Project Report CS 579

Analyzing and Predicting the areas/industries, which are grabbing investor's attention

Arpit Sheth(A20341089), Tirth Patel(A20320187)
Illinois Institute of Technology, Chicago

***Introduction*:**

- There are a large number of inexperienced Investors around the world who have a considerable amount of money to Invest, but they have a lack of guidance and not sure where to invest so they don’t have any idea about where to Invest. They are regularly reviewing past pricing history, and past investments and using it to influence their future investment decisions. Most of the times it results in failure. These Investors do not have the idea about whom to ask for their Investments, so they join some professional organizations who had done a lot of research in investing and charge them a lot amount of money just to give guidance.

- To help these amateur investors, our approach is to follow top-60 Investors/ Investing firms who are using Twitter a lot and analyze their twitter data and show which sector is trending so they can Invest and get profit.

- Nowadays Twitter is the best place for famous personalities to get engaged with other people. They regularly post their recent activities to inform their followers what’s happening around them. So we evaluate mutual interest of top-60 investors by analyzing their tweets and retweets and show common topic they are discussing on Twitter.

***Data*:**

- As a part of data, we look upto screen names of top-60 Investors/Investing firms who use Twitter on a daily basis as initial Input. Using Twitter API, We collected 140,541 tweets from all these 60 users. When we get a tweet from Twitter API It contains many fields. But for these experiment we only use ‘text’ and ‘created_at’ fields and dumped all these tweets with these fields to a *pickle[4]* in chronological order for future use. This works as our Training dataset.

- We collect dataset by unique keywords for specific 16 Industry Identification. We gathered more than 300 unique words and distributed them to relevant industries. This will work as our Testing dataset.

Methods:

- Twitter API and Twitter Authentication are used to fetch all tweets of users which are included in input file for experiment.

- First step in our experiment is to find the Top 60 Investors/ Investing companies who use Twitter on a daily basis and add their screen name in input.txt file. Twitter API is used to fetch all tweets of users which are include in Input file with their timestamps which are further dumped in data.txt file in chronological order and serialized as pickle for further use.

- As amount of tweets are high, we use Pickling and Unpickling for serializing and Deserializing fetched dataset respectively. Pickling is a method where python's object hierarchy dataset is converted into byte stream and Unpicking is inverse process of Pickling.

- Additionally, we create our training dataset which is stored in industrylist.txt which include more than 300 unique keywords which are used by 16 individual Industries frequently. Each line of this file belongs to individual Industry and the first word of line define Industry name and words followed by that Industry name(first word) are unique words of one Industry. We used tokenize method to tokenize this file to get dictionary of industry name and their unique keywords.

- We vectorize our training data using Counter vectorizer.Counter vectorizer is used to convert our tweets in a pickle to draw a sparse matrix of unique token counts in which rows are tweets and columns are unique terms of those tweets. While building this vocabulary we consider values of min_df = 2 and max_df = 7. For each column we check whether is it in our Industry list or not. If It's in Industry List then It increase the counter of relevant industry. At the end of these It gives most famous industry.

