****

**Text Analytics**

**Assignment 2 – Word Frequency Analysis**

**Yanzhe Li**

**02/22/2017**

**Word Frequency Analysis on Reviews from Vogue During 2015 Fashion Weeks**

Summary Report

**Abstract**

Words frequency can reflect the how people care about the concepts and in what level they are discussing. The frequency over time will carry information about topic trend. This report will go over four methods on analyzing the crawled review text data from Vogue during 2015 fashion week and discuss the results and performance on each method.

**Simple bag-of-words approach**

In this approach, a simple word frequency analysis was used. In addition, to improve the performance, the python model also allow user to set n for n-grams in case the user wants to see not only words but also phrases. In this case, we set n = 1.

Below chart “Words Frequency – Simple Bag of Words Approach” shows the top 30 frequent features. Since the simple bag of words doesn’t remove stop-words, 99% of the keywords here are meaningless. “Collection” might the the most relevant one about fashion domain.

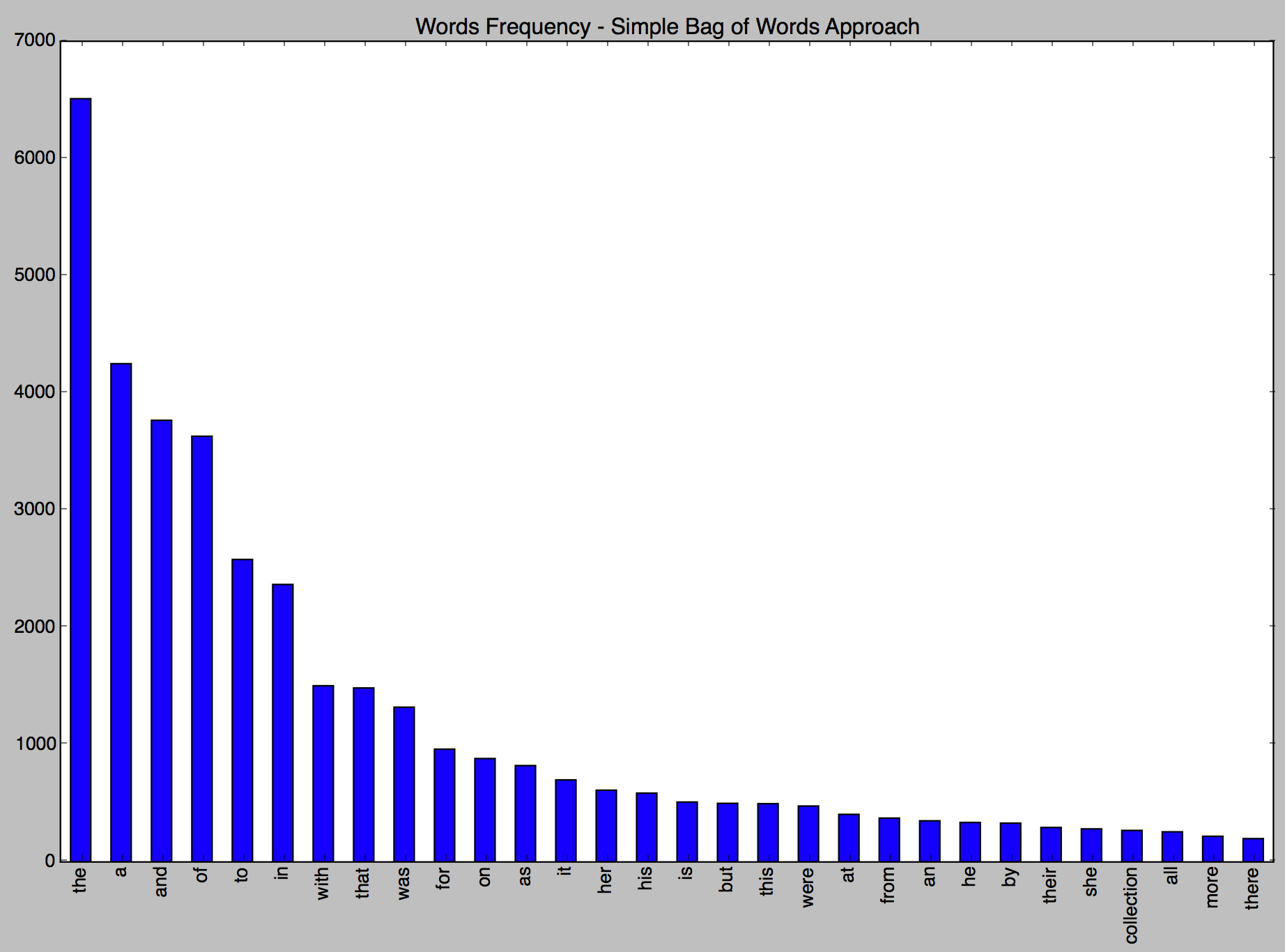


