, *Erkenntnis*, *International Studies in the Philosophy of Science*, le *Journal for General Philosophy of Science*, *Philosophy of Science*, *Studies in History and Philosophy of Science Part A* et *Synthese*. Il s'étend de 1930 à 2017 et comprend 16 917 documents. Le corpus a été nettoyé et prétraité de manière standard (les textes en langue étrangère ont été traduits mécaniquement en anglais). Seuls les noms, les verbes, les adverbes et les adjectifs ont été conservés après étiquetage POS et lemmatisation (TreeTagger package [12] avec les jeux d'étiquettes de Penn TreeBank [13]) et les mots apparaissant dans moins de 50 phrases du corpus ont été supprimés. Tous les documents ont ensuite été vectorisés, ce qui a permis d'obtenir une matrice termes-documents

avec fréquences de mots.

La matrice termes-documents a ensuite été soumise à CFMc, CFMf et LDA. Pour comparer quantitativement CFMf et CFMc, les deux méthodes ont été utilisées pour construire des modèles à k = 25. Ces modèles ont ensuite été comparés en termes de cohérence calculée avec différents nombres de mots principaux (de w = 5 à 100). L'objectif était ici d'évaluer quelle méthode donnait de meilleurs résultats pour des valeurs de w relativement petites (étant donné que les petits ensembles de mots principaux sont généralement plus faciles à interpréter à condition qu'ils soient bien formés). Dans une étape ultérieure, des modèles de thèmes ont été construits à la fois avec CFMf et LDA pour différents nombres de thèmes allant de k = 5 à 50 (par incréments de 5 de 5 à 20, et par incréments de 1 au-delà de 20). La modélisation LDA a été réalisée conformément à [1] et [4] par l'intermédiaire d'une API Python. Les modèles thématiques obtenus ont ensuite été comparés à l'aide de trois mesures permettant d'estimer la consistance ou la cohérence thématique. Tout d'abord C_{PMI}, (également appelée C_{UCI}) suivant [15] qui ont proposé d'évaluer la qualité des thèmes en termes de cohérence telle qu'elle est comprise par les lecteurs humains ; cette mesure compte la cooccurrence des mots dans une fenêtre glissante et calcule, pour chaque paire de mots, son PMI (information mutuelle ponctuelle) ; C_{PMI} est la somme (ou la moyenne arithmétique selon les implémentations) des valeurs PMI. Deuxièmement, C_{NPMI}, tel que proposé par [14] est une version améliorée de C_{PMI} utilisant l'information mutuelle ponctuelle normalisée (NPMI). Troisièmement, la mesure de cohérence souvent utilisée C_V proposée par [8] et mise en œuvre dans le package populaire Gensim en Python ; C_V compte les cooccurrences d'un certain nombre de mots principaux (généralement 10 à 20) dans une fenêtre glissante (généralement de taille 110) ; les cooccurrences sont utilisées pour calculer la NPMI entre les mots principaux, produisant des vecteurs pour chacun d'entre eux ; la moyenne arithmétique des similitudes cosinus entre chaque vecteur de mots principaux et la somme de tous les vecteurs de mots principaux est ensuite calculée. Les noms de thèmes de l'étude [7] ont été utilisés pour étiqueter les thèmes LDA, tandis que les thèmes CFMf ont été nommés à l'aide de leurs premiers mots et de la connaissance des experts, et les thèmes des deux modèles ont été comparés qualitativement. Pour comparer davantage les résultats des modèles LDA et CFMf, la distance Hellinger entre les 25 thèmes de chaque modèle (représentés sous forme de vecteurs de mots) a été calculée. Les thèmes d'un modèle ont ensuite été alignés sur ceux de l'autre.

4 Résultats

Les résultats de cohérence comparant les performances de CFMf et CFMc en fonction du nombre de mots principaux (pour un nombre donné de thèmes k=25) montrent que les modèles CFMf avec moins de mots principaux sont nettement plus performants que les modèles CFMc (figure 1A). Cela signifie que la mesure F1 donne des ensembles de mots principaux plus cohérents que le contraste. Compte tenu de l'objectif d'interprétabilité du sujet (basé sur un ensemble relativement restreint de mots principaux ordonnés), l'utilisation de la mesure F1 au lieu du contraste apporte une amélioration méthodologique significative, ce qui justifie l'utilisation de CFMf par rapport à CFMc dans des contextes similaires.

