Project Report **Name: Omar Sharif**

Course: Information Retrieval **ID: 7492161**

Instructor: Dr. Santosh KC

Text Mining using Term Indexing and Grouping

Overview

In this project, I did text mining using document matrices. There are four text files from which I needed to text mining. We need several python packages to work on this project. They are -

Glob: pathname and pattern expansion library

Pandas: data manipulation and analysis library

String: a library to perform a various operation with strings

Os: A library to use operating system dependent functionality

Nltk: natural language processing toolkit

Sklearn: scikit-learn is a machine learning library contain classification, regression and clustering algorithms

Wordcloud: A library to work with the word cloud

Scipy: a library for scientific and technical computing

And *Matplotlib*: a plotting library for drawing graphs and charts

Methodology

I first extracted the text files into the workspace using glob library. Then I did some preprocessing tasks such as removing punctuation and stop words, stemming. Then I used vectorizer function from scikit-learn library to transform the nature of data into vectors. The transformation will be useful for document indexing and clustering. I then arranged my data by terms and frequencies.

Project Report **Name: Omar Sharif**

Course: Information Retrieval **ID: 7492161**

Instructor: Dr. Santosh KC

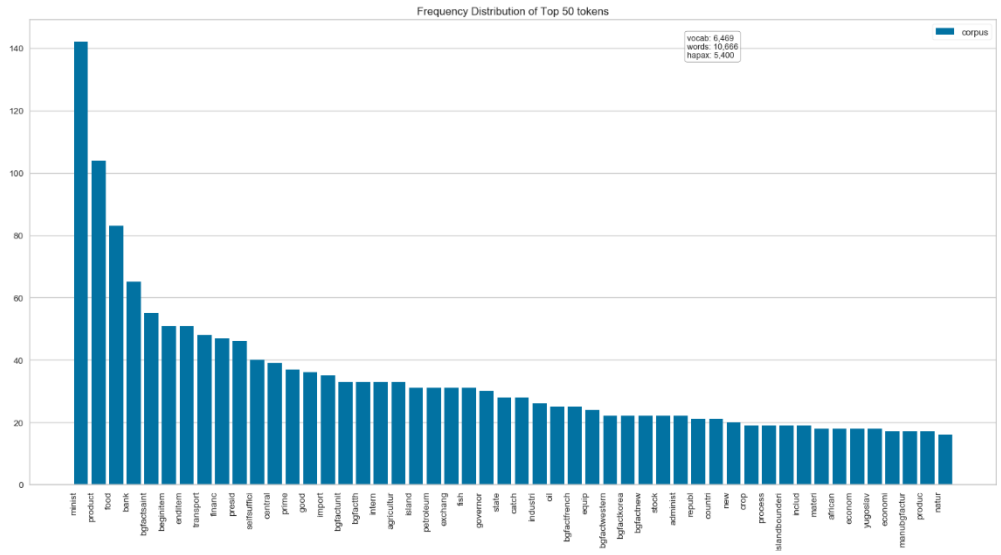

Figure1: Frequency Distribution

I used four kinds of evaluation to visualize text mining results. I used the bar diagram of the frequency matrix showing the number of the appearance of specific words. I have used word cloud to show more frequent words

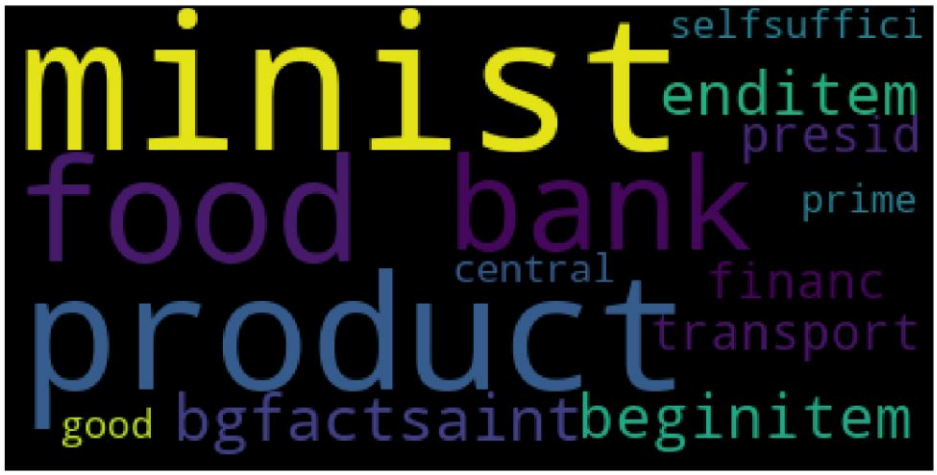

Figure2: Word Clouds

in the data. I also used a machine learning algorithm k-means clustering to observe classes of words with similar frequencies. Lastly, I plotted a hierarchical clustering to have a structural view of frequent words.

