2017

Assignment 10 Report

ROSS REELACHART

PROFESSOR MICHAEL L. NELSON

Question 1:

I started with the blog term matrix file, *blogTermMatrix.txt*, from assignment 8. I needed to change the data in that file into a vector format that can then be used to get the K-nearest neighbor, which was accomplished in the python program *getVector.*py. The output of that program was named blogTermVector.

I then used the knnestimate code from the PCI book[1] in the program *getKnnEstimate.py* in order to find the nearest neighbors for the selected blogs. I needed to run the program a total of ten times, for each K-value for each blog. The results of each run can be found on the github[2].

The results are shown in the tables below.

For 'http://f-measure.blogspot.com/' @ K = 1

Blog	Distance
PALMIRA A PISTA TRES	0.0160565786974

For 'http://f-measure.blogspot.com/' @ K = 2

Blog	Distance
PALMIRA A PISTA TRES	0.0160565786974
Boggle Me Thursday	0.0171782044484

For 'http://f-measure.blogspot.com/' @ K = 5

Blog	Distance
PALMIRA A PISTA TRES	0.0160565786974
Boggle Me Thursday	0.0171782044484
Unexpectedly Bart (King!)	0.0230328178092
Fran Brighton	0.0231957980049
Parish Radio	0.0247776979083

For 'http://f-measure.blogspot.com/' @ K = 10

Blog	Distance
PALMIRA A PISTA TRES	0.0160565786974
Boggle Me Thursday	0.0171782044484
Unexpectedly Bart (King!)	0.0230328178092
Fran Brighton	0.0231957980049
Parish Radio	0.0247776979083
Abu Everyday	0.025624902688
IoTube :)	0.0313823670467
Room 19's Blog 2016	0.0322352581205
|[Tu queres ver isto]|	0.0335574483771
Azul Valentina	0.0345469000882

For 'http://f-measure.blogspot.com/' @ K = 20

Blog	Distance
PALMIRA A PISTA TRES	0.0160565786974
Boggle Me Thursday	0.0171782044484
Unexpectedly Bart (King!)	0.0230328178092
Fran Brighton	0.0231957980049
Parish Radio	0.0247776979083
Abu Everyday	0.025624902688
IoTube :)	0.0313823670467
Room 19's Blog 2016	0.0322352581205
|[Tu queres ver isto]|	0.0335574483771
Azul Valentina	0.0345469000882
Oh Yes Jónsi!!	0.0413451732917
What Am I Doing?	0.0414348051018
Hello	0.0439654050148
A H T A P O T	0.0456548114699
earenjoy	0.0466991805878
Me fala uma música boa aí	0.048317746845
She's mad but she's magic. There's no lie in her fire.	0.048991199787
the traveling neighborhood	0.0495984933977
One Stunning Single Egg	0.0499339527169
music of the moment	0.049957048262

For 'http://ws-dl.blogspot.com/' @ K = 1

Blog	Distance
Music Album Torrent Download Link Suggestions	0.0235168690854

For 'http://ws-dl.blogspot.com/' @ K = 2

Blog	Distance
Music Album Torrent Download Link Suggestions	0.0235168690854
A H T A P O T	0.0259338872125

For 'http://ws-dl.blogspot.com/' @ K = 5

Blog	Distance
Music Album Torrent Download Link Suggestions	0.0235168690854
A H T A P O T	0.0259338872125
Abu Everyday	0.0272925780471
earenjoy	0.029899114384
IoTube :)	0.0376426235079

For 'http://ws-dl.blogspot.com/' @ K = 10

Blog	Distance
Music Album Torrent Download Link Suggestions	0.0235168690854
A H T A P O T	0.0259338872125
Abu Everyday	0.0272925780471
earenjoy	0.029899114384
IoTube :)	0.0376426235079
Me fala uma música boa aí	0.0383741211117
music of the moment	0.0384612465964
Desolation Row Records	0.043023390171
|[Tu queres ver isto]|	0.0477445212815
One Stunning Single Egg	0.051227319962

For 'http://ws-dl.blogspot.com/' @ K = 20

Blog	Distance
Music Album Torrent Download Link Suggestions	0.0235168690854
A H T A P O T	0.0259338872125
Abu Everyday	0.0272925780471
earenjoy	0.029899114384
IoTube :)	0.0376426235079
Me fala uma música boa aí	0.0383741211117
music of the moment	0.0384612465964
Desolation Row Records	0.043023390171
|[Tu queres ver isto]|	0.0477445212815
One Stunning Single Egg	0.051227319962
The Themes of My Life	0.0560097664164
Oh Yes Jónsi!!	0.0601824348885
Who needs a TV?	0.0612235685422
Unexpectedly Bart (King!)	0.0626015182575
Lo importante es que estes tú bien	0.0633293781312
Stereo Pills	0.0634197105167
Out of my Mind	0.0636586425685
F-Measure	0.0655284964114
PALMIRA A PISTA TRES	0.0680257007904
tumbleweed	0.0685412492148

Question 2:

Unfortunately, not attempted

(Continued on next page.)