Lorsqu'il s'agit de comparer CFMf avec LDA, les résultats montrent que CFMf surpasse largement LDA en termes de mesures de cohérence (Fig. 1B, C, D). C'est le cas pour les trois mesures de cohérence que nous avons testées (C_V, C_{NMPI} et C_{PMI}), et pour une large gamme de modèles avec différents nombres de thèmes (de k = 5 à 50 thèmes). L'amélioration de la cohérence apportée par CFMf par rapport à LDA est très significative, puisqu'elle va d'environ 50 % pour C_V à plus de 200 % pour C_{PMI}. Dans tous les cas, la cohérence augmente considérablement de 5 à 20 thèmes, puis plus lentement de 20 à 40 thèmes, et approche un plateau au-delà de 40 thèmes. Un nombre idéal semble donc se situer autour de 30-35 thèmes, en fonction des objectifs du modèle thématique. L'aspect le plus significatif des résultats est la surperformance quantitative constante du modèle CFMf sur le modèle LDA en termes de cohérence. Les résultats de l'analyse qualitative effectuée sur les 10 premiers mots des thèmes du CFMf montrent un type de couverture thématique similaire à celui de la LDA. Rappelons que, pour des raisons de commodité, cette comparaison a été effectuée pour k = 25 thèmes, étant donné qu'un modèle LDA antérieur a été examiné en détail pour k = 25 [7]. A l'époque, k = 25 avait été choisi pour des raisons pragmatiques, afin d'avoir un modèle thématique avec un nombre de thèmes relativement faible. Or, comme nous venons de le voir, des valeurs de k plus élevées montrent des mesures de cohérence plus élevées, ce qui implique que k = 25 n'est pas optimal à cet égard. Néanmoins, les 25 thèmes résultant du modèle CFMf semblent facilement interprétables sur la base des 10 premiers mots et de la connaissance du domaine par les experts.

Les distances Hellinger entre ces thèmes et ceux du modèle LDA montrent une assez bonne correspondance des sujets, mais l'appariement est loin d'être parfait, ce qui montre que les thèmes des deux modèles ont encore des différences notables. Nous avons examiné et comparé manuellement les 10 premiers mots des thèmes des deux modèles. Des exemples sont présentés dans la table 1 : certains sujets ont des mots du top-10 très similaires, tandis que d'autres semblent avoir été en quelque sorte fusionnés ou divisés.

5 Discussion

Cette première expérience de comparaison quantitative et qualitative donne des résultats préliminaires intéressants. La méthode CFMf, qui est une extension de la méthode CFMc utilisant désormais la mesure F1 plutôt que le contraste, semble bien plus efficace que la méthode LDA au regard de trois mesures de performance ainsi qu'en termes d'interprétation (au moins en ce qui concerne les

