Fabian Hansch Mauritzson

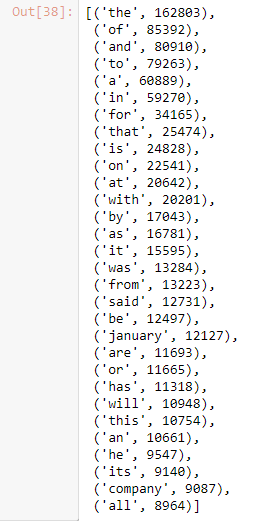
Text Analytics

Assignment 1

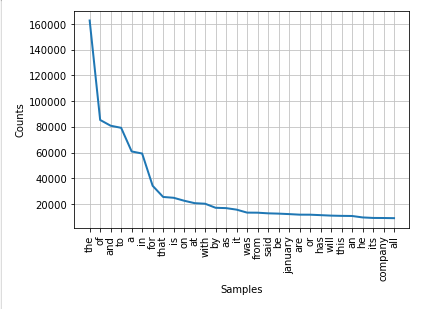
March 12, 2021

**Problem 1:**

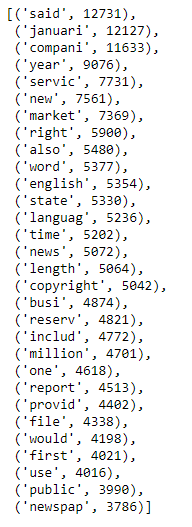
1. Frequency:



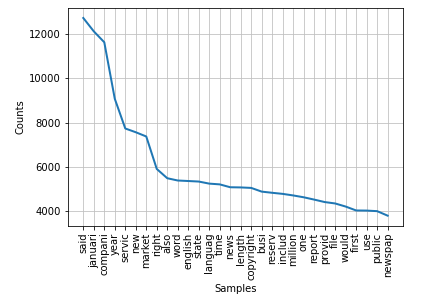
Plot:



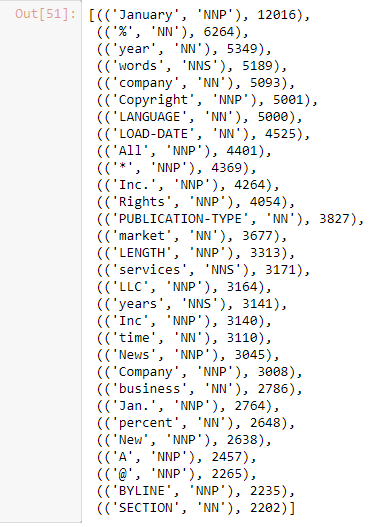
1. Frequency with Porter Stemmer:



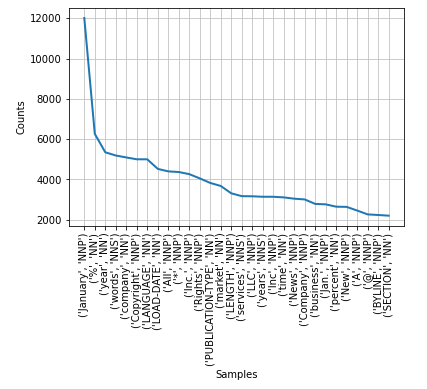
Plot:



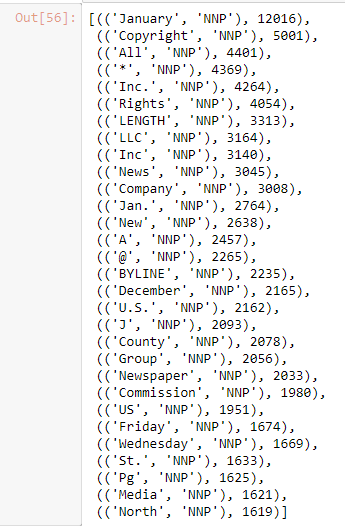
1. Frequency:



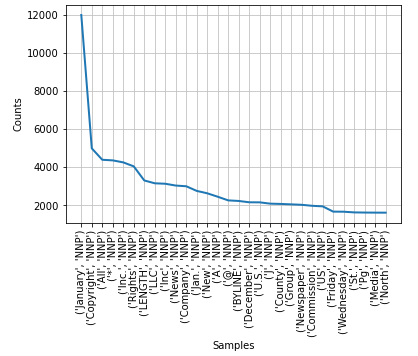
Plot:



1. Frequency:



Plot:



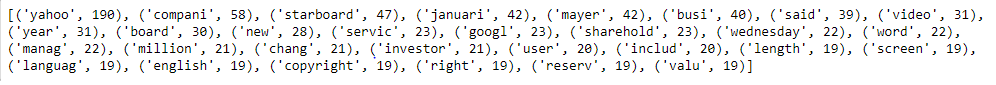
Analysis:

The first thing which can be pointed out is the difference when using stop words and stemming compared to just a regular bag-of-words. The most used words when including everything are the stop words. This makes sense since there is a reason why an analysis should be made without them. It also shows that the stop words do not add any value to the overall text analysis of these articles. Looking at the analysis with the stemming, it is clear that the texts in questions are articles since the most common stem is said. This means that someone is referring to another source, which is common in articles. The third most common stem is referring to different companies, which makes sense since it is articles about companies. What is interesting to note is that there are clues to make it possible to figure out that these articles are written early in a year. The second most common stem refers to January, and there are also stems referring to new year high up in the frequency.

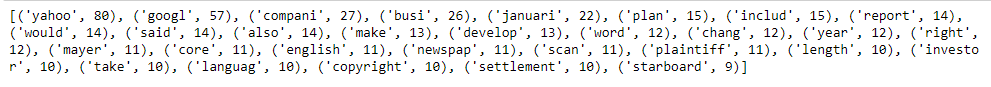
When only the nouns are analyzed, there is a lot of content that is missing. It is not as easy to understand what the articles are about. It is also since there are some questionable “words” considered to be nouns, which is taking up space in the top 30. In order to use the POS analysis better, the cleaning of what should have been considered a noun should have been more robust. While there is a significant improvement between the regular bag-of-words to the stemming and stop words removal, the same increase in performance is not seen between all the nouns and only the NNP’s. This means that the best performing text representation technique is bag-of-words with stemming and removing stop words.

**Problem 2 Google and Yahoo:**

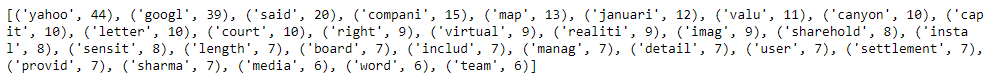
Week 1:



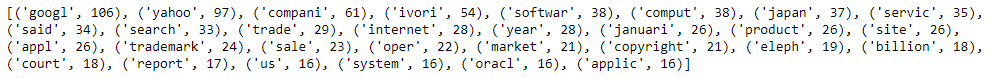
Week 2:



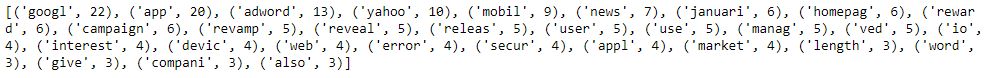
Week 3:



Week 4:



Week 5:



Analysis:

What is striking to see right away is the variance of number of times each company got mentioned from week to week. This can be seen in how many times each company got mentioned. It also seems that Yahoo had overall more mentioning than Google over the month, which is interesting since Google is such an influential company with a market share of their search engine magnitudes higher than Yahoo. Even though Yahoo is a larger player in other areas, for example Yahoo Finance, it is still a surprise that they are more prone to be in the media compared to Google. The different weeks had a great deal of difference in terms of content as well. There seems to be some discussion regarding the board of director for one of the companies in week one. While in week two there is some legal issues because of the use of the word “plaintiff”, and there also seem to be something developing as well. In the third week, the talks about legal issues and board of directors are continuing. Both these topics are present again this week. Week four talks about some legal battle again, while at the same time discussing intellectual property as well. Since week five only had a limited number of articles, it is harder to gain an understanding of what was crucial that week since the amount of data points is so low. However, there seem to be some reveal of a new homepage for either of the companies, and since Google got mentioned the most, it is fair to assume they were the ones who launched it.