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

Group 50: Junyan Dai

Feiyu Zheng

Analyzing Rutgers Google Scholar

Contents

- Our Goal
- Data Collection
- Data Exploration
- Modelling Data

Common Problems

If students want to start their research journey, they probably may meet some the following problems:

1. They may not know which research areas could be connected.

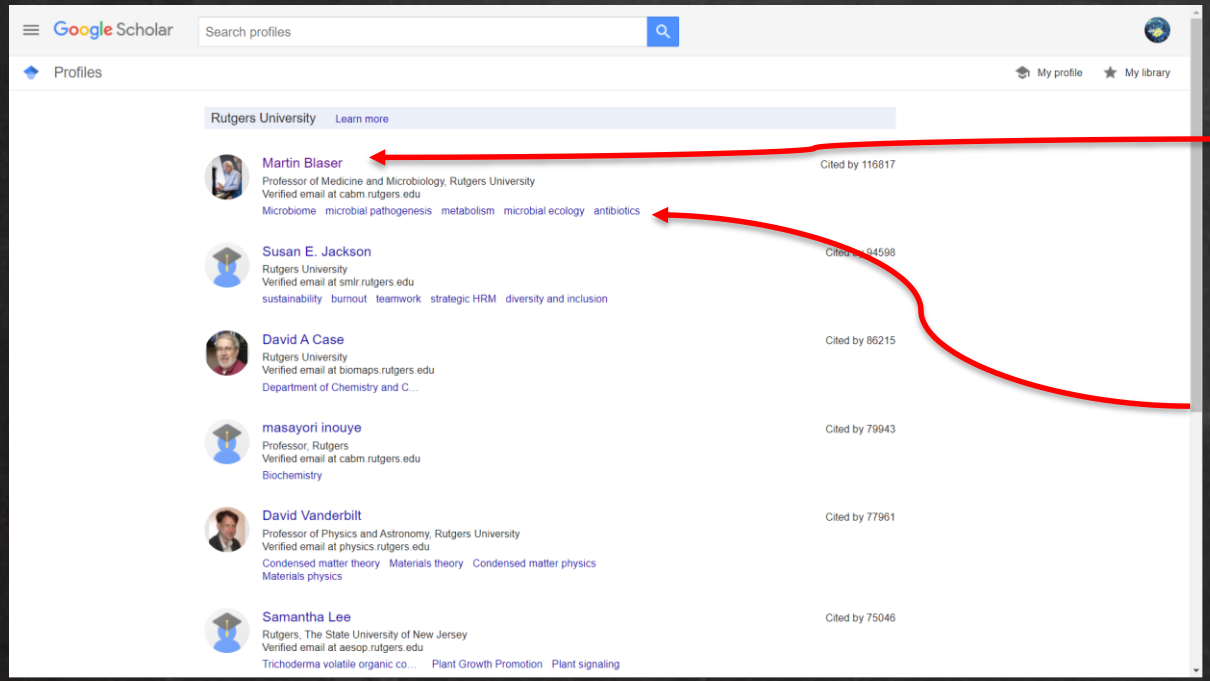
2. They may not know which professors are in those research areas and could help them start their research journey.

Our Goal

1. We want to find which areas are often related together so that it can give students a common sense about possibly intersected areas.

2. Also, if students are interested about several research areas, we may also want to find out which professors at Rutgers are working on these areas so that students could know who they could contact to start their research.

