A Modular Approach to Topic Modeling for Heterogeneous Documents

Discussion Paper

Giovanni Toto[2], Emanuele Di Buccio[1,2,*]

[1]Department of Information Engineering, University of Padova, Via G. Gradenigo 6/b, 35131, Padova, Italy
[2]Department of Statistical Sciences, University of Padova, Via C. Battisti, 241, 35121, Padova, Italy

Abstract

Topic Modeling algorithms help unveil the latent thematic structure from large document collections. Previous works showed that traditional approaches could be less effective when applied to short texts, e.g., tweets; however, that can be mitigated by assuming that each document is about a single topic, as done in Twitter-LDA. In this work, we relax this assumption and propose a new model where a document can be about single or multiple topics. Our model allows the generation of diverse types of descriptors from latent topics, e.g., words and hashtags, similarly to Hashtag-LDA. Moreover, words/hashtags can be generated from topics or a background/global distribution. The proposed model is modular, and our goal is to tailor it to collections that can be heterogeneous both in the presence of single or multiple-topic documents and in the adoption of diverse topic representations.

Keywords

Topic Modeling, Text Mining, Heterogeneous Text Topic Modeling, Topic Modeling for Microblogs

1. Introduction

Topic Modeling algorithms are Machine Learning approaches introduced to unveil the latent thematic structure from unstructured document corpora. In Probabilistic Topic Models [1, 2], whose most representative technique can be considered Latent Dirichlet Allocation (LDA) [3], a theme is represented through a *topic* which is a probability distribution over the entire corpus vocabulary. Documents in the corpus can be represented as a mixture of topics. One of the benefits of this representation is interpretability: the weights (probabilities) of the words in a topic help the interpretation of the topic, i.e., associate a topic to a theme, for instance, by looking at the words with the highest weights; moreover, the extracted topics allow users to have a preliminary idea of the themes covered in a possibly large document corpus; finally, each document can be represented in terms of topics, thus obtaining a more dense representation than when words are used as descriptors.

Topic Modeling has been adopted in many tasks and settings [4]. Previous works showed that when applied to short texts, e.g., Microblog posts, the lack of word co-occurrence information can

IIR2022: 12th Italian Information Retrieval Workshop, June 29 - June 30th, 2022, Milan, Italy

*Corresponding author.

toto@stat.unipd.it (G. Toto); emanuele.dibuccio@unipd.it (E. Di Buccio)
0000-0002-7444-5702 (G. Toto); 0000-0002-6506-617X (E. Di Buccio)

negatively affect the effectiveness of traditional approaches [5]; therefore, ad-hoc solutions were proposed. Twitter-LDA [6] assumes that each tweet is generated by a single topic, moving the topic mixture from document to the user; experimental results suggested that this assumption is promising. We hypothesize that this assumption might be too restrictive for generic short texts and also on Twitter after the extension of the maximum number of characters per tweet. Our approach aims at relaxing the assumption of single-topic short text.

Besides text length, another issue is the heterogeneity of the descriptors. For instance, Twitter allows the use of hashtags: an *hashtag* is a sequence of characters – not including punctuation or spaces – starting with "#" which "is used to index keywords or topics on Twitter".[1] Hashtag-LDA [7] relies on the same assumption of single-topic tweets, but differently from Twitter-LDA, not only words but also hashtags are generated by the topics. Even if previous works [8, 9, 10] explicitly include metadata/tags/labels, in Hashtag-LDA tags are generated by the latent topics, and not vice-versa. Our model shares the same intuition underlying Hashtag-LDA but relax the single-topic assumption and explicitly considers the possible generation of words and hashtags from a background/global distribution, as Twitter-LDA does for the words.

2. Modeling Single and Multi-Topic Documents and Heterogeneous Descriptors

The overall model in plate notation is depicted in Fig. 1. The model can be considered an extension of LDA, Twitter-LDA, and Hashtag-LDA: the latent structure of LDA is used to model multi-topic documents, while the latent structure of the other two is used to model single-topic documents. We will describe the model in the context of Microblog, e.g., Twitter; however, the model is modular and we plan to apply it to heterogeneous document collections constituted by diverse types of documents, e.g., news, forum posts, blog posts, and tweets.

Our model can be decomposed in four conceptual blocks depicted in different colors in Fig. 1.

The first block, highlighted in red, models the key idea underlying our approach: two *types* of documents can be distinguished, those about a single topic and those about multiple topics. In the model, each user u has her own inclination to write document on a single topic or multiple topics; this inclination is encoded in the probability π_u^T, which affects the type x_{ud} of document written by u. This is a simplifying assumption since aspects other than the user might affect the choice of writing on single or multiple topics. Our approach allows diverse types of users – in terms of their inclination on single or multiple topics – to be modeled. For instance, influencers or politicians, through their official accounts, usually write long and elaborated posts to express their point of view; other users publish very concise messages, e.g., for answers to other tweets.

