UIUC CS410 Project Report Zhou Wu zjwu2

I was able to successfully create a web app that could be used to compare LDA and Top2Vec, as well as potentially other topic modeling methods, by inspecting the coherence of topics and the relevance of documents returned by the models. My study suggests that 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. Top2Vec's python package is also 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 *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 I had to investigate. 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.

Due to the aforementioned setback, regrettably, I ran out of time to implement my plan to build a web scraper to offer custom testing datasets based on user-defined timeframe and themes from sites such as Wikipedia and Reddit. The total time spent on this project was approximately 40 hours.

Dear reviewers:

To my knowledge, I have met all requirements listed in the project instructions on Coursera and Google Doc. If you believe I have missed something, kindly let me know and I will address it immediately. Thank you for taking the time to review my project.