Course Work – Level 4
Social Media Analytics

Total Marks – 100 marks & Weightage – 20%
Course work deadline – Monday 9th March 2020 4:30PM

INTRODUCTION

This is an individual assignment!

The objective of this course work is to develop a Twitter crawler for data collection in English and to conduct social media analytics. You should use Python programming language and also MongoDB for data storage. It is very important that students provide working version of the software, as we need to validate them.

Students submit their **sample data**, **code and report** on or before the specified deadline. In addition, students provide a sample of data set.

Code and Sample data should be provided on the github. Report submission is through the Moodle page for the Web Science course.

The coursework will be marked out of 100. Course work will have 20% weight of the final marks. As the usual practice across the school, numerical marks will be appropriately converted into bands. Final written exam will have 80% weightage, which will be in April/May 2020.

Specific tasks to do

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

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
    a. Describe your method for grouping [5 marks]
        i. Extract important usernames; hashtags and entities/concepts from the group [10 marks]
        ii. Provide statistics on data and the resulting groups [10]

3. [25 marks] Capturing & Organising User and hashtag information

a. [15 marks] Develop a method to capture user information. Users occurring together in general data as well as on the groups. Differentiate between different kinds of networks like retweet network; quote tweets etc.
b. [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.

4. [25 marks] Network Analysis
    a. Analyse the data to generate network-based measures like ties, triads. Definition to follow in next lecture.
    When a user retweets, quotes, replies a link/tie is formed.
    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.

5. [10 marks] Report – Organisation, completeness, literate,

Report structure

Report should be organised the following way:

0: Front page
Title –
Student Name & Matriculation Number
Source code – github information
Data – github information for sample data

1. Section 1: Introduction
    a. Describe the software developed with appropriate details; if you have used code from elsewhere please specify it
    b. Specify the time and duration of data collected
    c.
2. Section 2: Data crawl
    a. Use Twitter Streaming API for collecting 1% data
        i. Specify the APIs used
            1. Please do not include entire code here; just main description of the function
            2. Along with a short description/justification
    b. Enhance the crawling using the hybrid architecture of Twitter Streaming & REST APIs
        i. Specify the APIs used
            1. Please do not include entire code here; just main description of the function
            2. Along with a short description/justification

3. 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
    a. Describe your method for grouping & statistics on groups. Provide in a tabular fashion
    Total data; groups; average size of a group; min size; max size etc.

    Describe the method for Username and Hashtag identification. Provide data on a tabular fashion and contrast for entire data and just for grouped data.

4. Describe the method for Capturing & Organising User and hashtag information
    a. Develop a method to capture user information. Users occurring together in general data as well as on the groups. Differentiate between different kinds of networks like retweet network; quote tweets etc. Provide information on data structure used.

    b. 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. Provide information on data structure used.
    Provide tabular data and contrast the data.
5. Network Analysis Information.
    Analyse links; path of 2 and Triads (closed loops) in all data and contrast with groups.

Submission

Deadline: Monday 9th March 2020 4:30PM

What to submit

1) Report as a pdf file. (Please submit this jus for the report link)
2) A github information on data and code
    a. Software (runnable version, readme info, and also properly commented). It is important that software is runnable with minimum effort for the markers.
    b. Data – provide a sample data. Use a CSV format or MongoDB. Importantly your software should be able to run on this sample data, without much hassle.

Where to submit

1) Moodle
2) Code and sample data on github