Modélisation thématique à l'aide des plongements lexicaux issus de Word2Vec

Svitlana Galeshchuk[1] Bruno Chaves[1]

[1] PSL Université Paris, Governance Analytics

{svitlana.galeshchuk,bruno.chavesferreira}@dauphine.fr

Résumé

Le papier étudie diverses approches pour la modélisation thématique et, en particulier, la méthode améliorée basée sur la paramétrisation des thèmes à partir d'une distribution continue sur l'espace des plongements lexicaux afin de tenir compte des interdépendances sémantiques. Ainsi, nous incorporons les représentations vectorielles des mots entrainés avec le réseau de neurones Word2Vec dans le processus génératif de la modélisation thématique. Nous proposons une approche alternative avec une approximation bêta de la distribution de l'information mutuelle et la comparons aux méthodes LDA standard et LDA Gaussien.

Mots Clefs

Modélisation Thématique, Latent Dirichlet Allocation, LDA Gaussien, Information Mutuelle, Plongements Lexicaux, Word2Vec.

Abstract

This paper discusses approaches for static topic modeling, in particular an improved method based on topic parametrization from a continuous distribution over the space of word embeddings. Word embeddings corpora proves to reflect semantic interdependences. Thus, we incorporate vectorized word representations trained with Word2Vec neural network in a generative process of topic modeling. The alternative approach with beta approximation of mutual information distribution over embeddings is proposed and compared with vanilla LDA and Gaussian LDA methods.

Keywords

Topic Modelling, Latent Dirichlet Allocation, Gaussian LDA, Mutual Information, Word Embedding, Word2Vec.

1 Introduction

La modélisation thématique est devenue une méthode de choix pour la fouille de données textuelles non structurées. De nombreux papiers (voir partie 2) se consacrent à l'implémentation de cette méthode, usuellement fondée sur la LDA (*Latent Dirichlet Allocation*), dans des domaines variés allant des textes juridiques aux papiers scientifiques. La méthode LDA définit les probabilités de mots sur une loi de Dirichlet qui appartient à la famille des distributions discrètes. Toutefois, l'usage de distributions discrètes empêche la découverte de mots nouveaux émergeant des thématiques. Pour cela, il est recommandé d'employer des distributions continues permettant au modèle d'attribuer à un mot nouveau une probabilité élevée d'appartenance à une thématique simplement parce que ce dernier est similaire à un mot existant représentatif de la thématique en question. Dans ce contexte, nous étudions la méthode LDA statique et ses formes modifiées avec des distributions de mots continues suivant des lois normales et bêta.

Le papier a ainsi pour objectif d'améliorer l'algorithme du modèle LDA standard à l'aide de l'approche continue, notamment la loi bêta, sur les plongements lexicaux (*word embeddings*).

La suite du papier est structurée de la manière suivante : La section 2 présente une brève revue de la littérature de la modélisation thématique. La section 3 décrit l'approche LDA standard. La section 4 présente la méthodologie de l'approche faisant appel à une distribution continue avec les plongements lexicaux. La section 5 introduit les données utilisées. La section 6 décrit le dispositif expérimental. Enfin, la section 7 conclut en présentant les résultats et des pistes pour des recherches futures.

2 Revue de la littérature

La modélisation thématique, fondée sur la méthode LDA, est devenue l'une des principales méthodes de fouille textuelle de la dernière décennie. Elle fait partie de la famille des méthodes d'apprentissage non supervisées destinées à extraire les structures thématiques latentes des corpus textuels.

Cette approche a été appliquée avec succès à des domaines de recherche variés : le journalisme, pour analyser les structures et tendances thématiques des articles de presse [7], les corpus de brevets [5], ou encore, pour classifier les textes issus de la littérature scientifique [14]. Toutefois, la méthode LDA n'est pas exempte de défauts. En particulier, la représentation des thématiques définies comme des distributions discrètes de mots empêche la prise en considération de mots nouveaux. Cette limitation peut être contournée en mobilisant, à la place, une distribution continue de mots sur des plongements lexicaux. Ces derniers définissent la représentation vectorielle des mots basée sur le contexte de leur utilisation au sein du corpus.

