# CAIM Lab, Session 1:
# ElasticSearch and Zipf's and Heaps' laws

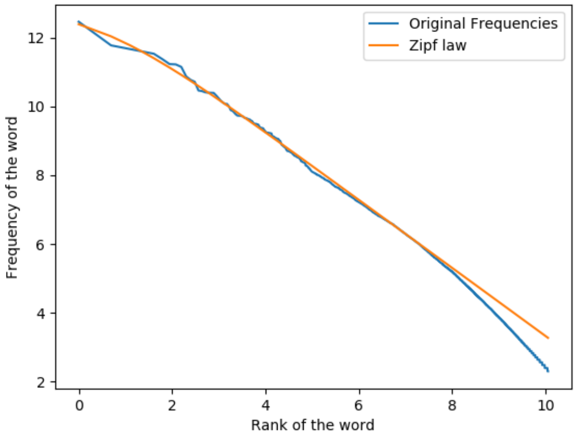
David Carballo Montalbán

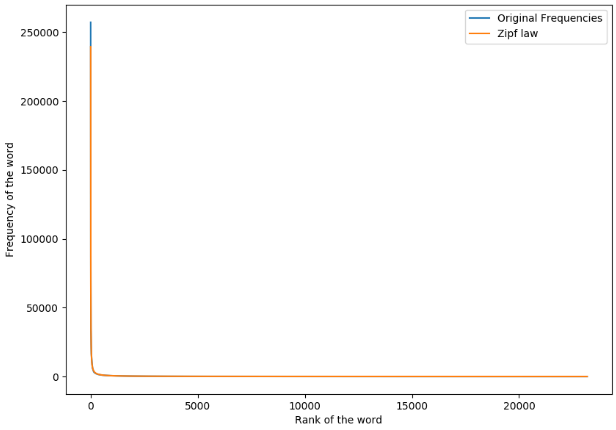
David Sànchez Peña

## Zipf's law

The main goal of this experiment is to find if the rank-frequency distribution of the words in the novels corpus follows Zipf's power law ( $f = \frac{c}{(rank+b)^a}$ ) and then adjust the parameters so that the formula described the distributions.

Before starting the experiment, we filter all the URL's and numbers, as well as the words that have a frequency smaller than 10.





After filtering the data, the plot showed a decreasing curve when represented on a log scale. Using scipy, we found the values that best adjusted to the representation, with the exception of the least frequent words, that diverged from the power law representation. Perhaps we should have been more strict regarding the minimum frequency threshold.

For this experiment, the values that give the best fit are a = 0,99, b = 1, c = 554541.

## Heap’s law

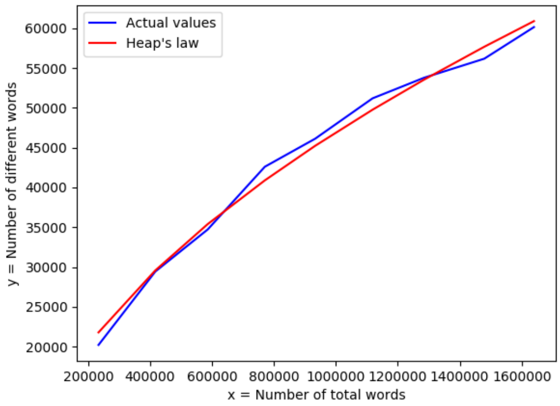
Heap's law (Herdan) describes the behavior of words in a text. That is, the relationship between the number of words and the number of different words.

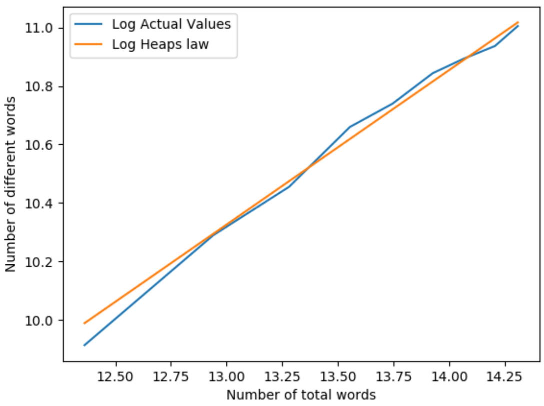
$$W = k\,x\,N^{\beta}$$

W: number of different words
N: number of words in the text
k,$\beta$: free parameters

Our goal is to obtain the values of k and $\beta$, to verify if Heap’s law is fulfilled.

First of all, we need to create indexes that contain different amounts of text. Therefore, we have been eliminating a part of the novels (2MB) to end up creating 9 indexes, half the size of all the files (18MB).





As we can observe in the previous plots, we have approximated the function and obtained the following values for k and $\beta$:

k = 32 $\beta = 0$

In conclusion, we confirm that it follows the law of Heaps.