TECHNISCHE
UNIVERSITÄT
WIEN

Group 03

Topic Modeling on Podcast Short-Text Metadata

Anabel Dautović, Medak Mirta, Sharma Hrithik, Bosse Behrens, Felix Neuwirth

Paper Overview

- Paper‘s goal is to explore feasibility of discovering relevant topics from podcast metadata (titles and descriptions)
- Proposes Named-Entity-informed Corpus Embedding (NEiCE)
- Leveraging Named Entities (NEs) within a Non-negative Matrix Factorization framework (dimensionality reduction for analyzing data and extracting features)
- seeks to improve topic coherence over state-of-the-art methods by addressing challenges in short-text metadata
  - data sparsity
  - noise

Strategy

- Goals
  - Reproduce experimental setup, execution, and results from paper
  - validate findings
  - identify possible inconsistencies

- Steps
  - Become familiar with paper/methods used
  - Get data if possible
  - Follow provided steps for preprocessing
  - Follow provided steps for environment setup/execution
  - Compare results to provided results in paper
  - Run statistical tests on results
  - Draw conclusions, check quality of experimental setup in paper
  - Evaluation

Setup

1. Entity Linking	2. Data Preprocessing	3. NEiCE Model	4. Evaluation
Extraction of named entities from the podcast metadata and their linking to Wikipedia entities.	Preprocessing of the podcast metadata considering the identified Nes from the previous step.	Apply NEiCE to the extended preprocessed data.	Compute CV score to evaluate coherence of the topics extracted with NEiCE.
Requirements (issues) not mentioned in the readme and not up to date in the code:			
▪ 60GB+ memory for the pretrained Entity Linker models ▪ gensim package	▪ Merging the obtained json files into one file (merge script added) ▪ Changes to the preprocessing functions due to deprecated methods	▪ Deprecated methods ▪ Too few arguments specified in the example command	▪ Very long execution time ▪ No other metrics than CV score ▪ No script for varying parameters to obtain more results (as in the paper)

Conclusion so far / Remaining Work

- Challenges
  - Partly missing data (Spotify dataset)
  - Outdated versions of libraries and models
  - Modifications necessary to run Code
  - Computationally heavy and time-consuming experimental process
- Intermediary results
  - First t-tests show some significant difference between paper’s results and ours
- Remaining work
  - Further statistical testing
  - Final validation of results
  - Inspection of practices, plausibility of metrics/models used
  - Evaluation