N-gram model in matlab

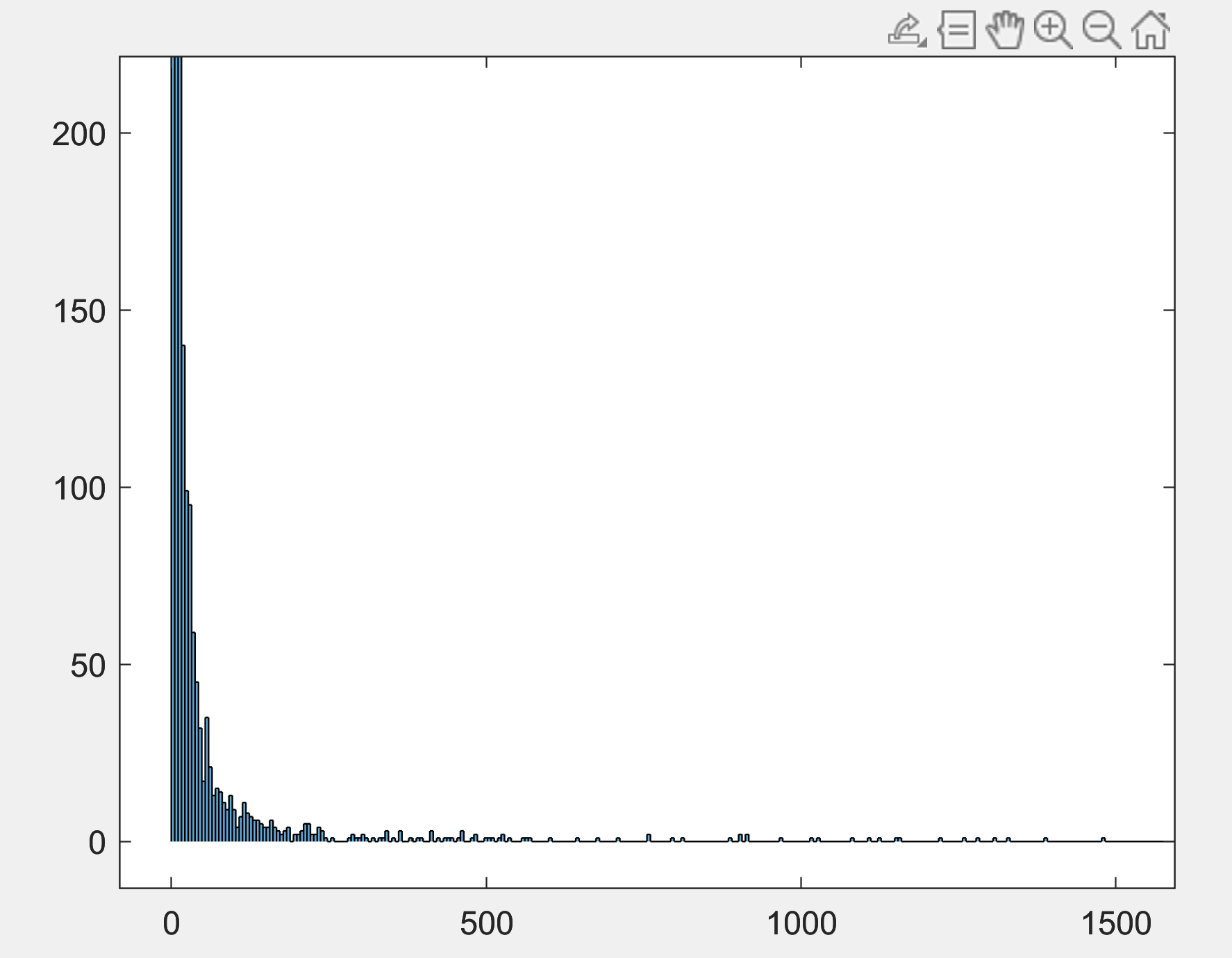
1.

Abstract

In this project, an n-gram model in matlab was trained with large set of text to be able to predict the word given a previous word. I was able to obtain all words, in order minus most special characters. I was then able to extract all unique words as well as their occurance count from the training text. I then attempted to extract all pairs of words and their respective occurance count. I then attempted to calculate the probability of the pairs of words.I lastly plotted the occurrences of the unique words on a histogram.

2.

The book, or the data, used to train the N-gram model is "A Room With A View", by E. M. Forster. I was able to obtain all words, in order minus most special characters. I was then able to extract all unique words as well as their occurance count from the training text. I then attempted to extract all pairs of words and their respective occurance count. I then attempted to calculate the probability of the pairs of words. I lastly plotted the occurrences of the unique words on a histogram.

3.

This is a Plot of the frequency of each unique word. The most common word is ‘a’ with 22934 occurrances. I had to zoom in on the graph or else the only thing that could be seen would be the small tall bar that represents the occurrence of the letter ‘a’. The result I found were that some words were much larger than others, it looks almost exponential.

4.

Additional information can be used to train the n-gram model to help it perform more effectively. If the eight parts of speech such as nouns, verbs, adverbs, etc are tought to the algorithm, It could be used to know how to stucture probable word pairs. A dictionary of synonyms could also be used to better predict the next word because this could extend the to the words(synonyms) that other text may choose to use, that the training text did not. A dictionary could be used to to reduce the probablity of words such as pronouns from a training text because pronouns are usually used in specific contexts.