Rohini Shrivastava

IST 736

Homework 8

**Introduction:**

News is [information](https://en.wikipedia.org/wiki/Information) about events happening currently. It can be provided in multiple ways such as [word of mouth](https://en.wikipedia.org/wiki/Word_of_mouth), printing, postal systems, broadcasting, or electronic communication using the internet and social media. Common topics for news reports include war, government, politics, economy, health, etc. It presents new information to the viewers. Originally news travelled by word of mouth or through handwritten letters that were passed between people. The printing press caused a major advance in the transmission of news. With the printing press, people were able to write “newspapers” a lot faster and were able to distribute it to more of the population.

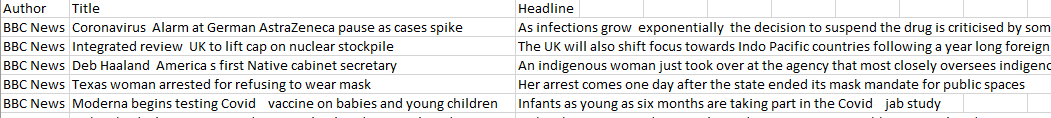
Since the printing press, the biggest project has been moving to online news. In 1983 a newspaper company Knight-Ridder designed a system to send electronic news to readers. It allowed the customers to access the stories before the paper got to their doorstep in the morning. Since the invention of the World Wide Web, more and more companies have been launching their own sites. Some of the pioneers to start the trend were CNN, Chicago Tribune, and a local paper in Raleigh N.C.

BBC (British Broadcasting Corporation) is the world’s leading public service broadcaster. Originally created in the UK, it is translated into more than 40 different languages. It started publishing their papers online in 1997 Along with news articles that are posted online, they also broadcast radio news and televise the news on BBC One. It is UK’s most watched channel and though it is UK-based, the BBC delivers news from all around the world.

**Analysis:**

About The Data:

The data was collected by using a key provided by *NewsAPI,* a simple REST API that returns the results as a JSON. The headlines were collected from BBC (British Broadcasting Corporation). The name of the news source, title of the article and headline of the article were all imported as a JSON and then converted into a list for further consumption. The top trending articles written in English were collected. All the punctuation was removed such as ‘,’, ‘!’, ‘@’, ‘\’, etc. Once the punctuation had been removed, the listed was appended to a CSV file called “News”.



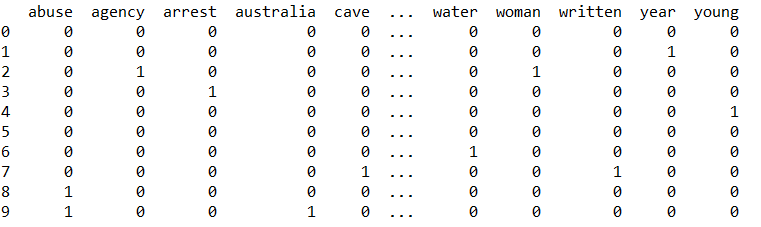
*Figure 1: Sample Headlines Imported into CSV File*

CountVectorizer:

SKLearn is a python programming library that is used for classification, regression, and clustering algorithms. It uses both supervised and unsupervised algorithms. Inside there is a module called CountVectorizer. CountVectorizer converts documents into a matrix of tokens and their counts.

CountVectorizer was used to create a matrix of the CSV News file. The file was imported through the OS library, which looks at the operating system interface, and then cleaned up. The CountVectorizer function was used on the files, and then was converted into a data frame.

The data frame had labels of the CSV as the columns. Once CountVectorizer had been run, a total was then calculated for all the columns and all the rows. Each row contained the number of times the word from the column was present in the review. The column total calculated the total of words used in each review, and the row total calculated the number of times each word was present in the entire corpus. Then columns were sorted based on row totals in increasing sequence (*Figure 2)*. The restaurant reviews had 10 rows and 83 columns.



*Figure 2: Sample of Vectorized Data with CountVectorizer*

LDA:

Latent Dirichlet Allocation(LDA) is a statistical model for topic modeling. Topic modeling is a method of unsupervised classification for documents. It tries to find natural groups of items by organizing, understanding, and summarizing the document.

