188.992 - Experiment Design for Data Science

Topic Modeling on Podcast Short-Text Metadata

Group 26-B

Terezia Olsiakova 12331438
Navya Velagaturi 12307012
Trevor Calvin Baretto - 12332204
Francesca Ricter 12331433

GOAL

The objective of the task is to assign a topic label to podcast via metadata using NEiCE approach, leveraging named entities often present in podcast titles and descriptions to derive additional context information

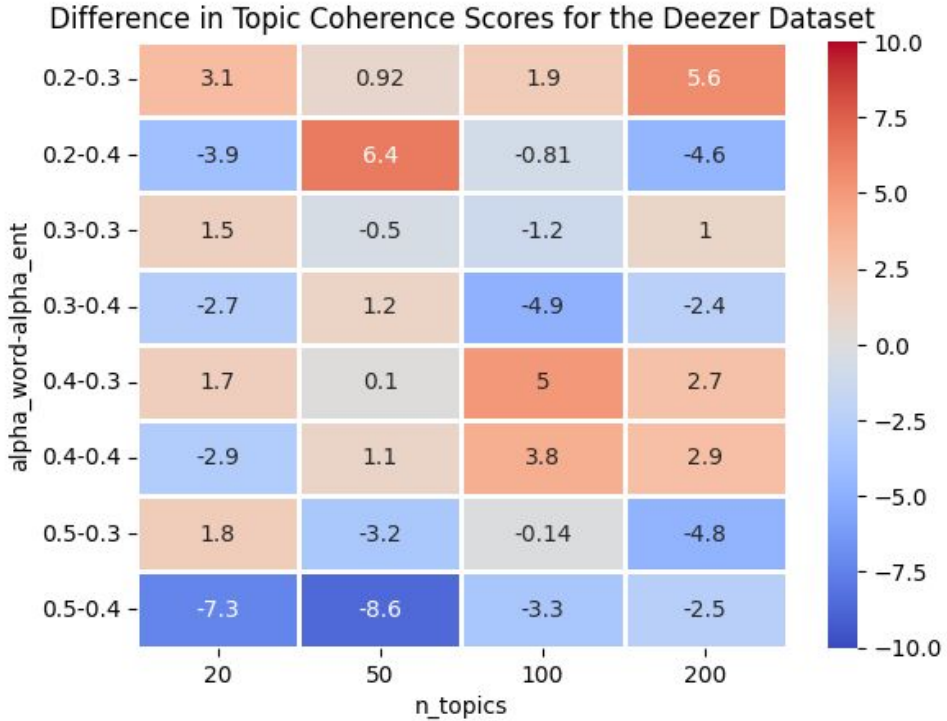
Challenges

1. **UNAVAILABLE DATA**: The Spotify dataset is no longer accessible
2. **REQUIRED SOFTWARE**: we use Docker
3. **UPDATES**: modifications in entity linking packages, model versions and in 'names_dataset'. This leads to variations in topic coherence scores, however the overall trends and conclusions should remain the same as stated in the GitHub repository
4. **ADDITIONAL PREPROCESSING FOR ITUNES**: The iTunes dataset requires further manual preprocessing to meet the desired format, as mentioned in the paper.
5. **ENSURING CONSISTENT ENTRIES IN THE ITUNES DATASET**: experiment using language detection methods (fastText, CLD3)

Workflow

1. Docker Environment setup for Ubuntu
2. Data Preprocessing for Deezer Dataset
3. Applied the NEiCE strategy to Deezer Dataset
4. Computed Evaluations for Deezer Dataset
5. Attempts to achieve same Data Preprocessing for iTunes Dataset

Results



- NEiCE outperforms the best baselines except *n_topics=200.*
- Parameter combinations do not follow the same trend.
- The best parameters per number of topics differ significantly:
  - 1st best (20 topics) -> result 3rd best
  - 1st best (50 topics) -> result 5th best
  - 1st best (100 topics) -> result 6th best
  - 1st best (200 topics) -> result 6th best