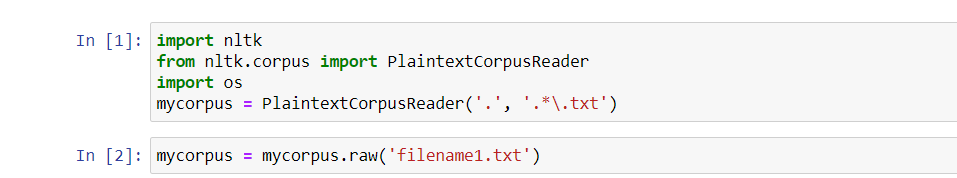
**Assignment – 2**

**Karan Ashar,** [**kashar@syr.edu**](mailto:kashar@syr.edu)**, 3/17/2020**

**Contents:**

1. **Data Processing**
2. **Descriptive Statistics**
3. **Unigram Analysis**
4. **Bigram Analysis**
5. **Sentimental Analysis**
6. **Data Processing: -**

I first imported my reviewText file which was submitted at the end of Assignment 1.



I then sentences tokenized it using nltk.sent\_tokenize(). I decided to approach the tasks individually to avoid complications.



**Question Sentences-**

I first started with extracting the question sentences. I filtered all the sentences having a question mark ‘?’ in the end using regular expression. I collected the filtered data in a list called ‘questions’. I then imported the treebank to conduct my tagging process. I will be using t2 (BigramTagger) as my first preference. t1 (Unigram Tagger) as my second preference and a ‘NN’ tag as my default. I then word tokenized my sentences using nltk.word\_tokenize(). I tagged the words after tokenizing them and stored them in a list called ‘taggedtext’. I set my grammar and parsed it using the nltk.RegexParser(). I then ran a loop to iterate through all elements in the taggedtext list. I then parsed a single element of the taggedtext which is nothing but a sentence tokenized and tagged. I then check each subtree in the tree created by the previous step to see if it contained the label for the adjective phrase defined by me. If the label existed I would store the sentence in the list called ‘question\_adj’. I also checked the list to avoid any duplication of data. I had my data ready for the question sentences in the ‘question\_adj’ list.



**Imperative Sentences-**

For the Imperative sentences I carried out the same procedure as above. I extracted the sentences with an exclamation mark at the end from the sentenced tokenized list. I then word tokenized them and tagged them. I then defined my grammar to check for imperative sentences. I then parsed it and stored the sentences containing the defined grammar tag in a list called ‘imperative’. I then carried out the same procedure to extract sentences which contain an adjective phrase and stored it in a list called ‘imperative\_final’. I had my data ready for the imperative sentences.



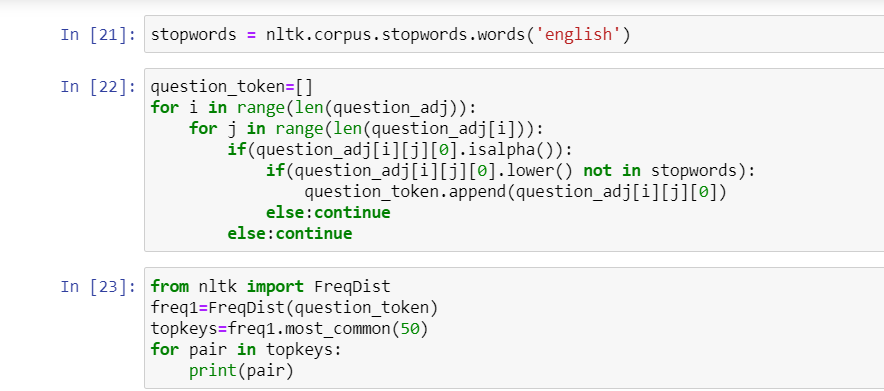
1. **Descriptive Statistics:**

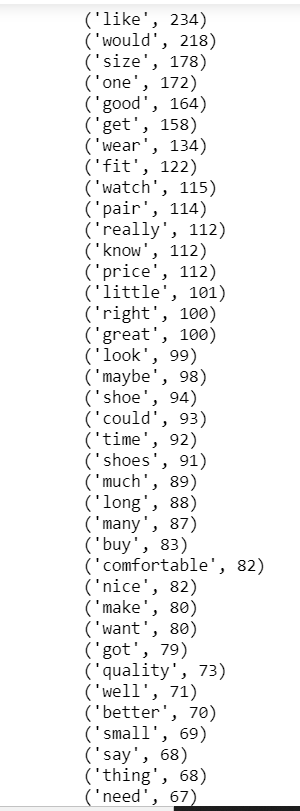
|  |  |  |
| --- | --- | --- |
|  | Question Sentences | Imperative Sentences |
| Number of sentences | 2250 | 2170 |
| Average Length of each sentence. | 24.668 | 18.779 |

Result – The number of extracted sentences is somewhat similar. What we can conclude is question sentences like longer than the imperative sentences. People tend to write more while questioning.

