1. Section 1: Introduction

- Crawlers were developed using Tweepy API for accessing Twitter data with attempt to overcome REST call restriction from Twitter.
- Crawlers were developed in a multithreading manner for running time control
- There are 5 data collections all of which was run for 60 minutes
    - Fat: All of the data
    - Basic: gardenhoseAPI
    - EnhancedStream: enhanced stream API with topic ("Christmas") as it's currently trendy on twitter
    - REST: data collected via REST API

2. Section 2: Data crawl

a. Use streaming API for collecting 1% data (5 marks)

Used API: Tweepy for twitter data collecting

Main functions: Stream Listener API for listening to the live tweet stream with "sample" or "filter" method for different criteria

b. Enhance the crawling using Streaming and REST API (10 marks)

Used API: Tweepy for twitter data collecting

Main functions: Tweepy API with "search" method to request twitter data via REST calls

c. Grab as much geo-tagged data for Glasgow/Singapore for the same period (5 marks)

Used API: Tweepy for twitter data collecting

Main functions: Stream Listener API for listening to the live tweet stream via "filter" method with the locations parameter which takes in a bounding box coordinate

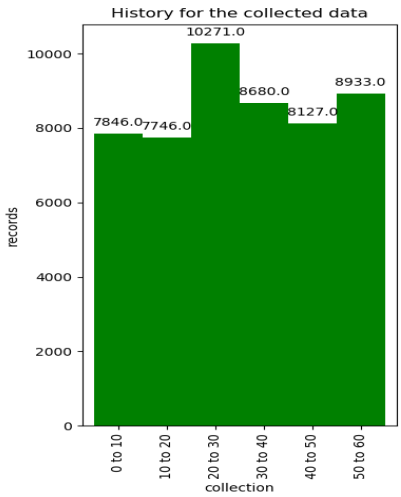
d. Discuss your data access strategies and how did you address Twitter data access restrictions (5 marks)

Twitter restrict REST calls to: 30 calls for 15 mins with each call return of 100 records

Strategy: make the thread going to sleep for 15 mins after 10 calls.

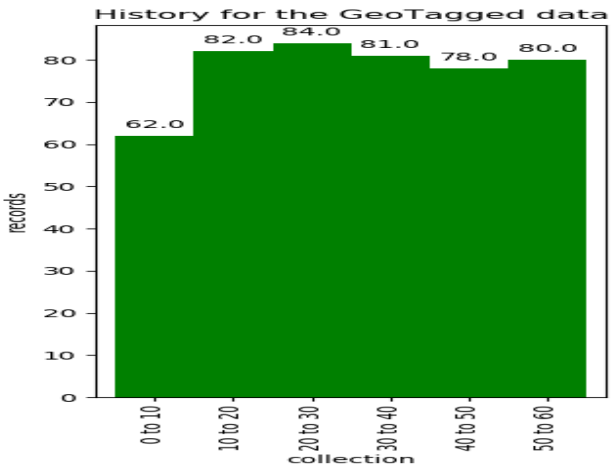
3. Basic data analytics (Total 20 marks)

a. Count the amount of data collected (5 marks)



This is the total amount of the collected tweets for each of the 10 minutes period

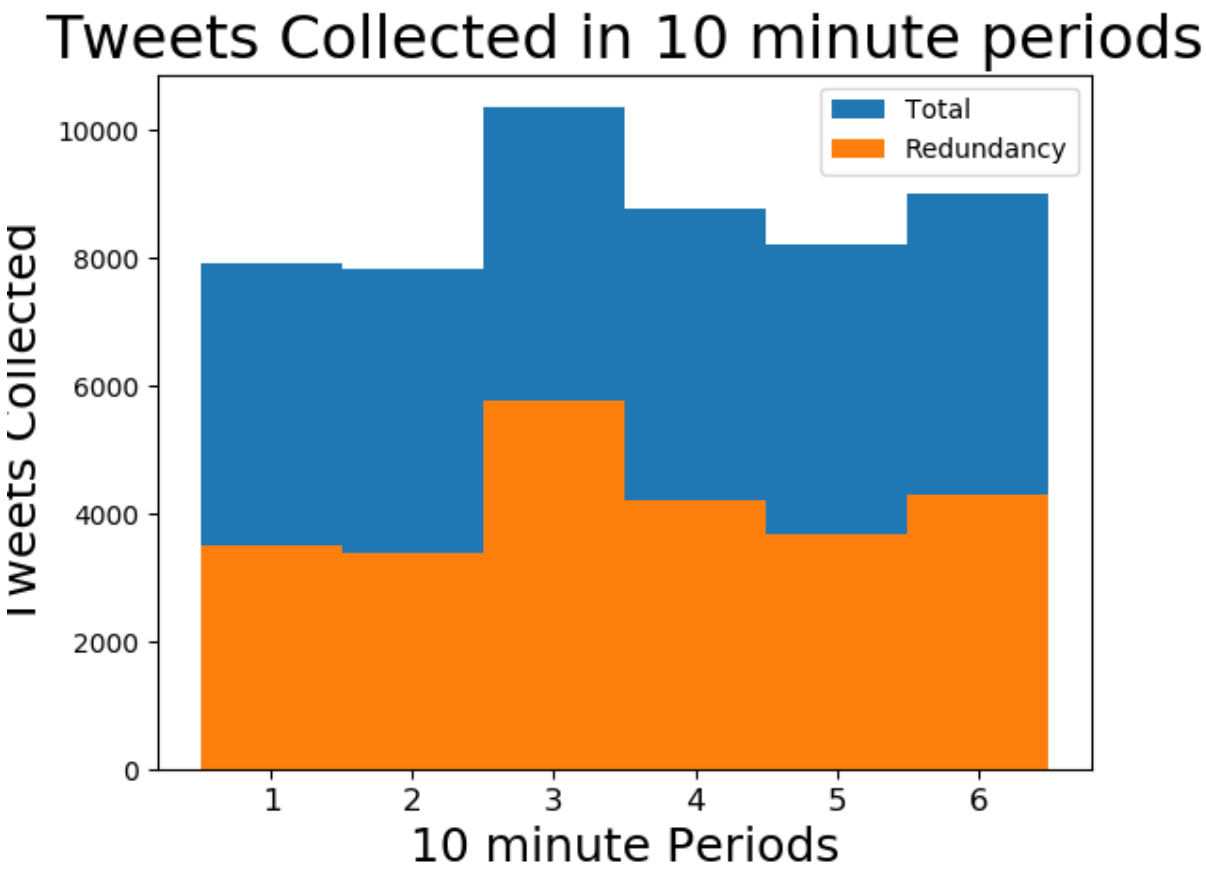
b. Count the amount of geo-tagged data from Glasgow / Singapore



