

DATA 620 – FINAL PROJECT

ANALYSIS OF REDDIT FEEDBACK DATA

BY RAJAGOPAL SRINIVASAN & MUTHUKUMAR SRINIVASAN

1)

Description

Reddit (/ˈrɛdɪt/) is an American social news aggregation, web content rating, and discussion website. Registered members submit content to the site such as links, text posts, and images, which are then voted up or down by other members. Posts are organized by subject into boards called "subreddits", which cover a variety of topics including news, science, movies, video games, music, books, fitness, food, and image-sharing. Submissions with more up-votes appear towards the top of their subreddit and, if they receive enough votes, ultimately on the site's front page. Despite strict rules prohibiting harassment, Reddit's administrators spend considerable resources on moderating the site.

As of 2017, Reddit had 542 million monthly visitors (234 million unique users), ranking as the #4 most visited website in U.S. and #8 in the world. Across 2015, Reddit saw 82.54 billion pageviews, 73.15 million submissions, 725.85 million comments, and 6.89 billion upvotes from its users.

Reddit was founded by University of Virginia roommates Steve Huffman and Alexis Ohanian in 2005. Condé Nast Publications acquired the site in October 2006. Reddit became a direct subsidiary of Condé Nast's parent company, Advance Publications, in September 2011. As of August 2012, Reddit operates as an independent entity, although Advance is still its largest shareholder. Reddit is based in San Francisco, California. In October 2014, Reddit raised \$50 million in a funding round led by Sam Altman and including investors Marc Andreessen, Peter Thiel, Ron Conway, Snoop Dogg, and Jared Leto. Their investment saw the company valued at \$500 million at the time. In July 2017, Reddit raised an additional round of \$200 million at a \$1.8 billion valuation, with Advance Publications remaining the majority stakeholder.

2)

URL

<https://en.wikipedia.org/wiki/Reddit>

3)

Data Source

Source 1

https://www.reddit.com/r/datasets/comments/3bxl7/i_have_every_publicly_available_reddit_comment/

DATA 620 – FINAL PROJECT

ANALYSIS OF REDDIT FEEDBACK DATA

BY RAJAGOPAL SRINIVASAN & MUTHUKUMAR SRINIVASAN

Source 2:

A smaller subset of some comments from May 2015 has been posted on Kaggle:

<https://www.kaggle.com/reddit/reddit-comments-may-2015> which is in Jason Format

Source 3

Reddit also provide API to access the data and is given below. Reddit API available to access the data:

<https://www.reddit.com/dev/api/>

4)

CONVERT JSON DATA TO MYSQL

Converted the JSON data through Ipython. We ran into several issues. one of them is permission issue. We have given Everyone access in Windows. But still, we could not do. Windows, we have this challenge. So we have gone ahead with (5) whereby we wrote a program to read and load it in MySQL

```
In [2]: import json

In [5]: cd "C:\temp\MSData\DATA-620\FinalProject\114"
C:\temp\MSData\DATA-620\FinalProject\114

In [10]: filedir="C:/temp/MSData/DATA-620/FinalProject/114/"

In [13]: json_data=open(filedir).read()

-----
PermissionError                                Traceback (most recent call last)
<ipython-input-13-a765c9bcc52e> in <module>()
----> 1 json_data=open(filedir).read()

PermissionError: [Errno 13] Permission denied: 'C:/temp/MSData/DATA-620/FinalProject/114/'
```

5)

CONVERT JSON DATA TO MYSQL

How did we convert the JSON data to MySQL Database?

1. We downloaded the data from Kaggle first.
2. Then unzipped it in our local drive.
3. Setup a small XAMPP webserver.
4. Then created a MySQL Database
5. Then we wrote a small PHP program to read JSON file and insert into MYSQL Database.
6. Have done few manipulations to get Edges, users and comments Only.

