Project 1 - TFIDF Report

Shamita Yedlapalli

October 2023

1 Summary

I sought to discover how indicative the words with the highest TFIDF values in these books are to the plot of the stories. Through this investigation, I discovered that whether a main plot point could be figured out through analyzing TFIDF values varies on a case-by-case scenario and is dependent on more than just the words used in the novel. I also wanted to compare the differences in the most common words used in a book and the most common words specific to that book. The words with the highest TFIDF values are much more indicative of the plot of the book than the most frequent words. However, they should not be the only resource used to learn about a book.

2 Introduction

Classic literature is a popular genre for students in high school. Of course, these students are often reluctant to take the time to read these books and opt for summary websites online. I wanted to explore whether the most common words used in a book or the words with the highest TFIDF values could provide insight into what a story might be about, similar to short summary websites. Going into this project, I figured the most common words (words with the highest term frequency) were not going to be the most insightful, but I wanted to use these terms as a control to compare to the highest TFIDF terms. Going into this project, I predicted that the terms with the highest TFIDF values would somewhat allude to the plot of the story, enough so that a student could understand the story or at least a section of the story.

3 Methodology

The very first thing I did was brainstorm books I could use. While I struggled to pick my selection, I decided to choose Classic Literature stories that I read in high school. I originally chose two books, *Frankenstein* by Mary Shelley, and *The Great Gatsby* by F. Scott Fitzgerald. I completed most of the whole process with these two books until I realized 2 books may be too small of a set to get major results. As a result, I added a third story, “The Yellow Wallpaper” by Charlotte Perkins Gilman. I created functions to make certain processes, especially those that may be repeated easier, but many of the later steps and analysis was done through the main function.

3.1 Parsing

In a function called get_words(), I read the text files and split them using the split() function.

This created a list of all the words in the selected book. Then I went through and split words that are separated by dashes. This was a necessary step that I did not consider originally. A common pattern in these books is a dash separating two unrelated words. For instance, in The Great Gatsby, a dash separated the words matter and tomorrow. Then as I created dataframes with words as columns, I had columns with two words joined together. For example, matter-tomorrow became mattertomorrow. In order to fix this, I went through the stored words and added them to a new list. Words joined with a dash was split by the dash and added to the new list separately. With this new updated list, I lowercased all of the words and removed all non-alphanumeric characters through a regex command. In this step, I could've removed the stop words, but I wanted to compare the words with the highest TFIDF with the words with the highest TF.

3.2 Vectorization

The next step was vectorization. Using the word list created in the previous step, I created dictionaries to keep track of the occurrences of each word in a story. I created a dictionary for each story. Then I merged the dictionaries together to create one big dictionary that keeps track of the union of the words in the separate dictionaries. This process does not duplicate words. Using this merged dictionary, I created a dataframe with the words as columns. The individual dictionaries were used to update the number of occurrences of each word in each of stories in the dataframe. The last column of this dataframe was used to store the number of words in each of the stories. This column was then used to calculate the Term Frequency (TF) values of each of the words by dividing the number of occurrences by the number of words. Next, I calculated the Inverse Document Frequency (IDF). I utilized the dictionaries I made above to check whether or not a particular word was used in each of the stories. I stored the IDF values for each word in a list. By the end of this step, I had a dataframe keeping track of the TF values and a list keeping track of the IDF values.

3.3 TF-IDF

Since I had a data frame with the TF values and a list with the IDF values, getting the TFIDF values was relatively easy. I simply created a new dataframe (called tfidf_data) that multiplied the old dataframe containing the TF values and the list containing the IDF values. This created a dataframe containing the TDIDF values for each word for each story. This is the end of calculating values. Finally, I attempted to extract some of the highest TFIDF values and their corresponding words. I did this by using the built-in max function, then used other built-in functions to get the name of the column that value was in. Then I replaced this value with a NaN value and repeated this process about 5 times to get the 5 highest TFIDF valued words in each book. I made sure to execute this process on a copy of the TFIDF dataframe to prevent losing any data. Since I sought to compare these words with the words with the highest TF values, I repeated this process on a copy of the dataframe containing TF values. Except this time, I collected the 10 highest TF valued words.

