CS 532: Assignment 10

Dinesh Kumar Paladhi

Spring 2016

Contents

1 Problem 1

1. Using the data from A8:

- Consider each row in the blog-term matrix as a 500 dimension vector,
corresponding to a blog.

- From chapter 8, replace numpredict.euclidean() with cosine as the
distance metric. In other words, you'll be computing the cosine between
vectors of 500 dimensions.

- Use knnestimate() to compute the nearest neighbors for both:

http://f-measure.blogspot.com/
http://ws-dl.blogspot.com/

for k={1,2,5,10,20}.

1.1 Solution

1. My task here is to find out the nearest neighbors for "http://f-measure.blogspot.com/" and "http://ws-dl.blogspot.com/" blogs.

2. In order to find this I took blogdata matrix from my assignment 8 and processed it using the code in listing1. Sample blogdata matrix can be found in fig1.

3. This code creates a vector for each blog which can be given as input to the my next code in listing2.

4. I have taken this code from "Programming Collective Intelligence" textbook and made modifications to it.

5. I have deleted Euclidean function and inserted cosine function as distance metric. So, this is used to find the cosine between vectors of 500 dimensions.

6. Knnestimate() function is to find the neighbors for a particular blog which takes input as k=1 or 2 or 5 or 10 or 20.

7. Each time we give a k value it gives the respective k number of neighbors for that particular blog.

8. The nearest neighbors for "F-Measure" blog can be found in fig2.

9. The nearest neighbors for "Web Science and Digital Libraries Research Group" blog can be found in fig3.

1.2 Code Listing 1

```
import json

input=open('blogdata.txt','r')
output=open('rowassign','w')
flag=0
row=[]
for i in input:
        flag=flag+1
        if flag >1:
                dictionary={}
                drow=i.strip().split('\t')
                name=drow[0]
                drow.pop(0)
                rowassign=drow
                # print name
                # print rowassign
                dictionary[name]=rowassign #assigning each row vector to a blog
                row.append(dictionary)
output.write(json.dumps(row))
```

Listing 1: Python Code for creating a list of all words for a particular blog

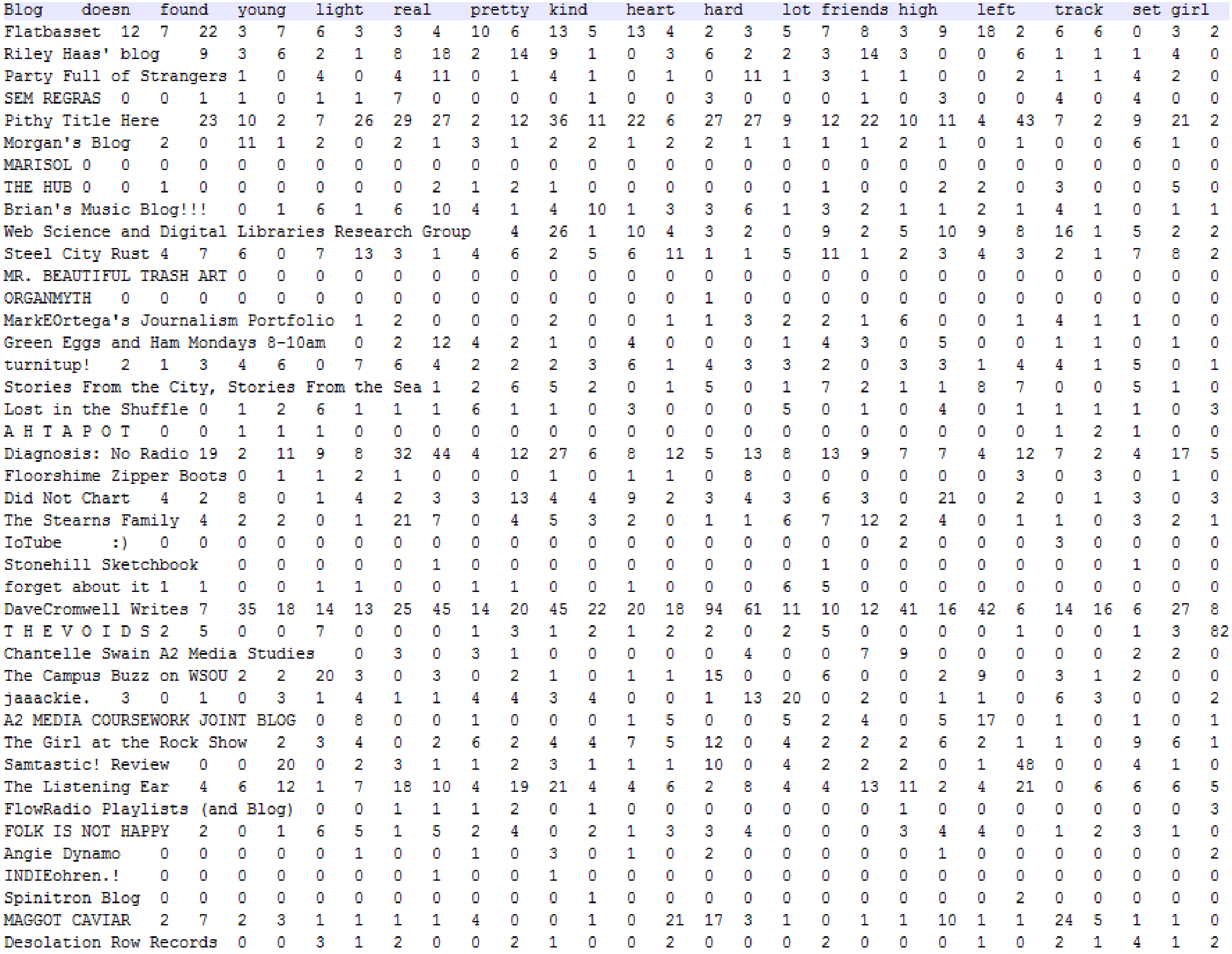
1.3 Code Listing 2

```
from random import random,randint
import math
import json



input= open('rowassign','r')
blogdata = json.load(input)
for line in blogdata:
        for nline in line:
                if nline == 'Web Science and Digital Libraries Research Group':
                        vec1= line[nline]
                        knnestimate(blogdata,vec1)



def getdistances(blogdata,vec1):
  distancelist=[]

  # Loop over every item in the dataset
  for i in blogdata:
    for nline in i:
      if nline != 'F-Measure':
        vec2= i[nline]

    # Add the distance and the index
    distancelist.append((cosineDistance(vec1,vec2),i))

  # Sort by distance
  distancelist.sort()

  return distancelist


def cosineDistance(v1,v2):
  "compute cosine similarity of v1 to v2: (v1 dot v2)/{||v1||*||v2||)"
  sumxx, sumxy, sumyy = 0, 0, 0
  for i in range(0,len(v1)-1):
    x = int(v1[i]); y = int(v2[i])
    sumxx += x*x
    sumyy += y*y
    sumxy += x*y
  return sumxy/math.sqrt(sumxx*sumyy)



def knnestimate(data,vec1,k=20):
  # Get sorted distances
  print 'k=20'
  print "Twenty neighbours for Web Science and Digital Libraries Research Group are"
  dlist=getdistances(data,vec1)
  avg=0.0
  # print dlist
  # Take the average of the top k results
  for i in range(k):
    idx=dlist[i]
    value = idx[0]
    for item in idx[1]:
                blogname= item
    print blogname +'\t'+ str(value)
```

