

INTRODUCTION TO WEB SCIENCES: Assignment 10

Manoj Chandra Kompalli

29 April 2016

Contents

1	Question 1:	2
1.1	Approach	2
1.2	Code Listing	3
1.2.1	makelist.py	3
1.2.2	numpredict.py	3
1.3	Output	5
1.3.1	F-Measure blog output	5
1.3.2	Web science blog output	6
2	Question 2:	7
2.1	Approach	7
3	Question 3:	8
3.1	Approach	9
3.2	Code Listing	9
3.2.1	memcount.py	9
3.3	Generating graph	10
3.3.1	Rcommands.R	10
3.4	Output Files	10
3.4.1	Output graph	10
4	Question 4:	11
4.1	Approach	11
4.2	Code Listing	11
4.2.1	For extracting raw and processed URIs (extract.py)	11
4.2.2	To find the size of the size of either raw or processed files(size.py)	12
4.2.3	To find the number of URIs exiting with status 200 (statuscode.py)	12
4.2.4	Generating graph	13
4.3	Output	13

1 Question 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 `knestimate()` 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 Approach

- I have used my previous blog matrix which was in tab separated format and made a similar one in JSON format
- I had used `numpredict.py` from Programming Collective intelligence and added some more functionality like cosine similarity between vectors.
- By using sorted distances `knestimate()` method computes the nearest neighbours
- I have used the `knestimate()` to find the nearest neighbours of both the F-Measure blog and web science blog
- Output gives the nearest neighbours of F-measure and Web science blogs

1.2 Code Listing

1.2.1 makelist.py

```
1 import json
2 file1= open('blogmatrix.txt','r')
3 file2 = open('bloglist ','w')
4 output = []
5 count = 0
6 for line in file1:
7     count=count+ 1
8     if count >1:
9         mydict = {}
10        line = line.strip()
11        input = line.split('\t')
12        # print input
13        name= input[0]
14        input.pop(0)
15        generated = input
16        # print name
17        # print generated
18        mydict[name] = generated
19        output.append(mydict)
20 file2.write(json.dumps(output))
```

1.2.2 numpredict.py

```
1 from random import random, randint
2 import math
3 import json
4
5 def cosine_similarity(v1,v2):
6     "compute cosine similarity of v1 to v2: (v1 dot v2) / (||v1||*||v2||)"
7     sumxx, sumxy, sumyy = 0, 0, 0
8     for i in range(0,len(v1)-1):
9         x = int(v1[i]); y = int(v2[i])
10        sumxx += x*x
11        sumyy += y*y
12        sumxy += x*y
13    return sumxy/math.sqrt(sumxx*sumyy)
14
15
16 def getdistances(data, vec1):
17     distancelist=[]
18
19     # Loop over every item in the dataset
20     for i in data:
21         # print item
22         for subitem in i:
23             if subitem != 'F-Measure':
24                 vec2= i[subitem]
25
26         # Add the distance and the index
27         distancelist.append((cosine_similarity(vec1, vec2), i))
28
29     # Sort by distance
30     distancelist.sort()
```

```

31
32     return distancelist
33
34 def knnestimate(data, vec1, k=20):
35     # Get sorted distances
36     print 'For Web Science blog and k=20'
37     dlist=getdistances(data, vec1)
38     avg=0.0
39     # print dlist
40     # Take the average of the top k results
41     for i in range(k):
42         idx=dlist[i]
43         value = idx[0]
44         for item in idx[1]:
45             blogname= item
46             print blogname + '\t' + str(value)
47
48 def main():
49     file1= open('bloglist', 'r')
50     data = json.load(file1)
51
52     for data1 in data:
53
54         for data2 in data1:
55             if data2 == 'Web Science and Digital Libraries Research Group':
56                 vec1= data1[data2]
57                 knnestimate(data, vec1)
58
59 main()

