UIUC CS410 Fall 2021 Project Report Zhou Wu zjwu2

I was able to successfully create a web app that could be used to compare LDA and Top2Vec by inspecting the coherence of topics and the relevance of documents returned by the models. The app can be extended to other topic modeling methods too. I optimized the app for sklearn's 20news dataset by loading pretrained models and, additionally, built a web scraper that can create custom test datasets from Wikipedia based on a user-defined timeframe. The total time spent on this project was approximately 60 hours.

The results of my study suggest that for relatively large datasets such as sklearn's 20news dataset, Top2Vec takes less time to train before achieving a useful model compared to LDA. In addition, it is easier to use because unlike LDA, it works well without requiring the user to figure out how many topics are in the dataset or how many passes/iterations should be run for training to converge. However, for a smaller collection of documents – e.g., less than 200 Wikipedia articles – LDA tends to return more coherent topics. Top2Vec's python package is well-built and contains many useful functionalities, whereas some functions for LDA in the genism package seemed buggy.

There were no significant blockers during the project, though troubleshooting bugs with gensim and the Streamlit web app took up more time than expected. Most notably, gensim's functions such as *dictionary.filter_extremes()*, *lda.get_topic_terms()* did not return expected results – such as *sorted* lists, contrary to its documentation; in addition, some functions' filtering mechanism based on parameters such as *minimum_probability* and *keep_n* also had unexpected behavior that required investigation. On the other hand, Streamlit did not have many bugs, however, due to its design, caching data was not straightforward: whenever a change to user input is detected, the app is re-run, which normally would result in the data processed again and the models retrained even if the user input does not concern the data and models. To reduce wait time for users, I spent a significant amount of time testing Streamlit's caching functionalities, which were still in beta phase.

https://github.com/wujameszj/CourseProject

Appendix: Notable design and implementation decisions

- For sklearn’s 20news dataset, a top2vec model is pre-trained so that it can be loaded right away. This is done to improve user experience because training could take more than 10 minutes depending on the machine’s hardware.
- A ‘default’ LDA model is trained with 2 passes and 500 iterations, which is sufficient to produce a fairly decent model for fair comparison. However, due to git repo size consideration, it was not included in the final version.
- Lemmatization is not performed for the vocabulary. This is because Top2Vec is designed to work well without lemmatization, though it allows the user to provide a lemmatized vocabulary if they really wanted to. Hence, for fairness’s sake, LDA’s vocabulary is also not lemmatized. An option could be added to future releases to allow lemmatization.