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

Group 50: Junyan Dai & Feiyu Zheng

1. Our Goal

When undergraduate Rutgers students want to start their research journey, they probably may meet 2 problems: They may not know which research areas they could connect to produce new thought. They may not know which Rutgers researchers are in those research areas and could help them start their research journey. Therefore, we want to find a possible solution to these two problems based on data on Google Scholar to on the one hand, show those often related research areas at Rutgers; on the other hand, build a system for students to type in their possible interests and find specific Rutgers researchers who are related to those research areas and can probably help them start their research journey.

2. 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 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.

3. Data Format Description

We collected data from web pages and stored them in local MySQL database. Currently, as the following Figure 1 shows, we have designed six tables: Authors, Interests, Organizations, Publications, Authors_to_Interests, and Authors_to_Publications.

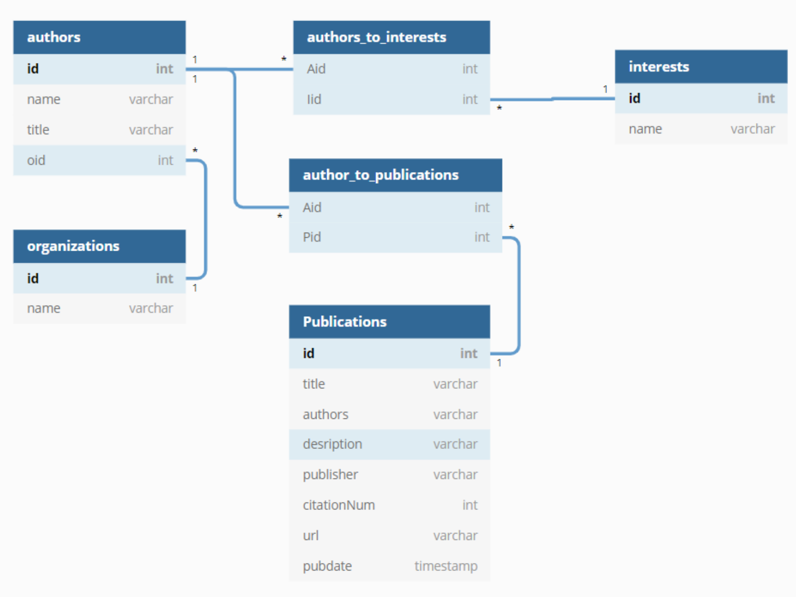


Figure 1 Table Relation

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 during the process of crawling the publication, the anti-crawling system of google keeps banning our IP addresses and hence makes it much harder for us to get the publication data. Therefore, instead of stopping and fixing the crawling problem for a long time, we finally decide to use all the data we get to start the data analysis and leave the publication part for a future work. 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 are not able to collect the publication data, 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.

4. Descriptive Statistics

After we observe the data and filter out invalid ones, we finally get 4597 valid different types of interests (totally 7919 valid interests) and 2567 valid researchers. We visualize the Interests data as the Figure 2 below.

