WEB SCIENCE (H) COMPSCI4077

H Level Course Work

GUID: 2330306S
Name: Yatin Sharma

Github Repository:

https://www.github.com/2330306s/web_science

1. Introduction

a. The software development process started with getting approval for a Twitter developer account. Following which credentials and tokens were awarded to access data using appropriate APIs. The next step naturally was identifying the correct library to use which led to the discovery of tweepy. A crawler was developed using official documentation on the official website http://docs.tweepy.org/en/latest/ which processed data on a MongoDB server running locally and stored data in a database. Data was read in using the pandas library documentation from https://pandas.pydata.org/docs/ and clustered using scikit learn library from https://scikit-learn.org/stable/user_guide.html. Matplotlib guide from https://matplotlib.org/ was referred to generate a plot.

b. Tweet data namely tweet id, tweet date, tweet user id, tweet text and tweet hashtags, was collected on 20 March 2020 from 14:15 GMT till 15:15 GMT.

2. Data Crawl

a. Tweepy API was used due to its popularity, ease of use and availability of widespread support.
Firstly, authentication credentials and tokens are passed to the API to gain required access. Then the data was streamed according to the information collected using the streaming aspect of the API, parsed for required data, duplicates were ignored using the errors encountered by the API and the final dictionary was stored as a tuple in the appropriate collection in the database on the locally running MongoDB server.

b. The error generated by tweepy on encountering a tweet with the duplicate id was used to filter out duplicates.

3. Grouping of tweets

a. The pandas library was used to read in data from the csv generated from the collection on the database. The scikit learn library allowed for the use of a vectorizer to extract meaningful information from the three groups of data the clustering was carried out upon - tweet user id, tweet text and tweet hashtags. The KMeans class which is based on the principle of solving the k-means problem using either Lloyd's or Elkan's algorithm was used for generating the clusters.

b. As stated above the KMeans class was used for identification of clusters. It did so by generating centroids and assigning them to the nearest cluster.

4. Capturing and Organising User and Hashtag Information

5. Network Analysis Information