Andrew Plum

Professor Ma

CS 479

2/7/2024

Data Collection Reflection

1. Data Management Practical Experience
   1. Goal and Modes of Collection

The goal of my data collection was to explore the correlation between the words in the speeches made by Federal Reserve Chair Jerome Powell and the performance of the S&P 500 on the same day the speech was given. The mode of collection consisted of obtaining initial datasets available online. Once collected, the datasets were combined into a joint dataset that I created.

* 1. Data Acquirement, Conversion, Problems Encountered, and Lessons Learned

The S&P 500 data was downloaded from the website fred.stlouisfed.org in .csv file format; downloading this was quick and straightforward. The transcripts of the speeches made by Federal Reserve Chair Jerome Powell were manually collected from the website federalreserve.gov by clicking on the relevant web pages where the speech transcripts were contained and copying and pasting the text into its own .txt file for each speech made. I did this for each of the 58 speeches that Jerome Powell gave ever since he was appointed as Chair of Federal Reserve at the beginning of 2018. This was easy although it was tedious and took a while. If this was a larger project where more data needed to be compiled, I probably would have tried to automate this process by coding a program.

* 1. Physical and Logical Organization of Data and Metadata, Successes, and Failures

After attaining the initial data, I created a joint dataset which consisted of all the .txt speech files and a joint .csv file. In the joint .csv file, I copied the S&P 500 data into it which consisted of the “DATE” column which contained all the dates the S&P 500 was open for trading and a column called “SP500” which contained all of the daily close index values which corresponded to its respective date in the “DATE” column. The next column I created, was called “CHAIR\_POWELL\_FED\_SPEECH\_NAMES” which consists of data on the names of .txt files I created which have the speech transcript data in them; the data in this column corresponds to its respective date in the “DATE” column. One thing to note is there are a lot of null values in this column because there were only 58 speeches given over the span of six years. One other problem I ran into in creating this joint dataset was the raw dataset of the speech transcripts collected didn’t have a lot of initial metadata elements I was looking for, so I had to create them myself. A lot of the metadata for the S&P 500 was easily found on the website the data was hosted on.

* 1. Metadata and Documentation

The metadata for the joint dataset I created was collected if already existing in the raw datasets and created if missing from those datasets. The metadata elements encompassed key details from the Dublin Core metadata standard such as the dataset title ("SP 500 Data and Federal Reserve Chair Powell Speech Transcripts"), creator (Andrew Plum), description, date last updated (February 7, 2024), date downloaded (January 29, 2024), file format (.txt and .csv), publisher (federalreserve.gov and fred.stlouisfed.org), source (S&P Dow Jones Indices LLC and Federal Reserve), language (English), rights information, the URLs to where the raw datasets can be found, and any relevant citations. In terms of the data provenance identified by the metadata, the original sources of the data as well as the websites the data was published on were listed as metadata elements; the data lineage also indicates that I interacted with the data by having me listed as the creator of the dataset as a metadata element. The metadata collection was compiled and created with the goal of making the dataset I created more identifiable.

* 1. Access to the Created Dataset

A copy of the dataset will be included in the submission as a .zip file. Here are some screenshots of the data in the dataset:

A screenshot of a table

Description automatically generatedA screenshot of a computer

Description automatically generated

1. Data Management Plan Experience
   1. What Worked and What Didn’t from My Plan

The creation of logical collections, physical data handling, interoperability, and security of the data was easy and went according to my plan. It did not take a lot of effort to organize the data in the way I said I was going to organize. Even though it maybe took more time than I was expecting to complete, the physical data handling was also simple as it consisted of downloading and copying and pasting the relevant data from the host website. The storing of the data in .csv and .txt files also ensured interoperability. Security measures were also minimal due to the public nature of raw datasets. The only part that wasn’t as straightforward was the creation of a metadata collection to describe the dataset I created because there was quite a lot of metadata elements missing with the Fed Chair speeches dataset.

* 1. What Would I Do Differently Next Time

If I was to do this again, I think I would explore ways to automate the speech data extraction though methods such as web scraping. However, this technique might be more useful for a large project. Another thing I could have done was spending more time looking for alternative data sources. There are probably datasets available online that are very similar to the one I created; the potential tradeoff is the data I might collect by doing this might be less useful as it might be less credible than the data I have now because the data I have now came directly from the sources of the data and not some intermediary. If I was to pursue this option, there are many open data websites that I could go to in search for a dataset like this. I could also ask the data community on public forums if they know of any datasets similar to the one I’m looking for.

* 1. Effects on the Data and Metadata Collected

The data collected should be unaffected and of very high quality since there was very little physical data handling involved. As for the metadata collection of the dataset, since a lot of the metadata elements were missing for one of the raw datasets, I had to create much of the metadata myself, and although I did the best I could to identify the dataset I was creating, it might not be as optimal as it could be in doing so. The metadata corresponding to the other raw dataset should be high quality since it was created by the Federal Reserve.

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