Blei et al. [1] proposent d'utiliser une distribution gaus-

sienne dans le processus génératif de la modélisation thématique dynamique afin de suivre l'évolution des thématiques au cours du temps. Dans cette étude, nous nous focalisations sur l'usage de distributions continues pour l'amélioration de la modélisation thématique statique plutôt que dynamique. En poursuivant un objectif similaire au notre, certains travaux [6], [11] et [12] ont proposé l'utilisation de la distribution gaussienne. Toutefois, dans notre papier, nous nous fondons plutôt sur les contributions de Das et al. [3] et Xun et al. [13]. Ces auteurs font émerger les thématiques d'une distribution gaussienne sur des plongements lexicaux en utilisant le modèle Word2Vec. De manière similaire, nous utilisons le modèle Word2Vec pour présenter les mots sous la forme de vecteurs et déduire les thématiques de distributions continues. Toutefois, nous justifions l'usage de la distribution bêta plutôt que gaussienne sur la base des résultats de Levy et Goldberg [8]. En effet, ces auteurs montrent que le modèle Word2Vec estime de manière implicite les informations mutuelles des paires de mots.

3 Modélisation thématique LDA standard

Le modèle LDA (*Latent Dirichlet Allocation*) est un modèle Bayésien faisant partie de la famille des modèles non supervisés génératifs où les observations sont générées par des variables latentes. Dans le contexte de la modélisation thématique, on cherche à découvrir des thèmes latents, à partir d'une collection de documents (articles, ouvrages, etc.) considérés comme des « sacs de mots » (*bag-of-words*) dans le sens où l'on ne tient pas compte de l'ordre des mots. Chaque document est modélisé par un mélange de thèmes qui génère ensuite chaque mot du document. Blei et al. [2] décrivent le processus génératif de LDA de la manière suivante :

1. Pour $k = 1$ à K :
   (a) Déduire la $\phi^{(k)} \sim Dirichlet(\beta)$
2. Pour chaque document d dans le corpus D :
   (a) Déduire la distribution de thèmes $\theta_d \sim Dirichlet(\alpha)$
   (b) Pour chaque index de mots n de 1 à N_d :
      i. Déduire le thème $z_n \sim Multinomiale(\theta_d)$
      ii. Déduire $w_{d,n} \sim Multinomiale(\phi^{z_n})$

Où $\phi^{(k)}$ est la distribution de mots dans le vocabulaire du k^{ime} thème, θ_d est la distribution de thèmes dans le document d et z_n est le thème n associé au mot $w_{d,n}$.

4 Modélisation thématique à partir d'une distribution continue

Cette partie de l'article présente l'approche gaussienne de la modélisation thématique fondée sur le plongement lexical et le modèle Word2Vec. Nous commençons par décrire Word2Vec et son utilisation dans le cadre de la modélisation thématique. Ensuite, nous discutons de la possibilité de représenter les thématiques à partir d'une distribution continue plutôt que discrète.

4.1 Le plongement lexical

Le modèle de Word2Vec est la représentation interne à partir d'un modèle de réseau de neurones de séquences de mots. Word2Vec utilise le perceptron monocouche pour apprendre le plongement lexical des mots ; c'est-à-dire que les mots sont appris à partir du contexte où ils sont mentionnés. Deux approches de Word2Vec sont proposées : CBOW et skip-gram. Nous implémentons le modèle skip-gram avec un échantillonnage négatif. Dans le processus d'apprentissage de Word2Vec, les mots avec des significations similaires convergent de manière graduelle vers les zones voisines de l'espace vectoriel [13]. Nous enrichissons les mots du corpus en les remplaçant par les mots correspondants de Word2Vec comme dans l'approche définie par Xun et al. [13].