```

1.3 Output

1.3.1 F-Measure blog output

```
1 atria:~/wsprograms/a10/q1> python numpredict.py
2 k=1
3 INDIEohren.!      0.0284258547722
4 atria:~/wsprograms/a10/q1> clear
5 atria:~/wsprograms/a10/q1> python numpredict.py
6 For F-Measure blog and k=1
7 INDIEohren.!      0.0284258547722
8 atria:~/wsprograms/a10/q1> python numpredict.py
9 For F-Measure blog and k=2
10 INDIEohren.!      0.0284258547722
11 MR. BEAUTIFUL TRASH ART 0.035023925836
12 atria:~/wsprograms/a10/q1> python numpredict.py
13 For F-Measure blog and k=5
14 INDIEohren.!      0.0284258547722
15 MR. BEAUTIFUL TRASH ART 0.035023925836
16 MARISOL 0.0421254990846
17 How to be an artist and still pass for normal 0.0615408984138
18 ORGANMYTH      0.063728972963
19 atria:~/wsprograms/a10/q1> clear
20 atria:~/wsprograms/a10/q1> python numpredict.py
21 For F-Measure blog and k=10
22 INDIEohren.!      0.0284258547722
23 MR. BEAUTIFUL TRASH ART 0.035023925836
24 MARISOL 0.0421254990846
25 How to be an artist and still pass for normal 0.0615408984138
26 ORGANMYTH      0.063728972963
27 IoTube      :)      0.076906363689
28 theindiefriend 0.0798683749435
29 A H T A P O T  0.118655531931
30 adrianoblog   0.123239578098
31 Rod Shone     0.133212524679
32 atria:~/wsprograms/a10/q1> python numpredict.py
33 For F-Measure blog and k=20
34 INDIEohren.!      0.0284258547722
35 MR. BEAUTIFUL TRASH ART 0.035023925836
36 MARISOL 0.0421254990846
37 How to be an artist and still pass for normal 0.0615408984138
38 ORGANMYTH      0.063728972963
39 IoTube      :)      0.076906363689
40 theindiefriend 0.0798683749435
41 A H T A P O T  0.118655531931
42 adrianoblog   0.123239578098
43 Rod Shone     0.133212524679
44 What Am I Doing?      0.137021765128
45 Stonehill Sketchbook  0.137508143489
46 "DANCING IN CIRCLES"  0.146334368545
47 sweeping the kitchen  0.14683876125
48 F-Measure      0.151098106517
49                  0.151098106517
50 FlowRadio Playlists (and Blog) 0.154577490878
51 Spinitron Blog  0.15802910721
52
53
```

```

54 0.161885545609
55 Boggle Me Thursday 0.174330461107

```

1.3.2 Web science blog output

```

1 atria:~/wsprograms/a10/q1> python numpredict.py
2 For Web Science blog and k=1
3 Rod Shone 0.0248283154805
4 atria:~/wsprograms/a10/q1> python numpredict.py
5 For Web Science blog and k=2
6 Rod Shone 0.0248283154805
7 adrianoblog 0.0601947926141
8 atria:~/wsprograms/a10/q1> python numpredict.py
9 For Web Science blog and k=5
10 Rod Shone 0.0248283154805
11 adrianoblog 0.0601947926141
12 INDIEohren.! 0.0609646307293
13 IoTube :) 0.063726921167
14 MARISOL 0.0773923984209
15 atria:~/wsprograms/a10/q1> python numpredict.py
16 For Web Science blog and k=10
17 Rod Shone 0.0248283154805
18 adrianoblog 0.0601947926141
19 INDIEohren.! 0.0609646307293
20 IoTube :) 0.063726921167
21 MARISOL 0.0773923984209
22 sweeping the kitchen 0.0837138528161
23 If You Give a Girl a Camera... 0.0874223257725
24
25
26 0.109439536637
27 theindiefriend 0.111459097808
28 "DANCING IN CIRCLES" 0.128160527359
29 atria:~/wsprograms/a10/q1> python numpredict.py
30 For Web Science blog and k=20
31 Rod Shone 0.0248283154805
32 adrianoblog 0.0601947926141
33 INDIEohren.! 0.0609646307293
34 IoTube :) 0.063726921167
35 MARISOL 0.0773923984209
36 sweeping the kitchen 0.0837138528161
37 If You Give a Girl a Camera... 0.0874223257725
38
39
40 0.109439536637
41 theindiefriend 0.111459097808
42 "DANCING IN CIRCLES" 0.128160527359
43 How to be an artist and still pass for normal 0.135218576611
44 GLI Press 0.14290805507
45 Azul Valentina 0.14693001748
46 F-Measure 0.149696250775
47 0.149696250775
48 Boggle Me Thursday 0.150102315139
49 Lo importante es que estes t bien 0.150707640184
50 MR. BEAUTIFUL TRASH ART 0.151538277269
51 What Am I Doing? 0.154256917386
52 Spinitron Blog 0.155694593625

