INLP Assignment 1 Report

Sean O'Sullivan, Johannes Heidecke

October 31, 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 word's meaning without any contextual information. For example "die", "dies", "died", and "dying" should all be mapped to a single root, "die" in

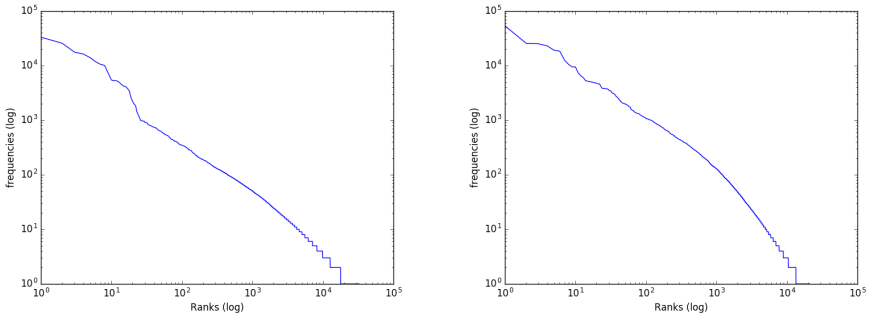
this case. We then computed the K values, the product of the rank and frequency, ignoring letter case and non-words. We plotted the logs of the ranks against the frequencies and histograms of the K values.

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. Ignoring letter case and punctuation does not artificially split up the same word into several tokens such as "go", "go!", and "Go".

We also applied the same process when looking at the corpora at a character level, without stemming of course.

3 Results

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

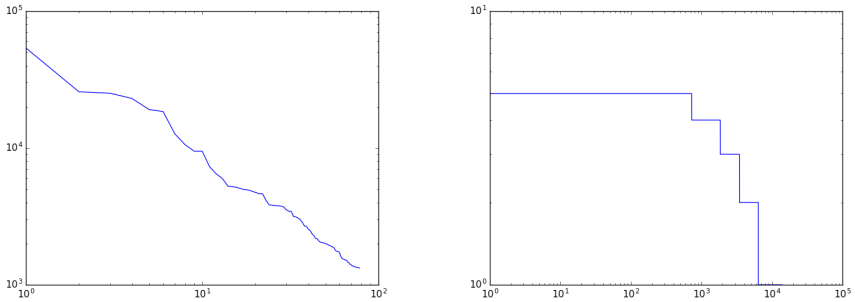


(a) English (b) Spanish

Figure 1: Rank vs Frequency, Log-log Scale

Our results seem to suggest that word frequencies are distributed in an approximation of the Zipfian distribution. Rank plotted against frequency should result in a mostly linear curve on a log-log plot and this is what we have observed (see figure 1).

When looking only at the 100 most frequent words of the English corpora, the variance of K drops significantly to 0.112. Similarly, the variance of K is only 0.324 for the last 15,000 words. We plotted the relationship between rank and frequency for those two cases in figure 2. For the last 15,000 words we observed a step-like behavior which is due to a high number of words having the same frequency. This applies mainly to words that appear infrequently. A larger corpus would help to spread their frequency over a larger range of occurrences and bring the sample frequency closer to the actual frequency in the language.



(a) First 100 words

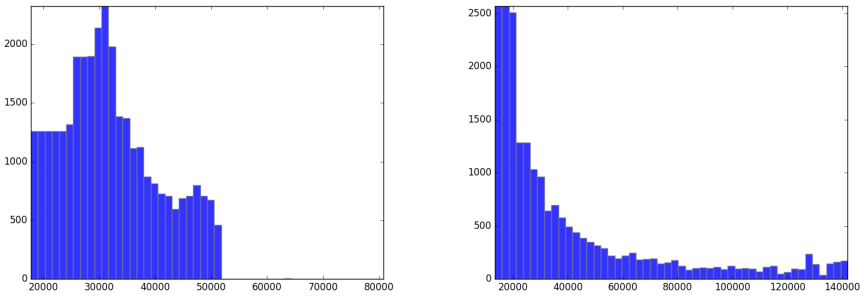
(b) Last 15,000 words

Figure 2: Rank vs Frequency, Log-Log scale

The K values in an optimal Zipfian distribution should all be the same. We observed a rather high variance of K values, with a coefficient of variation (CV) of 0.7928 for the English corpora and 0.2697 for the Spanish corpora. A histogram of K values for both corpora can be seen in figure 3.

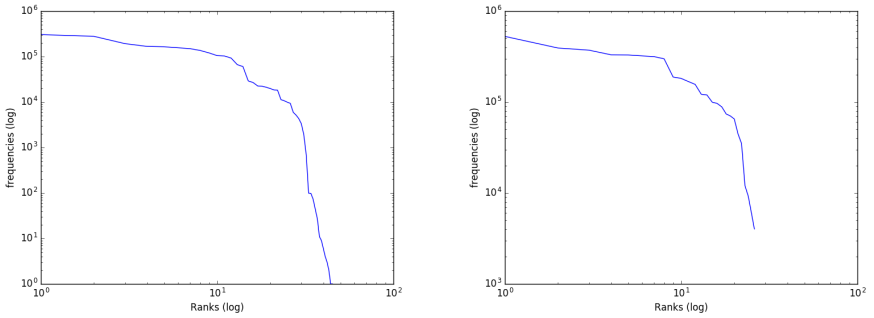
Our results for the characters did not conform to Zipf's law. They were curved on the Rank vs Frequency log-log plot instead of being linear and the histograms were similarly uneven, see figure 4.

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



(a) English (b) Spanish

Figure 3: K values in Histogram



(a) English Characters (b) Spanish Characters

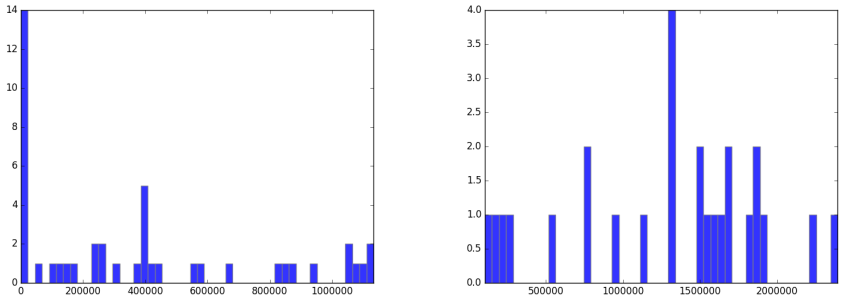
Figure 4: Rank vs Frequency Characters, Log Scale

4 Conclusions

Our results for the words were encouraging and suggested that word frequencies are indeed distributed in an approximately Zipfian manner when looking at the large-scale structure. The word frequencies do, however, also show statistical deviations in their distribution that could not be described with a model as simple as Zipf's law.

There are a number of possible improvements we could make to the experiments.

The most important and obvious improvement would be to use a much larger



(a) English Characters (b) Spanish Characters

Figure 5: K values in Histogram for Characters

corpora. By doing this we could bring our sample frequencies closer to the actual distribution of the language. This effect is most pronounced in words with a low relative frequency in the real language.

We could also use lemmatisation instead of stemming as it may potentially lead to better results. Another source of error that could be addressed is the correlation between frequency and rank introduced by using a single corpora that might be biased in its choice of included words. This could be avoided by using separate corpora to calculate rank and frequency.

5 Attached Files

The code and data for this assignment is provided in the following files:

- `zipfs.py`: The high level code running the experiment
- `zipfs_functions.py`: Functions implementing single steps of the experiment
- `corpus/`: contains the corpora used for this deliverable: `en.txt` and `es.txt`.