Lab 1

Zipf’s and Heap’s law

Bartomeu Perelló Comas
David Carballo Montalbán

15/10/2020

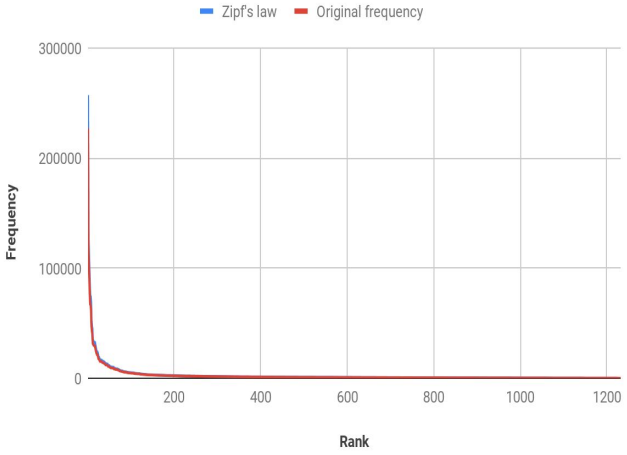
CAIM - GEI FIB

Zipf's law

The main goal of this lab is to check if the rank-frequency distribution seems to follow a power law, so we must find the values (a,b,c) that are best adjusted.

First of all, we used the given python functions to index the files we wanted to use for the session and count the number of words that the index had, we created an small script in python that given the number of total different words and the number of apparitions of each word it gave us the rank of each word in order.

In order to create the plots and manage all the data we used a mix of Calc and Google Drive spreadsheets.



We have used the 20newsgroups to follow Zipf's power law ($f = \frac{c}{(rank + b)^a}$).

By trial and error we found the following values for each parameter:

a = -0,99 ; b = 1 ; c = 1

In the first plot we can see that the distribution seems to follow a power law. Once we remove the higher frequency values we can see the tail more or less as a descending linear function. Because of the redundancy in the lower frequency we only plotted the first 5000 words which almost had 1 frequency.

Heap's law

Heaps' law describes how words behave in the text. That's the relationship between the number of words and the number of different words.

$$W = k\,x\,N^{\beta}$$

W: number of different words
N: number of words in the text
k, β: free parameters

On this occasion, the goal was to obtain the values of k and β, to verify if Heap's law is fulfilled.

In order to experiment, we created three indexes by splitting the novels between them so we could see how the equation behaves with different amounts of total words.

In the end we obtained the following results:

k = 0,0325 ; β = 1