The second and third blocks are highlighted in blue and green; they are responsible for the topic assignment to documents, words and hashtags. The assignment depends on the document type identified in the first block: if the document is about multiple topics – blue block –, assignment is very close to that proposed in LDA, where a single topic is associated to each textual element, e.g., a word or an hashtag; if the document is about a single topic – green block –, topic assignment follows Twitter-LDA and Hashtag-LDA, where a single topic is assigned to each document — we will refer to such topic as the *main topic*.

[1] https://help.twitter.com/en/using-twitter/how-to-use-hashtags

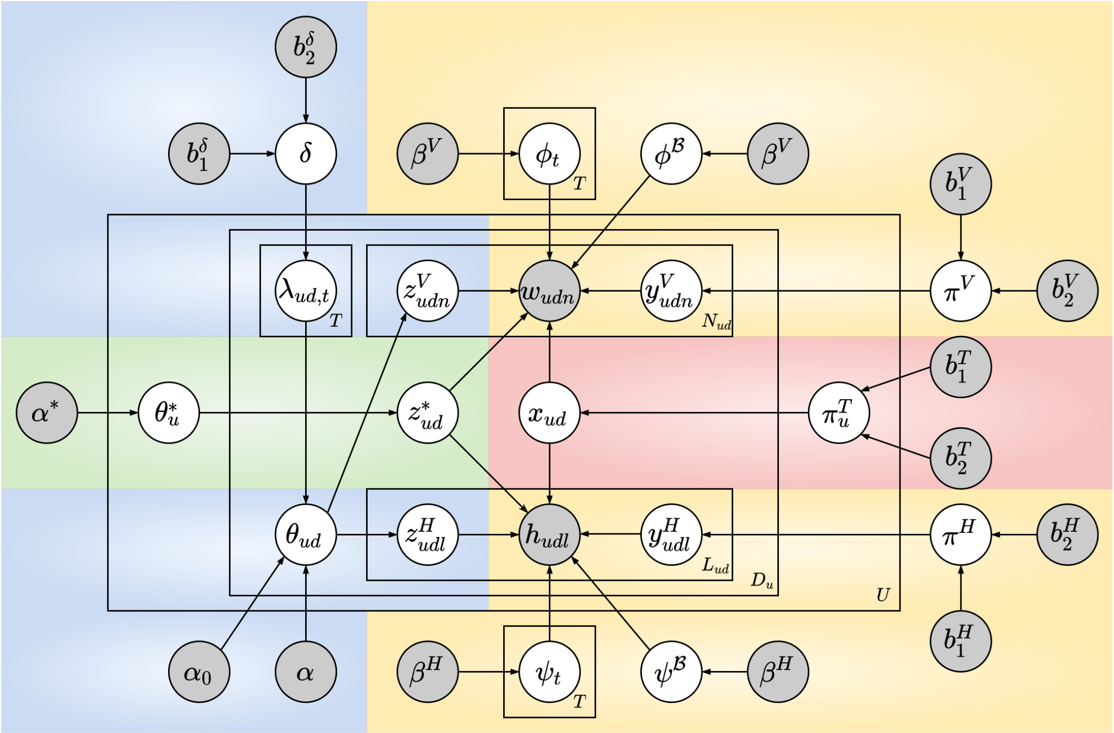

Figure 1: Full model in plate notation.

When a document is about multiple topics – blue block in Fig. 1 –,

- a topic proportion, $\boldsymbol{\theta}_{ud}$, is assigned to each document *ud*, where the *t*th element denotes the importance of topic *t* for document *ud*;
- a vector $\boldsymbol{\lambda}_{ud}$ of active topics is assigned to each document *ud*: a non-active topic will have a very low weight, thus making very unlikely the observation of words or hashtags associated with that topic; δ denotes the probability that a topic is active;
- a topic z^V_{udn} is assigned to each word *udn* and topics with a larger weight in the document will generate words more frequently; similarly a topic z^H_{udl} is assigned to each hashtag *udl* and topics with a larger weight will generate hashtags more frequently.

In the event of tweets, longer documents will be focused on a limited number of topics and the vector $\boldsymbol{\lambda}_{ud}$ will introduce sparsity in the representation of documents as mixture of topics.

When a document is about a single topic – green block in Fig. 1 –,

- a topic proportion, $\boldsymbol{\theta}^*_u$, is assigned to each user and its *t*th element denotes the preference of the user to select the *t*th topic as the main topic;
- a main topic, z^*_{ud}, is assigned to each document *ud* and topics with larger weight in $\boldsymbol{\theta}^*_u$ are assigned more frequently.

In this case, no topic is associated to words and hashtags since they are generated from the main topic. The idea underlying this block is that more simple and concise documents are focused on a single topic and the selection of the topic depends on the personal preference of the user.

The last block is highlighted in orange and is the one responsible for the generation of words and hashtags from the topics. The model considers:

- a double representation of topics: there is a fixed number of topics and each topic is represented both as a distribution over words and as a distribution over hashtags;
- *background words* common to all the topics: this group of words is modeled as a "dedicated" topic and therefore is represented as a distribution over the word vocabulary; similarly, *global hashtags* are used independently from the topic and are modeled as a dedicated topic and represented as a distribution over the hashtag vocabulary.

