Text Mining

Manipulating text data with R

The following packages are required for this lab: `text2vec`, `Matrix`. In order to install them with the `install.packages` command, you need to update to the latest version of R (as of now, 3.5.1).

Exercices

Part 1 - Analyzing a corpus of movie reviews

1. Load the corpus from "reviews.csv" into a data frame, identify the descriptors and print the number of reviews for each class.
2. Compute the vocabulary, then print the 10 most common words and plot the word frequency distribution (limit word frequency to 20).
3. Plot the word frequency vs rank, both with linear and logarithmic scales (only consider the 200 most frequent words).
4. Fit Zipf's law with the `lm` function.
5. Prune the vocabulary, then vectorize the corpus.
6. Write a function that returns the cosine similarity between two documents.
7. Measure the similarity between some pairs of documents; apply tf-idf weighting and measure the similarity again.

Part 2 - Analyzing random text

1. Write a function that returns a fixed length sequence of characters drawn from a uniform distribution.
2. Modify this function so that words don't exceed a given length.
3. Add the possibility to limit the number of distinct letters.
4. Fix the number of characters to 4, and consider the following distribution: $P(\text{a}) = 0.5, P(\text{b}) = 0.13, P(\text{c}) = 0.1, P(\text{d}) = 0.07, P(\text{space}) = 0.2$
5. Generate a sequence of 10^5 characters and fit Zipf's law.

Documentation

- Plotting histograms: https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/hist.html
- Fitting linear models: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html
- Matrix multiplication: https://stat.ethz.ch/R-manual/R-devel/library/base/html/matmult.html
- Applying tf-idf weighting: http://text2vec.org/vectorization.html#tf-idf

Oct. 2018 – Maestría en Ingeniería Matemática, Universidad de la República, Uruguay
Instructor: Adrien Guille