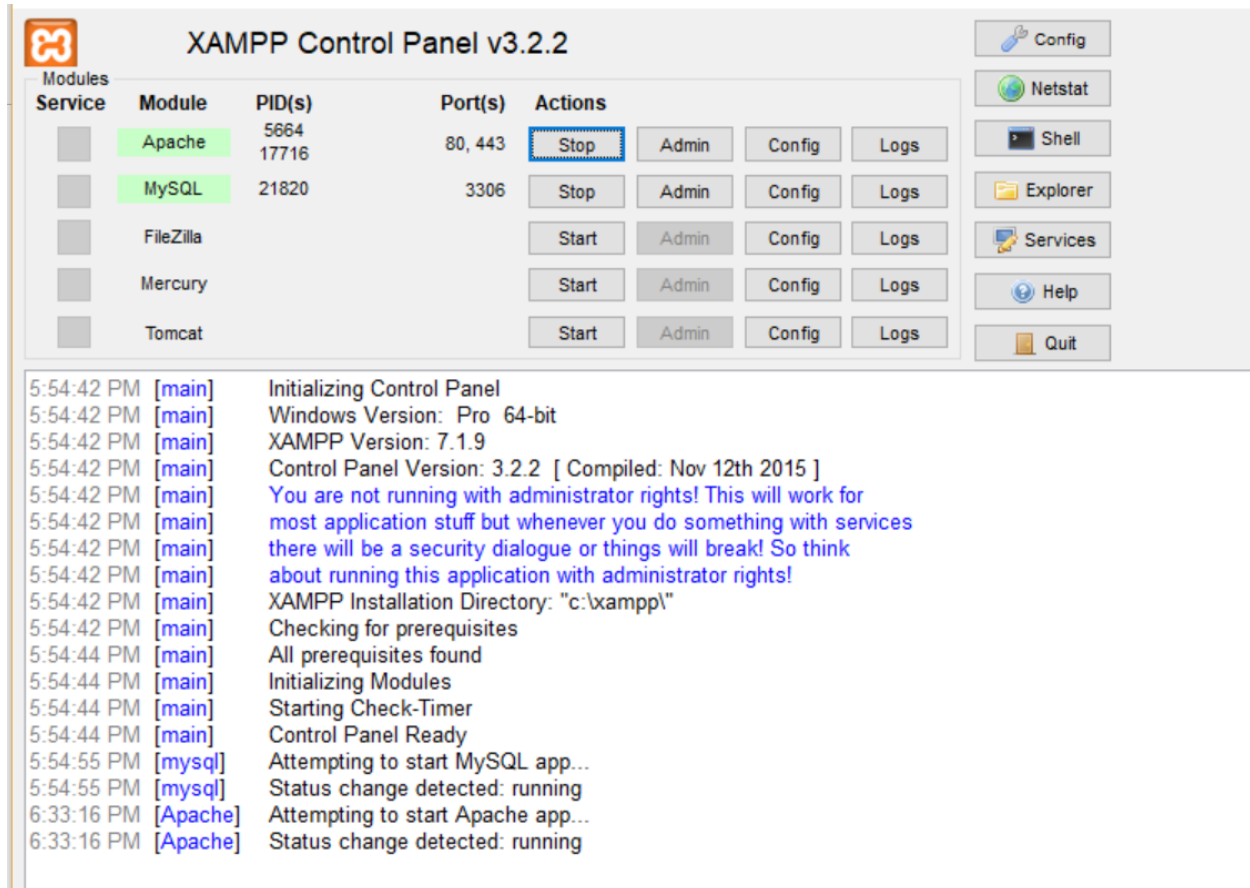
<http://www.kodingmadesimple.com/2014/12/how-to-insert-json-data-into-mysql-php.html>

