Assignment 2

Web Science CS595

Name: Amara Naas

Question 1 In the first Python program (mytweets.py)[1] will extract 1000 unique links from Twitter by accessing Twitter API and retrieve my Twitter data. First thing to do is to create an application, and generate your own access token. Secondly, register the new application, get the consumer key and consumer secret, and save them at the mytweets.py[1]. Last step to get the 1000 URIs is generating the OAUTH tokens by run mytweets python file, set Oauth_Token and Oauth_Token_Secret in the URL. The program will search by list of words in all tweets a get the URIs that contain the word. The For loop will filter out undesirable URIs such images etc, make sure of the one that responds with a 200, and save them in a file called res.txt[1]. The code was built on top of Thomas Sileo code [2].

Question 2 The second Python program (time_map.py)[1] downloads the TimeMaps for each of the list of URIs collected in question 1 by using the ODU Memento Aggregator[3], generates a list of mementos for each URI in the list and save it in finalresult.txt[4]. I used this list and Create a histogram of number of URIs vs. number of Mementos as shown in figure 1 a and b. By looking at these histogram and analysing them we may observe following trends:

1- Most of the URIs have no mementos.

2- Very few and most popular web site have a large number of mementos.

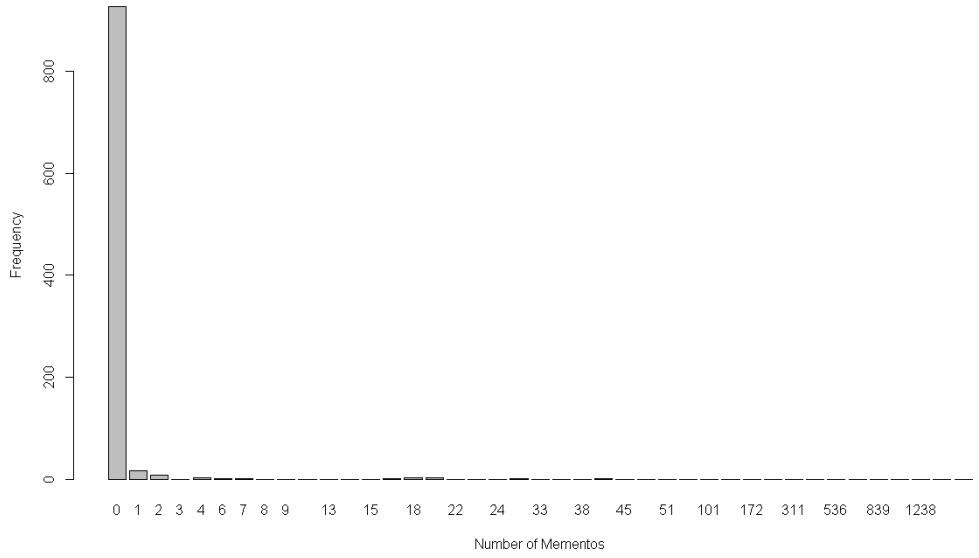
3- We may say that the curve follows an Inverse Gaussian Distribution, which has a [heavy right tail].

[1]File uploaded to github

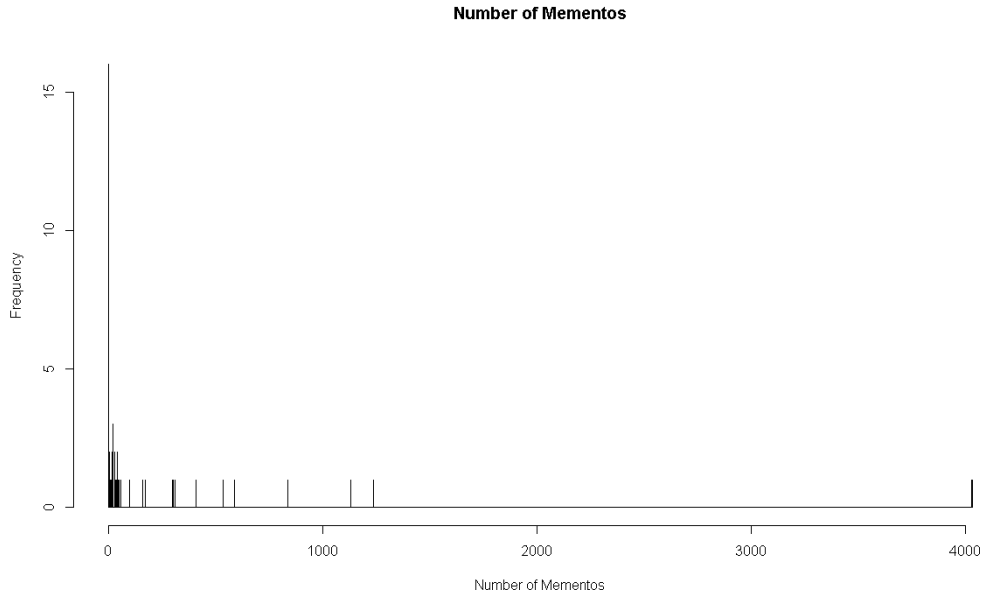
[2]http://www.thomassileo.com/blog/2013/01/25/using-twitter-rest-api-v1-dot-1-with-python/

