# CS 24200 Project Report

Matthew Sindac

September 29, 2023

# 1 Summary

This technical report presents a Term Frequency-Inverse Document Frequency (TF-IDF) analysis that was used on two documents by Mark Twain which were "The Adventures of Huckleberry Finn" and "The Adventures of Tom Sawyer," and two documents by Charles Dickens which were "Great Expectations" and "Oliver Twist."

The goal of this analysis was to find the significance of terms within each document relative to the entire list of terms which would allow us to identify key terms and patterns that differentiate the writing styles of these two authors. In the writings of Charles Dickens, the TF-IDF analysis of his novels revealed the prevalence of Victorian-era societal themes along with intricate character development. For Mark Twain, the TF-IDF analysis of his novels revealed terms related to American humor and colloquial language.

Based on these analyses, readers interested in American humor and satire should prioritize Mark Twain's works while those interested in exploring themes of social class and character depth should look into Charles Dickens' literary works. TF-IDF is a very powerful text analysis technique to gather insights and differentiate between authors. Readers and researchers can use these findings to find works that align more closely with their own personal preference.

# 2 Introduction

TF-IDF, a text analysis technique, is used to determine the importance of words within a collection of documents. By considering the amount of times a term appears in a document and how unique a term is across all documents, it is able to identify significant vocabulary and recurring motifs that characterize each text.

Within this analysis, we focus on the literary works of Mark Twain and Charles Dickens. It seeks to attribute authorship and uncover unique writing styles, focusing on patterns and vocabulary usage, including slang. By comparing their TF-IDF values, we want to gain insight into their distinctive writing styles.

## 3 Documents

These are the following documents we are using to analyze the writing styles of Charles Dickens and Mark Twain:

- Mark Twain's: "The Adventures of Huckleberry Finn"
- Mark Twain's "The Adventures of Tom Sawyer"
- Charles Dickens' "Great Expectations"
- Charles Dickens' "Oliver Twist"

# 4 Methodology

## 4.1 TF-IDF Calculation

For each document, there were a couple of steps that needed to be done in order to calculate the TF-IDF which were the following:

### 4.1.1 Tokenization

Each document was tokenized into words, only gathering relevant text and removing punctuation along with converting all text to lowercase.

- In this process, I created code that utilized pandas and regular expression, sorting and organizing all the relevant terms in each individual document. I removed things such as table of contents, bibliographic information, licenses and other things in order to gather the raw text data. Since we are analyzing the raw text itself and not the irrelevant things listed, tokenization is needed in order to sort the relevant terms from the irrelevant terms.

- The formatting of the books varied, which made it challenging to consistently remove unwanted sections. Through various custom regex patterns, I was able to separate all the unwanted information from the text that we wanted to analyze. The raw text data was then put into a pandas dataframe where it could be further analyzed.

#### 4.1.2 Term Frequency (TF)

The term frequency of each term in a document which is calculated by getting the amount a term appeared within the document and dividing it by the total amount of terms in the document.

- In order to calculate the TF-IDF, we first need to find the term frequency (TF) of every single term in each document. This is done in order to find document specific insights where we are able to distinguish the importance of terms within individual documents. For example, within biology documents, terms such as "life" and "organisms" may have high TF values while in a document about food, words like "savory" and "sweet" may have high TF values.
- Example of TF Values in the document "Great Expectations"

| Words | TF Values |
|---|---|
| 'click' | 1.627445236467793e-05 |
| 'myself' | 0.0012639824669899858 |
| 'the' | 0.044054942551183154 |
| 'paused' | 2.1699269819570573e-05 |

- From this table, the higher the value, the more it appeared throughout the novel. In this case, 'the' had the highest term frequency within this document which could possibly be a key term. However, we cannot assume since we don't know how unique it is across all the documents and how special it is to the novel "Great Expectations".
  - High TF values seem to indicate the importance of a term within a specific document.

#### 4.1.3 Inverse Document Frequency (IDF)

The inverse document frequency of each term is calculated by the natural log of the total number of documents divided by the occurrence of each term across all documents.

- IDF essentially helps identify terms that are unique across multiple documents and assigns a value based on that uniqueness. This is to rank the importance of terms within documents by increasing the score of terms that are both frequent within an individual document and rare across the collection of documents.
- Example of IDF Values within the four document corpus:

| Words | IDF Values |
|---|---|
| 'click' | 1.3862943611198906 |
| 'irresistibly' | 0.6931471805599453 |
| 'paused' | 0.28768207245178085 |
| 'the' | 0.0 |

- From the table, the higher the value, the more unique it is across documents. In this case, 'click' is the highest term out of the four documents selected, with only the document "Great Expectations" having the word click.
  - Low IDF values indicate that a term is common across many documents, reducing its importance.

