Francesca Giovanucci and Brian Collins
Problem Solving & Software Design
Assignment 3
October 30, 2016

1. Project Overview

The data source we used were movie reviews from the IMDB website. We scraped the text from the Lone Survivor review and saved it as a .txt file in Python, as we have previously done in class. The techniques we used were Characterizing by Word Frequencies, computing summary statistics to find the top 100 most frequent used word, and Natural Language Processing to assess the opinions of the user who wrote the review. We used the techniques to give us an overall impression on the positivity/negativity (sentiment) of the review. Additionally, we wanted to provide context of the review to someone who was completely unfamiliar with the film.

2. Implementation

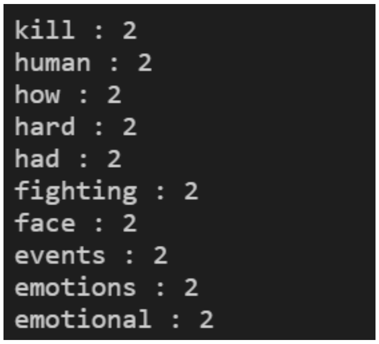
For the design, we wanted to provide the user with a clear description of what data they were being presented. For this reason, after all of the functions are defined, we have printed statements at the end of the code to communicate the significance of the information. We wanted to make the visual results to be clean and readable, focusing on using spacing appropriately. Initially the code will show the user words used within the text. Then they will be shown the overall sentiment of the chosen text.

The major components of our assignment compose of file imports for text file, the count of total/unique words, and lists. A histogram is created to find the words used, and the quantity they are used at. Additionally, lists/temporary lists were created and utilized within for-loops to extract certain information about the words used (for example for-loop that appends the most frequent words used and orders in descending manner). The structure we proceeded with was consistent to what we've done in the course so far, we noticed that we were able to understand the code better and run it effectively with structure we are familiar with (trying/exploring new things was giving us errors – which is valuable for learning, but for execution purposes we stuck with what we've seen in the course).

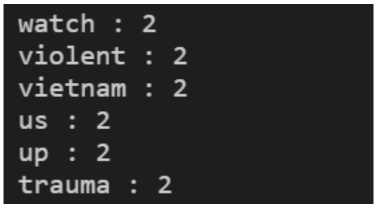
For the most part we kept our design pretty consistent and uniform. We noticed that a lot of code didn't run properly when each of our designs were not the same. We witnessed the importance of keeping a consistent coding environment, especially in a collaborative environment. However, we also learned that with different designs, come different capabilities for the code you are writing.

3. Results

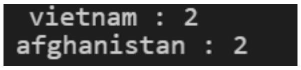
We found many interesting results with the text analysis techniques we used, starting by finding certain words that are strong indicators of the overall context of the movie. We found that it could be beneficial to identify a list of key identifier words (after getting rid of articles, characters, etc), and present the user with all of the words that appeared within the review (for example: 'violent' 'trauma' 'kill'). From these words, the user has a good idea of what kind of movie they should expect to see. Additionally, by seeing 'Afghanistan' and 'Vietnam' appear, we saw a potential opportunity to create a list of all the countries in the world, and define a function to present any country mentioned in the review – this could provide the movie's geographical context to the user without having to read anything. Certain words had strong correlation to the sentimental aspect of the movie. We found that presenting the user with key emotional indicators (for example: 'sad' 'crying' 'touching') gives the user an idea of the emotional reaction they can expect to have watching this movie (whether they are in the mood to laugh, or cry, or be scared).



```
kill : 2
human : 2
how : 2
hard : 2
had : 2
fighting : 2
face : 2
events : 2
emotions : 2
emotional : 2
```



```
watch : 2
violent : 2
vietnam : 2
us : 2
up : 2
trauma : 2
```



```
 vietnam : 2
afghanistan : 2
```

One of the methods we chose to do was the Natural Language Processing method (NLTK). NLTK tells the code to use sentiment analysis and full sentence parsing to establish the mood of the context within the desired piece of writing. In this user review of the movie Lone Survivor we analyzed the type of wording used to understand if the review was biased in a positive or negative way or if it was more of a neutral review. In this review the language is roughly 67% neutral with slightly more negativity. This however, knowing the context of the movie could be a misrepresentation of message the user is trying to get across. This method is so useful because if you have a document that is heavily biased in one way or the other you can start to learn more about the meaning behind it and along with other text analysis you can decipher when words are common or emphasized what kind of meaning they most likely convey.

4. Reflection

What went well was the importing of the text data. Scraping and creating a txt file within Python was most fluid way to conduct text analysis. Something to improve is removing the “useless” words within a text in order to make text analysis more efficient. That is something I would like to improve in, going forward in text analysis: mastering the functions used to clean text thoroughly (before mining). Going forward, I’ve noticed that proper nouns can give great insight, and it is important to ignore punctuation/capitalization.

We planned to do the assignment together, with equal distribution of work, while assigning tasks based on our interests. After going through the assignment requirements we understood that each of us would need to contribute to the overall write up. We agreed on what each of us wanted to get out of this project, and we agreed on a subject to explore (movie reviews). After making sure the text data was properly imported, we chose text analysis techniques that we wanted to work on. After choosing the techniques, we each did our work and then met together to combine and explain the results. There were no real issues besides different schedules, but we were able to meet, be on the same page, and complete the work. If there was anything different to do next time it would be to explore different packages that could helped us for more advanced analysis (pandas, numpy).

"I pledge my honor that I have neither received nor provided unauthorized assistance during the completion of this work."