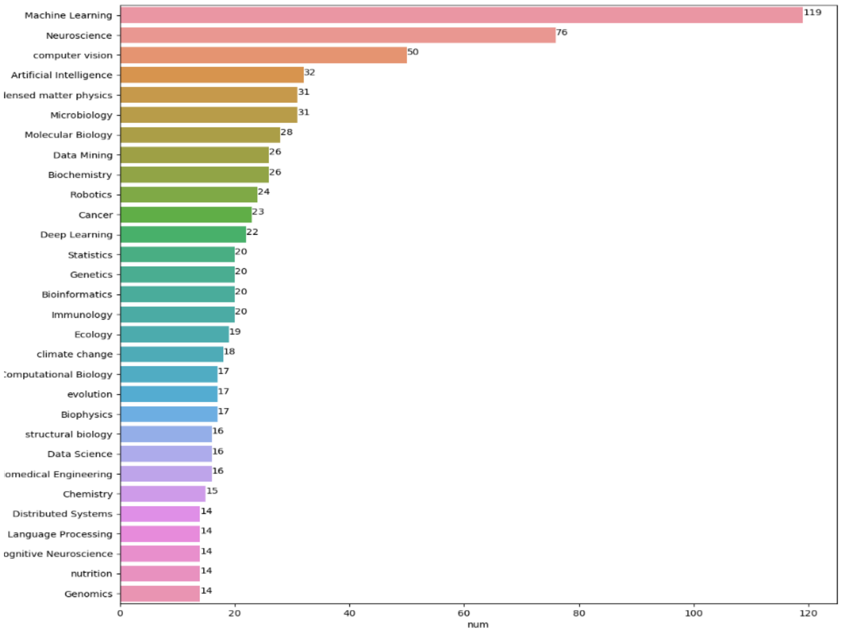


Figure 2: Top 30 Interests Among 2908 Rutgers Scholars

It shows the top 30 interests or focused research field among all 2567 Rutgers scholars we have found from Google Scholar. The top three topics are Machine Learning, Neuroscience, and Computer Vision. Since the 30th interest, Genomics, is focused by 14 researchers, it means that each of the rest 4567 different types of interests may only be focused by few researchers.

The following boxplot shows the distribution of the number of interests each researcher has. The mean is 3.08. The standard deviation is 1.4783. The maximum is 9 while the minimum is 1. The following boxplot also shows that the majority of researchers have 2-4 different interests.

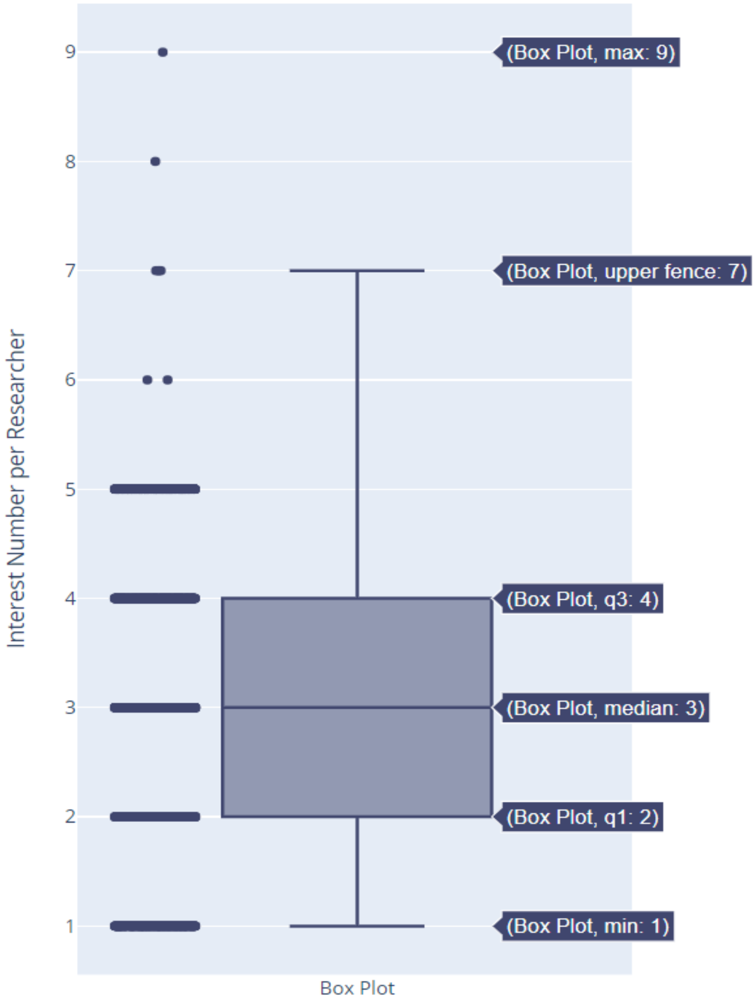


Figure 3: Distribution of Number of Interests Each Researcher Has

5. Data Analysis, Visualization, and Insights

As we have found out the mean topic number per scholar interested in is 3.08, 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. Figure 3 below is a graph which shows the relations among 2567 Rutgers scholars. For the most crowded part on the bottom of the graph, it shows the scholars who are 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.