4 Results

The most common words in *Frankenstein* were elizabeth, feelings, clerval, misery, justine, and cottage. 3/6 of the most common words are names (Elizabeth, Clerval, Justine). However, the remaining three words indicate the main points of Frankenstein. Frankenstein's monster was created and released into the word and one of the main concerns throughout the book were its feelings and the lack of feelings it received throughout its life. The word misery being one of the most common words also alludes to similar themes of the book. Frankenstein's monster was constantly in misery as it could not fit into the world because it could not escape its appearance. The last most common word is cottage which refers to Frankenstein's monster's place of being in some of the novel and really captures that the monster was in solitude. Overall, these three words show the main points and moods of the novel.

	TFIDF VALUE	TF VALUES
ELIZABETH	0.0005001579296136748	0.001234
FEELINGS	0.0004495801614504941	0.001109
CLERVAL	0.0003315653690697394	0.000818
MISERY	0.0003090863609972147	0.000762
JUSTINE	0.0003090863609972147	0.000762
COTTAGE	0.00028660735292469	0.000707

Table 1: Highest TFIDF values in *Frankenstein* and corresponding terms and TF

The most common words in *The Great Gatsby* were gatsby, tom, daisy, car, and gatsbys. 4/5 of the most common words were character names. This book was told in 1st person and mainly revolved around these characters and their lives, specifically Gatsby's life. Gatsby was wealthy and the main subject of the book, so Gatsbys (and Gatsby) being among the words with the highest TFIDF values could show Gatsby's possessiveness and his overall importance. Since this book revolved around Gatsby and his main focus was on Daisy, it makes sense that Daisy was also on this list. Car being one of the most common words in The Great Gatsby is quite surprising considering it is an undescriptive/generic word. However, considering one of the most important scenes involved a car, it makes sense that the word car is high up on the list. It is also possible that the author could have been foreshadowing this scene by bringing up cars throughout the book.

	TFIDF VALUE	TF VALUES
GATSBY	0.0016368931243913537	0.004037
TOM	0.0014782147092717835	0.003646
DAISY	0.001252724329891342	0.00309
CAR	0.0006096591738804531	0.001504
GATSBYS	0.0005595502006847994	0.00138

Table 2: Highest TFIDF values in *The Great Gatsby* and corresponding terms and TF

The most common words in "The Yellow Wallpaper" were jennie, creeping, nursery, smell, arbors, color, daytime, queer, bedstead, shines, and fungus. Of the three stories, "The Yellow Wallpaper" had the most descriptive words that alluded to the plot of the story. The main premise of the story is a young lady in post-partum told to stay in a romm and her descent to

madness as a result. This summary can somewhat be seen through some of the most common unique words such as wallpaper, creeping, color, bedstead, smell. The main character was obsessed with the wallpaper in the room she was trapped in and talked a lot about it, including the specific color it was (because she disliked the color). The bedstead was nailed down and later there were gnaw marks found on it. Arbor was also one of the common words and this alludes to part of the scene that the main character sees outside the window in her room. The main character sees women creeping in the wallpaper and she really dislikes seeing these women in the wallpaper. However, by the end of the story, she feels herself become one of the women in the wallpaper and creeps around. Throughout the story, she complains about the smell of the room and the wallpaper. The beginning of the story was mostly the main character disliking the room (which is seen through words like smell, color, and wallpaper). The story transitions to be the main character's crazy actions (which is seen through the word creeping).

	TFIDF VALUES	TF VALUES
WALLPAPER	0.0007925690336044914	0.001955
JENNIE	0.0007925690336044914	0.001955
CREEPING	0.0005944267752033685	0.001466
NURSERY	0.0003302370973352047	0.000814
SMELL	0.0003302370973352047	0.000814
ARBORS	0.0002641896778681638	0.000652
COLOR	0.0002641896778681638	0.000652
DAYTIME	0.0002641896778681638	0.000652
QUEER	0.00019814225840112285	0.000489
BEDSTEAD	0.00019814225840112285	0.000489
SHINES	0.00019814225840112285	0.000489
FUNGUS	0.00019814225840112285	0.000489

Table 3: Highest TFIDF values in "The Yellow Wallpaper" and corresponding terms and TF

The words with the highest TFIDF values are a very different set of words from the words with the highest TF values. The words with the largest TF values (and the order they were in) varied across the three books, but the sets were very similar. Some of the words that appeared in all three books were the, and, I, a, of, that, in (commonly known as stop words).

