Homework 1 Written Report

TEXT ANALYTICS 5:45PM

Bao Ngoc Dinh

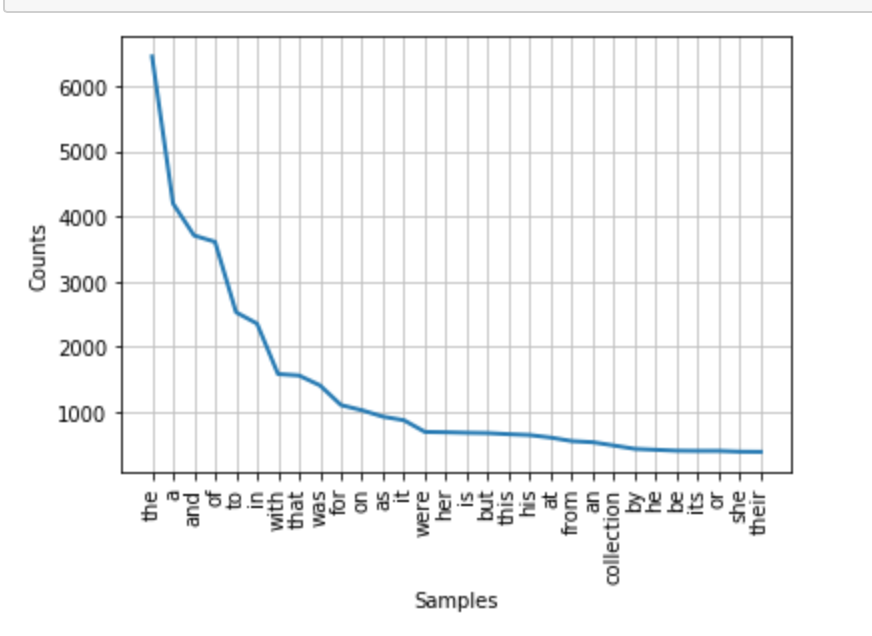
2020

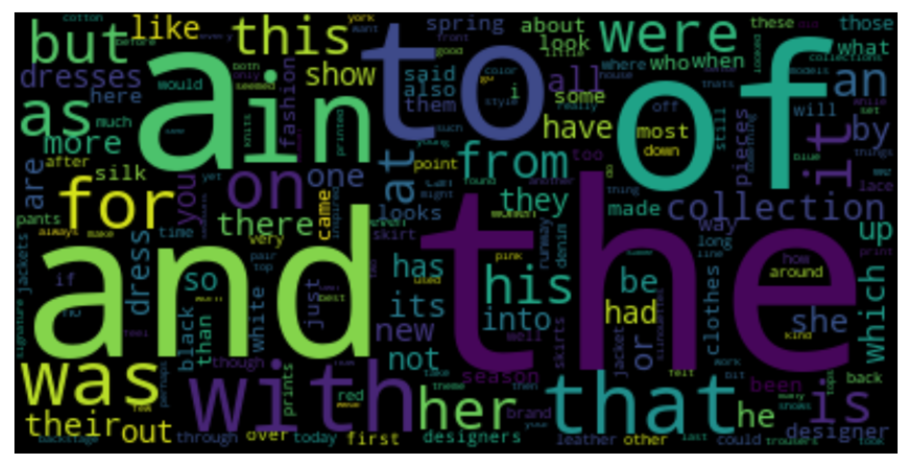
**Pre-processing**

Firstly, the dataset was tokenized. The ‘review-text’ portion of the data is the only part I will focus on. The reviews were tokenized. All words are then transformed into lowercase move and special characters, non-alphabetic characters were removed.

**Bag of Words**

In our simple bag of words approach, I did not reach any meaningful conclusion. The bag of words simply took all the words without much pre-processing so the word frequency list returned was full of stopwords: the, a, that, this….. The top 30 frequency plot provided us with no insight because most of the top words were stopwords

