CS 751: Assignment 1

Avinash Gosavi

Spring 2015

Contents

1 Question 1

Write a program that extracts 10000 tweets with links from Twitter. Reference - http://thomassileo.com/blog/2013/01/25/usin twitter-rest-api-v1-dot-1-with-python/ Other resources are available Note that only Twitter API 1.1 is currently available; version 1 code will no longer work.

- Save the tweets URIs, and the mapping to the link(s) each tweet contains
- For each `t.co` link use "curl I L" to record the HTTP headers all the way to a terminal HTTP status (i.e. chase down all the redirects)
- How many unique final URIs? How many duplicate URIs?
- Build a histogram of how many redirects (every URI will have at least 1) `http://en.wikipedia.org/wiki/Histogram` Build a histogram of HTTP status codes encountered (youll have at least 20000: 10000 301s, and 10000+ more)

1.1 Solution

The following steps were taken to extract tweets from twitter:

- Consumer key, consumer secret key, OAUTH token, OAUTH secret token are collected by registering an application in Twitter.
- Using above credentials and requests package for python I retrieved tweets from twitter
- Twitter response was saved as json in a local mongodb database.
- All redirects are calculated using the requests api and saved in a separate table other then the one we used for saving in tweets.
- Unique URI's, duplicate URI's are counted using set function in python.

2 Question 2

Write a Python program that:

- Uses “Carbon Date” to estimate the age of each links(s) in a tweet -See:“http://ws-dl.blogspot.com/2013/04/2013-04-19-carbon-dating-web.html”

- Create a histogram of (Age$_{\text{tweet}}$ - Age$_{\text{link}}$) .Many (most?) deltas will be 0,but there should be many¿0

- For these, compute: median, mean, std dev, std err

- Use wget to download the text for all the links. Hold on to those, we’ll come back to them later. -See :
“`http://superuser.com/questions/55040/save-a-single-web-page-with-background-images-with-wget`”
“`http://stackoverflow.com/questions/6348289/download-a-working-local-copy-of-a-webpage`”

2.1 Solution

- Phantomjs and Casperjs were the required for using Carbon Date library.

- I created a python scripts to retrieve and save details of links created date by using carbon date library.

- As it was taking time to retrieve the data using the library. I started running the code on multiple cloud server connected to a single mongodb server.

- To find an age of a link, I have written a python program ‘create$_{d}ata_{f}iles.py'whichwilltakethedateandreturndays.Mean, med$