Twitter Trend Summarizer – TweetLyze

DISSERTATION

Submitted in partial fulfillment of the requirements of the

MS Software Engineering Degree programme
By

NISCHAL HP
2010HS70022

Under the Supervision of

Dr. Ashok Veilumuthu
Senior Researcher

Dissertation work carried out at

SAP-Labs, Bangalore

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

June, 2014

Twitter Trend Summarizer – TweetLyze

DISSERTATION

Submitted in partial fulfillment of the requirements of the

MS Software Engineering Degree programme
By

NISCHAL HP
2010HS70022

Under the Supervision of

Dr. Ashok Veilumuthu
Senior Researcher

Dissertation work carried out at

SAP-Labs, Bangalore

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

June, 2014

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the Dissertation entitled `Twitter Trend Summarizer – TweetLyze` submitted by **Mr.Nischal HP**, IDNo.**2010HS70022** in partial fulfillment of the requirements of SESAP ZG629T Dissertation, embodies the work done by her under my supervision.

Place: Bangalore
Date:

Signature of the Supervisor
Name: **Dr.Ashok Veilumuthu**
Designation: Senior Researcher

ACKNOWLEDGEMENT

I have taken a lot of efforts to build this project. However, it would not have been possible without the kind support and help of many individuals and I would like to extend my sincere thanks to all of them.

I am highly thankful to my mentor Dr. Ashok Veilumuthu for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project.

I would like to specially thank my batch mates Raghotham. S and Shrayas Rajagopal for pitching in every time I had a problem and helping me solve them leading to the progress and completion of my project.

I would like to express my special gratitude and thanks to Mr. Ajit Joshi for giving me such attention and time.

I would like to express my special gratitude and thanks to SAP Labs Pvt Ltd, Bangalore for providing all the necessary infrastructure, software and great support for accomplishing this project.

I would like to express my gratitude towards my family for their kind co-operation and encouragement which helped me in completion of this project.

I am highly thankful to the Vocational Training Team here at SAP Labs Pvt Ltd, Bangalore for providing the opportunity to be a part of the wonderful scholar program.

Last but not least a special thanks to Birla Institute of Technology for opening a door to complete my post-graduation.

Nischal HP

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

SECOND SEMESTER 2013-2014

SESAP ZG629T DISSERTATION

Dissertation Title : Twitter Trend Summarizer - TweetLyze

Name of Supervisor : Dr. Ashok Veilumuthu

Name of Student : Nischal HP

ID No. of Student : 2010HS70022

Abstract

"Twitter Trend Summarizer - TweetLyze"

When a lot of tweets are talking about one thing in particular, they get categorized into trends. And each trend has millions of tweets associated to it and at any given point it is very hard to understand what the trend is actually talking about.

So the idea here was to build a summarizer that would let people understand what the trends meant, thereby helping people to make business decision based on the trends.

Social Media analytics is now being widely used to understand people across the world and help companies make better business decisions and devise new strategies.

Table of Contents

INTRODUCTION

Twitter is a micro-blogging social network. People all around the world use it extensively for varied reasons.

I can keep going on and on about twitter but nothing speaks as clearly as statistics.

Twitter Company Statistics	**Data**
Total number of active registered Twitter users	645,750,000
Number of new Twitter users signing up everyday	135,000
Number of unique Twitter site visitors every month	190 million
Average number of tweets per day	58 million
Number of Twitter search engine queries every day	2.1 billion
Percent of Twitter users who use their phone to tweet	43 %
Percent of tweets that come from third party applicants	60%
Number of people that are employed by Twitter	2,500
Number of active Twitter users every month	115 million
Percent of Twitters who don't tweet but watch other people tweet	40%
Number of days it takes for 1 billion tweets	5 days
Number of tweets that happen every second	9,100
Twitter Annual Advertising Revenue	**Revenue**
2013	$405,500,000
2012	$259,000,000
2011	$139,000,000
2010	$45,000,000

Looking at the current statistics we realize the data that is being generated every day is huge and Twitter has become one of the fastest growing software giants in the world.

Tweets end up creating **trends** if a large set of tweets are talking about the same subject, topic, event etc. in a given region.

Trends are very abstract as only a word is considered to be a trend. For example if millions of people are tweeting about Sachin Tendulkar and using #ThankYouSachin in their tweets, then the trend is #ThankYouSachin.

Now anybody will get curious as to why people are Thanking Sachin. There are 2 reasons why anybody would be interested to know what this trend means.

1) Curiosity
2) Business

You may now ask me how it affects business. Think about it, if the whole of India at a given point in time is talking about sachin and you start selling sachin tshirts for lesser price on that day, wouldn't you end up building a small business or expanding your horizon?

This is exactly what I am trying to solve. Understanding millions of tweets related to a trend and to build a summary around the trend, thereby empowering consumers to make decisions based on what the trend is all about.

Analysis and Proposed Solution

To achieve building a summarizer, I would first have to build a data collection framework to extract trends from different locations and their respective tweets.

As this needs to be a temporal system, I would need to collect tweets and trends over time pertaining to a location, in order to build a system that can understand trends and how trends change with regards to time for a given location.

Once data from twitter has been extracted, there are different means to understand the trends. The most common being, to read the text of the tweet related to a trend. I chose to not do it this way, as there has already been some research done in this area.

I propose to solve this problem by looking at the entities mentioned in all the tweets pertaining to a trend and understand the significance of these entities in the trend. I will do this by providing entity significance using measures like TFIDF.

I would also be performing clustering methods on all the trends for a given location, which in turn provides the understanding of trend similarity.

This would now result in grouping multiple similar trends into topics and understanding trends in a completely different way.

I would also be running a sentiment analyzer on all the tweets that is specific to a trend, thereby providing a different dimension to understanding trends.

Tools & Methodologies

The Tools and Techniques used in the project are listed in detail below:

Flask

A python based micro web framework to support the application.

HTML5 and CSS3

To build a web based User interface to visualize the summary for the trends.

JavaScript

To provide dynamic controls for the web app.

D3

JavaScript based Visualization library used to generate tag clouds

Python

I have used python for the analytical purposes where I use libraries such as:

1. Psycopg
2. Text blobs
3. PyCluster
4. Flask

Java

Data collection framework for twitter has been written using Java.

Postgres

Using postgres as the database to store relational data and also as JSON store for tweets.

Architecture

Before I started writing the data extraction framework I ran into a lot of problems understand the twitter rest api and how to work with rate limits.

There have been many a time when I have wanted to do some analysis on twitter but then ended up pushing it because I didn't have one important thing - The Data itself. When I set out to build a Trend Summarizer, I knew that I would have deal with this problem and wanted to solve it for myself. As I started building the framework, I faced a big hurdle of having to deal with rate limits.

The trend summarizer is a temporal and spatial system that required tweets belonging to trends in different locations. This meant that I had to have the ability to add locations and trends on a runtime basis without running into Twitter rate limit problems. Building a temporal and spatial system came with its own challenges of having to fetch data continuously and reduce redundant data as much as possible. To accomodate this, I had to spread out the calls to get data evenly over time and stay within limits. Also, given all these constraints, I had to collect large amounts of data for the summarizer to make sense.

This talk is about how I went around understanding the Twitter API along with its constraints and built a data collection framework. I built this out using Java with PostgreSQL as the backend.

Once this was done, I chose python to perform text analytics as I had good handlers to write data to the backend from python and also libraries like PyCluster ,Flask, textblobsetc to help me perform analytics and build a web app with least effort.

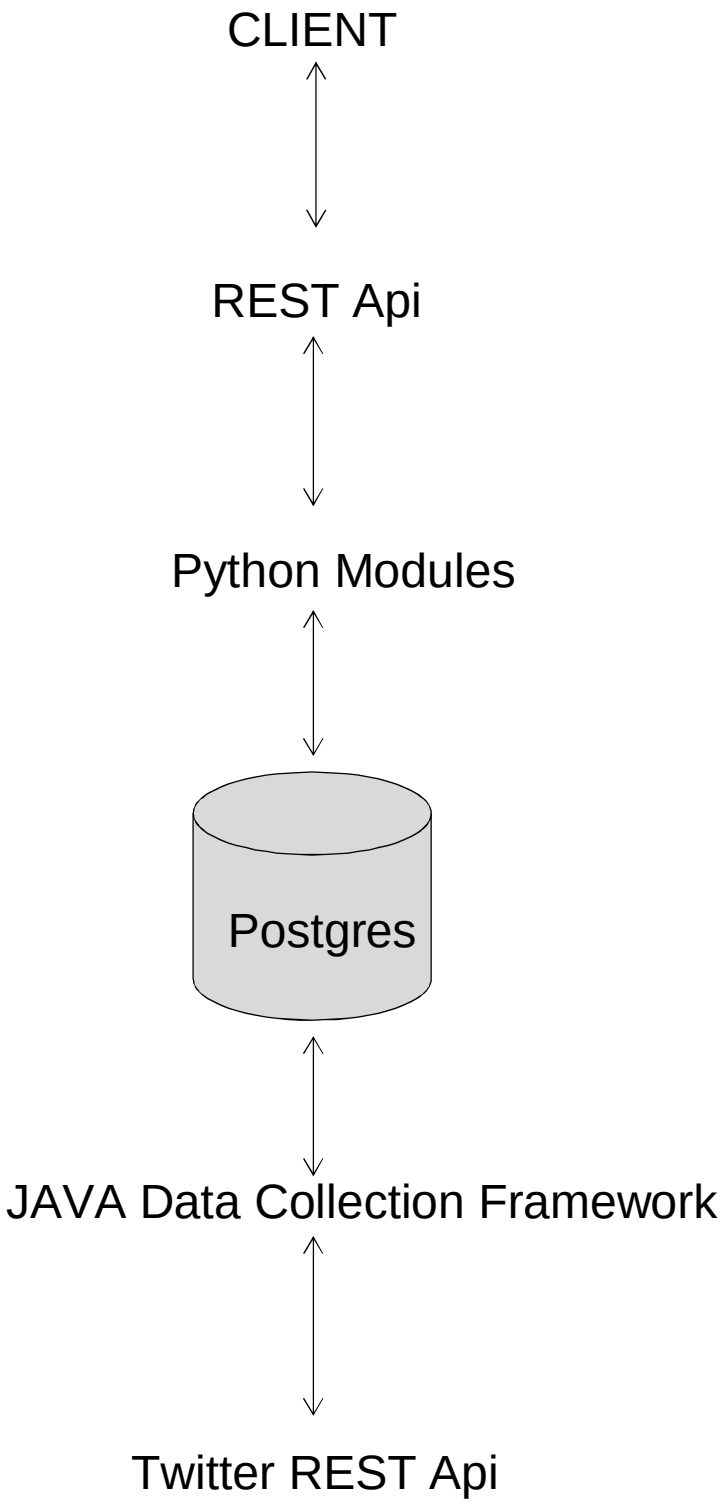
CLIENT
REST Api
Python Modules
Postgres
JAVA Data Collection Framework
Twitter REST Api

Figure 1

Implementation

Code Snippet From Data Extraction Framework

```
while (true) {

                    Stack<MrUrl> jobStack = MrMestri.buildJobs();
                    int jobToken = 0;
/*
* Removed all the complicate shit of timer
* and other conditions
* and finally figured out that all it needs is a basic if
* to get trends
* jobStack in this loop is empty if and only if there are no trends
* for the day
* So when that happens
* Boom ! You go and get trends
*
*/
if (jobStack.size() == 0){
          OAuthConsumer consumerObj = JuliusCaesar
                                      .getConsumerObject();
        GetTrends.retrieveTrends(consumerObj);
        JuliusCaesar.putConsumerObject(consumerObj);
}
while (jobStack.size() > 0) {
/*
* Now that the job stack is there we need to pick 80 jobs
* at a time and based on number of apps wait to run the
* process again
*/
MrRunnable worker = new MrRunnable(jobStack.pop());
jobToken++;
executor.execute(worker);
if (jobToken% Integer.parseInt(PropertyHandler.getProperty("totalTrends")) == 0) {
          log.info("All jobs for the clock cycle complete , waiting for next clock cycle to
start. Number of jobs completed "+ jobToken);
          log.info("Job stack size" + jobStack.size());
          Thread.sleep(milliseconds);
}
if (jobStack.size() == 0) {
          log.info("Now waiting to re energize my apps ");
          Thread.sleep(120000);
          }}
```

Code Snippet From App Server

```
from flask import Flask,send_file,jsonify,make_response
from flask import request
from Pipeline import Pipeline
app = Flask(__name__,instance_relative_config=True)
app.config.from_pyfile('config.py')


@app.route('/')
def hello_world():
        return send_file('static/index.html')

@app.route('/locations/',methods=['GET'])
def get_locations():
        pipeline_obj = Pipeline();
        locations_list = pipeline_obj.get_locations()
        json_dict = {}
        json_dict['data'] = locations_list
        print jsonify(json_dict)
        return jsonify(json_dict)

@app.route('/trends/',methods=['GET'])
def get_trends():
        location_id = request.args.get('locationid')
        min_date = request.args.get('min_date')
        max_date = request.args.get('max_date')
        pipeline_obj = Pipeline()
        trends_list = pipeline_obj.get_trends(location_id,min_date,max_date)
        json_dict = {}
        json_dict['data'] = trends_list
        return jsonify(json_dict)

@app.route('/dates/<location_id>/',methods=['GET'])
def get_dates(location_id):
        pipeline_obj = Pipeline()
        dates_list = pipeline_obj.get_dates_location(location_id)
        json_dict = {}
        json_dict['data'] = dates_list
        print json_dict
        return jsonify(json_dict)
```

Code Snippet From Text Analytics Module

```
def get_trends(self,location_id,start_date,end_date):
          trends_list = []
          try:
                    conn = PostgresConnector().get_connection()
                    cursor = conn.cursor()
                    query = """
                    select c,trend from
                    (select count(*) as c,trend from trends where
                    locationid = %s and date between %s and %s group by trend)as t1 order by c
desc limit 15
                              """
                    cursor.execute(query,(location_id,start_date,end_date))
                    trend_column = 1
                    count_column = 0
                    for row in cursor:
                              trend_count = {}
                              trend_count["trend"] = row[trend_column]
                              trend_count["count"] = row[count_column]
                              trends_list.append(trend_count)
                    except Exception as e:
                              print e

                    return trends_list
```

Conclusions / Recommendations

In the last few pages, I have tried to summarize the entire work done on Trend Summarize- Tweetlyze. Understanding the twitter social network, extracting data from it and then transforming all of the raw data into impactful information were the major tasks done during the course of the project including management of data and making sense of the flood of data—a task that requires sophisticated and complex analytic models.

Bibliography / References

Flask :

http://flask.pocoo.org/

Twitter Bootstrap:

http://getbootstrap.com/2.3.2/

D3:

http://d3js.org/

Postgres:

http://www.postgresql.org/

Text blobs:

http://textblob.readthedocs.org/en/dev/

PyCluster:

https://pypi.python.org/pypi/Pycluster

Twitter Rest Api:

https://dev.twitter.com/docs/api/1.1

Duly Completed Checklist

a) Is the Cover page in proper format? Y
b) Is the Title page in proper format? Y
c) Is the Certificate from the Supervisor in proper format? Has it been signed? Y
d) Is Abstract included in the Report? Is it properly written? Y
e) Does the Table of Contents page include chapter page numbers? Y
f) Does the Report contain a summary of the literature survey? NA
    i. Are the Pages numbered properly? Y
    ii. Are the Figures numbered properly? Y
    iii. Are the Tables numbered properly? Y
    iv. Are the Captions for the Figures and Tables proper? Y
    v. Are the Appendices numbered? Y
g) Does the Report have Conclusion / Recommendations of the work? Y
h) Are References/Bibliography given in the Report? Y
i) Have the References been cited in the Report? Y
j) Is the citation of References / Bibliography in proper format? Y

The above Checklist has been verified

Signature of the Student
Nischal HP
2010HS70022

Signature of the Supervisor
Dr.Ashok Veilumuthu
Senior Researcher,
SAP Labs, Bangalore.

Place: Bangalore
Date: