19ZO02-Social and Economic Network Analysis

Project Report

Team Members:

19Z308 BHARATH S

19Z323 SANDEEP KUMAR K

19Z330 MAHESH BOOPATHY M

19Z335 NIRANJAN S

19Z344 SANJAI K

19Z351 SURTIK S

BACHELOR OF ENGINEERING

Branch: COMPUTER SCIENCE AND ENGINEERING

Of Anna University

Problem Statement:

If a person enjoys Chinese cuisine, There's a good chance that the person would also like Korean cuisine, although it might be easy to find this out for a specific individual, finding preferences and recommendations for millions of customers that want to explore and order new food is quite complex. This is where data science comes into place. Typical, commonly used algorithms consider only the customer's past behavior to offer recommendations. This heavily limits the number of potential restaurants that can be recommended leading to sparsity and redundancy challenges, Thus we make use of knowledge gained from graphs and appropriate tools to recommend a wider variety of restaurants by comparing similarity between customers to show new restaurants that the customer may not have tried.

Dataset Description:

The dataset we have used here contains a multitude of hotels and a set of reviews given to each hotel by the customers. The dataset is in CSV format. Additional information like hotel ID for each hotel, user ID for each user, user rating are also given. The ratings given by each user for a hotel can range anywhere between 0 to 1. With this dataset, It is possible for us to create a smaller dataset with only two parties, Users and their preferred restaurants. This acts like a bipartite graph using which we can infer different information to conclude recommendations by algorithms like link prediction.

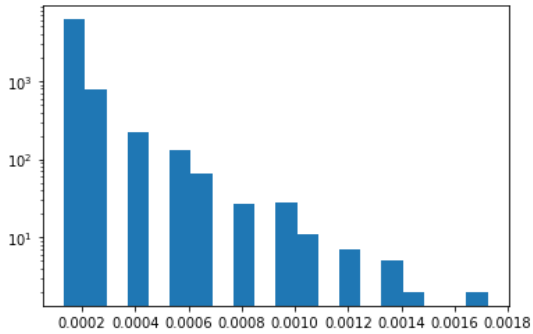


Figure.1. Degree distribution of hotels

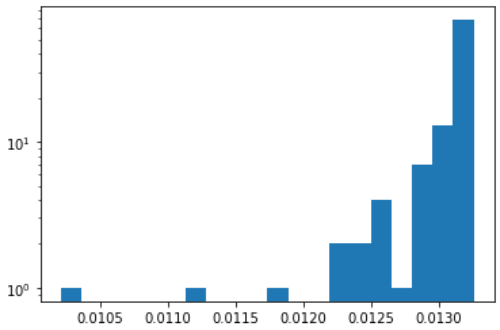


Figure.2. Degree distribution of users

Tools Used:

1.Gephi:

Gephi is the best software for visualisation and analysis of all types of graphs and networks. It is a Java-based open-source software and is built on the NetBeans platform. It helps data analysts to discover patterns or trends and identification of outliers.

2.Google Colab:

Colab is particularly well suited to machine learning, data analysis, and education. It enables anyone to create and execute arbitrary Python code through the browser. With Colab notebooks, users can include executable code, rich text, graphics, HTML, and more in one document.

3.Networkx:
NetworkX is a Python module for building, modifying, and researching the composition, dynamics, and purposes of complicated networks. NetworkX is a framework for network and social network analysis that is comparatively effective, relatively scalable, and portable.

4.Matplotlib:
Matplotlib is a Python library that allows you to create static, animated, and interactive visualisations.

5.Numpy:
It provides efficient multi-dimensional array objects as well as various operations to work with these array objects.

6.Pandas:
A Statistical analysis and data manipulation software written for python.

7.Scikit-learn (Sklearn):
It is Python's most useful and robust machine learning library. It offers a set of efficient tools for machine learning and statistical modelling, such as classification, regression, clustering, and dimensionality reduction, via a Python interface.

Challenges Faced:

1.Constructing a bipartite graph from the dataset was quite challenging as the dataset was huge
2.It took some time for us to incorporate our idea into the model.
3.Installation and understanding of the network packages were quite difficult.
4.With regard to the code, dataset, and collection method, we encountered some challenges during implementation.

Contribution of Team Members:

Roll no	Name	Contribution
19Z308	Bharath S	Hotel recommender ,User Similarity,Link Prediction Algorithms
19z323	Sandeep Kumar K	Hotel recommender ,User Similarity,Link Prediction Algorithms
19Z330	Mahesh Boopathy M	Hotel recommender ,User Similarity,Link Prediction Algorithms
19Z335	Niranjan S	Dataset Encoding
19Z344	Sanjai K	Centrality measure and Network Visualisation
19Z351	Surtik S	Hotel recommender ,User Similarity,Link Prediction Algorithms

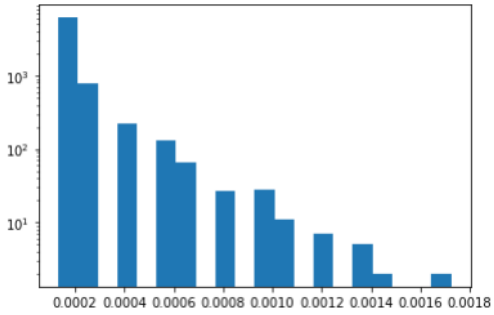
Annexure I: Code

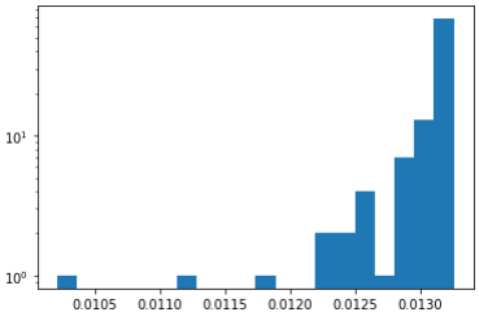
Code Link:

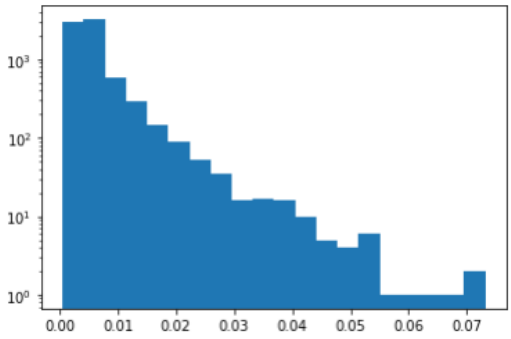
https://github.com/surtik48/Hotel-recommendation-based-on-user-similarity-using-Zomato-dataset

