INLP Assignment 1 Report

Sean O'Sullivan, Johannes Heidicke

October 28, 2016

Abstract

This report details our investigation of Zipf's law and its applicability to text corpora in the cases of word and character frequency. Our results confirmed that word frequencies do approximately obey Zipf's law, however our analysis of character frequencies was inconclusive.

1 Introduction

Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. To test this theory we need to extract all the "natural language utterances" from a text corpus and calculate the frequency of each.

2 Methodology

We used the Natuaral Language Tool Kit to tokenise the corpora. After reading the files into python we applied the WordPunctTokenizer which uses a regular expression to separate the text into words, punctuation, symbols and numbers. We used the PorterStemmer to stem the text, this is the process of reducing variations of the same word to their common root which retains the words meaning without any contextual informatioonn. For example die, dies, died, and dying should all be mapped to a single root, the stemmer

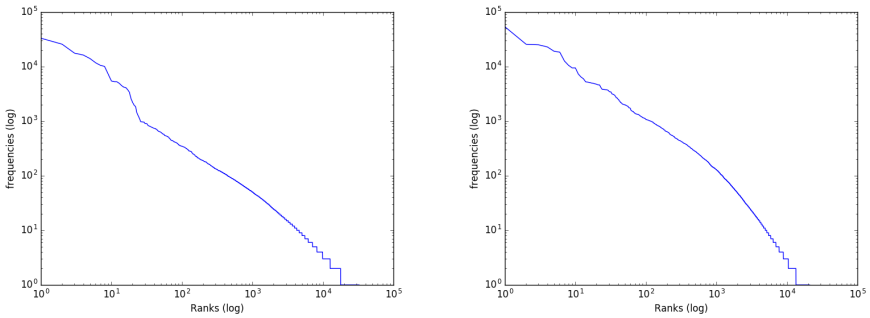
fails in the case of "dying", however. We then computed the K values, the product of the rank and frequency not including letter case and non-words. We plotted the logs of the ranks against the frequencies and histograms of the K values.

We decided to use these methods to separate the text into "utterances" as it was straightforward to implement and also effective. Stemming the text gives a more accurate count of word occurrences as it is not informative to consider different conjugations of the same word as different tokens.

We also applied the same process to the corpora for characters, without stemming of course

3 Results

The results we obtained on the two corpora for words are as follows.

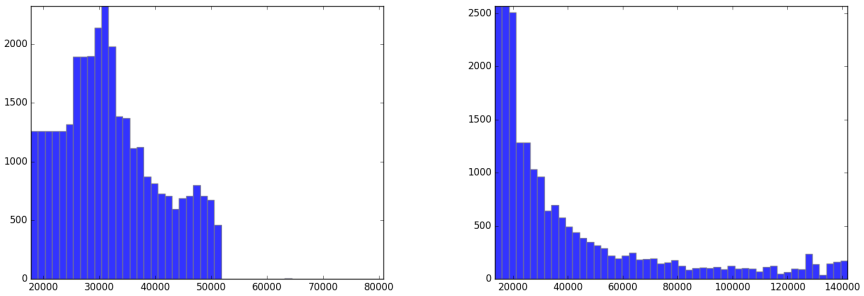


(a) English (b) Spanish

Figure 1: Rank vs Frequency, Log Scale

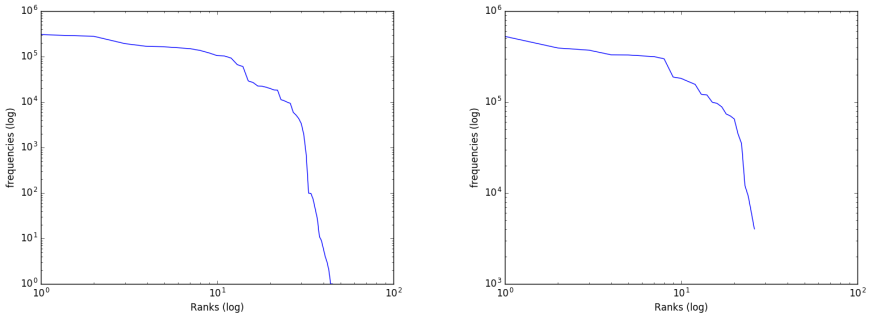
We did try using a more simple tokeniser that was case sensitive and included punctuation, however, as expected the results did not conform to zipf's law as well as those with stemming etc.

Our results seem to suggest that word frequencies are distributed in an approximation of the Zipfian distribution. Rank plotted against frequency



(a) English (b) Spanish

Figure 2: K values in Histogram

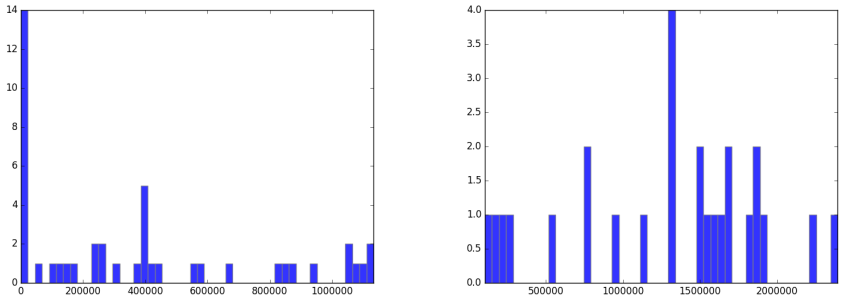


(a) English Characters (b) Spanish Characters

Figure 3: Rank vs Frequency Characters, Log Scale

should result in a mostly linear curve on a log-log plot and this is what we have observed. The K values on the other hand should result in a histogram of uniform height, this was not the case.

Our results for the characters did not conform to zipf's law. They were curved on the Rank vs Frequency log plot instead of linear and the histograms were similarily uneven.



(a) English Characters (b) Spanish Characters

Figure 4: K values in Histogram for Characters

4 Conclusions

Our results for the words were encouraging and suggested that word frequencies are indeed distributed in Zipfian manner, however, there are a number of things we could do to make the experiments more objective.

The corpora could be larger, this would help to alleviate the variance observed in lower frequency words. We could also use lemmatisation instead of stemming as it is more accurate albeit more difficult to implement. Another source of error that could be addressed is the correlation between frequency and rank introduced by using a single corpora. This could be avoided by using separate corpora to calculate rank and frequency.