StudentDetails:
Name : Ashwini Singh
UIN: 522006254

Platform Used : Linux(Ubuntu) and python 2.7

Solution 1

Processing Statistics::
('Size of full graph =', 65121)
('Size of weekly Conn graph =', 1526)
('Number of Iterations =', 136)

Top 20 Hubs:
****
1 . BaughnIsabella [8.26452267558e-14]
2 . 1D_EternalLove [8.26452267558e-14]
3 . IsaHoranT1D [8.26452267558e-14]
4 . Vanessa_Vieiira [8.26452267558e-14]
5 . Hey_ThereRomeo [7.91500875922e-14]
6 . NarryBites_ [7.83605283985e-14]
7 . WenYein [7.83605283985e-14]
8 . norosaprotested [7.83605283985e-14]
9 . Gnarlynialll [7.83605283985e-14]
10 . stehanielou [7.83605283985e-14]
11 . Narrys_1Derfull [4.34286886916e-14]
12 . nvrrystvgrvm [4.34286886916e-14]
13 . xnarrygirl [4.34286886916e-14]
14 . TeenageBitchbag [3.80715957487e-14]
15 . Upallnialls [3.7667552291e-14]
16 . criesluke_ [3.7667552291e-14]
17 . leela_brooks [3.7667552291e-14]
18 . weyheydacraic_ [3.7667552291e-14]
19 . RamosPCarla [3.52426050091e-14]
20 . 5secondsftbooks [2.65099032354e-14]

Top 20 Authorities:
****
1 . Harry_Styles [2.43943305519e-14]
2 . NiallOfficial [1.90343581397e-14]
3 . Louis_Tomlinson [1.32732217391e-14]
4 . Real_Liam_Payne [1.08482744572e-14]
5 . zaynmalik [1.08103435106e-14]
6 . onedirection [4.28469835733e-15]
7 . lewisxjones [2.11557268349e-15]
8 . Luke5SOS [1.53048799813e-15]
9 . Michael5SOS [1.25389766833e-15]
10 . Ashton5SOS [1.13609413721e-15]
11 . Calum5SOS [9.09592216855e-16]
12 . HeyHarryHoran [7.89559193739e-16]
13 . justinbieber [6.60124955342e-16]
14 . colourlouis [6.01642880955e-16]
15 . halfmoonlouis [6.01642880955e-16]
16 . loulust [6.01642880955e-16]
17 . BadGirlRiRi [5.87213963607e-16]
18 . MileyCyrus [4.39947091447e-16]
19 . 5SOS [4.31639108617e-16]
20 . edsheeran [4.04043457726e-16]

Solution 2

I have used different values for different folder of data set and used the C value that provides maximum match score.

1. **Folder1 (C=0.04)**

   Loading test file for Training!!
   Classification of training data !!
   Processing Test Data !!
   Match % = 92.6952141058
   **********Top 10 Feature**********
   1 . Feature23 [1.40053300736]
   2 . Feature39 [1.03957034034]

3 . Feature21 [0.665021636925]
4 . Feature37 [0.471024840596]
5 . Feature18 [-0.411048120532]
6 . Feature41 [0.405519901419]
7 . Feature19 [0.401820361062]
8 . Feature28 [0.349524954181]
9 . Feature25 [0.325456437049]
10 . Feature32 [0.308620284341]

2. <u>Folder 2(C=0.1)</u>

Match % = 82.7848101266
**********Top 10 Feature**********
1 . Feature23 [1.71032726595]
2 . Feature21 [1.3131900528]
3 . Feature39 [1.0597808219]
4 . Feature37 [0.916780266274]
5 . Feature32 [0.603912544776]
6 . Feature41 [0.592200883355]
7 . Feature16 [0.570584711796]
8 . Feature20 [0.556565883833]
9 Feature14 [-0.478823611848]
10. Feature40 [-0.467486730566]

3. <u>Folder 3(C=0.104)</u>

Match % = 87.8281622912
**********Top 10 Feature**********
1 . Feature23 [0.52261795212]
2 . Feature39 [0.479669094833]
3 . Feature21 [0.337948376511]
4 . Feature37 [0.324129054033]
5 . Feature25 [0.218348971513]
6 . Feature18 [-0.21790268539]
7 . Feature42 [0.202829891496]
8 . Feature46 [-0.198517326963]
9 . Feature41 [0.194728701393]
10 . Feature38 [0.173202184013]

Directory Structure:

hw2_522006254

|-------- src

|-----part1.py (**The entry point to the application)

|-----part2.py (contains some util function)

|-----tweet.py (contains some util function)

|-------- README.pdf (containing the execution instruction and Result on sample data)

Design Architecture:

For part 1, the design uses the networkx to create graph and find the weekly connected components and use this data to get the HUBS and AUTHORITY scores. The error value used is $10^{13.}$ At each iterations, I am updating HUBS/ AUTHORITY score at each iterations as HUB[i]=HUB[i]*3/squrt(len(HUB)) . This is to make values in the limit of the float.

For part 2, data for SVC is created from the train file. I have tested with different values of 'C' on the train data and used the one provides maximum match value. For the data set provided, the values of 'C' are 0.04, 0.1, 0.104 for fold1, fold2, fold3 respectively. The smaller value of 'C' gives a more fine-grained classification and thus provides better results on clearly separated dataset.

**Currently the code has C=0.1. Result may vary from the presented above.

How To Execute:

1. **Part 1**

step1: Go to src folder.

step2: Execute cmd "python part1.py --datapath=<location of data folder>

Sample Output:

Processing Statistics::

****

('Size of full graph =', 65121)

('Size of weekly Conn graph =', 1526)

('Number of Iterations =', 136)

Top 20 Hubs:

****

<List of Hubs>

Top 20 Authorities:

****

<List of Authorities>

2. **Part 2**

step1: Go to src folder.

step2: Execute cmd "python part2.py --datapath=<location of data folder>

**The date folder should contain train and test data file. We will have to run this for each three folders in our sample data.

Sample Output:

Loading test file for Training!!

Classification of training data !!

Processing Test Data !!

Match % = <match_%>

**********Top 10 Feature**********

<List of Features>