Using Social Media Posts to Predict Stock Price

DN & Drexel’s Big Data

**Data Science Capstone Project   
Data Acquisition and Pre-Processing Report**

Date:

[4/27/2021]

Team Members: Dhruvit Naik, Nelson Blickman

Name: Dhruvit Naik

Name: Nelson Blickman

[The purpose of this report is to describe the data of your project. It includes three major sections: Data Sources, Data-Processing, and Appendix]

**Identifying Data**

**Data Sources:**

[Identify the data sources of your project. It may have more than one data source. Describe each of them and explain why you select the data sources.]

Our main data source will be a Reddit API which contains a large amount of data, related to the comments and posts on Reddit. Such a data source will be ideal, because we have the ability to vary the parameters of our API call in order to efficiently pull the specific comments and posts that we want to experiment with. We will also be able to quickly manipulate the quantity of the posts and comments that we want to grab, in real time.

Another data source will be Twitter data using cashtag searches (ex. $GE). The search is only limited to the previous 7 days and does have rate limits using the free-tier API. The plan is to run our script every few days to get a large corpus of Twitter data for a few tickers.

We have also pulled Yahoo Finance Data for our selected companies for every day from December 2020 to March 2021.

**Acquisition Process:**

[Describe the data acquisition process. Is the dataset ready for download? How do you download the data? Do you need to write your own code to acquire the data from a public or private source? Describe how you do it. Are there multiple data sources? How do you integrate the data from multiple sources? Any other process involved in the acquisition process?]

As mentioned we have written code in order to acquire data from Reddit’s API. In our code, we test this by putting specific parameters into the API request. For example, reddit comments that have the word “GE” in it, which stands for General Electric. Other parameters include the start and end dates of such posts and comments, along with a the subreddit group that is associated with the posts and comments. We store the data we retrieve from this request in to a Python object, which we then turn into JSON format.

**Issues:**

[Are there any potential issues in data acquisition that have not be solved yet?]

There are no pressing issues for the reddit api. This is not the cleanest data object that we are requesting. For example, it is not in a ready to go csv file. Therefore, there will be a significant amount of pre-processing required. We have started this process and is discussed more below.

For the twitter api, we have a pipeline in place to grab data every few days to get the complete data for a couple months, since the free-tier does not offer historical data.

**Data-Processing**

[Examine the data you have acquired and understand the data properties. Is there any pre-processing you need to do before you can start analyzing the data? For example, missing data, sparsity, noise, veracity, ambiguity, interoperability, etc. Describe each data issue in a sub-section and explain how you clean up the data.]

The first preprocessing task we had involved iterating through our json object. Each iteration contained a Data Post. We pulled out the unique id in this post, and then searched to find all of the Data Comments that are associated with this particular Reddit post.

Next, since each comment has a number of fields or features, we are going to need to extract out the field “body” which contains the comments “text”. Next we will compile of all the text into cleaner data objects, creating different objects for each of the different companies (or tickers) that the text is associated with.

Challenges are likely to include sparsity and noise. For example, the amount of Reddit comments we have access to are plentiful. However, the “sparsity” is in regards to the fact that not all of the comments will be relevant. This is why I also use the word “noise” to describe this phenomena. Yes, there is plenty of data, but there are a lot of irrelevant comments. Applying “stop words” and using other pre-processing techniques to eliminate some of this noise will be considered.

**Appendix**

[Provide the code or pseudo code, data definition, sample data, and any other information in the appendix here.]

The below code is our central data acquisition code. In this case our parameters include the company ticker “q”, the time range “after” and “before”, the “subreddit”, and the “url”. This will allow for high flexibility and scalability in regards to acquiring the varying data which we desire.



The output below contains text of the comments that are related to the posts we pulled above.

{'all\_awardings': [],

'approved\_at\_utc': None,

'associated\_award': None,

'author': 'kamikazejesus',

'author\_flair\_background\_color': None,

'author\_flair\_css\_class': None,

'author\_flair\_richtext': [],

'author\_flair\_template\_id': None,

'author\_flair\_text': None,

'author\_flair\_text\_color': None,

'author\_flair\_type': 'text',

'author\_fullname': 't2\_1jreekhd',

'author\_patreon\_flair': False,

'author\_premium': False,

'awarders': [],

'banned\_at\_utc': None,

'body': 'One of my hypebeast reselling cousin stocked heavy on PS5s cause he caught the sleeper on Kohl’s... idk I’m looking at Halliburton',

'can\_mod\_post': False,

'collapsed': False,

'collapsed\_because\_crowd\_control': None,

'collapsed\_reason': None,

'comment\_type': None,

'created\_utc': 1606867268,

'distinguished': None,

'edited': False,

'gildings': {},

'id': 'gebcvpx',

'is\_submitter': False,

'link\_id': 't3\_k4rktg',

'locked': False,

'no\_follow': True,

'parent\_id': 't1\_geary1h',

'permalink': '/r/wallstreetbets/comments/k4rktg/on\_december\_11\_sp\_is\_going\_to\_announce\_a\_company/gebcvpx/',

