****

**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.

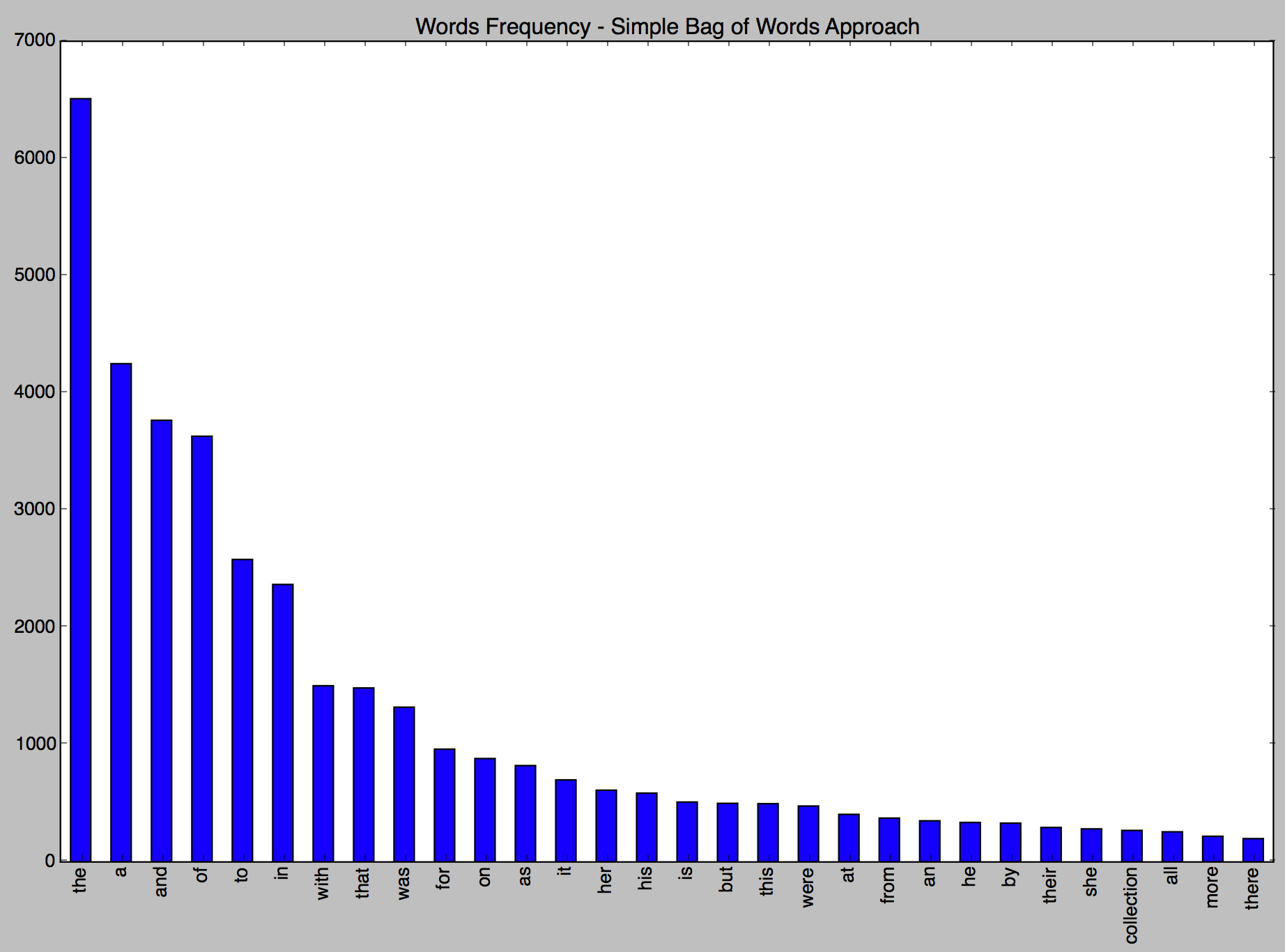


Figure 1. Words Frequency – Simple Bag of Words Approach

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.

