CS 535: Assignment #9

Due on 11:59pm April 21, 2018

Alexander C. Nwala 16:20

David Sinclair

April 20, 2018

Contents

1 Problem 1

Support your answer: include all relevant discussion, assumptions, examples, etc.

1.1 Question 1

(10 points)
1. Using the data from A7:

- Consider each row in the blog-term matrix as a 1000 dimension vector, corresponding to a blog.

- Use knnestimate() to compute the nearest neighbors for both:
http://fmeasure.blogspot.com/
http://wsdl.blogspot.com/

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

Use cosine distance metric (chapter 8) not euclidean distance. So you have to implement numpredict.cosine() instead of using numpredict.euclidean() in:
https://github.com/arthure/ProgrammingCollectiveIntelligence/blob/master/chapter8/numpredict.py

1.2 Answer 1

I attempted to load the data from assignment 7 blogterm matrix into the knnestimate from numpredict.py.

The numpredict program did not like the blog-term matrix data in the raw format. I then attempted to modify the data. This did not work out either. I then attempted to determine if I could find what line of data the FMeasure produced. That I was able to find and a saved the output of that row of data into a file called FMeasure. I then attempted to have the output of the row be vec1. This did not produce the results I was looking for and the program gave the error that vec was not suppose to be a string. I then used the k-mod programs in assignment 7 and found that clusters.py was actually doing vectors. I attempt to determine how it was making the vectors. Which in kcluster and hcluster it was creating the vectors. I attempted to use this section and determined that row 46 in my blog matrix was actually the FMeasure row and data. I attempted using the numpredict program again with 46 being vec1. This again failed.

2 Problem 2

==========
========The questions below is for 3 points extra credit==========
==========

2.1 Question 2

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.

2.2 Answer 2

I got the data from assignment A2 Q2. I then redownloaded the Timemaps or memtos using programs from assignment 2. The prorams were 07_memto.py, That got the data memto information. The 09_json.py program got took the memto data and extracted the number of total memtos and inputed into the histograph1.data.csv file. I then took the histograph.data.csv file from the assignment two and compared them in a histographsum.data.csv. I then took that data and added to the 10_memhist.py program. That created the below graphs. The first one was done with all data. Because one memto had over 7000 changes it made the other chagnes to small to see. So I removed that one and redid the graph called Figure_2.png