
Topic Modeling with 20 Newsgroups Dataset

Harshit Rawal

Department of Computer Science
Technical University of Kaiserslautern
Germany
`rawal@rhrk.uni-kl.de`

Abstract

In statistics and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. In this portfolio exam, there are two datasets, 20 Newsgroup original and a modified version. There are 20 distinct topics in the original dataset, but a modification in the later dataset must be identified using topic modeling techniques. There are two techniques to be utilized, a parametric and a non-parametric topic model.

1 Introduction

20Newsgroups text corpora is a dataset comprising around 18000 newsgroups posts on 20 topics. In the problem statement, there is also a variation of this dataset. Both datasets will be referred to as *ori* and *mod* respectively, from now on. Analysis of the two datasets and, thereby, identification of the core difference between the two need to be done using topic models. Topic models are a form of unsupervised learning technique employed to get the hidden semantic structures within the texts to provide helpful information relating to the task at hand.

Each parametric and non-parametric topic model must be used to identify the hidden difference between *ori* and *mod*. The Latent Dirichlet Allocation or LDA has to be used for the parametric topic model. For the non-parametric model, the chosen model is Principal Component Analysis or PCA. In order to make both the models consistent and comparable, the scikit-learn python library was used. The code and implementation of this portfolio exam can be found here at

<https://gitlab.rhrk.uni-kl.de/rawal/pgm-exam-2022>

Instructions on the usage and documentation of the code can be found within the repository.

2 Problem Setup

On initial inspection of the dataset, finding out the total count of documents in each topic reveals that the significant difference between the two datasets is that two of the topics, namely *rec.sport.baseball* and *rec.sport.hockey*, both have 80% fewer articles/documents than original. So, ideally, the topic model analysis needs to concisely identify this difference within both datasets. Hence, in this approach, the assumption is that there would be a metric calculated within the data of the two topics that must have a high deviation between *ori* and *mod*. In other words, both datasets will be passed through two models, parametric and non-parametric, and the comparison metric shall display the modified topics as outliers to the general trend.

3 Experiment

Preprocessing The documents in the dataset have headers mainly in email form and contain many other metadata elements that are not needed for the model performance. Hence these metadata blocks are removed in text cleaning before performing any meaningful transformation to the dataset. Other text cleaning transformations include lowering cases, stripping unnecessary space, tokenization, removing stopwords, and lemmatization. Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma or dictionary form. So, for example, 'dance', 'dancing', and 'dancer' will all transform into the lemma 'dance'.

TF-IDF The TF-IDF[3] transformation technique vectorizes the text, which means the term frequency-inverse document frequency. Term frequency, $tf(t,d)$, is the relative frequency of term t within document d ,

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

where $f_{t,d}$ is the raw count of a term in a document, i.e., the number of times that term t occurs in document d . Although a simple dictionary from corpora can also be used to give higher importance to the tokens that matter, TF-IDF is used.

Latent Dirichlet Allocation [2] is a generalization of Fisher's linear discriminant, a method used to find a linear combination of features that characterizes or separates two or more classes of objects or events. The LDA is used as the parametric model for topic modeling. LDA focuses on finding a feature subspace that maximizes the separability between the groups.

Principal Component Analysis [1] The principal components of a collection of points in a real coordinate space are a sequence of p unit vectors, where the i -th vector is the direction of a line that best fits the data while being orthogonal to the first $i-1$ vectors. The PCA is used as the non-parametric model for the topic modeling. PCA ignores the class label and focuses on capturing the direction of maximum variation in the data set.

Metrics The idea is first to process the data using the two models and get the latent space as the features to determine any abnormality between the topics based on specific criteria. Here topic variance and cosine similarity[4] are the metrics based on which the hope is to see the outliers when comparing *ori* and *mod*.

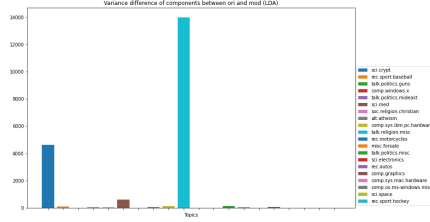
Here topic variance is defined as the total average variance of each latent space component within that particular topic group. Similarly, the cosine distance is calculated amongst the generated features within the topic group for cosine similarity. Then topic variance and cosine similarity are compared for *ori* and *mod*. Again, the expectation is that the modified topics would appear as outlier attributes.

Topic modeling depends on data, and if the aim is to identify the topic groups with variable data sizes, then the assumption is that the topic group with the smaller data size will have high variance compared to having an equal number of data points. It is because feature representation of that topic will be low since other features will overcome it from other topic groups due to their size.

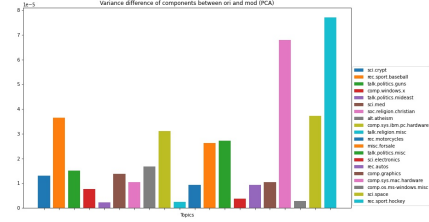
Similarly, cosine variance will also act as an ideal metric to detect topic groups with variable data sizes than the norm. The idea here is that due to the scarcity of relevant features due to their lower data size, the cosine distance will vary significantly compared to a dataset where the same topic group has a higher data size.