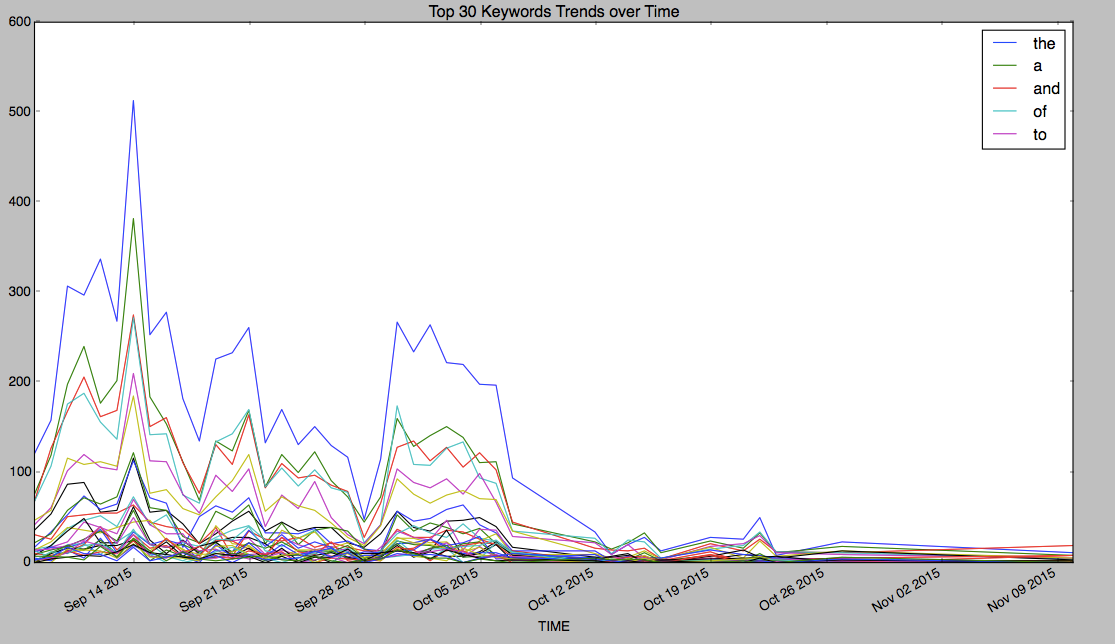


Figure 2. Top 30 Keywords Trend for Simple Bag-of-Words Approach

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 have

