Michael McClanahan

NetID: mjm31

CS410: Text Information Systems

Project Proposal

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

Captain: Michael McClanahan (NetID: mjm31)

This project will be completed individually.

1. Which paper have you chosen?

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

1. Which programming language do you plan to use?

Python (3.6.1 via Anaconda 4.4.0)

1. Can you obtain the datasets used in the paper for evaluation?

No. The dataset from the LDC requires a license for access:

<https://catalog.ldc.upenn.edu/LDC2008T19>

1. If you answer “no” to Question 4, can you obtain a similar dataset (e.g. a more recent version of the same dataset, or another dataset that is similar in nature)?

I could build a scraping engine to capture all of the text from the same articles here:

<https://spiderbites.nytimes.com/2000/>

Ideally though, I would already have access to the data in XML file format from the LDC.

1. If you answer “no” to Questions 4 & 5, how are you going to demonstrate that you have successfully reproduced the method introduced in the paper?

N/A