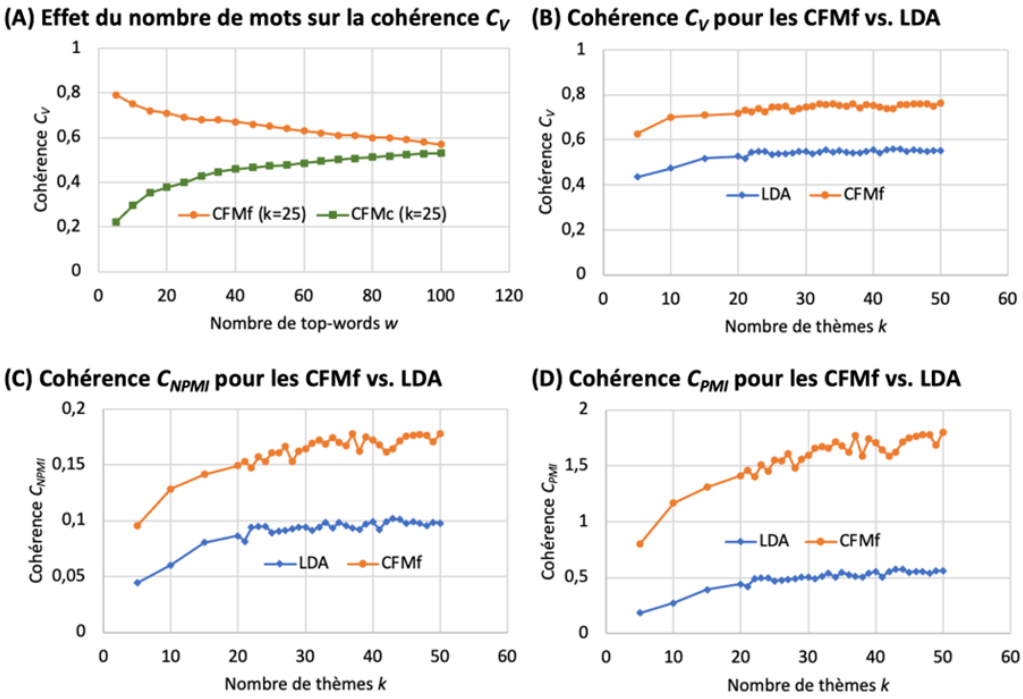

FIGURE 1 – Comparaisons entre modèles thématiques. (A) Cohérence C_V pour CFMf et CFMc en fonction du nombre de top-words w (pour k = 25 thèmes). (B, C, D) Cohérences C_V, C_{NPMI} et C_{PMI} pour les modèles CFMf et LDA en fonction du nombre de thèmes k (pour w = 10 top-words).

CFMf	*Top-10 words*	*LDA topics*	*Top-10 words*
Knowledge (11)	belief; epistemic; justified; justification; epistemically; doxastic; reliable; testimony; proposition; agent	Knowledge (21)	belief; knowledge; epistemic; believe; know; case; evidence; reason; justification; true
Neurosciences (20)	neural; brain; processing; cognitive; input; computational; neuron; cognition; mechanism; visual	Neurosciences (13)	system; information; process; cognitive; level; mechanism; state; representation; structure; function
Causation (7)	causation; causal; cause; counterfactual; intervention; causally; woodward; probabilistic; event; counterfactuals	Causation (19)	causal; cause; event; effect; causation; condition; case; variable; time; occur
		Explanation (16)	model; explanation; explain; account; explanatory; phenomenon; use; case; system; provide
Quantum mechanics (18)	quantum; measurement; particle; mechanic; observables; spin; operator; wave; probability; bell	Quantum (14)	time; state; space; quantum; system; theory; particle; physical; field; point
Relativity (23)	spacetime; relativity; inertial; metric; einstein; velocity; motion; coordinate; frame; tensor		

TABLE 1 – Comparaison des top-10 mots pour un échantillon de thèmes (CFMf et LDA).

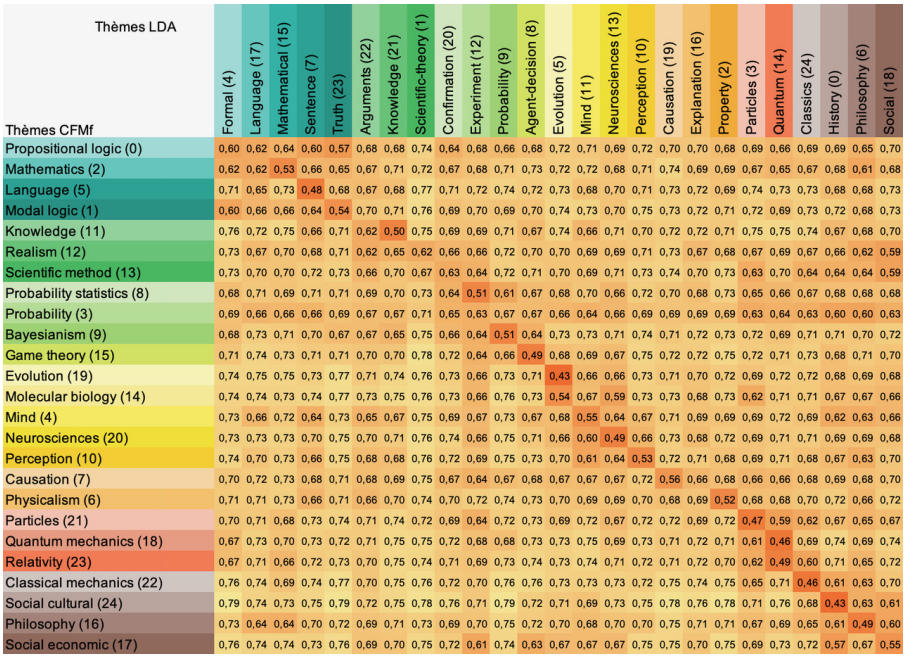

FIGURE 2 – Distances Hellinger entre thèmes issus du modèle LDA (rangée du haut) et ceux du modèle CFMf (colonne de gauche). Distances mesurées entre thèmes représentés comme des vecteurs de probabilités sur le lexique du corpus.