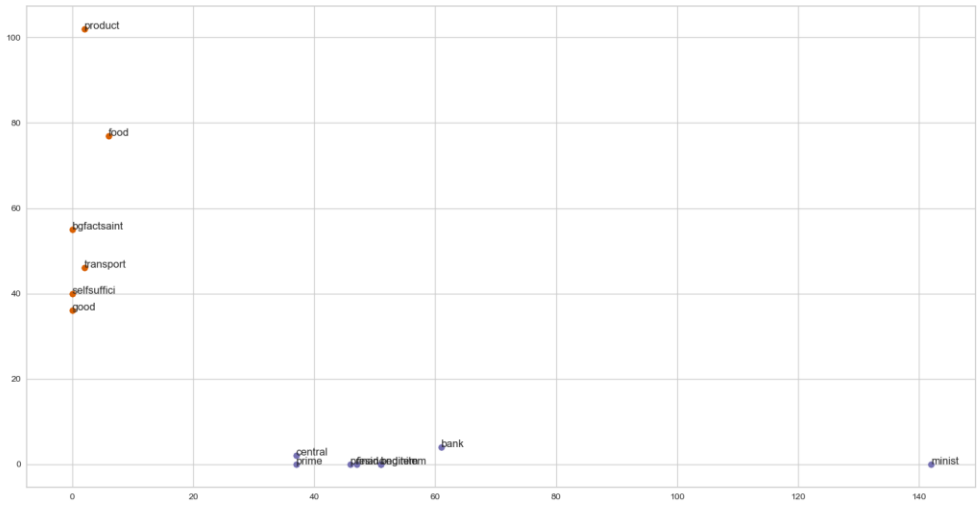

Figure3: K- Means Clustering

The course materials and lectures helped me a lot to complete this project. With this project, I got the chance to see the implementations of information retrieval and text mining theories. Nowadays, machine learning and natural language processing is a promising field of industry and research. This project will a good step to learn and practice those concepts to progress further in this field.

Project Report **Name: Omar Sharif**

Course: Information Retrieval **ID: 7492161**

Instructor: Dr. Santosh KC

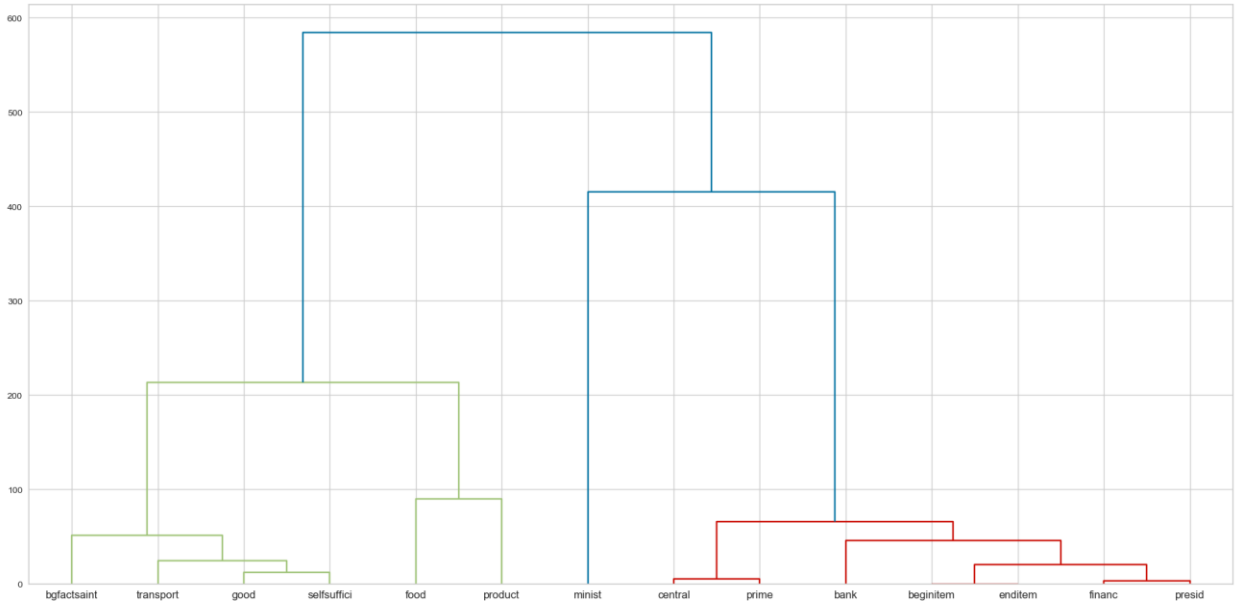

Figure 4: Hierarchical Clustering