CSYE7200
PROJECT

Sentiment Analysis on Tweets

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

PROJECT GOAL

- PROCESS REAL TWITTER DATASETS
- UTILIZE TWITTER API
- EXTRACT MEANINGFUL DATA
- NLP(NATURAL LANGUAGE PROCESS)
- PERFORM SENTIMENT ANALYSIS

USER INPUT

- A CITY NAME

  OR

- A COMPANY NAME

SYSTEM OUTPUT

- SENTIMENT SCORE FOR
- THE CITY WEATHER IN LAST 7 DAYS

  OR

- THE COMPANY STOCK IN LAST 7 DAYS

USER INPUT

- NOTHING

  OR

- A KEYWORD

SYSTEM OUTPUT

- TOP 10 POPULAR HASHTAGS FROM UPCOMING TWITTER CONTAINING THAT KEYWORD IN A 5MIN WINDOW
- THE COUNT OF TWEETS AND SENTIMENT SCORE FOR EACH HASHTAG
- A BAR CHART SHOWING THE RESULT*

USE CASE 3

USER INPUT

- NOTHING

  OR

- A KEYWORD

SYSTEM OUTPUT

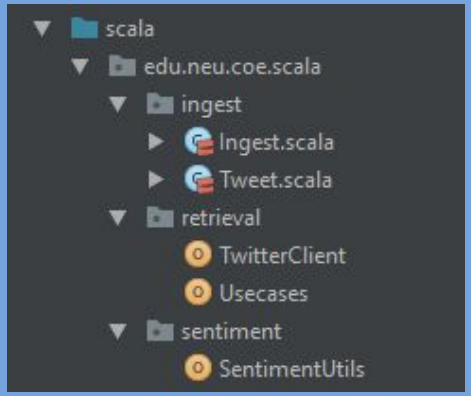
- TOP 10 POPULAR LOCATIONS FROM UPCOMING TWITTER CONTAINING THAT KEYWORD IN A 10HOUR WINDOW
- THE COUNT OF TWEETS AND SENTIMENT SCORE FOR EACH LOCATION
- A MAP SHOWING THE RESULT*

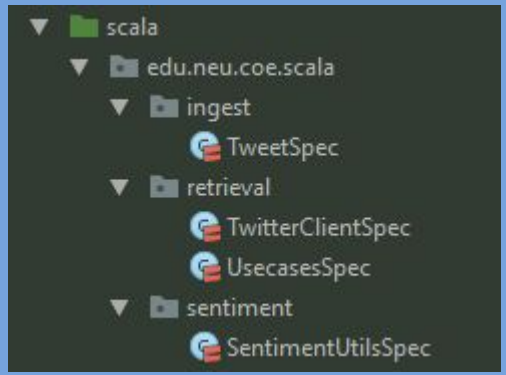
METH-ODOLOGY

- ACQUIRING (SEARCH API & STREAMING API)
- PARSING (JSON FORMAT)
- FILTERING (LANGUAGE & GEOLOCATION)
- CLEANING (SPECIAL CHARACTERS)
- CALCULATING SENTIMENT (STANFORD NLP)
- MAPPING & REDUCING (SPARK)
- VISUALIZATION (APACHE ZEPPELIN)

PROGRAM MING

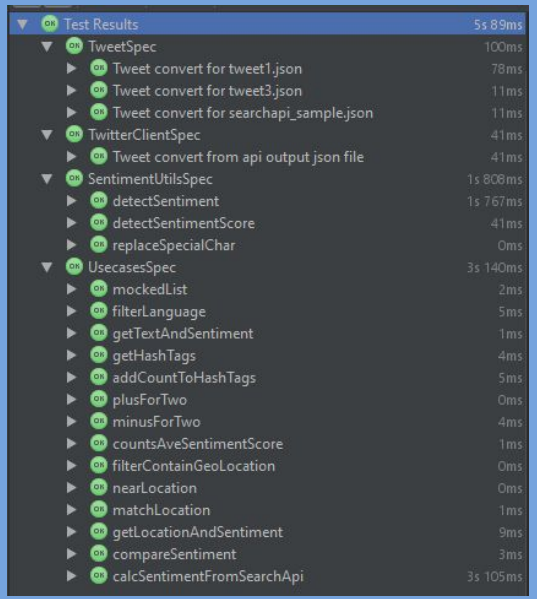
- Ingest.scala
- Tweet.scala
- TwitterClient
- Usecases
- SentimentUtils
- Unit Test





UNIT TEST

- TOTAL 43 UNIT TESTS
- 74% CLASSES COVERED
- 60% LINES COVERED



ACCEPTANCE CRITERIA

- VERIFY ANALYSIS RESULT WITH TEST TWEETS
  - Included in unit test
  - The accuracy should reach 90%
- VERITY ANALYSIS RESULT WITH KEYWORD RELATED TO WEATHER JUDGEMENT
  - Compare 2 city's weather in 7 days
  - The accuracy should reach 70%
- VERIFY ANALYSIS RESULT WITH KEYWORD RELATED TO STOCK ASSESSMENT
  - Compare 2 company's stock in 7 days
  - The accuracy should reach 80%

WEATHER

Boston vs. New York

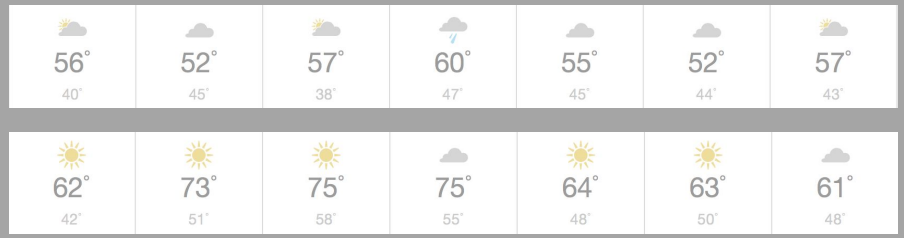
11 >>> 17

BOSTON GRADE: 1.35555556

NEW YORK GRADE: 1.41587301

PREDICTION: New York better than Boston

REAL WEATHER REPORT:



TRUTH: New York more pleasant than Boston

RESULT: accurate

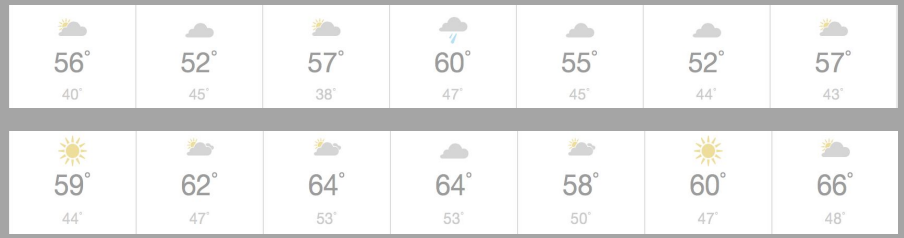
WEATHER

Boston vs. SF

11 >>> 17

BOSTON GRADE: 1.35555556
SF GRADE: 1.20634920
PREDICTION: Boston better than San Fran

REAL WEATHER REPORT:



TRUTH: San Francisco more pleasant than Boston

RESULT: in-accurate

WEATHER ANALYSIS SUMMARY

Among 10 cities over 7 days

TOTAL PREDICTION ACCURACY:
8/10 = 80%

SUCCESSFULLY ACHIEVED

ⓘ ACCEPTANCE CRITERIA reach 80%

STOCK

Apple Inc. vs. GNC

APPLE INC. GRADE: 1.09797297

GNC GRADE : 1.06976744

PREDICTION: Apple Inc. better than GNC

REAL STOCK REPORT:

DATE	OPEN	HIGH	LOW	CLOSE	
04/17/2017	141.48	141.88	140.87	141.83	16,529,130
04/13/2017	141.91	142.38	141.05	141.05	17,775,510
04/12/2017	141.6	142.15	141.01	141.8	20,320,420
04/11/2017	142.94	143.35	140.06	141.63	30,341,520

04/17/2017	7.31	7.38	7.08	7.24	2,968,706
04/13/2017	7.26	7.33	7.015	7.3	2,624,080
04/12/2017	7.44	7.44	7.21	7.27	2,404,252
04/11/2017	7.14	7.51	7.07	7.42	3,792,926

TRUTH: GNC better than Apple Inc.

RESULT: in-accurate

STOCK

Apple Inc. vs. Netflix

APPLE INC. GRADE: 1.09797297

Netflix GRADE : 1.41774193

PREDICTION: Netflix better than Apple Inc.

REAL STOCK REPORT:

DATE	OPEN	HIGH	LOW	CLOSE	
04/17/2017	141.48	141.88	140.87	141.83	16,529,130
04/13/2017	141.91	142.38	141.05	141.05	17,775,510
04/12/2017	141.6	142.15	141.01	141.8	20,320,420
04/11/2017	142.94	143.35	140.06	141.63	30,341,520
04/17/2017	144.43	147.32	144.43	147.25	15,993,460
04/13/2017	144.25	144.55	142.76	142.92	3,676,127
04/12/2017	144.85	145.74	143.55	143.83	4,383,524
04/11/2017	144.28	144.54	141.98	144.35	4,619,263

TRUTH: Netflix better than Apple Inc.

RESULT: accurate

STOCK ANALYSIS SUMMARY

Among 10 cities over 7 days

TOTAL PREDICTION ACCURACY:

8/10 = 80%

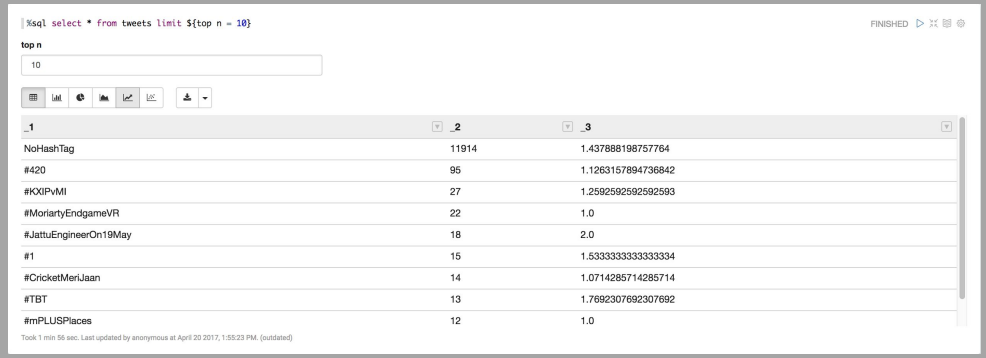
SUCCESSFULLY ACHIEVED

ⓘ ACCEPTANCE CRITERIA reach 80%

MOST POPULAR HASHTAG

5mins duration

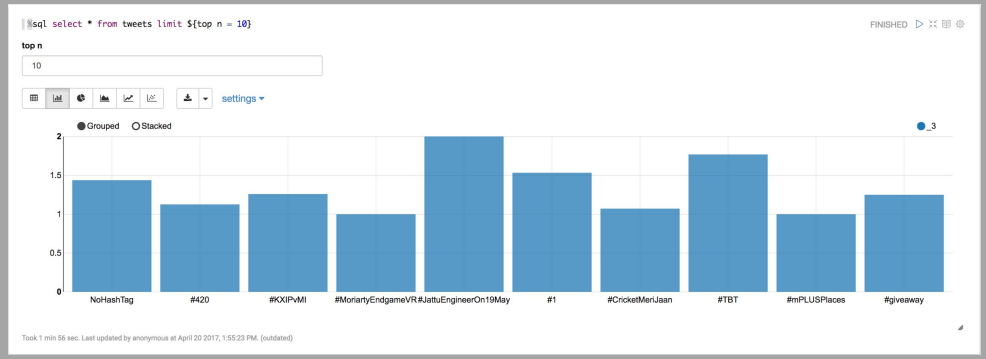
TABLE:



_1	_2	_3
NoHashTag	11914	1.437888198757764
#420	95	1.1263157894736842
#KXIPvMI	27	1.2592592592592593
#MoriartyEndgameVR	22	1.0
#JattuEngineerOn19May	18	2.0
#1	15	1.5333333333333334
#CricketMeriJaan	14	1.0714285714285714
#TBT	13	1.7692307692307692
#mPLUSPlaces	12	1.0

Took 1 min 56 sec. Last updated by anonymous at April 20 2017, 1:55:23 PM. (outdated)

CHART:



TOTAL DIFFERENT HASHTAG COUNT: 2279

SENTIMENT MAPPING

10 hour duration

http://35.166.178.247:8090/#/notebook/2CFBTGPRZ

CODE REPOSITORY

https://github.com/yingy4/CSYE7200_FinalProject_Team2_Spring2017

REFERENCE

- Course Repo
- http://140dev.com/twitter-api-programming-tutorials/aggregating-tweets-search-api-vs-streaming-api/
- https://www.udemy.com/apache-spark-with-scala-hands-on-with-big-data/learn/v4/overview
- https://community.hortonworks.com/content/kbentry/90320/add-leaflet-map-to-zeppelin-notebook.html
- https://dev.twitter.com/rest/public
- https://github.com/vspiewak/twitter-sentiment-analysis/blob/master/src/main/scala/com/github/vspiewak/util/SentimentAnalysisUtils.scala