Problem Set 6 - Text Analysis

August 2022

Depending on your level and your interests, you can either start with completing the A tasks or go straight to the B tasks.

A. Preprocessing and basic statistics with tm

1. Load the tm package
2. Generate a corpus object from a set of character vectors directly in Rstudio by copying four short texts online like quotes, the first lines of Wikipedia entries or tweets.
3. Preprocess the corpus in any way you find fitting for you documents with the help of the tm_map command but see to remove stop words related to the language in which your quotes are written.
4. Construct a document-term matrix from your corpus object and take out the 10 most frequently occurring terms for each document.
5. Look into what words correlate with the 10th most frequently used word in each document.
6. Generate a tf-idf object (i.e. term frequency–inverse document frequency) to inspect what words are most distinct for each document.

B. Text analysis with tidy text

1. Load the tidy text package and the gutenbergR package. Download the four books by the author Woolf, Virginia by only using Rstudio.
2. Convert your set of books into the tidy text format and remove stop words with the list provided in tidytext package.
3. Look up the 10 most frequently occurring terms for each book.
4. Construct a tf-idf object (i.e. term frequency–inverse document frequency) and look up the five words with the highest score for each document. Compare how and think about why they differ from the 5 most frequently used words (task B.3).
5. Generate a word cloud on the corpus level (make sure to only include the most frequently occurring terms to not overload your system and make the graph interpretable). Note if any words not present in the results given from task B.3 are present.
6. Run a LDA topic model with 6 topics and inspect the prevalence of topics in the documents (the gamma). Looking into the words constituting each topic (the beta) relates to words in the tasks above.