Şeyda Portillo Palma
Laila Rahmawati

Introduction

Zipf's Law of Abbreviation qualitatively states that the more often a word is used, the shorter that word tends to be, and vice versa; the less a word is used, the longer it tends to be.

According to Zipf's law, which was initially developed in terms of quantitative linguistics, the frequency of every word is inversely related to its rank in the frequency table for a corpus of natural language utterances. As a result, the most common term will appear roughly twice as often as the second most common word and three times as frequently as the third most common word.

For this work, two poetry books in two different languages, English and Spanish are retrieved to be processed and then compared by Zipf's Rule of Abbreviation. The main reason why the corpora was chosen from English and Spanish is that they are the common languages in which both group members have a good command. Moreover, both are languages that use the Latin alphabet and both their word boundaries are marked by whitespace which is thought to make the tokenization relatively easier. We expect that, even in poems, shorter words are still more frequent, compared to longer words.

Material and methods

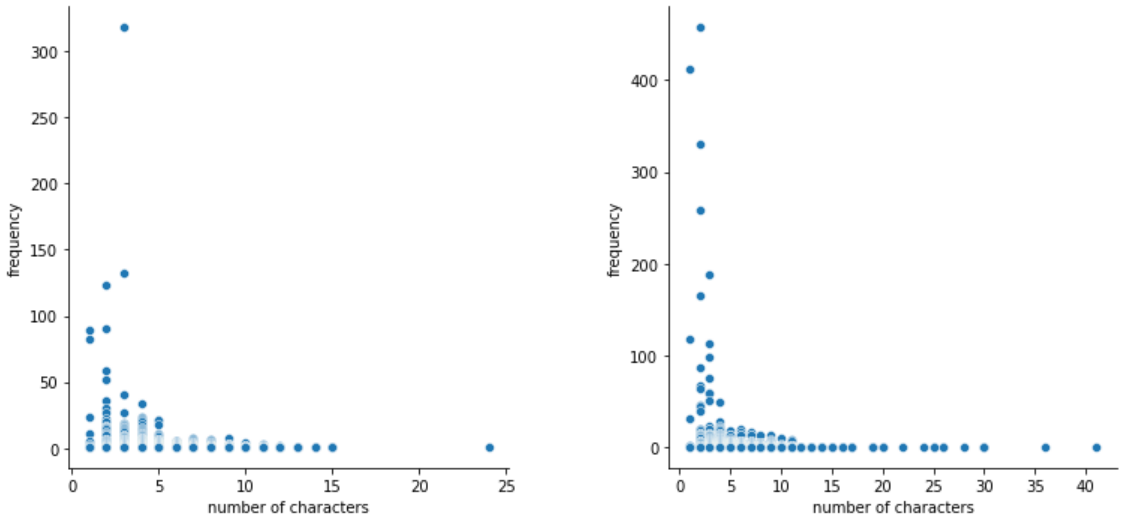
The proposed method begins with data collection and corpus building. Two poetry books in similar size and different languages were collected from the following website: https://www.gutenberg.org/. Our choices are *Poema del Otoño y otros poemas Obras Completas Vol. XI* by Rubén Darío (Spanish) and *The Waste Land* by T.S. Eliot (English).

The overall text analysis is simply divided into two phases as pre-processing and the application of Zipf's rule. Necessary packs are imported on Python.

Pre-processing is realized starting with punctuation removal. Since both files contain some description at the start and end of the file, a split function is also used to eliminate irrelevant parts of the corpora and clean the text. Tokenization is known as separating the text strings into smaller units. In both corpora, normalization steps are implemented and followed by word-level tokenization. The tokenization is done using a split function.

In the second part, word frequency and word length are counted. We use Counter to get the word frequency. After that, the data is put in a data frame and visualized through charts between word length and frequency of words. Same process is followed to extract the tokens on two corpora.

Results



The relation between number of characters and frequency in English (left) and Spanish (right) poems

	word	frequency	rank
	the	318	1
	and	132	2
	of	123	3
	in	90	4
	i	89	5
	...	...	...
	ended	1	1842
	meal	1	1843
	guesses	1	1844
	propitious	1	1845
	word	1	1846

	word	frequency	rank
	de	457	1
	y	412	2
	la	331	3
	el	258	4
	que	188	5
	...	...	...
	oiga	1	2594
	coro	1	2595
	amado	1	2596
	anunciáis	1	2597
	gutenbergs	1	2598

5 most and least frequent words in English (left) and Spanish (right) poems

Based on the data, indeed the most frequent words are the shorter ones both in English (e.g. *the, and, of, in, I*) and Spanish (e.g. *de, y, la, el, que*). The 5 most frequent words in both languages consist of less than 4 characters/alphabets and they are monosyllabic words. In terms of frequency, there is a significant difference of the frequency between the most frequent word (*the*) and the other words in English (see table above), while in Spanish, the frequency difference between the first three most frequent words are very small (see table above).

There are also some words that are relatively short among the least frequent ones (e.g. *word* and *meal* in English; *oiga* and *coro* in Spanish). However, in general, based on the charts above, we can see that shorter words occur more frequently than longer words, which means the zipf's law applies for poems in English and Spanish.

Code

We made our code publicly available on github:
https://github.com/lailarahma93/zipf-s-law/blob/6130dcef5972889073a54563a586712953b5b692/nlp_exercise_1_(zipf's).py

List of contributions

Investigation: Deciding on the corpus to be used and retrieve it using relevant sources

Şeyda Nur Portillo Palma

Conceptualization: Ideas; formulation or evolution of overarching research goals and aims.

Laila Rahmawati

Data Curation: Pre-processing of the selected corpora

Şeyda Nur Portillo Palma

Formal Analysis: Processing of the selected corpora

Laila Rahmawati

Writing: Preparation, creation and/or presentation of the results

Şeyda & Laila