#### 4.1.4 TF-IDF Score

The TF-IDF score for each term in each document was calculated by multiplying its TF and IDF values.

- This is calculated for each document and for each term of that document. It calculates the importance of that term for that document over several different documents and returns a value that measures its importance.
- Example of TF-IDF scores for "Great Expectations":

| Terms | TF-IDF Values |
|---|---|
| 'pip' | 0.0023839648497597093 |
| 'herbert' | 0.0021733938220206816 |
| 'wemmick' | 0.0019101800373468964 |
| 'havisham' | 0.0018048945234773826 |

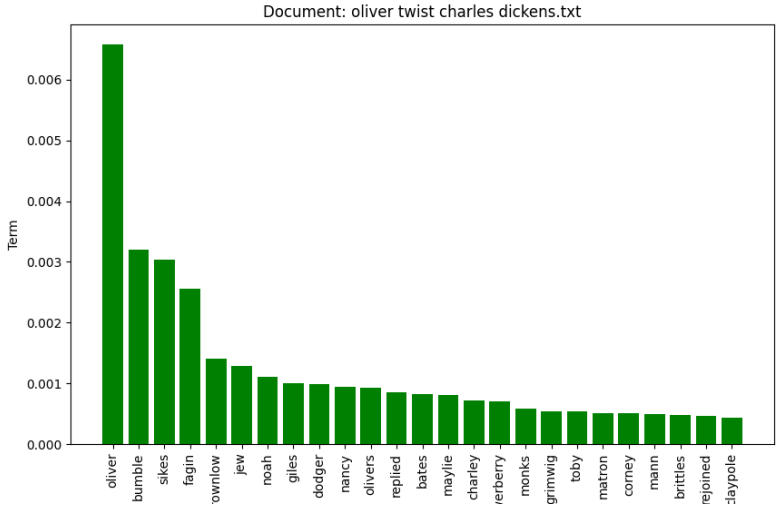
- From this table, we can determine how important these terms are to the document in question. In this case, since all these terms are relatively high, they must play a significant role in conveying the document. This is how the importance of each term is assigned a value within these individual documents.
  - The product of TF and IDF values, TF-IDF, highlighted terms that are both frequent in a document and rare across the collection of documents.

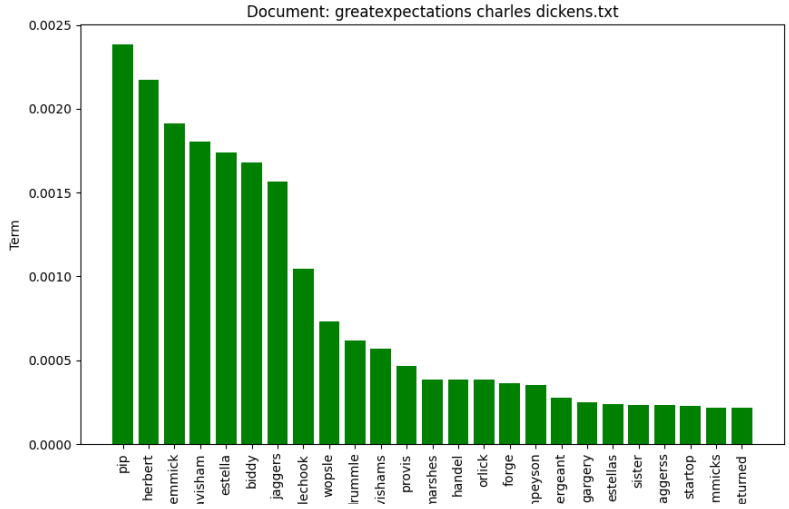
# 5 Results

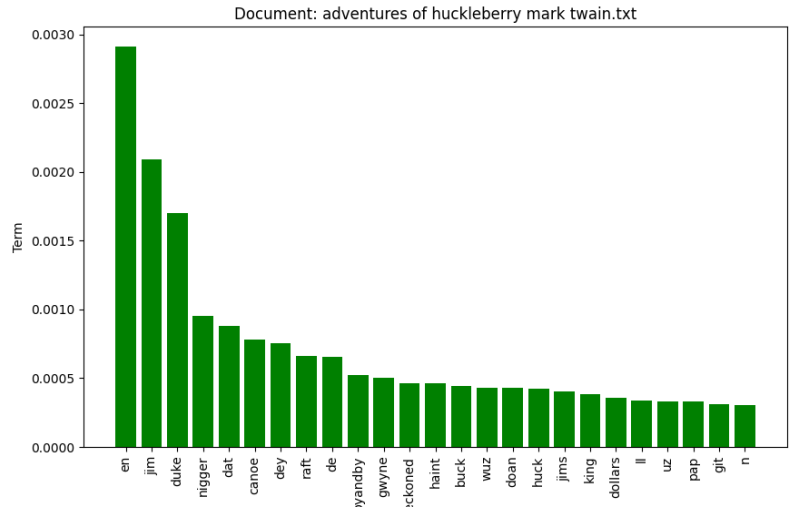
The TF-IDF values were calculated for each term and document pair using the formula tfidf(t, d, D) = tf(t, d) * idf(t, D). This analysis revealed the following key findings amongst the ones with the highest TF-IDF values:

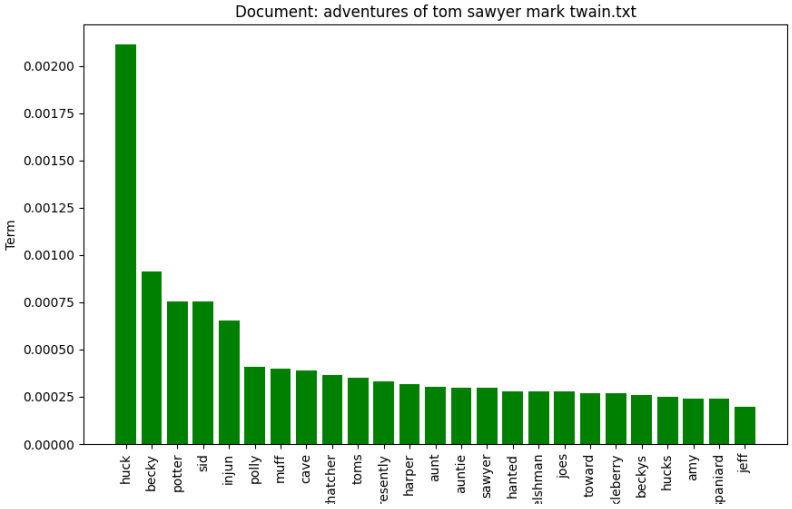
- Charles Dickens' Texts:
  - In "Oliver Twist," some notable terms with high TF-IDF values included "rejoined", "replied", "yer", "workhouse", "companion", "porochial," along with many character names such as "oliver", "bumble", "noah".
  - In "Great Expectations," significant terms with high TF-IDF values included "marshes", "forge", "returned", "blacksmith", "guardian", "convicts," in addition to various character names such as "pip", "herbert", "havisham".
- Mark Twain's Texts:
  - In "The Adventures of Huckleberry Finn," prominent terms with high TF-IDF values included "en", "dat", "dey", "canoe", "raft", "gwyne", "reckoned", "haint", "wuz", "doan", "dollars", "git," along with some character names such as "jim", "jane", "silas".
  - In "The Adventures of Tom Sawyer," some key terms with high TF-IDF values included "cave", "presently", "hanted", "toward", "cardiff", "pirates", "lookyhere," with many character names included such as "huck", "becky", "sid".

## 5.1 Visualizations






These terms selected from each document may seem random, however, they are essential to the document's message. Each of these terms were calculated using TF-IDF and can also be seen within the graphs above. From these scores, we can evaluate the following:

- Charles Dickens' Texts ("Oliver Twist" and "Great Expectations")
  - In both "Oliver Twist" and "Great Expectations", character names such as "oliver" or "pip" have high TF-IDF values. This suggests

that the characters play significant roles in the narratives and have an important purpose.

  - In "Oliver Twist", words with high TF-IDF values such as "workhouse", "companion" and "porochial" suggest the setting that "Oliver Twist" takes place in with "yer" being seen as a distinct language feature in the document as well.
  - In "Great Expectations", terms with high TF-IDF values such as "marshes", "forge", "blacksmith", "convicts", "guardian", "returned" seem to emphasize locations like the marshes and the presence of convicts along with some character specific elements.

- Mark Twain's Texts ("The Adventures of Huckleberry Finn", "The Adventures of Tom Sawyer")
  - In "The Adventures of Huckleberry Finn", words with high TF-IDF values such as "en", "dat", "dey", "canoe", "raft", "gwyne", "reckoned", "haint", "wuz", "doan", "dollars", "git", all have quite a significance to the document. These terms reflect a distinct language used within the narrative along with references to key characters such as "jim", "jane", "silas" and activities such as canoeing and rafting.
  - In "The Adventures of Tom Sawyer", words with high TF-IDF values such as "cave", "presently", "hanted", "toward", "cardiff", "pirates", "lookyhere," suggests an exploration of adventures, caves, and interactions among the key characters such as "huck", "becky", "sid".