	TF	TFIDF
THE	0.0552044352044352	-0.015881
AND	0.0399168399168399	-0.011483
I	0.0386001386001386	-0.01110
A	0.017948717948717947	-0.005164
OF	0.03426195426195426	-0.009857
THAT	0.014192654192654192	-0.004083
IN	0.0146500346500346	-0.004215

Table 4: Highest TF values in *Frankenstein* and corresponding terms and TFIDF values

	TF	TFIDF
THE	0.04935118434603501	-0.014197
AND	0.03227600411946447	-0.009285
I	0.02455200823892894	-0.007063

A	0.02898043254376931	-0.008337
OF	0.022966014418125645	-0.006607
THAT	0.01223480947476828	-0.003520
IN	0.016601441812564368	-0.004776

Table 5: Highest TF values in *The Great Gatsby* and corresponding terms and TFIDF values

	TF	**TFIDF**
THE	0.03909431503502199	-0.011247
AND	0.04691317804202639	-0.013496
I	0.049845251669653035	-0.014340
A	0.023293696041700604	-0.006701
OF	0.017429548786447303	-0.005014
THAT	0.017918227724385078	-0.005155
IN	0.015149047076071022	-0.004358

Table 6: Highest TF values in "The Yellow Wallpaper" and corresponding terms and TFIDF values

While each of these words has a high TF value in each book, the TFIDF values are actually negative, conveying that while it is a prevalent word, it is not an important word to the book (which is the effect of the IDF value).

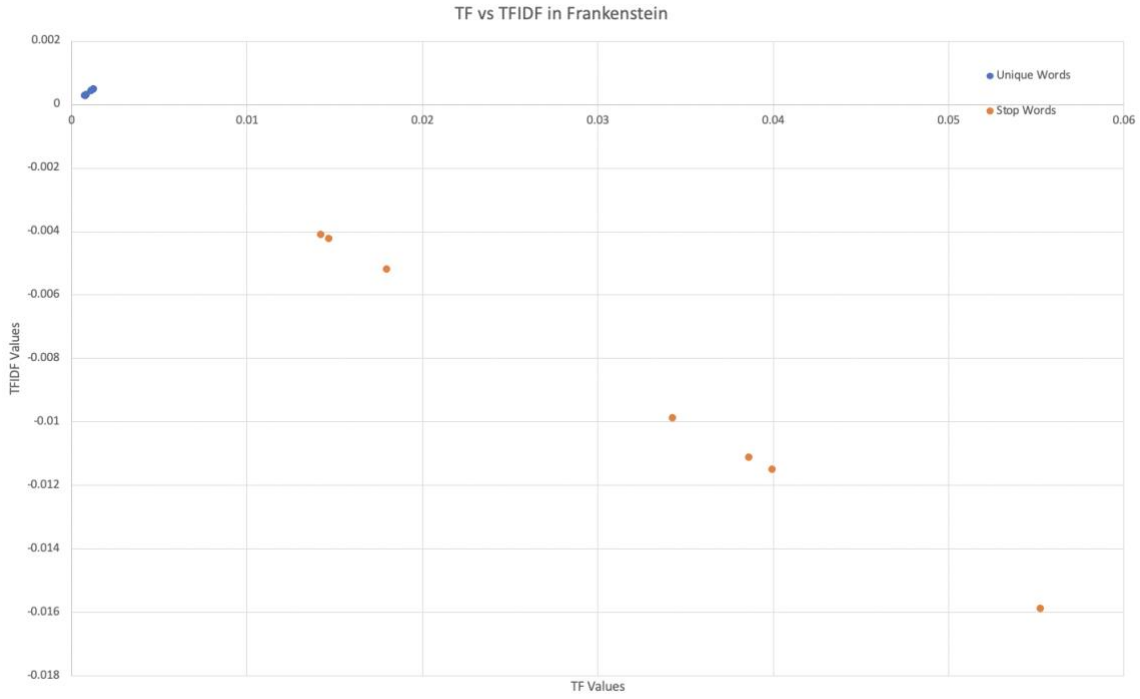

Figure 1: TF values vs TFIDF values of Stop Words and Unique Words in *Frankenstein*