[3]curl "http://mementoproxy.cs.odu.edu/aggr/timemap/link/+URI"

[4]File uploaded to github



(a) Num of URIs vs Num of Mementos where [Mementos with zero URI are not shown]



(b) Num of URIs vs Num of Mementos where [zero memento not in full scale]

Figure 1: URIs vs. number of Mementos (as computed from the TimeMaps)

Question 3 The second Python program (time_map.py)[5] estimates the age of each of the 1000 URIs by using the "Carbon Date" package that has been developed by Hany SalahELdeen, from OLD Dominion University, Norfolk, Virginia[6][7]. This program will give a list of URI's ages in days. Now I gathered the list of mementos from Q2 and the URI's ages list to create a graph of Archive Density as shown in Figure 2 which nothing but the frequency vs. size comparison. The data scattered apart from each other that is way I used logarithmic scale.

By looking at the graph we may observe that the data has no trends to follow and its very hard to analyse. In general we may say that URI has been in the web for a while, the probability of having large number of memento is high.

[5]File uploaded to github

[6]https://github.com/HanySalahEldeen/CarbonDate

[7]http://ws-dl.blogspot.com/2013/04/2013-04-19-carbon-dating-web.html

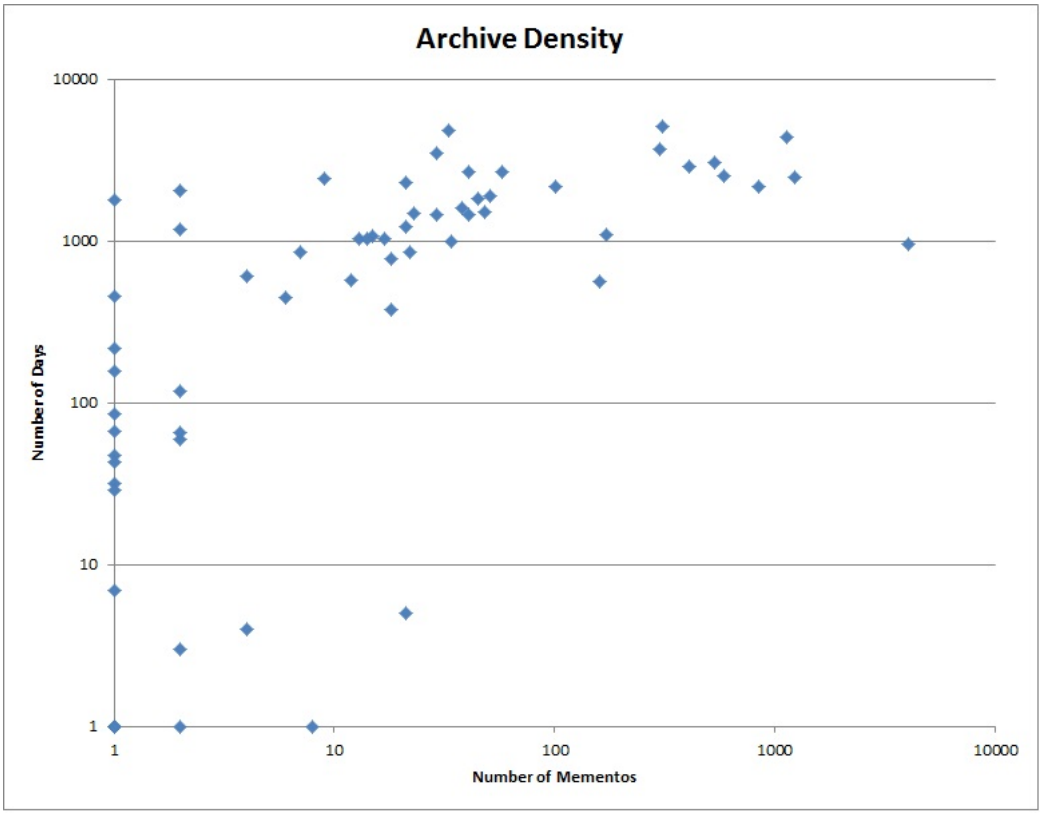


Figure 2: Archive Density [Num of age (in days) vs Num of Mementos]