Through these results, we are able to identify characteristics and thematic elements specific to each document and highlight the authors writing styles. High TF-IDF values indicate terms that are both frequent within a document and distinctive to that document, scoring its importance. All these terms are likely to be significant keywords or topics for the individual document. The TF-IDF values could be used to identify key themes in each document and the different elements as well.

# 6 Exploration

TF-IDF is a very useful text analysis tool that can be used in a variety of ways. Some of the ways that TF-IDF can be used are the following:

- Authorship attribution
  - For example, I analyzed two documents from Mark Twain and two documents from Charles Dickens. If there was an unknown document that was possibly one of the two authors, through the use of TF-IDF, we can determine a high possibility of who wrote it and make an educated guess on who the author is.
- Similar Writings
  - Since we know the text writing style of the two documents from Mark Twain and the two documents from Charles Dickens, if we wanted a story similar to one of them, we could use TF-IDF values and compare them to other documents with similar TF-IDF values in order to determine a relevant document.
- Similar Topics
  - By focusing on terms with high TF-IDF values within a document, we can discover texts that discuss similar subject matter, even if they are written by different authors or have distinct writing styles.

There are other ways to build upon TF-IDF and create an even better concise analysis and gather more information such as:

- Translation
  - If we are attempting to compare documents that are in different languages, translating the terms into a common language using machine learning could let us analyze even more texts from around the world
- Machine Learning
  - We could possibly create a machine learning model that could classify the documents from TF-IDF values into their own categories, where they be further used for specific

TF-IDF can be used to compare how similar documents are and allows us to place documents into different groups based on the various values.
Other methods of TF-IDF include the following:

- Bigram TF-IDF
  - This captures words not as individual terms but instead pairs consecutive words as terms. This could capture words that are more identifiable together rather than alone, making it a better way to analyze a text where word combinations matter.
- Trigram TF-IDF
  - This is an extension of bigram but instead of pairs, it takes triplets of consecutive words. This can be more useful for capturing important phrases and other key information rather than individual terms.

Overall, each n-gram TF-IDF method has its own use. Utilizing each method on a singular document can reveal information on many different levels. TF-IDF is a very powerful tool that is useful to various degrees.

# 7 Conclusions:

In this TF-IDF analysis, we delved into the works of two renowned authors, Mark Twain and Charles Dickens, to unveil distinctive writing styles and recurring motifs. The primary objective was to attribute authorship and identify unique vocabulary usage, including slang, by comparing TF-IDF values across their respective texts.

- Charles Dickens' Texts:
  - The TF-IDF analysis of Dickens' works, "Oliver Twist" and "Great Expectations," revealed the significance of terms related to specific settings, character names, and interactions. This suggests that Dickens' unique style is characterized by detailed descriptions of locations and relationships within his narratives.
- Mark Twain's Texts:
  - Analyzing Twain's "The Adventures of Huckleberry Finn" and "The Adventures of Tom Sawyer" highlighted a distinct vernacular and colloquial language, with an emphasis on character names and adventurous activities. Twain's style is marked by the use of unconventional language and a focus on character-driven adventures.

Our TF-IDF analysis successfully demonstrated the power of this text analysis technique in uncovering unique writing styles and recurring motifs within literary works. The results not only attributed authorship but also provided insights into the thematic elements that define each author's narratives.

Our results are grounded in quantitative analysis, which objectively evaluates the importance of terms within a text corpus. TF-IDF, as a widely accepted technique, allowed us to identify patterns and distinctive features in the authors' writing styles, such as Dickens' detailed descriptions and Twain's colloquial language.

In summary, our TF-IDF analysis offers a compelling glimpse into the distinctive writing styles of Charles Dickens and Mark Twain. Further research and projects in this domain have the potential to deepen our understanding of literature and authorship attribution techniques.

# 8 References

Dickens, C. (1996, November 1). *Oliver Twist.* Project Gutenberg.
https://www.gutenberg.org/cache/epub/730/pg730.txt

Dickens, C. (1998, July). *Great Expectations* Project Gutenberg.
https://www.gutenberg.org/files/1400/1400-0.txt

Twain, M. (2004, June 29). *Adventures of Huckleberry Finn.* Project Gutenberg.
https://www.gutenberg.org/cache/epub/76/pg76.txt

Twain, M. (2004, July 1). *The Adventures of Tom Sawyer, Complete.* Project Gutenberg.
https://www.gutenberg.org/cache/epub/74/pg74.txt