Homework 3

Spring 2022 MTH 9796

A common application of NLP in finance is to assess how people feel about particular entities related to tradable assets -- for example companies and their stocks, or countries and their currencies. How investors feel about an entity often corresponds to their willingness to buy or sell assets related to that entity. We explore that idea in this homework by testing two hypotheses:

- **Hypothesis-1:** The frequency of mentions of a corporate entity is correlated with trading activity of that corporation's stock.

- **Hypothesis-2:** The sentiment of the discussion around a corporate entity is correlated with the direction of movement in the related corporation's stock.

Assignment

Create a Colab Notebook that does the following:

1. **Retrieve at least two social-media/news headlines datasets.**
   To get you started, you can find a NY-Times headlines dataset and a WallStreetBets submissions dataset shared in Google Drive. See the Materials section below for details.

2. **[10 points] Test-Hypothesis-1 on both datasets --**
   a. **[2 points] Produce a time-series of (publication-time, entity-name)-tuples:**
      i. **Choose a Named Entity Recognition (NER) tool that is well suited for the specific dataset you are analyzing.** You may want to try several and/or adjust the parameterization of the NER tool to see which performs the best.
      ii. **Explain which technique the NER tool uses** (e.g. word-list, HMM, CRF, CRF-LSTM hybrid, Transformer, etc.)
      iii. **Explain why you found this to be a good tool-choice for each dataset.**

   b. **[3 points] Choose 5 corporations that your NER was able to identify and measure the following time series:**
      i. *Mentions per day* -- as measured by your NER.
      ii. Daily high minus low (as an approximation for volatility.) -- from market data.
      iii. Daily trading volume -- from market data.
   c. **[5 points] Does the data support hypothesis-1?** Analyze the relationship between the mentions-per-day and the volatility and trading-volume measures. Use appropriate hypothesis testing techniques to decide.

3. **[10 points] Test-Hypothesis-2 on both datasets --**
    a. **[2 points] Produce a time-series of (publication-time, entity-name, sentiment-index)-tuples by...**
        i. **Choosing a Sentiment Analysis (SA) tool that is well suited for the specific dataset you are analyzing.** You may want to try several and/or adjust the parameterization of the SA tool to see which performs the best.
        ii. **Explain which technique that SA tool uses** (e.g. word-list, Vader, Naive-Bayes, Transformer, Zero-shot learner, etc.)
        iii. **Explain why you found this to be a good tool-choice for each dataset.**

    b. **[3 points] For 5 corporations on which your NER and SA performed well, identify and measure the following time series:**
        i. Positive Mentions per day -- as measured by your NLP tools.
        ii. Negative Mentions per day -- as measured by your NLP tools.
        iii. Daily stock return -- taken from market data.

    c. **[5 points] Does the data support hypothesis-2?** Analyze the relationship between the positive/negative mention rates and the direction of the stock returns. Use appropriate hypothesis testing techniques to decide.

4. **[5 points]** At TA's discretion for extra attention to delivering a quality homework:
    a. **Scientific Method.** You test a hypothesis with a well designed repeatable experiment and draw a conclusion based on the data. *If the data doesn't support the hypothesis, say so. A negative result is still a valuable result.*
    b. **Working readable code.** Restart Kernel and Run-All should complete and produce the desired result. Document your code. Use MarkDown to organize and document your notebook.
    c. **Tool choice with explanations.** You have given thought to which NER and SA tools you used for different datasets and explain the reasons for your choices.
    d. **Creative Problem Solving.** If you are not getting conclusive results, you look at other datasets, other models, other training techniques, other measures of market activity, other measures of correlation, etc.

Expectations

- **You'll complete this homework in your homework group.**
- **Feel free to discuss any aspect of this on the MFE Forum** -- especially if you're getting stuck. We will be checking the forum periodically and helping where we can.

Dates

Assigned 2022-04-06
Due 2022-04-27.

Materials

You should be able to access the Homework-3 Google-Drive Folder in which you should find:

- **HW3-Datasets.ipynb** -- A Colab notebook demonstrating how to access the New York Times headlines dataset, the WallStreetBets submissions dataset, and stock-prices and volumes.
- **nyt-20200401_20210401** -- Collection of files comprising the New York Times dataset.
- **wsb-20200401_20210401** -- Collection of files comprising the WallStreetBets dataset.

Grading

This assignment is worth 25 points.

Handing In Your Homework

You will hand in a Colab Notebook with the following sections:

- **Authors** --Include the names of everyone who participated.
- **Abstract** -- A brief paragraph summarizing the hypotheses, experiments, and your conclusion.
- **Methods and Data** -- An explanation of the datasets, tools, and methods you used to test the hypothesis. This is where you also explain why you chose the tools you did.
- **Implementation** -- Runnable Python code cells where you apply the NLP techniques to data and perform the analyses.
- **Results** -- A summary of the results and analysis.
- **Conclusions** -- An explanation of whether or not the hypotheses were supported by the experiment.

To hand it in:

1. Name your homework as follows: **MTH9796-HW3-Group-X.ipynb (where Group-X is your homework group.)**
2. Share your homework with: **baruch.mfe.mth.9796@gmail.com**