Andrew Plum

Professor Ma

CS 479

1/22/2024

Data Collection Proposal

1. Data Collection

The data I will collect and observe will be the historical daily stock price data of the S&P 500 recorded by the Federal Reserve (link: https://fred.stlouisfed.org/series/SP500) and speeches made by current Federal Reserve Chair Jerome Powell in written text format also provided by the Federal Reserve (link: https://www.federalreserve.gov/newsevents/speeches.htm). I am choosing S&P 500 data over other stock market index data because it is a popular stock market index, and it is weighted by the market capitalization of the companies in it. The goal of the data collection will be to explore if there is any correlation between the specific words said in each speech and how the S&P 500 performed on the same day the speech was given. I would also like to try and create a model that can determine the overall sentiment of a speech which is then used to predict if the S&P 500 was up or down that day and potentially by how much. The speeches made by the Fed Chair seem to affect the stock market on the day that they are given and knowing how and to what degree could help to predict its performance in the future.

1.1 Creation of Logical Collections:

The dataset I will create will be a logical collection of S&P 500 daily close data and the speeches made by the Fed Chair. Both will be organized by date.

1.2 Physical Data Handling:

I will keep copies of the original raw datasets that are used to create the combined formatted dataset. Each of the speeches will be in their own text file but their file name will be listed in a .csv file with an associated date. I will record when the datasets are downloaded. The S&P 500 dataset will be downloaded in a .csv file and the words of each speech made by the Fed Chair will be copied and pasted into its own .txt file.

1.3 Interoperability Support:

I will store the S&P500 daily close data and the file names of the .txt files that contain the speeches in the same .csv file. The .txt files that contain the speech text will be in the same directory, so that the path of files is the same when accessing the data files in python. Using .csv and .txt files supports interoperability because most programming languages and software applications can read and write to both file formats.

1.4 Security Support:

Since the raw datasets are publicly available and the project when completed will be publicly available, very few security measures will be taken to safeguard the data. Because of license concerns, I won't be making the combined dataset I created public.

1.5 Data Ownership:

The Federal Reserve and Standard & Poor's are responsible for the quality of the data, and I will be responsible for interpreting the meaning.

1.6 Metadata Collection, Management, and Access:

There is original metadata for the two raw datasets. The metadata for the combined dataset will be created by me, and it will include metadata from the raw datasets and new metadata recorded by me such as time downloaded.

1.7 Persistence:

My project findings will be uploaded to a GitHub repository. The data will be stored in .csv and .txt files because these file formats have been widely supported for a long time. I will also store a private copy of the dataset on another computer of mine.

1.8 Discovery:

People can find my project's findings and the corresponding code of the project in a GitHub repository I will be creating where they can get access to it on request. Because of the license of the S&P 500 data, I currently do not have permission to share the combined dataset I created. On the S&P 500 dataset download page, it says I can email index_services@spdji.com to get permission to share the dataset. Because of this, I do not plan to share the data and allow people to contribute to the project.

1.9 Data Dissemination and Publication:

I want anyone who is interested and wants to access the project to be able to access it which is why the project's findings will be uploaded to GitHub and if I'm notified through the contact information I have listed in the GitHub repository, I will request for permission to share the project dataset with those who ask. If my results are well-founded and interesting, I will post my findings on a social media forum.

2. Survey of Data Storage / Formats

My data collection will be stored in a .csv file and .txt files. I will store the S&P 500 daily close data and the file names of the .txt files that contain the speeches in the same .csv file. The .csv is a good format because it allows for easy import into tools like pandas in Python or Excel for initial exploratory data analysis. Likewise, the .txt format is also a versatile format because it can be opened in software like Microsoft word. The .csv and .txt files that contain the speech texts will be in the same directory, so that the paths of files are similar when accessing the data files in python. In the .csv file the data will have a "Date" column, a "S&P 500 Daily Close" column, and "Speech File Name" column and the rows of the dataset will be ordered by the date.

3. Survey of Metadata Conventions

The metadata of the combined dataset will be produced entirely by me; however, I will grab available metadata from the S&P 500 dataset and the Fed Chair speeches dataset. The Dublin Core metadata standard should be an adequate metadata standard used for my dataset. From the Dublin Core metadata standard, there are several metadata elements I can describe. For both the stock data and the speech data, I can describe the title, creator, description, date last updated, date downloaded, file format, source, language, rights, etc. The metadata will clarify the context of the dataset used for my project.

Citation

S&P Dow Jones Indices LLC, S&P 500 [SP500], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, January 22, 2024.

https://www.federalreserve.gov/newsevents/speeches.htm