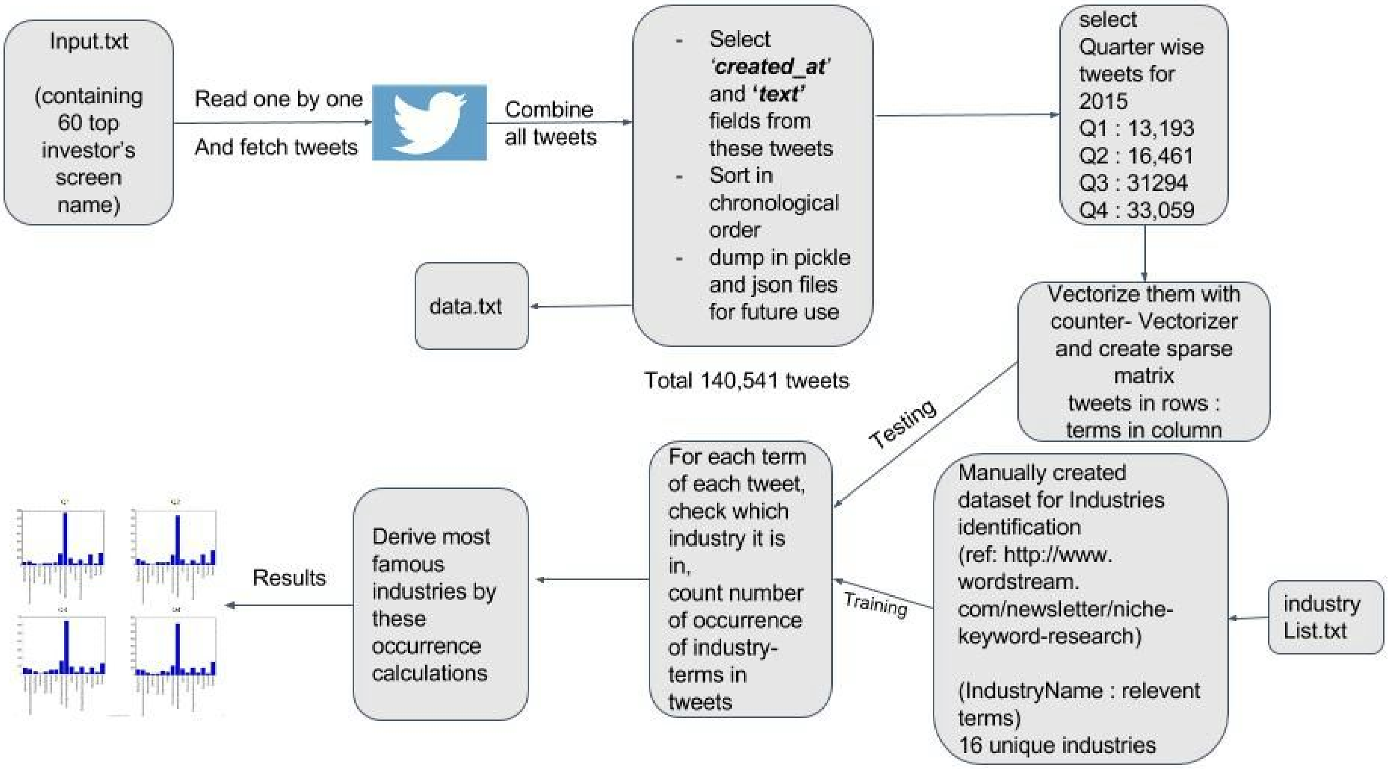
Input.txt
(containing 60 top investor's screen name)
Read one by one
And fetch tweets
Combine all tweets
- Select '*created_at*' and '*text*' fields from these tweets
- Sort in chronological order
- dump in pickle and json files for future use
data.txt
Total 140,541 tweets
select Quarter wise tweets for 2015
Q1 : 13,193
Q2 : 16,461
Q3 : 31294
Q4 : 33,059
Vectorize them with counter- Vectorizer and create sparse matrix
tweets in rows : terms in column
Testing
For each term of each tweet, check which industry it is in,
count number of occurrence of industry-terms in tweets
Training
Manually created dataset for Industries identification
(ref: http://www.wordstream.com/newsletter/niche-keyword-research)
(IndustryName : relevent terms)
16 unique industries
industry List.txt
Derive most famous industries by these occurrence calculations
Results

Figure.1) Whole Experiment

Experiments:

- Figure-1 explains our whole experiment for Investors' trend during the year 2015.

- We test our fetched testing data i.e. tweets only for year 2015 to check Investors’ trend during each quarter of 2015.
- From our quarterly results of year 2015, we can see most-common sector, investors are talking about is “ideological and single issues”, which is kind of intuitive (talking about social issues and ideology).
- Other than “ideological and single issues”:
    - Q1: Finance> IT> Education> labor> energy
    - Q2: Finance> Education> IT>Agriculture> labor
    - Q3: IT> Finance> labor> energy > Education
    - Q4: Finance> IT> energy> Education> labor

