Team EastToWest

Akshaya Jagannadharao (akshaya2) | Heidi Touissant (heidist2) | Hari Venkitaraman (hv4)

1) An overview of the function of the code (i.e., what it does and what it can be used for).

We are reproducing a paper in this project: Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback. We are only reproducing the first experiment where the authors examine the 2000 U.S. Presidential Election.

The code we implement here can be used to find causal relationships between two different datasets. To be more specific, we are trying to find a causal relationship between new articles and stock market prices.

2) Documentation of how the software is implemented with sufficient detail so that others can have a basic understanding of your code for future extension or any further improvement.

Data:

New York Times: There are two methods to collect articles published by New York Times from May-Oct 2000. One is to gain access to the LDC corpus. You should parse the documents and only collect relevant articles (i.e. keep paragraphs that contain “Bush” or “Gore). If the corpus is unavailable to you, it is still possible to collect the data by scraping the New York Times website. Details on how to do so can be found in the README file in the Data folder.

Iowa Electronic Market (IEM) Time Series:

The procedure to collect the IEM data is detailed in the README file in the Data folder.

Code:

We are running everything inside a Jupyter Notebook (with Python version 3.7.5) so that we can easily create and display visualizations. The code can be broken down into 3 sections (we have taken the initiative to section it off into three cells). The first cell contains all the import statements. We have detailed the usage of software in section 3 of this document for further perusal. The second cell loads the New York Times dataset and the Iowa Electronic Market Time Series Data into data tables and creates BOW (bag of words) representations by article and date. You should only have to run this cell once unless you want to change the BOW representation. The third cell contains the code to run the topic modeling with time series feedback. There is an example use case documented in the README file as well as an example all in the jupyter notebooks.

Challenges + Improvements + Differences between our results and theirs:

- Delay in obtaining dataset from NYT corpus
  - We were unsure if we were able to obtain the dataset because of copyright permissions. Due to this, we were unable to start our project until 1 month before the deadline (and one week was technically break).
- No efficient way to write code as a group remotely
  - We struggled to find a collaborative Jupyter Notebook environment. Jupyter Notebook is an important tool for our use to easily visualize the different outputs and

data structures we were generating throughout the code (i.e. visualization for LDA model, visualization of results).

- Missing dates in Time Series
    - There are multiple ways to handle missing dates in time-series data. Namely, take the last price, take the logistic regression, take an average, drop the missing date.
        - We took the easiest route and simply used the previous day's prices for the missing dates as there were only two (6/7 & 6/8)
- Calculating mu remained unclear after referring to the referenced paper, attending multiple office hours, and consulting with other classmates.
    - We decided to go with a similar parameter already integrated with Gensim's LDA model, decay.
- Some material was outside the scope of the course. We needed to understand how Granger Causality worked and the constraints of a time series feedback (namely the data should be stationary).
- A typographical error in the paper was discovered while attempting to reproduce Table 1 (p. 3, table 1)
    - This caused confusion amongst the group and we spent some time trying to understand it. After attending office hours, the professor, a co-author of the paper, ultimately decided this was a typographical error
- Our results do not replicate the results achieved in the paper. There is a clear difference in the progression of confidence and purity levels through multiple iterations.

3) Documentation of the usage of the software including either documentation of usages of APIs or detailed instructions on how to install and run a software, whichever is applicable.

Additional Instructions can be found in the README files in the GitHub repository. You can find a short demo on Mediaspace on how to set up the code: CS410_TeamEastToWest_FinalProject - Illinois Media Space.

To view a run of our code with the graphs generated at the end, visit CS410_TeamEastToWest_FinalProject.

If you would like to see the code in action, please reach out to the contributors (Akshaya, Heidi, or Hari) for a demo/tutorial. Due to copyright restrictions on the data, it is not possible to run this code yourself.

We use the following package and import statements:

```
pip install pyLDAvis
import re
import numpy as np
import numpy.linalg as la
import pandas as pd
from pprint import pprint
import datetime

# NLTK
```

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk import ngrams

# Gensim
import gensim
from gensim import models
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
from gensim.models import CoherenceModel
from gensim.models import Phrases # TODO: to create bigrams with

# spacy for lemmatization
import spacy

# Plotting tools
import pyLDAvis
import pyLDAvis.gensim  # don't skip this
import matplotlib.pyplot as plt

import statsmodels
from statsmodels.tsa.stattools import grangercausalitytests

import warnings
warnings.filterwarnings('ignore')
```

References:

Mei, Q, Ling, X, Wondra, M, Su H., and Zhai, C. Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of the 16th international conference on World Wide Web, pages 171-180, New York, NY, USA, 2007. ACM.

Mupfururirwa, W. (2019, September 19). Retrieved from https://stackoverflow.com/questions/58005681/is-it-possible-to-run-a-vector-autoregression-analysis-on-a-large-gdp-data-with

Prabhakaran, S. Time Series Analysis in Python, Retrieved from https://www.machinelearningplus.com/time-series/time-series-analysis-python

Sarit, M. (2019, October 7) Time Series Forecasting using Granger's Causality and Vector Auto-regressive Model, Retrieved from https://towardsdatascience.com/granger-causality-and-vector-auto-regressive-model-for-time-series-forecasting-3226a64889a6

4) Brief description of the contribution of each team member in case of a multi-person team.

We initially tried to code in real-time as a group across various notebook environments but could not find an efficient solution that would have allowed us to access the data at the same time (we could have potentially hosted the data on the cloud but that would have required additional charges). We ultimately decided that Akshaya would share her screen via Zoom video call while she coded and Hari and Heidi worked through understanding the steps of the paper together as well as communicated with TAs and Prof during office hours. Essentially pair programming with one person as driver and two people as navigators. During the week, Heidi and Hari would go through the paper and figure out the next steps. Akshaya would code during the week. During the meeting, we would review what Akshaya did and find any gaps in logic or deviations from the paper/catch for mistakes. After reviewing the code, Hari and Heidi would go through the next steps that Akshaya would code up later in the week. We split up the work like this because the LDA model would take a significant time to run during the meeting. If there was time remaining in our meeting, we would take some time to code as a team. If there was a report due, we would complete it as a team during the meeting. To make the best use of our time since the algorithm takes a significant amount of time to run, we decided on this strategy. When issues came up that we could not agree on a solution as a team, Hari and Heidi would take the time during the week to talk with the TAs during office hours. A good example of this situation in our group would be that Hari found a typo in the paper that caused a lot of confusion. Heidi and Hari went to office hours to clarify with the TA. And when there was no resolution, Heidi and Akshaya spoke with the professor during office hours and finally received an answer.