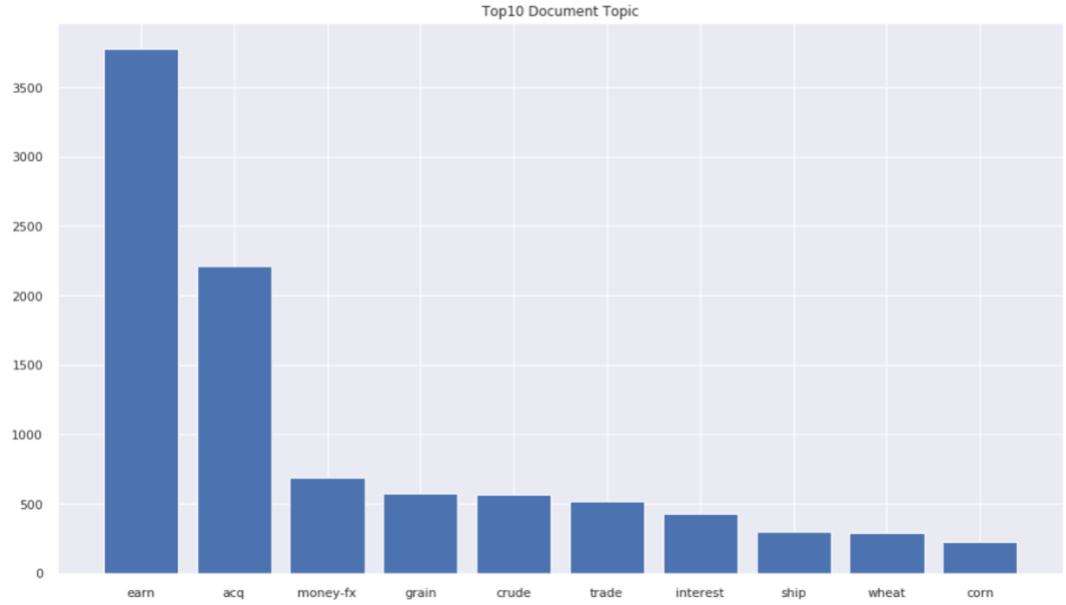
# Data Analytics Assignment – 2

Ufuk Dogan

19.05.2019

## General Information About Data:

I was given text files from Reuters and my topic is clustering them using their text's. In order to achieve this goal, I should have to organized my dataset as a dataframe since my documents were XML files. I have 135 different topics for approximately 20.000 text documents. Top 10 topics of documents are; earn, acq, money-fx, grain, crude, trade, interest, ship, sheat and corn.




## Pre – Processing Part:

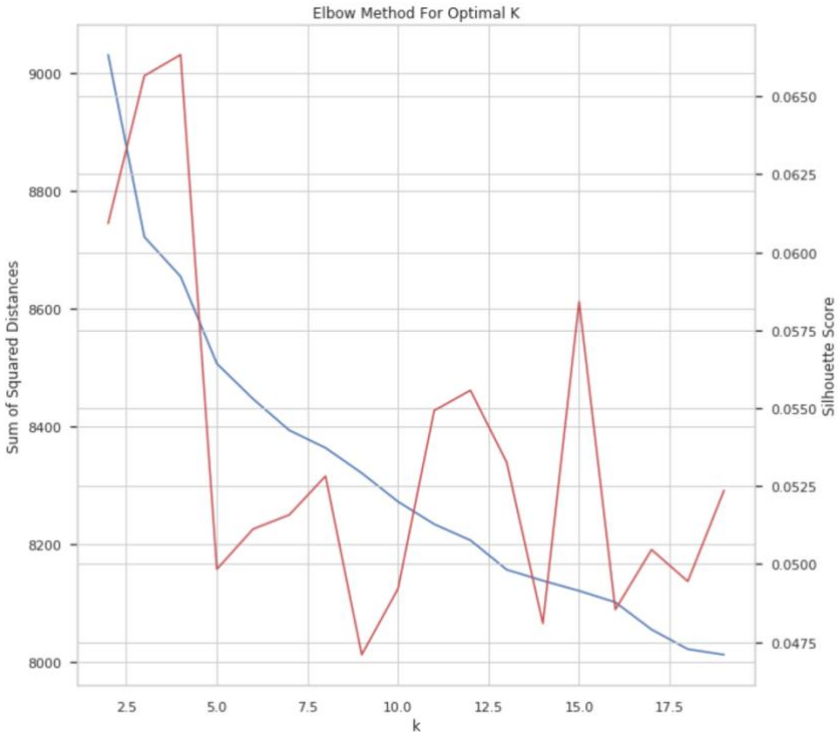
I have used BeautifulSoup to access the attributes of XML file and then I have created the dataframe.

