Readitopics: Make Your Topic Models Readable via Labeling and Browsing

Julien Velcin[1], **Antoine Gourru**[1], **Erwan Giry-Fouquet**[1],
Christophe Gravier[2], **Mathieu Roche**[3], **Pascal Poncelet**[4]
[1] Université de Lyon (ERIC, Lyon 2)
[2] Université de Lyon (LHC, St-Etienne)
[3] Cirad (TETIS, Montpellier)
[4] Université de Montpellier (LIRMM)
{julien.velcin, antoine.gourru, erwan.giry-fouquet}@univ-lyon2.fr,
christophe.gravier@univ-st-etienne.fr, mathieu.roche@cirad.fr, Pascal.Poncelet@lirmm.fr

Abstract

Readitopics provides a new tool for browsing a textual corpus that showcases several recent work on topic labeling and topic coherence. We demonstrate the potential of these techniques to get a deeper understanding of the topics that structure different datasets. This tool is provided as a Web demo but it can be installed to experiment with your own dataset. It can be further extended to deal with more advanced topic modeling techniques.

1 Introduction

Topic Modeling is a powerful tool for monitoring information flow. It is therefore a cornerstone for visualization platforms, allowing users to browse huge volumes of textual data. Interpreting and understanding them is usually left to the human on the basis of a ranked list of likely words. To address this issue, automatic topic labeling [Mei *et al.*, 2007] emerged as a task of the utmost practical interest – that is, to provide a textual expression sufficient to quickly grasp the topic informational content. Besides, several work designed automatic measures to assess the quality of topics, in particular by evaluating their semantic coherence [Röder *et al.*, 2015]. Several visualization tools have been designed on top of topic models [Chaney and Blei, 2012; Liu *et al.*, 2012; Sievert and Shirley, 2014] but even recent work such as [Kim *et al.*, 2017] do not integrate an advanced topic labeling tool for providing a deep understanding of the underlying topic meaning, or the possibility to estimate their coherence.

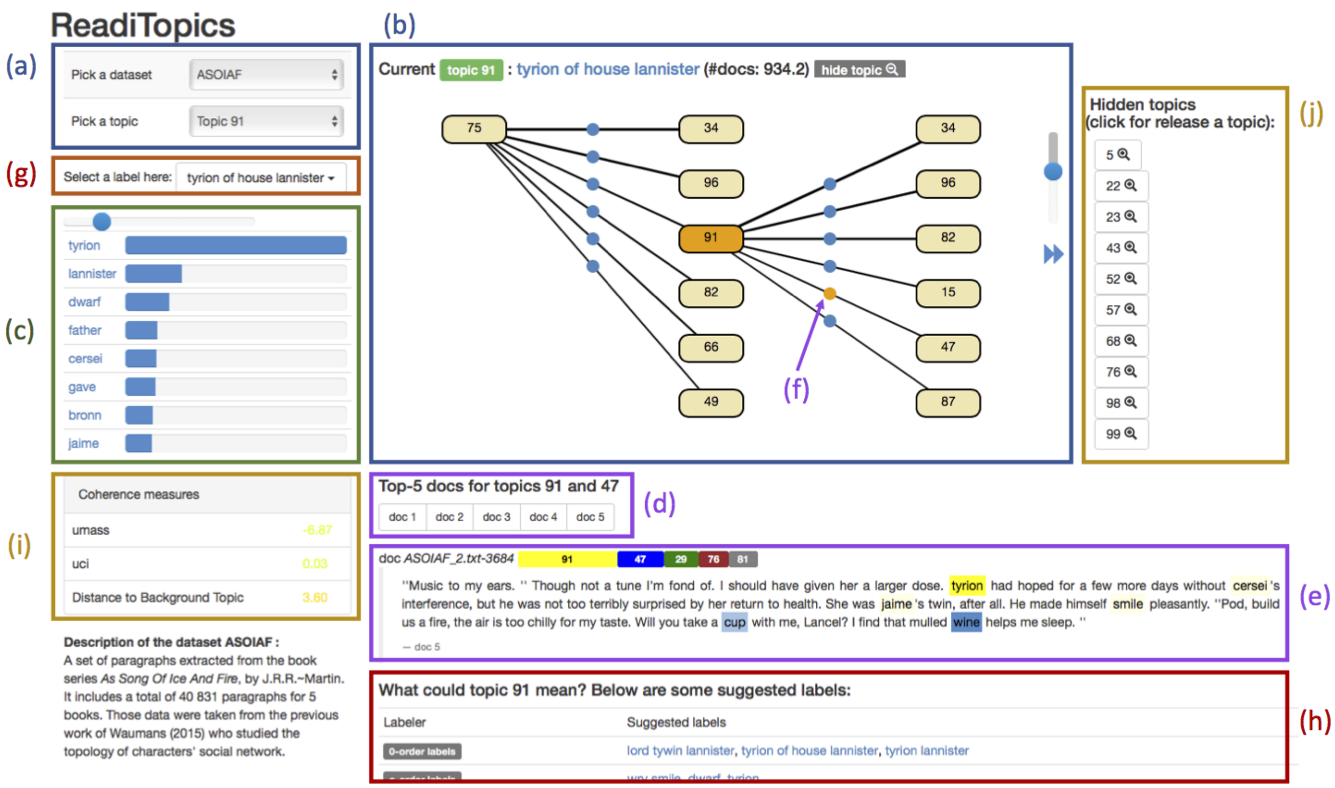
In this paper we showcase Readitopics, a Web interface to grasp topic informational contents via topic labeling and browsing. It gives the opportunity to get a better understanding of the underlying meaning of the topics that pervade their corpus. We experiment on several case studies with the well-known LDA model [Blei *et al.*, 2003] since it has been observed that it leads to “concise and coherent topics” outperforming SVD and NMF [Stevens *et al.*, 2012]. However, our system is intended to be compatible with any kind of (flat) topic models and it can be extended to document clustering.