'retrieved\_on': 1606873451,

'score': 2,

'send\_replies': True,

'stickied': False,

'subreddit': 'wallstreetbets',

'subreddit\_id': 't5\_2th52',

'top\_awarded\_type': None,

'total\_awards\_received': 0,

'treatment\_tags': []},

**Twitter:**

def create\_url():  
 tweet\_fields = **"tweet.fields=author\_id"** query = urllib.parse.urlencode({**'query'**: **'#GE'**})  
 start\_time = datetime.datetime.fromisoformat(**"2021-05-03T00:00:00-04:00"**).isoformat()  
 end\_time = datetime.datetime.fromisoformat(**"2021-05-04T00:00:00-04:00"**).isoformat()  
 max\_results = 10  
  
 url = **f"https://api.twitter.com/2/tweets/search/recent?**{query}**&start\_time=**{start\_time}**"** \  
 **f"&end\_time=**{end\_time}**&max\_results=**{max\_results}**"** return url  
  
  
def create\_headers(bearer\_token):  
 headers = {**"Authorization"**: **"Bearer {}"**.format(bearer\_token)}  
 return headers  
  
  
def call\_endpoint(url, headers):  
 response = requests.request(**"GET"**, url, headers=headers)  
 print(response.status\_code)  
 return response.json()  
  
  
def handler():  
 bearer\_token = auth()  
 url = create\_url()  
 headers = create\_headers(bearer\_token)  
 json\_response = call\_endpoint(url, headers)  
 print(json\_response)

**{'data': [{'id': '1389427392360288259', 'text': "$VPT shorts better cover soon IMO. #GE has over 2B scans on their machines/year. If they are making just $50 for 1% of those scans that's $1B/year off the scans alone. @goldmemb1 https://t.co/Ieo4EhIpyv"}, {'id': '1389426133595303940', 'text': '$GE seeing sustained chatter on twitch over the last few days\r\rVia https://t.co/aWUyIvrbLo\r\r#ge #twitch https://t.co/5R7Yap1r1r'}, {'id': '1389417784044445703', 'text': '@edzardhika #Gë\n\nTHANKKCHOOO!🥺😭🤝🥳'}, {'id': '1389417729145380864', 'text': 'Vea cómo el grupo digital de #GE Aviación redujo el tiempo de desarrollo de #FlightPulse con soluciones de @Microsoft como #GitHub y #Azure. Vea el video aquí: https://t.co/JD8Gzu1oMS'}, {'id': '1389389947791159299', 'text': 'RT @LeonardoRN\_: Provável Flamengo! #GE\n\nDiego Alves\n\nIsla\nWilliam Arão\nGustavo Henrique\nFilipe Luís\n\nDiego\nJoão Gomes \n\nArrascaeta\nÉvert…'}, {'id': '1389387289026043904', 'text': 'Provável Flamengo! #GE\n\nDiego Alves\n\nIsla\nWilliam Arão\nGustavo Henrique\nFilipe Luís\n\nDiego\nJoão Gomes \n\nArrascaeta\nÉverton Ribeiro \nBruno Henrique \n\nGabi'}, {'id': '1389380359142318082', 'text': '#vach \n【真空管MOD】CHECK #ヴァーチ #櫻康 #vachtomo #champion600 #12AX7 #SOVTEK #Phillips #GE #ElectroHarmonix\nhttps://t.co/dFInGhO4Xx\n◼️YouTubeアップロード致しました〜\n\n動画とは関係ありませんが、一番キツイ時に「逃げる人」「手を差しのべる人」が居ますね〜後者を目指して頑張ります❣️'}, {'id': '1389368861443563522', 'text': '【定期】GE2RBインフラ仲間募集中！#GE'}, {'id': '1389344698674950144', 'text': 'mey🏴\U000e0067\U000e0062\U000e0073\U000e0063\U000e0074\U000e007f&amp;r54Wx5💂🏿\u200d♀️Á&lt;w&lt;8&lt;&gt;𝄞&gt;1👨\u200d❤️\u200d💋\u200d👨mtÁo xca𝄞mjuuu𝄞👨\u200d❤️\u200d💋\u200d👨dlv⃑q𝄞Áko🙌🏽d&amp;Áz⃑u𝄞j wfv 🏄🏼\u200d♀️Á&lt;lcuhibwdt⃑5 FOQPRNTp #ge'}, {'id': '1389341242291408897', 'text': "#GE settled ATO audit for $273m in UK 'tax fraud' case\n\n#GEHealthcare #GeneralElectric #GEEnergy #GECapital #GEPower #GEAviation $GE https://t.co/OishG7B7Ii"}], 'meta': {'newest\_id': '1389427392360288259', 'oldest\_id': '1389341242291408897', 'result\_count': 10, 'next\_token': 'b26v89c19zqg8o3fostukqekxyey1v5bj3mwxnijde1dp'}}**

Table of Contributions

The table below identifies contributors to various sections of this document.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Data Sources** | **Nelson/Dhruvit** | **Nelson/Dhruvit** |
| **2** | **Data Pre-Processing** | **Nelson/Dhruvit** | **Nelson/Dhruvit** |
| **3** | **Appendix** | **Nelson/Dhruvit** | **Nelson/Dhruvit** |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.