4.2 La méthodologie de l'algorithme de LDA Gaussien et l'approche développée

La modélisation thématique de corpus textuels avec LDA est fondée sur les fréquences de types de mots. L'approche que nous utilisons est fondée sur l'idée selon laquelle les textes représentent des séquences de plongement lexicaux. Word2Vec transforme les mots en des vecteurs. Les mots, usuellement représentés par des valeurs discrètes, sont alors modifiés en des valeurs continues. Das et al. [3] font émerger les thématiques d'une distribution gaussienne sur ces plongements lexicaux et placent les *a priori conjugués* sur les valeurs suivantes : loi normale centrée à zéro pour la moyenne et la covariance.
Ils considèrent chaque document comme un mélange de thèmes de la loi de Dirichlet et décrivent le processus génératif de LDA Gaussien suivant :

1. Pour $k = 1$ à K :
   (a) Déduire la covariance du thème $E_k \sim W^{-1}(\phi, v)$
   (b) Déduire la moyenne du thème $\mu_k \sim N(\mu, \frac{1}{K} E_k)$
2. Pour chaque document d dans le corpus D :
   (a) Déduire la distribution de thèmes $\theta_d \sim Dirichlet(\alpha)$
   (b) Pour chaque index de mots n de 1 à N_d :
      i. Déduire le thème $z_n \sim Multinomiale(\theta_d)$
      ii. Déduire $v_{d,n} \sim N(\mu_{z_n}, E_{z_n})$

Ici $v_{d,n}$ est la représentation vectorielle du mot dans le document. W^{-1} est la loi de Wishart inverse pour la covariance.
Les auteurs justifient le choix de la paramétrisation gaussienne par les observations de Hermann et Blunsom [4] selon lesquelles les distances euclidiennes entre les plongements lexicaux sont corrélés avec la similarité sémantique. Pourtant, Levy et Goldberg [8] démontrent que le modèle

de Word2Vec factorise une matrice de contexte de mots (*co-occurrence matrix*) de manière implicite. Ses cellules sont les informations mutuelles des paires de mots et de contextes respectifs décalés d'une constante globale. Ainsi, les vecteurs de mots sont déduits de la distribution des informations mutuelles. Zaffalon et Hutter [15] montrent que la meilleure approximation de la loi de l'informations mutuelles conditionnelles est la loi bêta. Elle appartient à une famille de lois de probabilités continues. Dans notre approche nous suivrons les résultats de Zaffalon et Hutter [15]. Par suite nous proposons le processus génératif de LDA :

1. Pour $k = 1$ à K :
   (a) Déduire la covariance du thème $E_k \sim W^{-1}(\phi, v)$
   (b) Déduire la moyenne du thème $\mu_k \sim N(\mu, \frac{1}{K} E_k)$
2. Pour chaque document d dans le corpus D :
   (a) Déduire la distribution de thèmes $\theta_d \sim Dirichlet(\alpha)$
   (b) Pour chaque index de mots n de 1 à N_d :
      i. Déduire le thème $z_n \sim Multinomiale(\theta_d)$
      ii. Déduire $v_{d,n} \sim bêta(\alpha_n, \beta_{z_n})$

Où α et β sont les paramètres de forme de la distribution bêta.

5 Les données utilisées

Dans notre étude, nous utilisons un corpus composé des titres et résumées des articles présentés à la conférence SIOE (*Society for Institutional & Organizational Economics*) de 2008 à 2017. SIOE est une société savante internationale sur l'économie des institutions et organisations. Elle organise chaque année la principale conférence internationale consacrée à la recherche sur ces thématiques. Les données ont été récupérés à partir de la base de données MySQL du site web de la conférence (www.sioe.org).