While I was checking my data, I realized some of my text files do not have topics. I knew that, I should not use these files in order to check my K-Means cluster model in the future. That's why I dropped these documents from my dataframe. Some of my text files include more than one topic. Because of this, I have to used list for storing each topic related my text document.

The writing of the words are so important when you are trying to cluster text files, because if you do not standardize them "today" and "Today" can be keywords of two different clusters at the end. That's why, I have converted my text files lower case, I have removed unnecessary blanks and also, I have removed English stop words such as: I, you, we, the. After all, I could start clustering my text files.

## 1. K-Means:

K-Means is an unsupervised algorithm, so I cannot train it. I have fed the model with the data I have prepared before. In K-Means I should have to specify the cluster number which can be 2,3 or 15. It all depends on the dataset, in order to understand optimal cluster number for the dataset I have used two methods. The one is elbow method and the other one is silhouette score. However, reaching the optimal cluster number, I should have calculated sum of squared distances and also silhouette score for each cluster number in my cluster number range which is between 2 and 20. Below, I added the output of the elbow method and silhouette score.




With this graph, firstly I have chosen the number 11 as a cluster number however, the result was not good enough. I could have seen clusters were overlapped. That's why I have changed my cluster number as 8. According to my researches for silhouette score, I should choose the small one. This score is calculated using the mean intra-clusters distance and the mean nearest-cluster distance for each sample.

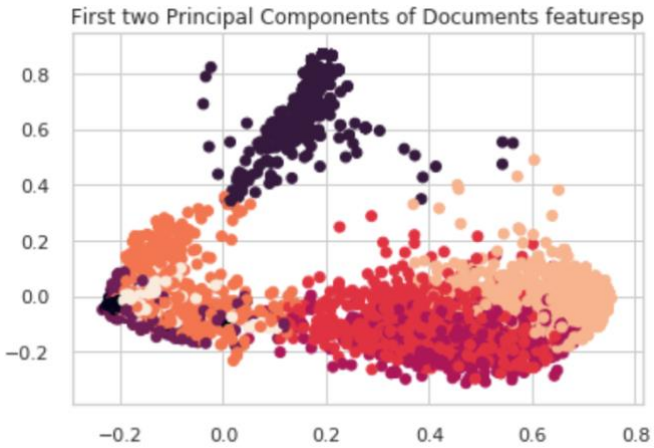
For K-Means algorithm, I have to vectorized the words for each document. So, model can cluster the documents using the vector. At the beginning I had 82.345 unique words and after vectorization process, I received a matrix shaped 10.377 x 82.345.

As we can see, in the beginning I had 135 different topics, but K-Means suggests me 8 clusters, so I understood that my topics are similar to each other.

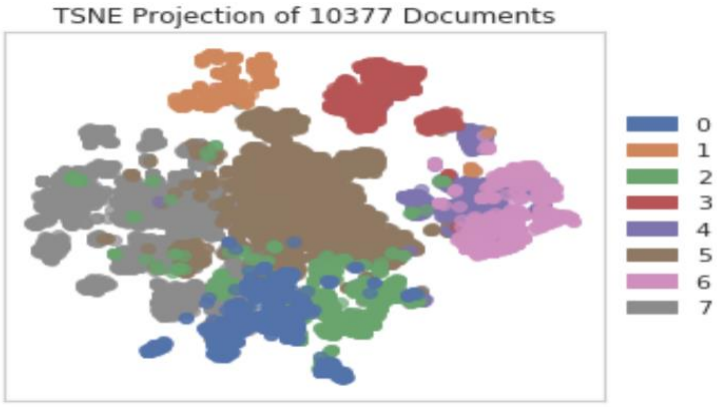
## K-Mean Algorithm Result:

After feeding the model, I have wanted to see the results as a graph however, I could not plot them since my matrix dimension was not 2. So, firstly I should have reduced the dimension of my matrix. For this, I have used 2 different approach, one of them is PCA and other one is TSNE. Below, I have added results of my K-Means clustering for both of them.