Listing 2: Python Code for finding neighbors

1.4 Input

Sample Blogdata

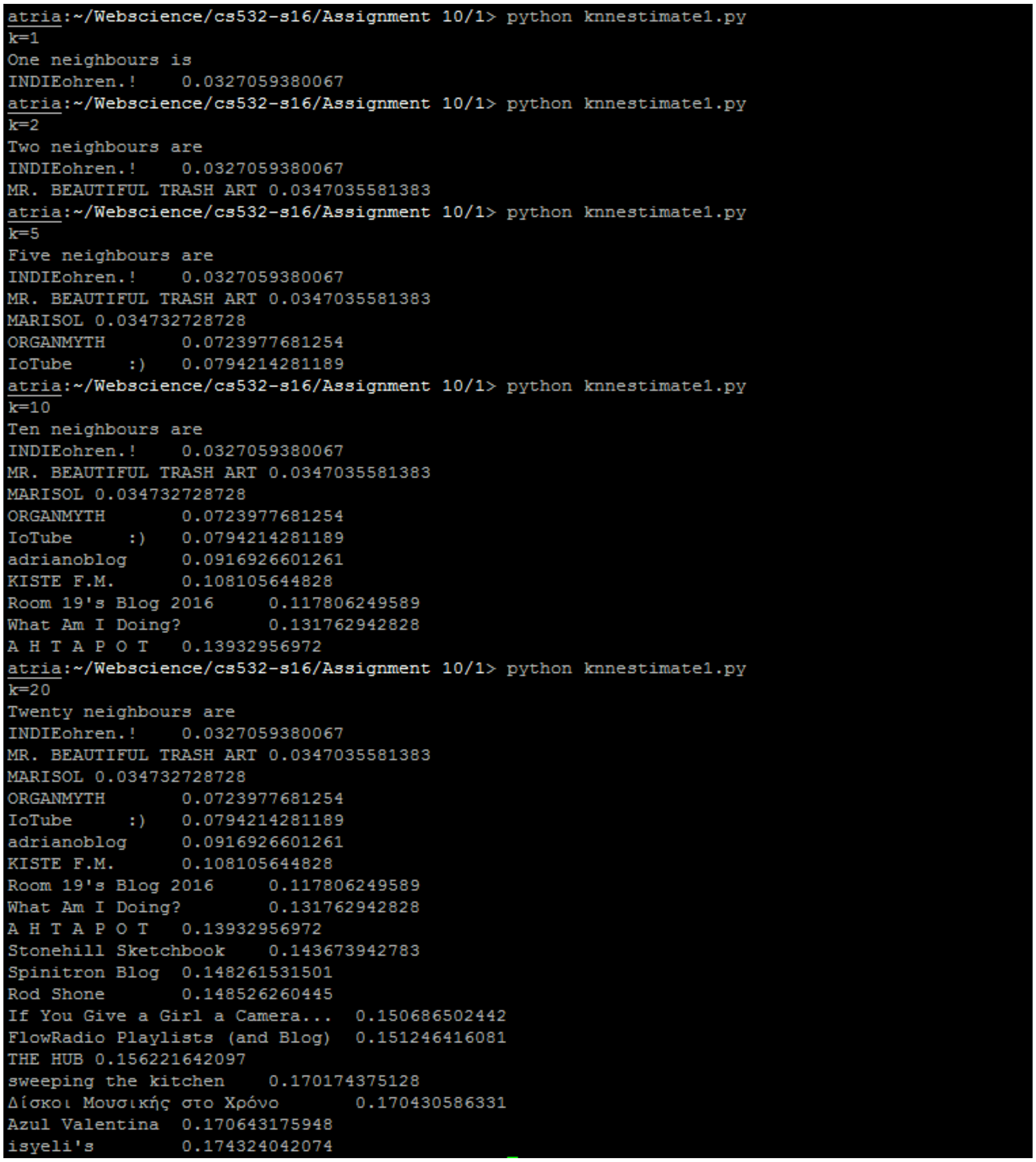


```
Blog    doesn   found   young   light   real    pretty  kind    heart   hard    lot friends high    left    track   set girl
Flatbasset  12  7   22  3   7   6   3   3   4   10  6   13  5   13  4   2   3   5   7   8   3   9   18  2   6   6   0   3   2
Riley Haas' blog    9   3   6   2   1   8   18  2   14  9   1   0   3   6   2   2   3   14  3   0   0   6   1   1   1   4   0
Party Full of Strangers 1   0   4   0   4   11  0   1   4   1   0   1   0   11  1   3   1   1   0   0   2   1   1   4   2   0
SEM REGRAS  0   0   1   1   0   1   1   7   0   0   0   0   1   0   0   3   0   0   0   1   0   3   0   0   4   0   4   0   0
Pithy Title Here    23  10  2   7   26  29  27  2   12  36  11  22  6   27  27  9   12  22  10  11  4   43  7   2   9   21  2
Morgan's Blog   2   0   11  1   2   0   2   1   3   1   2   2   1   2   2   1   1   1   1   2   1   0   1   0   0   6   1   0
MARISOL 0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
THE HUB 0   0   1   0   0   0   0   0   0   2   1   2   1   0   0   0   0   0   0   1   0   0   2   2   0   3   0   0   5   0
Brian's Music Blog!!!   0   1   6   1   6   10  4   1   4   10  1   3   3   6   1   3   2   1   1   2   1   4   1   0   1   1
Web Science and Digital Libraries Research Group    4   26  1   10  4   3   2   0   9   2   5   10  9   8   16  1   5   2   2
Steel City Rust 4   7   6   0   7   13  3   1   4   6   2   5   6   11  1   1   5   11  1   2   3   4   3   2   1   7   8   2
MR. BEAUTIFUL TRASH ART 0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
ORGANMYTH   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
MarkEOrtega's Journalism Portfolio  1   2   0   0   0   2   0   0   1   1   3   2   2   1   6   0   0   1   4   1   1   0   0
Green Eggs and Ham Mondays 8-10am   0   2   12  4   2   1   0   4   0   0   0   1   4   3   0   5   0   0   1   1   0   1   0
turnitup!   2   1   3   4   6   0   7   6   4   2   2   2   3   6   1   4   3   3   2   0   3   3   1   4   4   1   5   0   1
Stories From the City, Stories From the Sea 1   2   6   5   2   0   1   5   0   1   7   2   1   1   8   7   0   0   5   1   0
Lost in the Shuffle 0   1   2   6   1   1   1   6   1   1   0   3   0   0   0   5   0   1   0   4   0   1   1   1   1   0   3
A H T A P O T   0   0   1   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   2   1   0   0
Diagnosis: No Radio 19  2   11  9   8   32  44  4   12  27  6   8   12  5   13  8   13  9   7   7   4   12  7   2   4   17  5
Floorshime Zipper Boots 0   1   1   2   1   0   0   0   1   0   1   1   0   8   0   0   0   0   0   0   3   0   3   0   1   0
Did Not Chart   4   2   8   0   1   4   2   3   3   13  4   4   9   2   3   4   3   6   3   0   21  0   2   0   1   3   0   3
The Stearns Family  4   2   2   0   1   21  7   0   4   5   3   2   0   1   1   6   7   12  2   4   0   1   1   0   3   2   1
IoTube      :)  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0   0   3   0   0   0   0
Stonehill Sketchbook    0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   1   0   0
forget about it 1   1   0   0   1   1   0   0   1   1   0   0   1   0   0   0   6   5   0   0   0   0   0   0   0   0   0   0
DaveCromwell Writes 7   35  18  14  13  25  45  14  20  45  22  20  18  94  61  11  10  12  41  16  42  6   14  16  6   27  8
T H E V O I D S 2   5   0   0   7   0   0   0   1   3   1   2   1   2   2   0   2   5   0   0   0   0   1   0   0   1   3   82
Chantelle Swain A2 Media Studies    0   3   0   3   1   0   0   0   0   0   4   0   0   7   9   0   0   0   0   0   2   2   0
The Campus Buzz on WSOU 2   2   20  3   0   3   0   2   1   0   1   1   15  0   0   6   0   0   2   9   0   3   1   2   0   0
jaaackie.   3   0   1   0   3   1   4   1   1   4   4   3   4   0   0   1   13  20  0   2   0   1   1   0   6   3   0   0   2
A2 MEDIA COURSEWORK JOINT BLOG  0   8   0   0   1   0   0   0   1   5   0   0   5   2   4   0   5   17  0   1   0   1   0   1
The Girl at the Rock Show   2   3   4   0   2   6   2   4   4   7   5   12  0   4   2   2   2   6   2   1   1   0   9   6   1
Samtastic! Review   0   0   20  0   2   3   1   1   2   3   1   1   1   10  0   4   2   2   2   0   1   48  0   0   4   1   0
The Listening Ear   4   6   12  1   7   18  10  4   19  21  4   4   6   2   8   4   4   13  11  2   4   21  0   6   6   6   5
FlowRadio Playlists (and Blog)  0   0   1   1   1   2   0   1   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   3
FOLK IS NOT HAPPY   2   0   1   6   5   1   5   2   4   0   2   1   3   3   4   0   0   0   3   4   4   0   1   2   3   1   0
Angie Dynamo    0   0   0   0   0   1   0   0   1   0   3   0   1   0   2   0   0   0   0   0   1   0   0   0   0   0   0   2
INDIEohren.!    0   0   0   0   0   0   0   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
Spinitron Blog  0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   2   0   0   0   0   0
MAGGOT CAVIAR   2   7   2   3   1   1   1   1   4   0   0   1   0   21  17  3   1   0   1   1   10  1   1   24  5   1   1   0
Desolation Row Records  0   0   3   1   2   0   0   2   1   0   0   2   0   0   0   2   0   0   0   1   0   2   1   4   1   2
```

