Using Machine Learning techniques in Qualitative Data Analysis

Zsolt Takacs

247886T

Proposal

Motivation

Qualitative data analysis is a long and tiring process that is extremely difficult to scale up. Utilizing machine learning techniques to help researchers in this area can not only minimise their workload but can also enhance the subjectivity and reproducibility aspects that qualitative research is often under scrutiny.

Aims

This project aims to develop an accessible web application to allow users to analyse their textual data with machine learning techniques. The current scope of the project includes LDA Topic Modelling and sentiment analysis. However, further techniques should be easily integrated with the end system.

Progress

- Language and major libraries chosen: **Python** 3 to develop a **Django** webapp using **Gensim**, **Spacy** and **NLTK** for the analysis of data
- Design, system architecture and requirements confirmed with supervisor with wireframing and prototype
- Django container webapp developed to fully map the logic of the site
- Underlying scripts created to enable data analysis
    - Reading of files
    - Cleaning and processing of texts
    - Executing LDA topic modelling and finding optimal number of topics
    - Sentiment analysis on a sentence basis
- Compiling results as a PDF documents
    - Most important words for each identified topic is returned
    - Original text annotated by highlighting sentences that contain at least one of the important words (every topic has a separate result file)
    - Sentence-based sentiment scores visible next to highlighted sentences and aggregated for each document and the overall corpora
- Continuous testing of developed features

Problems and risks

Problems

- One of the libraries (pyLDAvis) that is used to create interactive visualizations, builds on deprecated packages.
- Correctly install and use LaTeX compiler on the CI pipeline
- Somewhat difficult and time consuming to deploy product from scratch.

Risks

- No current substitute for pyLDAvis → Look for possible alternatives or how to resolve issue locally
- No deployment for newer versions → Research how to make deployment as easy as possible
- Too many suggestions in pilot evaluation → Be clear about scope and limit features to implement, while keeping the rest as part of the future work section.
- Struggling with dissertation write-up → Start writing up early, to allow plenty of time

Plan

Semester 2

- Week 1-2: Finalise product for pilot evaluation
  - Deliverable: Working deployed project with all feedback incorporated
- Week 3: Pilot evaluation and start working on write-up
  - Deliverable: Notes on evaluation and advance with write-up notes
- Week 4: Collect results from pilot evaluation and start implementing suggestions
  - Deliverable: List if new features to be implemented as suggested by pilot evaluators
- Week 5: Finish implementing suggestions from pilot evaluation.
  - Deliverable: Working deployed project that implements newest fedback
- Week 6-7: Start final evaluation and start write up concurrently
  - Deliverable: Start working on draft submission
- Week 8: Finish evaluation and submit draft dissertation
  - Deliverable: Write up draft evaluation section
- Week 9: Work on presentation and start refining dissertation
  - Deliverable: Draft presentation and notes on improving draft dissertation
- Week 10-11 Finalise dissertation and presentation.
  - Deliverable: Finished dissertation and presentation.