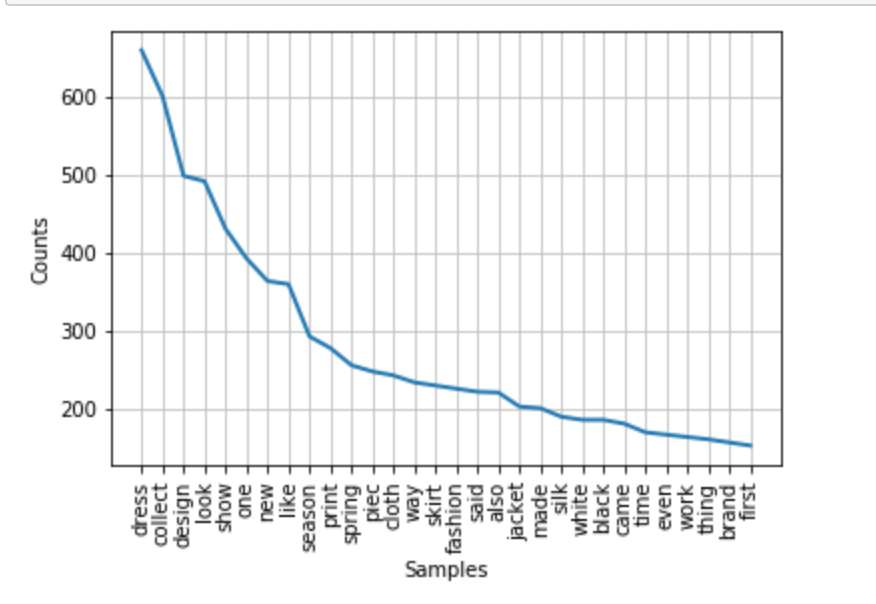


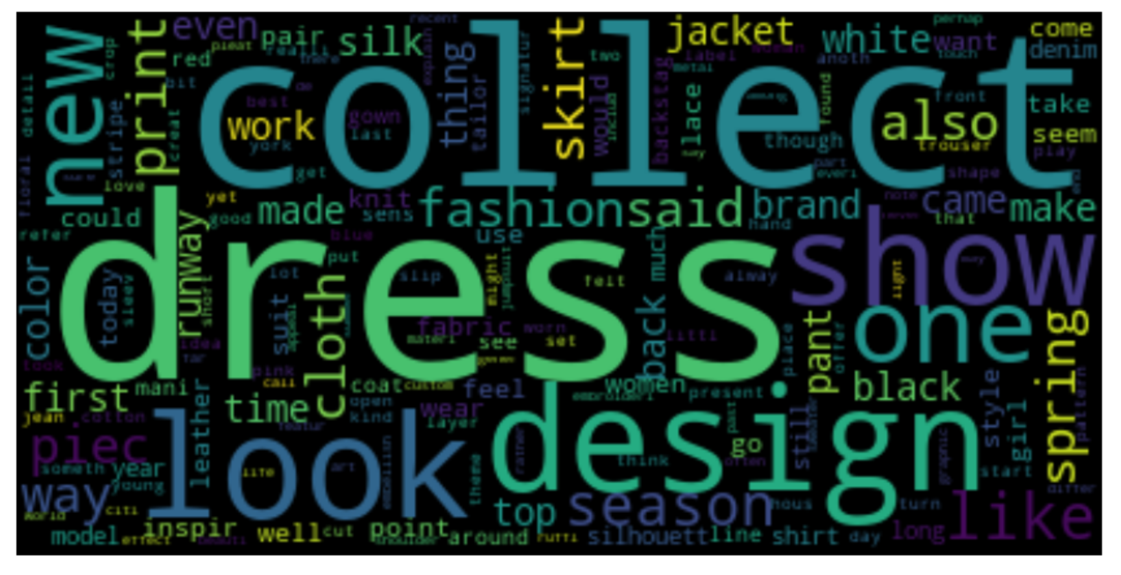


With the word cloud, we can capture more words outside the top 30. We can see in a very small font that ‘dresses’, ‘silk’ and ‘white’ were mentioned. However, those words are in small font so to say that those are some trends in 2016 would be an over-reaching conclusion.

