Allison Fernández Prof. Zhi Li

November 12, 2018 MIS 3640

Assignment #2

Project Overview

For this assignment I used IMDB as the data source. I retrieve the first review of the movies included in a dictionary. The program will analyze each review and determine the top words in each review. When I decided to work on this I hope to be able to analyze each review individually and then analyze all the reviews as a whole to see any relationships or discrepancies among them.

Implementation

For this assignment I used several components and data structures. I used functions, strings, dictionaries, external documents, lists, histograms, among others. The program is pulling movie reviews from IMDB which are bodies of text in the form of strings. So, in order to analyze the words within the review like I wanted to, I had to first process the review. Because the review consisted of many words with a single string, I had to strip the entire string into individual words so that I could proceed to analyze the text.

As mentioned before, I wanted to see which were the top words used in each review and the frequency of these words. However, I did not want to count stop words because they are used very frequently and they do not say anything about the content of the text. To prevent this, I imported a document containing the stop words that I did not wanted to count in the function where the most common words would be determine, and inserted an "if statement" instructing to review the histogram of words in the review and adding to a new list those words not included in the stop words document.

After this all that was left to do was to print the most common words and their frequencies in a way that was easy to read and understand. For the printing the data, I decided to use '\t' to format the numbers because even thought I could not get all the numbers perfectly aligned this way due to the length of the words, it was still better than using '{:>}' to align the text. At the end, in the main() function I created a for loop so that I could run each of the movies in the dictionary through the entire analysis process.

Results

After running the program I was able to see the top five words for the first IMDB review of each movie in the Harry Potter series and how many times in the review each of these words appeared. If I want to run this program with other movies, all I have to do is change the dictionary at the beginning of the program. This dictionary contains the movie title as the key and the IMDB movie id as values. I chose to do a series of movies but it can be any unrelated movies and in the amount I choose.

In two of the reviews the top word is a blank space. I tried to get rid of it but could not because I was not able to figure out what was causing it although I suspected that the cause was the blank lines created between paragraphs by hitting ‘enter’. At first, I tried checking if it was a space by adding a space to the stop words document but that did not work. I tried other things but could not fix it.

Although I was able to determine that the top words used among the reviews were “film” and “harry,” I could not use the program to determine it. The program produces the list of words and frequencies for all of the reviews and by going through them I was able to see this. However, I wanted to be able to take all of the lists of words and frequencies for each review and produce new data determining the top words for the reviews were as a whole. I was not able to find a way to do this. Maybe the architecture I used was not ideal for this kind of analysis.

Reflection

I think that something I did well was that before starting to write the code I thought about the different components that my program needed to perform the way I wanted it to. This way I was able to write down in a piece of paper all the components and then piece the program together by organizing the steps in an order that would be logical for the program to run properly. As I thought about the different components I wanted, I would also research ways to perform these steps. Something I could’ve done better was to ask for help on how to use the imdb functionality in python because I was learning as I went and made some compromises along the way just because I was not familiar enough with the process to do what I really wanted to do. This and other factors influenced in the scope of the project. I think that the scope is ok, but with my initial idea it could’ve been a little broader and useful.