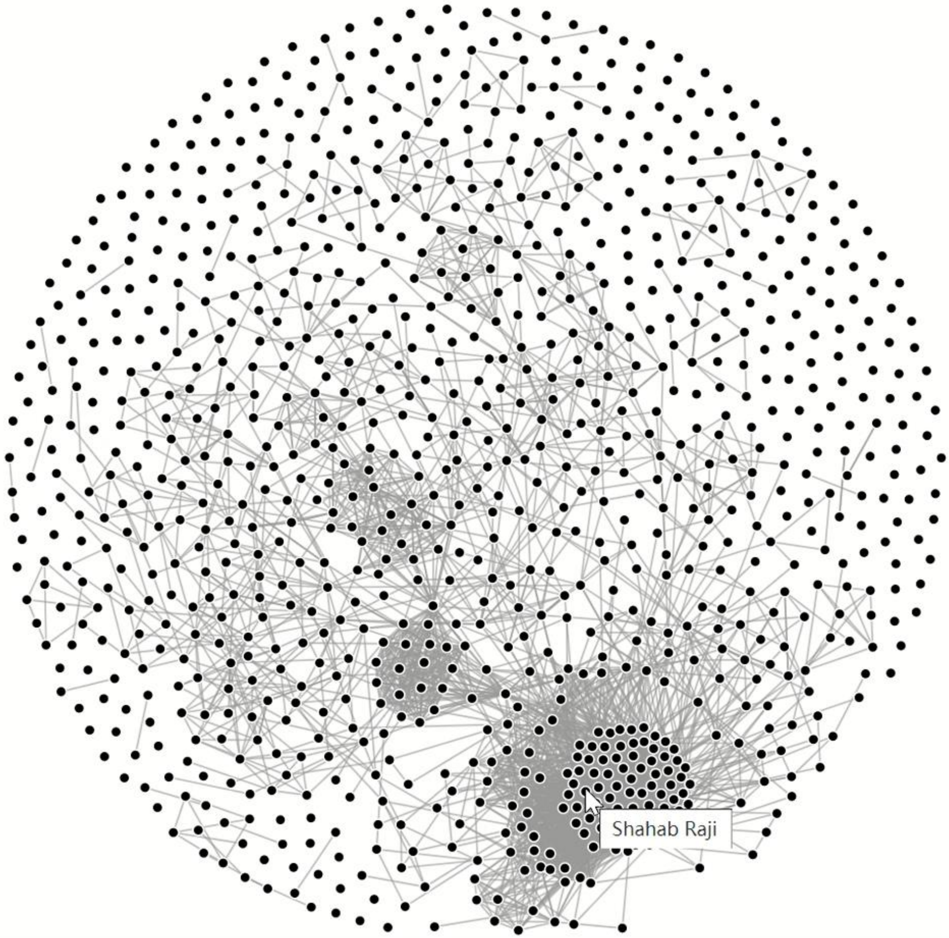
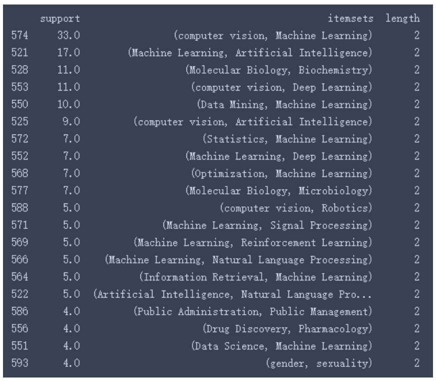


