Project 1 Report

Steven LaCroix

October 5, 2023

1 – Summary

- How do you find if a word is significant when compared to chapters within the book or even other books? What words used in a book like Treasure Island are unique within a single chapter? Are there any words frequently present that aren't often used anymore? All of these will be answered throughout this report.

- We can take a book, like Treasure Island, split it up into a list of words, and calculate a few things to analyze how unique a word is. Things like term frequency, inverse document frequency, and TFIDF scores are all helpful tools for seeing how unique a word is within the chapter of a book.

- It was found that occurring in every chapter of the book or having a low term frequency can hurt a word's TFIDF score. We found that The word "bill" is the most unique in the book, having the highest TFIDF score.

2 – Introduction

Treasure Island was an adventure novel published back in 1883 telling a story of buccaneers and buried gold. Since language evolves over time and given that the book is so old there are a lot of words used back then that aren't so popular anymore. I wonder if any of those weird words are used a lot throughout the book? Which words are most important to their respective chapter compared to the other chapters? How many words are just kind of throwaway words (used everywhere and all the time)? After running a TFIDF analysis on Treasure Island, these and many more questions will be answered. A TFIDF analysis is essentially finding a score for each word that takes into account the frequency it appears in a chapter combined with how many chapters it exists in in the book. This gives us a score for each word in each chapter. The higher the score, the more significant a word is to that chapter, especially compared to the rest of the book. This analysis will begin with laying out the methodology of the analysis. Explaining the processes used and strategies implemented. After that we'll cover results to see what the analysis revealed. Finally, some conclusions will be made. My goals for this analysis are to efficiently and effectively determine the most significant words in each chapter and to present that in a clear and concise way. With that being said, let's move on to the methodology.

3 – Methodology

3.1 Data Collection and Pre-Processing

The goal of this first part is to obtain the book and convert that to a list of all the words. To do that, a .txt file was downloaded from Project Gutenberg and brought into the coding environment ready to be analyzed. This file was a bit messy to begin with, though. To fix this in the easiest way possible, a few things were removed by hand which would allow parsing the book in the later stages to be much easier. Those were: the Project Gutenberg notes, authors note, and the table of contents. Now that we have a clean file, it was read into the coding environment and immediately converted to all lowercase (again for simplicity of parsing). After that, all of the chapters were split up so that they can be analyzed individually. We then went ahead and removed anything other than a lowercase letter from these chapters. Any punctuation, quotations, whitespaces, special symbols, etc. were removed. We don't want these to get in the way of looking at word frequency later down the line. At the same time all of the remaining words were split up into a list. This was so that words can be looked at individually rather than a combined chapter. When analyzing word trends there are certain words we don't want to include. Words like "the" and "a" which are so common that they oftentimes just take up space during an analysis. Because of this, a list of about 130 of the most

common English words were removed from all chapters. Now, our book is split into individual chapters without any punctuation, uppercase letters, or unnecessary text to get in our way.

3.2 Processing and Storage

To continue, we want to further divvy up our chapters so we can start looking at the individual words within. The next step in this process is to find a list of every single unique word across all chapters. Not every word is in every chapter, so we need to combine our lists to get one complete list. Once that's done, we then can go through each chapter and count the frequency of words within. All this information is stored in a table that has each chapter as the rows and each word as the columns. So, the number of occurrences of a word are stored in the correct row and column according to the chapter it's from and the word it is. After this, we want to find how often a word occurs. The first step of this process means finding the total number of words in the book. Once that's done, we can go through each chapter and word and divide the occurrences of that word by the total number of words. Not only do we want to find how often a word occurs in the book, but we also want to know how many chapters its present in. To do this we want to find the inverse document frequency for each word in the book. Inverse document frequency takes in the number of chapters a word appears in, combined with the number of chapters in the book and spits out a value saying how unique the word is. If a word is present in more chapters, the lower it's inverse document frequency will be. That's because the word is less unique across the book. Lastly, we want to find the TFIDF for each word in each chapter. This measures how relevant a word is to a given chapter compared to the rest of the chapters. To find this we simply multiply the term frequency by the inverse document frequency for each word in each chapter and store that in another table. Now we have a list of inverse document frequency values for each word, a table of term frequencies for each word in each chapter, and a table of TFIDF scores for each word in each chapter.

