Topic Modeling for Single Documents

Isabela Lago
isabelal@buffalo.edu

1 Introduction

Topic modeling is the process of finding groups of words that accurately describe the contents of a corpus, then classifying the documents in the corpus according to the found topics. Latent dirichlet allocation, or LDA, is the current standard of topic modeling. LDA is very accurate, but some limitations exist. The number of topics to be found has to be decided beforehand, and it may be more difficult to implement in cases where the size and variety of the corpus are unknown, or if there is no corpus to work with. Also, the outputs of LDA are unlabeled topics, so it takes extra work to create a clear interpretation of the topic.

Few studies have applied topic modeling on single documents, but it may be an appropriate alternative if the problems described occur. In this project, topic modeling is applied on single documents using LDA and two original approaches: k-medoids and average cosine similarity. The goal is to retrieve words that accurately describe the contents of single documents.

2 Approaches

The purpose of topic modeling is to find words that most accurately describe the contents of a document, so these two new approaches find the words in a document that are semantically the most similar to the rest of the text. To do this, a word2vec model is created for the document, and the new approaches vary in how the word2vec models are used.

2.1 K-Medoids

K-medoids are a clustering algorithm that use members of the data set as cluster centers. This is necessary when working with word embeddings because the exact centers of the clusters will not correspond to word vectors. The k-medoids approach only works in 2-dimensional space, so Principal Component Analysis is used to transform the multi-dimensional word vectors into 2 dimensions. The output is two words corresponding to the vectors in the cluster centers.

2.2 Average cosine similarity

Cosine similarity is used for comparing semantic similarity between two word vectors. The words that have the highest cosine similarity compared to all the other words in a document may be the most descriptive. For each vector in the word2vec model, the cosine similarity is found between that vector and every other vector in the model, and the average of the cosine similarities is taken. The two words with the highest average cosine similarity are the found topics.

2.3 LDA

LDA is the current standard in topic modeling. It is a statistical model that organizes words into a user-specified number of topics. Instead of a word2vec model, a doc2bow model is used. For this project, two topics are found, and the top word from each topic is used as the topic label.

3 Evaluation

The three methods were tested on the CMU Movie Summary Corpus (Bamman et al., 2013), which contains over 42,000 movie plots. This corpus is used because of the variety of the documents. The texts were tokenized with the NLTK and lemmatized with the WordNet Lemmatizer before the methods were applied, and the word2vec and doc2bow models were made with gensim.

The following example shows a document from the CMU Movie Summary Corpus and output for each method:

A Miss Galaxy competition is hijacked

by a gang who take a number of hostages and demand diamonds as a ransom. However, the host, Sharon Bell, a kick-boxing actress, is determined to stop them.

K-medoids: ransom, stop

Cosine similarity: number, diamond

LDA: diamond, take

	Mean	Coherence
K-medoids	2.36	.116
Cosine similarity	2.83	.105
LDA	**3.66**	**.464**

3.1 Human Judgement

For a subset of 30 summaries, volunteers rated the accuracy of each approach on a 5-point Likhert scale (1 = not very relevant, 5 = very relevant). The LDA model had the highest average rating, and the Mann-Whitney U test showed there was no significant difference between the k-medoids and cosine similarity models (p = .089).

3.2 Topic Coherence

Topic models can be evaluated automatically using topic coherence measures. A subset of 1500 summaries and results are evaluated with the normalized pointwise mutual information (NPMI) method, using the toolkit from Han Lau et al. (2014). LDA performed the best on this method, and the other two approaches were roughly similar.

4 Discussion

Based on visual inspection, LDA used of characters as topics more often than k-medoids and cosine similarity. The success of LDA could depend on how well human judges think character names are as topics.

5 Conclusion

The LDA model worked best for single documents, based on the results of both the human evaluation and the topic coherence measure. The k-medoids and cosine similarity methods performed similarly in both evaluation metrics.

This study could be expanded by adding a baseline of two random words in the summary. That way, the effectiveness of k-medoids and cosine similarity can be measured. It would also be interesting to see a comparison of how corpus-level LDA would perform on the CMU Movie Summary Corpus. However, the topics need to be labeled before it can be evaluated.

These methods could be applied to other areas in text mining such as document summarization and theme extraction.

References

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 530–539.