Analyzing Rutgers Research Topics and Finding Potential Correlation Based on Google Scholar Profile Records

Group 50: Junyan Dai & Feiyu Zheng

0. Note

After analyzing the situation of our original proposal and discussion about our project, we decide to change our project topic from analyzing coronavirus data to this topic. Please see the following for our current works and future plans.

1. Data Collection

All data used in this project are collected from Google Scholar. We designed a web scraping program using Python framework Scrapy to automatically send requests to scholar.google.com and then by analyzing the responses, we extracted totally 2908 profiles labelled with organization Rutgers University. From those profiles, we extracted 4597 valid interests or focused research fields. To get profile data, we first collected all profile urls from google scholar profile list filtered by organization--Rutgers University. Then, our program requested HTML page of each url and located and extracted useful data by their XPATHs.

2. Data Format Description

We collected data from web pages and stored them in local MySQL database. Currently, we have designed six tables: Authors, Interests, Organizations, Publications, Authors_to_Interests, and Authors_to_Publications. The Authors table contains the primary key (id), name, title(i.e Professor of Computer Science, Phd Candidate) of each scholar, and a foreign key (Oid) referring to organizations, which for now only include Rutgers University. The Interests table records all scholars' interests or focused research topics displayed on their profile pages with the primary key of each unique topic. The Organizations table as we mentioned above stores only Rutgers University so far. We created this table in case we may extend our project from Rutgers scholars only to various organizations. The Publications table is designed to store the information of all published papers authored by Rutgers scholars, but for now we have not finished collecting this part of data because of Google's robust anti-scraping system. The Authors_to_Interests table is a many-to-many relation between Author id(Aid) and Interests id (Iid), as we have found out that a scholar may focus on multiple research fields and many scholars may interested in the same topic. The Authors_to_Publications table is designed to store many-to-many relations between Authors and their publications, but since we have not finished Publications table, this table is also empty so far. In addition, during analysis of our collected data, we used pandas DataFrame to preprocess raw data and stored them in Json format for future use. By the time of this report, we only analyze Authors data and Interests data. Hopefully, we will involve other data mentioned above in our final version.

3. Descriptive Statistics

We visualize the Interests data as the Figure 1 below. It shows the top 30 interests or focused research field among all 2908 Rutgers scholars we have found from Google Scholar. The top three topics are Machine Learning, Neuroscience, and Computer Vision. Since there are 4597 interests among 2908 scholars, the average number of

topics each scholar interested in is 1.58.

