Weighted coherence as topic models' interpretability measure

A Preprint

Zhgutov K. D. (zhgutov.kd@phystech.edu)
Alekseev V. A. (wasya.alekseev@gmail.com)
Vorontsov K. V. (vokov@forecsys.ru)

Moscow Institute of Physics and Technology

Abstract

Topic modeling is very useful for analyzing text data. It can be used to analyze large collection of text data such as articles, reviews, social media, and others. This helps in clusterization documents by topic, extracting keywords, and identifying patterns in the data. There are a lot of automatically calculated criteria of informativeness of topic models. One of these criteria is coherence. But the problem with coherence is that it does not take into account most of the text in the calculation but only 10 most popular words, which makes evaluating the quality of the topic by this critera unreliable. Alternative approach called The aim is to propose a new method for calculating coherence that takes into account the distribution of the topic throughout the text.

Keywords topic modeling · topic coherence · topic interpretability · topic model · BigARTM · text analysis

1 Introduction

Topic modeling is a text data analysis method that automatically identifies hidden topics in large collections of text data. Topic models are used in information retrieval [9], documents' categorization [11], social networks' data analysis [8], recommendation systems [10], exploratory search [3] and other areas not related to texts [5].

Interpretability is a key characteristic of an useful topic model [2]. But interpretability of the topic model is a poorly formalized requirement. Informally, it means that according to the lists of the most frequent words of the topic, the expert can understand what this topic is about and give it an adequate name. Expert approaches are necessary at the research stage, but they make it difficult to automatically build good topic model.

It was previously shown [6] that among the quality criteria calculated automatically from a collection, coherence or consistency correlates best with expert estimates of interpretability. However, the existing methods for calculating coherence have a "fundamental limitation" [1]. These methods consider the distribution of only a small subset of words, resulting in a significant loss of accuracy. As an alternative, an approach known as Intra-text coherence has been proposed [1], which considers all words but rejects the concept of PMI.

This study aims to enhance coherence calculation techniques by investigating new quality criteria for topic models that incorporate both the distribution of topics across the text and the concept of PMI as a measure of word coherence. The research compares these new criteria with existing methods and proposes a methodology for assessing the interpretability of topics.

2 Problem statement

2.1 Introduction to topic modeling [12]

Let D denote a set (collection) of texts and W denote vocabulary, which is a set of all terms from these texts. Each term can represent a single word as well as a key phrase. Each document $d \in D$ is a sequence of n_d terms $(w_1, \ldots, w_n)$ from the vocabulary W. Each term might appear multiple times in the same document.

Assume that each term occurrence in each document refers to some latent topic from a finite set of topics T. Text collection is considered to be a sample of triples (w_i, d_i, t_i), $i = 1, \ldots, n$ drawn independently from a discrete distribution $p(w, d, t)$ over a finite space $W \times D \times T$. Term w and document d are observable variables, while topic t is a latent (hidden) variable. Following the "bag of words" model, we represent each document by a subset of terms $d \subset W$ and the corresponding integers n_{dw}, which count how many times the term w appears in the document d.

Conditional independence is an assumption that each topic generates terms regardless of the document: $p(w \mid t) = p(w \mid d, t)$. According to the law of total probability and the assumption of conditional independence

$$p(w \mid d) = \sum_{t \in T} p(w \mid t)\, p(t \mid d) \tag{1}$$

The probabilistic model (1) describes how the collection D is generated from some distributions $p(t \mid d)$ and $p(w \mid t)$. Learning a topic model is an inverse problem: to find distributions $p(t \mid d)$ and $p(w \mid t)$ given a collection D. This problem is equivalent to finding an approximate representation of counter matrix

$$F = \big(\hat{p}_{wd}\big)_{W \times D}, \quad \hat{p}_{wd} = \hat{p}(w \mid d) = \tfrac{n_{dw}}{n_d} \tag{2}$$

as a product $F \approx \Phi\Theta$ of two unknown matrices—the matrix Φ of term probabilities for the topics and the matrix Θ of topic probabilities for the documents:

$$\begin{array}{llllll} \Phi & = (\phi_{wt})_{W \times T}, & \phi_{wt} & = p(w \mid t), & \phi_t & = (\phi_{wt})_{w \in W} \\ \Theta & = (\theta_{td})_{T \times D}, & \theta_{td} & = p(t \mid d), & \theta_d & = (\theta_{td})_{t \in T} \end{array}$$