6 Démarche expérimentale

Les résultats issus du modèle LDA standard, le modèle thématique reproduit à partir de Das et al., 2015 [3] et le modèle que nous avons développé sont présentés, respectivement, dans les tables 1, 2 et 3. Par ailleurs, la visualisation des résultats LDA avec librairie Python pyLDA est présentée dans la figure 1 ci-dessous. Les trois modèles sont présentés comme des clusters de mots sur 4 thématiques.
Ces dernières sont assez proches dans les 3 modèles. On peut les représenter par les termes suivants : « Management », « Institutional framework », « Legal framework » et « Market environment ».
L'évaluation est l'un des principaux défis de la modélisation thématique. Des méthodes qualitatives et quantitatives peuvent être mobilisées comme dans [3] et [13]. Nous avons décidé d'utiliser une méthode quantitative et, plus particulièrement, celle proposée par Röder et al. [10]. Les auteurs élaborent une méthodologie permettant de mesurer la cohérence thématique qui consiste à mesurer l'ajustement entre des paires les mots ou sous-ensemble de mots. L'algorithme commence par effectuer une segmentation par paires de mots. Ensuite, chaque paire de mots est évaluée à l'aide d'un score d'information mutuelle spécifique (*pointwise mutual information*) normalisée et les probabilités des mots sont calculées. La cohérence résulte de l'agrégation de la concordance des paires sur la base des probabilités calculées. Pour cela, nous avons utilisé la librairie Python Palmetto qui permet de calculer la cohérence thématique des ensembles de mots ci-dessous. Les résultats (arrondis) obtenus sont présentés dans la dernière ligne des tables 1, 2 et 3. Notre approche obtient le score agrégé, sur les 4 thèmes, le plus élevé (1.342). L'approche LDA standard arrive en second (1.315) et le LDA gaussien en dernier (1.246). Ces résultants n'en restent pas moins très proches. Par conséquent, nous envisageons d'utiliser d'autres méthodes qualitatives et quantitatives dans des recherches futures.

Management	**Institutional framework**	**Legal framework**	**Market environment**
firm	institution	policy	contract
market	country	state	cost
law	level	law	agent
industry	development	model	transaction
innovation	growth	government	model
investment	government	court	governance
incentive	effect	decision	market
cost	state	election	property
organization	sector	party	system
model	impact	case	party
0.442	0.311	0.316	0.250

TABLE 1: Modélisation thématique LDA standard

Management	**Institutional framework**	**Legal framework**	**Market environment**
firm	state	innovation	business
market	corruption	patent	analysis
cost	development	property	decision
performance	industry	patent	market
industry	market	regulation	right
quality	governance	judge	datum
procurement	institution	law	change
incentive	policy	crime	innovation
strategy	regime	rule	governance
agent	economy	firm	capital
0.390	0.433	0.323	0.098

TABLE 2: Modélisation thématique de Das et al. [3]

Management	**Institutional framework**	**Legal framework**	**Market environment**
firm	government	enforcement	firm
contract	country	law	market
cost	level	patent	country
price	state	property	land
market	market	right	investment
governance	tax	system	capital
procurement	institution	rule	innovation
transaction	agent	crime	price
agent	decision	judge	level
strategy	policy	firm	change
0.467	0.263	0.321	0.290

TABLE 3: Notre Modélisation (*distribution bêta*)