Figure 1: Sample Blogdata

1.5 Outputs

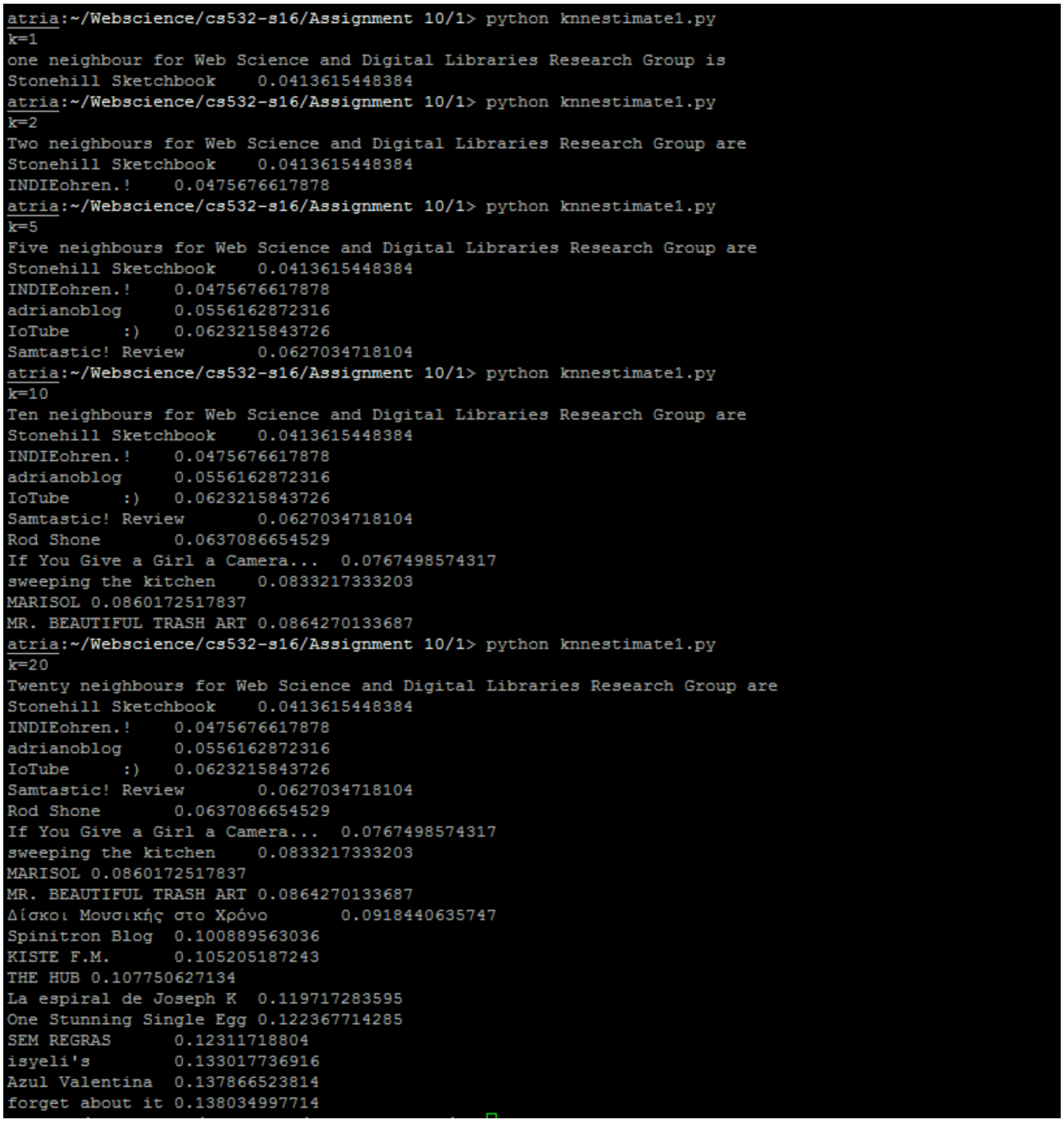
Neighbors of F-Measure



```
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=1
One neighbours is
INDIEohren.!    0.0327059380067
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=2
Two neighbours are
INDIEohren.!    0.0327059380067
MR. BEAUTIFUL TRASH ART 0.0347035581383
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=5
Five neighbours are
INDIEohren.!    0.0327059380067
MR. BEAUTIFUL TRASH ART 0.0347035581383
MARISOL 0.034732728728
ORGANMYTH       0.0723977681254
IoTube     :)   0.0794214281189
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=10
Ten neighbours are
INDIEohren.!    0.0327059380067
MR. BEAUTIFUL TRASH ART 0.0347035581383
MARISOL 0.034732728728
ORGANMYTH       0.0723977681254
IoTube     :)   0.0794214281189
adrianoblog     0.0916926601261
KISTE F.M.      0.108105644828
Room 19's Blog 2016     0.117806249589
What Am I Doing?        0.131762942828
A H T A P O T   0.13932956972
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=20
Twenty neighbours are
INDIEohren.!    0.0327059380067
MR. BEAUTIFUL TRASH ART 0.0347035581383
MARISOL 0.034732728728
ORGANMYTH       0.0723977681254
IoTube     :)   0.0794214281189
adrianoblog     0.0916926601261
KISTE F.M.      0.108105644828
Room 19's Blog 2016     0.117806249589
What Am I Doing?        0.131762942828
A H T A P O T   0.13932956972
Stonehill Sketchbook    0.143673942783
Spinitron Blog  0.148261531501
Rod Shone       0.148526260445
If You Give a Girl a Camera...  0.150686502442
FlowRadio Playlists (and Blog)  0.151246416081
THE HUB 0.156221642097
sweeping the kitchen    0.170174375128
Δίσκοι Μουσικής στο Χρόνο       0.170430586331
Azul Valentina  0.170643175948
isyeli's        0.174324042074
```

Figure 2: Neighbors of F-Measure

Neighbors of Ws-dl



```
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=1
one neighbour for Web Science and Digital Libraries Research Group is
Stonehill Sketchbook    0.0413615448384
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=2
Two neighbours for Web Science and Digital Libraries Research Group are
Stonehill Sketchbook    0.0413615448384
INDIEohren.!    0.0475676617878
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=5
Five neighbours for Web Science and Digital Libraries Research Group are
Stonehill Sketchbook    0.0413615448384
INDIEohren.!    0.0475676617878
adrianoblog     0.0556162872316
IoTube     :)   0.0623215843726
Samtastic! Review       0.0627034718104
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=10
Ten neighbours for Web Science and Digital Libraries Research Group are
Stonehill Sketchbook    0.0413615448384
INDIEohren.!    0.0475676617878
adrianoblog     0.0556162872316
IoTube     :)   0.0623215843726
Samtastic! Review       0.0627034718104
Rod Shone       0.0637086654529
If You Give a Girl a Camera...  0.0767498574317
sweeping the kitchen    0.0833217333203
MARISOL 0.0860172517837
MR. BEAUTIFUL TRASH ART 0.0864270133687
atria:~/Webscience/cs532-s16/Assignment 10/1> python knnestimate1.py
k=20
Twenty neighbours for Web Science and Digital Libraries Research Group are
Stonehill Sketchbook    0.0413615448384
INDIEohren.!    0.0475676617878
adrianoblog     0.0556162872316
IoTube     :)   0.0623215843726
Samtastic! Review       0.0627034718104
Rod Shone       0.0637086654529
If You Give a Girl a Camera...  0.0767498574317
sweeping the kitchen    0.0833217333203
MARISOL 0.0860172517837
MR. BEAUTIFUL TRASH ART 0.0864270133687
Δίσκοι Μουσικής στο Χρόνο       0.0918440635747
Spinitron Blog  0.100889563036
KISTE F.M.      0.105205187243
THE HUB 0.107750627134
La espiral de Joseph K  0.119717283595
One Stunning Single Egg 0.122367714285
SEM REGRAS      0.12311718804
isyeli's        0.133017736916
Azul Valentina  0.137866523814
forget about it 0.138034997714
```

Figure 3: Neighbors of Ws-dl

2 Problem 2

2. Rerun A9, Q2 but this time using LIBSVM. If you have n categories, you'll have to run it n times. For example, if you're classifying music and have the categories:

metal, electronic, ambient, folk, hip-hop, pop

you'll have to classify things as:

metal / not-metal
electronic / not-electronic
ambient / not-ambient

etc.

Use the 500 term vectors describing each blog as the features, and your mannally assigned classifications as the true values. Use 10-fold cross-validation (as per slide 46, which shows 4-fold cross-validation) and report the percentage correct for each of your categories.

2.1 Solution

1. Not Attempted

3 Problem 3

```
3. Re-download the 1000 TimeMaps from A2, Q2.  Create a graph where
the x-axis represents the 1000 TimeMaps.  If a TimeMap has ‘‘shrunk’’,
it will have a negative value below the x-axis corresponding to the
size difference between the two TimeMaps.  If it has stayed the
same, it will have a ‘‘0’’ value.  If it has grown, the value will be
positive and correspond to the increase in size between the two
TimeMaps.

As always, upload all the TimeMap data.  If the A2 github has the
original TimeMaps, then you can just point to where they are in
the report.
```

3.1 Solution

1. For this question I need to get data from my Assignment 2.
2. I have taken the code from the assignment 2 and executed it again which gives me a complete different set of TimeMaps. So, Now I have old and new TimeMaps which should be used to get solution for this question.
3. Python code for getting the new TimeMaps can be found in listing3.
4. The input I gave to the above code can be found in fig4.
5. So, Now I subtracted the old TimeMaps from the New TimeMaps which gives me the difference between both of them.
6. This difference in the TimeMaps is then plotted using the following R code which can be seen in listing4.
7. The plotted graph can be seen in the fig5.
8. So by this we can know that there have been positive increase and negative increase from new and old TimeMaps data.

3.2 Code Listing 1

```python
import re
import urllib2
import json
import sys

def getmementos(url):
        mem_prefix = 'http://mementoproxy.cs.odu.edu/aggr/timemap/link/1/' + url   #memento aggregator is concatenated with the url for which mementos should be found out
        try:
                response = urllib2.urlopen(mem_prefix)
                time_map = response.read()
        except urllib2.HTTPError:
                time_map = None
        return time_map

find_memento = re.compile(r'rel.*?=.*?"memento".*?')  # To find memento using regular expression
my_urls = open('my_json_data','r+')  #This file contains 1000 urls their tweets,tweet ids and created dates
output_file = open('mem_and_links.json','a')  # This file stores number of mementos for each url
output_file2 = open('only_count.csv','a')
output_file_carbon = open('mem_grt0.json','a')
one_element={}
count_of_mems = [] #array is created to store count
for line in my_urls.readlines(): #reads line by line
        each_line = json.loads(line)
        url = each_line['url']
        memento_data = getmementos(url)

        #print memento_data
        if memento_data == 'Null':
                count = 0
                one_element['num_of_mems'] = count
                one_element['url'] = url
                output_file.write(json.dumps(one_element)+'\n') #adding each element into json file
                #print count,"    ",url
        else:
                count = len(find_memento.findall(str(memento_data))) #forms an array where "memento"" is found and finds the length of that array
                # a=find_memento.findall(str(memento_data))
                # print a
                one_element['num_of_mems'] = count
                one_element['url'] = url
                output_file.write(json.dumps(one_element)+'\n') #adding each element into json file
                output_file2.write("%s\n" % (count))
                if one_element['num_of_mems'] != 0:
                        output_file_carbon.write(json.dumps(one_element)+'\n') # for getting urls and mementos for mementos > 0
                #output_file2.write('\r\n')
                #print count,"    ",url
output_file.close()
output_file2.close()
output_file_carbon.close()
```