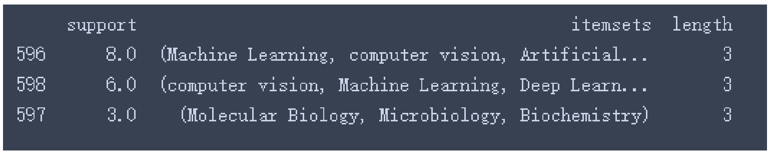
Figure 3: Relation Graph of 2567 Rutgers Scholars

Figure 4 and 5 below are two tables of frequent interest sets with the length of 2 and 3 respectively. We get these two tables by apriori algorithm which means that these interest sets in the table are frequently focused by 2567 Rutgers researchers.



```
     support                                          itemsets  length
574     33.0           (computer vision, Machine Learning)       2
521     17.0    (Machine Learning, Artificial Intelligence)       2
528     11.0           (Molecular Biology, Biochemistry)         2
553     11.0              (computer vision, Deep Learning)       2
550     10.0              (Data Mining, Machine Learning)        2
525      9.0     (computer vision, Artificial Intelligence)      2
572      7.0               (Statistics, Machine Learning)        2
552      7.0            (Machine Learning, Deep Learning)        2
568      7.0             (Optimization, Machine Learning)        2
577      7.0           (Molecular Biology, Microbiology)         2
588      5.0                   (computer vision, Robotics)       2
571      5.0         (Machine Learning, Signal Processing)       2
569      5.0  (Machine Learning, Reinforcement Learning)         2
566      5.0 (Machine Learning, Natural Language Processing)     2
564      5.0      (Information Retrieval, Machine Learning)      2
522      5.0 (Artificial Intelligence, Natural Language Pro...   2
586      4.0  (Public Administration, Public Management)         2
556      4.0              (Drug Discovery, Pharmacology)         2
551      4.0              (Data Science, Machine Learning)       2
593      4.0                         (gender, sexuality)         2
```

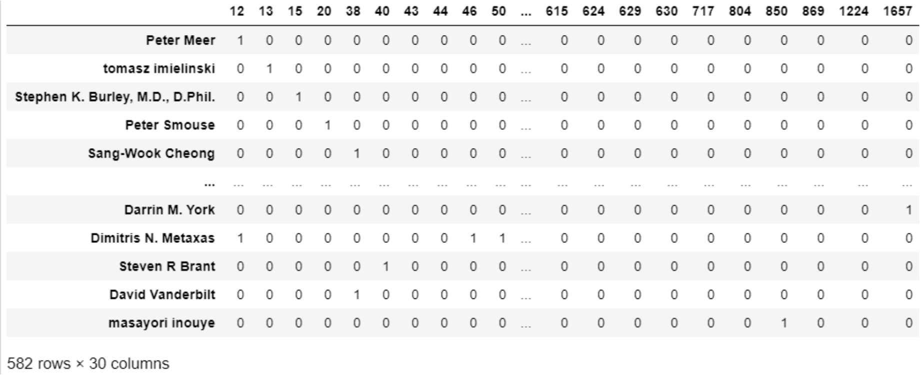
Figure 4: Frequent Interest Sets (Length = 2, minSup = 3)



```
     support                                           itemsets  length
596      8.0  (Machine Learning, computer vision, Artificial...       3
598      6.0  (computer vision, Machine Learning, Deep Learn...       3
597      3.0    (Molecular Biology, Microbiology, Biochemistry)       3
```

Figure 5: Frequent Interest Sets (Length = 3, MinSup = 3)

Since most interests in frequent interest sets come from top 30 interests, we choose researchers whose interests contains one or more of these top 30 interests to classify them into several categories. Since we do not know what exact categories they are, we propose an unsupervised machine learning method to first separate them into different clusters and then figure out the category based on the common interests of each researcher in each cluster. Figure 6 below shows the one hot encoding table for researchers who have one or more than one top 30 interests. Since there are 30 interests and hence 30-dimension, it is hard to visualize each point on a graph. Therefore, we use Principle Components Analysis (PCA) method to reduce the dimensionality and make the point visible in 3D- or even 2D-graph. Figure 7 and 8 below are the graphs of clusters in 3D and 2D respectively.



