Using a Latent Dirichlet Allocation (LDA) Model for Topic Modeling On Caption Contests

Alex Cheung

August 2023

1 Introduction

Natural language processing (NLP) is is the ability of a computer program to understand human language as it is spoken and written and with the rapid accumulation of electronic data, requires techniques and tools that deals with automatically organize, summarize and understand a large collection of textual information. In recent years, topic modeling has grown in popularity due in part because of its ability as a statistical model to find hidden topical patterns of words in a document collection. Topic modeling can be described as a technique for finding a collection of words (topic) from a group of documents that represents the information in the group.

Topic modeling algorithms allow us extract hidden knowledge or meaningful information from text data and in turn is a powerful tool that can model objects as latent topics that can reflect meaning of the collection of document (Barde and Bainwad, 2017). The two most common approaches for topic analysis with machine learning are topic modeling and topic classification. Topic modeling because it is an unsupervised method of inferring topics allows us to unlock semantic structures within texts to extract insights and help make data-driven decisions (Meddeb and Romdhane, 2022). On the other hand, classification is a supervised approach that leverages pre-defined text classifiers to assign a label to a document based on its content, so there is no additional insight to be gained.

One method of applying topic modeling to text data is through using the Latent Dirichlet allocation (LDA) model which is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random vectors over latent topics, where each topic is characterized by a distribution over words (Xie et al., 2018).

2 Aims of the analysis

The goal of this analysis is to use a LDA topic model to extract topic vectors from captions of multiple contests from the New Yorker Caption Contest.

Specifically, we want to see if there are any common topics across various contests and find cluster of words that appear frequently within the documents. By using a LDA model, we can visualize how the topics are clustered around each other and what words constitute the meaning of a single topic.

3 Design

We decided to perform topic modeling in Python by using the spacy, gensim, and pyLDAvis libraries because it is the most efficient computational method to extract topic vectors from the data and visualize them. We pulled down data of 5 contests from a SQL database. Before doing any modeling or text analysis, the action of pre-processing text is important because it creates a uniform text database. By removing non-meaningful information at the beginning such as punctuation and numbers, it becomes easier to extract semantic meaning from the text without the noise. Afterwards, we lemmatize the text to their base form instead of stemming because we want to keep the semantic meaning of the words. Because we lemmatize the text before tokenizing and removing stopwords, the analysis process becomes more efficient due to having a smaller vocabulary. We then fit our LDA model onto the data and extract its topic vectors which we can visualize the top 50 words in each topic through wordclouds.

4 Conclusion

By performing topic modeling on a subset of caption contests, we are able to see individual topic vectors and the words that comprise those topics. We can also see if there is commonality between certain topics. Topic modeling has allowed us to extract valuable information about the semantic meaning behind the text which can be used to identify patterns among words.

There are certain limitations to this analysis in that we only worked with a small sample size of the contests and did not fit the model to the whole dataset purely due to computational limits. If one were to try to perform topic modeling on the entire dataset, they would need to have greater computational power. Another limitation would be this is only one type of topic model, there are many other models that can used to get various results such as Top2Vec, Doc2Vec for caption vectors, and Latent Semantic Indexing (LSI). Future directions of this analysis should include more contests for greater generalizability purposes and use different models to see how topic vectors differ in each model.