2.2 Coherence

A topic is called coherent if the most frequent topics of a given topic are often found side by side in collection documents. The average coherence of topics is considered a good measure of the interpretability of a thematic model [6].

$$n(u, v) = \sum_{d=1}^{|D|} \sum_{i=1}^{N_d} \sum_{j=1}^{N_d} [0 < |i - j| \le k][w_{di} = u][w_{dj} = v]$$

$$n(u) = \sum_{w \in W} n(u, w)$$

$$n = \sum_{w \in W} n(w)$$

$$p(u) = \tfrac{n(u)}{n}$$

$$p(u, v) = \tfrac{n(u,v)}{n}$$

$$\mathrm{PMI} = \log_2 \tfrac{p(u,v)}{p(u)p(v)}$$

$$\mathrm{coh}_{t_0} = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} \mathrm{PMI}(w_i, w_j)$$

Despite the fact that this approach shows good correlation results with human scores, we will improve it by increasing the number of words that are taken into account when calculating indicators.

2.3 Weighted coherence

To generalize previous method let us define weighted coherence as

$$\text{coh}_{t_0} = \frac{\sum_{u,v} \text{rel}_{t_0}(u,v)\text{coh}(u,v)}{\sum_{u,v} \text{rel}_{t_0}(u,v)}$$

Our objective is to identify functions $\text{rel}_t(u,v), \text{coh}(u,v)$ that exhibit the strongest correlation with human evaluations of topic consistency with topic interpretability.

2.4 Consistency with word chains

As previously stated, topic interpretability is a vaguely defined concept. For the purposes of this article, as measure of interpretability we will use consistency with marked chains.

Let C_{di} represent i-th word chain from document d. A word chain is a subset of a document consisting of connected words from the same topic.

$$p(t|C) = \sum_{w \in C} p(t|w)p(w|C) = \text{mean } p(t|w)$$

,

where $p(t|w) = p(w|t)\frac{p(t)}{p(w)} = \phi_{wt}\frac{n_t}{n_w}$

$C(t) = \{C_{di} \mid t = \text{argmax}_t p(t|C_{di})\}$ - set of word chains that are consisted to topic t.

$$\text{cons}_{t_0} = \text{mean}_{C_{di} \in C(t_0)}\ p(t|C_{di})$$

So main idea of consistency with word chains is that instead of estimate each topic of topic models we build perfect topic model and compare it's text markup with markup of our model.

3 Computational experiment

3.1 Data and frameworks

1. 20 Newsgroups Dataset: This dataset comprises documents from 20 different newsgroups, with each file containing one document per newsgroup.
2. Small dataset of word chains builded manually from previous dataset.

Additionally, it is essential to identify and extract word segments from certain documents to assess topic interpretability. Acquiring these specific data segments poses a challenge for the experiment due to the lack of a defined source.

3.2 Plan of experiment

1. Construct a topic model.
2. Compute the metrics $\text{coh}_t^{(i)}$ for analysis.
3. Determine the Spearman correlation between the consistency of word chains cons_t and the values of metrics $\text{coh}_t^{(i)}$.

We will create topic model using the TopicNet library, which is a high-level framework for BigARTM. Model consists of 11 topics (10 main and 1 background). Modality which was used to train model and compute metrics is 'lemmatized'.

3.3 Theory

The main part of our experiment is to explore different options of rel_t (relevance), coh (coherence).

Coherence is responsible for the connectedness of words, that is, the non-random occurrence of these words in the same context. Let's take PMI [6] as the basis for coherence. PMI (especially in its positive pointwise

mutual information variant) has been described as "one of the most important concepts in NLP" [4]. Originally, PMI has "a well-known tendency to give higher scores to low-frequency events" [7]. So we will use some modidications of it.

- PPMI = $(\text{PMI})_+$
- NPMI = 1 - $\frac{\text{PMI}}{log_2(p(u,v))}$
- $\text{PMI}^k = \log_2 \frac{p^k(u,v)}{p(u)p(v)}$, where $k = 2, 3$

Relevance is an indicator that describes how well a pair of words (u, v) correspond to a certain topic. So we will look for a symmetric function depending on the probabilities ϕ_{ut} and ϕ_{vt}.