DATA 620 – FINAL PROJECT

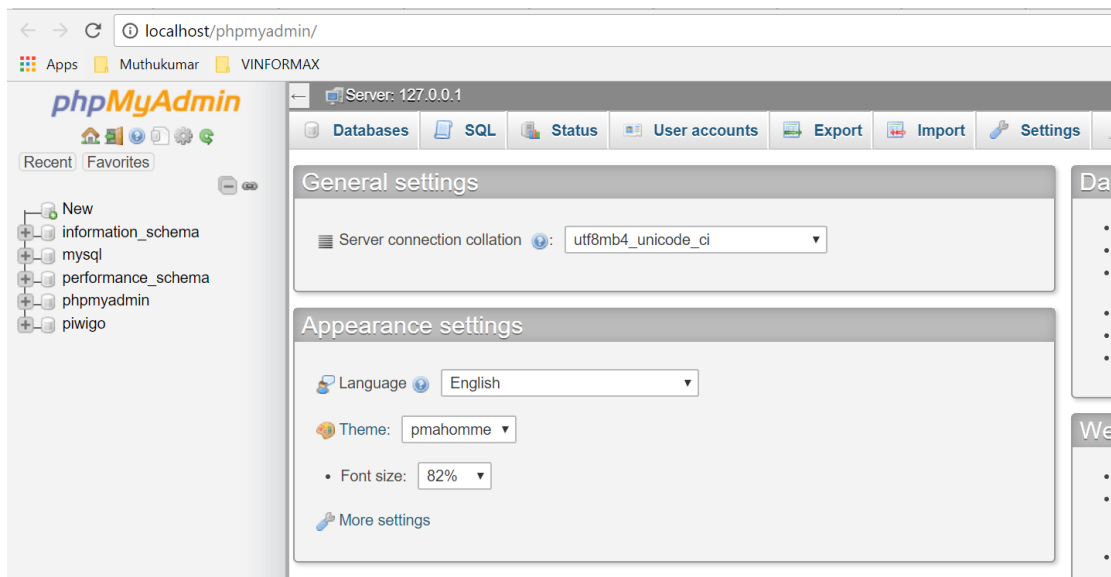
ANALYSIS OF REDDIT FEEDBACK DATA

BY RAJAGOPAL SRINIVASAN & MUTHUKUMAR SRINIVASAN

Our XAMPP version screen shot is given below



Our Phpmyadmin is running in Localhost and screen shot is below



DATA 620 – FINAL PROJECT

ANALYSIS OF REDDIT FEEDBACK DATA

BY RAJAGOPAL SRINIVASAN & MUTHUKUMAR SRINIVASAN

6)

DATA TRANSFER FROM LAPTOP

For Individual work, we have first loaded the data into Muthukumar laptop XAMPP server. then migrated and transferred to Rajagopal laptop. Please see below

1 schema transferred.

=====
Schema: reddit
=====

Tables: 3
Views: 0
Routines: 0

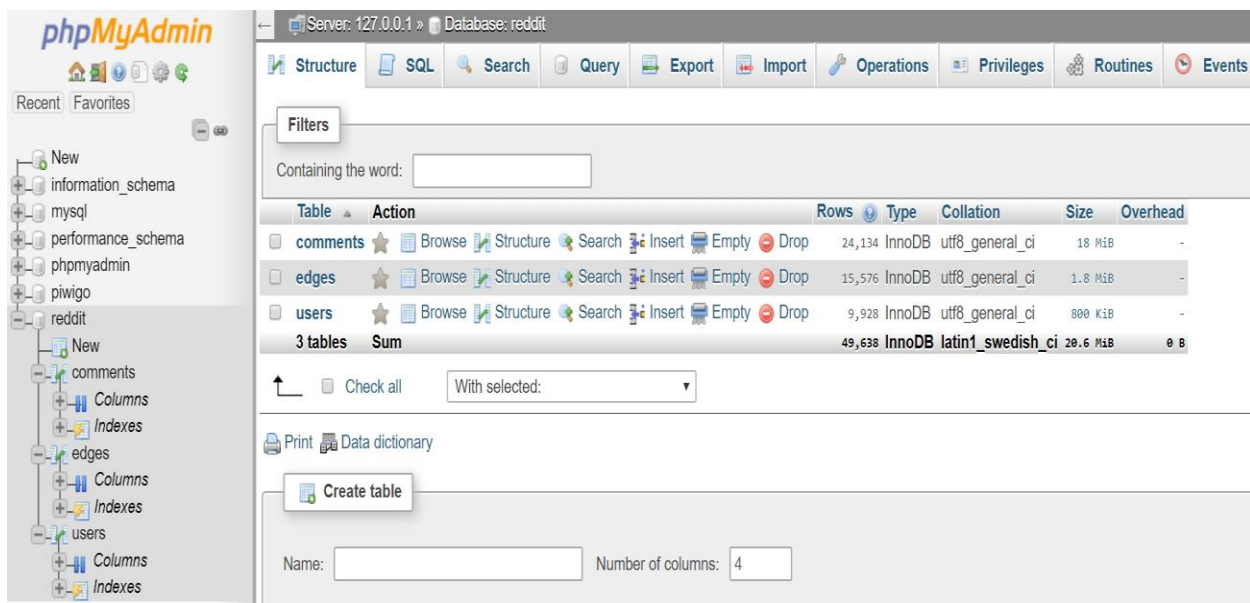
Data copy report:

Succeeded: copied 15576 of 15576 rows from `reddit`.`edges`
Succeeded: copied 9928 of 9928 rows from `reddit`.`users`
Succeeded: copied 24134 of 24134 rows from `reddit`.`comments`

7)

DATA READY

a) Database is ready. Please see the screen shot below which is taken from localhost/phpmyadmin



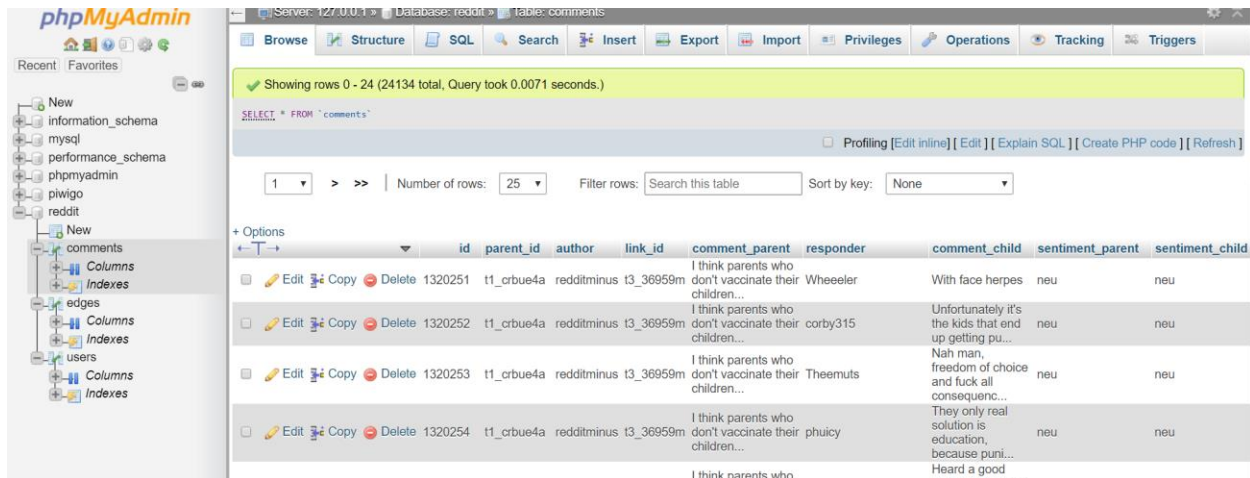
b) Tables are ready: see the screen shot below