```

2 Question 2:

2.1 Approach

Not Attempted

3 Question 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 Approach

- I have re-downloaded all the 1000 time maps from the previous assignment
- I have compared the mementos generated previously with the new mementos for each URI
- the generated graph shows that in general mementos have only increased and only in rare cases decreased by 1 or 2

3.2 Code Listing

3.2.1 memcount.py

```
1 #!/usr/local/bin/python3
2 import re
3 import sys
4 import urllib2
5 import json
6
7 mymementos = re.compile(r'rel.*?=.*?"memento".*?')#use regular expressions to find
   mementos
8 file3=open('abovezerocounts2.json','w')
9 file4=open('abovezerourls2.json','w')
10
11 def getTimeMap(url):
12     mem_url = "http://mementoproxy.cs.odu.edu/aggr/timemap/link/1/" + url #plug in the
   url to a timemap
13     try:
14         response = urllib2.urlopen(mem_url)
15         timemap = response.read()
16     except urllib2.HTTPError:
17         timemap = None
18     return timemap
19
20 def countMementos(mem_url):
21     time_map = getTimeMap(mem_url)
22     if not time_map:# if no time maps
23         count=0
24     else:
25
26
27         count=len(mymementos.findall(str(time_map)))#finds the count of all mementos per
   url
28         if count>0:
29             file3.write("%s\n"% count)
30             file4.write("%s\n"% time_map)
31         #print count
32     return count
33
34 if __name__=="__main__":
35     file1=open('output.json','r')# input a json file that contains 1000 urls
36     file2=open('memcount2.json','w')
37     #memcountlist=[]
38     for line in file1.readlines():
39         one_line = json.loads(line)# loads a json object
40         link = one_line['link']
```

```

41 counter=countMementos(link)# counter has count of the urls
42 file2.write("%s"% counter)#outputs count of mementos of each url to a json file
43 file2.write("\r\n")
44 #for item in memcountlist:
45
46
47 file1.close()
48 file2.close()

```

3.3 Generating graph

3.3.1 Rcommands.R

```

1 d = read.table('dif2.json',col.name=c("mementos"))
2 plot(d$mementos,xlab="Number of URI's",ylab="Difference between Old and New Mementos",
    xlim=c(0,1000),ylim=c(-2,50),type="l")

```

3.4 Output Files

3.4.1 Output graph

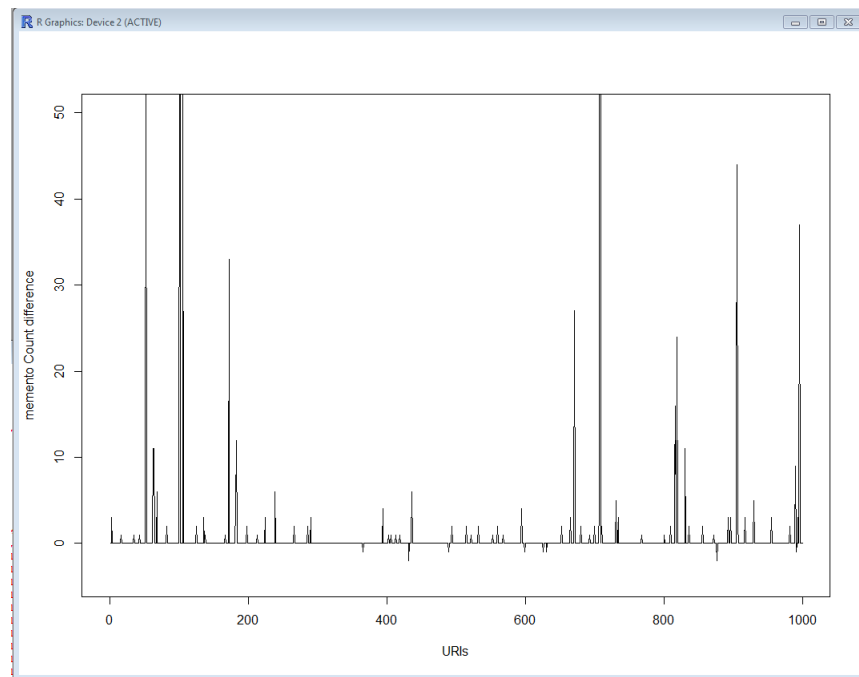


Figure 1: Memento difference

4 Question 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 Approach

- I have extracted raw and processed URIs again which returns status 200.
- I have found the size of old raw data, new raw data, old processed data, new processed data using size.py.
- I have found the difference in sizes of old raw data and new raw data and generated a graph.
- Similarly, I have found the difference in sizes of old processed data and new processed data and generated a graph .
- I wrote a program statuscode.py to find the status of the existing URIs.
- I have found that only 784 out of entire 1000 URIs had exited with status code 200.
- I had found the top three maximum distances of the old and new processed URIs using vim -d command.
- The comparison is shown in three seperate screenshots