Figure 1. Words Frequency – Simple Bag of Words Approach

In addition, Figure 1.2, 30 keywords were plotted over time. There were two peaks on Sept. 14 2015 and Sept. 30, 2015. Although the words don’t mean anything, but since the frequency are much higher on those two days, we may guess there were large number of reviews posted or discuss on those days, indicating some big events may have happened.

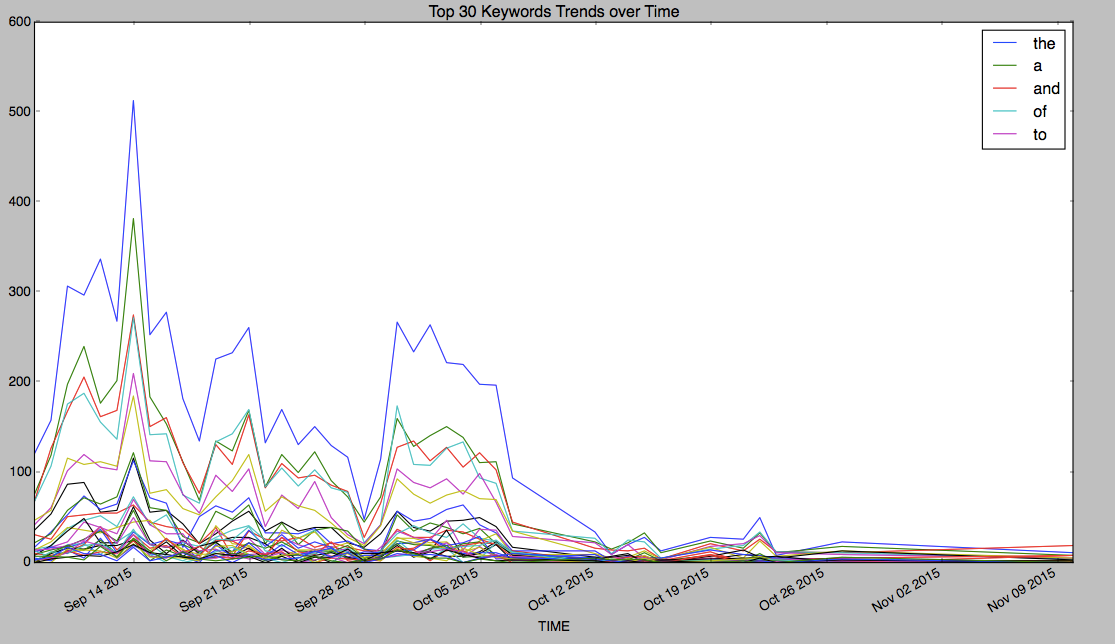


Figure 2. Top 30 Keywords Trend for Simple Bag-of-Words Approach

Simple bag of words approach can be barely helpful due to the large amount of noisy words. Next step will remove stop words and stem the words.

**Bag-of-words approach with stemming and stop words removal**

In this part, the new approach will remove stop words and stem words using WordNet Lemmatizer. The result is showing below in Figure 3. As the top concept is “dress” which is more meaningful than the previous approach’s “the”. On the other side, there is a clear frequency group displayed in the chart that “collection” and “dress” are in a group, from “new” to “jacket” is another group, and the rest are the third group. They may be similar topic or refer to similar things.

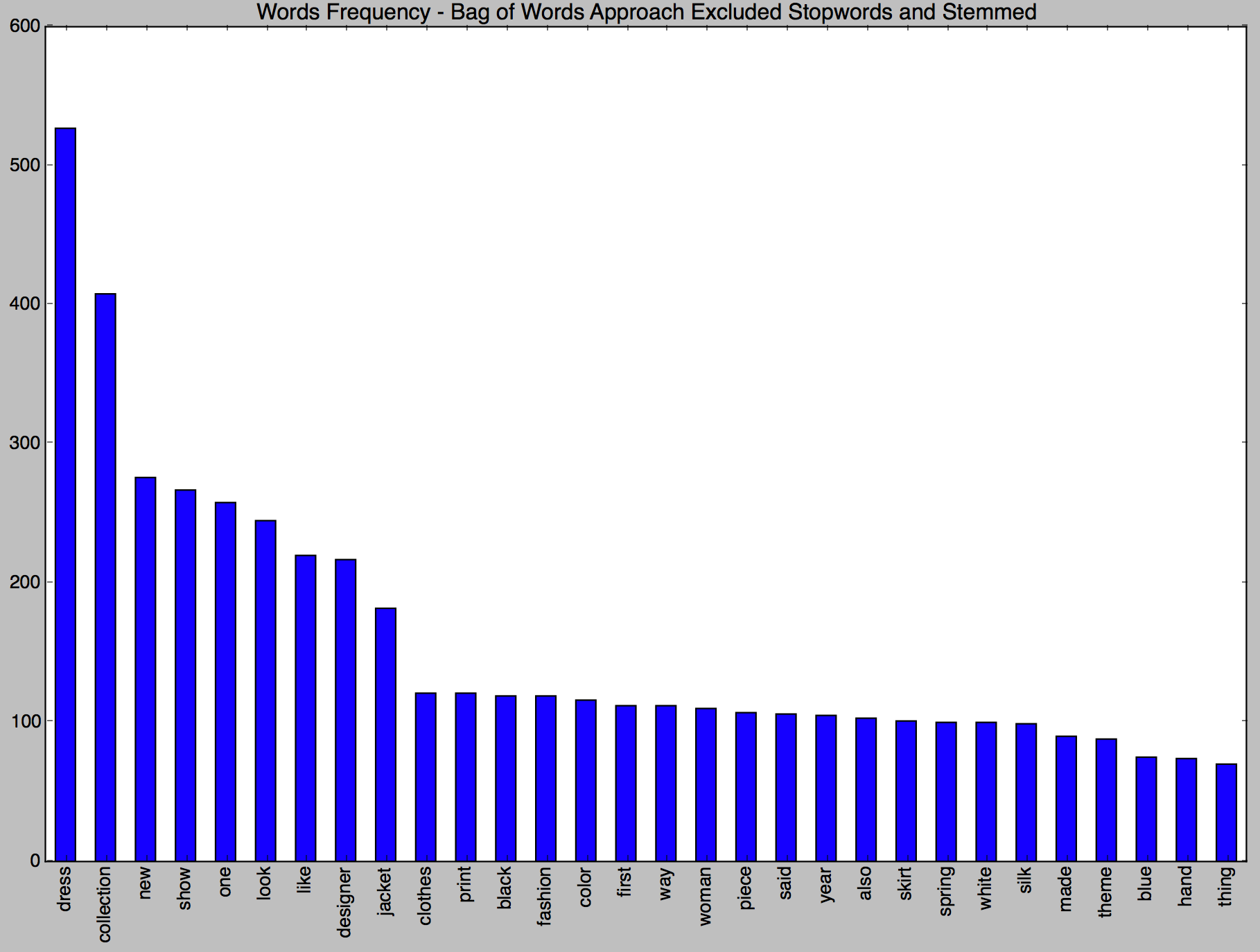


Figure 3. Words Frequency – Bag-of-Words with Stemming and Stop-words Removal Approach

Regarding the results, “dress” and “collections” are the most two important concepts, following by “new”, “show”, and “designer” is also a good catch.

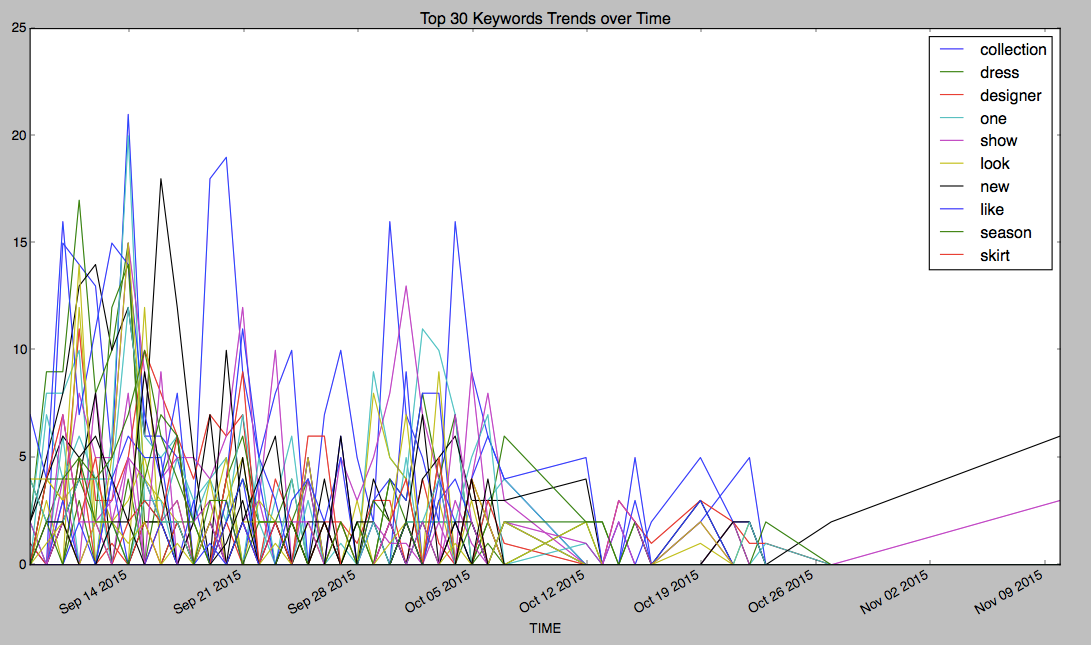


Figure 4. Top 30 Keywords Trend for Bag-of-Words with Stemming and Stop-words Removal Approach

In Figure 3, it plots the top 30 frequent words over time. At the beginning of the fashion weeks, “collections”, “dress”, “new”, “like” appeared frequently. As the middle of the fashion weeks, a lot of words disappeared and only few left, and at the end of fashion weeks, only two words: “show” and “new”, left. One reason may because the total amount of review has decrease over time and the other reason is that at the end of the fashion weeks, there was a “show” for “new” season and people focused on talking about these stuff.

**POS approach and focus on all the noun forms (NN, NNP, NNS, NNPS)**

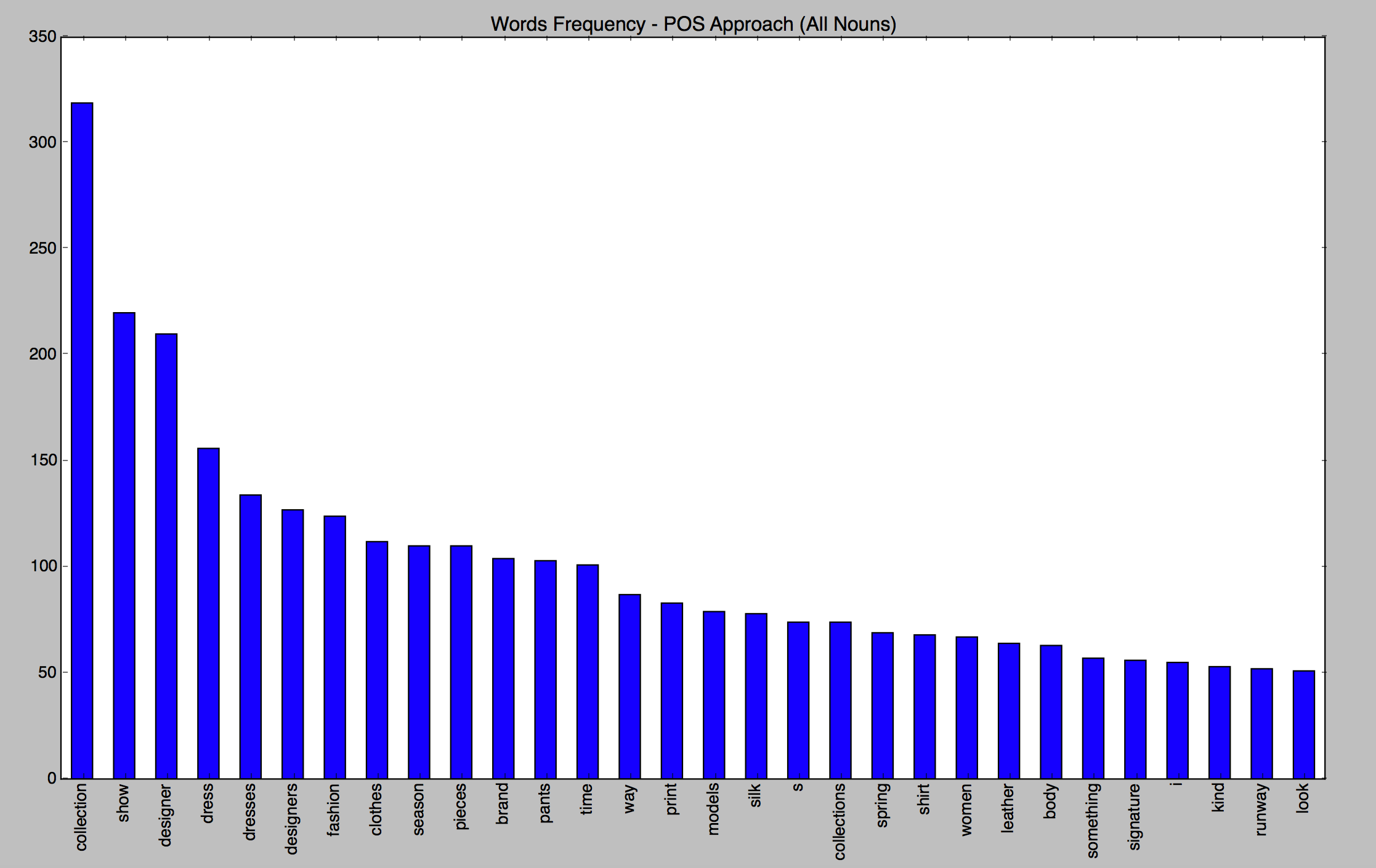


Figure 5. Words Frequency – Part of Speech Approach Using All Nouns

Usually nouns carry more information than other words. Using part of speech tags, the words will be labeled and nouns can be selected for analysis individually. The result is better than previous bag of words with stemming and stop-words removal approach, since a lot of verbs or adverbs have been removed. Nouns are more like concepts that other words.

Therefore, in Figure 5, top three concepts are “collections”, “clothes” and “Spring”. Actually the 2015 fashion weeks is for 2016 spring season. The words in Figure 5 are very relevant.

However, since the POS approach didn’t remove stop words, there are also some words without so much meaning included. Overall, the performance is acceptable.

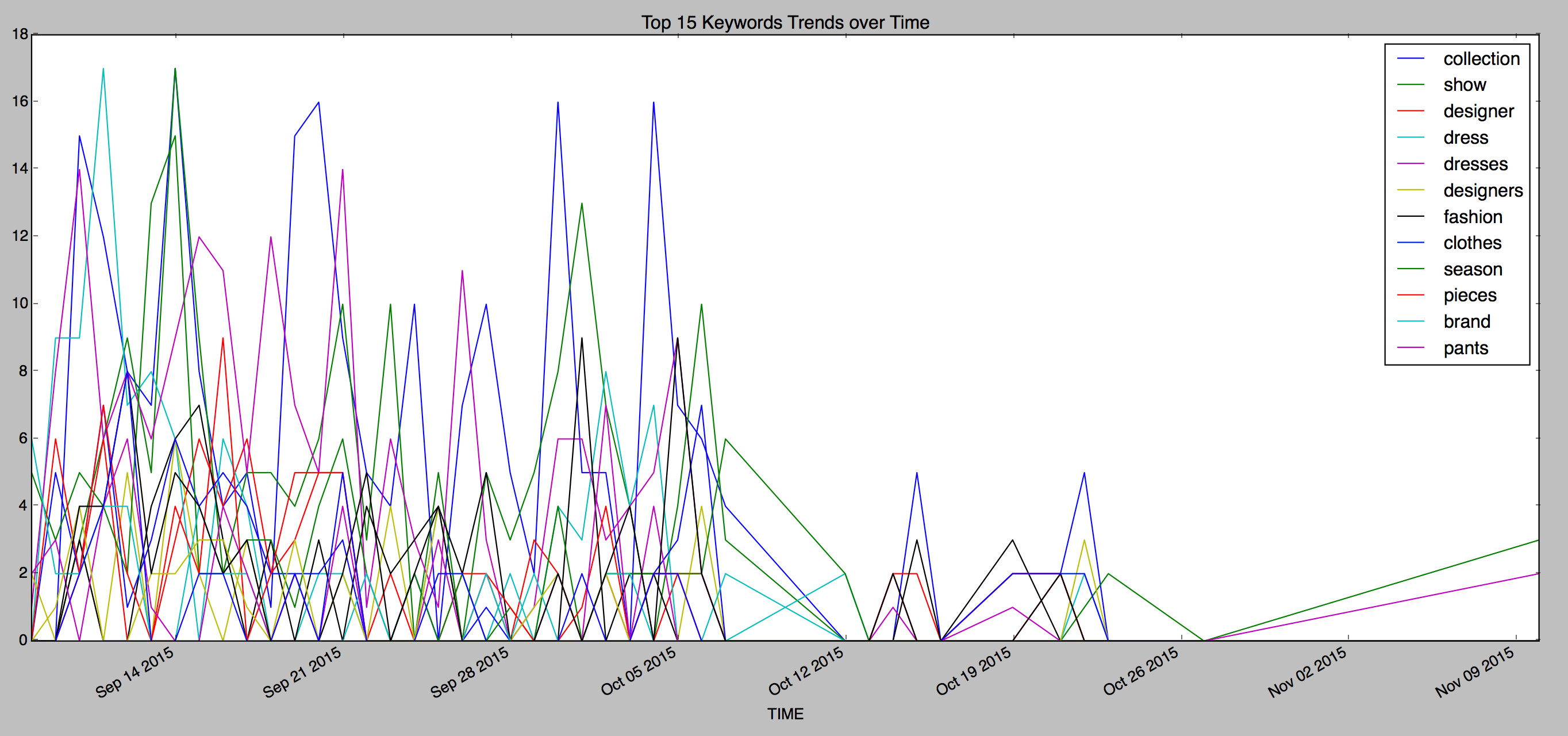


Figure 6. Top 15 Keywords Trend for Part of Speech Approach Using All Nouns

Figure 6 plots the most frequent 15 words among above 30 keywords for trend analysis. Similar to previous methods, there are many discussions on all kinds of things at the beginning of the fashion weeks, while, in the middle of the fashion weeks, some words disappeared and at the end of the event, two words have the highest frequent: “show” and “dress”.

**POS approach and only focus on NNP**

Using NNP only, the method will count NNP nouns which displayed in the Figure 7. As shown in the figure, the words are more like special nouns that contains special meanings. For example, “kitsch”, which appears the most times, means a low-brow style of mass-produced art or design using popular or cultural icons, according to Wikipedia; “bomber” jacket is a style of jacket. So this approach can locate how the fashion trend goes and what fashion styles that the reviewers are caring about. Although these words didn’t show a really high frequency, but they are indeed very important to be extracted and analyzed.

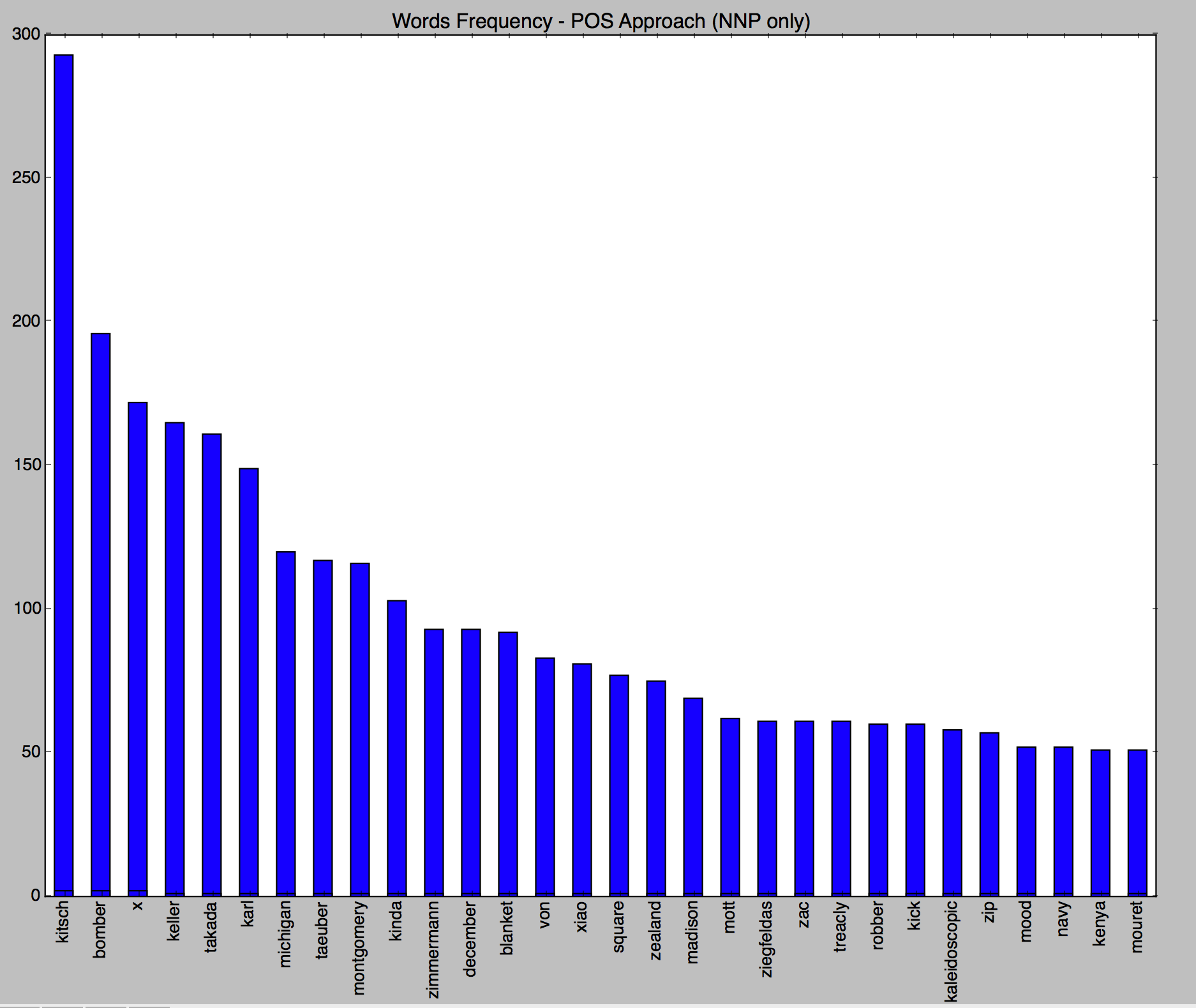


Figure 7. Words Frequency – Part of Speech Approach Using NNP Only

**Appendix:**

[1] Assign2.ipynb - detailed steps and analysis written in IPython Notebook

[2] ReviewAnalyzer.py - Core module for four approaches