CS 495 - Introduction to Web Science

Fall 2014

Assignment 2

by

Eric Littley
UIN: 00821698

September 28, 2014

Instructor

Dr. Michael Nelson

Department of Computer Science
Old Dominion University

Honor Pledge

I pledge to support the Honor System of Old Dominion University. I will refrain from any form of academic dishonesty or deception, such as cheating or plagiarism. I am aware that as a member of the academic community it is my responsibility to turn in all suspected violations of the Honor Code. I will report to a hearing if summoned.

Signed: Eric Littley

Contents

1 Introduction

Assignment 2 consists of three parts: downloading 1000 unique links from Twitter, Downloading the timemap for each link, and using the CarbonDate tool to estimate the age of each link. To complete each part a combination of python programs and bash scripts were used. The explanation of how these programs work together can be found below, along with a discussion on the data that was retrieved from Twitter.

2 Design

The first part of Assignment 1 is downloading 1000 unique links from Twitter. This was accomplished by writing a python program (extractLinks.py) that uses the Twitter search API. The program stores a list of search queries for various topics. The program uses each item in the list to query the Twitter API. The program queries twitter with a filter that only returns tweets with a link in them. The first 200 tweets for each query is saved and each link from each of those tweets are followed to their final URI using urllib2. The resulting URIs are saved in a set to ensure that there are no duplicates. The program loops through the queries until 1000 unique URIs have been downloaded. These links are saved in to a file. The figure 1 below shows the flow of data from programs to files up though part 3.

The second part of the assignment was to download the time tables for each URI and store the number of mementos for each URI. This was done using mementoweb.org. A python program(downloadTimetables.py) looped through the file of stored URIs and queried mementoweb.org for each of them. The resulting number of mementos along with the corresponding url was saved to another file.

The third part of the program was to estimate the creation time of each URI. The Carbon Date program was used to estimate the age of each URI. One draw back to using the Carbon Date tool was the length of time it took to return the results for each URI. A simple bash script(getCreation.sh) was written to loop through of file containing the URI and call the Carbon Date tool on it, then the results were saved in a temporary file to be parsed for useful information later. The list of URIs was split up into five files with a script running on each file independently. This helped reduce the time it took to get results. A better alternative would have been to write a multi-threading program. After these scripts finished executing the results were merged with a script(stripTime.sh) that parsed the raw results finding the estimated creation date for each URI and saving it to a file. The resulting file was used by a final script (memvstime.sh) to generate a file that contains URIs that had more than 0 mementos and a creation time this file was used to generate graphs. This script also generated a file that contained all the URIs with their memento count and creation date. Figure 1 shows the data flow of the overall assignment.

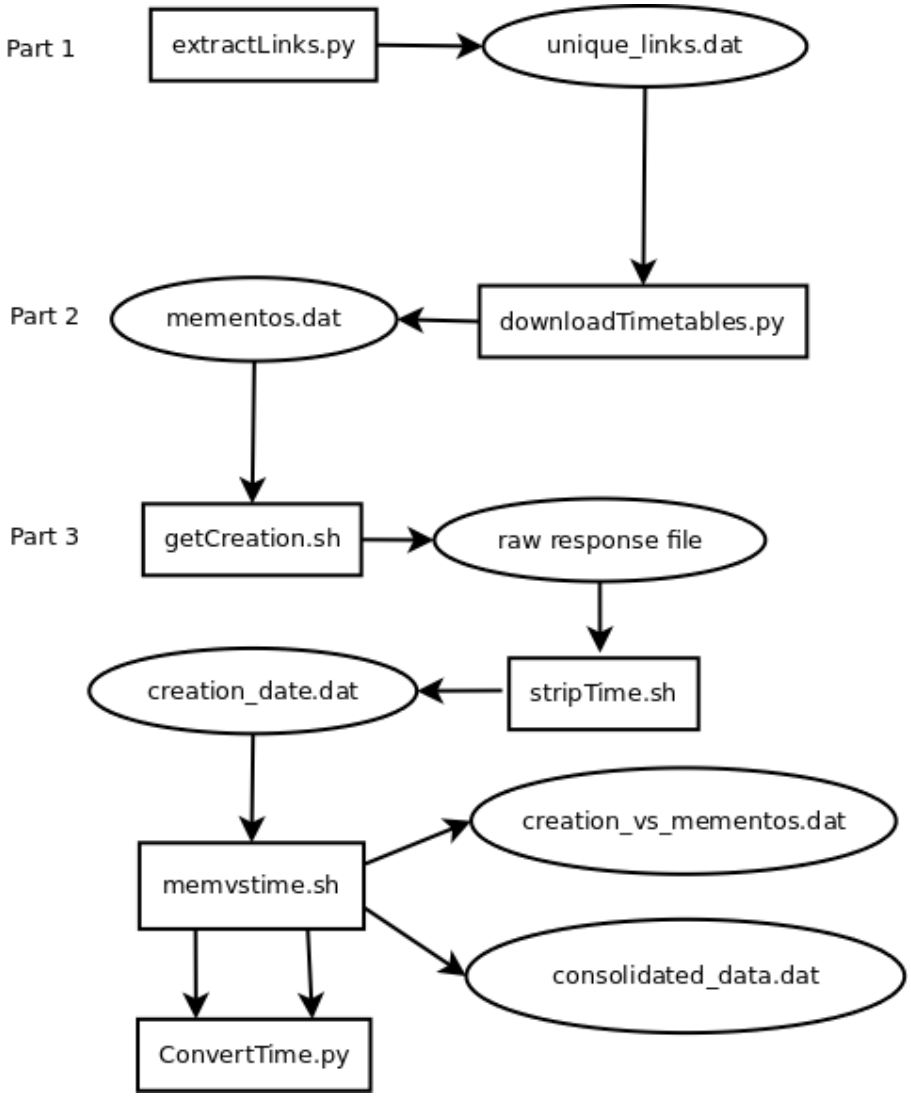
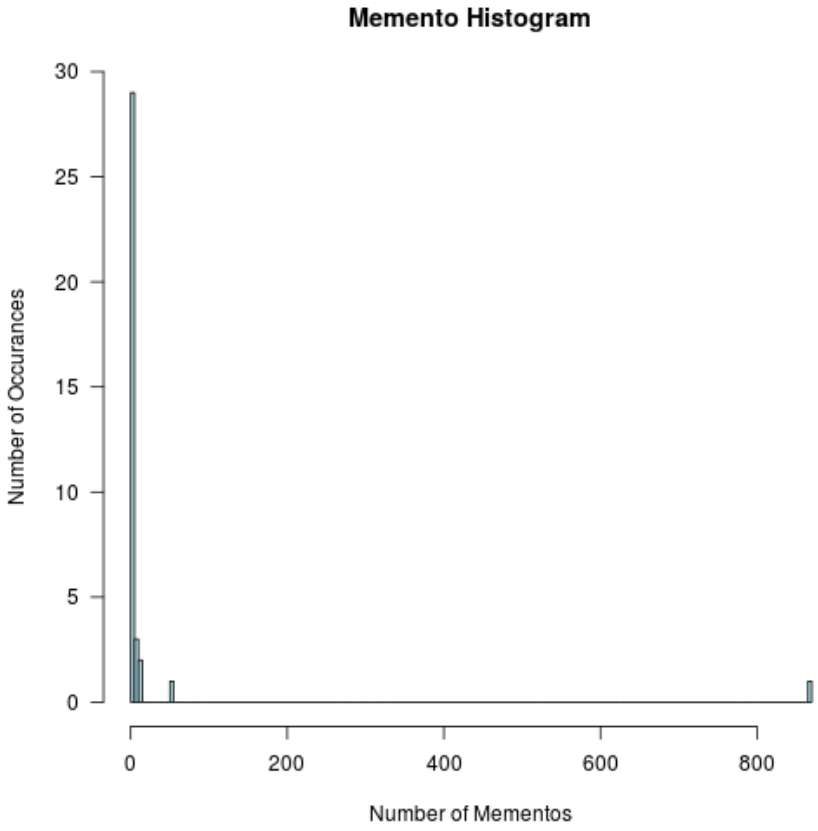


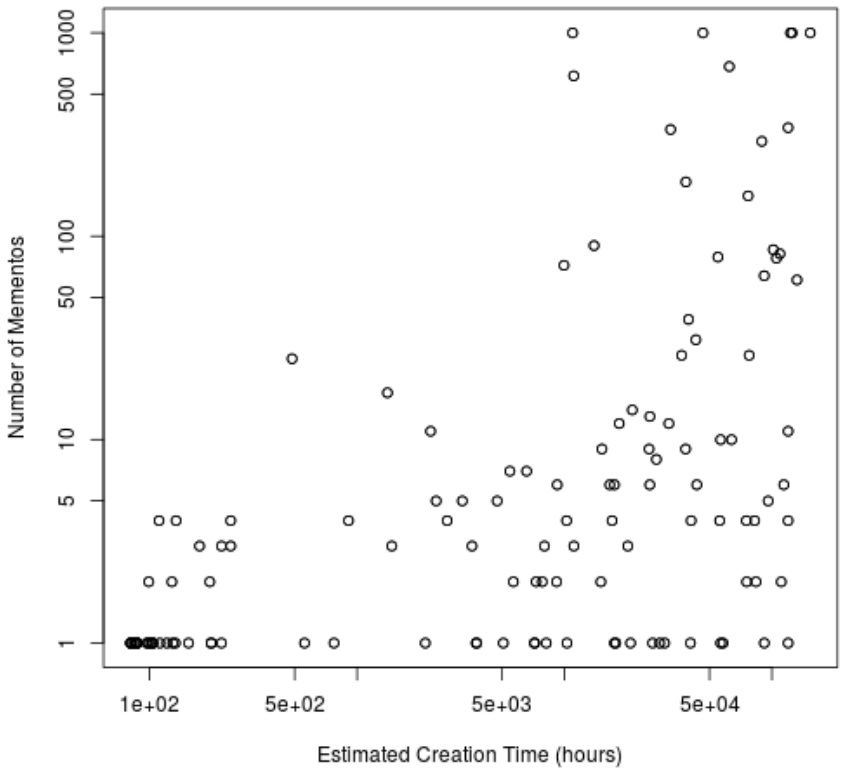
Figure 1: Dataflow

3 Results

Displayed below are the graphs developed in R.



The lack of pages with any mementos can be contributed to the fact that the program that extracted the links only extracted the most recent 200 tweets with each query.



The graph is log vs log and the time is in hours counting back from when the information was extracted.