2 Topic Labeling and Coherence

Topic labeling aims at finding a relevant label or title that provides a better understanding of what constitutes the homogeneity of a given topic [Mei *et al.*, 2007; Danilevsky *et al.*, 2014]. In the following we consider that a given topic z is associated to a distribution $p(w|z)$ over a vocabulary of words w, among which we can extract the top k words, and every document d is associated to a distribution $p(z|d)$ over topics. Several measures have been proposed to associate either a term or a phrase based on the top-k words with a given topic. [Lau *et al.*, 2011] and [Bhatia *et al.*, 2016] used external resources (e.g., Wikipedia) to find a title and introduced some supervision. [Kou *et al.*, 2015] explored new solutions relying on letter trigram vectors and word embeddings. Another option is to use multiple measures to increase the chance to find the correct phrase [Gourru *et al.*, 2018].

Using representative sentences has been successfully integrated into topic-modeling oriented applications [El-Assady *et al.*, 2017]. The system we present in this demo lets the user select the best possible labels built by a selected number of (unsupervised) labeling techniques. We focus on techniques based on n-gram scoring, such as the 0-order (see Section 3.2), since it has been shown that keyphrases are more understandable than word lists or even images [Aletras *et al.*, 2017].

Topics are not equal when it comes to their relevance with respect to the global informational content of the entire corpus. We therefore integrated a set of coherence measures for each topic, serving as a relevance metric. For a given topic z, DBT measures the distance of $p(w|z)$ from the background word distribution: a small value means a broad topic with general (contextual) words [AlSumait *et al.*, 2009]. UCI [Newman *et al.*, 2010] measures the quality by calculating the average Pointwise Mutual Information (PMI) of the top words on an external corpus (in our case, a dump of Wikipedia). UMass [Mimno *et al.*, 2011] estimates the quality based on document frequencies of the original documents used for learning the topics. We showcase that these metrics are marginally correlated, which leaves topic coherence still an open field of research.



Pick a dataset: ASOIAF

Pick a topic: Topic 91

Select a label here: tyrion of house lannister

Coherence measures	
umass	-6.87
uci	0.03
Distance to Background Topic	3.60

Description of the dataset ASOIAF :
A set of paragraphs extracted from the book series *As Song Of Ice And Fire*, by J.R.R.~Martin. It includes a total of 40 831 paragraphs for 5 books. Those data were taken from the previous work of Waumans (2015) who studied the topology of characters' social network.

Current topic 91 : tyrion of house lannister (#docs: 934.2)

Hidden topics (click for release a topic):
5
22
23
43
52
57
68
76
98
99

Top-5 docs for topics 91 and 47

doc *ASOIAF_2.txt-3684*

"Music to my ears. '' Though not a tune I'm fond of. I should have given her a larger dose. tyrion had hoped for a few more days without cersei 's interference, but he was not too terribly surprised by her return to health. She was jaime 's twin, after all. He made himself smile pleasantly. ''Pod, build us a fire, the air is too chilly for my taste. Will you take a cup with me, Lancel? I find that mulled wine helps me sleep. ''

— doc 5

What could topic 91 mean? Below are some suggested labels:

Labeler	Suggested labels
0-order labels	lord tywin lannister, tyrion of house lannister, tyrion lannister
	wry smile, dwarf, tyrion

Figure 1: Overview of Readitopics (online demo available at http://mediamining.univ-lyon2.fr/readitopics/)

3 Browsing a Corpus With Readitopics

3.1 Studied Datasets

We showcase Readitopics on four different corpora: scientific articles (SA), news articles (NA) Harry Potter (HP) and a song of ice and fire book series (ASOIAF). SA is a set of 18,465 scientific abstracts gathered by [Tang *et al.*, 2012] over a period of 16 years ; NA is a set of 12,067 news we gathered automatically from the Huffington Post RSS feeds (US version). This set spans a period of 3 months (from June the 20th until Sept. the 8th, 2016) ; HP and ASOIAF are two sets of paragraphs extracted from the book series Harry Potter, by J.K. Rowling, and As Song Of Ice And Fire, by J.R.R. Martin. It includes a total of 38,997 paragraphs for 7 books and 40,831 paragraphs for 5 books, respectively, taken from the previous work of [Waumans *et al.*, 2015].