- Results of PCA



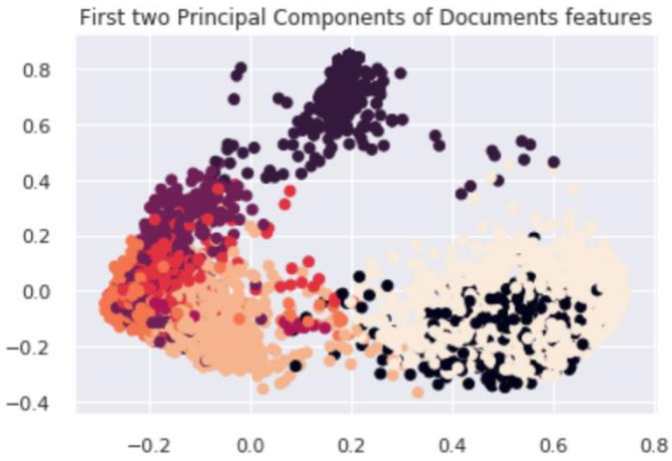

- Results of TSNE



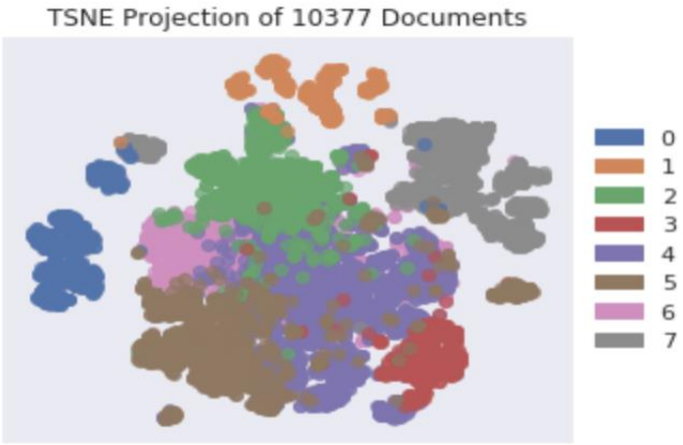

TSNE output was much better than PCA output since PCA has shown to me some clusters were already overlapped. For preventing this I have changed some useful and important details in my vectorization. I have vectorized all unique words which are included in my documents, but I have understood that, not all of them useful for clustering since sometimes one word can appear in all texts. So, I should remove these kinds of common word. That's why, I have specified in my model, I do not want to use the words which appears more than %85 of the documents and also which appears low than %.02 of documents.

After removing these words, my results are better now.

- Results of PCA after modification



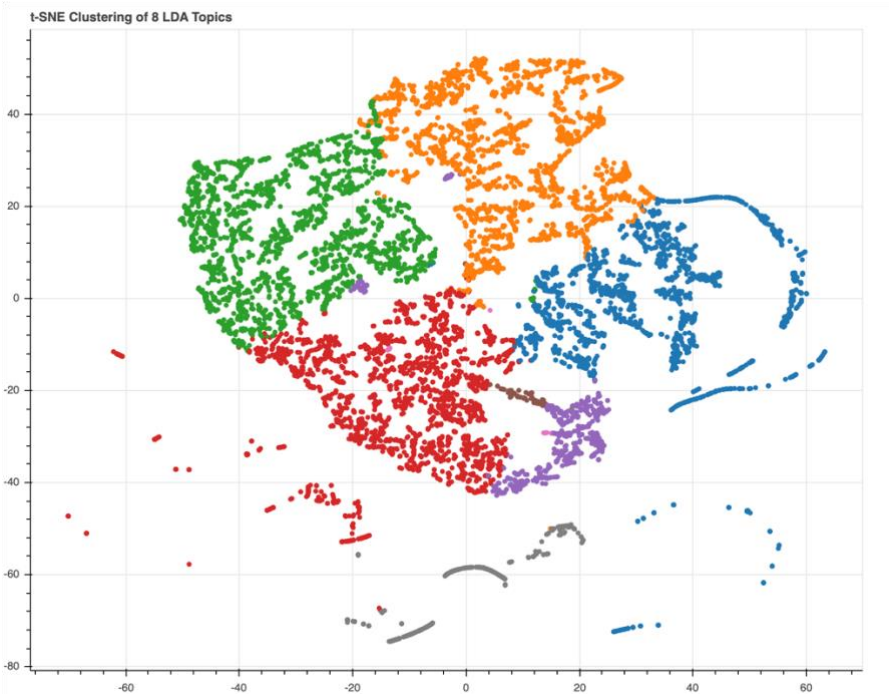

- Results of TSNE after modification




## 2. Latent Dirichlet Allocation (LDA):

I have used another algorithm named LDA which means, Latent Dirichlet Allocation. LDA is a topic model that generates topics based on word frequency from a set of documents. LDA is particularly useful for finding reasonably accurate mixtures of topics within a given document set. I have applied the same pre-process for also LDA algorithm. I have used same number of clusters.

- TSNE Result of LDA




- Most Frequent Words of Each Cluster