TU Dortmund University
Department of Statistics
Dr. J. Rieger
M. Sc. K.-R. Lange

Winter Semester 2024/25
12/03/2024

Exercise sheet 8

Natural Language Processing

Hand-in (voluntarily): 12/09/2024 until 11:59 p.m. via Moodle
Please submit a `.py`, `.ipynb`, `.R` or `.rmd` file!

Task 1

In moodle you will find the file `biden.csv`. It contains every tweet in the month prior to the U.S. presidential election in 2020 containing the hash tag #joebiden. Load the file into your console. It contains each tweet in the "tweet" column and the date of the tweet's creation in the "created_at" column.

We are interested in how the topics of the tweets develop over time. For this, we will train a dynamic topic model called RollingLDA on the speeches and compare the resulting topical changes in the following tasks.

Task 2

Remove unwanted fragments that are not relevant for our analysis. Preprocess the texts so that they are fit for an analysis. Argue the use the preprocessing steps you take for the given analysis.

Task 3

Train a normal LDA on the entire corpus with $K = 30$.

Task 4

Train a RollingLDA on the corpus. Set the time chunk length to three days and choose $K = 30$. If this takes a lot of time, chose `prototype=1` and lower the epoch count.

Task 5

Compare the evolution of topics within the model: does the content of any topic change in particular between the time chunks? Would you prefer this model or the model you used in task 3?

Recommended packages & functions

R: `RollingLDA, topWords(getTopics(RollingLDA))`
Python: `https://github.com/K-RLange/ttta, ttta.methods.rolling_lda`