Ai Labs Data Science Intern Assessment Write Up

Hilary Melroy

The data provided for this assessment is a csv file comprised of titles, genre categorization, summaries, and a few other features of ~42,000 films. From this data, the top five genres was obtained, the words characterizing each of these summaries was assessed, and the summaries for all the movies in the dataset was investigated for evidence of Zipf’s law.

To begin, the dataset was read into a pandas dataframe. A dictionary comprised of the genres as keys and the frequency of occurrence as values was constructed by iterating through each of the entries in the ‘genres’ column of the dataframe. While the data appeared to be structured as a list of strings, each entry in the ‘genres’ column was actually one large string; thus each string was stripped of non-alphanumeric characters and then split on the commas separating each word. Once the genre data was in a list format, the dictionary was constructed. A simple dictionary counter function was used to return the top 5 genres in the list, which are as follows:

{'Drama': 19134, 'Comedy': 10467, 'Romance Film': 6666, 'Thriller': 6530, 'Action': 5868}

Next, the words characterizing the movies with each of the top 5 genres was analyzed by visualizing the top 100 words with a word cloud. Similar to the procedure for obtaining the top 5 genres in the dataset, a list was constructed with all the words in each of the summaries of the movies whose ‘genre’ column contained the genre of interest. The summaries were first stripped of non-alphanumeric characters, and were then split on the spaces between words. It should also be noted that the function that produced the word clouds corrects for stop words, so that words such as ‘of’, ‘the’, etc, were not included in the word clouds. Figures 1-5 represent the words characteristic of each of the top 5 genres.

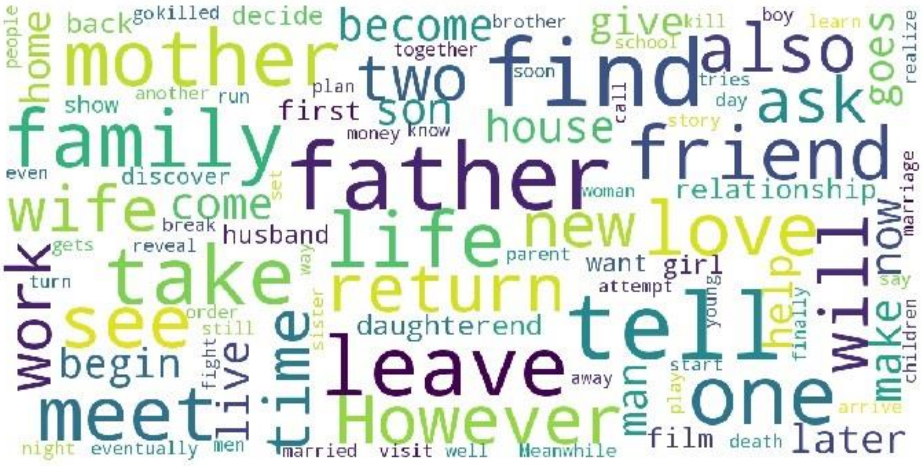

Figure 1: Word cloud representing the top 100 words of summaries whose movies are categorized as ‘Action’

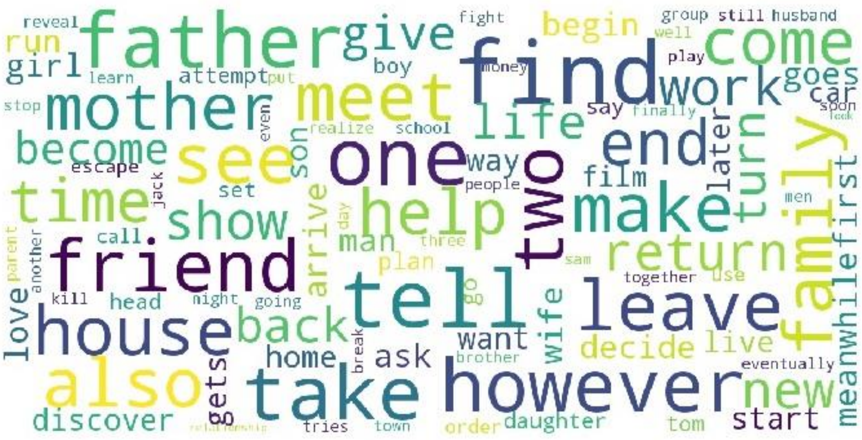

Figure 2: Word cloud representing the top 100 words of summaries whose movies are categorized as 'Comedy'