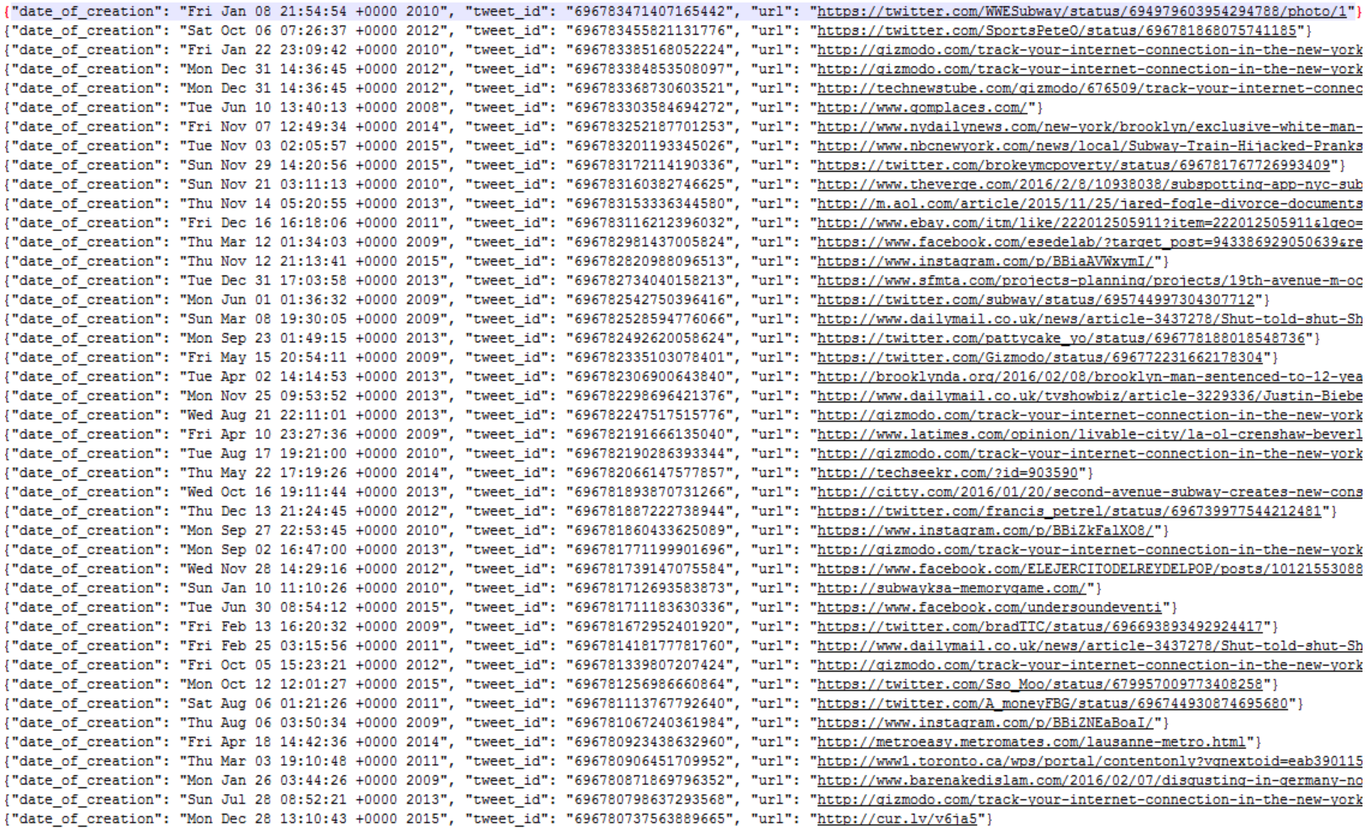
Listing 3: Python Code for counting the number of mementos for each URI

3.3 Code Listing 2

```
data = scan(”difference_new−old.csv”)
plot(data,xlab=”Number of URI's”,ylab=”Difference in bytes between New and Old Raw data”,
     main=”Differences in the number of Mementos for Old and New data for 1000 URI's”,xlim=c
     (0,1000),ylim=c(−2,20),col=”blue”,type=”l”)
```

Listing 4: R Code for for plotting graph

3.4 Input



```
{"date_of_creation": "Fri Jan 08 21:54:54 +0000 2010", "tweet_id": "696783471407165442", "url": "https://twitter.com/WWESubway/status/694979603954294788/photo/1"}
{"date_of_creation": "Sat Oct 06 07:26:37 +0000 2012", "tweet_id": "696783455821131776", "url": "https://twitter.com/SportsPete0/status/696781868075741185"}
{"date_of_creation": "Fri Jan 22 23:09:42 +0000 2010", "tweet_id": "696783385168052224", "url": "http://gizmodo.com/track-your-internet-connection-in-the-new-york
{"date_of_creation": "Mon Dec 31 14:36:45 +0000 2012", "tweet_id": "696783384853508097", "url": "http://gizmodo.com/track-your-internet-connection-in-the-new-york
{"date_of_creation": "Mon Dec 31 14:36:45 +0000 2012", "tweet_id": "696783368730603521", "url": "http://technewstube.com/gizmodo/676509/track-your-internet-connec
{"date_of_creation": "Tue Jun 10 13:40:13 +0000 2008", "tweet_id": "696783303584694272", "url": "http://www.gomplaces.com/"}
{"date_of_creation": "Fri Nov 07 12:49:34 +0000 2014", "tweet_id": "696783252187701253", "url": "http://www.nydailynews.com/new-york/brooklyn/exclusive-white-man-
{"date_of_creation": "Tue Nov 03 02:05:57 +0000 2015", "tweet_id": "696783201193345026", "url": "http://www.nbcnewyork.com/news/local/Subway-Train-Hijacked-Pranks
{"date_of_creation": "Sun Nov 29 14:20:56 +0000 2015", "tweet_id": "696783172114190336", "url": "https://twitter.com/brokeymcpoverty/status/696781767726993409"}
{"date_of_creation": "Sun Nov 21 03:11:13 +0000 2010", "tweet_id": "696783160382746625", "url": "http://www.theverge.com/2016/2/8/10938038/subspotting-app-nyc-sub
{"date_of_creation": "Thu Nov 14 05:20:55 +0000 2013", "tweet_id": "696783153336344580", "url": "http://m.aol.com/article/2015/11/25/jared-fogle-divorce-documents
{"date_of_creation": "Fri Dec 16 16:18:06 +0000 2011", "tweet_id": "696783116212396032", "url": "http://www.ebay.com/itm/like/22012505911?item=22012505911&lgeo=
{"date_of_creation": "Thu Mar 12 01:34:03 +0000 2009", "tweet_id": "696782981437005824", "url": "https://www.facebook.com/esedelab/?target_post=943386929050639&re
{"date_of_creation": "Thu Nov 12 21:13:41 +0000 2015", "tweet_id": "696782820988096513", "url": "https://www.instagram.com/p/BBiaAVWxymI/"}
{"date_of_creation": "Tue Dec 31 17:03:58 +0000 2013", "tweet_id": "696782734040158213", "url": "https://www.sfmta.com/projects-planning/projects/19th-avenue-m-oc
{"date_of_creation": "Mon Jun 01 01:36:32 +0000 2009", "tweet_id": "696782542750396416", "url": "https://twitter.com/subway/status/695744997304307712"}
{"date_of_creation": "Sun Mar 08 19:30:05 +0000 2009", "tweet_id": "696782528594776066", "url": "http://www.dailymail.co.uk/news/article-3437278/Shut-told-shut-Sh
{"date_of_creation": "Mon Sep 23 01:49:15 +0000 2013", "tweet_id": "696782492620058624", "url": "https://twitter.com/pattycake_yo/status/696778188018548736"}
{"date_of_creation": "Fri May 15 20:54:11 +0000 2009", "tweet_id": "696782335103078401", "url": "https://twitter.com/Gizmodo/status/696772231662178304"}
{"date_of_creation": "Tue Apr 02 14:14:53 +0000 2013", "tweet_id": "696782306900643840", "url": "http://brooklynda.org/2016/02/08/brooklyn-man-sentenced-to-12-yea
{"date_of_creation": "Mon Nov 25 09:53:52 +0000 2013", "tweet_id": "696782298696421376", "url": "http://www.dailymail.co.uk/tvshowbiz/article-3229336/Justin-Biebe
{"date_of_creation": "Wed Aug 21 22:11:01 +0000 2013", "tweet_id": "696782247517515776", "url": "http://gizmodo.com/track-your-internet-connection-in-the-new-york
{"date_of_creation": "Fri Apr 10 23:27:36 +0000 2009", "tweet_id": "696782191666135040", "url": "http://www.latimes.com/opinion/livable-city/la-ol-crenshaw-beverl
{"date_of_creation": "Tue Aug 17 19:21:00 +0000 2010", "tweet_id": "696782190286393344", "url": "http://gizmodo.com/track-your-internet-connection-in-the-new-york
{"date_of_creation": "Thu May 22 17:19:26 +0000 2014", "tweet_id": "696782066147577857", "url": "http://techseekr.com/?id=903590"}
{"date_of_creation": "Wed Oct 16 19:11:44 +0000 2013", "tweet_id": "696781893870731266", "url": "http://citty.com/2016/01/20/second-avenue-subway-creates-new-cons
{"date_of_creation": "Thu Dec 13 21:24:45 +0000 2012", "tweet_id": "696781887222738944", "url": "https://twitter.com/francis_petrel/status/696739977544212481"}
{"date_of_creation": "Mon Sep 27 22:53:45 +0000 2010", "tweet_id": "696781860433625089", "url": "https://www.instagram.com/p/BBiZkFalXO8/"}
{"date_of_creation": "Mon Sep 02 16:47:00 +0000 2013", "tweet_id": "696781771199901696", "url": "http://gizmodo.com/track-your-internet-connection-in-the-new-york
{"date_of_creation": "Wed Nov 28 14:29:16 +0000 2012", "tweet_id": "696781739147075584", "url": "https://www.facebook.com/ELEJERCITODELREYDELPOP/posts/10121553088
{"date_of_creation": "Sun Jan 10 11:10:26 +0000 2010", "tweet_id": "696781712693583873", "url": "http://subwayksa-memorygame.com/"}
{"date_of_creation": "Tue Jun 30 08:54:12 +0000 2015", "tweet_id": "696781711183630336", "url": "https://www.facebook.com/undersoundeventi"}
{"date_of_creation": "Fri Feb 13 16:20:32 +0000 2009", "tweet_id": "696781672952401920", "url": "https://twitter.com/bradTTC/status/696693893492924417"}
{"date_of_creation": "Fri Feb 25 03:15:56 +0000 2011", "tweet_id": "696781418177781760", "url": "http://www.dailymail.co.uk/news/article-3437278/Shut-told-shut-Sh
{"date_of_creation": "Fri Oct 05 15:23:21 +0000 2012", "tweet_id": "696781339807207424", "url": "http://gizmodo.com/track-your-internet-connection-in-the-new-york
{"date_of_creation": "Mon Oct 12 12:01:27 +0000 2015", "tweet_id": "696781256986660864", "url": "https://twitter.com/Sso_Moo/status/679957009773408258"}
{"date_of_creation": "Sat Aug 06 01:21:26 +0000 2011", "tweet_id": "696781113767792640", "url": "https://twitter.com/A_moneyFBG/status/696744930874695680"}
{"date_of_creation": "Thu Aug 06 03:50:34 +0000 2009", "tweet_id": "696781067240361984", "url": "https://www.instagram.com/p/BBiZNEaBoaI/"}
{"date_of_creation": "Fri Apr 18 14:42:36 +0000 2014", "tweet_id": "696780923438632960", "url": "http://metroeasy.metromates.com/lausanne-metro.html"}
{"date_of_creation": "Thu Mar 03 19:10:48 +0000 2011", "tweet_id": "696780906451709952", "url": "http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=eab390115
{"date_of_creation": "Mon Jan 26 03:44:26 +0000 2009", "tweet_id": "696780871869796352", "url": "http://www.barenakedislam.com/2016/02/07/disgusting-in-germany-no
{"date_of_creation": "Sun Jul 28 08:52:21 +0000 2013", "tweet_id": "696780798637293568", "url": "http://gizmodo.com/track-your-internet-connection-in-the-new-york
{"date_of_creation": "Mon Dec 28 13:10:43 +0000 2015", "tweet_id": "696780737563889665", "url": "http://cur.lv/v6ja5"}
```

Figure 4: Sample Json data

3.5 Outputs

Output 1

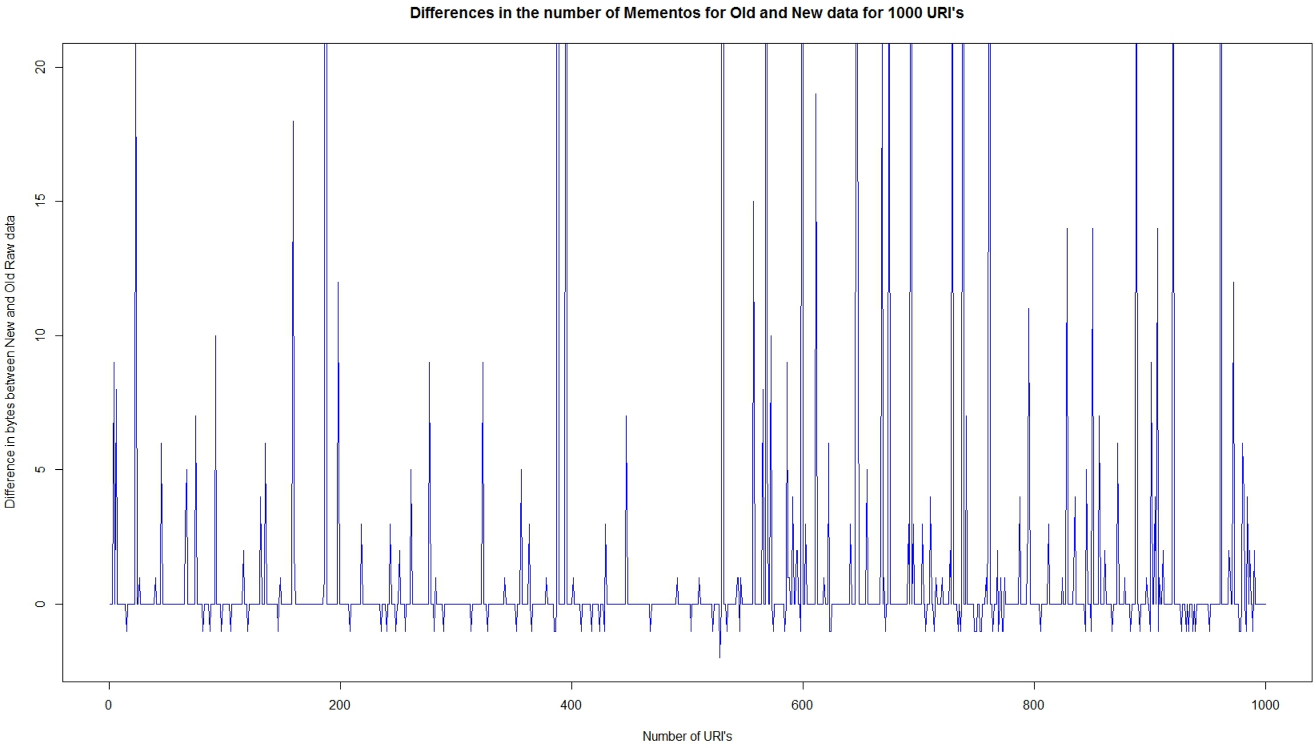


Figure 5: Line graph showing differences in the number of Mementos for Old and New data for 1000 URI's

4 Problem 4

```
4.  Repeat A3, Q1.  Compare the resulting text from February to
the text you have now.  Do all 1000 URIs still return a ‘‘200 OK’’
as their final response (i.e., at the end of possible redirects)?

Create two graphs similar to that described in Q3, except this
time the y-axis corresponds to difference in bytes (and not difference
in TimeMap magnitudes).  For the first graph, use the difference
in the raw (unprocessed) results.  For the second graph, use the
difference in the processed (as per A3, Q1) results.

Of the URIs that still terminate in a ‘‘200 OK’’ response, pick the
top 3 most changed (processed) pairs of pages and use the Unix
‘‘diff’’ command to explore the differences in the version pairs.
```

4.1 Solution

1. In this question I need to compare the resulting text from my 3rd Assignment and present resulting text.
2. In order to do this I have taken code from 3rd Assignment and executed it again and the code for this can be seen in the following listing5.
3. This gives me a new set of raw and processed data files with complete updated text. Now I need to find the difference in the file sizes in bytes for each URI.
4. In order to subtract the old file sizes from the new file sizes I wrote a code for it which can be seen in the listing6.
5. This was a tough task because I need to do it for new raw and processed data and old raw and processed data which is very confusing as the data files are pretty similar.
6. I also checked for the status codes of all the URI’s using the code in listing7. I have found out that there are 891 URI’s which give a status code of “200”.
7. The list of other status codes can be seen in the table below.
8. I have then plotted a line graph using R which shows the differences in bytes for the files in old and new data. This code for this can be found in the listing8.
9. Line graph showing the differences for the files sizes in bytes for raw data can be found in the fig6.
10. Line graph showing the differences for the files sizes in bytes for processed data can be found in the fig7.
11. Then my last task is to take a list of all URI’s which return status code as “200” and from that list I need to pick top 3 most changed data files. This is done only for processed data files.
12. The top 3 URI’s whose resulting text are mostly changed are “http://www.gaynycdad.com/2016/02/giveaway-25-walmartsams-club-gift-card.html’,“http://peanutbutterandwhine.com/februarys-50-your-way-giveaway-single-blog/” and “http://newsbunch.com/tech-news/track-cell-service-along-your-subway-route-with-this-new-app/”.
13. The changes for these particular URI’s are compared using “vim -d newdatafile olddatafile”.
14. When the above code in executed in putty it gives me the changes that occured in their text which are shown below.
15. The changes in the text files for 1st top most changed URI can be seen in the fig8.
16. The changes in the text files for 2nd top most changed URI can be seen in the fig9.
17. The changes in the text files for 3rd top most changed URI can be seen in the fig10.

Table 1: Status code and their count

Status codes	count
417	1
423	1
200	891
403	11
404	32
503	48
500	1
410	3

4.2 Code Listing

Code Listing 1

```
import json
import commands
import hashlib

file_1 = open('my_json_data.json','r')
count = 0
for each_line in file_1.readlines():
        dumy_line=json.loads(each_line)
        url = dumy_line['url']
        hash = hashlib.md5(url.encode())
        final_hash = hash.hexdigest()
        count = count +1
        file_name_1= "Raw"+'-'+str(count) + '-' + final_hash + '.txt'
        co_1 = 'curl -s -L ' + url + ' > ./Raw_Htmldata/' + file_name_1
        commands.getoutput(co_1)
        file_name_2= "processed"+'-'+str(count) + '-' + final_hash + '.txt' #Naming a file
        co_2 = 'lynx -dump -force_html ' + url + ' > ./processed_htmldata/' + file_name_2 # writes files into processed_htmldata folder
        commands.getoutput(co_2)
        #print url,' ',count,' ',file_name_1
```

Listing 5: Python Code for getting raw and processed files for each URI

Code Listing 2

```
input1=open(”raw_data_size.txt”,”r”)
input2=open(”raw_data_size_old.txt”,”r”)
input3=open(”processed_data_size.txt”,”r”)
input4=open(”processed_data_size_old.txt”,”r”)
output_raw=open(”raw_result.txt”,”w”)
output_processed=open(”processed_result.txt”,”w”)

array1=[]
array2=[]
array3=[]
array4=[]
raw_result=[]
processed_result=[]


for line in input1:
        array1.append(int(line))

for line in input2:
        array2.append(int(line))
for line in input3:
        array3.append(int(line))

for line in input4:
        array4.append(int(line))

raw_result = [new - old for new, old in zip(array1, array2)]
processed_result = [new - old for new, old in zip(array3, array4)]

for value in raw_result:
        output_raw.write(str(value)+’\n’)

for value in processed_result:
        output_processed.write(str(value)+’\n’)

# for i,j in array1,array2:
        # value=i-j
# raw_result=value
# print raw_result

# for k,l in array3,array4:
        # value1=k-l
# processed_result=value1
# print processed_result

        # for line1 in input2:
                # value=int(line)-int(line1)
                # output_raw.write(str(value)+’\n’)

# for line2 in input3:
        # for line3 in input4:
                # value1=int(line2)-int(line3)
                # output_processed.write(str(value1)+’\n’)
```

