

Pei Lun Hung
CS 410
Professor ChengXiang Zhai
October 25th, 2020

Project Proposal

Project Topic: Reproducing a Paper on Causal Topic Modeling

1. Team

Name: Pei Lun Hung (project coordinator/leader)

NetID: prhung2

Email address: prhung2@illinois.edu

2. Paper

I plan to reproduce the results from Hyun Duk Kim et al.'s paper, "Mining causal topics in text data: Iterative topic modeling with time series feedback."¹

3. Implementation Language

I plan to implement the project in Python.

4. Data Set

The paper uses *The New York Times* Annotated Corpus. As such, I have registered for an account with the Linguistic Data Consortium and am awaiting acceptance from admin.

5 Backup Data Set

There is a digital archive² of *The New York Times*, with articles from the 1850s to the present day. As backup, I can scrape articles from January 1st, 1987 to June 19th, 2007 from this digital archive in order to obtain a similar data set to the aforementioned corpus, sans annotations.

¹ Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas Rietz, and Daniel Diermeier. 2013. Mining causal topics in text data: Iterative topic modeling with time series feedback. In Proceedings of the 22nd ACM international conference on information & knowledge management (CIKM 2013). ACM, New York, NY, USA, 885-890. DOI=10.1145/2505515.2505612

² <https://spiderbites.nytimes.com/>