**Porter Stemmer and Removing Stop Words**

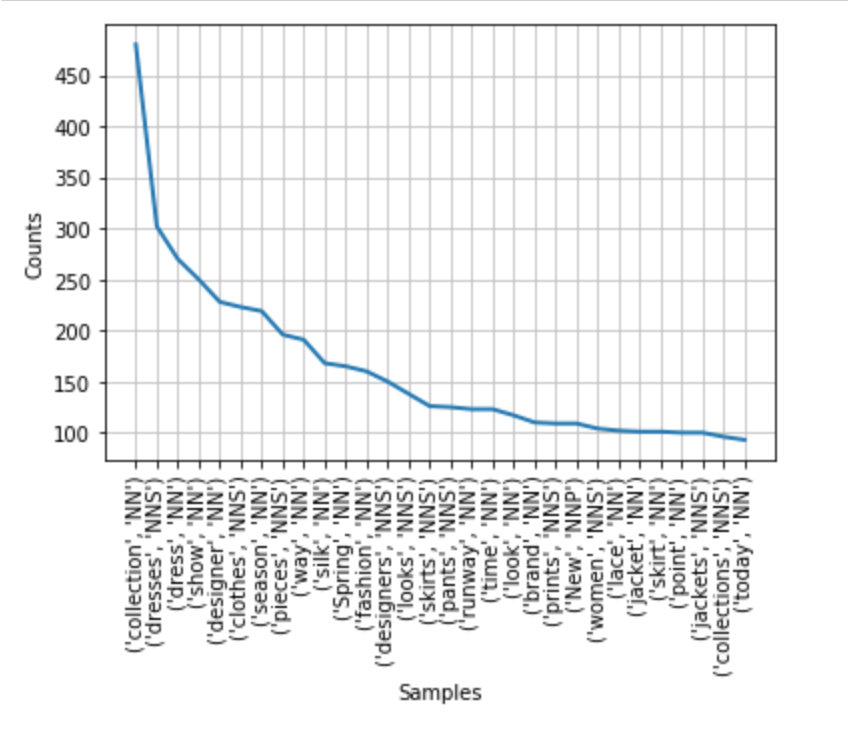
With the second approach I combined removing stop words and using the Porter Stemmer. With the porter stemmer I ‘stemmed’ the word down to its basic roots. With the second approach, we could identify some clear trends and top words. In the spring of 2016, the fashion world cared about dresses, print, skirt, jacket, silk, black and white clothing. Dress was the most frequently mentioned word at more than 600 counts. However, ‘dress’ is a very vague and not well-identified word. Dresses can have many styles or flares. The second approach helped us identify fashion-related nouns and adjectives, but single noun or adjectives aren’t as conclusive or specific enough to help us see clear trends.





**Focus on Noun Forms**

In the third approach, I used the part-of-speech tag and focused on all the noun forms. We could see that ‘dress’, ‘silk’, ‘skirts’, ‘pants’, ‘prints’, ‘lace’, ‘jacket’ are some of the items the fashion world frequently talked about in 2016. There were some inconclusive words, such as ‘point’. Point could refer to either point shoes or pointy skirts.



**Focus on Proper Nouns**

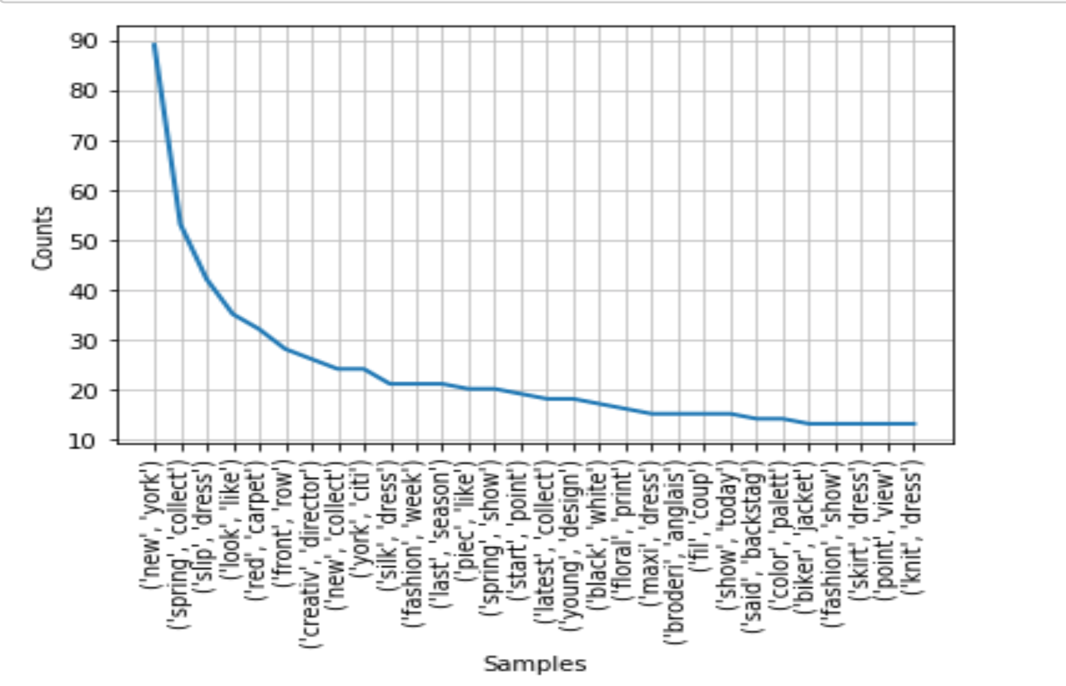
In the fourth approach, only proper nouns were plotted. The proper nouns could provide insights into the most talked-about brand or social media in the fashion world. As expected, the top frequented proper nouns were the names of fashion capital: New York, Paris, London, Milan. ‘Wang’ could be a reference to famous designer Alexander Wang, Chanel and Versace were two other fashion brands mentioned. The software is not perfect, so ‘I’m’ was actually considered a proper noun. ‘A-line’ was also mislabeled as proper noun because the a was capitalized, but it could simply referred to a-line skirts, a noun.