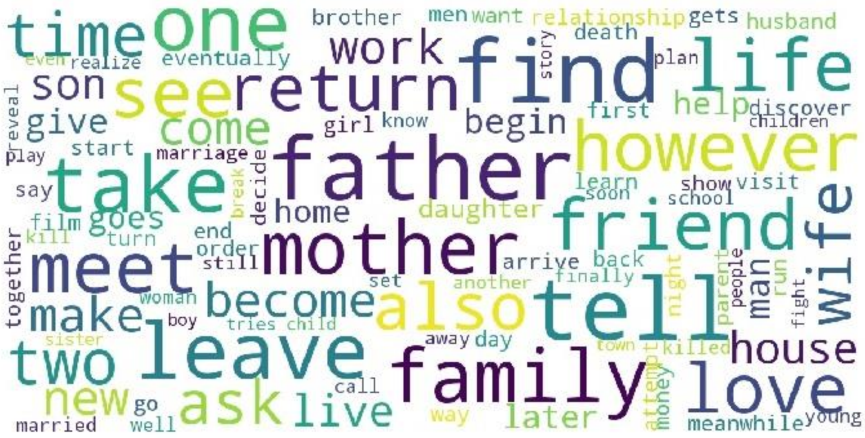

Figure 3: Word cloud representing the top 100 words of summaries whose movies are categorized as 'Drama'

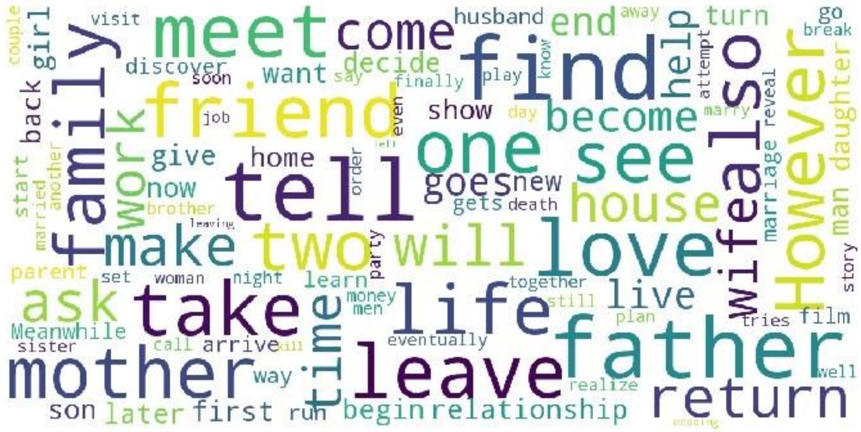

Figure 4: Word cloud representing the top 100 words of summaries whose movies are categorized as 'Romance Film'

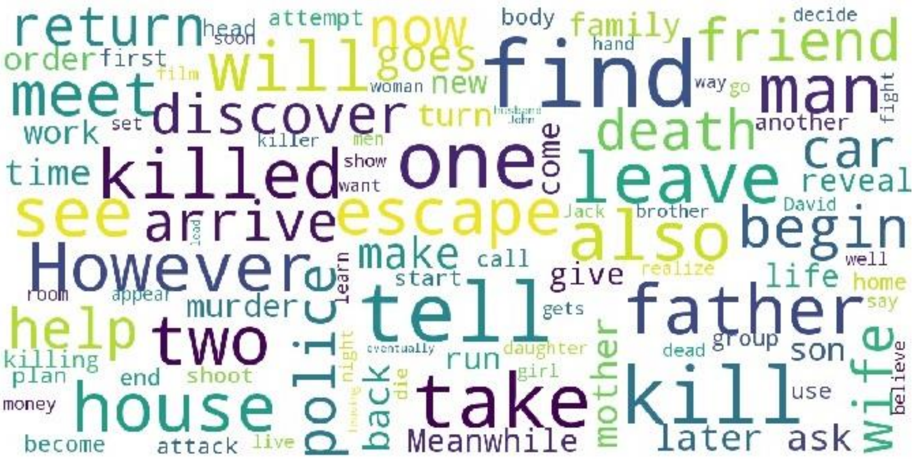

Figure 5: Word cloud representing the top 100 words of summaries whose movies are categorized as ‘Thriller’