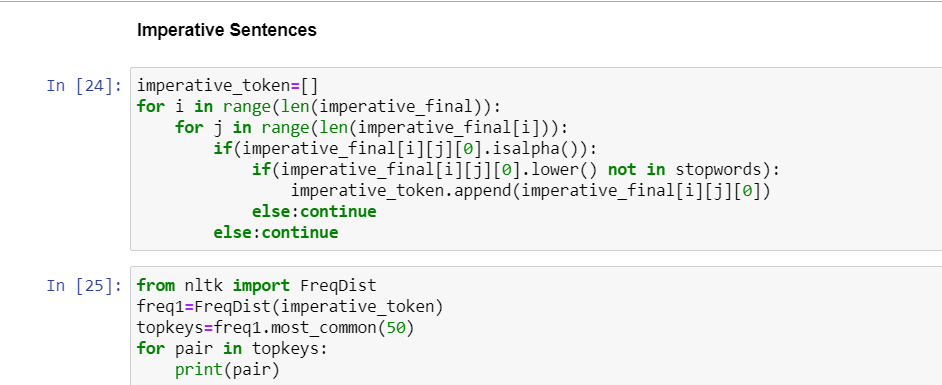
1. **Unigram Analysis:**

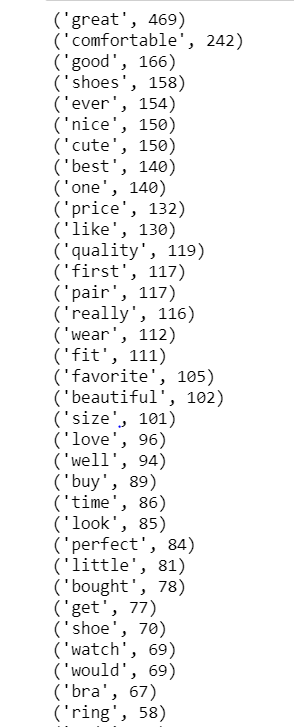
**Question Sentences-**  
I conducted the analysis for each list separately. I filtered the data first using isalpha() function and then seeing if they are not in the stopwords list. I carried a frequency analysis of these tokens. I found words like ‘like’, ‘good’, ’great’,’comfortable’ occurring many times. Which indicates a positive response. But since these are question sentences it is possible that these people are asking others who bought the product about it. But there were also words like ‘would’, ‘maybe’ occurring many times which indicates some kind of recommendation. They are certainly not negative words.





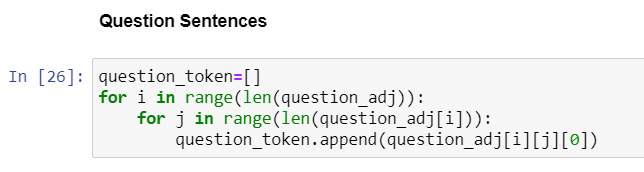
**Imperative Sentences-**   
I filtered the data in the same way as the question sentences. Here too we see words like ‘great’, ‘comfortable’, ‘good’, ‘nice’ occurring many times. This indicates a positive response.