10 premiers mots du modèle $k = 25$). Il serait intéressant d'évaluer ces méthodes avec d'autres mesures de performance, notamment la perplexité [16], UMass [17], la KL-divergence symétrique [18], la densité [19] ou les statistiques bayésiennes [1]. Les résultats peuvent également être comparés à d'autres implémentations de LDA (par exemple, les approches LDA avec Bayes variationnel [20] au lieu de l'échantillonnage de Gibbs), et à des modèles alternatifs (par exemple, Structural Topic Models (STM) [21]). Le comportement de CFMf sur les documents (comparé à LDA) devrait également être exploré, afin d'évaluer les changements dans les distributions de probabilité des thèmes. CFMf pourrait par ailleurs être testé sur d'autres corpus : il serait intéressant d'étudier si certains corpus se prêtent mieux à une méthode qu'à l'autre. Bien que CFMf suppose que le corpus se prête au clustering et que les clusters identifiés correspondent à des thèmes, elle ne requiert pas initialement l'hypothèse (faite dans LDA) que les documents sont des distributions de thèmes. En outre, en dehors du nombre de thèmes, CFMf ne nécessite aucune paramétrisation. Il serait également intéressant de vérifier la robustesse des résultats en répétant la même méthode sur le même corpus avec les mêmes paramètres. Nous pensons que CFMf est plus robuste que LDA, car cette méthode est moins sensible à l'ensemencement aléatoire et implique une coopération entre les prototypes pour opérer des regroupements thématiques.

6 Conclusion

Nous avons présenté une nouvelle approche prometteuse pour la construction d'un modèle thématique basé sur une combinaison de clustering neuronal, de maximisation des traits et de mesure F1 : CFMf. Nous avons également exposé les résultats de nos expériences comparatives. Cependant, malgré le potentiel évident de CFMf pour la réalisation d'études à grande échelle, telles que les études scientifiques présentées ici, plusieurs adaptations et expériences supplémentaires sont encore possibles. Nous prévoyons tout d'abord d'étendre notre comparaison à des méthodes connues pour être des alternatives efficaces à LDA, comme la méthode STM. Nous examinerons également si LDA peut être amélioré en utilisant des composants-clés de la présente méthode, notamment en exploitant la sélection de variables et l'approche FMax pour pondérer les termes dans les thèmes de LDA. Nous examinerons aussi l'effet de la réduction du lexique sur les deux types de méthodes, tout en comparant notre modèle avec des approches de modélisation de thèmes basées sur différentes représentations du corpus (binaire, fréquence, enchâssements). Nous envisageons également d'étudier la faisabilité du choix d'un nombre optimal k de thèmes, notamment avec une approche similaire à celle de [22]. Des stratégies spécifiques combinant la mesure F1 et le classement par contraste pourront être explorées afin d'optimiser davantage encore les descriptions de thèmes.

Remerciements

J.-C.L. remercie l'ANRT pour son soutien financier. F.L. remercie le Fonds de recherche du Québec Société et culture (FRQSC-276470) et la Chaire de recherche du Canada en philosophie des sciences de la vie de l'UQAM. C.M. remercie le Conseil de recherches en sciences humaines du Canada (subvention 430-2018-00899) et le programme des Chaires de recherche du Canada (CRC-950-230795) pour

leur soutien financier.

Références

[1] T.L. Griffiths et M. Steyvers, "Finding scientific topics.," *Proceedings of the National Academy of Sciences*. vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[2] E.M. Talley, D. Newman, D. Mimno, et al., "Database of NIH grants using machine-learned categories and graphical clustering.," *Nature methods*. vol. 8, no. 6, pp. 443–444, 2011.

[3] K. Börner, F.N. Silva, et S. Milojević, "Visualizing big science projects.," *Nature Reviews Physics*. vol. 3, no. 11, pp. 753–761, 2021.

