Statistical Language Models Analysis

Weston Hawes
California State University, Fresno
March 12, 2023

Abstract

The report presents an analysis of a statistical language model trained with a large set of text to predict the probability of certain words or phrases occurring in a given context. The text is preprocessed by removing special characters and numbers and converting it to lowercase. The report explains the concepts of statistical N-gram models and their application in natural language processing and machine learning. The report outlines the approach, including the dataset used for training, the code breakdown, and the results. The report also includes various plots to visualize the distribution of unique words and bigram occurrences. The project identified 69599 words in the text and generated a plot of the occurrence of each unique word and a plot of the frequency of each unique word. There were 7740 unique words and 40180 unique word pairs identified in this book. The most common word was 'the' with 3260 occurrences and a frequency of .047 and the most common word pair is ‘of the’ with 288 occurrences and a frequency of .0041. The report provides insights into the structure and meaning of natural language and demonstrates the effectiveness of statistical language models in natural language processing.

1 Introduction

This report presents an analysis of a statistical language model trained with a large dataset of text to predict the likelihood of certain words or phrases occurring in a given context. Unlike traditional methods of analyzing language, which relies on rules or dictionaries, statistical language models use patterns in data to learn and predict probabilities of language use. In this report, we explore the concepts of statistical N-gram models and their application in natural language processing and machine learning. We also discuss the dataset used for training, the code breakdown, and the results. The analysis includes various plots that illustrate the distribution of unique words and bigram occurrences, which provide insights into the structure and meaning of natural language. By demonstrating the effectiveness of statistical language models in natural language processing, this report aims to contribute to the development of more accurate and efficient language models for real-world applications.

2 Background Material

Key concepts utilized in this project will be defined and linked to this project in this section.

2.1 Statistical N-Gram Models

Statistical N-gram Models are widely used in natural language processing. These models focus on analyzing sequences of words or characters known as "N-grams" to predict the probability of specific words or phrases that could possibly occur in a given context. N-gram models have many practical applications, such as text classification, machine translation, and language modeling. Unigram and bigram models are also noteworthy examples of N-gram models, with one and two grams, respectively.

3 Approach

3.1 Dataset Used for Training

The dataset that the statistical language model will be trained on is the book "A Room with a View" written by British author E.M. Forster and first published in 1908. The novel has a total word count of approximately 80,000 words. The version of the book that was used in this project is on Project Gutenberg at the following URL: https://www.gutenberg.org/ebooks/2641.

3.2 Code Breakdown

The MATLAB code used to produce this project performs various text analysis tasks on the text file "room_with_a_view.txt", the name given to the text file storing the contents of the book. Here is a breakdown of the approach:

1. The text file is read and stored as a string using the function readTextToStr().
2. The text is preprocessed using the preprocessText() function. The text is converted to lowercase, special characters and numbers are removed, and extra whitespaces are removed.
3. The number of words in the preprocessed text is counted using the countNumWords() function.
4. The number of unique words in the preprocessed text is counted using the countNumUniqueWords() function.
5. The number of unique words with a minimum length is counted using the countNumUniqueWordsMin() function.
6. The occurrence of each word (uni-gram) in the preprocessed text is counted using the countUnigram() function.
7. The occurrence of each pair of words (bi-gram) in the preprocessed text is counted using countBigram() function.
8. The generateText() function generates text based on the probability of the words. The function takes the word pairs and counts them as input, and starts with a given word. It generates a sequence of words based on the probability of the next word given the previous word.
9. The generateTextImproved() function also generates text based on the probability of the words. In addition, it takes an input for the maximum word length and a list of end words.

The function generates a sequence of words such that the generated text ends with one of the specified end words and the maximum length of any word in the generated text is less than or equal to the specified maximum length.

10. Various plots are generated to visualize the distribution of unique words and the bigram occurrences.

4 Results

After cleaning the string of punctuation and other non-letter characters, this project found that there are a total of 69599 words in the book.

4.1 Unigram

Figure 1 is a Plot of the occurrence of each unique word. Figure 2 shows the frequency of each unique word; if all frequencies are added up, we get the sum of 1. The most common word is 'the' with 3260 occurrences and a frequency of .047. This is followed by 'and' at 1872 occurrences and a frequency of .027. These words are thirdly followed by 'to' at 1740 occurrences and a frequency of .025. There were 7740 unique words identified in the book.