3.3 Analyzation and Visualization

Now that we have our three main components, those being term frequency, inverse document frequency, and TFIDF, we can now start to compare the chapters. I'll go over some interesting things I decided to look at. There are many more possibilities, but these are the ones I went with. The top 3 term frequencies in each chapter. By simply sorting the term frequencies for each chapter and isolating the top 3 in each, a bar chart can be made to compare the most frequently occurring terms in each chapter. Another is seeing how many words occurred at each inverse document frequency value. This can be seen by summing the total number of words when grouping words by their inverse document frequency calculation. Showing this through a stem plot is the best course of action. Doing a similar thing as term frequency, but for TFIDF, we can see the most relevant words to their respective chapters.

4 – Results

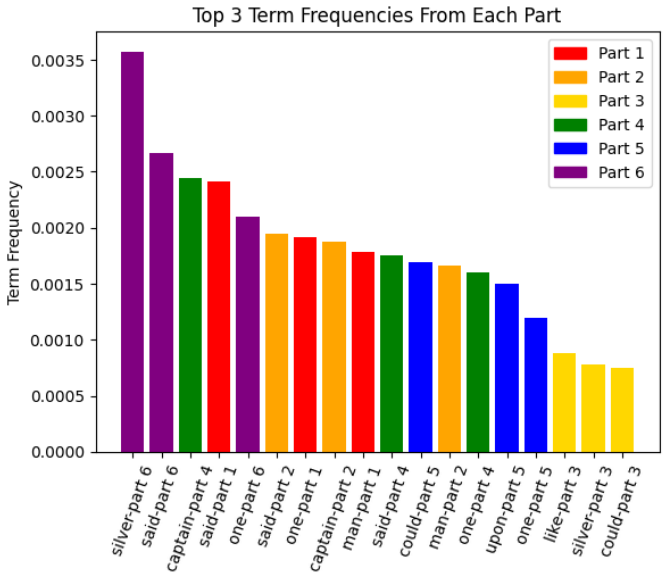

Figure 1: Top 3 Term Frequencies From Each Chapter

After creating some visualizations and delving into the TFIDF scores (and other components) of the various words in the book I have come to a few conclusions. Firstly, words that occur a lot in one chapter are very likely to occur in many more chapters at a high frequency. Because of this, many of our popular words end up having extremely low TFIDF scores. Their term frequencies are extremely high because they occur super often, but the inverse document frequencies are often low or zero. Once the TFIDF is calculated many of these words are not seen near the top because of that. As you can see in Figure 1 the words “upon” and “like” are the only ones to not be repeated on this bar chart across multiple chapters. However, both of those words have TFIDF scores of zero because they are present in every chapter regardless. Even the word “silver” has a TFIDF score of zero despite being the most popular word in chapter 6 by a large margin. Now, if we were comparing chapter 6 of Treasure Island to a futuristic sci-fi book, the TFIDF score of “silver” would probably be super high since it’s not very likely to be present in all chapters of that book. However, that is not the case here.