[4] D.M. Blei, A.Y. Ng, et M.I. Jordan, "Latent dirichlet allocation.," *Journal of Machine Learning Research*. vol. 3, no. Jan, pp. 993–1022, 2003.

[5] J.L. Boyd-Graber, Y. Hu, et D. Mimno, *Applications of topic models*. now Publishers Incorporated, 2017.

[6] J.-C. Lamirel, Y. Chen, P. Cuxac, S. Al Shehabi, N. Dugué, et Z. Liu, "An overview of the history of Science of Science in China based on the use of bibliographic and citation data : a new method of analysis based on clustering with feature maximization and contrast graphs.," *Scientometrics*. vol. 125, no. 3, pp. 2971–2999, 2020.

[7] C. Malaterre et F. Lareau, "The early days of contemporary philosophy of science : novel insights from machine translation and topic-modeling of non-parallel multilingual corpora.," *Synthese*. vol. 200, no. 3, p. 242, 2022.

[8] M. Röder, A. Both, et A. Hinneburg, "Exploring the Space of Topic Coherence Measures.," In : *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*. pp. 399–408. ACM Press, Shanghai, China (2015).

[9] J.-C. Lamirel, P. Cuxac, A.S. Chivukula, et K. Hajlaoui, "Optimizing text classification through efficient feature selection based on quality metric.," *Journal of Intelligent Information Systems*. vol. 45, no. 3, pp. 379–396, 2015.

[10] B. Fritzke, "A growing neural gas network learns topologies.," *Advances in neural information processing systems*. vol. 7, p. 1994.

[11] T. Prouteau, V. Connes, N. Dugué, et al., "SINr : Fast Computing of Sparse Interpretable Node Representations is not a Sin !," In : P.H. Abreu, P.P. Rodrigues, A. Fernández, and J. Gama, Eds. Advances in Intelligent Data Analysis XIX. pp. 325–337. *Springer International Publishing*, Cham (2021).

[12] H. Schmid, "Probabilistic part-of-speech tagging using decision trees.," In : *Proceedings of International Conference on New Methods in Language Processing*. pp. 44–49. , Manchester (1994).

[13] M.P. Marcus, M.A. Marcinkiewicz, et B. Santorini, "Building a Large Annotated Corpus of English : The Penn Treebank.," *Computational Linguistics*. vol. 19, no. 2, pp. 313–330, 1993.

[14] N. Aletras et M. Stevenson, "Evaluating topic coherence using distributional semantics.," In : *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. pp. 13–22 (2013).

[15] D. Newman, J. Han Lau, K. Grieser, et T. Baldwin, "Automatic evaluation of topic coherence.," In : *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 100–108 (2010).

[16] H.M. Wallach, I. Murray, R. Salakhutdinov, et D. Mimno, "Evaluation methods for topic models.," In : *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. pp. 1–8. ACM Press, Montreal, Quebec, Canada (2009).

[17] D. Mimno, H. Wallach, E. Talley, M. Leenders, et A. McCallum, "Optimizing semantic coherence in topic models.," In : *Proceedings of the 2011 conference on empirical methods in natural language processing*. pp. 262–272 (2011).

[18] R. Arun, V. Suresh, C.E. Veni Madhavan, et N. Murthy, "On finding the natural number of topics with latent dirichlet allocation : Some observations.," In : *Pacific-Asia conference on knowledge discovery and data mining*. pp. 391–402. Springer (2010).

[19] J. Cao, T. Xia, J. Li, Y. Zhang, et S. Tang, "A density-based method for adaptive LDA model selection.," *Neurocomputing*. vol. 72, no. 7–9, pp. 1775–1781, 2009.

[20] M. Hoffman, F. Bach, et D. Blei, "Online learning for latent dirichlet allocation.," *advances in neural information processing systems*. vol. 23, p. 2010.

[21] M.E. Roberts, B.M. Stewart, D. Tingley, et E.M. Airoldi, "The structural topic model and applied social science.," In : *Advances in neural information processing systems workshop on topic models : computation, application, and evaluation*. pp. 1–20. Harrahs and Harveys, Lake Tahoe (2013).

[22] N. Dugué, J.-C. Lamirel, et Y. Chen, "Evaluating clustering quality using features salience : a promising approach.," *Neural Computing and Applications*. p. 2021.