Please note: this report will be divided up into three sections:

**Section 1:** Problem #1 results and summary.

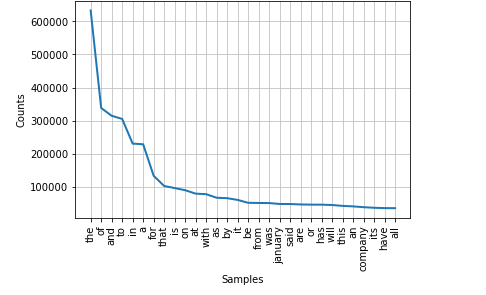
**Section 2:** Problem #2 results and summary.

**Section 3:** Comparison of all approaches and overall discussion of results.

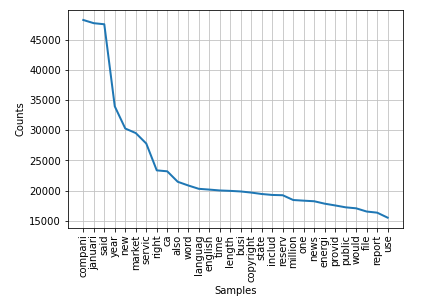
**Section 1: Problem #1 (please see submitted ipynb file for Python code)**

Please note: I have opted to first place the graphs and then write the analysis afterwards. Also, please note that I have opted to not tinker with the NLTK default stopwords list. As this is my first time using the NLTK package, I have decided to keep the results as “package standard” as possible.

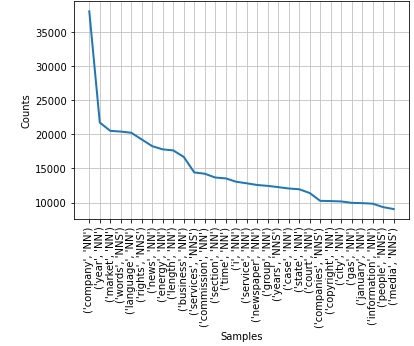
*Bag of Words Approach*



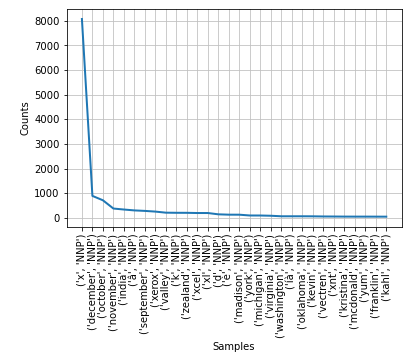
*Bag of Words Approach With Stop Words and Lemmatization*

**

*Part of Speech Analysis – Nouns Only*

**

*Part of Speech Analysis – NNP Only*

**

To begin with the most obvious finding, the simplistic “Bag of Words Approach” by itself yielded, by far, the greatest counts with many word frequencies in the hundreds of thousands. That said, the simplistic “Bag of Words Approach” also yields mostly junk. Comparing by eye with “Bag of Words with Stopwords and Lemmatization” reveals that only two words captured by the former, January and Company, are actually not stopwords. Thus, simplistic “Bag of Words Approach” without augmentation is likely near-useless as a foundation for Text Mining.

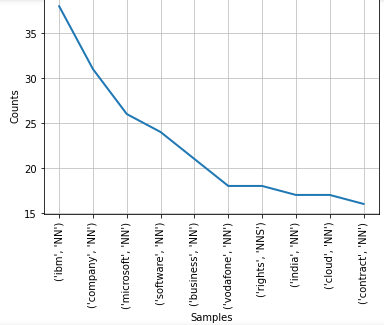
As would be expected, further refining the parameters used by the NLTK, specifically getting rid of all parts of speech except for nouns, reduces the number of words generated in the final result. Additionally, it seems to provide more focused analysis than the previously discussed methods. For example, “Bag of Words with Stopwords and Lemmatization” naturally captures verbs but does not allow us to contextualize those verbs by observing the nouns they are acting directly upon. However, zooming in on the nouns provides us somewhat more context as to what the common topics in these news articles might be. For example, newly appeared are legal words such as “case” and “court”, which hint to us that an appreciable amount of content in these news articles is concerned with legal proceedings or legal matters.

Finally, imposing highly restrictive constraints, that is, focusing on only proper nouns, dramatically cuts the overall word frequency. However, while this approach does reveal some names and locations, it actually ends up providing too little information to be of any use. We lose topically-specific words, such as “copyright” and instead gain some bizarre word fragments like “x”, “d”, and “e”. Therefore, this approach is likely too aggressive to be of much use, except in very specific scenarios where we are uniquely interested in learning names or places.