4.2 Bigram

Figure 3 is a Plot of the occurrence of each unique word pair. Figure 2 shows the frequency of each unique word pair. The most common word pair is 'of the' with 288 occurrences and a frequency of .0041. This is followed by 'in the' at 257 occurrences and a frequency of .0037. These words are thirdly followed by 'miss bartlett' at 186 occurrences and a frequency of .0026. There were 40180 unique word pairs identified in this book.

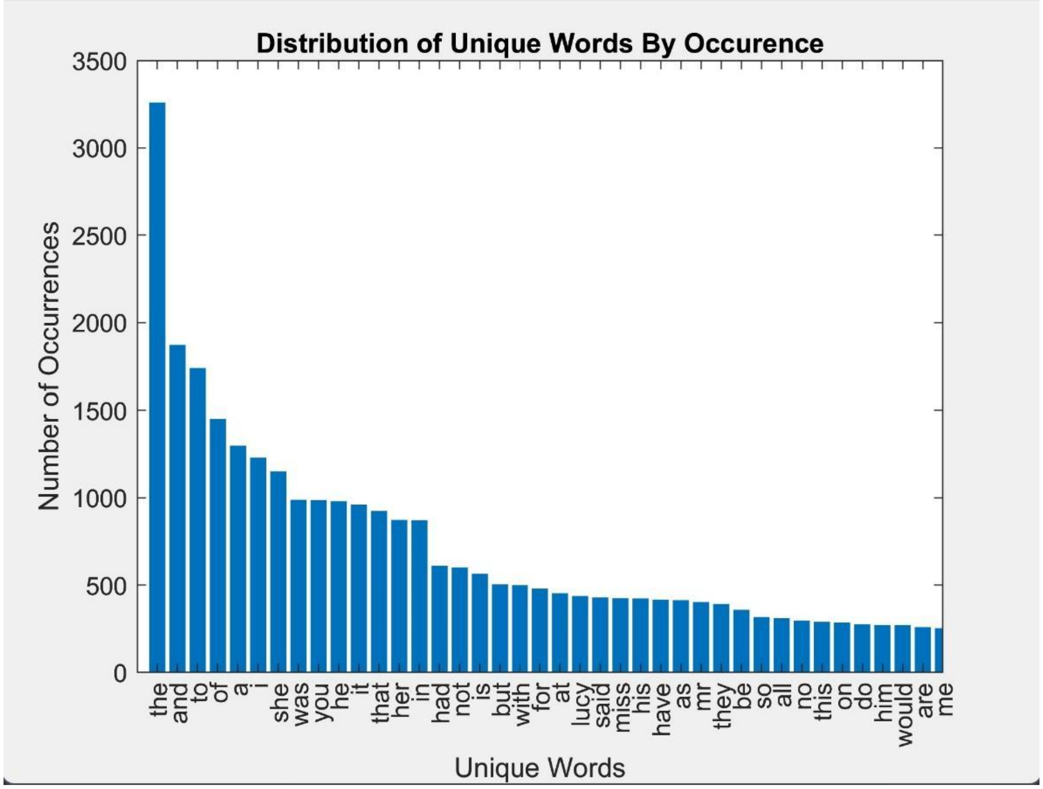

Figure 1

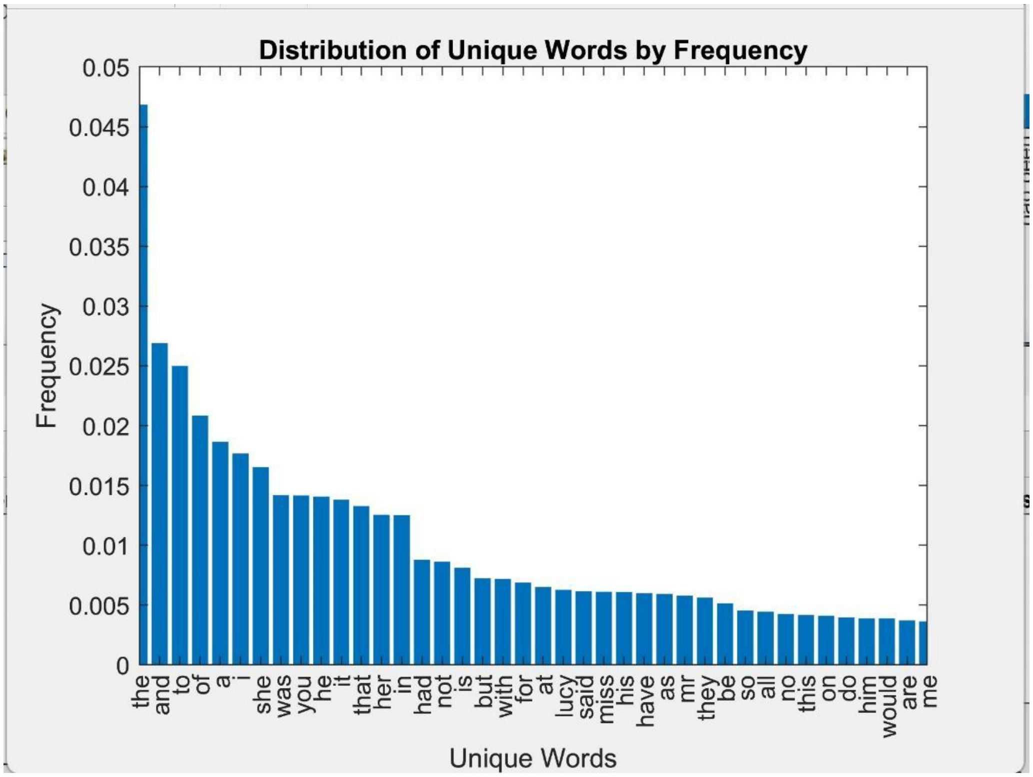

Figure 2

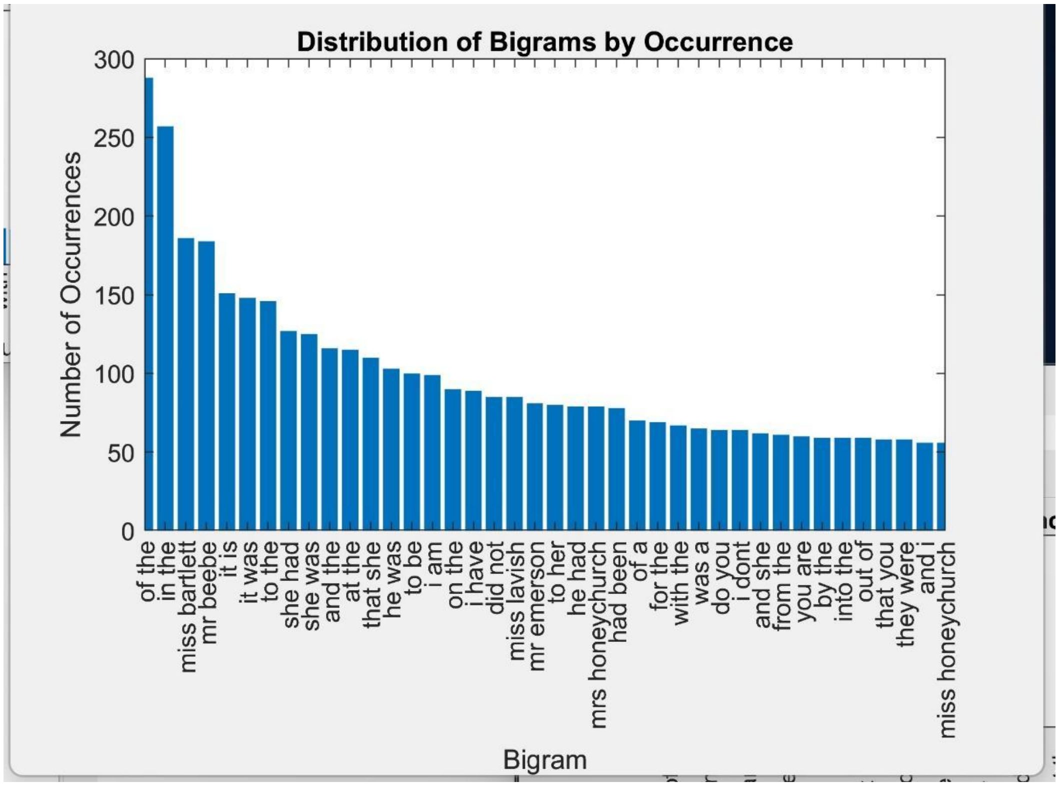

Figure 3

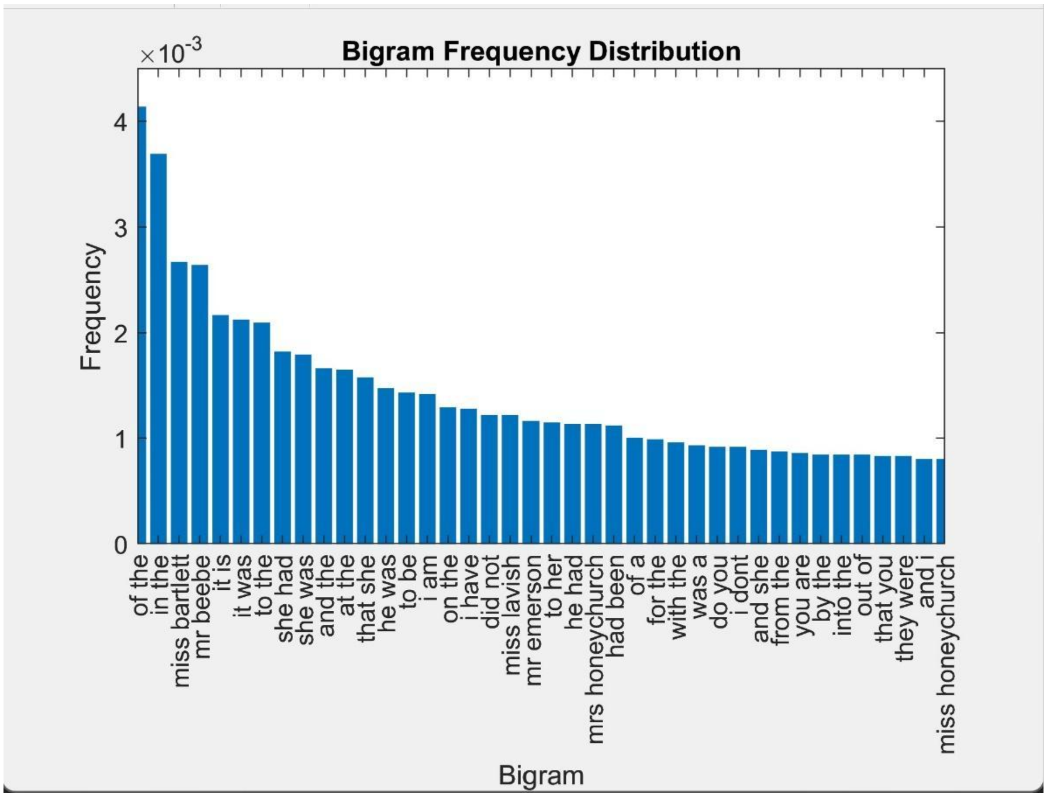

Figure 4

5 Discussion

To further improve this project, additional information can be used to train the n-gram model to help it perform more effectively. One of the limitations of the model is that it only considers the probability of the next word given the previous word or word pair, and does not take into account the overall context or meaning of the text. This means that the generated text may not always be coherent or make complete sense. If the eight parts of speech such as nouns, verbs, adverbs, etc. are taught to the algorithm, it could be used to know how to structure probable word pairs. A dictionary of synonyms could also be used to better predict the next word because this could extend to the words(synonyms) that other text may choose to use, that the training text did not. A dictionary could be used to reduce the probability of words such as pronouns from a training text because pronouns are usually used in specific contexts. Another limitation of the model is that it only considers up to bigram models. While this is sufficient for many applications, higher-order n-gram models may be needed to capture more complex patterns and relationships in the text. The improvement to address this limitation is to expand the model to consider higher-order n-gram models. This would require more computational resources and larger datasets but may provide more accurate predictions and insights into the structure and meaning of natural language.

6 Conclusion

In conclusion, this project successfully implemented an n-gram model in MATLAB to predict the likelihood of certain words or phrases occurring in each context. The model was trained on the book "A Room with a View" by E.M. Forster, and various text analysis tasks were performed on the preprocessed text, such as counting the number of words and unique words, analyzing the occurrence of each word (uni-gram) and pair of words (bi-gram), and generating text based on the probability of words. Additionally, various plots were generated to visualize the distribution of unique words and bigram occurrences. The project highlights the usefulness of statistical N-gram models in natural language processing and computational linguistics and how they can be used to develop highly accurate and effective natural language processing applications.

7 References

E. M. Forster, "A Room with a View," 1908. [Online]. Available: https://www.gutenberg.org/ebooks/264. [Accessed: May 12, 2023].