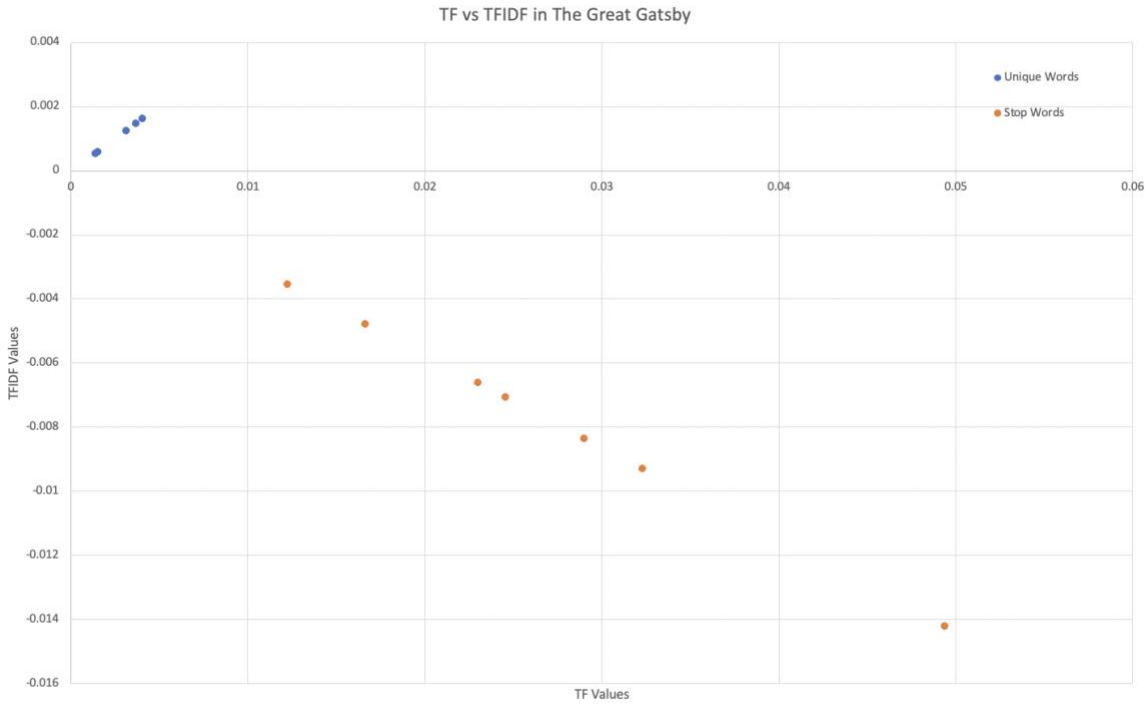

Figure 2: TF values vs TFIDF values of Stop Words and Unique Words in *The Great Gatsby*