The generative processes of words and hashtags are basically identical, and the difference lies in the latent variables and the parameters. In the case of words (hashtags), a *source*, y^V_{udn} (y^H_{udl}), is assigned to each word *udn* (hashtag *udl*) and indicates if it was generated from a topic or it is a background word (global hashtag). The observed word (hashtag) depends on the source, the type of document, and the topic: if it is a background word (global hashtag), the background distribution $\boldsymbol{\phi}^B$ ($\boldsymbol{\psi}^B$) is considered; otherwise, the main topic distribution, $\boldsymbol{\phi}_{z^*_{ud}}$ ($\boldsymbol{\psi}_{z^*_{ud}}$), or that of the topic associated to the word, $\boldsymbol{\phi}_{z^V_{udn}}$ ($\boldsymbol{\psi}_{z^H_{udl}}$), is considered, depending on the document type — respectively single or multiple-topic document. The two generative processes are identical and independent of each other – the presence of certain words does not affect the presence of hashtags in the same document and vice-versa –; therefore, our approach can be extended with topic representations based on additional vocabularies, e.g., emojis.

3. Ongoing and Future Work

We are currently focusing on the experimental evaluation of the proposed approach using Twitter datasets and "generic" short texts [5]. A first evaluation was carried out on a collection of tweets in Italian gathered by Twitter API. The collection is constituted of 8895 tweets about the COVID-19 published between Jan. 24 and Jan. 30, 2022. LDA, Twitter-LDA, and Hashtag-LDA were adopted as baselines. Since those methods are parametric in the number of topics, we selected the number of topics that maximized Topic Coherence (TC), specifically TC-PMI [11], for LDA. Topic Coherence was computed on the same collection, not on an external corpus. We used collapsed Gibbs sampling for learning the topic models.

Our approach achieved results comparable with Twitter-LDA, which was the most effective baseline in terms of Topic Coherence – TC-PMI and TC-NZ [12] – and Jensen-Shannon divergence between the distribution over the words of the topics and the distribution over the words of the collection. However, differently from Twitter-LDA, our approach provides two different representations of the same topic, one in terms of words and one in terms of hashtags; these representations might be beneficial for interpreting the topics.

The subsequent steps will be: (i) investigate in detail the effect of the number of topics on the proposed approach; (ii) investigate how to tailor the model to heterogeneous test collections [13] since it was initially designed for Microblogs; (iii) extend the set of adopted baselines, e.g., including the relevant ones among those surveyed in [13, 5]; (iv) evaluate the effectiveness in diverse tasks such as (hash)tag recommendation, text classification, and clustering; (v) perform a qualitative analysis through a case study, for instance, involving expert users.

References

[1] M. Steyvers, T. Griffiths, Probalistic Topic Models, in: Latent Semantic Analysis: A Road To Meaning, Lawrence Erlbaum Associates Publishers, 2007, pp. 427–448.

[2] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (2012) 77–84. doi:10.1145/2133806.2133826.

[3] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[4] J. Boyd-Graber, Y. Hu, D. Mimno, Applications of Topic Models, Foundations and Trends® in Information Retrieval 11 (2017) 143–296. doi:10.1561/1500000030.

[5] J. Qiang, Z. Qian, Y. Li, Y. Yuan, X. Wu, Short Text Topic Modeling Techniques, Applications, and Performance: A Survey, IEEE Transactions on Knowledge and Data Engineering 34 (2022) 1427–1445. doi:10.1109/TKDE.2020.2992485. arXiv:1904.07695.

[6] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, in: P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, V. Mudoch (Eds.), Advances in Information Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 338–349.

[7] F. Zhao, Y. Zhu, H. Jin, L. T. Yang, A personalized hashtag recommendation approach using lda-based topic model in microblog environment, Future Gener. Comput. Syst. 65 (2016) 196–206. doi:10.1016/j.future.2015.10.012.

[8] D. Ramage, D. Hall, R. Nallapati, C. D. Manning, Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, in: EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009, August, 2009, pp. 248–256.

[9] F. S. Tsai, A tag-topic model for blog mining, Expert Systems with Applications 38 (2011) 5330–5335. doi:10.1016/j.eswa.2010.10.025.

[10] Z. Ma, W. Dou, X. Wang, S. Akella, Tag-Latent Dirichlet Allocation: Understanding Hashtags and Their Relationships, in: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), volume 1, IEEE, 2013, pp. 260–267. doi:10.1109/WI-IAT.2013.38.

[11] D. Newman, J. H. Lau, K. Grieser, 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, HLT '10, Association for Computational Linguistics, USA, 2010, p. 100–108.

[12] J. Boyd-Graber, D. Mimno, D. Newman, Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements, CRC Handbooks of Modern Statistical Methods, CRC Press, Boca Raton, Florida, 2014.

[13] J. Qiang, P. Chen, W. Ding, T. Wang, F. Xie, X. Wu, Heterogeneous-length text topic modeling for reader-Aware multi-document summarization, ACM Transactions on Knowledge Discovery from Data 13 (2019). doi:10.1145/3333030.