Data Collection Overview



Name

Web crawler → Data

Research areas

Database

Data Exploration

Valid different interests = 4597

Valid total interests = 7919

Valid Researchers = 2567

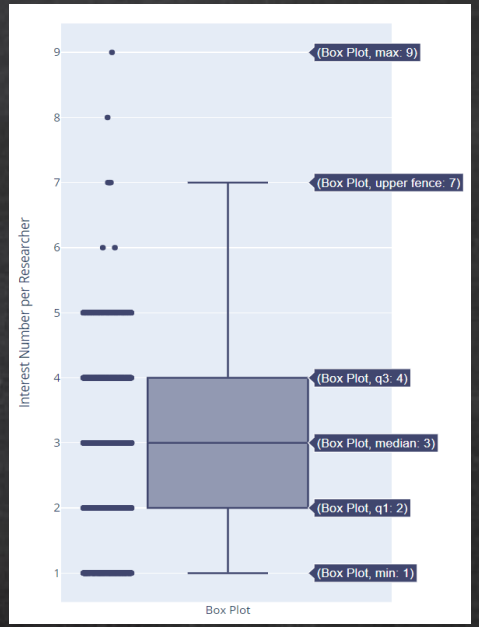
Average Interest/Researcher ≈ 3.08

Standard Deviation ≈ 1.4783

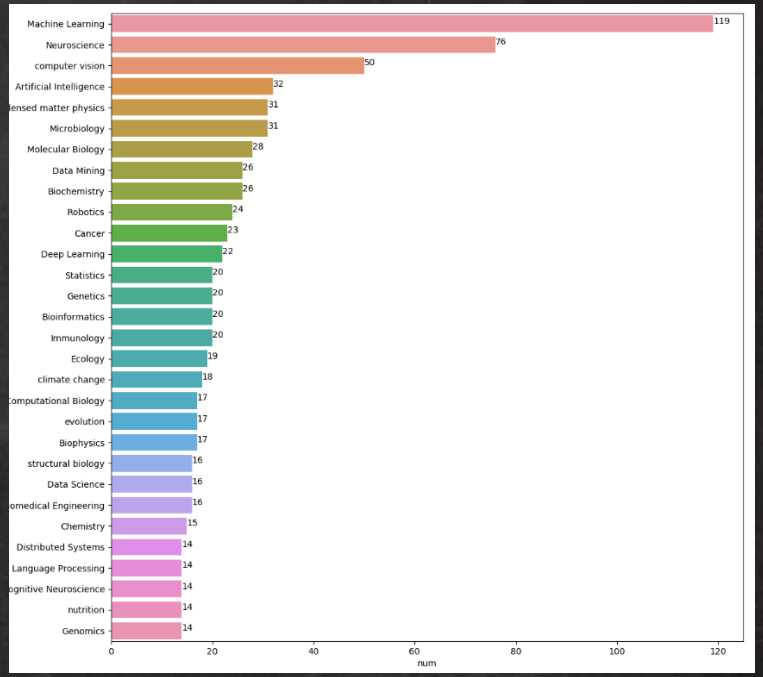
Maximum = 9

Minimum = 1

Boxplot



Top 30 Interests among 2908 Researchers:



Summary:

1. Most researchers have more than 1 interests (commonly 2-4).
2. Only a small amount of interests are popular among researchers.

Data Exploration

Top 20 Frequently related Interests (Apriori Algorithm, minSup=3):

	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

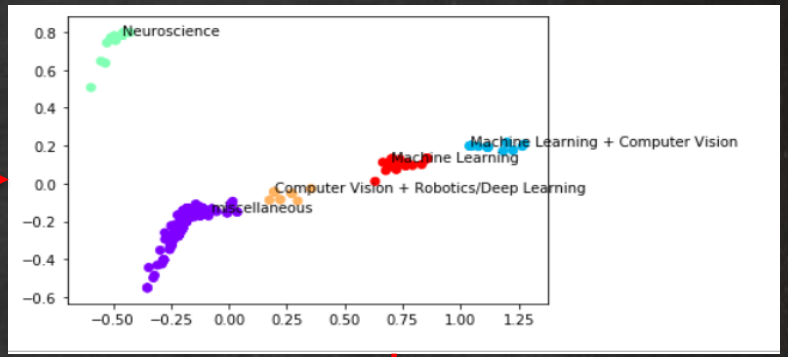
Length = 2

	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

Length = 3

Modeling Data

5 Classes



Unsupervised Machine Learning Method

Input Interest

Connected Interest

Recommended Researcher

Prediction Model

Supervised Machine Learning Method
(80% training, 20% test)

	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

Add classified label to each data

Modeling Data

```
{'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']]}
```

newone (computer vision) → Prediction Model → Computer Vision + Robotics/Deep Learning

Researcher that may help: Peter Meer Ayman Saleh Yunhe Gao

newtwo (data mining + computational biology) → Prediction Model → miscellaneous

Researcher that may help: Tomasz imielinski Stephen K... Peter Smouse

Acknowledgment

1. Crawler Tool: Scrapy
2. Third party packages: pandas, sklearn.decomposition, sklearn.cluster.AgglomerativeClustering, matplotlib, d3.js, pymysql, json, plotly, scipy, numpy, mlxtend, seaborn
3. All Used Data Comes From Google Scholar
4. 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/

Thanks For Watching!

Group 50: Junyan Dai

Feiyu Zheng