3.2 Demonstration Overview

The conference attendees will go through five different main tasks that can be performed on Readitopics interface (see Fig.1). The features introduced by Readitopics and not available in publicly available web interfaces counterparts are prefixed with a star (*). The full sources are available online through a git repository[1].

- Choose a dataset and a topic among the full list **(a)** or via the (partial) topic graph **(b)**. We can see in **(c)** the topic's k words (k can be customized) and the top p documents **(d)**. Each document is provided with the top 5 topics **(e)** and the words associated with the current topic are highlighted.

- Move from one topic to another by following the *correlation* between documents, such as in [Liu *et al.*, 2014]. The graph edges in **(b)** are sorted by their score in term of Pearson correlation between topics i and j. The number of top correlated topics can be customized. Once a topic is selected on the right side, we can click on the blue arrow to resume the browsing experience.

(*) Look at the documents at the frontier of two topics (blue circles over the edges), i.e. documents that maximize the use of two topics at the same time (e.g., 47 and 91 for **(f)**). The top documents in **(d)** targets two topics now.

(*) Choose recommended labels in **(g)** to explicit the meaning of a topic. Suggested labels given in **(h)** are based on n-grams scored by different measures, such as the 0-order and 1-order of [Mei *et al.*, 2007] and C-order of [Gourru *et al.*, 2018]. For instance, 0-order scores a set of label candidates (in our case, selected by the term identification tool of [Lossio-Ventura *et al.*, 2014]) by considering the sum of log probabilities of the words composing the term. Besides, extracted sentences are presented at the very bottom of **(h)** to help the user to figure out the meaning of a given topic (see [Gourru *et al.*, 2018] for technical details).

(*) See in **(i)** the coherence of the topic as calculated by several state-of-the-art measures. These measures, such as Umass or UCI, are calculated with the Palmetto library provided by [Röder *et al.*, 2015]. Based on this information, Readitopics allows the user to hide a subset of topics **(j)**.

[1] https://github.com/Erwangf/readitopics

References

[Aletras *et al.*, 2017] Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*, 68(1):154–167, 2017.

[AlSumait *et al.*, 2009] Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. Topic significance ranking of LDA generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer, 2009.

[Bhatia *et al.*, 2016] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. Automatic labelling of topics with neural embeddings. In *Proceedings of COLING*, 2016.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.

[Chaney and Blei, 2012] Allison June-Barlow Chaney and David M Blei. Visualizing topic models. In *Proceedings of the International Conference of Weblogs and Social Media (ICWSM)*, 2012.

[Danilevsky *et al.*, 2014] Marina Danilevsky, Chi Wang, Nihit Desai, Xiang Ren, Jingyi Guo, and Jiawei Han. Automatic construction and ranking of topical keyphrases on collections of short documents. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 398–406. SIAM, 2014.

[El-Assady *et al.*, 2017] Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Daniel Keim, and Christopher Collins. Progressive learning of topic modeling parameters: A visual analytics framework. *IEEE Trans. on Visualization and Computer Graphics*, 2017.

[Gourru *et al.*, 2018] Antoine Gourru, Julien Velcin, Mathieu Roche, Christophe Gravier, and Pascal Poncelet. United we stand: Using multiple strategies for topic labeling. In *Proceedings of the 23rd Conference on Natural Language & Information Systems (NLDB)*, pages 352–363, Paris, France, 2018.

[Kim *et al.*, 2017] Minjeong Kim, Kyeongpil Kang, Deokgun Park, Jaegul Choo, and Niklas Elmqvist. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):151–160, 2017.

[Kou *et al.*, 2015] Wanqiu Kou, Fang Li, and Timothy Baldwin. Automatic labelling of topic models using word vectors and letter trigram vectors. In *Asia Information Retrieval Symposium*, pages 253–264. Springer, 2015.

[Lau *et al.*, 2011] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT) - Volume 1*, pages 1536–1545. Association for Computational Linguistics, 2011.

[Liu *et al.*, 2012] Shixia Liu, Michelle X Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):25, 2012.

[Liu *et al.*, 2014] Shixia Liu, Xiting Wang, Jianfei Chen, Jim Zhu, and Baining Guo. Topicpanorama: A full picture of relevant topics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 183–192. IEEE, 2014.

[Lossio-Ventura *et al.*, 2014] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biotex: A system for biomedical terminology extraction, ranking, and validation. In *Proceedings of the 13th ISWC*, pages 157–160, 2014.

[Mei *et al.*, 2007] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 490–499. ACM, 2007.

[Mimno *et al.*, 2011] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.

[Newman *et al.*, 2010] David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 215–224. ACM, 2010.

[Röder *et al.*, 2015] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.

[Sievert and Shirley, 2014] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[Stevens *et al.*, 2012] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics, 2012.

[Tang *et al.*, 2012] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1285–1293, 2012.

[Waumans *et al.*, 2015] Michaël C Waumans, Thibaut Nicodème, and Hugues Bersini. Topology analysis of social networks extracted from literature. *PloS one*, 10(6), 2015.