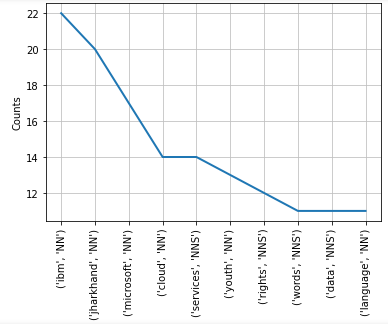
**Section 2: Problem #2 (please see submitted ipynb file for Python code)**

In order to better facilitate an analysis of the week-over-week differences in the Top 10 words which appear in the Microsoft and IBM data when using a “nouns only” approach, I have opted to create four frequency distributions; one for each week of January 2016.

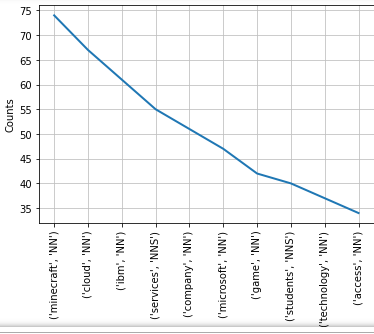
*Microsoft and IBM Nouns Frequency Distribution for Week 1 of January 2016*

**

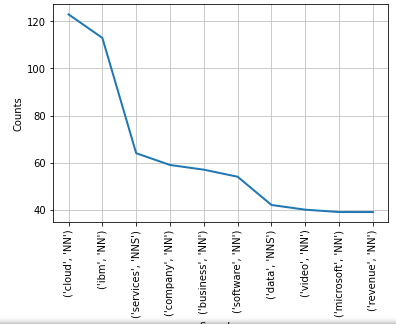
*Microsoft and IBM Nouns Frequency Distribution for Week 2 of January 2016*

**

*Microsoft and IBM Nouns Frequency Distribution for Week 3 of January 2016*

**

*Microsoft and IBM Nouns Frequency Distribution for Week 4 of January 2016*

**

When looking at these four graphs, the first observation that jumps out is that the total count of nouns in the Top 10 roughly increases as the weeks in the month progress. To be perfectly honest, given that the month here is January, I am not certain if this just the result of randomness or not. If, on the other hand, the month was a quarter-end month then I would feel comfortable hypothesizing that longer articles are being written towards the ends of those months discussing the companies’ quarterly performance, and that this could account for the upwards trend observed. In addition, it is worth noting that as the weeks of January 2016 went by, the noun “Microsoft” continued to slide farther and farther down the Top 10 list, whereas “IBM” consistently occupied a high-frequency slot. Interestingly, a quick internet search reveals that in January 2016, a few business media outlets were writing articles noting that January 2016 marked a consecutive 15 quarter sales decline for IBM. Therefore, it is reasonable to posit that although the noun “IBM” was often discussed in this corpus of documents, it may have been for a rather negative reason.

Finally, I would like to highlight the interesting appearance of “Minecraft” and “video”. On January 19th, 2016, the extremely popular video game known as Minecraft was announced to have a forthcoming “education edition” in development. Minecraft was developed in part by Xbox studios, which is a subsidiary of Microsoft. Given that it was, and still is, quite rare for a major gaming studio to bridge the gap between entertainment and education, it seems likely that media outlets picked up on this announcement and began writing stories covering the Minecraft video game and its newly announced expansion.

**Section 3: Overall Summary**

As implied in Section 1, the best approach for obtaining and analyzing text frequency distributions out of the four surveyed in this assignment is the third approach, *Part of Speech Analysis – Nouns Only*. This approach seems to yield words in the “Top N” which allow us to glean the most amount of context out of a relatively small amount of information.

Like I touched on above, a notable trend in January 2016 was the prevalence of legal and legal-related terms such as “case”, “commission”, “rights”, “state”, and “court”. Given this, it would appear that, vaguely speaking, companies in the S&P 1500 were dealing with notable legal activities.

Finally, in order to improve the results of this analysis, the Python code could be changed to perform either a sentiment analysis or a syntactic analysis. Stated differently, since the analysis done in this assignment was only very surface-level, any additional depth would likely reveal additional information and therefore additional meaning/insights. Beyond that, more work could be done to refine the stopwords list. It is not clear what amount of benefit this would confer to the analysis, but it would necessarily mean that the final frequency distributions would contain more signal and less noise.