WEIGHTED COCHERENCE 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 thematic 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, which makes evaluating the quality of the topic by this critera unreliable. 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 · machine learning

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, document categorization, social networks data analysis, recommendation systems, exploratory search and other areas.

Interpretability is a key characteristic of an effective topic model [1]. 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 [2] that among the quality criteria calculated automatically from a collection, consistency or coherence correlates best with expert estimates of interpretability. However, the previously proposed methods of calculating coherence have a fundamental limitation. They take into account the distribution of only a very small number of words, which leads to a significant loss of accuracy.

This study aims to advance coherence calculation techniques by exploring new quality criteria for topic models that take into account the distribution of topics throughout the text. 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

Let D denote a set (collection) of texts and W denote a set (vocabulary) 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

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

The probabilistic model (1) describes how the collection D is generated from the known 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

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

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:

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

2.2 Weithed coherency

Let us define weighted coherency as

$$coh_{t_0} = \frac{\sum_{u,v} rel_{t_0}(u,v) coh(u,v)}{\sum_{u,v} rel_{t_0}(u,v)}$$

Our objective is to identify functions $rel_t(u, v), coh(u, v)$ that exhibit the strongest correlation with human evaluations of topic interpretability. As previously stated, topic interpretability is a vaguely defined concept. For the purposes of this article, interpretability will be defined as follows.

Let C represent the set of word chains. A word chain is a subset of a document consisting of connected words from the same topic. Unfortunately, there is no more precise definition available.

So topic interpretability of t_0 will be

$$I_{t_0} = \sum_{c \in C} \left(\sum_{w \in c} \log \phi_{wt_0}\right) [t_0 = argmax_{t \in T} \sum_{w \in c} \log \phi_{wt_0}]$$

3 Computational experiment

3.1 Data

1. 20 Newsgroups Dataset: This dataset comprises documents from 20 different newsgroups, with each file containing one document per newsgroup.

2. Lenta.Ru News Dataset: This dataset consists of news articles sourced from the website Lenta.ru.

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. Utilize TopicNet to construct a topic model.

2. Compute the metrics $coh_t^{(i)}$ for analysis..

3. Determine the Spearman correlation between the interpretability of topics I_t and the values of metrics $coh_t^{(i)}$.

References

[1] Wang C. Boyd-Graber J. L. Blei D. M. Chang J., Gerrish S. Reading tea leaves: How humans interpret topic models.

[2] Grieser K.-Baldwin T. Newman D., Lau J. H. Automatic evaluation of topic coherence.

1. Аннотация сейчас не очень хорошо написана. Попробуйте переписать по плану,
который давался на занятии.
Подумайте, как эта аннотация поможет исследователям, которые читают вашу работу

2.Здесь нужна ссылка

3. На каждый пример применения также нужна ссылка

4. Место между секцией и подсекцией обычно заниамается мотивационным текстом:
как устроена секция, зачем читателю читать каждую часть

5. w

6. Старайтесь не разбивать формулы.
Если формула значима, ее можно выключить (отобразить по центру).
Это делается вот так: $$x=y$$ или так \[x=y\]

7. Это точно долнжа быть выключенная формула

8. Между словами и тире как минимум нужен пробел.
Но вообще в английском языке свои правила пунктуации и я не уверен что здесь можно ставить тире.
Проверьте по учебнику или замените на двоеточие (это более безопасный вариант)

9. Функции rel, coh пишите как rel и coh.
Формула - это элемент предложения, после нее обычно ставится запятая или точка.