	12	13	15	20	38	40	43	44	46	50	...	615	624	629	630	717	804	850	869	1224	1657
Peter Meer	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
tomasz imielinski	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Stephen K. Burley, M.D., D.Phil.	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Peter Smouse	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Sang-Wook Cheong	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Darrin M. York	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
Dimitris N. Metaxas	1	0	0	0	0	0	0	0	1	1	...	0	0	0	0	0	0	0	0	0	0
Steven R Brant	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
David Vanderbilt	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
masayori inouye	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0

582 rows × 30 columns

Figure 6: One Hot Encoding Table (Researcher Name—Interest ID)

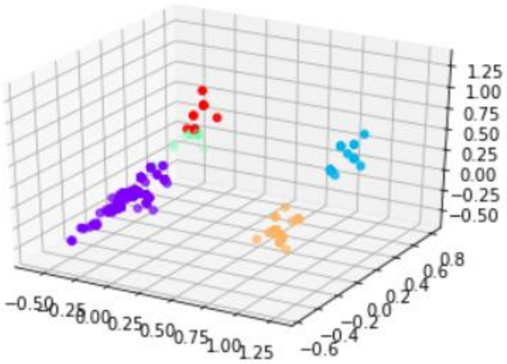


Figure 7: Clusters in 3D Graph

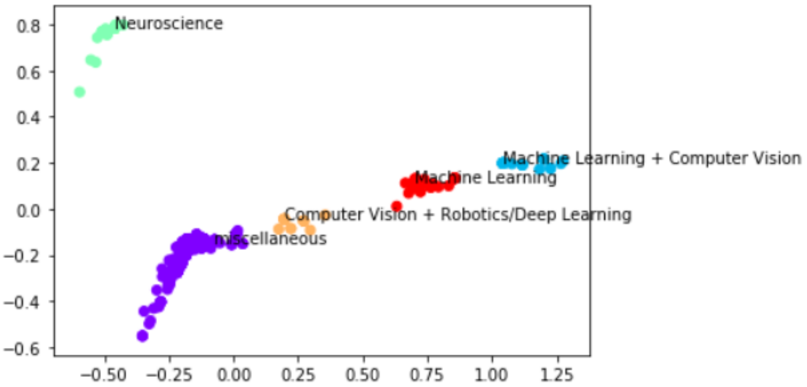


Figure 8: Clusters in 2D Graph

During the modelling process, we tried different hyperparameters and got different graph as Figure 9-12 below shows. We finally take 5 clusters since we think 5 clusters properly fit the training data while less-than-5 clusters may underfit the data and more-than-5 clusters may overfit the data. By analyzing the 5-clustered data, we figure out each category as Figure 8 above shows. Blue points represent researchers whose interests have a strong relation with both Machine Learning and Computer Vision. Red points stand for researchers whose interests have a strong relation with Machine Learning. Orange points represents researchers have a strong relation with both Computer Vision and Robotics/Deep Learning. Green points represent researchers have a strong relation with Neuroscience. Purple points represent researchers have a strong relation with other interests.

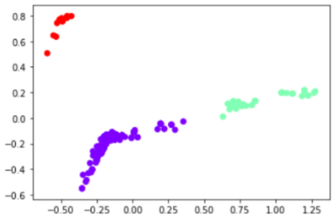


Figure 9: 3 Clusters (Underfit)

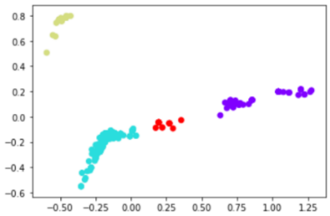


Figure 10: 4 Clusters (Underfit)

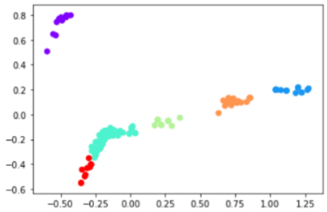


Figure 11: 6 Clusters (Overfit)

