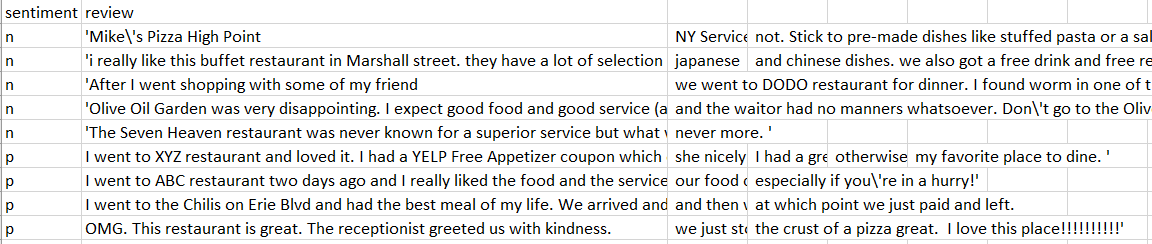
Rohini Shrivastava

IST 736

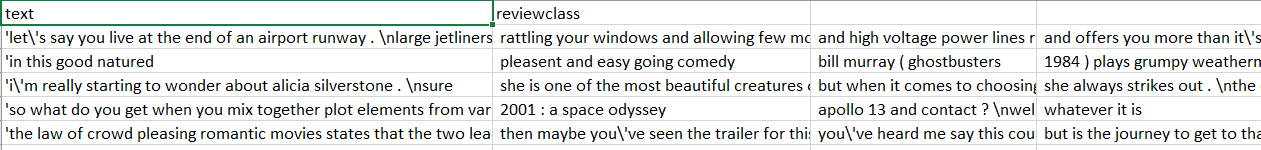
**Analysis:**

*About the Data:*

Two sets of data were provided as a Comma Separated Value (CSV) file. The first set contained information about restaurants and their reviews (*Figure 1).* This csv contained the sentiment (whether its positive or negative), the restaurant, and the reviews associated with that restaurant. The second file contained movie reviews (*Figure 2).* The first column contained text of the review, with the second column saying whether it was a positive or negative review. The restaurant data set did not need much preprocessing, however the movie data set had multiple errors inside the file that needed to be fixed.

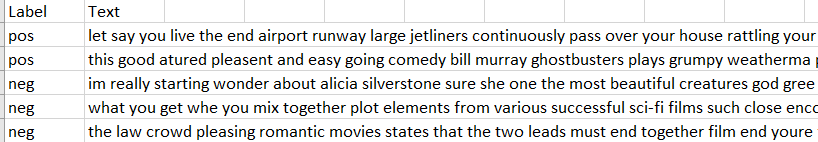


*Figure 1: Sample of Restaurant Reviews Before Pre-Processing*



*Figure 2: Sample of Movie Reviews Before Pre-Processing*

For each of the CSV files, a script was run to create a new text file and then clean up the data. The punctuation, such as “.”,”?”,”’”,”/” etc, were removed. Some of the words were capitalized so all those were converted to be all lowercase. The new line characters were also removed. The text file was then placed back into the CSV format with the appropriate labels and then made into a list so it can then be vectorized.

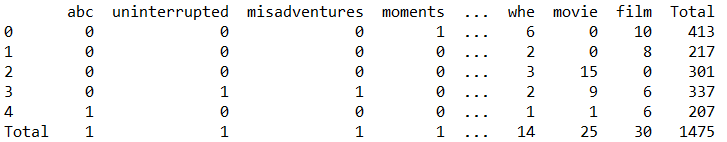


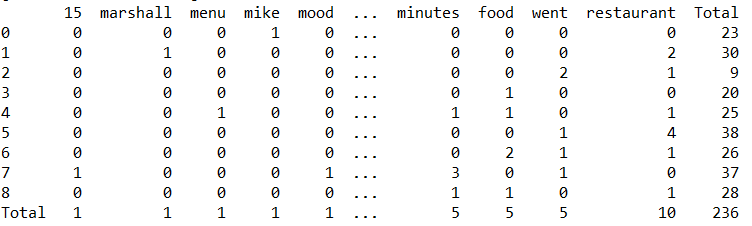
*Figure 3: Movie Reviews After Pre-Processing*

*CountVectorization*:

SKLearn is a python programming library that is used for classification, regression, and clustering algorithms. It uses both supervised and unsupervised algorithms. Inside there is a module called CountVectorizer. CountVectorizer converts documents into a matrix of tokens and their counts.

CountVectorizer was used to create a matrix of the two CSV files given. The files were imported through the OS library, which looks at the operating system interface, and then cleaned up. The CountVectorizer function was used on the files, and then was converted into a data frame. The data frame had labels of the CSV as the columns. Once CountVectorizer had been run, a total was then calculated for all the columns and all the rows. Each row contained the number of times the word from the column was present in the review. The column total calculated the total of words used in each review, and the row total calculated the number of times each word was present in the entire corpus. Then columns were sorted based on row totals in increasing sequence (*Figure 4)*. The restaurant reviews had 92 rows and 1256 columns, while the movie reviews had 5 rows with 912 columns.





*Figure 4: Sample of Vectorized Data and Counts*

**Results:**

All the reviews combined contained 1711 words, not including stop-words. The restaurant reviews had 236 words in total, whereas the movie reviews had 1475 words *(Figure 5)*. The top four most used words for restaurant reviews were “minutes”, “food”, “went”, and “restaurant”. These words follow in line with what is typically expected to hear about a food service. “Minute” may be relating to how long the wait was, “food” about what was eaten and the quality of it, “went” discussing when the customer visited, and “restaurant” referring to the place the customer ate. Movie reviews followed a similar pattern. “When” may be referring to a time, “just” is an adverb commonly used to explain one’s thoughts, “movie” and “film” referring to the what the user just watched (*Figure 4).*

*Figure 5: Comparing Total Words for Restaurant and Movie Reviews*

In the restaurant reviews, the word “restaurant” was used 10 times as the most used word. In the movie reviews, the most common word was “film” which was used 30 times. The rest of the three most common words for restaurant reviews were used 5 times each, while in movie reviews “movie” was used 25 times, “when” was used 14 times, and “just” was used 13 times (*Figure 6)*. While movie reviews had almost six times the amount of words compared to restaurant reviews, the most common words did not show a proportional difference.

*Figure 6: Comparing Number of Times Top Four Words Were Present*