A picture containing text

Description automatically generated

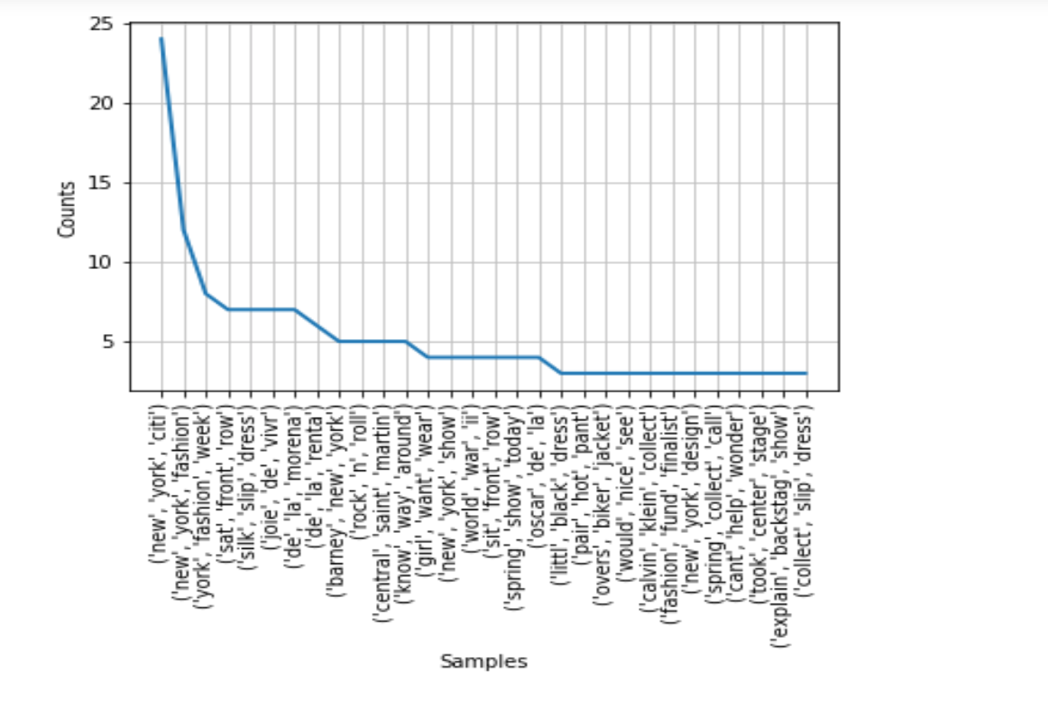
**Bi-grams**

For additional analysis, I used Bi-grams. Pair of words are put together and counted using freqdist. However, I opted to use the list of all the stemmed reviews instead of the reviews in its pure form. The frequency distribution showed some clearer identification of trends. For 2016, the fashion world talked a lot about slip dress, silk dress, floral print, maxi dress, biker jacket, knit dress and the colors black and white.



**Tri-gram**

I went a bit further and tried the tri-gram approach to see if we can find more insights. However, with the trigram approach, we hit a wall. The trio of words mostly refer to the city, collection or the designers. We did get one fashion term, little black dress, but with extremely low frequency. Thus, it might not be a frequented trend in spring 2016.



**Conclusion**

Overall, removing stopwords and porter stemmer increased the performance of our frequency distribution. The best performance was reached with the addition of using bi-grams. That’s because sometimes words go together in pair. The challenge of this project was finding industry-focused words. If whoever that wants to extract the trends just want to focus on the trending items and not the city or the designer names, they could create a list of designer names and cities they want to exclude. Hence, the frequency distribution will be cleared of terms such as ‘New York, ‘Paris’ or ‘Barneys’ that provided almost no insight to what trends were like in 2016.