Project Progress Report

Table of Contents

Project Progress Report 1

Project Topic: 2

Iterative Topic Modeling Framework with Time Series Feedback: On a high level, the paper implemented the below framework: 2

Brief Description of How Tasks were divided to Implement the Paper: 2

Project Progress: 3

Progress made thus far – *3*

Remaining tasks – *3*

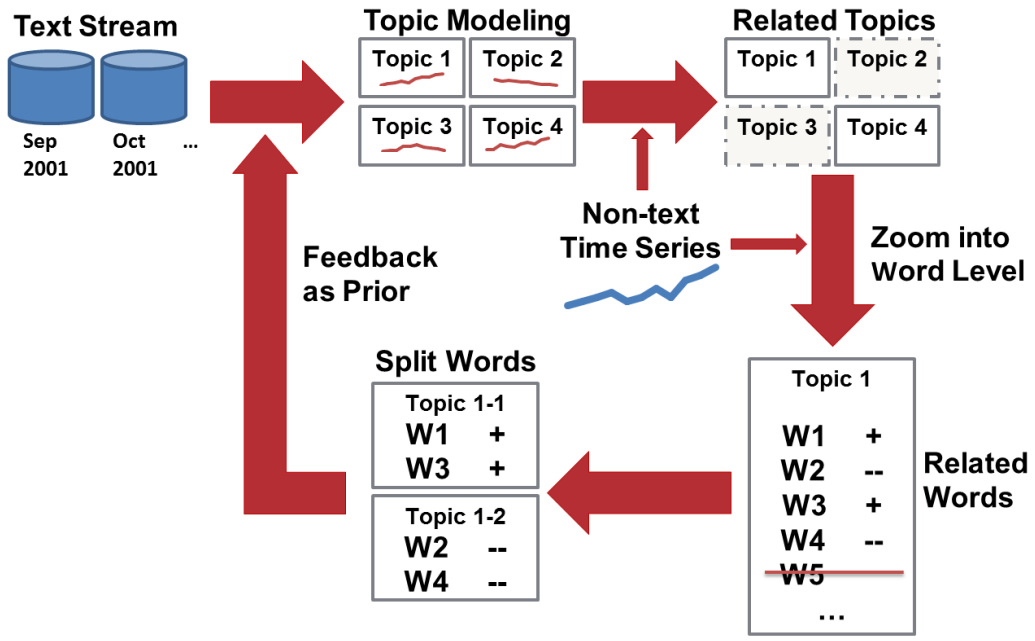
Project Topic:

Reproducing a paper - Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback. Mining causal topics in text data: Iterative topic modeling with time series feedback. ACM, New York, NY, USA, 885-890. DOI=10.1145/2505515.2505612

Team:

- Heta Desai (Group Leader): Net-id: hetahd2
- Harpreet Siddhu: Net-Id: hsiddhu2

Iterative Topic Modeling Framework with Time Series Feedback: On a high level, the paper implemented the below framework:



Brief Description of How Tasks were divided to Implement the Paper:

1. **Presidential Election Dataset Pre-processing –** This step involves selection of articles from New York Times dataset which has the collection of all the articles published in NYT from 1997 to 2007. For the implementation of this paper, select all the news articles of year 2000 from May to October where keyword “Bush” or “Gore” or both are mentioned in the news articles. This will be the Presidential election dataset which will be analyzed later with another time series dataset.

2. **Time-Series Dataset Pre-processing:** This step involves creating a time series data for Presidential Prediction Market prices from May 1st 2000 to Oct 30th 2000. b - https://iemweb.biz.uiowa.edu/pricehistory/pricehistory_SelectContract.cfm?market_ID=29)

3. **Topic Modeling:** Use presidential election dataset and run any Topic modeling algorithm.

4. **Related Topics (Topic Level Causality Analysis):** Use the topics form step:3 and the non-text time series dataset from step: 2 to determine the related topics. This step will use "Granger Test" for causality measure. This step will give the topics with highest significance.

5. **Related Words (Word Level Causality Analysis):** Use the most significance topics (high causality score) from step-4 and analyze the words of each topic. The first step in analyzing word level causality is to create a word count time series for each day and then use it to run the causality test to determine the most casual words.

6. **Feedback (Prior Generation):** From step 4 and 5, we get the casual topics and words. Generate a topic prior using the most significant words from the significant topics and feed that to the next topic modeling process. Before feeding, improve the topic quality by dividing the topic into positive and negative words. (if one of the group is smaller which is < 0.1*of other group, just keep one group topic).

7. **Iterative Model & Feedback:** Repeat the iterative process and remodel topics until no more significant topic changes or reached a pre-defined topic quality.

Project Progress:

This report is based on the following questions that we tried to answer to give the status of our progress by the end of 11/30/2020. If you need more information, please contact team leader.

Progress made thus far –

We have completed the following things from the above list –

1. Pre- processing of New York Times dataset and prepared the presidential election dataset using the data of year 2000 from 1st May 2000 to 30th October 2000.
2. Pre- processing of presidential prediction market prices and prepared a time series from 1st May 2000 to 30th October 2000 of stock prices.
3. Completed topic modeling using Genism's API for Latent Dirichlet Allocation. See Link: https://radimrehurek.com/gensim/models/ldamodel.html)
4. Competed mining related topics (the most casual high significant topics). This is accomplished by running the causality measure test using "**grangercausalitytests**" from statsmodels.tsa.stattools used for time series analysis.
5. Completed creating word count time series.

Remaining tasks –

We still need to work on word level casual analysis, prior generation and feedback and document results.

Any challenges/issues**:**

Few things in the paper are not clear. We attended TA sessions and also talking to other team members to get some our answers.