Final Project Progress Report

Reproducing a paper

*Mining causal topics in text data: Iterative topic modeling with time series feedback. (Zhai et al 2013)*

*Progress made thus far*

* *Read the paper (above)*
* *Obtained access to the New York Times Annotated Corpus from the Linguistic Data Consortium*
* *Downloaded the NYT annotated corpus*
* *Read through the overview document for the NYT corpus*

*Remaining Tasks*

* Loop through XML data files and create functions for text parser
* Parse XML data for the framework inputs: a time series dataset (list of timestamps) and a corpus consisting of a list of (Document,timestamp) tuples
* Select a topic modeling method (M) and implement using a standard library
* Implement the iterative topic modeling framework with time series feedback
* Train against the NYT dataset
* Implement the experiment in the paper and produce the same set of sample results

*Challenges*

* Corpus size is significant (3.06 GB), but it should fit in memory on my laptop (which has 32 GB RAM).
* This will be a significant undertaking to perform myself, but I’m taking the final early to give myself more time to get this done.