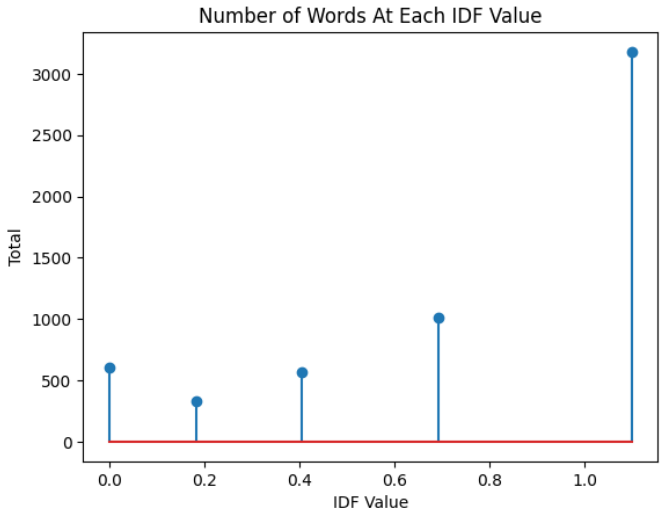

Figure 2: Word Count at Each IDF Value

There are a lot of unique words that only occur in one chapter throughout the book. The thing is a lot of these words may only appear once in the entire book. These words would have an extremely high inverse document frequency value corresponding to them, but a super low term frequency, thus making their TFIDF scores low as well. As you can see in Figure 2, there are well over 3000 words that appear in just a single chapter. Having a zero IDF value means the word was present in five out of six chapters or all six chapter. This means it's not unique in the book and so it's IDF score is low. There were about 600 such words, and all the words we covered in Figure 1 were in this list. This graph shows us that in the book there was a lot of variety between words but when they were repeated, they were repeated a lot.

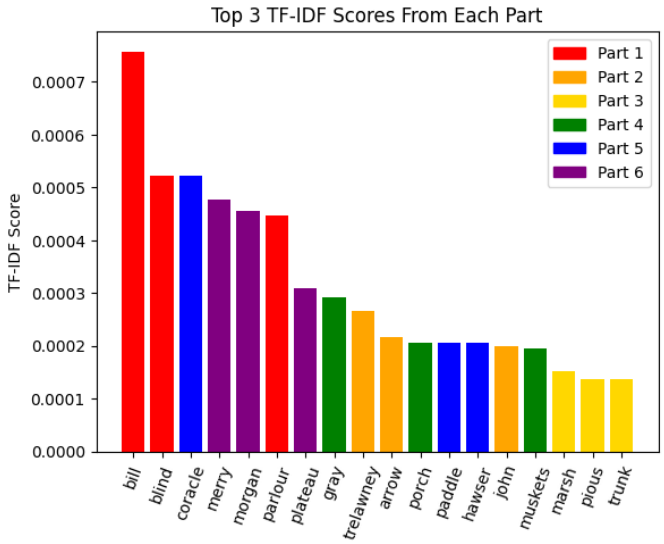

Figure 3: Top 3 TFIDF Scores From Each Chapter

Now that we've seen the two ways for words to get a low TFIDF score, here are the words that actually have a high score. As you can see in Figure 3, "bill" has by far the highest TFIDF score. To obtain that this word must have only been present in a singular chapter and been repeated a decently high number of times. When looking at the numbers, that is the case. This word appeared 22 times in the first chapter and nowhere else in the book. That means this word is super important to the first chapter but is not important at all anywhere else in the book. The word "blind" appeared 24 times in chapter one, 3 times in chapter two, and nowhere else. The word "coracle" appeared 24 times in chapter five, 1 time in chapter six, and nowhere else. These factors combined to give both of those words high TFIDF scores, too.

5 – Conclusion

For the book Treasure Island, we were able to take a .txt file and boil it down to a list of all the words in the book after cleaning up the extra junk that was present. We then found how often words occurred across the entire book, how many chapters they were present in, and eventually find each word's TFIDF score. Once we had this, we could create some visualizations and tables to help us explore the results. Those results showed us that there were two main ways for a word to hurt its TFIDF score. Those being: occurring in nearly every chapter of the book, or those having a super low term frequency. Once we calculated the actual TFIDF scores there was one word in particular that stuck out for occurring a lot within just a single chapter. Because of this analysis we were able to find the most unique words to their respective chapters and find a handful of those old words that aren't used very often anymore. Specifically, "coracle" which is a term referring to a small boat made of wickerwork propelled by a paddle. You always hear that language is an ever-evolving thing. Analyzing old texts like this can give us a quantifiable way to track this. Others should continue analyzing older texts and allowing us to see how language has developed over a range of topics and track trends over time.

6 – References

https://edstem.org/us/courses/42859/discussion/3489991
https://www.geeksforgeeks.org/matplotlib-pyplot-legend-in-python/#
https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.patches.Rectangle.html
https://stackoverflow.com/questions/18595686/how-do-operator-itemgetter-and-sort-work