4 Results

The code is implemented in python using the same library for as many tasks as possible to maintain consistency. Therefore, based on the assumptions, it should be clear that the topics where a modification occurred appear as outliers based on their respective metrics. However, executing the code showed that only one of the metrics could generate the expected result. Therefore, topic variance

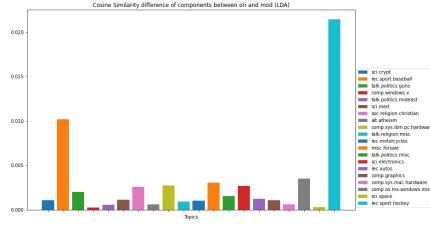


(a) LDA

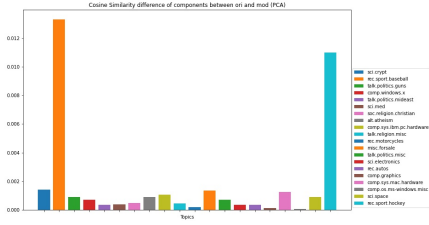


(b) PCA

Figure 1: Difference between component variance within topics of *ori* and *mod*



(a) LDA



(b) PCA

Figure 2: Difference between Cosine similarity within topics of *ori* and *mod*

could not be tested positively for the hypothesis for both models. However, the hypothesis shows successful results with cosine similarity as the metric.

Due to resource constraints, TF-IDF vectorization with a maximum of 10K words was used. However, the best results are to be expected at about 100K. Furthermore, for the LDA model, the number of components used is restricted to 19 as per the condition $\min(classes - 1, features)$. As a result, the training data's cumulative explained variance was close to 1 for both *ori* and *mod* datasets. Again, this is to be expected in the case of LDA. However, for the PCA model, the components used were 200 (which was empirically chosen along with resource constraints). The cumulative explained variance was lower than expected, close to 0.2% for both datasets.

4.1 Topic Variance

Here, as per the hypothesis, the modified topics must have become outliers, but this does not hold as seen. As seen on figure 1 (a) for LDA, the difference between the datasets incorrectly gives us *talk.religion.misc* and *sci.crypt* as the outlier topics. As seen on figure 1 (b) for PCA, the most significant outlier in the true outlier, i.e., *rec.sport.hockey*, but the rest is incorrect.

4.2 Cosine Similarity

The hypothesis does hold successfully for this metric. The cosine similarity within the topic groups for the *mod* dataset differs from the *ori* dataset. Hence the difference is much more evident, as seen in figures 2 (a) and (b). Both the desired topic groups appear as outliers i.e. *rec.sport.hockey* and *rec.sport.baseball*. On further analysis, the difference certainly has to do with different TF-IDF vectors learned due to differences in data sizes. Hence different priorities were given to features.

5 Further Improvements

This section proposes alternative methods to overcome the performance variations due to dataset size dissimilarity. Since topic models are unsupervised, they thoroughly depend on the data itself rather than the ground truth. Due to this, alternative techniques that improve the separability of the data must be considered.

TF-IDF is directly used on the dataset. Here further improvements can be made. For example, using TF-IDF on bigrams and trigrams may help push the more relevant part of the data further upstream to be considered essential for a particular topic group. Furthermore, generating TF-IDF based on inter topic dataset and weighing it with global TF-IDF may incorporate more detailed features that might have been previously discarded.

Besides this, providing weights to the individual topic elements based on their respective sizes may generate equally essential features. Here several libraries are provided to send this information to varying topic groups, at least for LDA.

Word2vec and other transformer-based models are recommended to generate features contributing to more relevant information being passed to the model. Transformer-based models are known for utilizing context instead of just frequency. Redundant data may have been sent using TF-IDF, where words with similar context still appear highly valued and used as features. On the contrary, even the rare words in vocabulary with similar context may still be learned in context-based models that fail to show up in TF-IDF-based features due to their low frequency.

6 Conclusion

Due to the unsupervised nature of topic models, the features matter significantly in providing an accurate result. If the topic groups are unevenly distributed, this may result in discarding features essential to topic groups with smaller dataset sizes. Due to this, both parametric and non-parametric models will learn features that might not be useful to provide the desired topic/class separability. Therefore, metrics like cosine distance amongst the varying size dataset can highlight the topic groups whose classes may have learned different features due to their training size to the global topic/class.

In the results, the topic variance does not perform well; it may be because if a topic group has a small data size, the features used to describe it will be less diverse. Another explanation could be that variance with the topic group is irrelevant to the data sizes. Perhaps, certain topic groups naturally have high variance, which overshadows the variable data sizes. Also, the notion of model learning completely different features and hence having high variability between its two versions of topic groups with variable data size seems to hold as seen with cosine similarity. Since cosine similarity depends on the feature vectors, having different weights for these vectors based on what was learned gives a high difference in this metric.

References

- [1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.
- [2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [3] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.
- [4] Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1, 2012.