Arbeitsgruppe "Social Computing"
Fakultät für Informatik
Technische Universität München

Survey of Topic Modeling Algorithm

Supervisor: Professor Dr. Georg Groh

Advisors: PhD Candidate (M.Sc.) Miriam Anschütz, PhD Candidate (M.Sc.) Ahmed Mosharafa

Short Description

The aim of the project is to evaluate different Topic Modeling algorithms on short/long text dataset. The outcome should include metric-based evaluation, as well as, human based evaluation to the algorithms. This includes drawing observations on the applicability of certain algorithms' clusters to different types of datasets.

Long Description (Milestones)

1. Datasets Exploration

Explore the provided datasets to unveil the inherent characteristics. The outcome of the milestone should be an overview of the statistical characteristics of the datasets. The proposed datasets are CRISIS NLP dataset and 20NewsGroups dataset.

2. Preprocessing steps

Preprocessing steps for the datasets should be applied. A list of pre-processing steps required should be reviewed in advance. A proper distinction should be made in terms of differences from short-text to long-text preprocessing algorithms (when relevant).

3. Applying Un-supervised Topic Modeling algorithms on short-text data.

Different algorithms should be applied on the short-text dataset and evaluated for the common metrics. Algorithms include LDA, NMF, Embedding-based (BERTopic, Topic2VEC). The outcome should include both metric-based evaluations, and human-based conclusions for the differences.

Optional: LSA/LSI, pLSA, BTM, DMM, WNTM.

4. Applying Un-supervised Topic Modeling algorithms on long-text data

Same as (3), but for long-text data.

Optional: LSA/LSI, pLSA, BTM, DMM, WNTM.

5. LDA-Bert implementation (Optional).

Implement the LDA Bert.

6. Apply comparisons points between short/long text with respect to the algorithms tried out.

The outcome of the milestone is to create a comparison between different algorithms with respect to the dataset approached (short/long text). This should include metric based evaluations, as well as human based evaluation on how well the algorithms work on the provided dataset.

7. Suggest the best approach for short/long texts

Following from (6), a proposal for the best set/family of algorithms for short/long text dataset should be provided.

8. Future Work

Future work indicates the directions (according to the research and experiments done) where possible changes could be tackled in the algorithms/pipeline for a better performance with respect to the datasets.

Deliverables

1. Code-base for
   a. Benchmarking different algorithms and datasets in (1) and (2)
   b. LDA Bert implementation code
   c. [Optional] Integrating the algorithm within our code-base of the Automated Journalist.
2. Lab Report.