Problem set 7

Exercise 1. Text mining with R

Download one book in .txt format that you like and analyze the word frequency of this book (or another social media dataset) by following the steps from the lecture.

(You can try Gutenberg Library https://gutenberg.org/). Do not forget to remove metadata, so that you have only plain text documents. For more information, read the Introduction to the **tm** Package Text Mining in R:

https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Plot Term-occurrence histogram and word cloud for your book. Choose the corresponding parameter that suits your data.

https://www.kaggle.com/code/ranamahmud/tweet-sentiment-extraction-sentiment-analysis-in-r/notebook

https://www.kaggle.com/code/rtatman/tutorial-sentiment-analysis-in-r/notebook

Exercise 2. Text mining with Python

In Python Notebook textJEWK.ipynb we compared word frequency from two books. Chose two of your data sources and compare the word frequency there. Add some comments to the code you find important. Find how to remove numbers from text. Add some new features to your code from the online book on NLTK: https://www.nltk.org/book/

Exercise 3. Text mining with RapidMiner

Build the process in RapidMiner to analyze the sentiment of a text, choose the text you would like to analyze, and analyze the sentiment of the chosen text.

https://rapidminer.com/glossary/sentiment-analysis/

https://www.meaningcloud.com/developer/sentiment-analysis/doc

https://www.meaningcloud.com/developer/rapidminer-extension/doc/2.1/getting-started