This is the total amount of the collected tweets for each of the 10 minutes period with geo tag

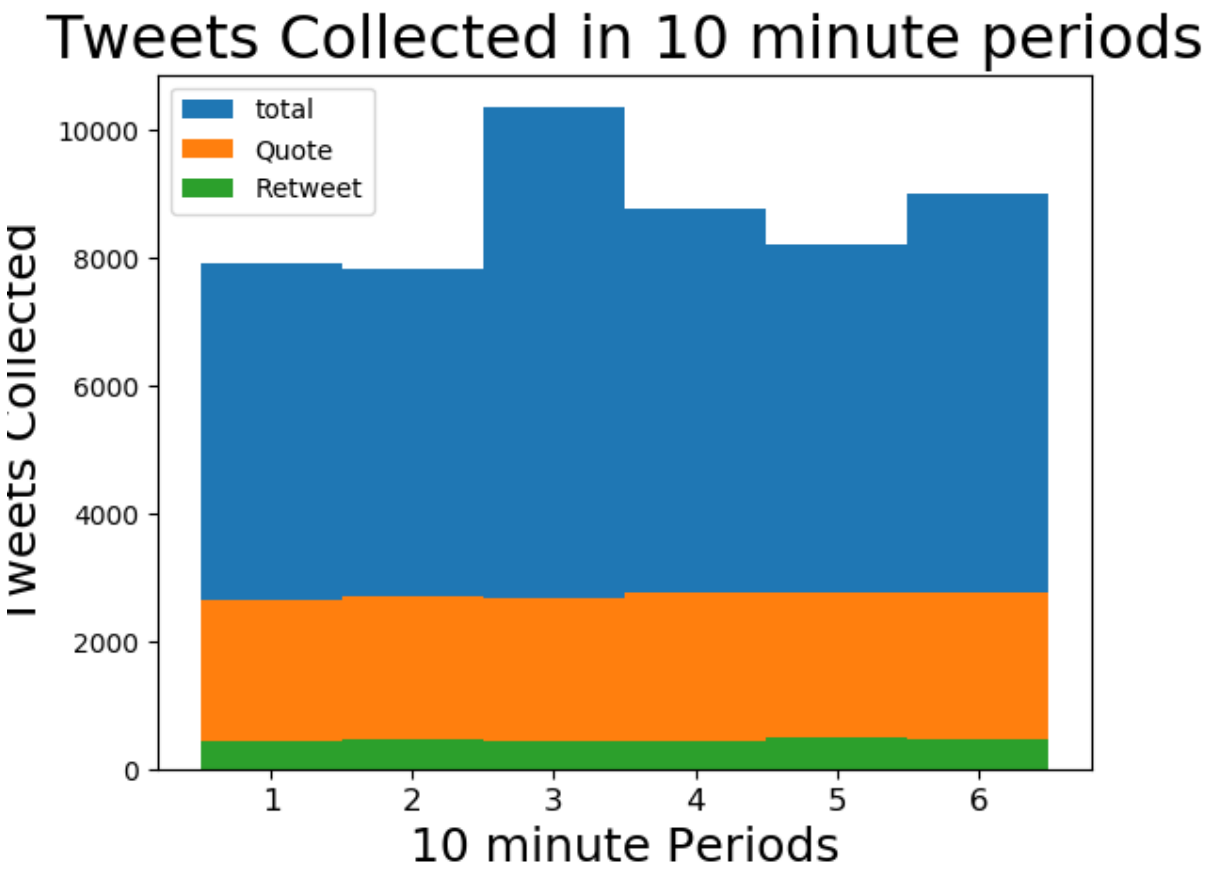
We can see there is very little of geotagged tweet in comparison to the total amount

c. Count redundant data present in the collection (you may end up collecting the same tweets again through various APIs) (5 marks)



There were lots of duplicated tweets during the data collection due to REST calls returning past data which should have been collected by streaming APIs.

d. Count the re-tweets and quotes (5 marks)



This is the total amount of the collected tweets for each of the 10 minutes period against the quotes and retweet. During this period, because the number of retweets and quotes is significantly lower than total amount of tweet, we can tell this maybe an indicator of new events.