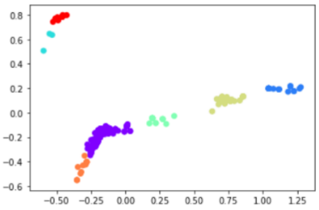
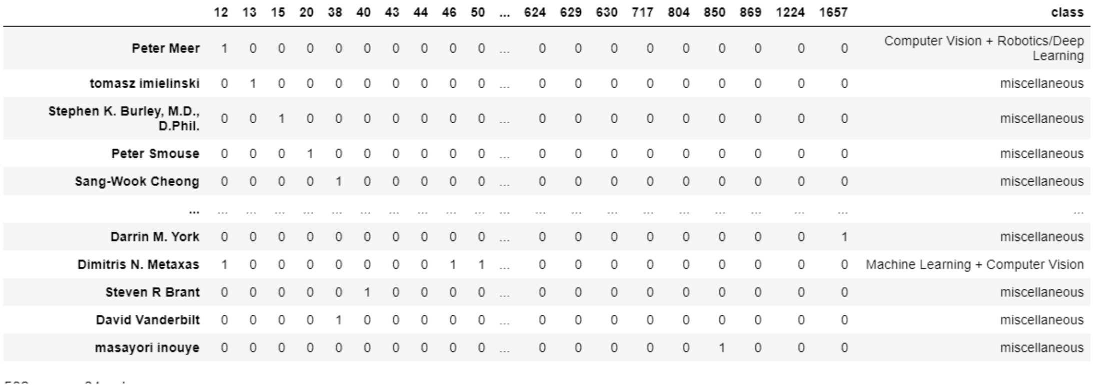


Figure 12: 7 Clusters (Overfit)

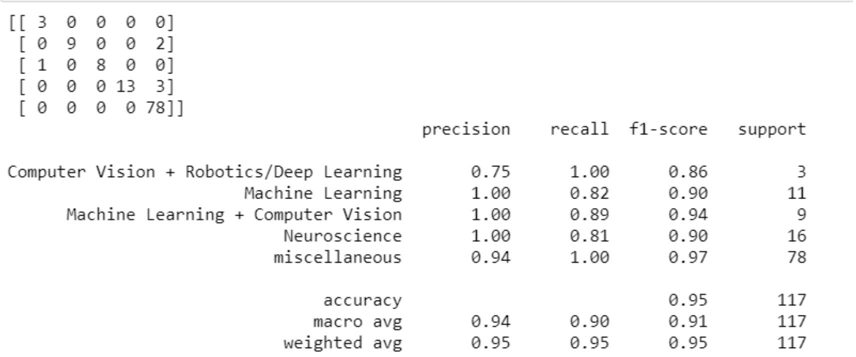
After we figure out the category of each professor, we add their corresponding classes to the data as the Figure 13 shows.



	12	13	15	20	38	40	43	44	46	50	...	624	629	630	717	804	850	869	1224	1657	class
Peter Meer	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Computer Vision + Robotics/Deep Learning
tomasz imielinski	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	miscellaneous
Stephen K. Burley, M.D., D.Phil.	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	miscellaneous
Peter Smouse	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	miscellaneous
Sang-Wook Cheong	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	miscellaneous
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Darrin M. York	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	miscellaneous
Dimitris N. Metaxas	1	0	0	0	0	0	0	0	1	1	...	0	0	0	0	0	0	0	0	0	Machine Learning + Computer Vision
Steven R Brant	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	miscellaneous
David Vanderbilt	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	miscellaneous
masayori inouye	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	miscellaneous

Figure 13: Data with Classified Labels

With labels on data, we are able to run a supervised machine learning method to build a model for prediction. Since we do not have enough data sample, before the learning process starts, we break our original data into two parts: 80% as the training data and 20% as the test data. After the model is built, we use the test data and get the following model with a proper statistic as Figure 14 shows.