Listing 6: Python Code subtracting the new byte count and old byte count for raw and processed data files for 1000 URI’s

Code Listing 3

```
import json
import requests

input= open("my_json_data.json","r")
output=open("status_codes.json","w")
dictionary={}
for uri in input.readlines():
        each_line = json.loads(uri)
        url = each_line['url']
        try:
                status=requests.get(url)
                print status.status_code
                if status.status_code in dictionary:
                        dictionary[status.status_code] +=1
                else:
                        dictionary[status.status_code] =1
        except Exception, e:
                print e
                continue
output.write(json.dumps(dictionary))
output.close()
```

Listing 7: Python Code for checking the status codes of all the URI's

Code Listing 4

```
data = scan(”raw_result.txt”)
plot(data,xlab=”Number of URI's”,ylab=”Difference in bytes between New and Old Raw data”,
    main=”Differences in bytes for raw data for 1000 URI's”,xlim=c(0,1000),col=”blue”,type=”
    l”)
```

Listing 8: R code to plot a line graph to show the differences in bytes for old and new data

4.3 Outputs

Output 1

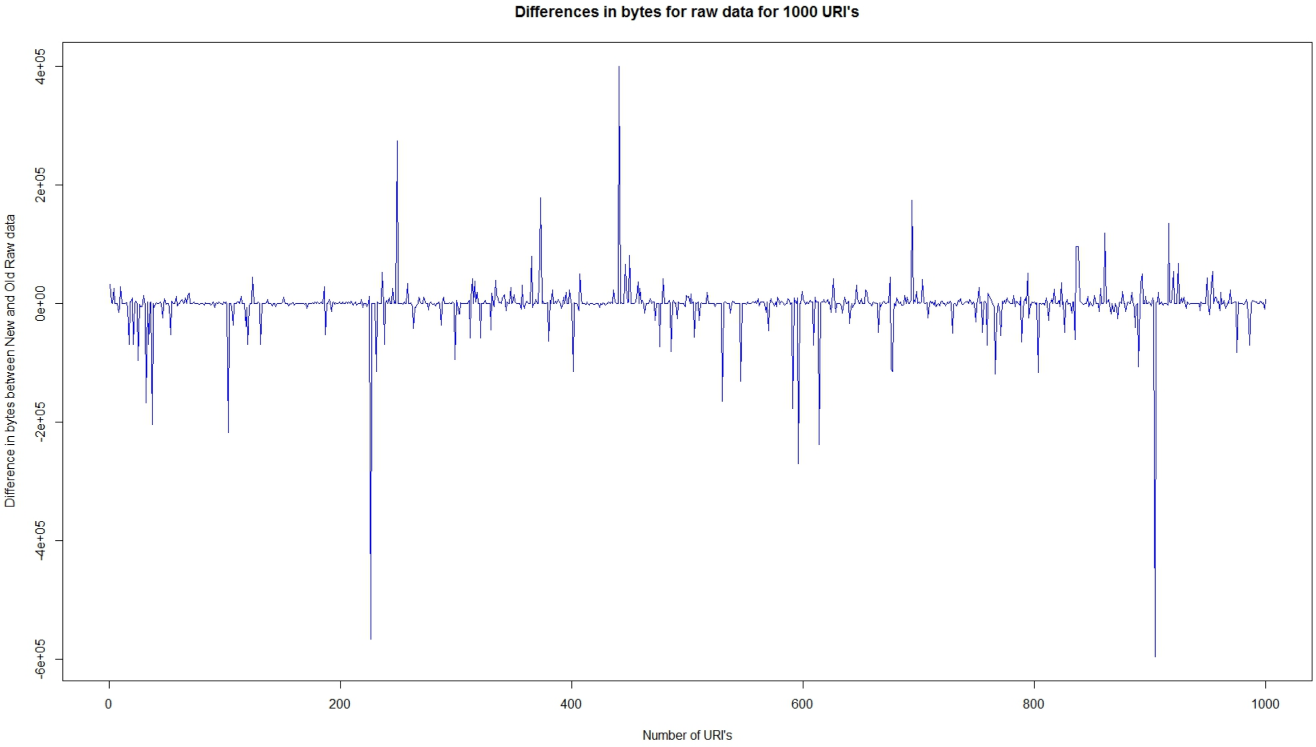


Figure 6: Line graph showing differences in bytes for raw data for 1000 URI's

Output 2

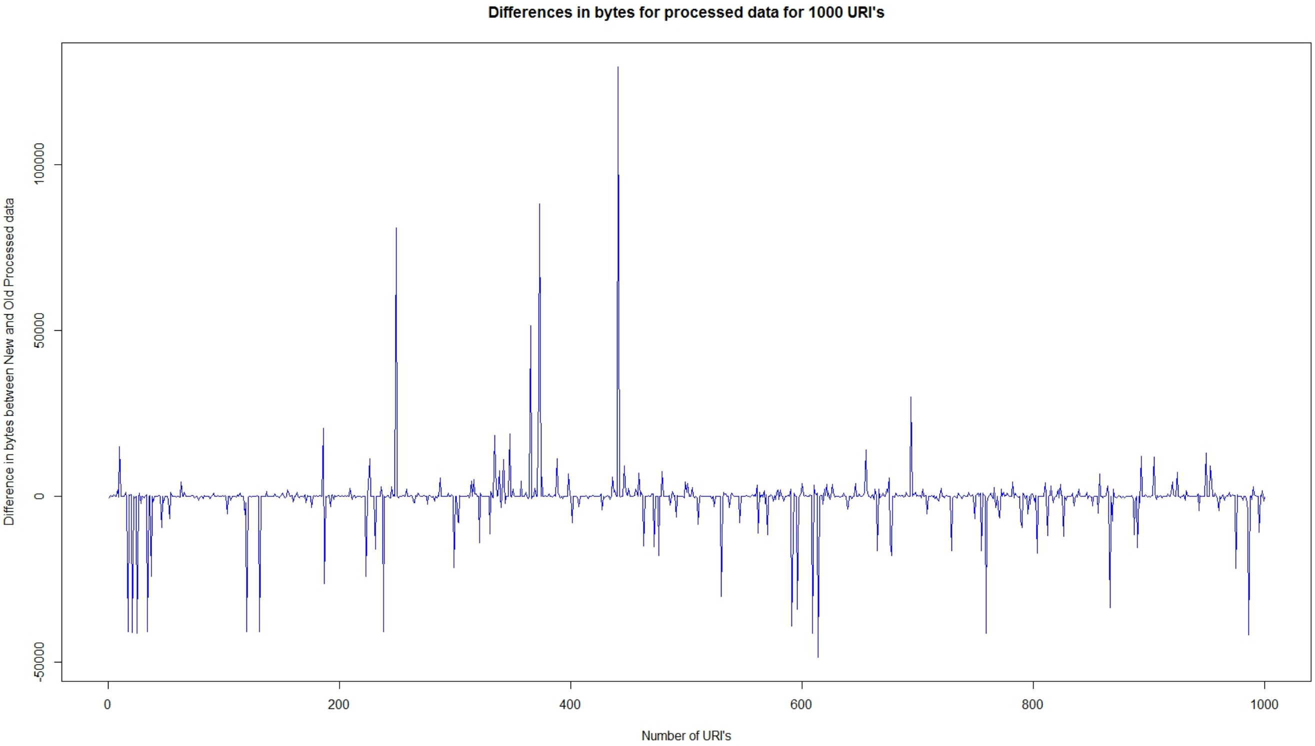
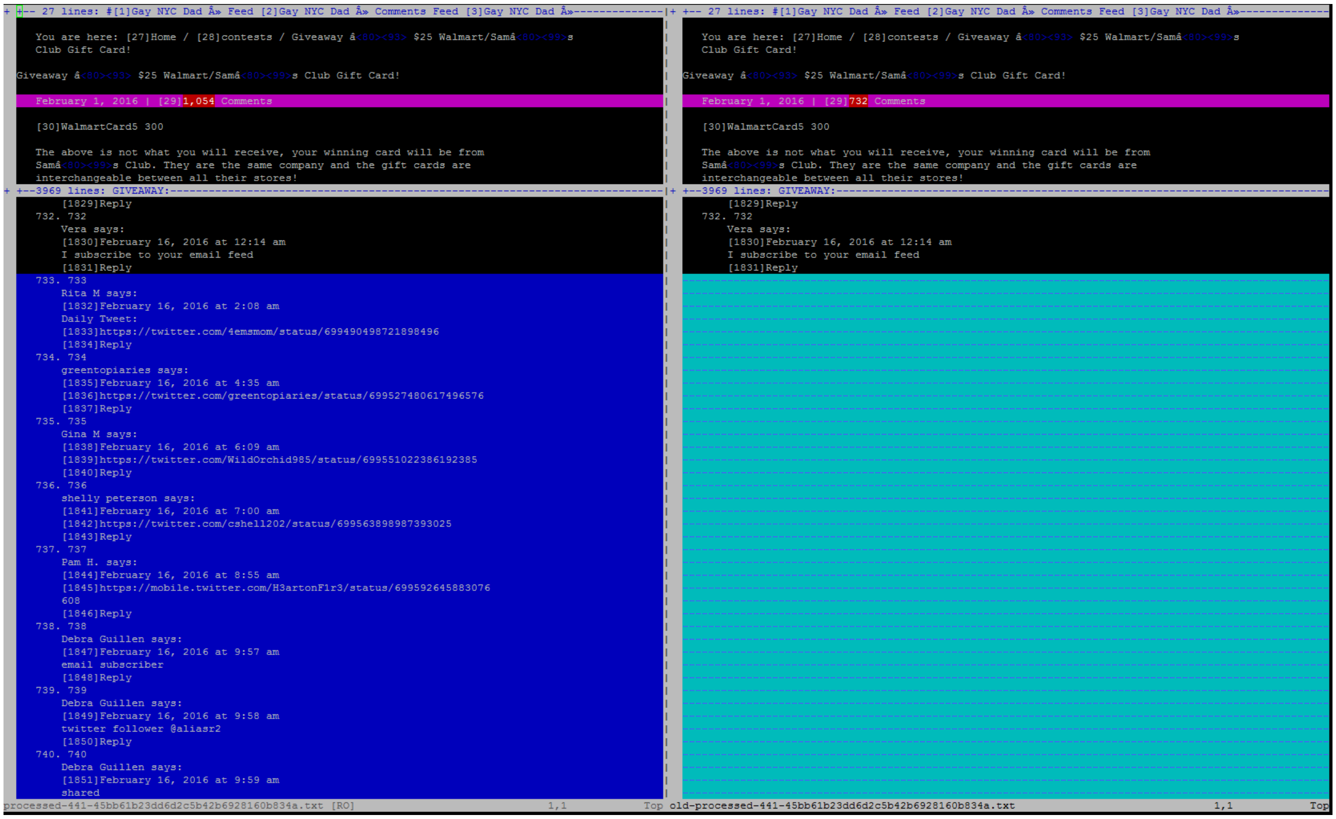


Figure 7: Line graph showing differences in bytes for processed data for 1000 URI's

Output 3

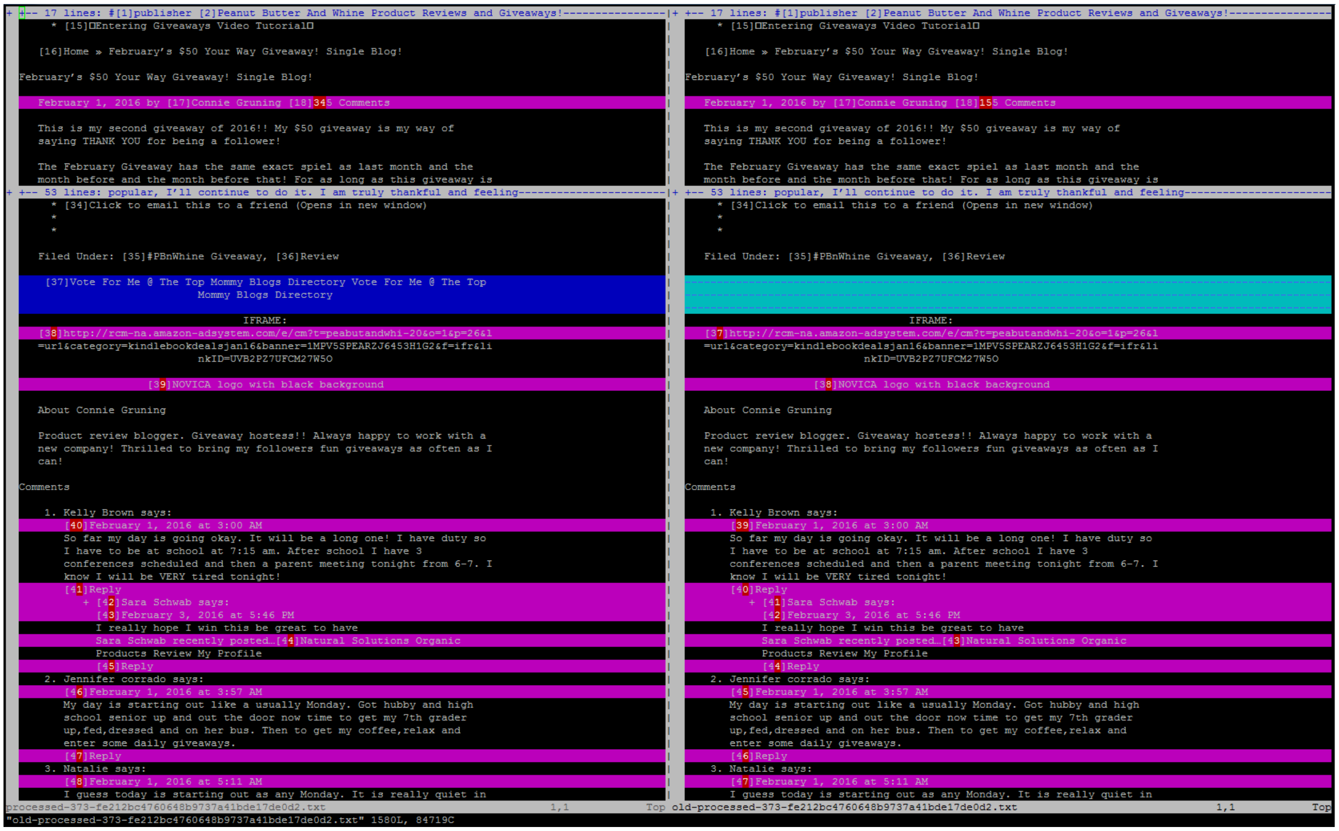


```
+ +-- 27 lines: #[1]Gay NYC Dad Â» Feed [2]Gay NYC Dad Â» Comments Feed [3]Gay NYC Dad Â»--------------

   You are here: [27]Home / [28]contests / Giveaway â<80><93> $25 Walmart/Samâ<80><99>s
   Club Gift Card!

Giveaway â<80><93> $25 Walmart/Samâ<80><99>s Club Gift Card!

   February 1, 2016 | [29]1,054 Comments

   [30]WalmartCard5 300

   The above is not what you will receive, your winning card will be from
   Samâ<80><99>s Club. They are the same company and the gift cards are
   interchangeable between all their stores!
+ +--3969 lines: GIVEAWAY:------------------------------------------------------------------------
       [1829]Reply
   732. 732
       Vera says:
       [1830]February 16, 2016 at 12:14 am
       I subscribe to your email feed
       [1831]Reply
   733. 733
       Rita M says:
       [1832]February 16, 2016 at 2:08 am
       Daily Tweet:
       [1833]https://twitter.com/4emsmom/status/699490498721898496
       [1834]Reply
   734. 734
       greentopiaries says:
       [1835]February 16, 2016 at 4:35 am
       [1836]https://twitter.com/greentopiaries/status/699527480617496576
       [1837]Reply
   735. 735
       Gina M says:
       [1838]February 16, 2016 at 6:09 am
       [1839]https://twitter.com/WildOrchid985/status/699551022386192385
       [1840]Reply
   736. 736
       shelly peterson says:
       [1841]February 16, 2016 at 7:00 am
       [1842]https://twitter.com/cshell202/status/699563898987393025
       [1843]Reply
   737. 737
       Pam H. says:
       [1844]February 16, 2016 at 8:55 am
       [1845]https://mobile.twitter.com/H3artonF1r3/status/699592645883076
       608
       [1846]Reply
   738. 738
       Debra Guillen says:
       [1847]February 16, 2016 at 9:57 am
       email subscriber
       [1848]Reply
   739. 739
       Debra Guillen says:
       [1849]February 16, 2016 at 9:58 am
       twitter follower @aliasr2
       [1850]Reply
   740. 740
       Debra Guillen says:
       [1851]February 16, 2016 at 9:59 am
       shared
processed-441-45bb61b23dd6d2c5b42b6928160b834a.txt [RO]                    1,1            Top
```

```
+ +-- 27 lines: #[1]Gay NYC Dad Â» Feed [2]Gay NYC Dad Â» Comments Feed [3]Gay NYC Dad Â»-------------

   You are here: [27]Home / [28]contests / Giveaway â<80><93> $25 Walmart/Samâ<80><99>s
   Club Gift Card!

Giveaway â<80><93> $25 Walmart/Samâ<80><99>s Club Gift Card!

   February 1, 2016 | [29]732 Comments

   [30]WalmartCard5 300

   The above is not what you will receive, your winning card will be from
   Samâ<80><99>s Club. They are the same company and the gift cards are
   interchangeable between all their stores!
+ +--3969 lines: GIVEAWAY:------------------------------------------------------------------------
       [1829]Reply
   732. 732
       Vera says:
       [1830]February 16, 2016 at 12:14 am
       I subscribe to your email feed
       [1831]Reply
old-processed-441-45bb61b23dd6d2c5b42b6928160b834a.txt                     1,1            Top
```

Figure 8: differences in old and new data plotted by vim -d for 1st top most changed URI

Output 4



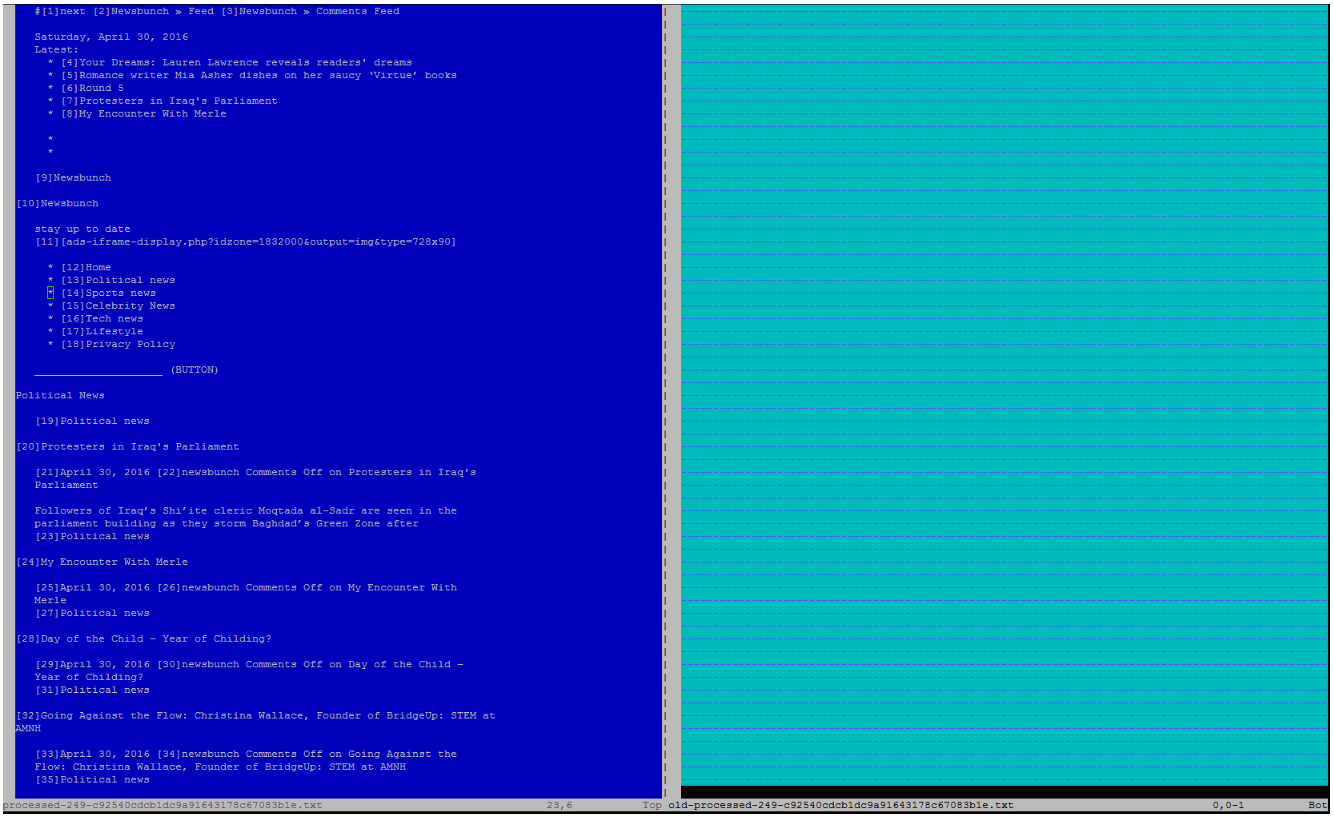
```
+ +-- 17 lines: #[1]publisher [2]Peanut Butter And Whine Product Reviews and Giveaways!----------------
     * [15]□Entering Giveaways Video Tutorial□

   [16]Home » February’s $50 Your Way Giveaway! Single Blog!

February’s $50 Your Way Giveaway! Single Blog!

   February 1, 2016 by [17]Connie Gruning [18]345 Comments

   This is my second giveaway of 2016!! My $50 giveaway is my way of
   saying THANK YOU for being a follower!

   The February Giveaway has the same exact spiel as last month and the
   month before and the month before that! For as long as this giveaway is
+ +-- 53 lines: popular, I’ll continue to do it. I am truly thankful and feeling-----------------------
     * [34]Click to email this to a friend (Opens in new window)
     *
     *

   Filed Under: [35]#PBnWhine Giveaway, [36]Review

    [37]Vote For Me @ The Top Mommy Blogs Directory Vote For Me @ The Top
                           Mommy Blogs Directory

                                    IFRAME:
   [38]http://rcm-na.amazon-adsystem.com/e/cm?t=peabutandwhi-20&o=1&p=26&l
   =ur1&category=kindlebookdealsjan16&banner=1MPV5SPEARZJ6453H1G2&f=ifr&li
                           nkID=UVB2PZ7UFCM27W5O

                    [39]NOVICA logo with black background

   About Connie Gruning

   Product review blogger. Giveaway hostess!! Always happy to work with a
   new company! Thrilled to bring my followers fun giveaways as often as I
   can!

Comments

    1. Kelly Brown says:
       [40]February 1, 2016 at 3:00 AM
       So far my day is going okay. It will be a long one! I have duty so
       I have to be at school at 7:15 am. After school I have 3
       conferences scheduled and then a parent meeting tonight from 6-7. I
       know I will be VERY tired tonight!
       [41]Reply
          + [42]Sara Schwab says:
            [43]February 3, 2016 at 5:46 PM
            I really hope I win this be great to have
            Sara Schwab recently posted…[44]Natural Solutions Organic
            Products Review My Profile
            [45]Reply
    2. Jennifer corrado says:
       [46]February 1, 2016 at 3:57 AM
       My day is starting out like a usually Monday. Got hubby and high
       school senior up and out the door now time to get my 7th grader
       up,fed,dressed and on her bus. Then to get my coffee,relax and
       enter some daily giveaways.
       [47]Reply
    3. Natalie says:
       [48]February 1, 2016 at 5:11 AM
       I guess today is starting out as any Monday. It is really quiet in
processed-373-fe212bc4760648b9737a41bde17de0d2.txt                            1,1           Top
```

```
+ +-- 17 lines: #[1]publisher [2]Peanut Butter And Whine Product Reviews and Giveaways!----------------
     * [15]□Entering Giveaways Video Tutorial□

   [16]Home » February’s $50 Your Way Giveaway! Single Blog!

February’s $50 Your Way Giveaway! Single Blog!

   February 1, 2016 by [17]Connie Gruning [18]155 Comments

   This is my second giveaway of 2016!! My $50 giveaway is my way of
   saying THANK YOU for being a follower!

   The February Giveaway has the same exact spiel as last month and the
   month before and the month before that! For as long as this giveaway is
+ +-- 53 lines: popular, I’ll continue to do it. I am truly thankful and feeling-----------------------
     * [34]Click to email this to a friend (Opens in new window)
     *
     *

   Filed Under: [35]#PBnWhine Giveaway, [36]Review

----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
                                    IFRAME:
   [37]http://rcm-na.amazon-adsystem.com/e/cm?t=peabutandwhi-20&o=1&p=26&l
   =ur1&category=kindlebookdealsjan16&banner=1MPV5SPEARZJ6453H1G2&f=ifr&li
                           nkID=UVB2PZ7UFCM27W5O

                    [38]NOVICA logo with black background

   About Connie Gruning

   Product review blogger. Giveaway hostess!! Always happy to work with a
   new company! Thrilled to bring my followers fun giveaways as often as I
   can!

Comments

    1. Kelly Brown says:
       [39]February 1, 2016 at 3:00 AM
       So far my day is going okay. It will be a long one! I have duty so
       I have to be at school at 7:15 am. After school I have 3
       conferences scheduled and then a parent meeting tonight from 6-7. I
       know I will be VERY tired tonight!
       [40]Reply
          + [41]Sara Schwab says:
            [42]February 3, 2016 at 5:46 PM
            I really hope I win this be great to have
            Sara Schwab recently posted…[43]Natural Solutions Organic
            Products Review My Profile
            [44]Reply
    2. Jennifer corrado says:
       [45]February 1, 2016 at 3:57 AM
       My day is starting out like a usually Monday. Got hubby and high
       school senior up and out the door now time to get my 7th grader
       up,fed,dressed and on her bus. Then to get my coffee,relax and
       enter some daily giveaways.
       [46]Reply
    3. Natalie says:
       [47]February 1, 2016 at 5:11 AM
       I guess today is starting out as any Monday. It is really quiet in
old-processed-373-fe212bc4760648b9737a41bde17de0d2.txt                        1,1           Top
```

```
"old-processed-373-fe212bc4760648b9737a41bde17de0d2.txt" 1580L, 84719C
```

Figure 9: differences in old and new data plotted by vim -d for 2nd top most changed URI

Output 5



```
   #[1]next [2]Newsbunch » Feed [3]Newsbunch » Comments Feed

   Saturday, April 30, 2016
   Latest:
     * [4]Your Dreams: Lauren Lawrence reveals readers' dreams
     * [5]Romance writer Mia Asher dishes on her saucy ‘Virtue’ books
     * [6]Round 5
     * [7]Protesters in Iraq's Parliament
     * [8]My Encounter With Merle

     *
     *

   [9]Newsbunch

[10]Newsbunch

   stay up to date
   [11][ads-iframe-display.php?idzone=1832000&output=img&type=728x90]

     * [12]Home
     * [13]Political news
     * [14]Sports news
     * [15]Celebrity News
     * [16]Tech news
     * [17]Lifestyle
     * [18]Privacy Policy

   ____________________ (BUTTON)

Political News

   [19]Political news

[20]Protesters in Iraq's Parliament

   [21]April 30, 2016 [22]newsbunch Comments Off on Protesters in Iraq's
   Parliament

   Followers of Iraq’s Shi’ite cleric Moqtada al-Sadr are seen in the
   parliament building as they storm Baghdad’s Green Zone after
   [23]Political news

[24]My Encounter With Merle

   [25]April 30, 2016 [26]newsbunch Comments Off on My Encounter With
   Merle
   [27]Political news

[28]Day of the Child - Year of Childing?

   [29]April 30, 2016 [30]newsbunch Comments Off on Day of the Child -
   Year of Childing?
   [31]Political news

[32]Going Against the Flow: Christina Wallace, Founder of BridgeUp: STEM at
AMNH

   [33]April 30, 2016 [34]newsbunch Comments Off on Going Against the
   Flow: Christina Wallace, Founder of BridgeUp: STEM at AMNH
   [35]Political news
processed-249-c92540cdcb1dc9a91643178c67083b1e.txt        23,6        Top old-processed-249-c92540cdcb1dc9a91643178c67083b1e.txt        0,0-1        Bot
```

Figure 10: differences in old and new data plotted by vim -d for 3rd top most changed URI

Bibliography

[1] Alex Leach, Comparing two files in Vim, http://unix.stackexchange.com/questions/1386/comparing-two-files-in-vim, 2012

[2] Walter Traspadini, numpredict.py, https://github.com/uolter/PCI/blob/master/chapter8/numpredict.py, 2013

[3] Artsiom Rudzenka, Getting file size in Python?, http://stackoverflow.com/questions/6591931/getting-file-size-in-python, 2011

[4] Mike Housky, Cosine Similarity between 2 Number Lists, http://stackoverflow.com/questions/18424228/cosine-similarity-between-2-number-lists, 2013