CIS 6397 - TEXT MINING - MINI PROJECT 1: Analyzing Word Distributions in Two Corpora

Venkata Sai Bharath Thoranala: 2245697; Pradeep Reddy Mallepally: 2200020; Vamsi Krishna Reddy Batch: 221393

Abstract

This article analyzes the word distributions in great detail, to acquire a deeper understanding of the characteristics and content of the two separate corpora. The research focuses on evaluating the importance of the top 30 terms found in each corpus, creating word distribution profiles, and making deductions about the underlying topics or descriptors of each corpus through the analysis of word frequencies. It also investigates variations in the inclusion of frequent terms and dive into the examination of bi-gram distributions. It applies a stop word list in the corpora to assess its effect on the comprehension of the corpus descriptions in order to further refine the analysis. This publication explains the results of the trials, and provides a detailed overview of the research methodologies and the conclusions from the analytical work.

1 Background

A thorough investigation of word distributions within corpora is essential to gaining profound insights into the content, topics, and properties of textual data. This aspect, which is essential to the fields of text mining and natural language processing, is key to understanding the subtleties of a particular corpus.

In fact, the careful examination of the most common phrases and n-grams is the core of this study. Such an effort has the potential to provide light on the fundamental concepts and categories that form the basis of the corpus's essence, making it a worthwhile project.

This research attempts to establish a strong basis for interpreting the complex tapestry of information contained within textual data by diving into the subtleties of word distribution patterns.As a result, this article aims to be a significant resource for academics and industry professionals by providing an in-depth examination of the methods and learnings from this analytical journey.

2 Introduction

This paper focuses on the dynamic field of textual data analysis, especially on two different corpora, namely Corpus1 and Corpus2. Untangling the complex web of word distributions within these corpora is our main goal in this project; doing so is crucial to the field of text analysis.

In each of the aforementioned corpora, the 30 phrases that occur most frequently will be methodically identified and examined in this inquiry. By doing this, we hope to set out on a journey of semantic investigation, examining these terms' deep-seated meanings and connotations. Furthermore, we aim to identify the overarching themes, concepts, and descriptors that characterize the particular essence of each corpus as part of our investigation, which goes beyond merely examining word frequencies.

We enter the world of experimentation in the quest for a complete understanding. We attempt to reveal hidden patterns and nuances hidden within the corpora's textual fabric by adjusting factors like the amount of frequent terms and investigating the world of bi-grams.

We introduce the stop word list component to sharpen our analytical focus, and we'll examine its effects to see how they affect our ability to understand the descriptions that are embedded in the corpora.

This report establishes the groundwork for a scholarly investigation of textual data analysis, providing a thorough explanation of our methodology, showcasing the findings of our exhaustive research, and ultimately paving the way for wise conclusions to be drawn from this complex voyage of discovery.

3 Methodology

This section gives a thorough explanation of the technique used to conduct a thorough study of Corpus1 and Corpus2. The methodology consists of several crucial steps, each of which is carefully planned to shed light on the corpus's textual content, word distributions, and semantic nuances.

3.1 Data Collection:

Corpus1 and Corpus2 are two different sources of textual information that were used to start the investigation. They were kept in separate directories on the system.

3.2 Data Pre-processing:

The preparation of the texts found in both Corpus1 and Corpus2 is the first step in our investigation. This crucial stage entails a number of transformative procedures intended to make the textual material accessible for further study. Tokenization, punctuation mark removal, managing special headers or tags, and taking care of any text-specific nuances or extraneous characters that may be present are some of these activities. The main goal of this preparation stage is to make sure that our data are accurate and clean, enabling precise and insightful analysis.

3.3 Word Frequency Analysis:

The full analysis of word frequencies within each corpus is the next phase of our investigation. We identify and highlight the 30 most common terms in Corpus1 and Corpus2 by computing word distributions. By highlighting the most frequently used terms, this analytical endeavor enables us to get insightful knowledge into the text's content.

3.4 Topic or Descriptor Inference:

We set out on a mission to identify the overarching subjects or descriptors that define each corpus using the word frequencies as our guide. This inference is based on an analysis of the words that appear most frequently in the corpora, which gives us a better grasp of the main ideas and overarching themes present in the textual data.

