CS 6375

ASSIGNMENT 02 – Part II

Names of students in your group:

Dishant Khanna (2021438172)

Ekansh Goyal (2021474594)

Number of free late days used:  1

Number of free late days left:  3

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

**Project Report**

**Title:** Implementation of K-means unsupervised learning algorithm.

**Assumptions made:**

The user will provide valid name of the dataset file present in the same directory as the code.

**Accomplishments:**

·        Successfully implemented K-means unsupervised learning algorithm.

·        Successfully implemented Jaccard distance to find the distance between 2 tweets.

·        Successfully calculated the Sum of squares error for various values of k.

·        Successfully pre-processed the data, which includes

* Remove the tweet id and timestamp
* Remove any word that starts with the symbol @ e.g. @AnnaMedaris
* Remove any hashtag symbols e.g. convert #depression to depression
* Remove any URL
* Convert every word to lowercase.

**RESULTS**

* For Datasets 01(usnewshealth): Refer to the table below.

|  |  |  |
| --- | --- | --- |
| Value of K | Sum of Squared Errors | Size of Each Cluster |
| 5 | 1117.888001485695 | 1 : 129 tweets  2 : 122 tweets  3 : 324 tweets  4 : 432 tweets  5 : 393 tweets |
| 10 | 1073.6189225560306 | 1 : 72 tweets  2 : 375 tweets  3 : 121 tweets  4 : 112 tweets  5 : 247 tweets  6 : 39 tweets  7 : 50 tweets  8 : 195 tweets  9 : 166 tweets  10 : 23 tweets |
| 15 | 1053.733343153102 | 1 : 125 tweets  2 : 76 tweets  3 : 30 tweets  4 : 30 tweets  5 : 15 tweets  6 : 35 tweets  7 : 28 tweets  8 : 210 tweets  9 : 157 tweets  10 : 20 tweets  11 : 82 tweets  12 : 192 tweets  13 : 177 tweets  14 : 84 tweets  15 : 139 tweets |
| 20 | 1033.937260111873 | 1 : 14 tweets  2 : 163 tweets  3 : 80 tweets  4 : 149 tweets  5 : 28 tweets  6 : 121 tweets  7 : 155 tweets  8 : 13 tweets  9 : 81 tweets  10 : 64 tweets  11 : 114 tweets  12 : 36 tweets  13 : 10 tweets  14 : 24 tweets  15 : 59 tweets  16 : 186 tweets  17 : 18 tweets  18 : 21 tweets  19 : 38 tweets  20 : 26 tweets |
| 25 | 1025.9198276339928 | 1 : 86 tweets  2 : 24 tweets  3 : 19 tweets  4 : 59 tweets  5 : 12 tweets  6 : 133 tweets  7 : 32 tweets  8 : 35 tweets  9 : 7 tweets  10 : 53 tweets  11 : 43 tweets  12 : 135 tweets  13 : 183 tweets  14 : 57 tweets  15 : 38 tweets  16 : 132 tweets  17 : 17 tweets  18 : 24 tweets  19 : 40 tweets  20 : 44 tweets  21 : 133 tweets  22 : 68 tweets  23 : 8 tweets  24 : 10 tweets  25 : 8 tweets |

**Learning:**

* Learnt programming in python
* Learnt the implementation of the K-Means unsupervised learning algorithm.
* Analysed different data pre-processing methods and implemented several of them.

**Note:** Screenshots and complete output of a run of the program are present in the folder ‘output’ for reference.

‘Output’ for any run will be stored in the file generated automatically by the name ‘K\_MEANS\_OUTPUT.txt’