4.2 Code Listing

4.2.1 For extracting raw and processed URIs (extract.py)

```
1 import re
2 import os
3 import json
4
5 if __name__=="__main__":
6     f2name='raw.txt'
7     f3name='processed.txt'
8     count=0
9     count1=0
10    file1=open('links.json','r')# file which contains 1000 uris
11
12    for line in file1.readlines():
13        count=count+1
14        newfile=str(count)+f2name #concatenates counter value to a string
15        one_line = json.loads(line)
16        link = one_line['link']
17
```

```

18 cmd="curl -s -L "+ link+" >./rawurls/"+ newfile # shell script to print raw html
    content of each uri
19 os.system(cmd)
20 for line in file1.readlines():
21     count1=count1+1
22     newfile1=str(count1)+f3name #concatenates counter value to a string
23     one_line1 = json.loads(line)
24     link1 = one_line1['link']
25
26 cmd1="lynx -dump -force_html "+ link1+" >./processedurls/"+ newfile1 # shell
    script to print processed html content of each uri
27 os.system(cmd1)

```

4.2.2 To find the size of the size of either raw or processed files(size.py)

```

1 import os
2
3
4 path= "/home/mkompal/wsprograms/a10/q4/rawurls/"
5
6 file1 = open('newraw', 'w')
7 directory = os.listdir(path)
8
9 for file in directory:
10
11     final = path + file
12     print final
13     size= os.path.getsize(final)
14     file1.write(str(size)+'\n')

```

4.2.3 To find the number of URIs exiting with status 200 (statuscode.py)

```

1
2 import requests
3 import json
4
5 file2 = open("links.json", "r")
6 file= open("codecount.txt", "w")
7
8
9 count=0
10
11 for line in file2:
12
13     short=line.strip()
14     one_line = json.loads(line)
15     link = one_line['link']
16
17     try:
18         info=requests.get(link)
19         if info.status_code==200:
20
21             count=count+1
22             print count
23     except Exception, e:
24         print e
25         continue
26 file.write(count)

```

```
27 file.close()
```

4.2.4 Generating graph

```
1 #--For Raw Files Size difference-----
2 d = read.table('rawdifftxt', col.name=c("sizes"))
3 plot(d$sizes, xlab="URIs", ylab="Size difference in raw files", xlim=c(0,1000), ylim=c
4      (-700000,700000), type="l")
5 #--For Processed Files Size difference-----
6 d = read.table('processdiftxt', col.name=c("sizes"))
7 plot(d$sizes, xlab="URIs", ylab="Size difference in processed files", xlim=c(0,1000), ylim
8      =c(-200000,200000), type="l")
```

