Tasks completed:

- Experimented and became familiar with the top2vec algorithm/library, including its main dependencies UMAP and HDBSCAN (dimension reduction and clustering algorithms)
- Explored the genism python package for LDA
- Developed and deployed Streamlit web app; incorporated top2vec into web app

Tasks pending:

- Develop scripts to scape select websites for text data
- Learn advanced usage of Streamlit and incorporate additional features/algorithms into web app
- Record presentation and write final report

Challenges:

Once I actually incorporated topic modeling into the web app, I realized just how long it takes to run. Depending on the dataset, it could take over 20 minutes, which is probably too long for users/reviewers. One potential solution could be to precompute the models offline and then simply load them for the web app, however, that would not allow the users to select custom date ranges to scrape websites, which was one of my hopes for this web app. Another solution is to limit the size of the dataset, thereby reducing the most time consuming step of the top2vec topic modeling algorithm, which appears to be the embedding step.

The web app (incomplete) is hosted here:

https://share.streamlit.io/wujameszj/courseproject/main/main.py