Let $\text{pos}_t(u)$ denote number of ϕ_{ut} in variational series of ϕ_{*t} and $\text{pos}_u(t)$ denote number of ϕ_{ut} in variational series of ϕ_{u*}.

- $[\text{pos}_t(u) \leq k][\text{pos}_t(u) \leq k]$, where $k = 10, 20, 50, 100$. That variant responds to Newman's coherence.
- $\phi_{ut}\phi_{vt}$
- $\sqrt{\phi_{ut}\phi_{vt}}$
- $[\phi_{ut}\phi_{vt} \geq \varepsilon]$
- $[\phi_{ut} > 0][\phi_{vt} > 0](\phi_{ut} + \phi_{vt})$

3.4 Results of experiment

coh	rel	correaltion
PMI	$[\text{pos}_t(u) \leq 10][\text{pos}_t(u) \leq 10]$	0.62
PMI	$[\text{pos}_t(u) \leq 20][\text{pos}_t(u) \leq 20]$	0.49
PMI	$\phi_{ut}\phi_{vt}$	0.52
PPMI	$\phi_{ut}\phi_{vt}$	0.88
NPMI	$\phi_{ut}\phi_{vt}$	-0.52
PMI	$\sqrt{\phi_{ut}\phi_{vt}}$	0.54
PPMI	$\sqrt{\phi_{ut}\phi_{vt}}$	0.42
NPMI	$\sqrt{\phi_{ut}\phi_{vt}}$	-0.55
PMI	$[\phi_{ut} > 0][\phi_{vt} > 0](\phi_{ut} + \phi_{vt})$	0.82
PPMI	$[\phi_{ut} > 0][\phi_{vt} > 0](\phi_{ut} + \phi_{vt})$	0.68
NPMI	$[\phi_{ut} > 0][\phi_{vt} > 0](\phi_{ut} + \phi_{vt})$	-0.83
PMI	$[\phi_{ut}\phi_{vt} \geq \varepsilon]$	0.32
PPMI	$[\phi_{ut}\phi_{vt} > \varepsilon]$	0.33
NPMI	$[\phi_{ut}\phi_{vt} \geq \varepsilon]$	-0.32

Table 1: Correaltion of metrics

4 Conclusion

This paper introduce an alternative method of topic model's interpretability measuring. As computational experiment shows, metrics that based on weithed coherency demonstrate a higher correlation with the consistency of word chains compared to the traditional top word coherence.

References

[1] Vasily A Alekseev, Vladmir G Bulatov, and Konstantin V Vorontsov. Intra-text coherence as a measure of topic models' interpretability. In Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pages 1–13, 2018.

[2] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, 22, 2009.

[3] Anastasia Ianina, Lev Golitsyn, and Konstantin Vorontsov. Multi-objective topic modeling for exploratory search in tech news. In Artificial Intelligence and Natural Language: 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers 6, pages 181–193. Springer, 2018.

[4] Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

[5] Massimo La Rosa, Antonino Fiannaca, Riccardo Rizzo, and Alfonso Urso. Probabilistic topic modeling for the analysis and classification of genomic sequences. BMC bioinformatics, 16:1–9, 2015.

[6] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn, editors, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 100–108, Los Angeles, California, June 2010. Association for Computational Linguistics.

[7] François Role and Mohamed Nadif. Handling the impact of low frequency events on co-occurrence based measures of word similarity-a case study of pointwise mutual information. In International Conference on Knowledge Discovery and Information Retrieval, volume 2, pages 218–223. Scitepress, 2011.

[8] Devesh Varshney, Sandeep Kumar, and Vineet Gupta. Modeling information diffusion in social networks using latent topic information. In Intelligent Computing Theory: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings 10, pages 137–148. Springer, 2014.

[9] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. Information Retrieval, 16:331–368, 2013.

[10] Jian-hua Yeh and Meng-lun Wu. Recommendation based on latent topics and social network analysis. In 2010 Second International Conference on Computer Engineering and Applications, volume 1, pages 209–213. IEEE, 2010.

[11] Shibin Zhou, Kan Li, and Yushu Liu. Text categorization based on topic model. International Journal of Computational Intelligence Systems, 2(4):398–409, 2009.

[12] Константин Воронцов. Вероятностное тематическое моделирование: теория регуляризации artm и библиотека с исходным кодом bigartm. МВ Ломоносова, 2023.