COMP6235: Foundations of Data Science - CW 2 Specification

Coursework Title: Analysis of unstructured microblogging data (*Weighting 15%*)

Description: Based on the tutorials and teaching material provided in COMP6235, you are required to implement the first stages of the data processing pipeline, which will involve the transformation, storage, and querying of unstructured data.

In this assessment, you will be provided with an unstructured dataset containing records collected from a microblogging service. Each record will contain a timestamp, an identifier, and a short block of text. We wish for you to construct a data processing pipeline using a MongoDB datastore, and run create and compute a set of queries (given below).

Assessment: The assignment will consist of two handins: A 1-2 page document of the process required to import and query the data in MongoDB, along with a link to the code hosted on Github, and the results of the queries. The assignment will be awarded on the successful completion of the queries (80%), and the design and implementation of the processing pipeline (20%).

Requirements:
The following components are required to be used within the processing pipeline:

- A GitHub repository of the project. A URL to the project should be included in the project document.
- Program code that is compilable (Java, C, C++), or can be executed (Python/Javascript)
- A locally (or remotely) hosted MongoDB store.
- Results of the predefined queries.

Dataset Description:
The dataset contains the following fields:

- id - a unique identifier of the tuple
- id_member - a unique identifier of the user who posted the message
- timestamp - a UTC timestamp of when the message was published
- text - the microblog message that was published
- geo_lat - the latitude coordinate of where the message was posted from
- geo_lng - the longitude coordinate of where the message was posted from.

Queries to be performed:
Below are the queries to be completed:

1. How many unique users are there?
2. How many tweets (%) did the top 10 users (measured by the number of messages) publish?
3. What was the earliest and latest date (YYYY-MM-DD HH:MM:SS) that a message was published?
4. What is the mean time delta between all messages?
5. What is the mean length of a message?
6. What are the 10 most common unigram and bigram strings within the messages?
7. What is the average number of hashtags (#) used within a message?
8. Which area within the UK contains the largest number of published messages?
   Hint, the geographic latitude and longitude coordinates can be aggregated.

Dataset:

Compressed dataset: http://www.edshare.soton.ac.uk/15622/