Level 4 Network Analysis

Simple tips

- When you develop all components program work with small amount of data
  - Make sure the programs work and test
- Download a version of mongodb
  - Develop your programs on local db
- When you work with 1 hour data
  - Work locally if you can
  - If your system cant handle use school systems

School system

- Logon to sibu
    - Run your python code on sibu
- Mongodb is hosted on another server called marcus
- Use the code I gave
- It should work

mongodb

```
• client = pymongo.MongoClient('127.0.0.1',27017) #is assigned local port
•     print(client.list_database_names())
•     db = client.twitterStream
•     db1 = client.twitterDump
•     dbl = client.Logs
•     db3 = client.invertedIndex

•     print db.collection.stats()

• for item in db[c].find():
•             print(item)
```

Task 1

- [Total 15 marks] Develop a crawler to access as much Twitter data as possible.
    - [5 marks] Use Twitter streaming API for collecting 1% data.
    - [10 marks] Enhance the crawling using the hybrid architecture of Twitter Streaming & REST APIs.
        - For example, 1% streaming (provide justification for why you chose certain words or user to follow).
        - Keyword based and/or user based REST probes.

How to do?

- Run 1% streaming for an hour
    - If you can identify important topics happening
        - Then use these for REST probes
    - If you can identify power user, who are tweeting more often
        - Use them in REST probes
- OR
- Run 1% streaming for an hour
- Analyse the text through some clustering/grouping
- REST probes
    - Use busty keywords/topics etc. in REST probes
    - Use power users in REST probes
    - Fix the time period for the 1% sample
        - You can specify the time period in which you need tweets in REST API

Remove duplicates

- If your mongodb uses
- tweet = {'_id' : tweet_id, 'date': created, 'username': username, 'text' : text}
- It will index on _id
- This means if you try to put another tweet with the same tweeted
  - It will generate an error
  - Count these errors and you get number of duplicates removed

Task 2

- [25 marks] Grouping of tweets: Group the tweets based on content analysis, You can collect the data and then cluster them using any off-the shelf software; or use any locality sensitive hashing software; or build a content index and group them
  - Describe your method for grouping [5 marks]
    - Extract important usernames; hashtags and entities/concepts from the group [10 marks]
    - Provide statistics on data and the resulting groups [10]

How to do?

- Use any clustering algorithm
  - https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
  - Any of your choice
- Clusters won’t be accurate
  - That is expected
- If you are using K-Means
  - Try roughly 10% as k,
  - Ideally you should try multiple values of k and choose one with suitable clusters
- Just use tweeted and text for clustering
  - We are grouping semantically similar documents
- If your system can’t use all the text,
  - Partition them into 2 files

Task 3

- 25 marks] Capturing & Organising User and hashtag information
  - [15 marks] Develop a method to capture user interaction graph. Users occurring together in general data as well as on the groups. Differentiate between different kinds of networks like retweet network; quote tweets etc.
- User interaction Graph
- G(V,E)
- V is the list of users
  - Those who send the tweets and those mentioned in the tweets, e.g., @bbc?
    - This way you form the edges
- Overall tweet data (a structure like following will help)
- user i -> userj, freq; userk, freq, …..
  - Frequency is how ofetn they occur together
- Repeat this for
  - retweet data;
  - quotes/reply data;….

- [10 marks] Develop a mechanism to capture hashtag information occurring together in general data as well as on the groups. Differentiate between different kinds of networks like retweet network; quote tweets etc.
-
- hashtag-I -> hashtga-j; hashtag-k; …
- Pleaser note we don’t have direction information
  - Just co-occurrence information only

Task 4

- [25 marks] Network Analysis
  - Analyse the data to generate network-based measures like ties, triads.
- tie/link between nodes
  - This means if there is a link between nodes, then there is a tie/link
  - Tie is when two users connect
- As discussed in lecture 4 and 5
- How often a a link/tie is formed, when a user retweets, quotes, replies. Use the structure created above to generate these data
- Repeat same analysis for two
  - Overall data
  - Clustered data

Triad:

- Triad: a group of 3 users – node i, j and u forming a path of length 2 (i. e, node i is connected to node j; and node j is connected to node u); when node u is also connected to node i then the path is closed; forming a loop of length 3 or a triangle.
- Triad
  - A->B; B-> C;
  - A->B; B-> C; C->A
  - A<->B; B-> C; C->A
  - A<->B; B<-> C; C->A
  - A->B; B<-> C; C<->A
  - A->B; B<-> C;

- Compare and contrast between general data (ungrouped) and also from the grouped data (that is tweets grouped by clustering or hashing; section 2)
- Repeat same analysis for two
  - Overall data
  - Clustered data

Report

- [10 marks] Report – Organisation, completeness, literate,