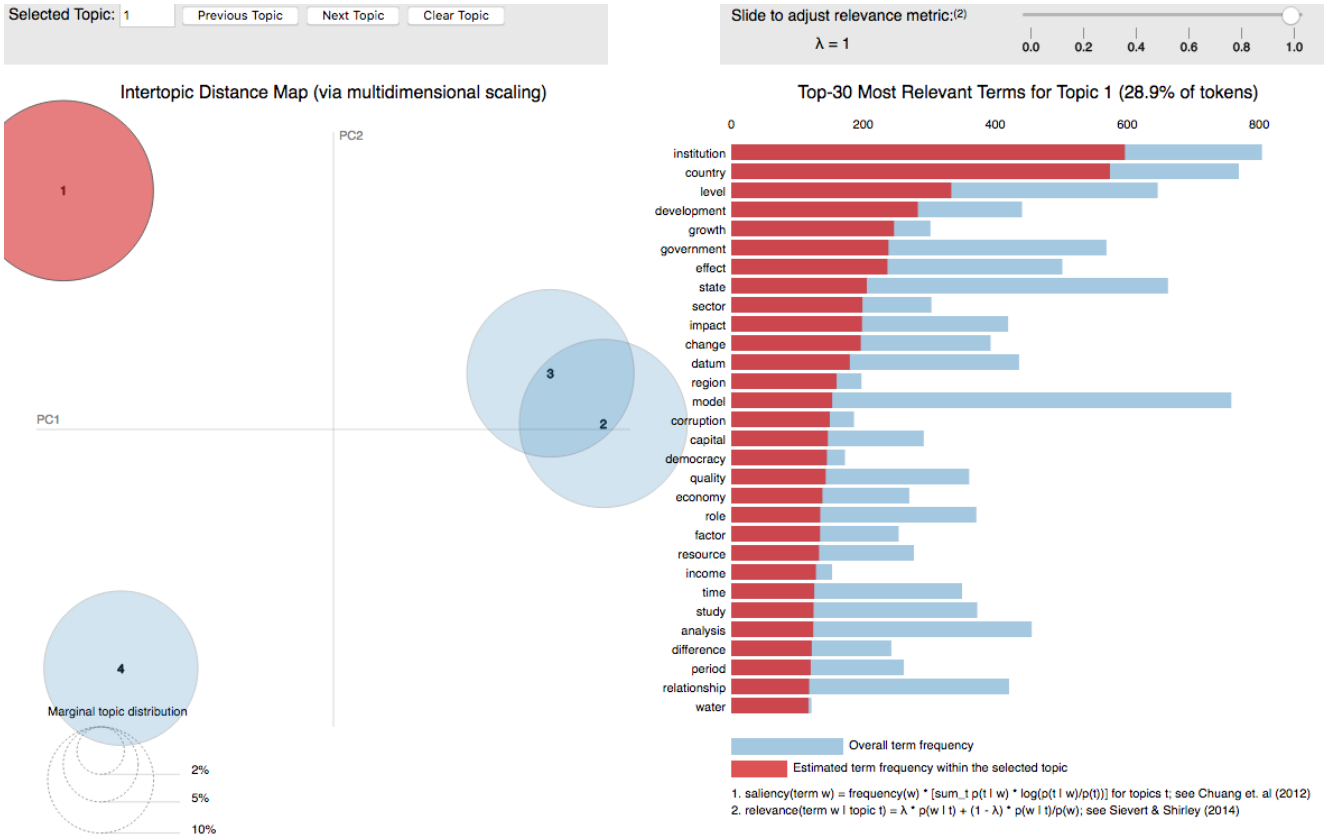

FIGURE 1 – Visualisation du thème « *Institutional framework* » avec le modèle LDA standard

Références

[1] Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. *In Proceedings of the 23rd international conference on Machine learning* (pp. 113-120). ACM.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

[3] Das, R., Zaheer, M., & Dyer, C. (2015). Gaussian lda for topic models with word embeddings. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1 : Long Papers) (Vol. 1, pp. 795-804).

[4] Hermann, K. M., & Blunsom, P. (2014). Multilingual models for compositional distributed semantics. arXiv preprint arXiv :1404.4641.

[5] Hu, Z., Fang, S., & Liang, T. (2014). Empirical study of constructing a knowledge organization system of patent documents using topic modeling. *Scientometrics,* 100(3), 787-799.

[6] Hu, P., Liu, W., Jiang, W., & Yang, Z. (2012, September). Latent topic model based on Gaussian-LDA for audio retrieval. *In Chinese Conference on Pattern Recognition* (pp. 556-563). Springer, Berlin, Heidelberg.

[7] Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism,* 4(1), 89-106

[8] Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *In Advances in neural information processing systems* (pp. 2177-2185).

[9] Naili, M., Chaibi, A. H., & Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science,* 112, 340-349.

[10] Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408). ACM.

[11] Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. arXiv preprint arXiv :1206.3298.

[12] Weinshall, D., Levi, G., & Hanukaev, D. (2013, February). LDA topic model with soft assignment of descriptors to words. *In International Conference on Machine Learning* (pp. 711-719).

[13] Xun, G., Gopalakrishnan, V., Ma, F., Li, Y., Gao, J., & Zhang, A. (2016, December). Topic discovery for short texts using word embeddings. *In Data Mining (ICDM), 2016 IEEE 16th International Conference on* (pp. 1299-1304). IEEE.

[14] Yau, C-K, Porter, A.L., Newman, N.C., and Suominen, A. (2014), Clustering scientific documents with topic modeling, *Scientometrics,* GTM special issue; 100 (3) 767-786.

[15] Zaffalon, M. & Hutter M. (2002). Robust feature selection by mutual information distributions. *In Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence* (UAI'02), Adnan Darwiche and Nir Friedman (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 577-584.