DATA 620 – FINAL PROJECT

ANALYSIS OF REDDIT FEEDBACK DATA

BY RAJAGOPAL SRINIVASAN & MUTHUKUMAR SRINIVASAN

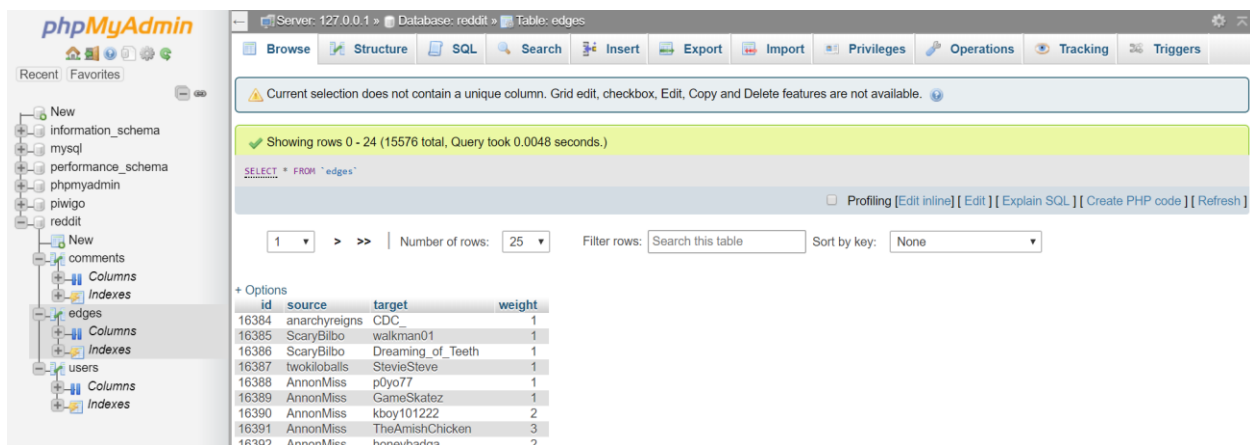
Comments table



The screenshot shows the phpMyAdmin interface for the 'comments' table in the 'reddit' database. The table has 9 columns: id, parent_id, author, link_id, comment_parent, responder, comment_child, sentiment_parent, and sentiment_child. The data shows comments on a post about vaccination, with sentiment values of 'neu' (neutral).

	id	parent_id	author	link_id	comment_parent	responder	comment_child	sentiment_parent	sentiment_child
	1320251	t1_crue4a	redditminus	t3_36959m	I think parents who don't vaccinate their children...	Wheeler	With face herpes	neu	neu
	1320252	t1_crue4a	redditminus	t3_36959m	I think parents who don't vaccinate their children...	corby315	Unfortunately it's the kids that end up getting pu...	neu	neu
	1320253	t1_crue4a	redditminus	t3_36959m	I think parents who don't vaccinate their children...	Theemuts	Nah man, freedom of choice and fuck all consequenc...	neu	neu
	1320254	t1_crue4a	redditminus	t3_36959m	I think parents who don't vaccinate their phucy children...		They only real solution is education, because puni...	neu	neu
					I think parents who		Heard a good		

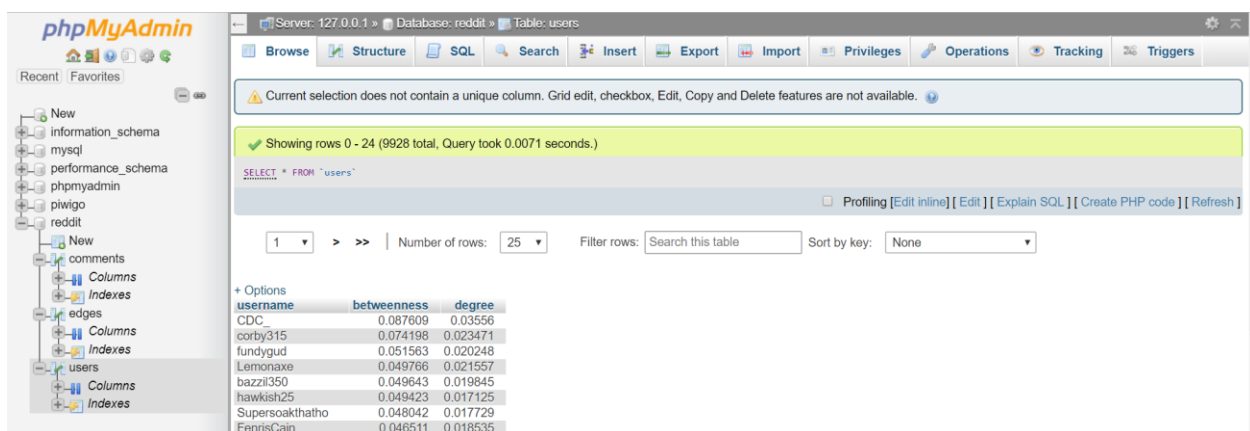
Edges table



The screenshot shows the phpMyAdmin interface for the 'edges' table in the 'reddit' database. The table has 4 columns: id, source, target, and weight. The data shows connections between users and entities like CDC, walkman01, and Dreaming_of_Teeth.