The word cloud represents the top 100 used words in each list of all the words in the collective summaries of a given genre. A larger word corresponds to that word having a higher frequency in the word summary list. It should be noted that very many of the movies in the dataset had multiple genres, so some of the word clouds look similar. For example, drama and romance have very similar word clouds because there is overlap in the summary lists for these two genres.

Finally, the empirical observation known as Zipf’s law was investigated using the words in the summaries of all of the movies in the dataset. Zipf’s law states that the most frequent word in a given text corpus will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. By constructing a dictionary of the words in the summaries of all of the movies (not corrected for stop words), the top 5 words were obtained and shown to be

{'the': 735984, 'to': 477724, 'and': 454993, 'a': 362058, 'of': 260834}

A bar chart was constructed to visualize the distribution of the top 5 words, which is shown in Figure 6. Examining the distribution of the top words in the summaries of the movies in the dataset, it appears that Zipf’s law is roughly observed. As illustrated in the figure below, the bulk of the list of words in the summaries are made up of the very common stop words such as ‘the’ and ‘of’, and the frequencies of the words quickly drops off as the rank decreases.

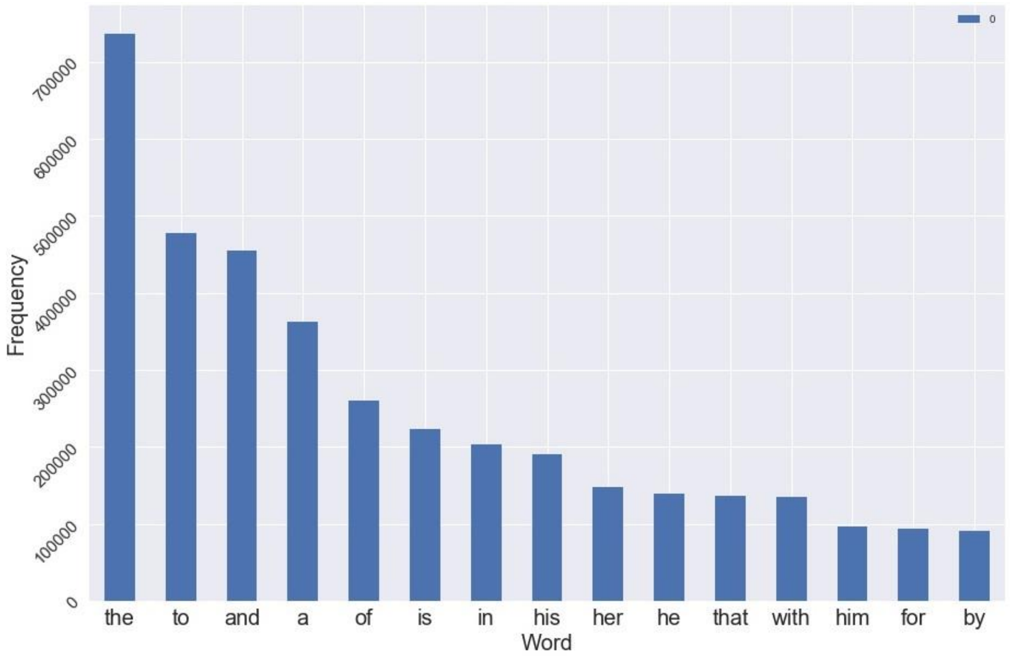

Figure 6: Distribution of the top 5 words in the summaries of all the movies in the dataset, illustrating Zipf’s law.