Stocks and Social Data

CMDA 3654
Jake Martinez

Q & A:

1. Are there any correlations between specific stocks? (Tech companies in particular)
    - All tech companies that were monitored exhibited similar behavior, having generally corresponding rises and falls in price. Refer to figure 5 at the end of this document.
2. Is there any correlation between the price of a stock and the amount of interest on a search engine?
    - Occasionally there were spikes in search interest and these spikes corresponded mostly to a sharp increase in stock price. However, this correlation was less conclusive than I had originally hoped. Refer to the red lines in figures 1-4 at the end of this document.
3. Is there any correlation between the price of a stock and the traffic of tweets mentioning the particular stock?
    - Similarly to the search interest, there were spikes in the number of tweets mentioning a particular stock and often correlated to a rise in prices. There were occasionally tweets about a stock without any noticeable rise or fall. This correlation was moderately stronger than the search trend, however. Refer to green lines in figures 1-4 at the end of this document.

Original Data

Data gathered for historical stock data was scraped from Yahoo Finance, while trends in interest were downloaded in csv format (the amount of cleaning necessary to scrape the trend data from the webpage was unrealistic for the time constraints I had, so the static csv's were downloaded instead). I initially wanted to collect twitter data as well, but I learned that the API is limited to 6-9 days prior and to find a real correlation much more data would be necessary (about 60 days is what I was going for).

Included is code that creates a series of graphs comparing the percent change in stock price to google search trends for four different tech stocks.

This preliminary analysis showed that there is occasionally a correlation between the spikes in interest and a spike in the stock price. Usually the spike in interest corresponds to a strong positive increase in the stock price. Also, I found that most of the trend data had some sinusoidal noise so when graphing I normalized and then scaled the trend data to more appropriately display any correlation that could've been present. In order to decrease the "noise" of the data, I squared the normalized data before scaling in order to isolate the peaks of the trend data so that a correlation could be more easily distinguished. Some of the trend data did not

provide very much information as it varied constantly and only within a small window of change. This preliminary analysis led me to try to find a work-around for gathering the tweet data.

New Data

In this portion of the project, I gathered the twitter data I had initially wanted (or at least as close as I could get). I gathered data between January 26th and now, up to 9000 tweets per user account (plus or minus 200). The data was gathered from a sample size of about 10 user accounts that are all reliable stock traders. Once I had gathered this data, I removed extra data so that the dates of the tweets matched the range of the previous data I had collected.

Data Cleaning and Organizing

The tweet data was gathered and saved according to the user account it came from. A set of punctuations was removed from the data in order to more easily find keywords. Every collection of data was then separated into 65 documents (to approximate the days that had gone by) with a variable number of tweets per document per user, because some users tweeted more frequently than others. For example, some users had close to the full 9000 limit while others had only about 1000.

Data Analysis

Once the documents had been created, the keywords for each stock (e.g. “apple” and “aapl”) were search for throughout the set of documents. A scoring matrix was generated with the total frequency of all keywords in each individual document. This scoring matrix was then normalized and scaled to be comparable to the stock price data, then added to the previous graphs from the last portion of the project.

To Run the Code

All analyses between different stocks (for Question 1) were analyzed with the python file “historicaltrends.py”. All analyses between tweet/search data and stocks (for Questions 2 & 3) were analyzed with the python file “socialtrends.py”. The “tweets.py” file is used by “socialtrends.py” and was simply created to reduce the clutter of “socialtrends.py”. The “gather_tweets.py” file was used to gather the twitter data; however, however all necessary data has been collected and deposited in the “data” folder.

Figures

Figure 1

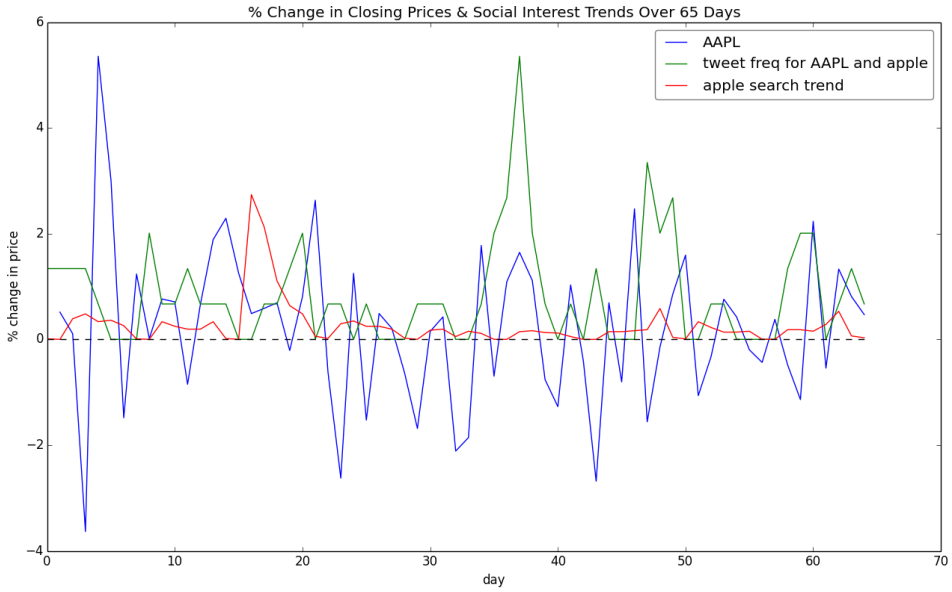


Figure 2

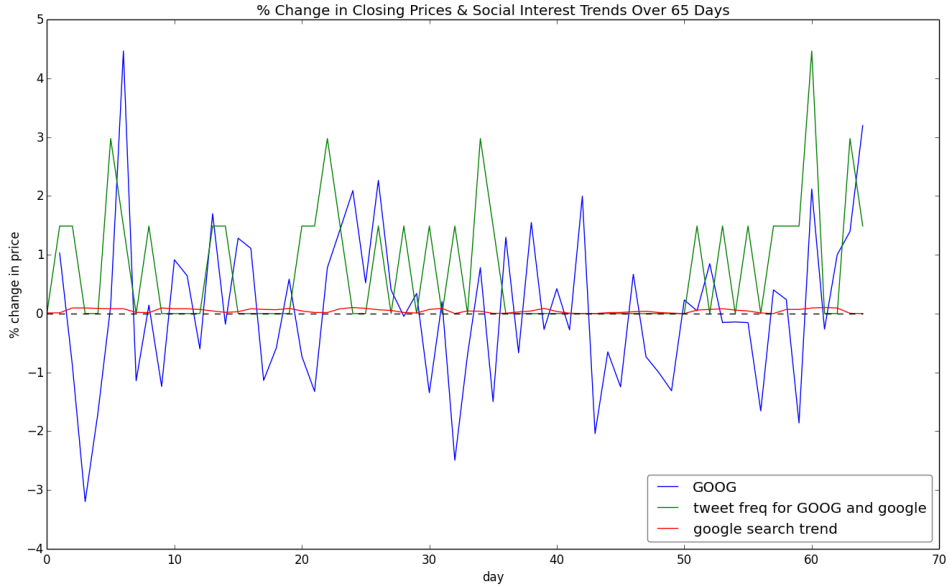


Figure 3

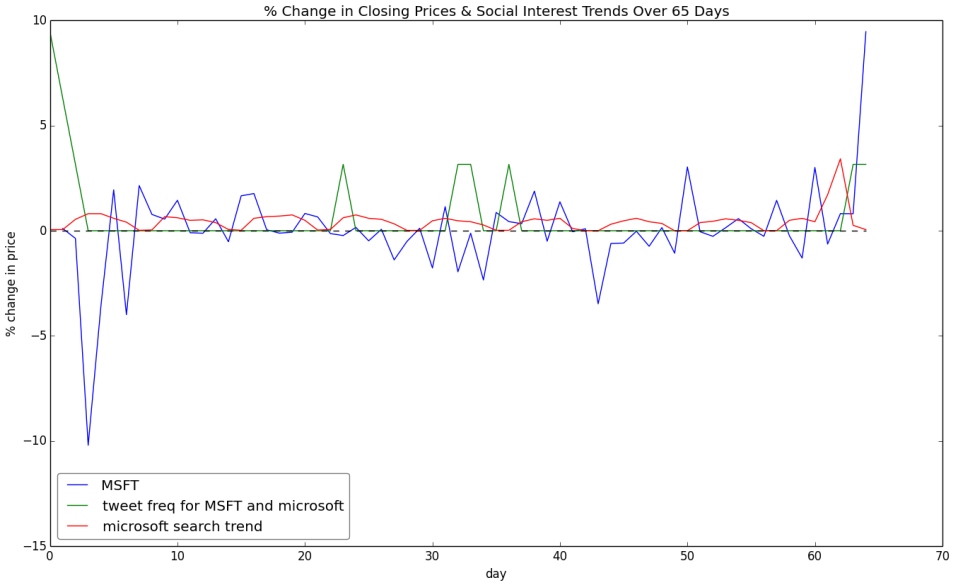


Figure 4

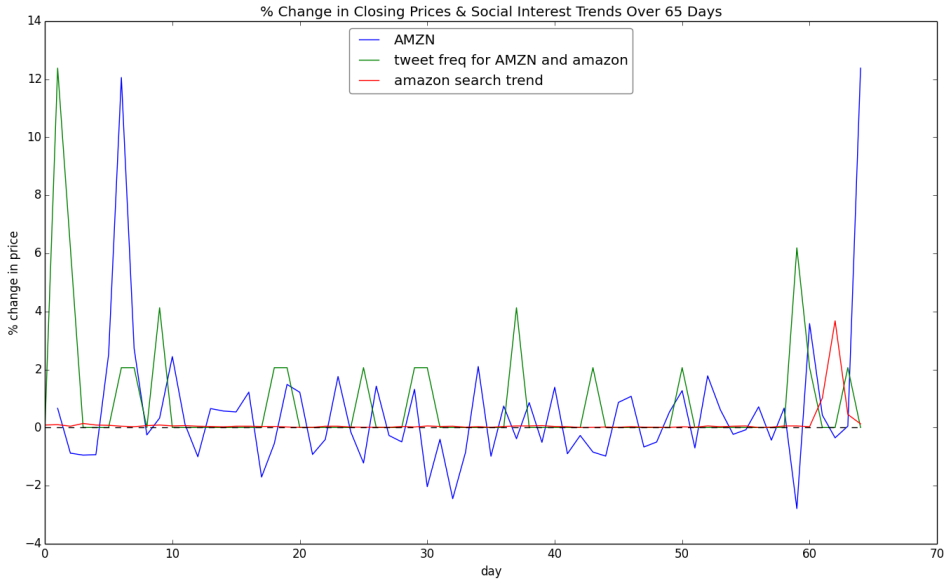
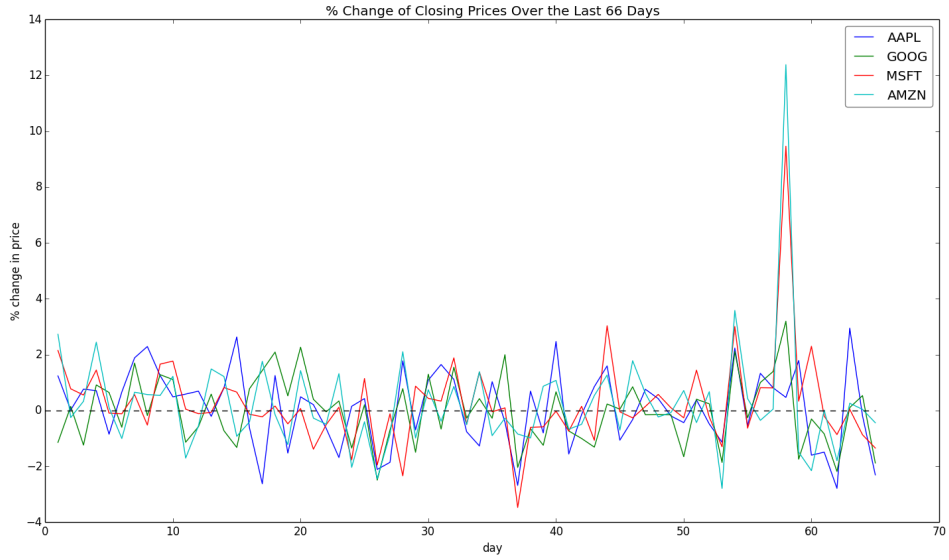


Figure 5


% Change of Closing Prices Over the Last 66 Days
AAPL
GOOG
MSFT
AMZN
% change in price
day
14
12
10
8
6
4
2
0
−2
−4
0
10
20
30
40
50
60
70