Stanley Zheng
CS495: Web Sciences
September 26th, 2013

1. Writing a program that extracts 1000 unique URLs from twitter.
The Approach, query a set of keyword terms until no more making sure they follow these criteria.

- Not a 404
- Extended full URL (not t.co)
- Is unique

Code is very verbose because of escaped characters in URL encoding was giving trouble.
Otherwise it will spit out based on the specified programming languages a list of related tweets.
IE about #go, #scala, #java etc..

The Urls are then chased down using the 'REQUESTS' library and full url returned. An alternative was using a subprocess or pycurl.

See Q1.py for full code:

Example output ..
http://us2.campaign-archive2.com/?u=e2e180baf855ac797ef407fc7&id=c4a3a2ab87&utm_content=buffere8f30&utm_source=buffer&utm_medium=twitter&utm_campaign=Buffer
http://blog.disqus.com/post/62187806135/scaling-django-to-8-billion-page-views
http://www.meetup.com/newhavenio/events/139225322/
https://pythonconquerstheuniverse.wordpress.com/2012/02/15/mutable-default-arguments/
http://www.reddit.com/r/Python/comments/1n6etg/python_quick_reference_for_v27/
http://gitat.me/
http://blog.schockwellenreiter.de/essays/basemap01.html
https://code.google.com/p/pybotwar/?utm_content=buffercd930&utm_source=buffer&utm_medium=twitter&utm_campaign=Buffer
http://www.infragistics.com/community/blogs/d-coding/archive/2013/09/20/fundamentals-of-python-functions-formatting-amp-assignment-statements-week-2.aspx
http://www.10news.com/entertainment/around-the-web/usda-gets-patent-to-test-new-python-trap-in-florida-everglades09262013

Stanley Zheng
CS495: Web Sciences
September 26th, 2013

2. Download the TimeMaps for each of the target URIs. We'll use the ODU Memento Aggregator, so for example:

URI-R = http://www.cs.odu.edu/
URI-T = http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://www.cs.odu.edu/

The List was ran through the mementoproxy @ odu. However the R requirements for the assignment I believe were relaxed so just a list of the URL's run through the momento aggregator with the time score is put.

- http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://www.cs.odu.edu/ 564
- http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://www.meetup.com/newhavenio/events/139225322/ 0

Brief code: Rest in pCurl.py

```
for each in l:
        try:
                query = odu_memento+ each
                #print each
                num = num + 1
                print num
                r= requests.get(query)
                if (r.status_code==200):
                        print r.url
                        search = r.text
                        count = [i for i in range(len(search)) if search.startswith('memento', i)]
                        t.write(each+" "+ len(count) +"\n")
                elif (r.status_code==404):
                        print each+" 0\n"
                        t.write(each+" 0\n")
        except:
                pass
```

Stanley Zheng
CS495: Web Sciences
September 26th, 2013

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

See the outprint of all the carbon
Use of Hany's updated code* in Hany.py

```
if(len(sys.argv)!=2):
        print "wrong format"
else:
        url = sys.argv[1]
        f = open('final_uri_list.txt', 'r')
        t = open('carbon.txt', 'w')
        l = f.read().splitlines()
        for url in l:
                try:
                        carbon= carbonDate(url)
                        t.write(carbon + "\n\n")
                except:
                        pass
```

Bit.ly Key would not work. Example output


```
{
  "URI": "http://www.meetup.com/newhavenio/events/139225322/",
  "Estimated Creation Date": "2013-09-09T14:43:22",
  "Last Modified": "",
  "Bitly.com": "Bitly Key has expired",
  "Topsy.com": "2013-09-09T14:43:22",
  "Backlinks": "",
  "Google.com": "",
  "Archives": {
    "Earliest": "",
    "By_Archive": {}
  }
}
```


sources

========

Filtering

Stanley Zheng
CS495: Web Sciences
September 26th, 2013

---
http://stackoverflow.com/questions/6257544/twitter-search-api-to-exclude-tweets-with-a-filter

Original Boiler Plate
---
http://thomassileo.com/blog/2013/01/25/using-twitter-rest-api-v1-dot-1-with-python/

how to run bash curl
http://stackoverflow.com/questions/5460923/run-bash-built-in-commands-in-python

how to follow uris
==========
google group Dr. Nelson's Hint
http://docs.python-requests.org/en/latest/

simple to count the number of mementos
---
http://stackoverflow.com/questions/4664850/find-all-occurrences-of-a-substring-in-python