Name: Morgan Gere

Course: IST 736- Text Mining

Date: 6/03/2022

**Topic Modeling**

**Introduction**

Latent Dirichlet Allocation (LDA) is an unsupervised classification. It is a way to find the main topics of a document in a set of documents. It works under the assumption that documents pertaining to the same topic will use the same words. It uses the probability distribution of words in a text to map the topics (“Latent Dirichlet Allocation.”). Using 429 documents from the 110th Congress debates floor (House only) 10 topics consisting of the top 10 words will be created. This will be accomplished by vectorizing the text and using sklearn’s LatentDirichletAllocation.

**Method**

The folders obtained were placed together manually outside of python. The folders were read into python with the code provided in the sklearn code. The number of files read in was printed to ensure all files were uploaded into the dictionary.

Text

Description automatically generated

The dictionary was broken into its values (the text) and its keys (the file name), each was placed into a separate list. The text was manually inspected for preprocessing requirements.

Text

Description automatically generated

There was still code that needed to be removed (/n, <DOC>, etc.) Using regular expressions these items were selected and replaced with nothing. A section of the fixed text is also shown below.

Text

Description automatically generated

Upon completing the first round of LDA it was found that many of the topic words selected were states, people’s names, and either democrat or republican. These did not seem to be what the overall topic should have been. Likely, Congressmen talking about the state they are from and how the actual topic effects that state. These were added a stop words list. These were then added to the sklearn English stop words.

Text

Description automatically generated

Count Vectorizer from sklearn was used to vectorize the text. The number of features selected as max was 1000, the stop words list created was used to remove the stop words, and the min document frequency was set to 2. The documents were fit and transformed, and the feature names were obtained and placed into a numpy array.

Text

Description automatically generated

The LDA model was selected for 10 topics. The max iterations were raised to 10. The LDA model was used to fit and transform the term frequency vectorized data.

Graphical user interface, text, application

Description automatically generated

The function from the sklearn learning code provided was used selecting the number of top words used in each topic to be 10. This could allow for a smaller more coherent understanding of what the model is doing. The function was then used on the model with the feature names numpy array obtained earlier.

Text

Description automatically generated

Each text was placed into a row in a pandas data frame. Each column is one of the topics and the probability distribution associated with that text. The last column added was the name of the text for an easy way to find the index.

Text

Description automatically generated with medium confidence

The pandas data frame was used to create a list for each row and these lists were placed inside another list to be used to visually inspect of any text and its probability distribution of the topics.

Graphical user interface, text, application

Description automatically generated

A function was created that took the index number from the row list (same as the pandas data frame) and plotted the topics along the x and the probability distribution on the y. It also displays the text name for a title. This will allow for easy visual inspection of the topics for each text.

Text

Description automatically generated

**Results**

Two examples of the graph will be displayed that were created using the function along with the topic top words.

Shape

Description automatically generated with medium confidence

Chart, bar chart

Description automatically generated

Text

Description automatically generated

**Conclusion**

The top words for each topic show somewhat overall ideas. Topic 0 is money, topic 1 is education/science, topic 2 resources (oil/gas/coal) topic 3 is rules, topic 4 veterans and their pursuits after military, topic 5 current military, topic 6 is medical, topic 8 is defense and military strategy, topic 9 economy and topic 7 is not relevant but may be some sort of filibuster. A more in-depth look would be required to gain information on topic 7.

Graph number 1 is text number 5 and shows the speaker is mostly discussing resources, this is easily seen. While graph number 2 (text 1) we can see a more even mix of the topics with a high emphasis on the defense and military strategy.

In conclusion with the correct preprocessing and vectorization LDA can be a powerful tool to extract insightful information into text topics. Lots of information can be gleamed and displayed in easily understandable ways.

**References**

“Latent Dirichlet Allocation.” *GeeksforGeeks*, 6 June 2021, https://www.geeksforgeeks.org/latent-dirichlet-allocation/.