Technische Universität Dortmund
Department of Statistics
Dr. J. Rieger
M. Sc. K.-R. Lange

Winter Semester 2023/24
12/04/2023

Exercise sheet 7

Text as Data

Hand-in (voluntarily): 12/08/2023 until 11:59 p.m. via Moodle

---

Task 1

In moodle you will find the file `ASoIaF.zip`. It contains the five books of the "A song of ice and fire" series in a plain txt-format. Load all files into your console.

We are interested in how the story and its themes develop over time. For this, we will train a topic model on each book and compare them. For Python-users, you will find a better function to show the top-words in `utils.py`.

Task 2

Remove unwanted fragments that are not part of the narrative. Then split the texts into individual chapters, resulting in one large (chronologically ordered) list of chapters for each book.

Task 3

Preprocess the texts so that they are fit for an analysis. Argue the use the preprocessing steps you take for the given analysis.

Task 4

Train five LDAs with $K = 10$ and 50 iterations on the very first book. Compare the resulting topics from these five models. What do you notice? Is "topic 1" the same topic in all models?

Task 5

Train an LDA with $K = 10$ and 50 iterations on each book separately. Compare the models by keeping your findings of task 4 in mind.

Additional information for Python users

The top words function of the gensim function outputs unweighted top words (which will be dominated by stop words). To get more meaningful top words, you can use the function we provide in the file `utils.py`, which takes a gensim LDA object as an input.

Recommended packages & functions

R: `tosca::LDAgen, tosca::LDAPrep, tosca::topWords()`

Python: `gensim.corpora.dictionary`, `gensim.models.ldamodel`, `gensim.models.ldamodel.top_words()`