Team Buddie Progress Report

Topic: Reproducing a Causal Topic Mining Paper

Team Members:

- Aman Gupta (@amang5)
- Katie Shin (@ks56)
- Venkata Sandeep Chillara (@vsc5)

Please upload your progress report to the Github repo shared on CMT. The progress report should give us an idea of how you're implementing your proposal. It should answer 3 main questions:

1) Which tasks have been completed?

- Implemented the initial PLSA algorithms using https://github.com/yedivanseven/PLSA. Modified the source code to include the prior. (Our initial implementation for PSLA can be found in main.py)
- Ingested and parsed stock data for the 2000 presidential election using daily markets in a Jupyter Notebook. Calculated normalized price for each candidate on a daily basis and stored this data in a data frame. This can be seen in (Stock Data Parsing & Normalization.ipynb with stock data in the stock_data folder)
- Explored using https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.grangercausalitytests.html for running granger tests to find casual topics based on a time series

2) Which tasks are pending?

- Organize the PLSA results in a Date-Topic-Probability format so that it can be used in the Granger test
  Run granger tests for topics using results from first task with time series PLSA output
- Understand and run pearson coefficients tests to find casual words from relevant documents
- Use output from pearson coefficient tests as prior for PLSA.
- Add README.md with steps to organize NYT corpus data in and run the code

3) Are you facing any challenges?

- PLSA process time for the NYT corpus times out when run on a local machine. One option is to look into https://colab.research.google.com/
  - Since we need the Date-Topic-Probability, we don't have to run PLSA on the entire dataset
- The version of PLSA with priors isn't readily available, we had to implement that on our own.