4.12 Optional Assignment

Exploratory Data Analysis: Learning Activity

Frequency-Word Distribution in Natural Languages

- **Data acquisition:**

  1. Reading List: Zipf's law

  https://en.wikipedia.org/wiki/Zipf's_law

  2. Select all the records for the three columns of “Rank”, “Word”, “Count” including the column names from https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/PG/2006/04/1-10000

  3. Copy and paste the 10000 selected records to a blank Excel worksheet

  4. Save the workbook as a Comma Separated Value file (say English.csv)

- Exploratory analysis in R

  1. In Rstudio, using setwd select your working directory as the one containing English.csv

2. Load the file with the function read.csv (for convenience study and use the option header = TRUE)

3. Convert the columns Rank and Count first to characters via the function as.character and then to real numbers via as.numeric

4. Explore the relationship between Count and Rank by plotting, e.g., using the linear, semilog, and log-log plots

5. Which plot is the best to characterize the dependence of Rank on Count and why?

Predictive analysis: Discovering Zipf’s law

1. Calculate the Pearson correlation coefficient between log(Rank) and log(Count). Interpret the obtained value.

2. Study the function lm for performing linear fits in R (to quantify the linear dependence)

3. Show that if the quantities y and x relate to each other by a power law: y = b*xa then log(y) = a * log(x) + log(b), where a and b are constants. Note that log(y) linearly depend on log(x) with a being the slope and log(b) is the intercept.

4. Using lm to find values a and b in the power law Count = b * Ranka

5. Confirm the found power law dependence by plotting it on top of the raw data

6. Does the power law dependence exist only for the English language?

Click here to download the list of word in csv file

Click here to download the code