Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Topic Modelling in Literary Studies

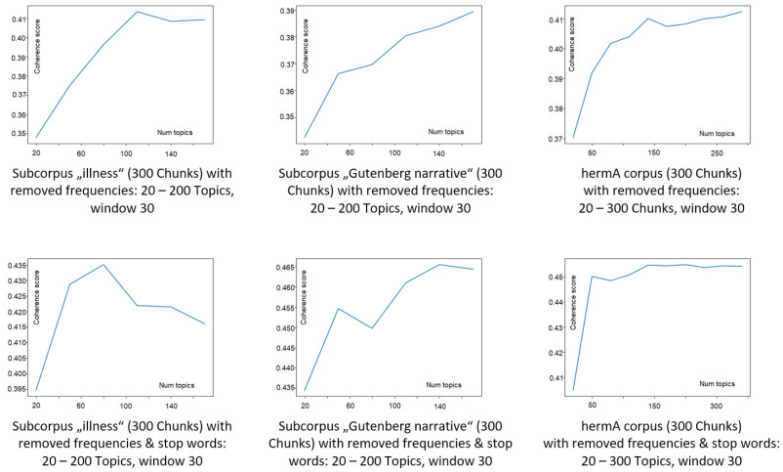

Figure 1

Introduction

The project aims to calculate and evaluate a series of topic models for a corpus of literary texts (a collection of about 1,800 German-language prose texts – circa 90 million words – from the period between 1870 and 1920) with various stages of data transformations (i.e. removing of high-frequency and low-frequency words, numbers as well as punctuations, partial segmentation into text chunks of 300 or 500 words, the restriction to certain word-formation classes) and with various combinations of parametrisation (number of topics, number of iterations, window size, etc.) to shed some light on the regularities between text features and the quality of the topic modelling performed for literary texts.

Methods

Topic modelling was performed using the LDA algorithm (Latent Dirichlet Allocation) implemented in the MALLET (MAchine Learning for LanguagE Toolkit) package. Topic modelling is a text-mining tool for automatically discovering a corpus's thematic structure. The content of any text can be characterised through a limited set of themes. All the words in a text are derived from it. From the observation it is known that the words belonging to similar topics usually appear in a similar context. That is what the LDA algorithm is based on. It calculates the probability of words appearing in a particular context and groups words into these thematic clusters. The quality of the models was evaluated using the coherency measure Cv. It measures the contextual dependencies between words within a topic and can be interpreted as a measurement of the meaningfulness of a theme. The higher the value obtained, the better a topic or model should be. This operation is one of the most energy-intensive because it is based on a stepwise increase of the number of topics using a sliding window algorithm. A coherence

Project Manager
Dr. Inna Uglanova

Researchers
Svenja Guhr

Principal Investigator
Prof. Dr. Evelyn Gius

Project Term
2020 - 2021

Clusters
Lichtenberg Cluster Darmstadt

Additional Software
MALLET (MAchine Learning for LanguagE Toolkit), NLTK (Natural Language Toolkit)

Institute
Institut für Sprach- und Literaturwissenschaft

University
Technische Universität Darmstadt

value is calculated for each step (window). For example, in one of the experiments, Cv was calculated from 20 to 300 topics with a step of 20. It took about 23 hours to complete a job on the Lichtenberg cluster and was utilised 806,24 GB of memory. In addition to the quantitative evaluation, a manual assessment of the interpretability of the topics was used. Finally, the visualisation package PyLDAvis was used to evaluate the distribution of topics.

Results

One main finding of the project is that each data configuration type has its own coherence profile:
• the ascending coherence type corresponds to non-segmented raw data;
• the descending coherency type corresponds to non-segmented cleaned data;
• the parabolic coherency type corresponds to segmented data, both raw and cleaned;
• the discontinuous type corresponds to non-segmented data and data with selected word classes, in both cases for a special subset of the corpus.
It means that the Cv measure can be used not only to find the best model but also to understand how the results of topic modelling differ if we manipulate the data structure (by removing certain frequency classes and sop words, by segmentation, and by word-class selection). The curve progressions and visualisation of distributions of topics by pyLDAvis-tool demonstrate what happens to the structure of a text in each case and how this affects the quality of the machine modelling. In general, our experiments with data of varying sizes have demonstrated that the quality of topics can be improved simply by increasing the size of a data set.

Discussion

The conducted experiments provide new insights into the relationship between topic modelling as a tool and a literary text corpus. Manipulations with text structure clarify the relationship between cohesive and coherent mechanisms in a literary text. Cohesive structures make thematic modelling difficult. Cohesion elements as pronouns or conjunctions are hardly meaningful from a thematic perspective, but they are the basic building blocks of the textual structure. A half of any text consists of this sort of unit. Getting rid of these elements through pre-processing improve the quality of topics. The problem is that structure in literary text (compared to the structure of a scientific text, for example) can have its own meaning. If the structure is changed, the meaning is also changed. Therefore, any of the manipulations changes structure and meaning. These changes are reflected in the coherence types. The fewer cohesion elements the structure contains, the more heterogeneous and complex it becomes in the thematic sense.

Outlook

The next step would be to test other modelling approaches with other evaluation metrics, comparing the literary corpus with a non-fiction corpus.

Publications

Uglanova, I.; Gius, E.: The Order of Things. A Study on Topic Modelling of Literary Texts. In: Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020: 57-76. http://ceur-ws.org/Vol-2723/

Last Update: 2022-04-27 17:06