	id	source	target	weight
	16384	anarchyreigns	CDC	1
	16385	ScaryBilbo	walkman01	1
	16386	ScaryBilbo	Dreaming_of_Teeth	1
	16387	twokkibalis	StevieSteve	1
	16388	AnnonMiss	p0y077	1
	16389	AnnonMiss	GameSkatez	1
	16390	AnnonMiss	kboy101222	2
	16391	AnnonMiss	TheAmishChicken	3
	16392	AnnonMiss	honevbadra	2

Users table



The screenshot shows the phpMyAdmin interface for the 'users' table in the 'reddit' database. The table has 3 columns: username, betweenness, and degree. The data shows user statistics for various usernames.

	username	betweenness	degree
	CDC	0.087609	0.03556
	corby315	0.074198	0.023471
	fundygud	0.051563	0.020248
	Lemonaxe	0.049766	0.021557
	bazzi350	0.049643	0.019845
	hawkish25	0.049423	0.017125
	Supersoakthatho	0.048042	0.017729
	FennisCain	0.046511	0.018535

DATA 620 – FINAL PROJECT

ANALYSIS OF REDDIT FEEDBACK DATA

BY RAJAGOPAL SRINIVASAN & MUTHUKUMAR SRINIVASAN

8)

ISSUES FACED – Initial Setup

- a) We imported pymysql in ipython. The screen error saying that there is no package. So we installed pymysql. Please see below. This package will help to connect to MySQL Database from Python

Microsoft Windows [Version 10.0.15063]

(c) 2017 Microsoft Corporation. All rights reserved.

C:\Users\Muthukumar>pip install pymysql

Collecting pymysql

Downloading PyMySQL-0.7.11-py2.py3-none-any.whl (78kB)

100% |#####| 81kB 975kB/s

Installing collected packages: pymysql

Successfully installed pymysql-0.7.11

C:\Users\Muthukumar>activate python36

(python36) C:\Users\Muthukumar>pip install pymysql

Collecting pymysql

Using cached PyMySQL-0.7.11-py2.py3-none-any.whl

Installing collected packages: pymysql

Successfully installed pymysql-0.7.11

(python36) C:\Users\Muthukumar>

9)

DATA IMPORT

```
In [5]: reddit = pd.read_sql('SELECT DISTINCT author, responder, COUNT(1) as weight FROM comments '
      'GROUP BY author, responder '
      'HAVING author<>"[deleted]" AND responder<>"[deleted]";', con=sql_con)

reddit_comment = pd.read_sql('SELECT * from comments;', con=sql_con)

# Import users with centrality measures previously calculated
# Just so we don't have to wait for them every time
users = pd.read_sql('SELECT username, betweenness, degree FROM users', con=sql_con)

sql_con.close()
```

```
In [6]: # Data test
      reddit.loc[0:5]
```

```
Out[6]:
```

	author	responder	weight
0	---DevilsAdvocate---	dannybtw	1
1	---DevilsAdvocate---	ThisFreaknGuy	2
2	-Covariance	DomeSlave	1
3	-Mountain-King-	ASK_ABOUT_STEELBEAMS	3
4	-Mountain-King-	AvoldNoiderman	2
5	-Mountain-King-	Bigfrostynugs	1

DATA 620 – FINAL PROJECT

ANALYSIS OF REDDIT FEEDBACK DATA

BY RAJAGOPAL SRINIVASAN & MUTHUKUMAR SRINIVASAN

10)

HOW TO BUILD NETWORK GRAPH

Nodes	Reddit users
Edges	Direct responses by one user to another user's comment
Edge Weights	Are number of responses between two nodes/users
Graph	The graph is undirected. Response to users means, communication between two users.

