Sentiment Analysis on Tweets

Team 2
Yuan Ying - ying.yua@husky.neu.edu
Mushtaq Rizvi - rizvi.m@husky.neu.edu
Wei Huang - huang.wei3@husky.neu.edu
Jinjin Zhang - zhang.jinj@husky.neu.edu

Goals of the project

- Process real Twitter datasets to extract meaningful analysis by performing Sentiment Analysis
- In this project, we utilize information available through the Twitter API to gather information about the tweets and their users

Use case

- User will input
  - A Keyword
  - Select one or more location
  - Pick a date (Optional)
- User will get
  - A visualization result such as a bar chart represent the sentiments of tweets with that keyword in those locations at that date

Methodology

- Parse tweets(JSON format)
  - Text field
  - User field
  - Created_at field - Date
  - Location field and etc.

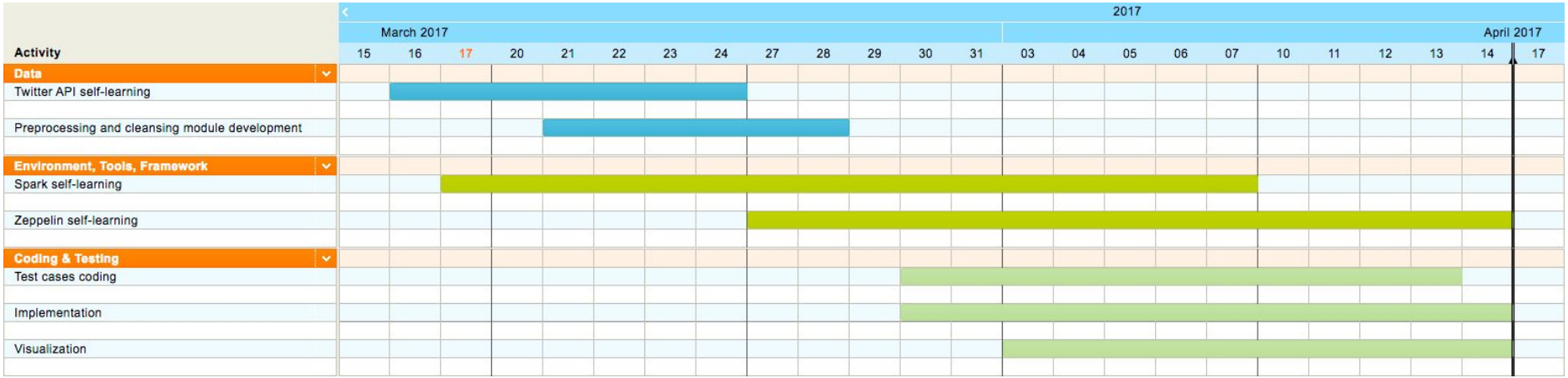

```
"created_at": "Tue Jan 27 01:52:28 +0000 2016",
  "id": 555661720180465617,
  "id_str": "555661720180465617",
  "text": "welcome DR0PTEAM @parkjin @GyoungD @OSJVNG @seolgim",
  "source": "\u003ca href\u003d\"http://twitter.com\" rel\u003d\"nofollow\"\u003eTwitter Web
Client\u003c/a\u003e",
  "truncated": false,
  "user": {
    "id": 348726398,
    "id_str": "348726398",
    "name": "joe",
```


- Clean and break down the text field into words
- Filter the tweets by keyword
- Identify and mark the word sentiment
- Calculate and catalog tweets using Stanford NLP
- Visualization using Apache Zeppelin

Data sources

- Tweets based on certain parameters like keyword, language, location, etc can be set to define what data to request
- Finding tweets by the location can be done either by the Streaming API or Search API
- Data size: around 100,000 tweets which cover last 7-day tweets
- Reference:http://140dev.com/twitter-api-programming-tutorials/aggregating-tweets-search-api-vs-streaming-api/

Milestones/sprints



Programming in Scala and code repository

- Most part of project will be programed in scala including
  - Parsing
  - Cleaning
  - Filtering
  - Identifying
  - Calculating
  - Also all the unit test will be programed in scala
- Code repository - GitHub
  - https://github.com/yingy4/CSYE7200_FinalProject_Team2_Spring2017
- Document - Google Slides

Acceptance criteria

- Verify analysis results with test tweets (created by ourselves).
  - The accuracy should reach 90%
- Verify analysis results with input like “weather” is “bad” in a certain location on a certain date (check against weather report).
  - The accuracy should reach 70%
- Verify analysis results with input like “Apple Inc stock” is “good” on certain date (check against Yahoo Finance).
  - The accuracy should reach 80%

Q&A

Thank you!