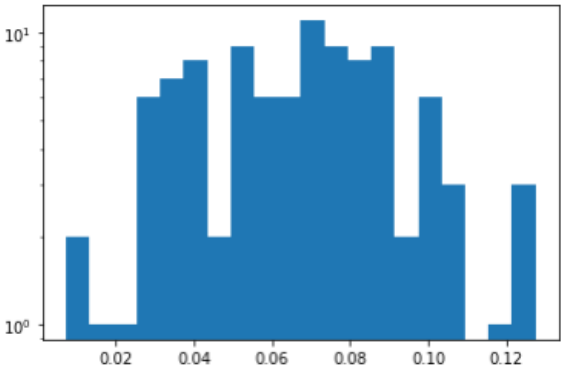
Annexure II: Screenshots

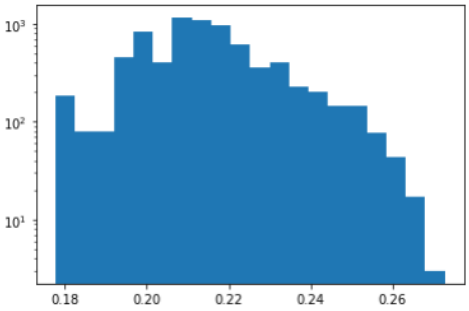
Plot the degree distribution of users

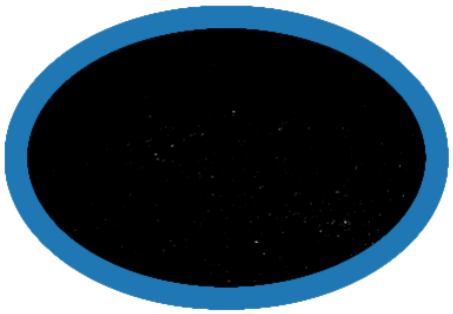






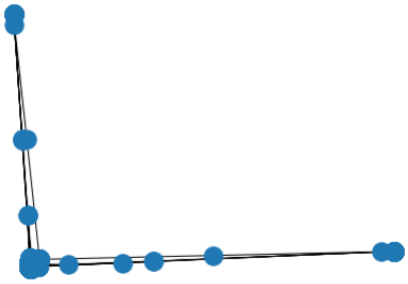




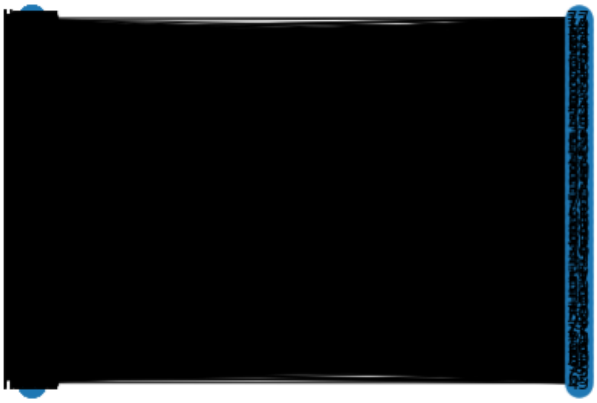


Circular Layout

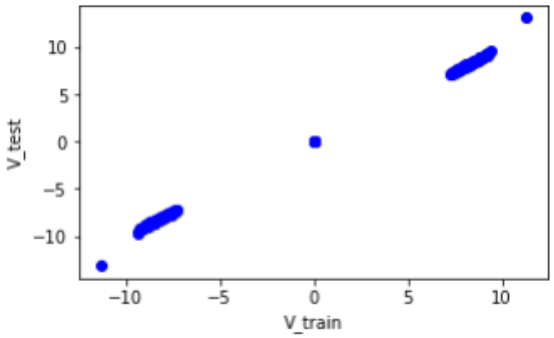
Random Layout



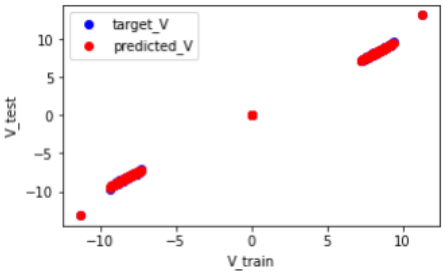
Spectral Layout



Bipartite Layout



OddNeumannPseudokernel



OddPathCountingKernal

References:

1.Dataset: https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants

2.https://www.extendoffice.com/documents/excel/670-excel-add-text-to-beginning-end-of-cell.hml

3. https://gephi.org/users/

4. https://networkx.org/

5. https://www.nltk.org/

6. https://scikit-learn.org/stable/

7.https://www.geeksforgeeks.org/networkx-python-software-package-study-complex-networks/

8.https://youtu.be/NyNqzDKcKG4

9.https://www.whitman.edu/mathematics/cgt_online/book/section05.04.html

10.https://www.zomato.com/blog/connecting-the-dots-strengthening-recommendations-for-our-customers-part-one