Once the data had been vectorized, the LDA model was run. The model created 5 different topics for the 10 news articles (*Figure 3*) Another model was run looking at how LDA would work if only 3 different topics were created (*Figure 4).*



*Figure 3: LDA Grouping for 5 Topics*



*Figure 4: LDA Grouping for 3 Topics*

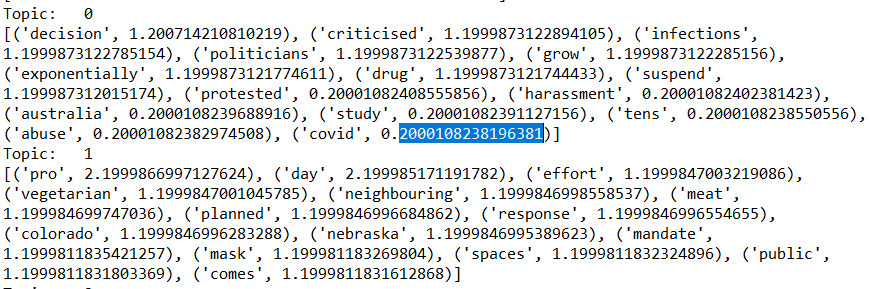
**Results:**

All the reviews combined contained 88 words, not including stop-words. The top used words for restaurant reviews were “day”, “pro” and “comes” at only being used twice. The least used words were all only used once and included words like “meat”, “protested” and “water”. There was little difference between the most and least used words.

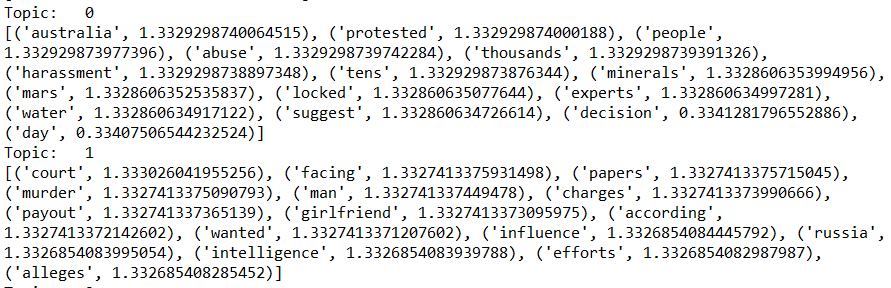
The LDA models only showed patterns but the were not very clear at times. When looking at the 5 Topic groupings (*Figure 3)*, Topic 0 seems to show it was possibly about traveling to mars. Topic 1 was about a sexual crime that was committed, Topic 2 was may have been about policies revolving around the environment, however it was harder to determine. Topic 3 was about politicians in Australia that involves protesting, and the last topic (Topic 5) was also about protests, potentially against new public policies. “Abuse” was a constant word seen in four of the five topics.

When it was run with only three groupings, Topic 0 was about foreign policy review and abuse. Topic 1 was more relating to the united states food industry, and Topic 2 was about COVID and history relating to vaccines. It was harder to determine the modeling of Topic 2.

When looking at the probabilities and comparing the two models, there is some definite overfitting. When changing the number of Topic from five to three, the average for each of the topics went down. In the five topic model, on average it had a value of 1.19 per word (Figure 5). In the second model with three topics, the average went up to 1.33 (Figure 6)



*Figure 5: Sample LDA for Five Models*



*Figure 6: Sample LDA for Three Models*

**Conclusion:**

Topic modelling is a method for finding a group of words from a collection of documents that best represents the information in the collection. In this case, the data is looking at headlines of BBC news articles. There is a large amount of data collected every day. As companies and news channels receive more information, it becomes difficult for users to find what they are looking for. By using topic modeling, these news companies can organize, understand, and summarize the articles they are writing. It can help in determining patterns of what the most accessed articles may be about, how often these patterns are present, and summarizing what is written.

When forming these topic for the headlines gathered, it was showing common themes. Many of the articles talked about abuse, whether it be political abuse or relationship abuse. It also discussed current event topics such as COVID-19 and the vaccinations that come with it. LDA modeling is able to show the different topics that arise in news articles. However, the more topics there are, the easier it is for the individual to understand what the articles are trying to relay.