Name: Faith Chung
NetID: faithc2
Captain: Faith Chung (individual project)
Project: Reproducing “Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback”

For this CS 410 project, I will be reproducing the Causal Topic Mining paper by Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas Rietz, and Daniel Diermeier. This will be an individual project. I am excited to implement the iterative topic model to discover causal topics that are hopefully similar to those found in the experiment.

For obtaining the datasets and implementing the model, the programming will primarily be done in Python. The 2 datasets referenced in the paper are the NYT articles from 5/2000-10/2000, as well as the IEM 2000 Presidential Winner Takes All Market time series data.

For the NYT articles, the link referenced in the paper for the NYT Annotated Corpus requires an account to download the data, which requires a $300 fee for non members. Assuming I cannot get access to this data via UIUC, I will go directly to the NYT archives and use a python web scraper to collect the text data, only selecting paragraphs that contain “Bush” and/or “Gore”.

Similarly, the IEM time series data will be retrieved, with a web scraper, from IEM’s monthly data sets for 5/2000 - 10/2000, which can be found here: https://iemweb.biz.uiowa.edu/pricehistory/pricehistory_SelectContract.cfm?market_ID=29.