Progress Report

CS 410 Text Information Systems - Fall 2020

Team: The Electric Moccasins

James Coffey, NetID: jamesfc2 – Captain and Praveen Bhushan, NetID: bhushan6

Goal

The function of the tool being implemented is to "combine probabilistic topic modeling with time series causal analysis to uncover topics that are both coherent semantically and correlated with time series data." (Kim et al., DOI=10.1145/2505515.2505612) This tool is being implemented in Python instead of the R implementation done in the paper to make it easier for software deployment. The techniques/algorithms being used are PLSA topic model and Granger testing.

Progress made thus far

- Implementation in Python Jupyter Notebook
  - Acquired needed datasets: New York Times Annotated Corpus (NYTAC), IEM 2000 U.S. Presidential Election ticker, and stock tickers for AAPL and AAMRQ
  - Wrote script for determining significant Granger causality at different lag values
  - Wrote function for calculating impact value using Granger causality
  - Wrote function for calculating topic purity
  - Wrote script to trim and organize NYTAC to data subset needed
  - Imported PLSA class from MP 3 and modified the build corpus to use NLTK and multiprocessing.
  - Wrote function for topic level causality.

Remaining tasks

- Compare results of Python implementation with reference paper.
- Write command line interface for tool.
- Write documentation and record tutorial presentation.

Challenges/issues being faced

- Even with multiprocessing, building corpus takes a while on 4 core Intel i7 CPU
- EM algorithm takes a while. Could this be sped up by using CuPy instead of NumPy so that it uses the GPU?
- Not sure if it is one iteration of EM for each update of the topic prior or if it is multiple EM iterations for one topic prior update.
- Not sure how to incorporate the topic prior into the PLSA algorithm.