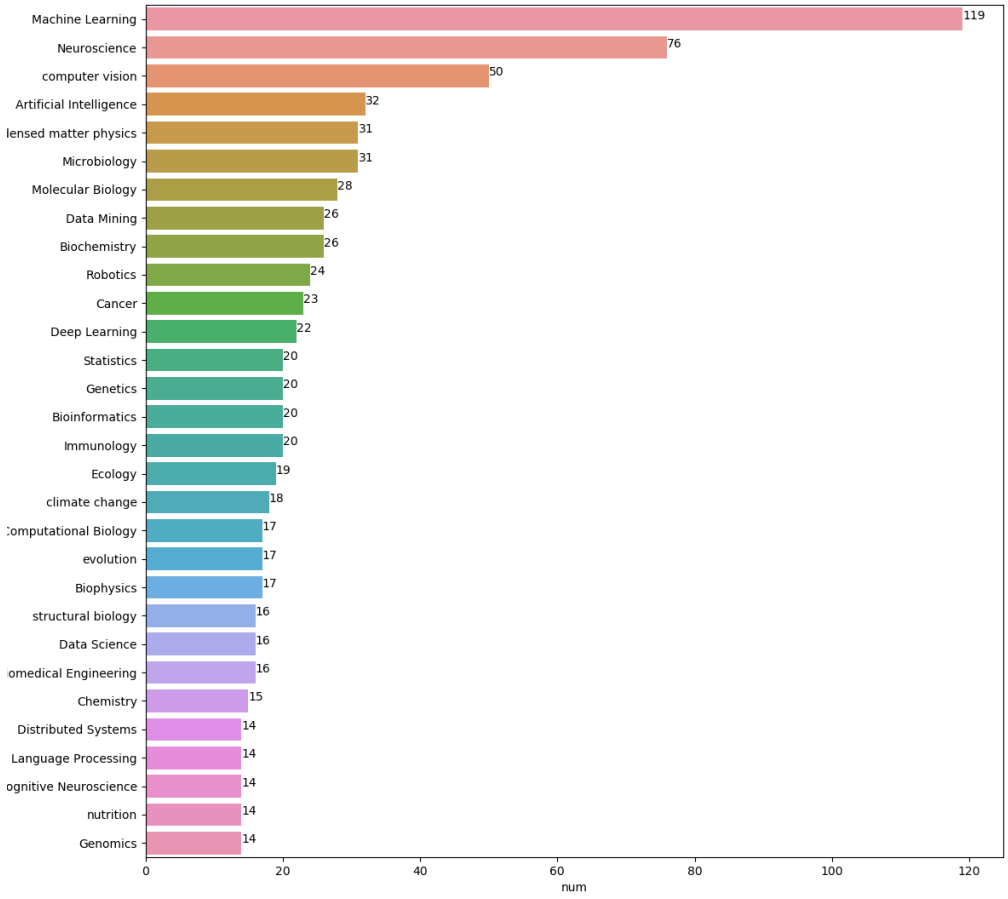

Figure 1: Top 30 Interests Among 2908 Rutgers Scholars

4. Data Analysis, Visualization, and Insights

As we have found out the mean topic number per scholar interested in is 1.58, it shows that one scholar may have multiple interests. With this in mind, we consider that different authors may have the same interests and therefore, there may be some relations between each author bonded by their focused topics. In the same manner, we also may find out some relations between each interest because the more two interests displayed on a scholar profile, the higher possibility of developing an interdisciplinary research they have. Based on these considerations, we are able to create a link between two scholars or two interests. For now, we've already drawn a graph that shows the links between each scholar. Figure 2 below is the attempted graph which shows the relations among 2908 Rutgers scholars. For the most crowded part on the bottom of the graph, it shows the scholars who connected by the topic Machine Learning. Besides that, there are many isolated nodes which means that many scholars do not have the same interest that can connect them with others based on our data. With this network, students or

professors at Rutgers can find potential collaborators in a more convenient and efficient way.

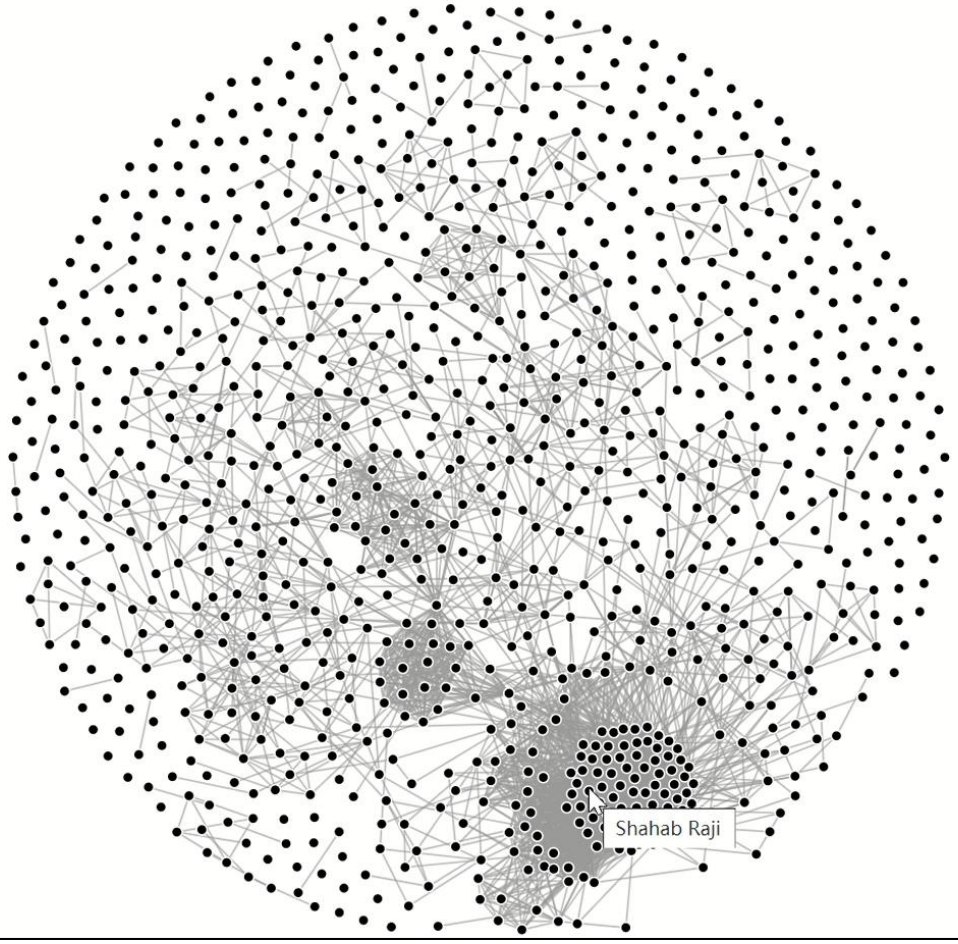

Figure 2: Relation Graph of 2908 Rutgers Scholars

5. Future Plans

As we mentioned in above sections, we are still working on collecting publication data, which contains title, published date, description, co-authors, and cited numbers. By the time we complete the data collection for publications, we plan to do more analysis such as computing and visualizing cosine similarity between scholars, and visualizing total number of publications each year. Also, since we have drawn the network of related scholars as Figure 2 shows but left the network of linked topics, we plan to finish the left graph and analyze it (i.e. what topic are connected most by others?)