Question 3:

The original TimeMaps from assignment 2 were re-downloaded for comparison to the new TimeMaps, and our included in the github under the new name *finalMementoList2.txt*[3]. So the new TimeMaps for the 1000 links were acquired in the same manner as assignment 2, starting with *1000Links.txt* and using *getMementos.py* get the new numbers which are stored in the file *finalMementoList10.txt*.

Using these two files, I compared them using the python program *getDifference.py* which looked at the memento column of each file and compared them against each other. The results were stored in the file *differenceMementos* which contained a single column of positive or negative number representing increased or decreases in mementos, respectively.

This last file was then used in R to create the following graph:

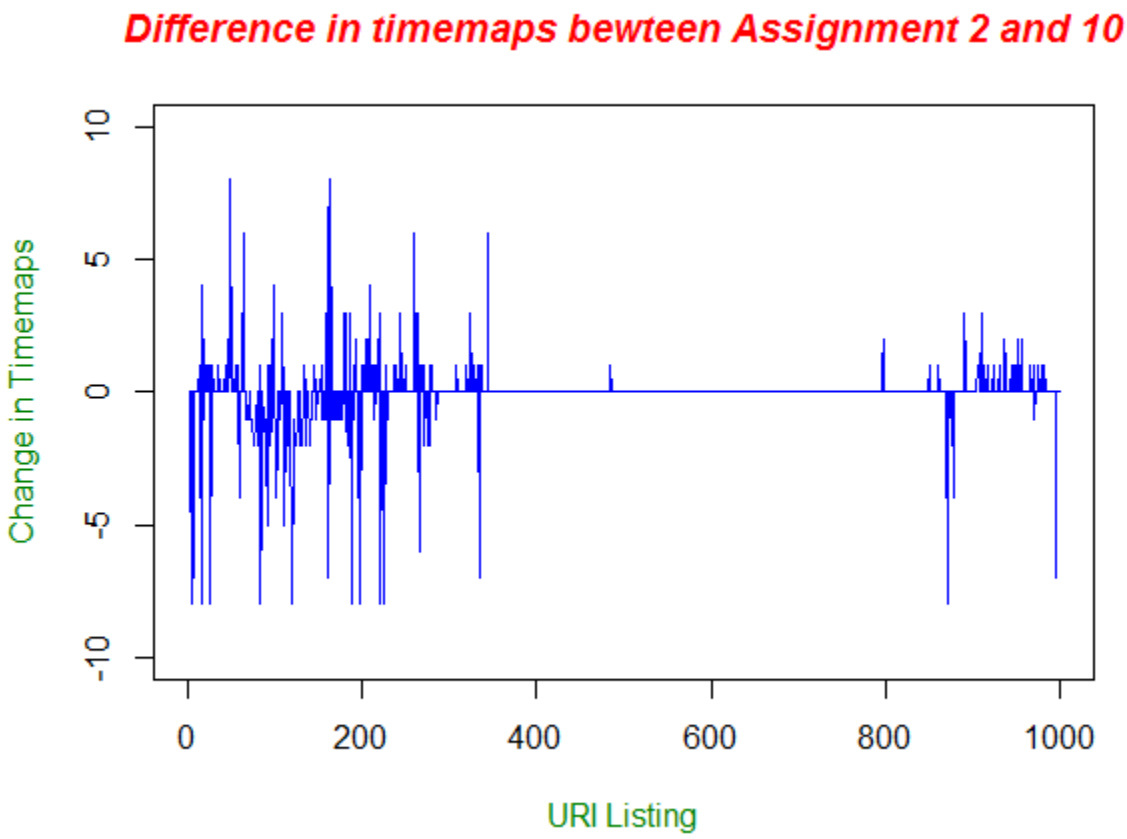

Figure 1 - The changes in the number of mementos for the 1000 links gathered in assignment 2.

The greatest changes seemed to occur in the URIs that were blogs or news sites, with larger sites gaining greater positive numbers and more niche sites like Red Letter Media losing numbers. The larger area in the middle with no changes were mostly Twitter status posts. The bits at the end were YouTube videos gaining a few numbers or a few pages that didn't end in a 200 code.

Question 4:

For this question, I was only able to really do the first part of the question. That is, I was able to find which of 1000 links still worked. Using the python program *getStatusCode.py*, I was able to iterate through my original list of 1000 URIs and test their connection. The final tally was stored in the file *statusCodeCount*, and can be seen in the table below.

Status Code	# of URIs
200	963
403	6
404	4
429	4

References

1.) https://github.com/uolter/PCI/tree/master/chapter8
2.) https://github.com/rreelachart/cs532-s17/tree/master/submissions/assignment%201
3.) https://github.com/rreelachart/cs532-s17/tree/master/submissions/assignment%2010/Question%203