**CSC 407/607 Network Analysis**

**Fall 2022**

Graduate Project

Sai Ram Ajay Krishna Gabbula

**Data Set Preprocessing:**

I have downloaded the data from the provided link and extracted different datasets present in them. Out of these datasets we need only two datasets namely business dataset (business.csv) and reviews dataset (reviews.csv).

Merging the two datasets extracted to get a super dataset which contains all the data required for the analysis. The dataset contains **6.99 Million rows and 16 columns** in it

Graphical user interface, application

Description automatically generated

From the above dataset selecting **California** state for further analysis.

Graphical user interface, text, application, email

Description automatically generated

From this dataset extracting all the cities present for further analysis.

Text

Description automatically generated

From the above cities list selecting ‘**Montecito**’ city as the first metropolitan area.

Graphical user interface, text, application, Word, email

Description automatically generated

After all the extraction we got a final dataset with 5042 rows with 16 columns.

1. **Identify the total number of reviews for each business in one metropolitan area.**

This step is done by using the ‘groupby’ function in pandas to extract the total number of reviews for each business in Montecito.

Graphical user interface

Description automatically generated

From this image it is clear that the data contains only 93 businesses for the city Montecito.

1. **Keep only those businesses with at least R reviews and remove all others.**

The number of reviews selected is 105. So from the above data frame extracting the data with reviews greater than or equal to 105.

Graphical user interface, text, application

Description automatically generated

The data extracted has only 3197 rows and 16 columns.

1. **Among the remaining businesses, when two businesses have been reviewed by the same reviewer, connect them.**

For this step I am excluding all the except **‘business\_id’** and **‘user\_id’** by dropping all the other columns.

Graphical user interface, text, application, email

Description automatically generated

So in this dataframe looking for the businesses which was reviewed by the same user. Now I am trying to iterate over all the user ids from this dataframe. From this I am trying to extract the data which have more than 1 review with an if condition because we are looking for businesses which have more than 1 review. For the data satisfying the conditions I have extracted the business ids to plot the graph and saving it as a node list.

Text

Description automatically generated

So for each business satisfying the above conditions I am trying to check for same reviews by iterating the data through a nested for loop where the first for loop will be one business and the inner loop iterates through all the data for the selected businesses there by generating an edge list. I am saving the extracted edge list into a list. This list has the data of the edge connections between the nodes.

After all the analysis we have a list of nodes which have been reviewed by more than 1 user and the list of edge connections between these nodes. So, from these two lists we can plot a graph to show the connections visually.

The above image shows the code written to fulfill the logic explained here.

Graphical user interface, text, application

Description automatically generated

The above figure shows the unique number of businesses i.e., 12 for the city ‘Montecito’ satisfying the condition.

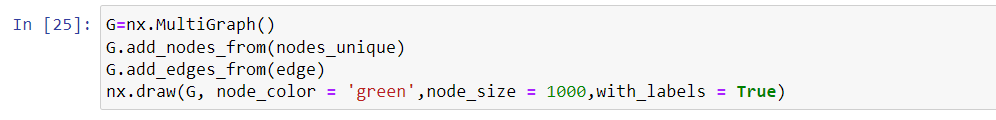
Text

Description automatically generated

The above figure shows the number of reviews between the businesses. There are 456 connections between the 12 businesses.

**d. Draw this graph and compute the degree centrality of all vertices. Compute the average degree centrality of all nodes.**

From the extracted data drawing a graph with the help of **networkx** library in python. The following image shows the code written to draw the graph.



Diagram

Description automatically generated

Above image is the graph generated from the conditions in ‘**C’.** The graph shown contains multiple edges between the nodes which is not been rendered by the IDE used while coding.

Graphical user interface, application, Word

Description automatically generated

The degree centrality for all the nodes present is calculated, which is shown in the above image. **The average degree centrality for the graph is 6.9090909**.

Text

Description automatically generated

The adjacency matrix for the above graph is shown above. This explains about the connectivity between the businesses.

**OUTCOMES:**

* The above graph shows us the total businesses as nodes and there related user ids as edges.
* State selected is **California** and the metropolitian city selected is **Montecito**. The number of reviews selected for the businesses is **105**.
* The total number of nodes/business extracted are **12**. Total number of connections/edges between these businesses are **456**.
* The average degree centrality for all the nodes in the graph is **6.90909**.

1. **Repeat for a second metropolitan area and compute the new average degree centrality of all nodes in the new metropolitan area.**

Selecting **‘Santa Barbara’** as the other metropolitan area in California for this step.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

In this city the number of businesses that have been reviewed by more than 1 reviewer is **550** and the number of connections between them is **177549**. The below graph shows these connections.

A picture containing plant

Description automatically generated

I have calculated the degree centrality and the average degree centrality of this graph.

Text

Description automatically generated

**OUTCOMES:**

* State selected is **California** and the metropolitan city selected is **Santa Barbara**. The number of reviews selected for the businesses is **105**.
* The total number of nodes/business extracted are **550**. Total number of connections/edges between these businesses are **177549**.
* The average degree centrality for all the nodes in the graph is **1.17601**.

**f. Do you see any difference in these two values? Explain.**

* For the same number of reviews and above (i.e.,105) for two different metropolitan area in California the total number of businesses which was reviewed by the users differs a lot.
* For **Montecito** the number of businesses extracted were **12** but whereas the number of businesses for **Santa Barbara** were **550**.
* The number of edges between these nodes were **456** for Montecito and **177549** for Santa Barbara.
* The average degree centrality for all the nodes for Montecito is **6.90909** but whereas average degree centrality for all the nodes for Santa Barbara is **1.17601**.
* The above statistics tells us that the businesses in Santa Barbara were reviewed more than that of Montecito.
* From average degree centrality it can be said that more users reviewed for similar businesses in Montecito than that of the businesses in Santa Barbara. It can also be said that the businesses in Montecito have more connections than Santa Barbara.
* The adjacency matrix for Santa Barbara is more sparse than that of the Montecito.

**The graph for GOLETA city in CALIFORNIA with 105 reviews:**

A picture containing diagram

Description automatically generated

**Average degree centrality is: 5.425716625716629**

**The graph for WEST BERLIN city in NEW JERSEY with 20 reviews:**

Diagram

Description automatically generated with low confidence

Average degree centrality is: 0.7662337662337663

**The graph for TURNERSVILLE city in NEW JERSEY with 20 reviews:**

Diagram

Description automatically generated

Average degree centrality is: 0.5656213704994192

**The graph for TURNERSVILLE city in NEW JERSEY with 105 reviews:**

Chart, radar chart

Description automatically generated

Average degree centrality is: 5.0

All code is written in **python** , used **jupyter notebook** for the same. Source code is present in file **src.ipynb .**

Note: I have collaborated with ‘Venkata Vijay Krishna Gabbula’ for the coding part of section C of the project. So there might be some similarities in the logic used to extract the graph.