Project Proposal: PLSA with Prior for MetaPy

Billy Li
zl20@illinois.edu
Department of Computer Science

October 25, 2021

1 Team Members

Billy Li (zl20@illinois.edu)

2 Topic

Theme: System extension
Subtopic: MeTA toolkit

3 Details

This project will aim to extend the PLSA algorithm written in MP3 with the added functions of including both a background model and priors and implement this extension as a new function of the MetaPy toolkit.

This extension will be implemented based on the formulations discussed in the Week 9 videos on PLSA and LDA, and we will perform a test integration using a clone of the MetaPy repository [1] locally. As the MetaPy toolkit is written in Python, this project will be implemented using Python.

Some particular features we plan on including are

- Reading the background model from an online source (e.g. [3]), as the large amount of text data required to construct a background model of reasonable accuracy limits the feasibility of reading files locally. Depending on available online sources, we may also allow the user to retrieve topic models online.
- Saving intermediate models. This facilitates operations such as incremental training, which allows the user to run a few more iterations of the EM algorithm should they be dissatisfied with the current model results.

We will test our project using the 20newsgroups dataset [2] in sklearn, as it is a real-world dataset and is of the appropriate size (11000 documents).

The workload of this project will be estimated to be around 20 hours:

- Extending the PLSA model to include Prior & background model: 6 hours
- Adding model saving/loading functionality: 2 hours
- Reading background/topic models from online sources: 6 hours
- Integrating the model with MetaPy: 6 hours

References

[1] https://github.com/meta-toolkit/metapy

[2] https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

[3] https://www.wordfrequency.info/