Jennie Lee
Michael Alaniz
Mirko Draganic
Steve Braich
Gennadii Styov

Data Science & Analytics Project Part 2 – Midpoint Report

Retrospective:

For our project, the things that went right are we were able to clean up data set so that we can perform analysis and we were able to explore different tools to find what worked and what did not work for our project. The biggest issue we had to tackle initially was cleaning up the data set. We were able to clean the data up and store the data into multiple files, which is useful for our beginning stages of developing metrics and analysis. We investigated and explored different tools to reinforce the tools that we have selected work well for our project. Since we are developing this project from scratch, we were not sure what tools are best to use, so we decided to check out a few different ones, such as the NLTK library in Python, Jupyter Notebook, MongoDB, and R and R Studio. From the exploration, we found some tools that worked for us, such as R and R Studio, and what did not work, e.g., MongoDB, so now we are set on the tools we want to use to develop metrics and for analysis.

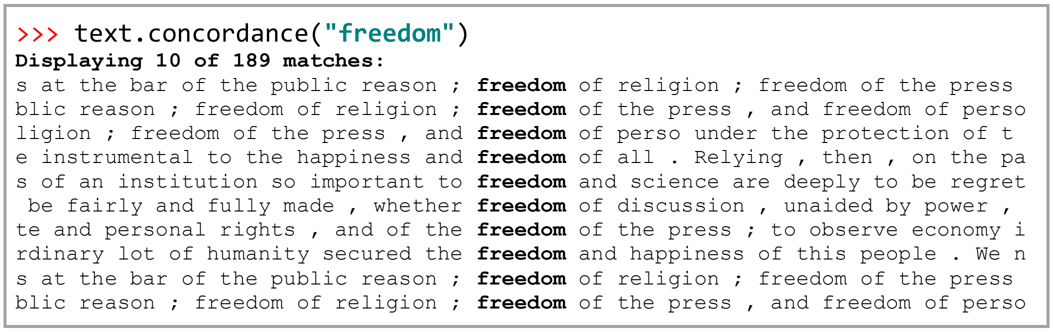
What could have gone better is cleaning and organizing the data in a better way and setting up a unified environment so that we all are working with the same environment. We were able to clean up the data, but that was challenging and more time consuming than expected. We should have formatted, preprocessed, and tokenized the data correctly from the start, but we did not. Instead, we kept having to go back and work on cleaning the data some more when these things could have been done early on, so we ended up spending more time on this part. We also spent a significant amount of time trying to figure out what tool we wanted to use and setting up a unified environment. We were all over the place trying to find the right tools because we could not come into an agreement about what we should use, so instead, maybe next time, we can just initially choose the tools that we need to use for the project to cut down on the time exploring other tools.

Project Design Issues / Adjustments:

We are on track for our project; we are actually a little ahead. We are at point of determining what metrics we want to produce and in the process of producing them. What was modified/added to the spreadsheet was determining metrics and producing the metrics. These were modified/added because we found that we needed to first determine what metrics we want and then we need to actually produce them, which is a two-step process. Another task that was added was producing the pipeline/predictive model. Our final goal is to produce a pipeline/predictive model that can predict sarcastic comments with pretty good accuracy. If time permits, we will attempt to produce a pipeline/predictive model, but this part is definitely based on time since we have to first come up and produce metrics and then perform analysis on the results and create visualizations for our results. Therefore, this part of the project is placed last on the task sheet, but it is something that we hope that we can work on if we are able to since it is the ultimate goal of our project.

Draft Analytics or Graphics:

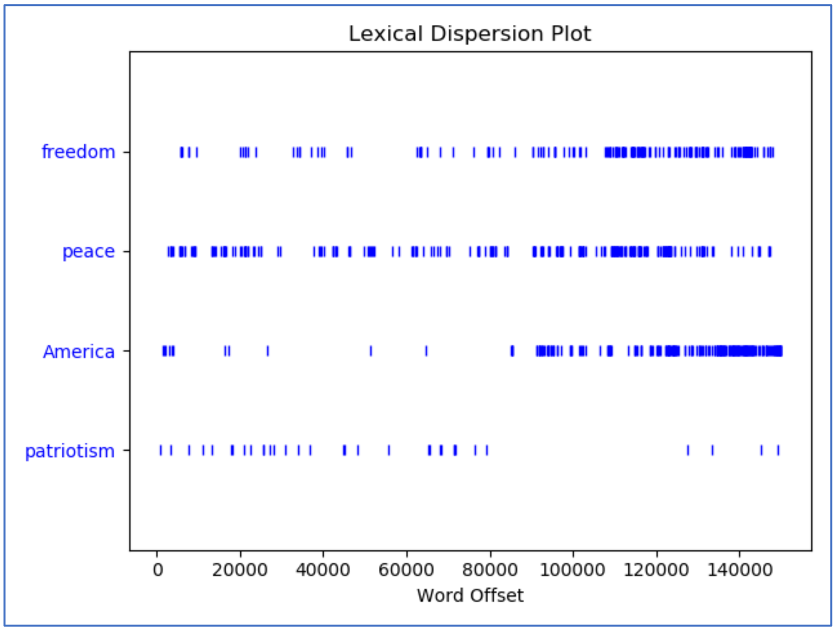
1.) Concordance Search



```
>>> text.concordance("freedom")
Displaying 10 of 189 matches:
s at the bar of the public reason ; freedom of religion ; freedom of the press
blic reason ; freedom of religion ; freedom of the press , and freedom of perso
ligion ; freedom of the press , and freedom of perso under the protection of t
e instrumental to the happiness and freedom of all . Relying , then , on the pa
s of an institution so important to freedom and science are deeply to be regret
 be fairly and fully made , whether freedom of discussion , unaided by power ,
te and personal rights , and of the freedom of the press ; to observe economy i
rdinary lot of humanity secured the freedom and happiness of this people . We n
s at the bar of the public reason ; freedom of religion ; freedom of the press
blic reason ; freedom of religion ; freedom of the press , and freedom of perso
```

The purpose of this analytic/graphic is to see how words are used in context. With this, we can also search for what other words appear in a similar range of contexts. We can also examine and compare the contexts shared by two or words that are related. The intended audience, in the case of our project, are those who are interested in finding the context surrounding a sarcastic word in order to predict whether the phrase is sarcastic or not.

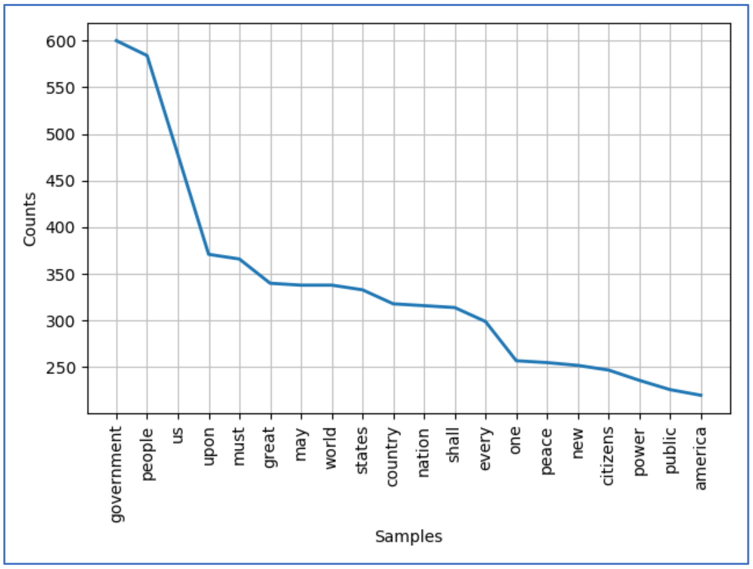
2.) Lexical Dispersion Plot



Lexical Dispersion Plot of words in the US Presidential Inaugural Address corpus.

The purpose of this analytic/graphic is to find the location of a word in text; more specifically, how many words from the beginning does a word appear. The purpose of this is to get a sense of where these words appear in a corpus and determine if the word is important in terms of the text. The intended audience, in the case of our project, are those who are interested in seeing how important a sarcastic and non-sarcastic word is and where it appears in a corpus.

3.) Word Frequency Distribution



Frequency Distribution of the top 20 words in the US Presidential Inaugural Address Corpus

The purpose of this analytic/graphic to get a sense of the overall topic and genre of a corpus. In the example above, just seeing the frequently used words, we get the sense that the corpus has something to do with something of the political nature. We intend to apply this to our project. Our intended audience is those who are interested in knowing what words are frequently used in sarcastic and non-sarcastic sets of corpuses.