Readitopics: Make Your Topic Models Readable via Labeling and Browsing
Julien Velcin, Antoine Gourru, Erwan Giry-Fouquet, Christophe Gravier, Mathieu Roche, Pascal Poncelet

To cite this version:

HAL Id: lirmm-01910611
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01910611
Submitted on 1 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Readitopics: Make Your Topic Models Readable via Labeling and Browsing

Julien Velcin¹, Antoine Gourru¹, Erwan Giry-Fouquet¹, Christophe Gravier², Mathieu Roche³, Pascal Poncelet⁴

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 for labeling topic models and estimating topic coherence. We will demonstrate the potential of these techniques to get a deeper understanding of the topics that structure different kinds of datasets. This tool is provided as a Web demo but it can be easily 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 [Allan, 2012]. 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.

In parallel, 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 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 (Section 2) and browsing (Section 3). Readitopics gives the opportunity for users 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 easily extended to document clustering. Besides, we released the sources of Readitopics so that the user can experiment with her own data.

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 [AILsumait 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.
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 (*).

- 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 to 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 mentioned earlier, 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).

Overall, the demonstration will focus on showing how i) all this information makes it easier for the human to quickly grasp the topic informational content, and ii) the open challenges in topic labeling and topic coherence. The full sources are available online through a git repository1.

1https://github.com/Erwangf/readitopics
References


4 Additional Requested Information

4.1 Type of Demonstration
We choose to present Readitopics in a traditional way. In addition to give an overview of our software at a demonstration booth, we can help the attendees to install it on their computer if it is equipped with Java 8 or above.

4.2 List of Requirements
For presenting Readitopics, we only need a table with a couple of chairs. The demonstration can be done on a simple laptop.