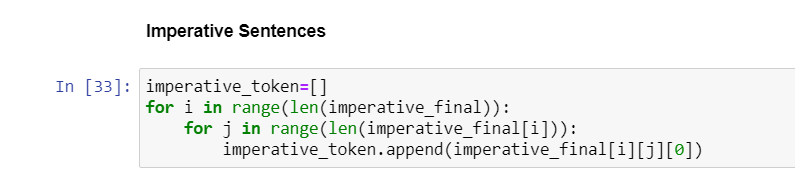




1. **Bigram Analysis:**

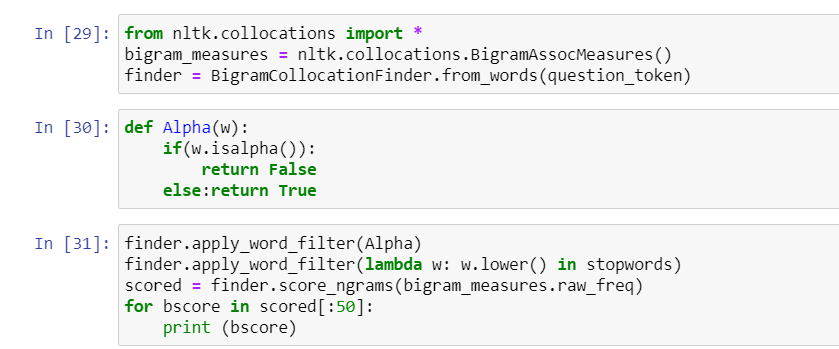
For the bigram analysis I first extracted the sentences from the ‘question\_adj’ and the ‘imperative\_final’ list. I then created bigrams using nltk.bigrams() function. I filtered the bigram using Alpha filter followed by removing stop words. I also filtered based on frequency using frequency filter before calculating the PMI score.

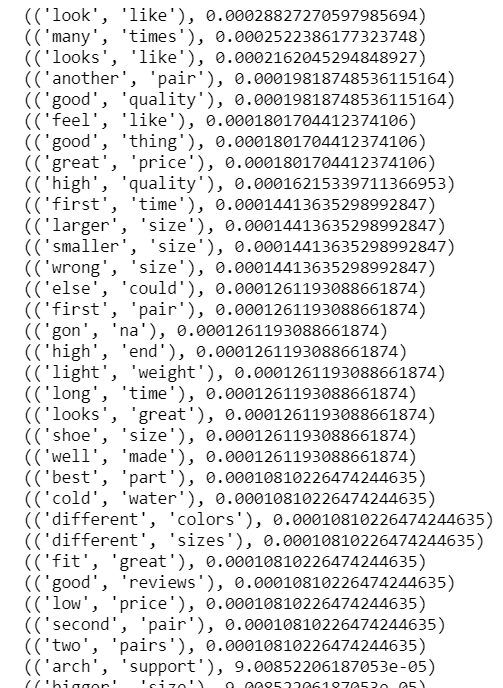




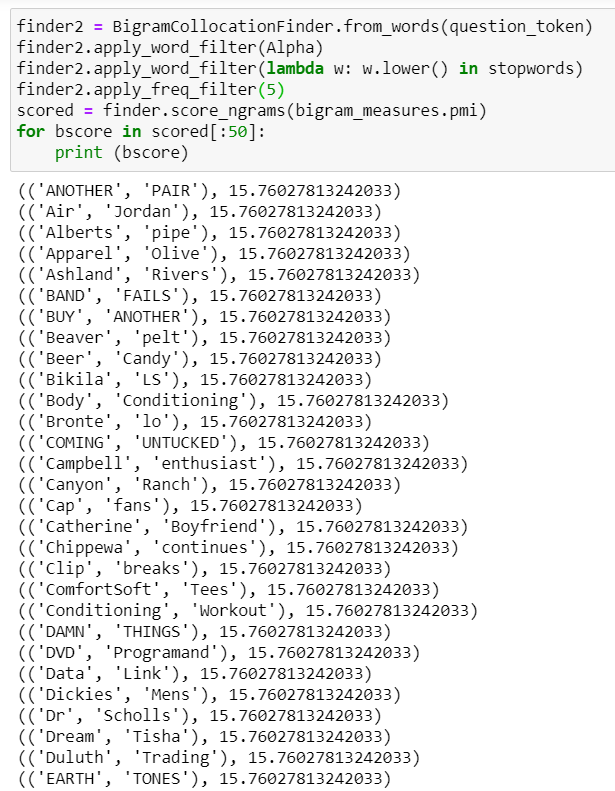
**Question Sentences-**   
For the question sentences we see bigrams like ‘good quality’, ‘good thing’, ’high quality’ many times. We also see bigrams with the word ‘size’ occurring many times. I think people are questioning ‘sizes’.  
The PMI score give out company names like ‘Air Jordan’.

Question Sentences Raw Frequency -





Question Sentences PMI -



**Imperative Sentences-**  
We see bigrams like ‘good quality’, ‘great quality’, ‘well made’, ‘really cute’, ‘comfortable shoes’. This indicates a very positive response. People seem happy and satisfied with the products they have purchased.   
The PMI score gives the name of the brand bought and mentioned by the customer.

Imperative Sentences Raw Frequency-



Imperative sentences PMI-



1. **Sentimental Analysis:**

We can give each word a score based on pre-defined dictionary of words and score. Then we can calculate the average score of each sentence. We can then set a threshold limit and then classify the sentence as a positive or negative sentence based on their average score (1 or 0). We can then count the positive sentences and negative sentences to find out the overall score of the text.