EM 624 Informatics for Engineering Management

Homework 5

Compare the New York Times and the Wall St. Journal comments on the Democratic political debate. The comments are in the 2 attached files (DemocraticDebate_NYT.txt and DemocraticDebate_WSJ.txt). Most of the non-relevant characters have been removed from the files, leaving plain text. Feel free to use other text data that you are interested in.

In order to perform the comparison, you will extract key information from the text:

- Remove from both the files the stopwords, using the attached *stopwords_en.txt* file
- Remove end-of-line (“/n”) and create lists of words
- For both the files/lists of words, calculate the 10 most frequent words
- Extract bigram. Consider bigrams 2 words appearing together more than 2 times in the whole text. Calculate then the most frequent bigrams
- Use this website: http://www.danielsoper.com/sentimentanalysis/ to calculate the sentiment for the 2 articles. You can just copy and paste the text to get the sentiment. You may use the original text files (the one with the stopwords)
- Using/modifying the attached cloud.py script, create word clouds for the 2 texts.

Once you performed all the above steps, write a brief report with your conclusions. The report should contain the output of the above analyses.

Submit the report and your python file (.py or .ipynb).