3.5 Parameter Validation:

We test the validity of our analysis by applying several values for the "k" most frequent terms to our script. With the help of this exploratory experiment, we can evaluate the sensitivity of our analysis and see how changes in the number of top words affect how we interpret the corpus descriptors.

3.6 Bi-gram Analysis:

We explore bi-gram analysis (n-grams with n=2) in addition to single-word frequencies (n-grams with n=2). This thorough investigation takes into account both word connections and specific word frequencies across the corpora. We gain important understandings into the complex network of word associations and their importance inside the textual data by analyzing these bi-grams

3.7 Stopword Removal:

The adoption of a stop word list is the final aspect of our investigation. We create or include such a list, and then systematically eliminate each term from the corpora. To determine how much the elimination of stop words affects how we perceive the corpus descriptors, we conduct the word distribution and bi-gram analysis.

4 Github Repository used for the Project:

The following is the link for the github repository used for this research. It contains the source code of the analysis done on the two corpora. GithubRepository

5 Experimental Results

Our research methodology yielded the following key findings:

Word Frequency Analysis before removing the stopwords:

- Initial analysis revealed a noticeable presence of common stopwords in the top ten lists of the most commonly occurring words in both Corpus1 and Corpus2. The linguistic building blocks "the," "and," "in," and "of" are examples of stopwords. They are crucial for proper sentence construction and grammatical coherence. However, their dominance among the top word frequencies shows that they might not have a significant impact on the text's central theme. They act more as grammatical scaffolding and connecting words.

- Both the top 30 and top 50 most often occurring words in each corpus were examined in the research. A more thorough analysis of the text's content was possible thanks to the

inclusion of the top 50 words. Despite this growth, the findings continually demonstrated stopwords' supremacy in the ranks. Although helpful for context and readability, these common terms frequently represent general language constructions and do not necessarily capture the main issues or themes of the text.

Bigram analysis before removing the stopwords:

- Both corpora's most prevalent bigrams are primarily made up of common word pairs like "of the," "in the," "to the," and "and the." These phrases frequently appear together and act as linguistic building blocks for the English language. Despite the absence of detail in these bigrams, they are extremely important for sentence structure and reading.

- Bigrams like "it was," "he had," and "he was" in the text offer narrative hints. These pairings frequently denote previous deeds, personality traits, or descriptions. For instance, the words "it was" and "he had" imply the beginning of a narrative or a descriptive statement, respectively, but "he was" and "he was" provide information on activities or character traits.

- We notice that several bigrams are used consistently in both corpora, pointing to ingrained linguistic patterns and conventions. These bigrams may not explicitly state the topic matter, but they provide the text cohesion and stability.

Word Frequency analysis after stopwords removal:

Corpus1 analysis: After removing the stopwords and analysing the top common words in both the corpus, the results were improved and more insightful on the corpus context. From the word frequency analysis of the final filtered corpus 1, the following are the most contextual words which are frequently repeated in the texts:

- ***money***
- ***bank***
- ***value***
- ***gold***
- ***stock***
- ***business***
- ***credit***
- ***market***
- ***per cent***
- ***years***

These terms frequently refer to subjects like banking, finance, and economic analysis and are strongly suggestive of a financial and economic environment. The Bigram analysis of filtered Corpus 1 revealed the following word combinations:

- ***"stock exchange"***
- ***"per cent"***
- ***"clearing house"***
- ***"electronic works"***
- ***"quantity theory"***
- ***"federal reserve"***
- ***"money market"***
- ***"paper money"***
- ***"national bank"***

The financial and economic topics included in the texts are further supported by these bigrams. The terms "stock exchange" and "federal reserve" are directly associated with financial organizations and ideas. Together, the word frequency and bigram results show that the analyzed texts mainly deal with banking, economics, and finance. Terms like "New York," "United States," and "Project Gutenberg" are present, implying a connection to financial activity in the US and perhaps engagement in digital or electronic publication. Corpus2 analysis: In normal speech, words like "said," "would," "could," "like," "know," "say," "much," and "think" imply a narrative or conversational style of writing. Words like "man," "old," "good," and "thought" are generalizations that do not specify a particular situation. Words like "mr," "us," and "shall" suggest that character names or formal language, possibly from literary works, may be present. The terms "electronic works" and "Project Gutenberg" could imply that the writings are available in digital or electronic version. There are digital or electronically accessible texts since words like "Project Gutenberg" and

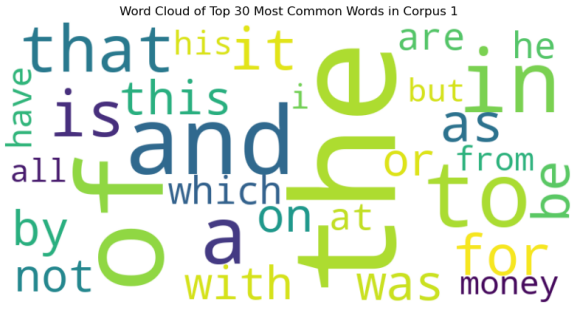

Figure 1: fig:Top 30 Common Words in Corpus 1 before removing Stopwords

Figure 2: fig:Top 50 Common Words in Corpus 1 before removing Stopwords

”electronic works” are used. Character names like ”Alexey Alexandrovitch,” ”Stepan Arkadyevitch,” ”Mr. Casaubon,” ”Captain Nemo,” ”Van Helsing,” ”Mr. Darcy,” ”Mr. Brooke,” ”Nastasia Philipovna,” ”Sir James,” and ”Darya Alexandrovna” suggest that the texts may be fictional works, possibly novels, with a varied cast of characters. A narrative or conversational setting is also suggested by words like ”let us” and ”said Mr.” It is noteworthy how words like ”united states” are used in this context because they may allude to locations or references in the texts.

The word frequency and bigram statistics alone cannot reveal the precise themes and narratives of

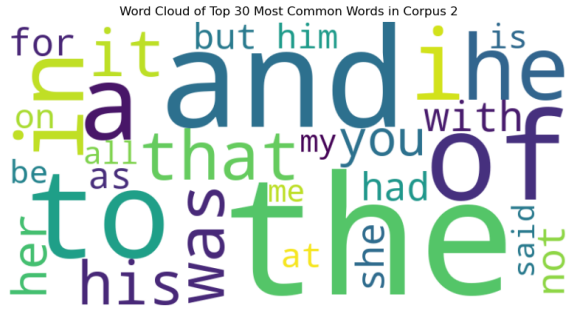

Figure 3: fig:Top 30 Common Words in Corpus 2 before removing Stopwords

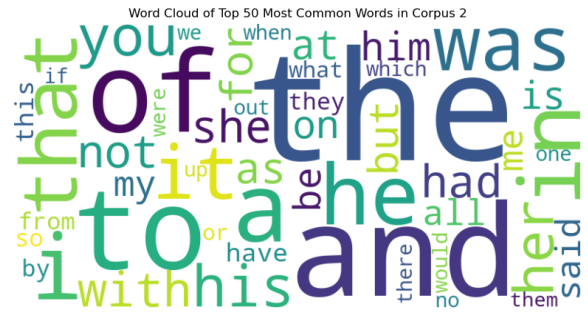

Figure 4: fig:Top 50 Common Words in Corpus 2 before removing Stopwords

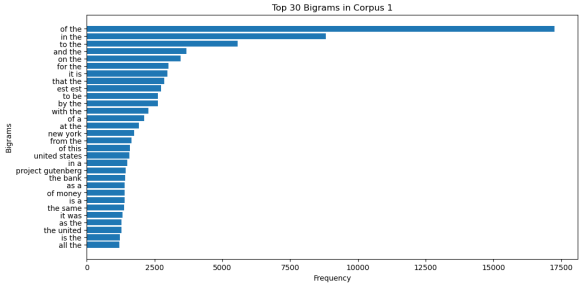

Figure 5: fig:Top 30 Bigrams in Corpus 1 before removing Stopwords

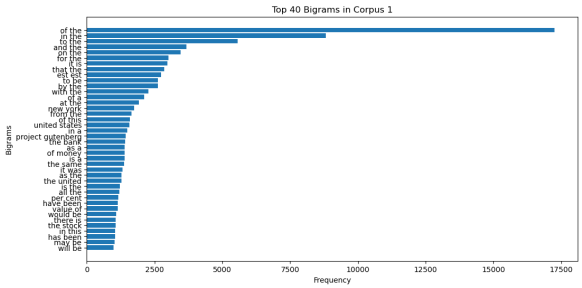

Figure 6: fig:Top 40 Bigrams in Corpus 1 before removing Stopwords

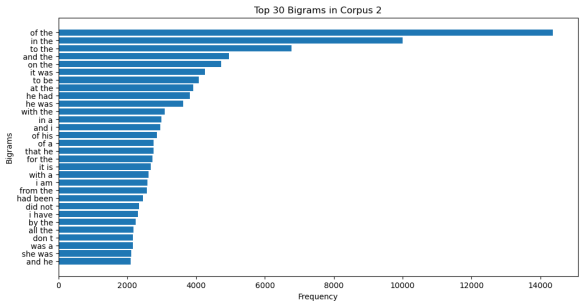

Figure 7: fig:Top 30 Bigrams in Corpus 2 before removing Stopwords

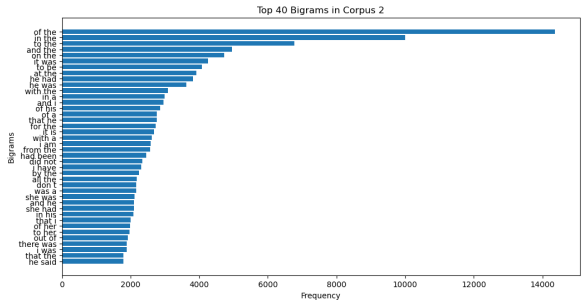

Figure 8: fig:Top 40 Bigrams in Corpus 2 before removing Stopwords

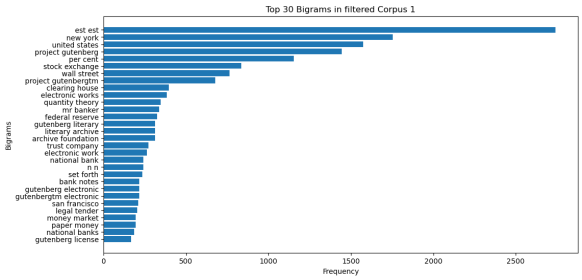

Figure 12: fig:Top 30 Bigrams in Corpus 1 after removing Stopwords

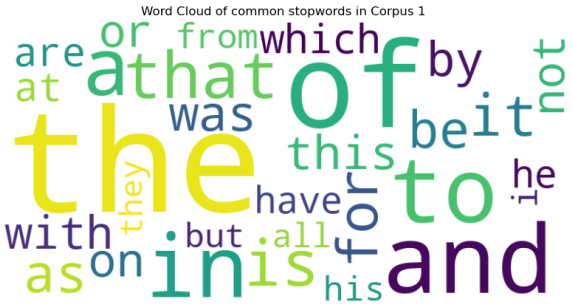

Figure 9: fig:Top Stopwords in Corpus 1

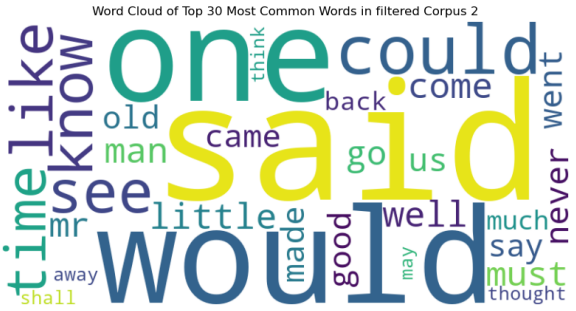

Figure 13: fig:Top 30 Common words in Corpus 2 after removing Stopwords

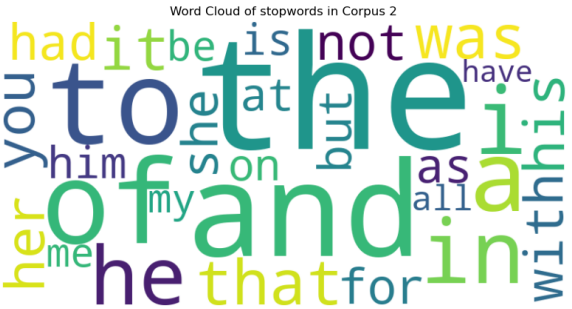

Figure 10: fig:Top Stopwords in Corpus 2

Figure 11: fig:Top 30 Common words in Corpus 1 after removing Stopwords

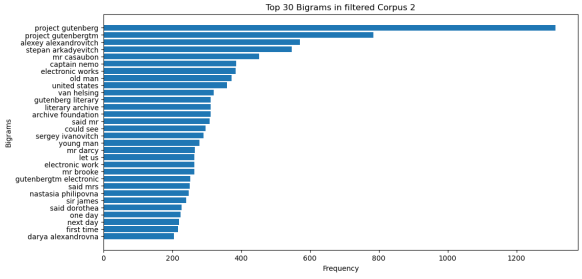

Figure 14: fig:Top 30 Bigrams in Corpus 2 after removing Stopwords

these texts, but it is obvious that they fall under the category of fiction or storytelling. It would be necessary to conduct additional research, such as text mining and content analysis, to determine the precise contexts and subjects of these literary works.

6 Conclusion

This study's thorough examination of two different text corpora exemplifies the power of text mining techniques in revealing information hidden in textual data. Following is a summary of the study's main conclusions and implications:

6.1 Key Observations and Findings:

The study of frequently occurring words and bigrams in Corpus 1 and Corpus 2 has shed light on the recurrent themes and subjects in these databases. Understanding the main ideas and objectives of each corpus can be aided by this knowledge. Stopwords' impact on textual analysis has been clarified through the examination of stopwords in corpora. Stopwords were eliminated, revealing several sets of frequent words and bigrams, highlighting the significance of careful preprocessing in text mining jobs. We better understood the significant material in the corpora by running filtered analyses that don't include stopwords and single-letter words. This method may reveal more subtle insights, particularly when stopwords predominate in the text.

6.2 Implications:

The importance of text preparation in text mining and analysis is shown by this study. For the purpose of getting relevant results and lowering noise in textual data, appropriate preprocessing methods, such as lowercasing, tokenization, and stopword removal, are required. A glimpse into the subject landscape of each corpus is provided by the study of common words and bigrams. Researchers, content producers, and data analysts that work with huge textual datasets might benefit greatly from this expertise.

6.3 Real World Applications:

By identifying common terms and bigrams, information retrieval algorithms can be improved, making it simpler to locate pertinent content in huge corpora. Personalized content recommendation systems can be created by examining user-generated material and using text mining to identify their preferences. These systems can make recommendations for content, goods, or services based on the preferences of specific users, increasing user happiness and engagement. Medical literature and electronic health records (EHRs) can both benefit from text mining approaches. Predicting disease outbreaks, patient diagnoses, and treatment outcomes can be the subject of future research. This can help medical practitioners make better judgments and provide better patient care. The use of text mining can be used to generate improved chatbots. Chatbots can offer more precise and contextually aware assistance by examining consumer questions and responses. This increases the effectiveness and satisfaction of client support. Text mining can be utilized in the realm of education to evaluate student performance and learning preferences. This data can help create individualized instructional materials and learning paths to enhance the learning process.

7 Future Scope

Despite the fact that this research has shed light on the investigated corpora, there is still need for more research and improvement: More sophisticated text mining methods, like sentiment analysis, topic modeling, and machine learning classification, can be explored in future research. These methods can bring to light latent topics, sentiment trends, and deeper insights within the corpora. Comparative examination of several corpora can provide information about general patterns and distinctions in language use and content. Exploring cross-corpus relationships can help us grasp textual data more thoroughly. Common words and bigrams can have context and semantic meaning, which can be a useful expansion. Word embeddings and semantic analysis are two examples of natural language processing (NLP) approaches that can assist in revealing the subtle subtleties of the text's context. Domain-specific corpora can be mined using text mining techniques to discover insights and patterns unique to that domain. One can make domain-specific discoveries, for instance, by studying medical literature, legal records, or social media data. Creating interactive visualization tools can make it easier for users to explore corpora. Interactive dashboards, word clouds, and bar charts can all improve user comprehension and engagement.

8 References

1. NLTK Official Website: http://www.nltk.org/
2. Scikit-learn Official Website: https://scikit-learn.org/
3. Word Cloud Generator: https://www.wordclouds.com/
4. Kaggle: https://www.kaggle.com/
5. Towards Data Science: https://towardsdatascience.com/
6. GitHub: https://github.com/
7. Gensim Official Website: https://radimrehurek.com/gensim/
8. ChatGPT by OpenAI: https://beta.openai.com/signup/
9. TextBlob: https://textblob.readthedocs.io/en/dev/
10. DataCamp: https://www.datacamp.com/