4.3 Output

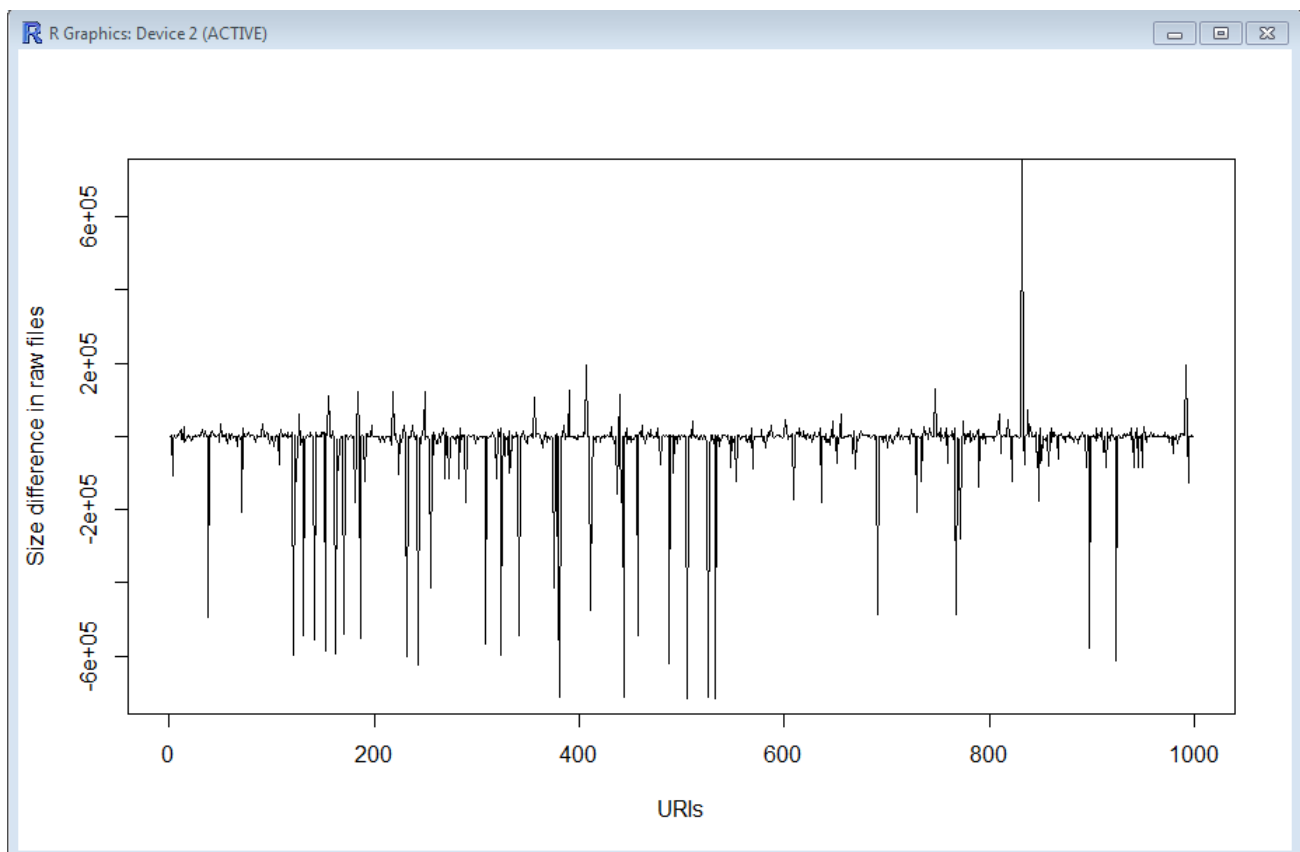


Figure 2: Raw Difference

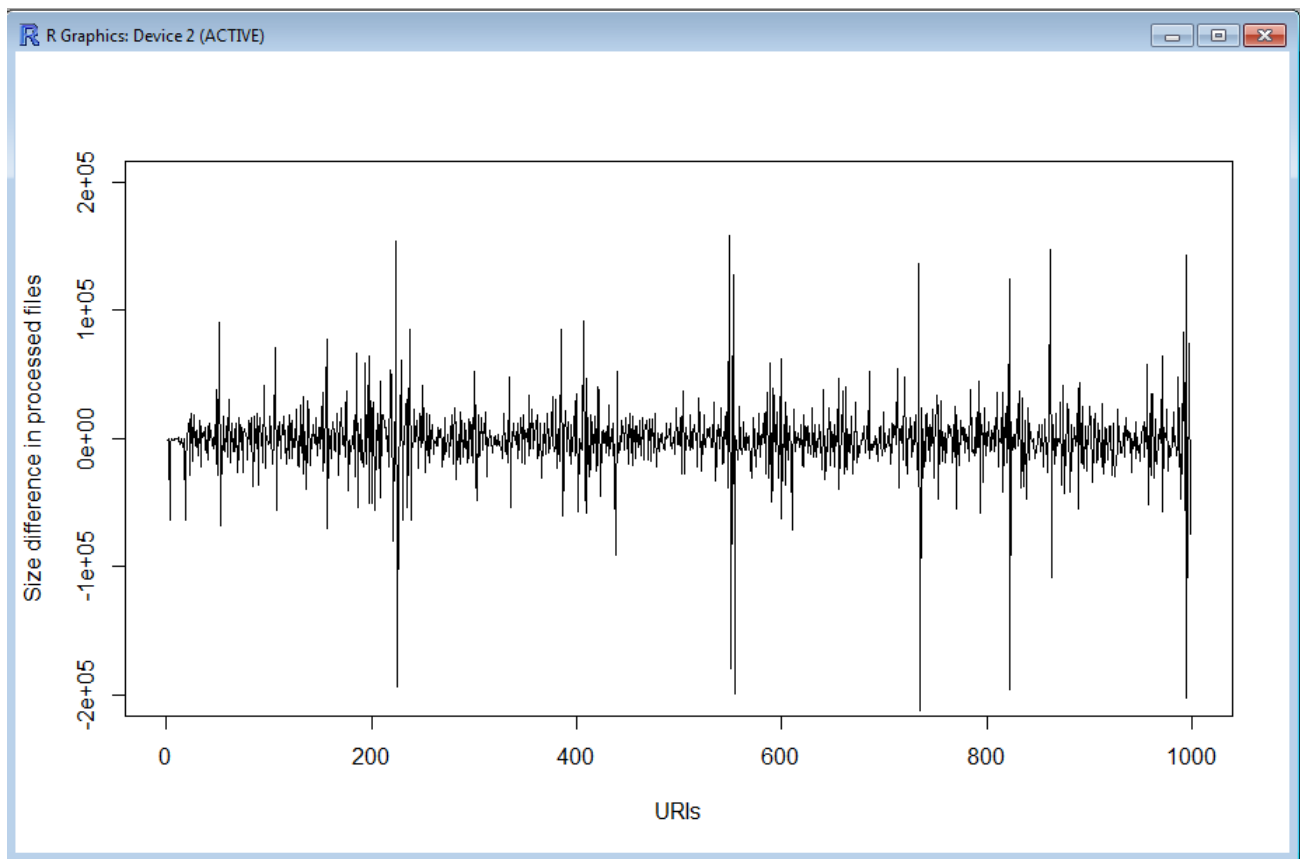


Figure 3: Processed Difference

```

sirius:~/wsprograms/a10/q4/topsrx> vim -d 735processedold.txt 735processednew.txt
2 files to edit
# [1]News Å» Feed [2]News Å» Comments Feed

mtv Menu
mtv
_____ M
* [3]mtv Home
* [4]Shows
+ [5]Full Episodes
+ [6]Shows A-Z
+ [7]TV Schedule
+ [8]App
+ [9]Shows News
+ [10]Music from the Shows
+ [11]Casting Calls
* [12]Music
+ [13]Music
+ [14]Artists A-Z
+ [15]Music Videos
+ [16]Artist to Watch
* [17]News
+ [18]Latest News
+ [19]Music
+ [20]Celebrity
+ [21]TV
+ [22]Movies
+ [23]Style
+ [24]Politics
+ [25]Life
+ [26]Issues
* TV Provider Signout
*
* [27]facebook [28]twitter [29]tumblr

[30]news
_____ M
* [31]Shows
+ [32]Full Episodes [33]Shows A-Z
+ [34]TV Schedule [35]App
+ [36]Shows News [37]Music from the Shows
+ [38]Casting Calls
* [39]Music
+ [40]Music
+ [41]Artists A-Z
+ [42]Music Videos
+ [43]Artist to Watch
* [44]News
+ [45]Latest News [46]Music [47]Celebrity
+ [48]TV [49]Movies [50]Style
+ [51]Politics [52]Life [53]Issues
* [54]facebook [55]twitter [56]tumblr

[57]News
GQ Magazine
[58]Style

```

```

# [1]News Å» Feed [2]News Å» Comments Feed

mtv Menu
mtv
_____ M
* [3]mtv Home

* TV Provider Signout
*
*
[4]news
_____ M
*
[5]News
+ [6]Music

GQ Magazine
[6]Style

```

Figure 4: Comparing Processed data of maximum size difference


```

sirius:~/wsprograms/a10/q4/topsix> vim -d 554processedold.txt 554processednew.txt
2 files to edit

