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

Arpit Sheth : A20341089

Tirth Patel : A20320187

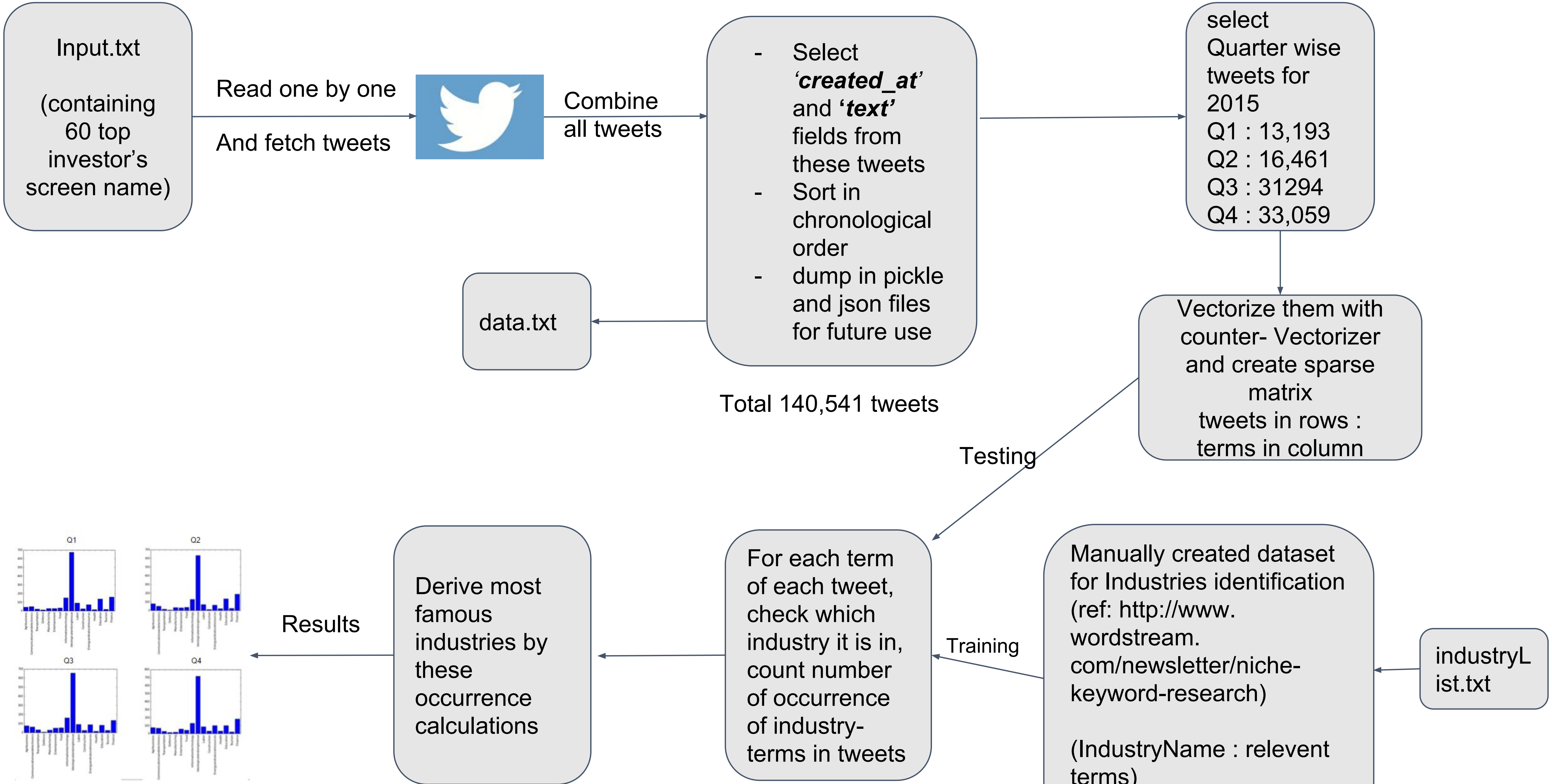
CS579 – Online Social Network Analysis,
IIT, Chicago

Problems and concerns

- Large number of inexperienced and immature investors
- They want to invest in stock market/industries/startups
- Where to invest?
- Whom to ask?
- How much do they charge just to give you this guidance?

Approach

- Why not follow top investors!
- Analyze recent activities of top individual investors and investing companies
- Looking into details about their
    - recent tweets
- Evaluate data to find common list of interest
- Predict trending industries
- Cross check findings by real time stock market data



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

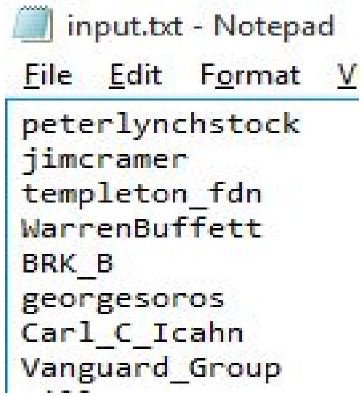
industryList.txt

Derive most famous industries by these occurrence calculations

Results

Data Collection

- Collected list of leading 60 investors (google: top investors, forbes list, South East Asia Investors etc.) and found their twitter screen-names. Stored in 'input.txt' file



```
peterlynchstock
jimcramer
templeton_fdn
WarrenBuffett
BRK_B
georgesoros
Carl_C_Icahn
Vanguard_Group
```

- Use twitter API to:
  - Collect all available recent tweets for them (max 3200 for a user)
    - Total 140,541 tweets collected
    - Took 'created_at' and 'text' part of the tweet as needed
    - Dumped all tweets in 'data.txt' file in json format and into pickle

Year	2009	2010	2011	2012	2013	2014	2015
Tweet Count	304	2218	3060	5305	9231	21889	95139

Q1	13,193
Q2	16,461
Q3	31294
Q4	33,059

- We will be focusing on year 2015 tweets. We will divide these tweets in 4 quarters.

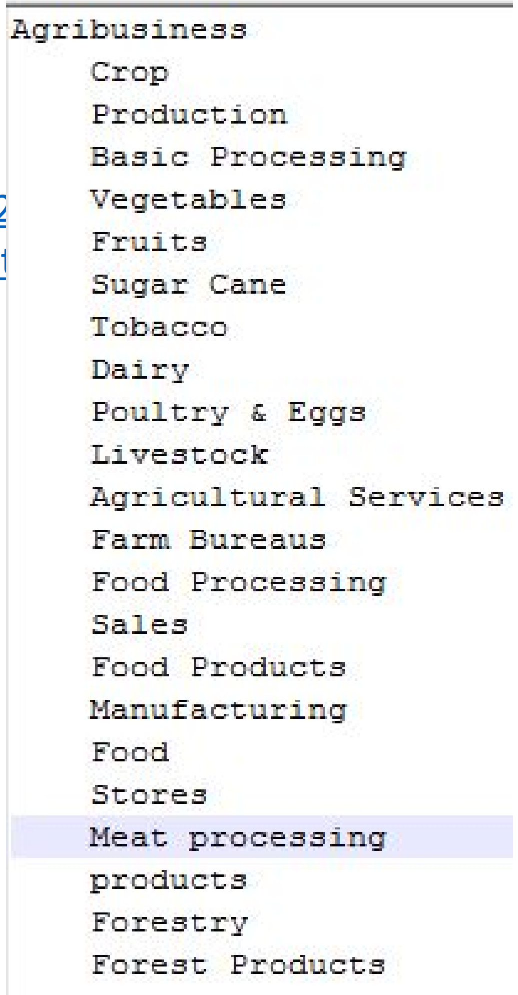
Data Collection

- We have created dataset for industry identification (ref: http://www.sos.la.gov/BusinessServices/PublishedDocuments/Industry%20Business%2 //www.opensecrets.org/industries/alphalist.php and http://www.wordstream.com/newslet research)
  - This dataset is like-> industryname: list of relevant terms
  - This is stored in 'industryList.txt' file

- Our this industry identification dataset is very primary with 16 industry types

- Industry relevant 330 words on social network are classified into these 16 industries

