Course Work – Level 4
Event Detection

Total Marks – 100 marks & Weightage – 20%
Course work Deadline – Monday 12th November 2018 4:30PM

INTRODUCTION

The objective of this course work is to develop a Twitter crawler to collect data (English only). Further to this, perform geo-tagging and conduct basic data analytics. We recommend students to use Python or Java programming languages and also MongoDb for data storage. It is very important that students provide workable version of the software as we need to run them. Use Twitter API for accessing data.

Students submit their **code and report** on or before the specified deadline. In addition, students provide a sample of data set. Submission is through the Moodle page for the Web Science course.

For more details, you need to attend the lecture on Monday, 24th September 2018 (Tuesday 25th for Singapore) & also 1st and 12th October 2018.

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 2019.

Collect data for 1 hour of any day. In addition, collect geo-coded data for Glasgow for the same time period. Singapore students should collect data for Singapore.

Specific tasks to do

1. Develop a crawler to access as much Twitter data as possible (Total 25 marks)
   a. Use streaming API (gardenhose api) for collecting 1% data (5 marks)
   b. Enhance the crawling using Streaming & REST API (10 marks)
      i. For example topic based or user based streaming
      ii. Keyword based/user based REST probes
   c. Grab as much geo-tagged data for Glasgow/Singapore for the same period (5 marks)
   d. Discuss your data access strategies and how did you address Twitter data access restrictions (5 marks)
2. Develop basic data analytics (Total 20 marks)
   a. Count the amount of data collected. Specify amount of geo-tagged data in this data set. (5 marks)

   b. Count the amount of geo-tagged data from Glasgow / Singapore. Measure if there is any overlap with 1% data. (5 marks)
   c. Count redundant data present in the collection (you may end up collecting the same tweets again through various APIs) (5 marks)
   d. Count the re-tweets and quotes (5 marks)

3. Enhance the geo-tagged data (Total marks 30)

The idea is to enhance geo-location information of tweets. Less than 6% tweets have geo-location information. You should group similar tweets, and then assign a tweet location to each member of the group. Any grouping method can be used, however, LSH based approach is a good option. More grouping techniques will be discussed on 12th October 2018 Event detection lecture.

Locality sensitive hashing (LSH) can be used for grouping similar tweets. LSH is an algorithm for grouping similar documents into a single bucket. LSH is a data independent hash method. It is easier to find an item corresponds to a nearest neighbor using traditional approaches like linear search but imagine if the database is big and if the item is complicated, it can lead to more cost and computational time as well. LSH can be used for clustering, nearest neighbor search, detecting near or exact duplicates. LSH is implemented in python using LSHash given by Kayzhu (https://github.com/kayzhu/LSHash.)

Next assign geo-location for each tweet in each of the groups. We will discuss a strategy on 1st October 2018.

   a. Grouping of tweets (10 marks). Provide statistics like how many groups and number of tweets per group. This can be given as a histogram or graph (for example, groups in x-axis and y-axis size of the group. Enrich this information with another graph showing number of geo-tagged tweets and also profile based geo-information.
   b. Geo-location assignment (10 marks). Explain your method and justify it. Enhance the above graph information providing the number of additional tweets with geo-information. Also comment on the number of tweets with no geo-information (you wont be able to assign geo-information if none of the tweets in group have geo-information.
   c. Conduct an evaluation of the method (10 marks). A method will be discussed on Monday 1st October 0218. One option would be take 50%of geo-tagged tweets and assume that they have no geo-information. Assign a geo-information to each of these using your above method. Measure the differences between assigned location and the actual location. Plot them!

4. Open Ended Problem (Total 25 marks)
   a. Develop a crawler for any of the following social media sites (facebook, Instagram, Google Plus, Tumblr, Flickr, any other);
   b. Conduct data analysis similar to Twitter
   c. We are not giving any more specific directions, as we want students to be creative.
   d. Analysis similar to Steps 1 and 2 expected.

Report structure & Mark Distribution

Report should be organised the following way & Mark distribution

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;
2. Section 2: Data crawl
    a. Use streaming API for collecting 1% data (5 marks)
        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 Streaming and REST API (10 marks)
        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

    c. Grab as much geo-tagged data for Glasgow/Singapore for the same period (5 marks)
        i. Specify the APIs used
            1. Please do not include entire code here; just main description of the function
    d. Discuss your data access strategies and how did you address Twitter data access restrictions (5 marks)
        i. Discuss how creative you are in collecting as much data
3. Basic data analytics (Total 20 marks)
    a. Count the amount of data collected (5 marks)
        i. Show a histogram of 10 minutes periods (x –axis duration of 10 minutes – y-axis count)
    b. Count the amount of geo-tagged data from Glasgow / Singapore (5 marks)
        i. Show a histogram of 10 minutes periods (x –axis duration of 10 minutes – y-axis count)
    c. Count redundant data present in the collection (you may end up collecting the same tweets again through various APIs) (5 marks)
        i. Show a histogram of 10 minutes periods (x –axis duration of 10 minutes – y-axis count) for both collected data and redundant data for same period
    d. Count the re-tweets and quotes (5 marks)
        i. Show a histogram of 10 minutes periods (x –axis duration of 10 minutes – y-axis count) for both collected data , re-tweets, quotes
        ii.
4. Enhance the geo-tagged data (Total marks 30)

a. Grouping of tweets (10 marks). Provide statistics like how many groups and number of tweets per group. This can be given as a histogram or graph (for example, groups in x-axis and y-axis size of the group. Enrich this information with another graph showing number of geo-tagged tweets and also profile based geo-information.

b. Geo-location assignment (10 marks). Explain your method and justify it. Enhance the above graph information providing the number of additional tweets with geo-information. Also comment on the number of tweets with no geo-information (you wont be able to assign geo-information if none of the tweets in group have geo-information.

c. Conduct an evaluation of the method (10 marks). A method will be discussed on Monday 1st October 0218. One option would be take 50%of geo-tagged tweets and assume that they have no geo-information. Assign a geo-information to each of these using your above method. Measure the differences between assigned location and the actual location. Plot them!

5. Crawling data from another social media
   a. Develop a crawler for any of the following social media sites (facebook, Instagram, Google Plus, Tumblr, Flickr, any other);
      i. Describe access restrictions and the approach developed
   b. Conduct data analysis similar to Twitter data analysis
      i. What level of comparative analysis given
   c. We are not giving any more specific directions, as we want students to be creative.

What to submit – non conformity to the submission instructions will lead to reduction in marks.

1) Report as a pdf file.
2) A zip file containing
   a. Software (runnable version, readme info, and also properly commented). It is important that software is runnable with minimum effort from the markers
   b. Data – provide a sample data for 5 minutes. Software should be able to run on this sample data.

Where to submit

1) Through Moodle link