**Text Analytics**

**Assignment 1 Text Representation**

This is an exercise to practice text representation methods and NLTK package.

**Input data –**

The input dataset is a sample of news articles related to S&P1500 in year 2016: <*News\_Article\_2016.csv>.* You also have the option to use a subset of the data that only contains news from January 2016 *<News\_Article\_2016\_Jan.csv>.*

The dataset contains the following information, where ***Name*** is the S&P1500 company related in the news. ***Content*** is the news articles.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Day | Name | Year | month | Content |

**Problem 1** – Your task is to analyze the dataset based on different text representation techniques to understand topic trend in 2016.

* You are required to analyze the data with 4 text representation approaches:

1. Use a simple bag-of-words approach
2. Use a bag-of-words approach with stemming and stop words removal (you can choose which stemming/lemmatization to use)
3. Use POS approach and focus on all the noun forms (NN, NNP, NNS, NNPS)
4. Use POS approach and only focus on NNP

* For each approach, present top 30 keywords (concepts) you extracted and plot them in a distribution chart using the NLTK tool.
* Compare the different text representation approaches and write an analysis report of their performance.

**Problem 2 –** Pick 2 companies preferably from the same sector of interest (try to pick companies with more data points, an example can be Microsoft and IBM). Perform topic trend analysis using the text representative method of your choice. This time, you do not need to show the distribution plot. Instead, pick top-N words from each month (or every week if you are using the small dataset). Observe the difference and describe what you have observed.

**Problem 3** – Bonus point (Optional). If you are interested in further analysis, you may implement the approach described in (Schumaker and Chen, 2008)

“Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System.” This will require that you get stock market data from Yahoo Finance for a particular period that you choose.

**Submission** – You need to submit the following document on the Blackboard:

1. A report in word document.

In the report, you need to present the results of 4 approaches (in terms of frequency and plot). Compare the results, and provide your insights about

1. Subjective analysis of each approach (which one give you best result?)
2. What do you see is a trend in 2016. Figures and tables are encouraged in the report.
3. Brief description of other improvements you can make.
4. Your Python code (Both .ipynb and .py files are accepted).