- We can grow this dataset to get more accurate results



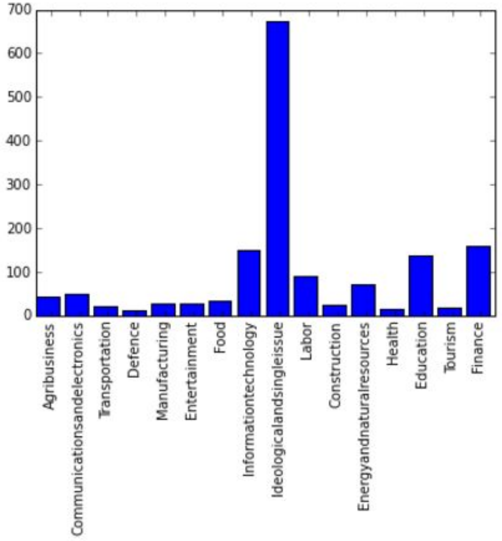
```
Agribusiness
    Crop
    Production
    Basic Processing
    Vegetables
    Fruits
    Sugar Cane
    Tobacco
    Dairy
    Poultry & Eggs
    Livestock
    Agricultural Services
    Farm Bureaus
    Food Processing
    Sales
    Food Products
    Manufacturing
    Food
    Stores
    Meat processing
    products
    Forestry
    Forest Products
```

Sample of our dataset

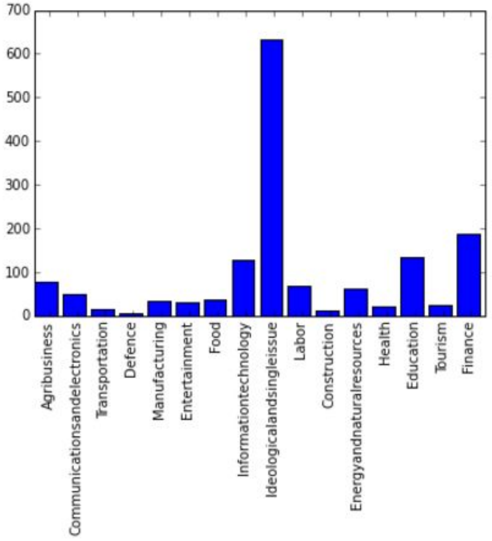
Results

- 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
    - Q1: IT> Finance> labor> energy > Education
    - Q1: Finance> IT> energy> Education> labor
- Our approach is based on Industry Identification dataset which we created. investor’s tweets can be classified more accurately between industries as we grow this training dataset
- We faced difficulties in cross verifying our results.

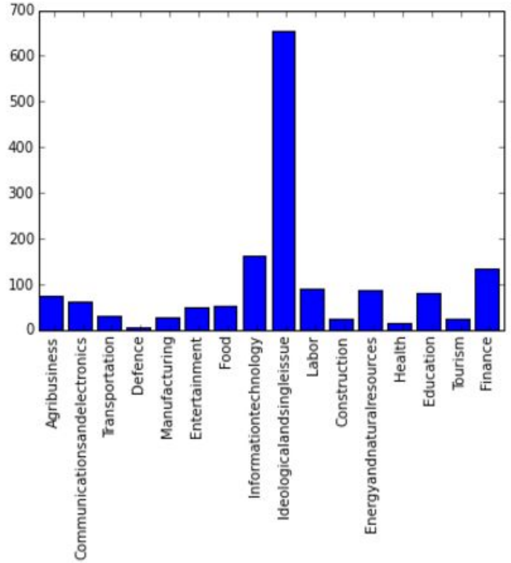
Q1



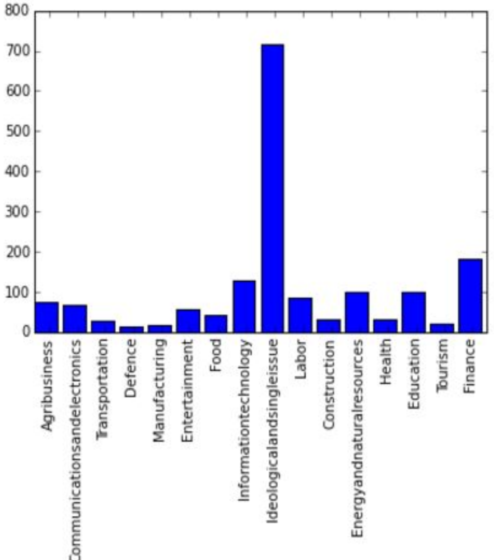
Q2



Q3



Q4



Conclusion

- We analysed top investor's tweets and concluded popular/trending sectors in each quarters of 2015
- 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.
- From our analysis one amatuer person can predict which industry is more popular among top Investors on Twitter.
- Future enhancements:
    - We can build a proper investing strategy based on our outputs and actual stock pricing for inexperienced and immature investors.