11)

ISSUES FACED – During Program Setup

```
In [12]: # Initialize graph
G = nx.Graph()

-----
NameError                                Traceback (most recent call last)
<ipython-input-12-41e2f12cd528> in <module>()
      1 # Initialize graph
----> 2 G = nx.Graph()

NameError: name 'nx' is not defined
```

The above is due to not importing the respective modules. This has been resolved by importing the right modules at the beginning of the program

12)

ANALYSIS OF A SPECIFIC USER

With close to 10,000 nodes the graph is too busy to decipher. We can note that even though there is a large number of long tentacles, the maximum distances between any two nodes is 16. There is a lot of interconnectedness in the hairball center.

High degree centrality identifies users that generate a lot of responses. Betweenness centrality may identify users who may not necessarily generate a lot of responses, but that start active discussions and, therefore, connect various clusters together.

Let's look at one user. Bt1222 is in the top 20 lists for degree centrality, but he or she just misses the top 20 list for betweenness centrality. Let us also measure similar for two more users.

User 2: tehweave

User 3: NicoUK

In Our Laptop, the processor was taking too long time to respond. We left it for more than 12 hours. still processing. See below the Kernel being very Busy.

DATA 620 – FINAL PROJECT

ANALYSIS OF REDDIT FEEDBACK DATA

BY RAJAGOPAL SRINIVASAN & MUTHUKUMAR SRINIVASAN

The screenshot displays the JupyterLab interface. At the top, there's a 'jupyter' logo and a 'Logout' button. Below this, a tab bar shows 'Files', 'Running', and 'Clusters'. A message says 'Select items to perform actions on them.' with 'Upload', 'New', and a refresh icon. The file browser shows a list of files: 'DATA620-FinalProject.ipynb' (Running, seconds ago), 'HowDidIgettheData.rtf' (4 hours ago), 'ref.txt' (a day ago), 'Week14-Finalproject.docx' (13 minutes ago), and 'Week14-FinalProjectProposal.rtf' (10 hours ago). Below the file browser, the JupyterLab header shows 'jupyter DATA620-FinalProject' and 'Last Checkpoint: 17 minutes ago (autosaved)'. The main area is a code editor with a menu bar (File, Edit, View, Insert, Cell, Kernel, Help) and a toolbar. The code in the editor is:

```
fig = plt.figure(figsize=(18, 18))

main = 'NicoUK'

# Select neighbors and neighbors of neighbors
subG = G.neighbors(main)
n_list = subG
```

On the right side of the code editor, there's a 'Trusted' button, a 'Python 3' button, and a 'Kernel Busy' button.

On right hand side, one can see Kernel showing as Busy.

13)

SENTIMENTAL ANALYSIS

Please see the ipython output

14)

CONCLUSIONS

In this project, we have determined that there is no correlation between sentiment scores and betweenness & degree centralities. It seems that sentiment of user comments, whether positive or negative, does not necessarily influence in how the discussion progresses. Interactions between users develop similarly whether posts are deemed negative or positive. It is important to note that our analysis is limited and illustrates just one side of the subject.

15)

CHALLENGES FACED

1. Windows version – there are lot of limitations.
2. Laptop is Windows 10 with 2 Core processor, 8 GB RAM – there is only so much we can do.
3. Used methodology to create a MYSQL Data from JSON format. We lacked in knowledge of JSON and PHP. So, it took long time to learn and develop a program.

DATA 620 – FINAL PROJECT
ANALYSIS OF REDDIT FEEDBACK DATA
BY RAJAGOPAL SRINIVASAN & MUTHUKUMAR SRINIVASAN

16)

OTHER REFERENCES

<http://www.kodingmadesimple.com/2014/12/how-to-insert-json-data-into-mysql-php.html>

<https://www.kaggle.com/reddit/reddit-comments-may-2015>

THANK YOU

BY

MUTHUKUMAR SRINIVASAN

(HEAD OF TECHNOLOGY/ ENTERPRISE ARCHITECT, VINFORMAX & KOHLS)

RAJAGOPAL SRINIVASAN

(HEAD OF ENTERPRISE ARCHITECTURE, TATA CONSULTANCY SERVICES)