Information Retrieval and Analysis
OCTOBER 2020

Lab Session 1: Power laws

PROJECT REPORT

Elías Abad Rocamora
Victor Novelle Moriano
Barcelona, UPC - FIB & FME

Exercise 1:

After plotting the frequency of the words depending on the rank, we notice that the curve decreases very fast, meaning that there are very few words with high frequency, and many more words with a very low frequency.

Exercise 2:

As we have already seen in class, the plot of the variables, using a logarithmic scale in both axes, should approximately follow a straight line if its distribution follows a power law. As we know, a power law is:

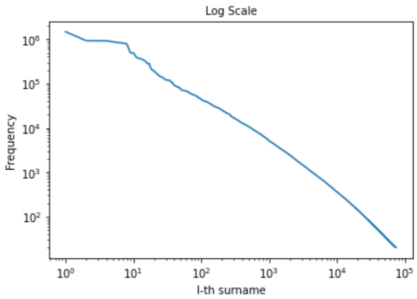
$$y = c \cdot (x + b)^a \quad (F.1)$$

where *a*, *b* and *c* are the constants that will determine the behavior of the function. If we apply logarithms to both variables and forget about the *b* parameter, we obtain the following:

$$log(y) = a \cdot log(x) + log(c) \quad (F.2)$$

Applying this logarithmic transformation, it is clear that $log(y)$ is a linear function of $log(x)$. This justifies the previous statement, which said that if our data follows a power law, the log-log plot should be a straight line, having a as slope and $log(c)$ as intercept.

Exercise 3:



As we mentioned in the previous exercise, if the log-log plot follows approximately a straight line, the data follows a power law. We can see that this is clearly that case, despite having little fluctuations on the *[5,25]* range.

Exercise 4:

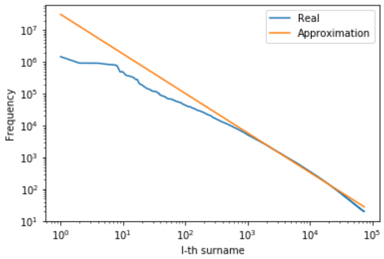
Now that we know that our data follows a power law, we can estimate the parameters of that law for this case. Our first approach will be to choose two points from the log-log line and solve for a and $log(c)$ the following system of equations (coming from F.2):

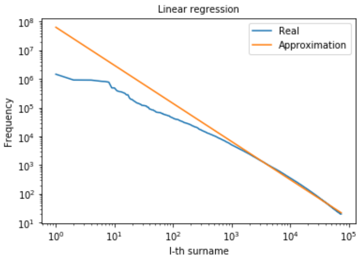
$$log(y_1) = a \cdot log\ (x_1) + log\ (c)$$
$$log(y_2) = a \cdot log\ (x_2) + log\ (c)$$

Leading to:

$$a = \frac{log(y_1) - log(y_2)}{log(x_1) - log(x_2)}\ ,\ log(c) = log(y_1) - a \cdot log(x_1)$$

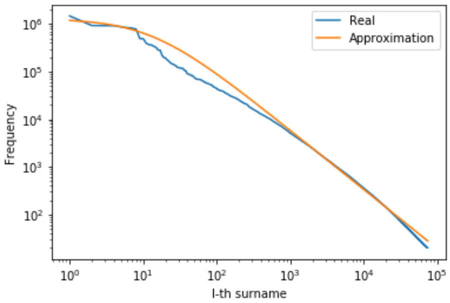
In this case, we chose $(x_1, y_1) = (2500, 1875)$ and $(x_2, y_2) = (20000, 141)$, so $a = -1.244$ and $log(c) = 17.272$





As we can observe on the left image, the adjustment of the line is not as good as we would wish. To try to get better results, we have done a linear regression, but the results aren't much better (right image).

The last thing that we can do is add the b parameter and try some values to improve the adjustment. After some attempts, we arrived at the following result with $b = 13$.

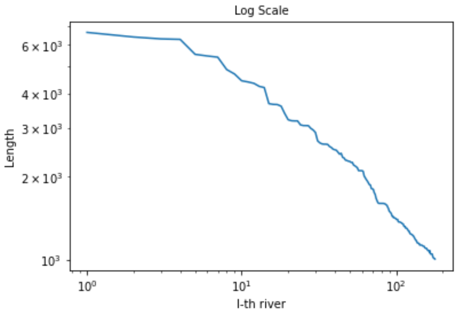


The main improvement found when adding the b parameter is that we get a better adjustment for lower rank values than before, while not affecting large-rank observations. The straight line failed to "explain" low-rank names, as the slope and intercept were calculated from two large-rank names.

Exercise 5:

This procediment can also be done using a spreadsheet, in our case *Google Sheets*. The importation of the csv is performed automatically by the software, and using the intuitive *GUI* the plot of the requested data can be easily done. Once the main graphic is generated, using the *customize* option, we can modify the scale of the axes (to apply logarithms) and also request a trendline of the plotted series, where we can also select which type of trend we want to use (*logarithmic, exponential, e.g.)*. This software is very useful because it allows people that may not have any program skills to perform the above exercises pretty easily.

Exercise 6:



It can be seen that the log-log plot of *Length* against the *I-th river* follows approximately, although is not as clear as in *Exercise 3*. We can differentiate two main regions, one made up by the first 50 rivers (*approx)* which has a slow decreasing slope, meaning that all of them have similar lengths, and the other is conformed by the rest of the rivers, having a steeper slope which indicates a bigger difference between their lengths. This slope difference will be corrected by the *b* parameter in the power law formula.

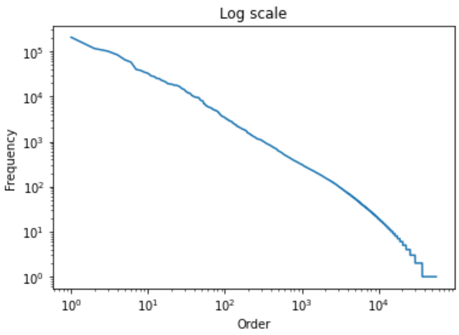
After plotting the basin area variable, we noticed a humongous value, which corresponds to the *Amazon Ucayall Apurímac* river, having a value of 7050000, an order of magnitude higher than the average value (766244).

Exercise 7:

The first step we performed in order to solve this exercise was a careful reading of the provided code to understand what was its functioning and what needed to be added.

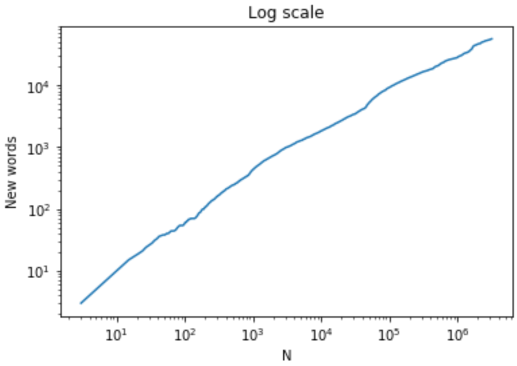
After understanding it, and with the help of the recommended links and other web pages, we managed without many difficulties to successfully complete the code and fulfill the imposed requirements.

Exercise 8:



Using the code generated in *Exercise 7* to compute the frequency of appearance of the words present on the *novels* folder we can see that its log-log plot follows approximately a straight line, indicating that it appears to follow a power law.

Exercise 9:



After changing the code and executing it with $k = 3$, we obtained this log-log plot, indicating that indeed, these texts follow Heaps law with an exponent (a) between 0 and 1.