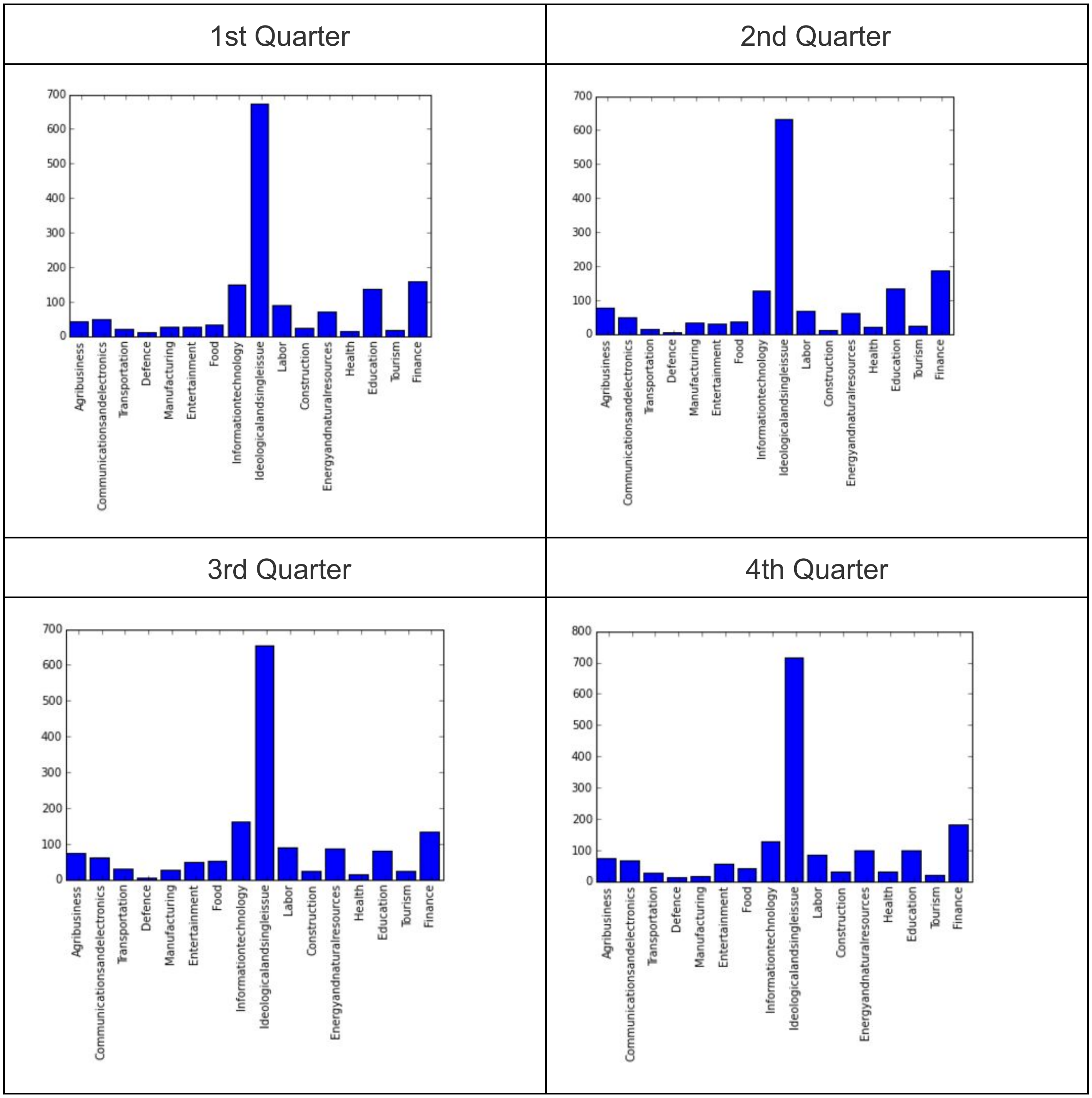

Figure.2) Result of each Quarter in year 2015

Related work:

- Converting people’s online social presence to financial analysis is relatively new field. We were able to find projects related to predicting stock market pricing based on people’s social activity, their positive or negative thoughts for specific company[8][9].

- Our approach is different as we are trying to defining investment strategy based on popularity of company/industry. As an initial step, we were able to find tending industries among the top investors based on their recent tweets.

***Conclusions and Future Work*:**

- Our approach is based on Industry Identification dataset which we created. Classification of these tweets can be done in more specified industries and accurately as we grow our training dataset. We can add more industries in relevant terms to enhance our dataset, which will result in accurate accurate industry trends.

- We will implement multiprocessing in our module to optimize running time.

- We can build a proper investing strategy based on our outputs and actual stock pricing for inexperienced and ammature investors.

***References*:**

1. http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
2. https://dev.twitter.com/rest/reference/get/search/tweets
3. Google: https://www.google.com/search?q=top+investors+in+the+world
4. Python Documentation: https://**docs**.**python**.org/
5. http://docs.scipy.org/doc/numpy/reference/generated/numpy.array.html
6. http://matplotlib.org/api/pyplot_api.html
7. stackoverflow.com
8. http://arxiv.org/ftp/arxiv/papers/1305/1305.7014.pdf
9. http://cs229.stanford.edu/proj2011/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf