B.Tech Major Project Progress Report

Happiness Index in Social Network (Social Network Analysis)

Under Supervision of :

Prof. Bhaskar Biswas
Department of Computer Science & Engineering
IIT(BHU)-Varanasi

By-

Swayam Bhardwaj(10100EN020)

Syed Wali Hamza(10100EN007)

Introduction

We aim at studying the happiness index pertaining to social networking sites likes of facebook , twitter , linkedin , google+ etc.

Here by happiness index we mean the happiness quotient of people's status or comments. As is often said “happiness is contagious ” and this becomes more true in case of social networking sites.

So we try to calculate degree of happiness of a person who's status , comment or tweet we are analysing.We just don't group people as belonging to either a happy class or sad class but rather we calculate their happiness index on some scale. The scale is yet to be determined but we are pretty sure that this is how we are going to do it.

Now the problem with analysing social networking sites is that people tend to deviate alot from the conventional language.Thus semantic analysis is quite difficult. Also the elements of puns , sarcasams etc tend to make this processing of strings more difficult.

In this midterm project report we'll baiscally cover following things:

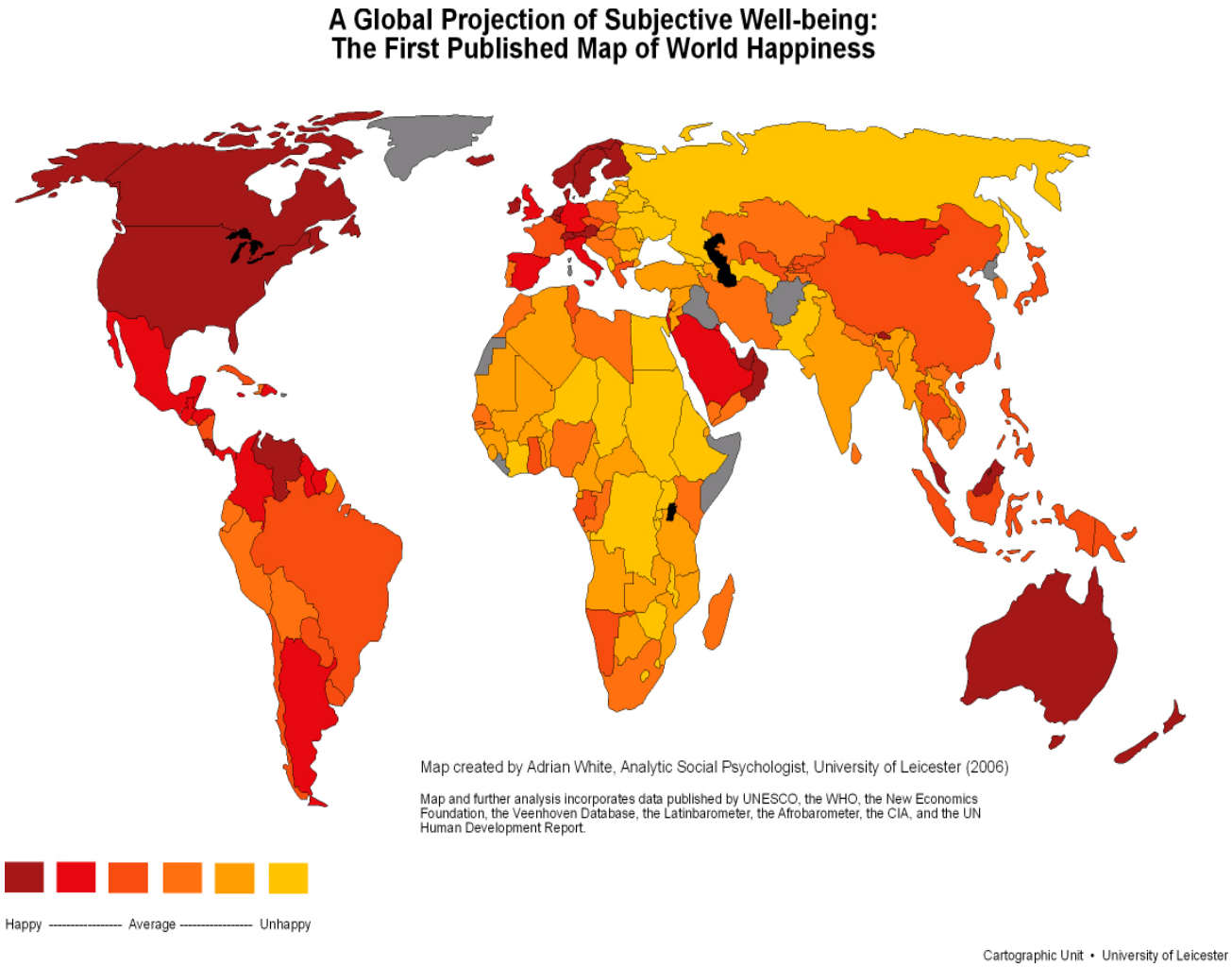
1. Scope of our project
2. Technologies used(to be used)
3. Challenges faced
4. Next phase plan

Scope

- Calculate happiness index based on different demographies.
- To calculate happiness pertaining to some event.
- To calculate happiness of a person .
- To predict mood swings and bipolar behaviors and analyze the duration swings.
- To predict whether a person whether a person is going through a phase of depression by analysing his status or comments.
- To determine how big an event was and how long its effect lasted.
- Calculating trending topics over social networking sites.
- Predict outcomes of some event like assembly poles or olympics or some random sports event.

However , since we will be concetrating mainly on happiness index but the approach which we will be following can be easily generalized for a lot of other similar fields and concepts.

An example of people's happiness across the globe generated by some NGO :



Technologies Used :

Python:

We started with Java and PHP but realized that using either of them for natural language processing is a bit difficult. Also the interfaces(end-points) of sites like twitter or facebook have python based libraries for extarcting tweets/posts and other data. Hence though we decided to learn python and apply it in our quest of finding happiness index of the world through social networking sites.

About Python -

Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C. The language provides constructs intended to enable clear programs on both a small and large scale.
Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library.

NLTK:

Python has a well developed and extensively used library called Natural language tool kit which helps in natural language processing . Though we haven't started implementing nltk based functions in our project yet but in the next phase we'll be using them extensively.

About NLTK-

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language. NLTK includes graphical demonstrations and sample data. It is accompanied by a book that explains the underlying concepts behind the language processing tasks supported by the toolkit,plus a cookbook.

NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning.NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems.

MySql :

We are maintaining a database called happiness-index having different tables primarily for storing tweets and raw data . We are using MySql as our DBMS.

About MySql:

Mysql is an open source RDBMS. It is one of the most used relational database management system used across the globe.

The default port of Mysql is 3306. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation.

MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack (and other 'AMP' stacks). LAMP is an acronym for "Linux, Apache, MySQL, Perl/PHP/Python." Free-software-open source projects that require a full-featured database management system often use MySQL.

Tweepy :

To get hold of the tweets we are using an interface library provied by twitter called tweepy. Its written in python we have written our script using this library to extract tweets from the twitter's stream and save it into our database.

About Tweepy:

The API class provides access to the entire twitter RESTful API methods. Each method can accept various parameters and return responses.

When we invoke an API method most of the time returned back to us will be a Tweepy model class instance. This will contain the data returned from Twitter which we can then use inside our application.

The API class contains all the methods for access the Twitter API.

RESTful services Representational State Transfer is an architectural style consisting of a coordinated set of constraints applied to components, connectors, and data elements, within a distributed hypermedia system to achieve desired architectural properties. REST ignores the details of component implementation and protocol syntax in order to focus on the roles of components, the constraints upon their interaction with other components, and their interpretation of significant data elements

Progress :

We have understood how python works and have become comfortable working in it. Now we are able to get hold of tweets from the public stream of twitter from across the world and save it into our db. We are also able to filter contents , tweets and people based on various factors like nationality , language , region , timezone . Also tweets can be filtered based on keyword . For instance if we want only those tweets which are from India and contain the keyword happy , we are able to do it. So this makes our analysis quite easy and localized in our are of interest.

Database schema :

```
mysql> show tables;
+---------------------------+
| Tables_in_happiness_index |
+---------------------------+
| Tweets                    |
+---------------------------+
1 row in set (0.00 sec)

mysql> desc Tweets;
+--------------+---------------+------+-----+---------+----------------+
| Field        | Type          | Null | Key | Default | Extra          |
+--------------+---------------+------+-----+---------+----------------+
| ID           | int(20)       | NO   | PRI | NULL    | auto_increment |
| USER_ID      | varchar(255)  | YES  |     | NULL    |                |
| USER_NAME    | varchar(255)  | YES  |     | NULL    |                |
| TWEET_STRING | varchar(3000) | YES  |     | NULL    |                |
| SCREEN_NAME  | varchar(255)  | YES  |     | NULL    |                |
+--------------+---------------+------+-----+---------+----------------+
5 rows in set (0.07 sec)
```

Script to extract tweets and publish to db :

```
import sys
import tweepy
import json
import MySQLdb as mdb
consumer_key="oNeZSSGb1bvZD156W3Cv9A"
consumer_secret="UApZt4Q4rufTsDmwaCfvQzDvhzF1nBOHmyB2UHbRxA"
access_key = "142917712-mgmfkAmPf9e5WQfYSsQfn0eCqhIAmDyorQNUYIBF"
access_secret = "IaMmeioBZNnEgllG4IHPd1evjaYzhRzhRq4Jv33v0"

con = mdb.connect('localhost' , 'root' , 'admin' , 'happiness_index');
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth)

class CustomStreamListener(tweepy.StreamListener):
    def on_status(self, status):
        print status.text

    def on_error(self, status_code):
        print >> sys.stderr, 'Encountered error with status code:',
```

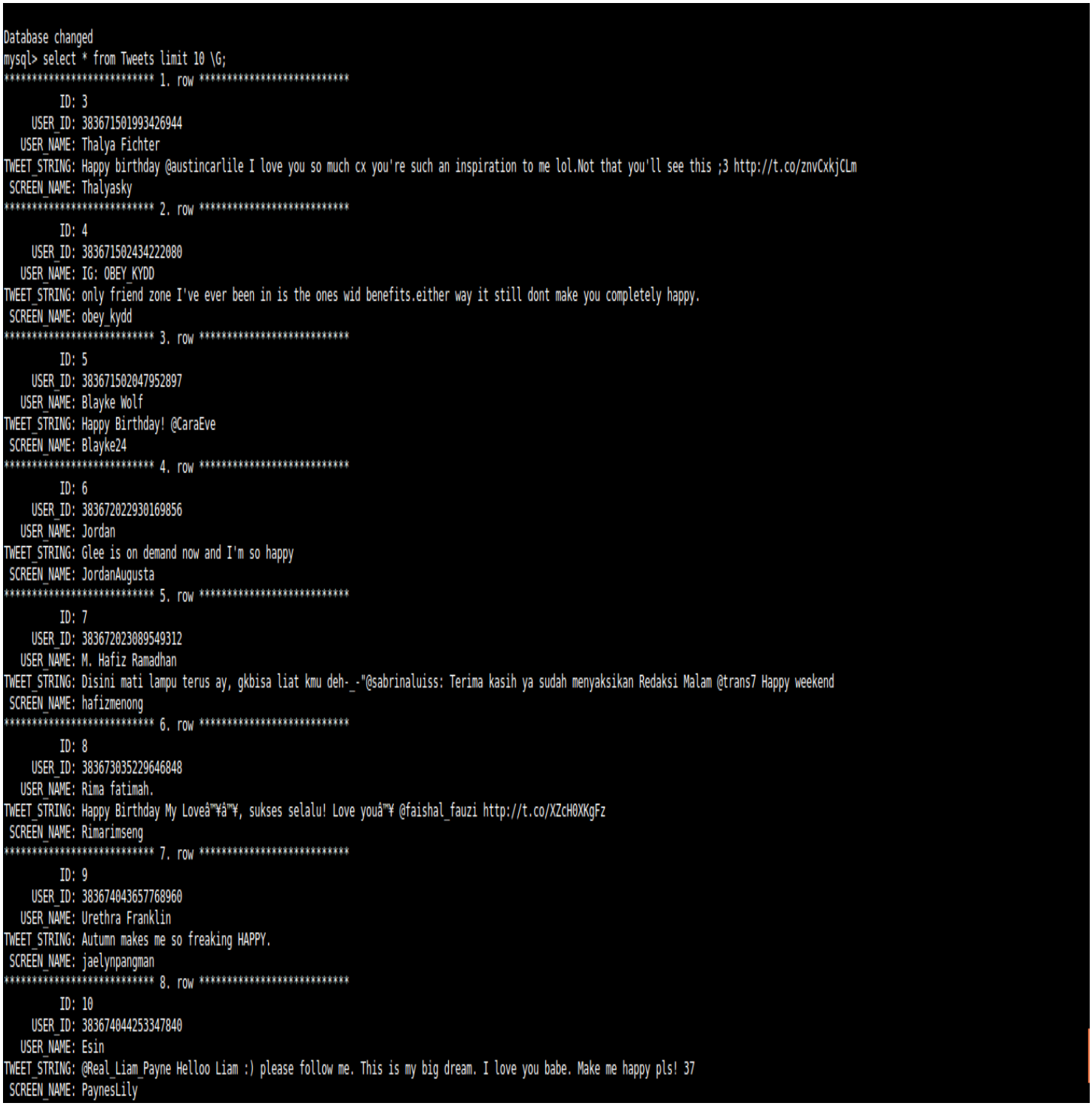
```
status_code
        return True # Don't kill the stream

    def on_timeout(self):
        print >> sys.stderr, 'Timeout...'
        return True # Don't kill the stream


    def on_data(self, data):
        print 'start'
        tweets = json.loads(data)
        print tweets
        print '----------------------------------------------'
        with con:
            cur =con.cursor()
            cur.execute("INSERT INTO Tweets(USER_ID ,
USER_NAME , TWEET_STRING , SCREEN_NAME) VALUES
(%s , %s , %s , %s)" , (tweets['id'] , tweets['user']
['name'].encode('utf-8') , tweets['text'].encode('utf-8') ,
tweets['user']['screen_name'].encode('utf-8')))
        print tweets['text']
        print tweets['id']
        print tweets['user']['screen_name']
        print tweets['user']['name']
        print 'success'


sapi = tweepy.streaming.Stream(auth, CustomStreamListener())
sapi.filter(track=['happy'])
```

Result screenshot :



```
Database changed
mysql> select * from Tweets limit 10 \G;
*************************** 1. row ***************************
        ID: 3
    USER_ID: 383671501993426944
  USER_NAME: Thalya Fichter
TWEET_STRING: Happy birthday @austincarlile I love you so much cx you're such an inspiration to me lol.Not that you'll see this ;3 http://t.co/znvCxkjCLm
 SCREEN_NAME: Thalyasky
*************************** 2. row ***************************
        ID: 4
    USER_ID: 383671502434222080
  USER_NAME: IG: OBEY_KYDD
TWEET_STRING: only friend zone I've ever been in is the ones wid benefits.either way it still dont make you completely happy.
 SCREEN_NAME: obey_kydd
*************************** 3. row ***************************
        ID: 5
    USER_ID: 383671502047952897
  USER_NAME: Blayke Wolf
TWEET_STRING: Happy Birthday! @CaraEve
 SCREEN_NAME: Blayke24
*************************** 4. row ***************************
        ID: 6
    USER_ID: 383672022930169856
  USER_NAME: Jordan
TWEET_STRING: Glee is on demand now and I'm so happy
 SCREEN_NAME: JordanAugusta
*************************** 5. row ***************************
        ID: 7
    USER_ID: 383672023089549312
  USER_NAME: M. Hafiz Ramadhan
TWEET_STRING: Disini mati lampu terus ay, gkbisa liat kmu deh-_-"@sabrinaluiss: Terima kasih ya sudah menyaksikan Redaksi Malam @trans7 Happy weekend
 SCREEN_NAME: hafizmenong
*************************** 6. row ***************************
        ID: 8
    USER_ID: 383673035229646848
  USER_NAME: Rima fatimah.
TWEET_STRING: Happy Birthday My Loveâ™¥â™¥, sukses selalu! Love youâ™¥ @faishal_fauzi http://t.co/XZcH0XKgFz
 SCREEN_NAME: Rimarimseng
*************************** 7. row ***************************
        ID: 9
    USER_ID: 383674043657768960
  USER_NAME: Urethra Franklin
TWEET_STRING: Autumn makes me so freaking HAPPY.
 SCREEN_NAME: jaelynpangman
*************************** 8. row ***************************
        ID: 10
    USER_ID: 383674044253347840
  USER_NAME: Esin
TWEET_STRING: @Real_Liam_Payne Helloo Liam :) please follow me. This is my big dream. I love you babe. Make me happy pls! 37
 SCREEN_NAME: PaynesLily
```

Difficulties faced :

First and foremost requirement was to get the working data in our db. We tried to build a php curl script to do so but no networking site would give access . All of them likes of twitter or facebook require that the script be authenticated. Hence we decided to make a crawl to extract the tweets but that also failed becuase or rate limiting algorithms implemented on most of such sites . After researching a bit we came across the twitter much celebrated public stream end point GET statuses/sample . So we used tweepy a library in python to help us in getting access to the tweets.

Next work phase involved writing the scripts and configuring our db .

Also we implemented filters to filter out only those kind of tweets that we are interested in.

1. Next problem was converting the format of tweets to utf-8. Because we are actually hitting on the twitter's public stream which gets tweets from any language in the world. ASCII doesn't support thus we did of brainstorming as why we are getting errors and finally we conculded this as being the source.

So now we are equipped with python , are already through the most difficult part of collecting workable data and next phase would invlove choosing right algorithm to process them.

Next phase time line (20th Dec onwards):

1. Apply first round of processing on the raw data.
2. Study relevant papers regarding NLP and social networking .
3. Build our happiness index on top of the raw data.
4. Apply NLP based approaches in other domains.