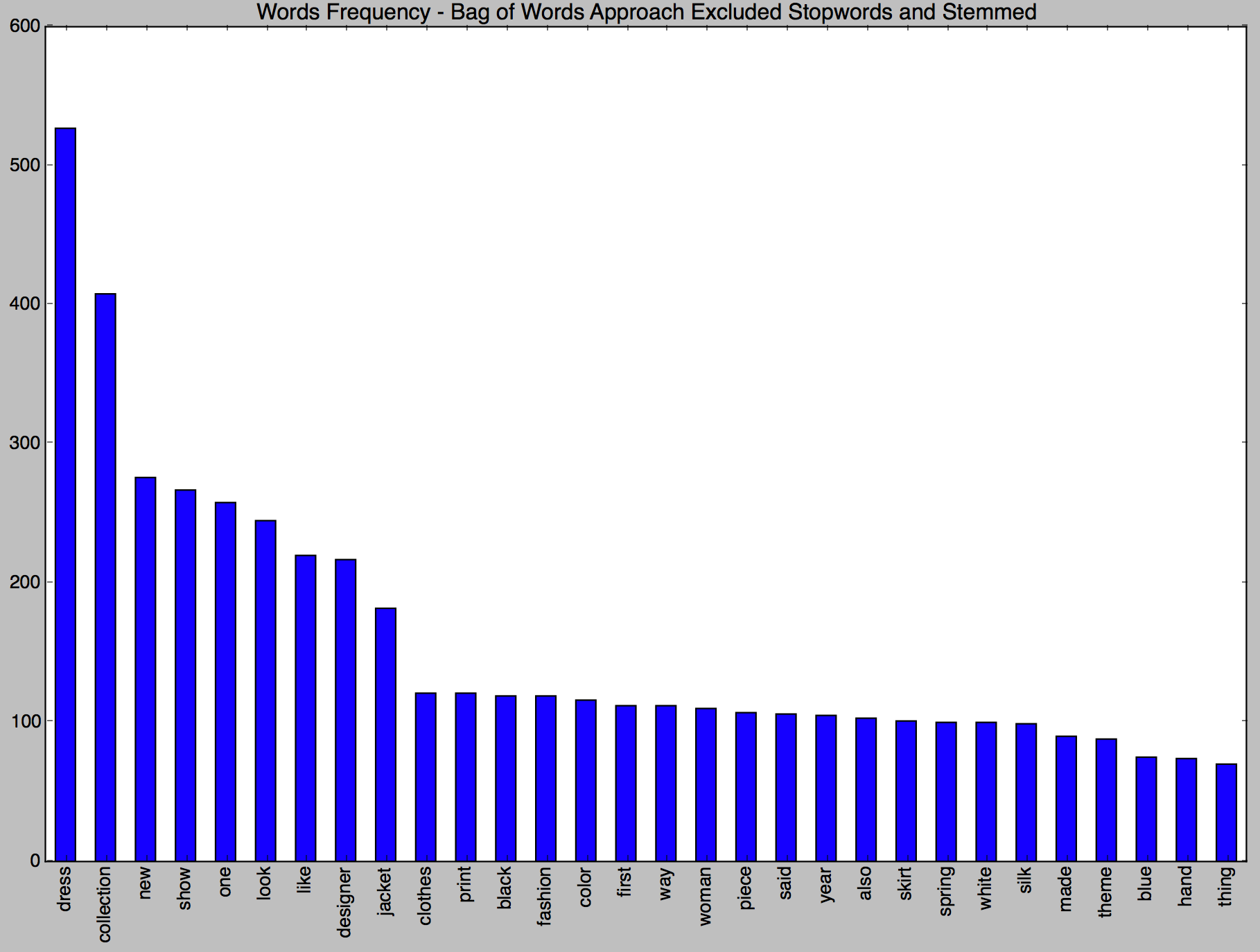
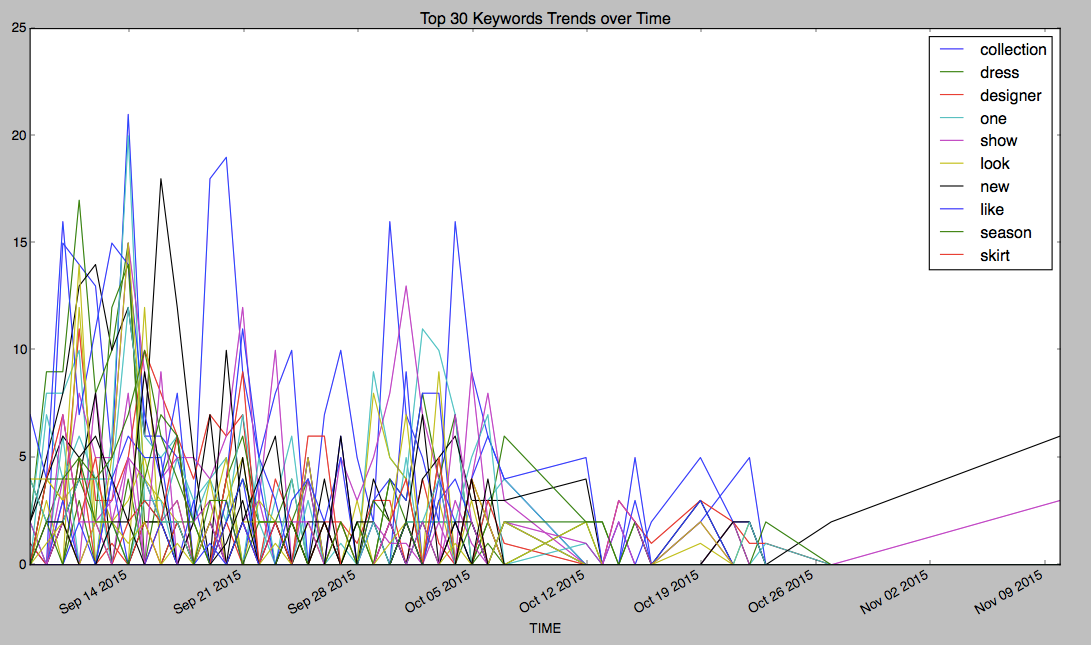
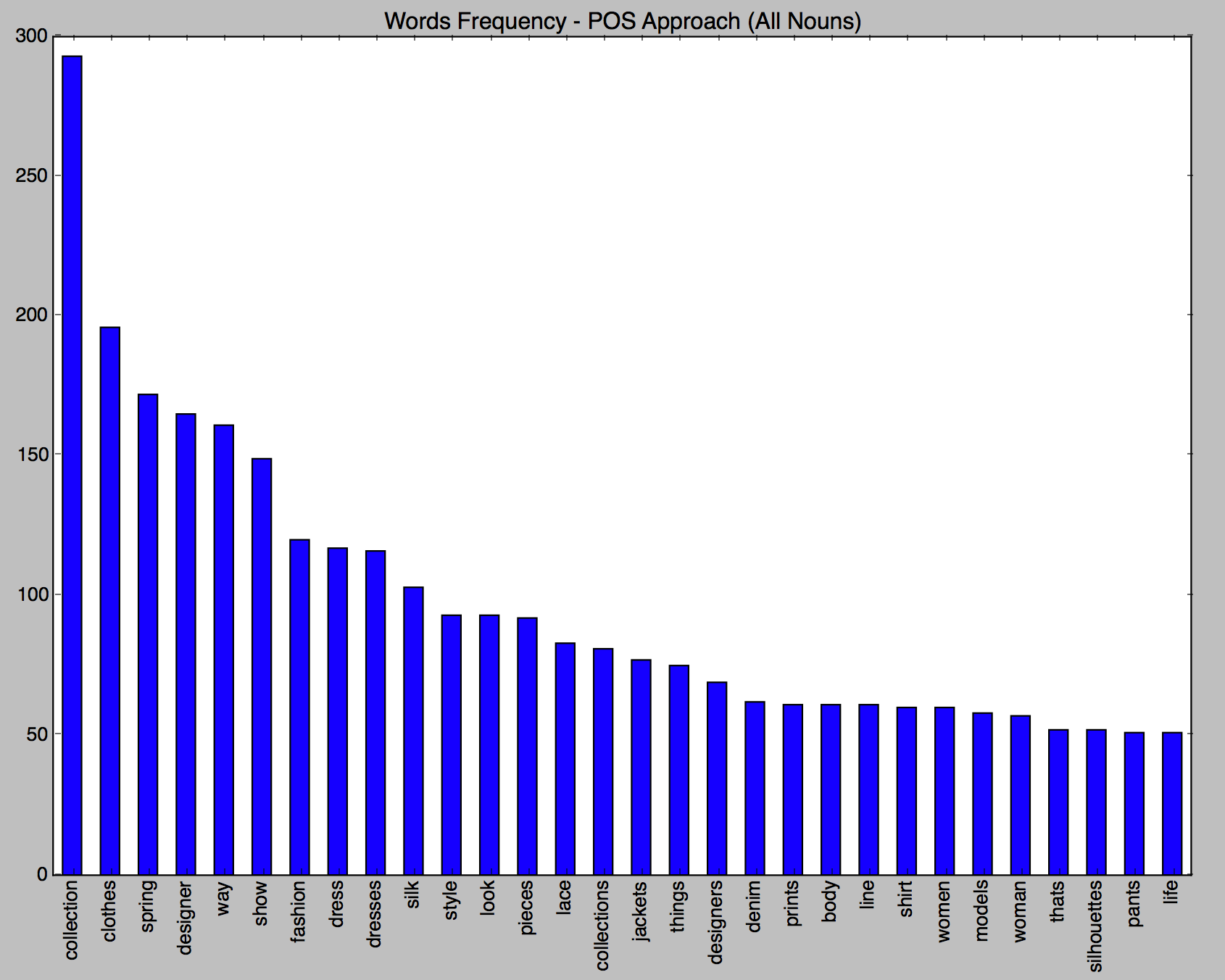
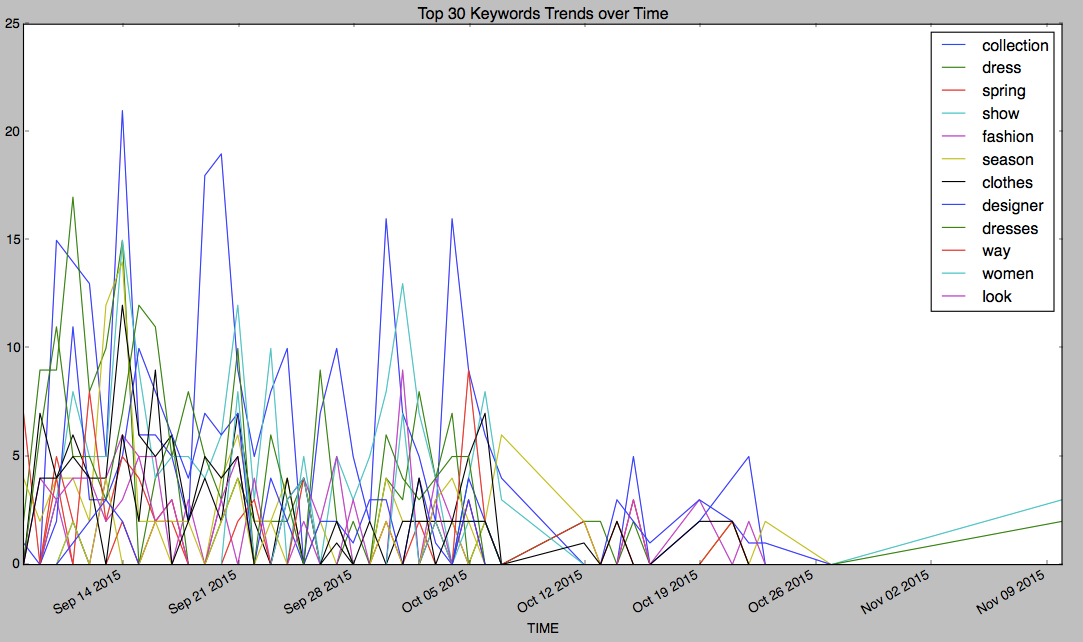


Figure 3. Top 30 Keywords Trend for Bag-of-Words With Stemming and Stop-words Removal Approach



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





**POS approach and only focus on NNP**

