Data Intensive Computing

Lab 2

Introduction:

This project aims to compare the word count of data from Twitter and NYTimes. For this purpose, we have prepared a comprehensive application that meticulously scrapes data from such sources and runs a Hadoop map reduce in order to produce a word count data. This data is being used to plot word counts using D3.js web interactive library.

Work Flow:

- It all starts with the scrapers. Because, we need the data to do anything. Two scrapers using Twitter API (Tweepy) and NYTimes JSON API. We just need the text data from these sources. Python has really helpful tools to assist in scraping and a lot of support online, so we took our programming language to be Python.
- We generated two text files through this process, one from Twitter and one from NYTimes.
- We, then, processed the data in the mapper (cleaning and stopwords removal). We followed the Michael Noll mapper and reducer to do this part.
- The reducer generated a file, which we then fed to the D3.js application that we developed.
- Just before that, the files were parsed to become Javascript files to be directly fed into the code.
- The website is a replaceable wordcloud, with options to select daily or weekly data and unigrams or bigrams. It loads on the first run and then, works seamlessly to plot wordclouds using D3.js.

Inference:

We tested the application with the keyword, Trump, since he has been a very hot topic in recent news and twitter. We found the wordclouds to be similar in many words. Even in bigrams, the hot topics of discussion were prevalent and found in bigger sizes. This concludes this big data endeavor. It proved to be a successful learning experience.

References:

https://github.com/casmlab/get-nytimes-articles
https://github.com/wvengen/d3-wordcloud
http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/
https://www.karambelkar.info/2015/01/how-to-use-twitters-search-rest-api-most-effectively./