|  |
| --- |
|  |
| 1. collect.py |
|  |  |
|  | Input: twitter data of two screen\_names |
|  | Output: clusterinput.py,classifyinput.py |
|  |  |
|  | This Assignment Collects twitter live data of Google and Tesla |
|  | This assignment aims to show how these social network groups are connected graph is not so dense or sparse who are common followers all these groups have, |
|  | we break the link between groups and how different communities (3 in this case) are formed |
|  |  |
|  | In collect.py |
|  | For clustering |
|  | A subset of 50 friends and followers of 2 communities, |
|  | For each common follower take their 50 friend list |
|  | For classification: |
|  | Separately taking around 50 mentions tweets and description of users for classification of gender where people have mentioned any of these 2 groups |
|  |  |
|  |  |
|  |
|  |
|  |
|  |
|  |  |
|  | 2) cluser.py |
|  | Code: |
|  | Input: clusterinput.py from collect.py |
|  | Output: clusteroutput.py |
|  |  |
|  | I am taking data of users/friends and followers that was stored in cluster.py and using partition girven newman to cluster the data into 3 communities. |
|  | Drawing a network graph of the same. |
|  | Also considering friends of common followers |
|  | Labelling only the three brand groups and its common followers |
|  | Storing the clusters that were obtained from partition to output file clusteroutput.py |
|  |  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |  |
|  |  |
|  |  |
|  | This data and community cluster tells us that how two nodes in graph are related which edge has high weightage betweeness and |
|  | How two different groups are connected how can we break link between them and common followers how they tend to link groups when data is partitioned |
|  |  |
|  |  |
|  | 3) Classify.py |
|  | Input: classifyinput.py |
|  | Output: classifyoutput.py |
|  |  |
|  | Classification based on gender. With male and female surnames taken from census. |
|  | This data gives us information about how many males and females are interested in technology (Google and Tesla) and mention them in their tweets and talk about the teams in twitter |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  | We are using logistic regression classifier model to fit and predict our data |
|  | We need to vectorize and create csr matrix based on features we decide to predict gender |
|  |  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |  |
|  |  |
|  |
|  |
|  |
|  |
|  |
|  |  |
|  | 4)Summarize.py |
|  |  |
|  | Input : classifyinput,pkl,classifyoutput.pkl,clusterinput.pkl,clusteroutput.pkl |
|  | Output: summary.txt |
|  | Summarize all data collected based on classification and clustering |
|  | We are getting count for no of communities detected no of users and tweets |
|  | We analyse how many people are classified and what their gender is |
|  | How many average users exist in one community |
|  |  |
|  | Assignment Analysis: |
|  | The whole assignment aims to perform community detection link prediction and classification using machine learning |
|  |
|  |
|  |  |
|  |  |
|  |
|  |
|  |
|  |  |