Assignment #1

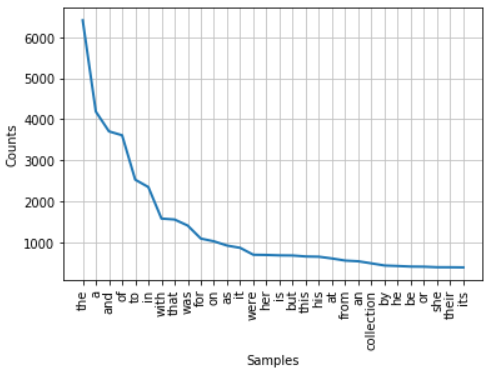
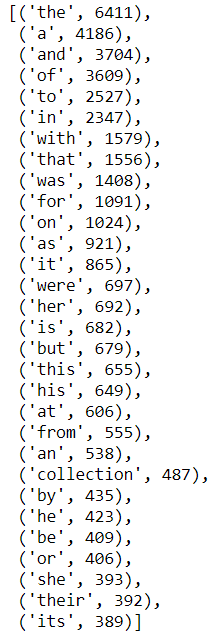
BYGB 7977 – section 1

Professor. Katsamakas

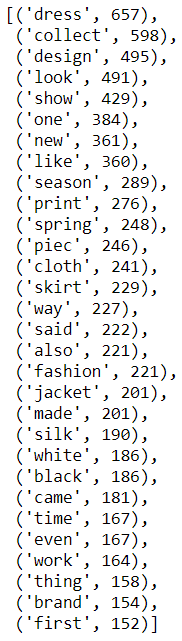
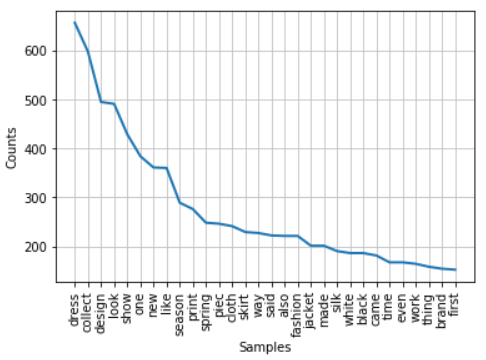
Hanzheng Wang

By analyzing the data with 4 approaches, the results of the top 30 keywords frequency (frequency table and plot) are as shown below:

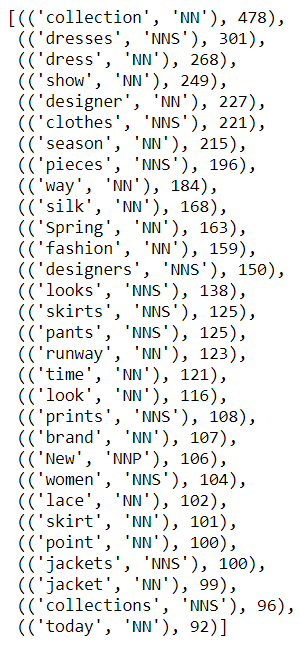
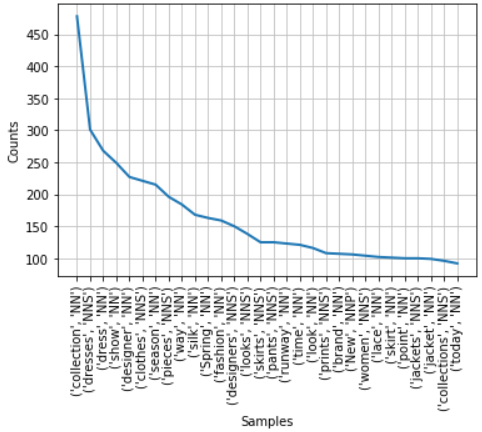
Approach 1: Simple bag-of-words



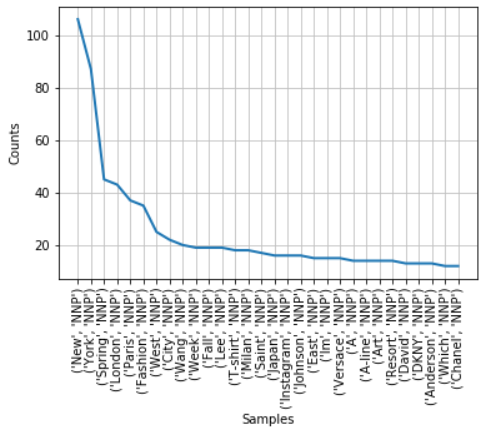
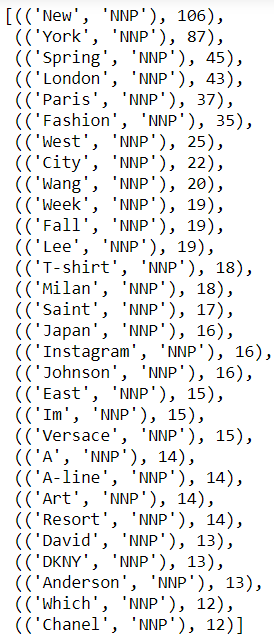
Approach 2: bag-of-words with stemming and stop words removal (Porter Stememer)

Approach 3: POS (NN, NNP, NNS, NNPS)

Approach 4: POS (NNP)



**Performance**

By comparing the results of the top 30 keywords frequency, approach 3 and 4 can provide more useful information than approach 1 and 2. Briefly, the rationale behind this is because that approach 3 and 4 provide the frequency distribution of nouns. And nouns can delivery most of the information in text documents. The task is to understand fashion trend instead of popular fashion brand, so I think approach 3 (POS approach which focus on all the noun forms) gives the best result.

**Potential fashion trend**

Based on the result of approach 3, the theme of fashion should be mainly around the design of spring (spring with 163 times) clothes. Customized product may be more appealing to the customer because the terms of frequency related to design are 337 (“designer” with 227 times and “designers” with 150 times).

For apparel categories, the result of approach 3 suggests us focus on gowns (“dresses” with 301 times), dress (“dress” with 268 times), skirt (“skirts” with 125 times), pants (“pants” with 125 times), skirt (“skirt” with 101 times), jacket (“jackets” and “jacket” with total 199 times), and etc.

For apparel material, silk may obtain a high popularity because “silk” shows up 168 times in the texts. What’s more, adding lace (“lace” with 101 times) design may also be the fashion trend of this spring season.

**Improvement**

In order to improve result of this case, we can apply the part-of-speech approach which focus on different type of tag. For instance, adjective (JJ, JJR, JJS) can provide and express more information about the selection and matching of colors.