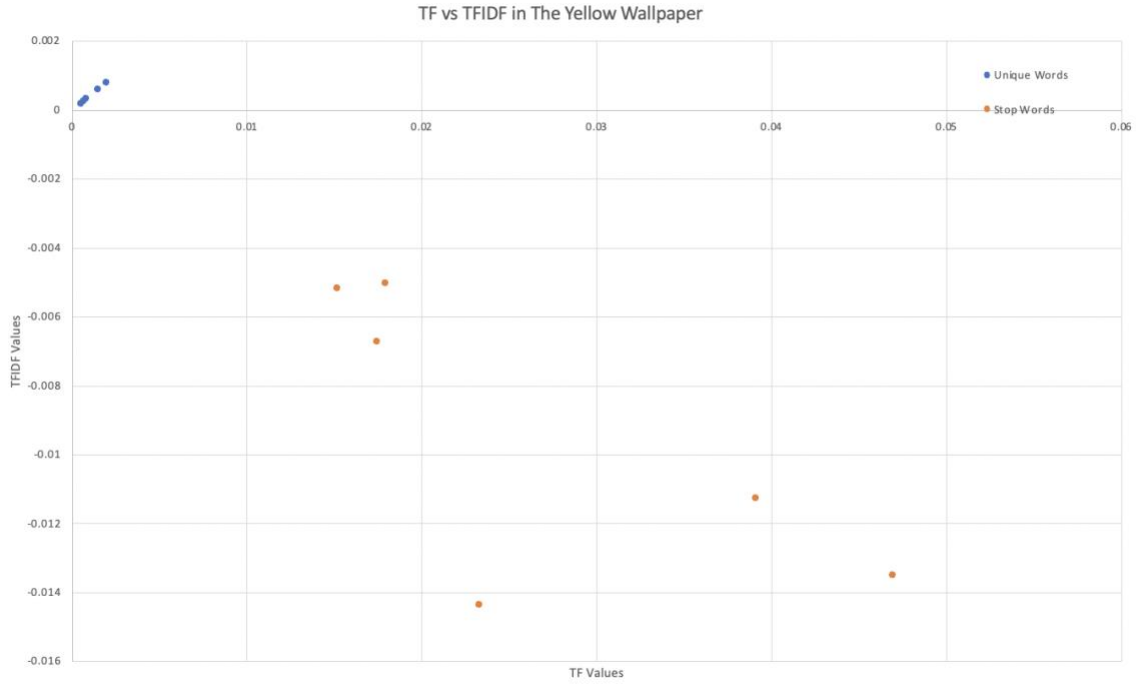

Figure 3: TF values vs TFIDF values of Stop Words and Unique Words in "The Yellow Wallpaper"

This data shows that the most common words across literature will probably have a relatively low TFIDF value while words more unique to a piece of literature will have higher TFIDF values. Overall, words with higher TFIDF values are somewhat indicative of most used unique words in a story. These words will be more accurate about a story's plot than words with the most occurrences.

5 Conclusion

The main topic this project sought to investigate is whether the top words based on TFIDF could be indicative of the plot of a story. It seems to be more likely with shorter books, like "The Yellow Wallpaper," probably because there are fewer words to go through and each word has a greater importance to the story. Shorter stories are more likely to only have a main plot while longer books may include several sub plots as well. This could make words more specific to a certain plot less important to the overall story. For such reasons, TFIDF might not be the best way to get the main points of a story. As seen with The Great Gatsby, most of the most common unique words were character names. The TFIDF words do not really indicate any main thing that happens in the story. All this data shows is that those characters exist and are most likely the main characters. While TFIDF is not indicative of the plot of a story, it shows the focus of these classic literature stories. In Frankenstein, the main focus is on the monster's solitude and misery. The main focus in The Great Gatsby are the main characters. In The Yellow Wallpaper, the main focus is the wallpaper and the narrator's dislike for this wallpaper. Overall, the TFIDF is better for getting a refresher on a story than it is for getting a summary of a story.

It would be interesting to see whether calculating TFIDF could work when the set only has 2 samples or books or if there is a better method. I originally was only calculating the TFIDF values with two books (*Frankenstein* and *The Great Gatsby*). I noticed that the TFIDF values were either 0 or negative. Mathematically, this makes sense since the only possible IDF values were 0 ($\ln(2/(1+1)) = 0$) and a negative number ($\ln(2/3) = -0.405465108$). However, I wonder if there is a way to get around this issue or if there is an entirely different method to follow.

6 References

https://stackoverflow.com/questions/14734695/get-column-name-where-value-is-something-in-pandas-dataframe

- Post by Nic Scozzaro

https://www.geeksforgeeks.org/how-to-move-a-column-to-first-position-in-pandas-dataframe/

https://stackoverflow.com/questions/17839973/constructing-pandas-dataframe-from-values-in-variables-gives-valueerror-if-usi

- Post by DSM