Assignment 2

Part 1: Gather 1k URIs from twitter

Following guide and using code found here: http://adilmoujahid.com/posts/2014/07/twitter-analytics/

Stream was set up using keywords “GTA” and “Grand Theft Auto” for raw data collection over two days, and URLs were extracted from data. (ass2stream.py)

Made script (ass2read.py) with help from instructor to filer through collected URLs to remove links that did not work (return 200)

Used script found at:

https://stackoverflow.com/questions/20475552/python-requests-library-redirect-new-url/20475639

To expand shortened twitter links to final URLs (expandURIs.py).

Wrote simple script to sort expanded URLs and remove duplicates (sortList.py). Also manually removed all youtube and twitter links, because they were not useful. Ended up with 427 links remaining. Will gather more links using more useful keywords for future assignments

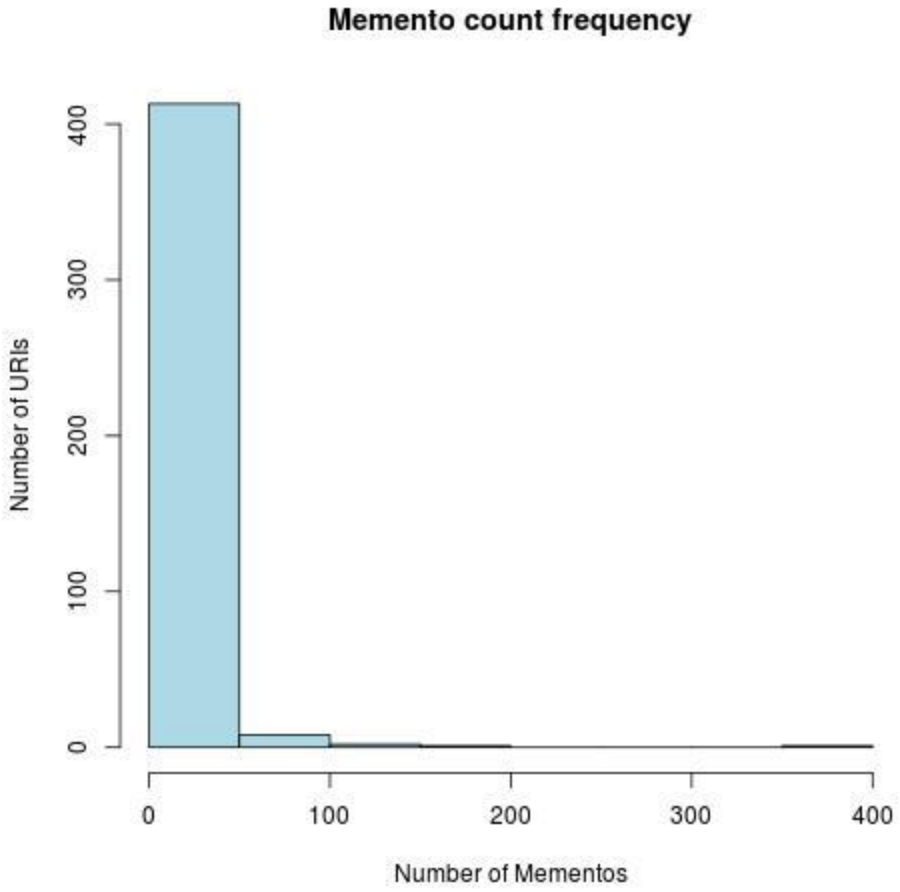
Part 2: Download the TimeMaps for each of the target URIs.

Wrote a script (dlTimeMaps.py) to download timemap info for each of the remaining links and save the info into a numbered file (ntimemap.txt). File number corresponds to line in which the url appears in the source file containing the URLs.

Wrote a script (mementoCount.py) to count the mementos in each timemap file, and save the counts into a separate text file memcount.txt

Wrote a script (simplecount.py) to create a text file simplecount.txt to match formatting requirements required for use with r.

Using r (with difficulty), managed to get a single histogram so successfully output.




It is obvious that the vast majority of web pages (caught by my twitter search keywords, anyway) do not have any mementos. I assume this is because of the quantity of video game content created in the form of streams, clips, replays, and screenshots. There is a ridiculous amount of content that is generated, probably at a rate that is to fast to track.

```
Part 3: Estimate the age of each of the 1000 URIs using the "Carbon
Date" tool:
```

Unable to complete. Carbon Date web tool is unresponsive, and I could not get Docker to function on my machine. Cannot spare any more time trying to complete this assignment.