Project Progress Report: PLSA with Prior for MetaPy

Billy Li
zl20@illinois.edu
Department of Computer Science

November 16, 2021

1 Completed tasks

- **Extending the PLSA model to include Prior & background model**: This was a pretty straightforward step as the equations are available on lecture slides and notes provided in MP3[4]. The input files are read in a fashion similar to SKlearn models with the training documents, labels for the training documents and testing documents read separately. The training documents and labels are used to compute the prior probabilities for each topic and the background model.

- **Adding model saving/loading functionality**: This functionality was implemented using the Pickles library in Python. The users can save and load computed class variables (i.e. $P(w|z)$) to both resume an uncompleted E-M algorithm or save time by not having to recompute static variables such as prior probabilities and the background model.

2 Pending tasks

- **Integrating the PLSA model with MetaPy**: This step is found to be harder than expected: the MetaPy library is written using C++ instead of Python and translated into Python using the PyBind library. This step will require both rewriting the PLSA model using C++ and becoming familiar enough with the PyBind library to integrate the model into MetaPy.

- **Reading background/topic models from online sources**: Cancelled. The dataset found at [3] is a paid dataset, and all free datasets for English word frequency are either samples of paid datasets or have a limit to the frequency of API calls. Simply using the overall distribution of words in all documents of all topics as proposed in [4] appears to give satisfactory results.

3 Challenges

- Understanding how C++ code is translated into equivalent Python code via the PyBind library.

- Implementing the PLSA model in C++ in a style that is consistent with other MetaPy functions.

4 Estimated progress

- Extending the PLSA model to include Prior & background model: **Done**
- Adding model saving/loading functionality: **Done**
- Reading background/topic models from online sources: **Cancelled**
- Integrating the model with MetaPy: **15 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/

[4] http://times.cs.uiuc.edu/course/598f16/plsa-note.pdf