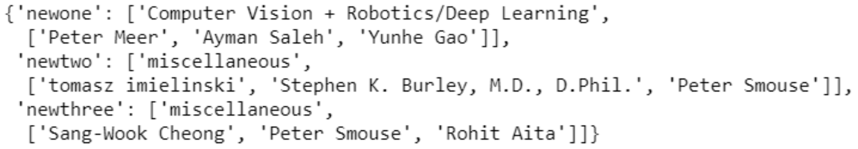
```
[[ 3  0  0  0  0]
 [ 0  9  0  0  2]
 [ 1  0  8  0  0]
 [ 0  0  0 13  3]
 [ 0  0  0  0 78]]
                                          precision    recall  f1-score   support

Computer Vision + Robotics/Deep Learning       0.75      1.00      0.86         3
                        Machine Learning       1.00      0.82      0.90        11
      Machine Learning + Computer Vision       1.00      0.89      0.94         9
                            Neuroscience       1.00      0.81      0.90        16
                           miscellaneous       0.94      1.00      0.97        78

                                accuracy                           0.95       117
                               macro avg       0.94      0.90      0.91       117
                            weighted avg       0.95      0.95      0.95       117
```

Figure 14: Prediction Model Statistics

With this model, we can then classify any new researchers into a specific class based on their interests. Figure 15 below gives an example of prediction.



```
{'newone': ['Computer Vision + Robotics/Deep Learning',
  ['Peter Meer', 'Ayman Saleh', 'Yunhe Gao']],
 'newtwo': ['miscellaneous',
  ['tomasz imielinski', 'Stephen K. Burley, M.D., D.Phil.', 'Peter Smouse']],
 'newthree': ['miscellaneous',
  ['Sang-Wook Cheong', 'Peter Smouse', 'Rohit Aita']]}
```

Figure 15: Prediction Example

As the example shows, 'newone' is an sample with an interest—computer vision. The model predicts that it belongs to 'Computer Vision + Robotics/Deep Learning' class and then lists three recommended researchers—Peter Meer, Ayman Saleh, and Yunhe Gao. “newtwo” is an sample with two interests—data mining and computational biology. The model predicts that it belongs to “miscellaneous” class and then lists three recommended researchers—Tomasz imielinski, Stephen K. Burley, M.d., D. Phil., and Peter Smouse. “new threee” is an sample with two interests—AI and Genetics. The model predicts that it belongs to “miscellaneous” class and then lists three recommended researchers—Sang-Wook Cheong, Peter Smouse, and Rohit Aita. So this example imitates the process that a student input some of their interested research areas and our model then processes the input to predict which class the student may be in and gives some recommended researchers that may help the student.

6. Future Plans

As we mentioned in above sections, due to google anti-crawling system, we do not manage to collect publication data, which contains title, published date, description, co-authors, and cited numbers. By the time we complete the data collection for publications, we can do more analysis such as computing and visualizing cosine similarity between scholars, and visualizing total number of publications each year. Besides, since the size of our data is small, we could find more Rutgers researchers from other platforms to expand our dataset. With complete dataset, we could build a recommendation system which can show recent research trend at Rutgers and also help not only students to start their research journey but also researchers to find collaborators.

7. Acknowledgment

(a) Crawler Tool: Scrapy

(b) Third party packages:
pandas, sklearn.decomposition, sklearn.cluster.AgglomerativeClustering, matplotlib, d3.js, pymysql, json, plotly, scipy, numpy, mlxtend, seaborn

(c) All Used Data Comes from Google Scholar

(d) References:

[1] Apriori Algorithm Code:
http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/

[2] Boxplot Code:
https://plotly.com/python/v3/basic-statistics/

[3] Unsupervised Machine Learning (Cluster) Code:
https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/

[4] Supervised Machine Learning (KNN) Code:
https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/

[5] Relation Graph Among Researchers:
https://d3js.org/