```

```

[1]Log In / [2]Register
[3]Log Out / [4]My Profile
* [5]HitFix
* Popular

Groot home_top_story 1
Aww look how cute Groot is in the first 'Guardians of the Galaxy
2' image!
Aquaman-costume home_top_story 2
Our first good look at Jason Momoa's full Aquaman costume comes
from ToyFair
Groot2 home_top_story 3
Guess who just joined the cast of 'Guardians of the Galaxy 2'?
Love! home_top_story 4
Judd Apatow: 'Love' was made for the Netflix binge, and 'Girls'
should be weekly
X-files-anderson-duchovny-babylon home_top_story 5
Outrage Watch: Last night's 'X-Files' is being accused of the very
thing it was attempting to critique
Predator-predators-2010-movie-14721624-800-1200 2 home_top_story
6
Your first look at the new 'Predator' is finally here
Better-call-saul-switch home_top_story 7
'Better Call Saul' pits Jimmy against one of Walter White's
enemies. Sort of.
Gaynes home_top_story 8
Bulky Brewster says good-bye to Henry
The-people-v-o-j-simpson-dream-team home_top_story 9
Review: 'The People v. O.J. Simpson' assembles 'The Dream Team'
The-walking-dead-car-theory home_top_story 10
This minor detail in 'The Walking Dead' premiere could spell doom
for (REDACTED)
Nakedandafraid home_top_story 11
9 burning 'Naked and Afraid' questions answered
Bvz home_top_story 12
Despite a great trailer, Warner Bros is worried about Batman V
Superman and Justice League may be in jeopardy
* [6]Videos
[7]Trending [8]Trailers [9]Fandemonium [10]Girls on Film [11]2
Steps Forward 1 Step Back [12]The Snap [13]2 Minute History [14]See
or Skip [15]Interviews [16]Ask Alan [17]See or Skip [18]The
Dartboard
* [19]Movies
+-- 12 Kings: [20]Motion Captured [21]Reviews [22]In Theaters [23]On Demand
* login
[44]Log In / [45]Register
[46]Log Out / [47]My Profile
* Go

[48]What's Alan Watching [49]Motion Captured [50]The Dartboard
[51]Fangirl [52]See or Skip [53]Streaming Genie

Grab the Embed Code
Follow HitFix

```

```

[1]Log In / [2]Register
[3]Log Out / [4]My Profile
* [5]HitFix
* Popular

Better-call-saul-klick-mike home_top_story 1
'Better Call Saul' creators: Don't automatically expect to see Gus
next year
042916s home_top_story 2
Ask Alan: What's the best TV lineup ever?
Suicide-squad-enchanted home_top_story 3
The Enchantress costume from 'Suicide Squad' is sexist...and worse?
Inaccurate
Person-of-interest-michael-emerson-jim-caviezel home_top_story 4
Why I gave 'Person of Interest' another try... and was glad I did
Harleyrobbiesuicidesquad home_top_story 5
Margot Robbie is not a fan of Harley Quinn's hot pants in 'Suicide
Squad'
Walking-dead-alan home_top_story 6
Why Alan Sepinwall is done with The Walking Dead
Ferrellreagan home_top_story 7
Why Will Ferrell's disappearing 'Alzheimer's comedy' may signal the
end of an era in comedy
The-americans-travel-agents home_top_story 8
Review: 'The Americans' zeroes in on the hunt for Martha in 'Travel
Agents'
Punisher home_top_story 9
Netflix loved The Punisher so much on 'Daredevil' that he's getting
his own series
Lemonade home_top_story 10
Seth Meyers had the best late night take on Beyonce's 'Lemonade'
Wheel of time home_top_story 11
'The Wheel of Time': Book series coming to TV according to author's
widow
Newghostbustersgroup home_top_story 12
New details from the female 'Ghostbusters' reboot; don't worry men,
the original still exists
* [6]Videos
[7]Trending [8]Trailers [9]Fandemonium [10]Girls on Film [11]2
Steps Forward 1 Step Back [12]The Snap [13]2 Minute History [14]See
or Skip [15]Interviews [16]Ask Alan [17]See or Skip [18]The
Dartboard
* [19]Movies
+-- 12 Kings: [20]Motion Captured [21]Reviews [22]In Theaters [23]On Demand
* login
[44]Log In / [45]Register
[46]Log Out / [47]My Profile
* Go

[48]What's Alan Watching [49]Motion Captured [50]The Dartboard
[51]Fangirl [52]See or Skip [53]Streaming Genie [54]Vote In Heroes vs.
Villains
Grab the Embed Code
Follow HitFix

```

Figure 6: Comparing Processed data of second maximum size difference

References

- [1] Cosine distance. <http://stackoverflow.com/questions/18424228/cosine-similarity-between-2-number-lists>.
- [2] Status Checker. <http://cairographics.org/pycairo>.
- [3] numpredict.py. <https://github.com/manojchandrak/Programming-Collective-Intelligence/blob/master/chapter8/numpredict.py>.

[]