COMP3210/COMP6210 - Big Data
Assignment 1

Semester 1, 2020
Macquarie University, Department of Computing

Due: Week 3 (Sunday 15 March)
Demonstration and Marking: Week 4, during the Practicals

Weighting: 5%

In this assignment you will acquire hands-on experience in designing, implementing and querying a NoSQL database, i.e. MongoDB.

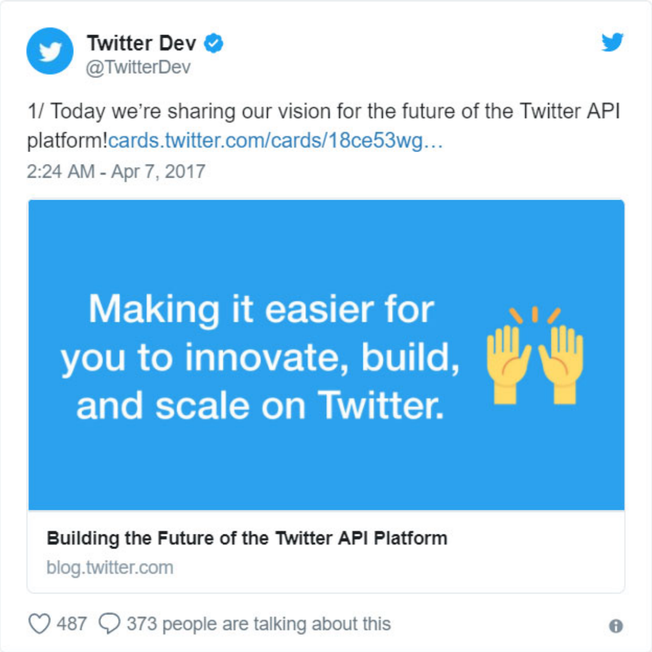
This Assessment Task relates to the following Learning Outcomes:
- Apply techniques for storing large volumes of data.

Dataset

Twitter[1] serves many objects as JSON[2], including *Tweets and Users*. These objects all encapsulate core attributes that describe the object. Each Tweet has an author, a message, a unique ID, a timestamp of when it was posted, and sometimes geo metadata shared by the user. Each User has a Twitter name, an ID, a number of followers, and most often an account bio.

With each Tweet, Twitter generates 'entity' objects, which are arrays of common Tweet contents such as hashtags, mentions, media, and links. If there are links, the JSON payload can also provide metadata such as the fully unwound URL and the webpage's title and description.

So, in addition to the text content itself, a Tweet can have over 140 attributes associated with it. Let's start with an example Tweet:



Twitter Dev
@TwitterDev

1/ Today we're sharing our vision for the future of the Twitter API platform!cards.twitter.com/cards/18ce53wg…

2:24 AM - Apr 7, 2017

Making it easier for you to innovate, build, and scale on Twitter.

Building the Future of the Twitter API Platform
blog.twitter.com

487 373 people are talking about this

[1] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json

[2] JSON is based on key-value pairs, with named attributes and associated values. These attributes, and their state, are used to describe objects.

The following JSON illustrates the structure for these objects and *some* of their attributes:

```
{
 "tweet": {
  "created_at": "Thu Apr 06 15:24:15 +0000 2017",
  "id_str": "850006245121695744",
  "text": "1/ Today we’re sharing our vision for the future of the Twitter API
platform!https://cards.twitter.com/cards/18ce53wgo4h/3xo1c … ",
  "user": {
   "id": 2244994945,
   "name": "Twitter Dev",
   "screen_name": "TwitterDev",
   "location": "Internet",
   "url": "https:\/\/dev.twitter.com\/",
   "description": "Your official source for Twitter Platform news, updates & events. Need technical help?
Visit https:\/\/twittercommunity.com\/ \u2328\ufe0f #TapIntoTwitter"
  },
  "place": {

  },
  "entities": {
   "hashtags": [

   ],
   "urls": [
    {
     "url": "https:\/\/t.co\/XweGngmxlP",
     "unwound": {
      "url": "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\/3xo1c",
      "title": "Building the Future of the Twitter API Platform"
     }
    }
   ],
   "user_mentions": [

   ]
  }
 }
}
```

Part 1 (30 Marks)

1) (3 mark) Create a collection called ‘Tweets’. We’re going to put some Tweets in it.
2) (3 marks) Add 50 Tweets to the database.
3) (3 mark) Write a MongoDB query that returns all the Tweets.
4) (3 marks) Write a MongoDB query to find one of your Tweets by name (e.g. "name": "Twitter Dev").
5) (3 marks) Update your two favourite Tweets to have two tags called ‘My number 1 Tweet’ and ‘My number 2 Tweet’. Show two ways to do this. Do the first using *update()* and do the second using *save()*. Hint: for save, you might want to query the object and store it in a variable first.
6) (3 marks) Write a MongoDB query that returns only Tweets that have tags. Not all of your Tweets should have tags, obviously.
7) (3 marks) Write a MongoDB query to display the first 5 Tweet which has the location ‘Internet’ ("location": "Internet").

8) (3 marks) Write a MongoDB query to find the Tweets whose id is greater than 2000000000 but less than 3000000000.
9) (3 marks) Write a MongoDB query to find the Tweet Id, name and location for those Tweets which contain 'Thu' as first three letters for its ‘created_at’.
10) (3 marks) Write a MongoDB query to find the Tweet Id for those Tweets which contain the keyword ‘health’ in their text.

Part 2 (20 Marks)

Write a program in Python, to read the Tweet Dataset (in part 1) from MongoDB. Then for each Tweet, extract keywords from the text of the Tweet. Then for each Tweet, add a new name/value pair to store the keywords in a comma-separated value (CSV) format; and update the original Tweet in the MongoDB.