Aravinda Kumar Reddy Yempada
1208601637

Project1 – P1

The folder contains two python source files, P1-a.py and P2-b.py and the results in .csv format.

P1-a.py is the code for scraping the 'Academia.edu' website. Academia.edu overhauled their website recently and have two kinds of designs for the follower's page that I am scraping. The code contains both techniques to scrape the website irrespective of the design. The script takes the URL of the initial node as a command line argument. To run the script use the following line in a command prompt or terminal:

python P1-a.py http://asu.academia.edu/AravindKumarReddyYempada

The script produces two outputs, output.csv and count.csv which are renamed as '1 - Raw Dataset.csv' and '2 - UserCount Mappings.csv' respectively in the deliverables. '1 - Raw Dataset.csv' file contains the Non-anonymized dataset and the '2 - UserCount Mappings.csv' file contains the user and number of followers mapping.

P1-b.py is the python code to anonymize the data and perform uniform sampling. It takes two parameters as command line arguments. The first parameter is the output.csv file generated by P1-a.py and the second parameter is a Boolean datatype to let the program know if we need to sample the data or not. Use the following commands to execute the script in a command prompt or terminal:

- *python P1-b.py output.csv 0*

  This gives the edge list, node list and anonymized edge list of the entire raw data. The output files are renamed to '1a - Edge List.csv', '1b - Nodes List.csv' and '3 - Anonymized Edge List.csv' in the deliverables.

- *python P1-b.py output.csv True*

  This gives the edge list, node list and anonymized edge list of a subset of raw data and samples the nodes that contain more than 1000 nodes by selecting 1000 nodes uniformly. The output files are renamed to '4a - Sampled Edge List.csv' , '4b - Sampled Nodes List.csv' and '5 - Sampled Anonymized List.csv' in the deliverables. Only a subset of raw data containing 2000 nodes are returned which will be used as the input for the remaining problems of this project.

Aravinda Kumar Reddy Yempada
1208601637

Summary:

- P1-a : code used to crawl Academia.edu
- P1-b : Code used to anonymize the data
- '1 - Raw Dataset.csv', '1a - Edge List.csv' and '1b - Nodes List.csv' : non-anonymized dataset
- '2 - UserCount Mappings.csv ' : usernames -> count mapping
- '3 - Anonymized Edge List' : anonymized dataset
- '4a - Sampled Edge List.csv' and '4b - Sampled Nodes List.csv' : non-anonymized sampled dataset
- '5 - Sampled Anonymized List.csv' : anonymized sampled dataset