Dishti Dave CWID- 10451780

Dyuti Dave   CWID- 10453480

Mayur Kotian CWID- 10449990

CS 557 Natural Language Processing

Final Project

**Latent Dirichlet Allocation on NewsGroups**

**Abstract:**

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. To the best of our knowledge, it was originally collected by Ken Lang, probably for his Newsweeder: Learning to filter netnews, though he does not explicitly mention this collection. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

**Approach and Uniqueness:**

Understanding conversations found in news groups is a challenging task as they are often interwoven between other speakers and other topics. A solution is to focus on the topics that are found within each conversation.

In natural language processing, latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar [2]. In layman’s terms, LDA is a method of automatically discovering topics and their probabilities from a given corpus.

After running this analysis on 20 groups, the top 10 most occurring words within each topic as well as the probability of finding these topics in the corpus are given. By using an analyst’s intuition and experience in the detection process, topics that were not understandable were eliminated.

LDA is used to classify text in a document to a topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

* Each document is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words.
* LDA assumes that every chunk of text we feed into it will contain words that are somehow related. Therefore, choosing the right corpus of data is crucial.
* It also assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution.

Finally, the probabilities found in each topic were mapped over time. Trends were discovered and were found to be significant after comparing major events found in the document.

**Data Collection and Analysis:**

This involves the following:

* **Tokenization**: Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.
* Words that have fewer than 3 characters are removed.
* All **stopwords** are removed.
* Words are **lemmatized** — words in third person are changed to first person and verbs in past and future tenses are changed into present.
* Words are **stemmed** — words are reduced to their root form.

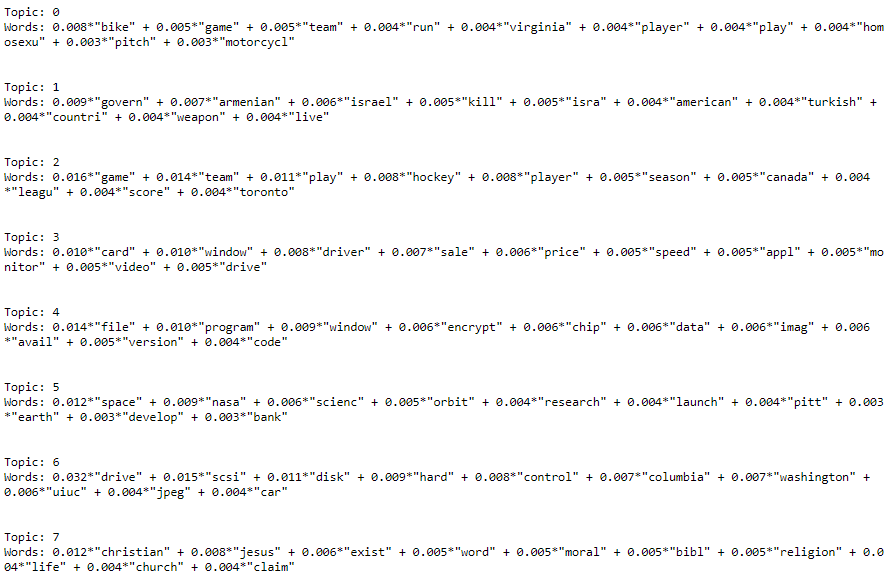
Prior to topic modelling, we convert the tokenized and lemmatized text to a bag of words — which you can think of as a dictionary where the key is the word and value is the number of times that word occurs in the entire corpus.

We can further filter words that occur very few times or occur very frequently.

Now for each pre-processed document we use the dictionary object just created to convert that document into a bag of words. That is for each document we create a dictionary reporting how many words and how many times those words appear.

**Result:**

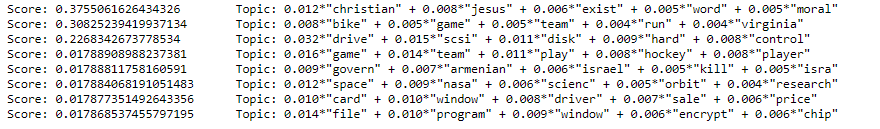
After running the file, the topic scores over given as well as the top 8 words associated with that topic. Table 1 shows the clear and well-defined topics that were extracted. Below is the list of topics on their probabilities associated with it.



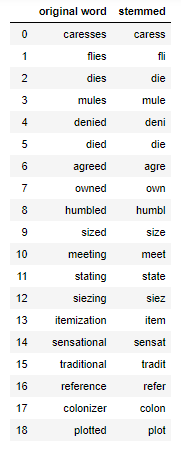
Below are the images of two different text documents(text) and their scores associated where one document is more close so a higher score is observed compared to other.

Figure 1

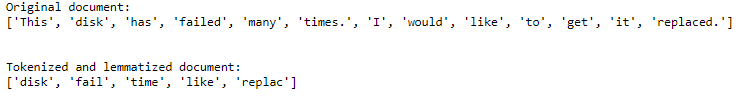
Figure2



Further, training document of the 20 newsgroup is being stemmed and tokenize. Below is the image of stemmed words



The extraction of tokenized and lemmatized word list is show below .



**References:**

**[2]Wikipedia contributors. "Latent Dirichlet allocation." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 14 Oct. 2019. Web. 31 Oct. 2019.**

**[3]    Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica, 57, 357–384.**