TEAM happy data BIGDATA

DSCI550

Group Name: Happy Data

Students:

Jinshan Yang
Weiqian Zhang
Bingqing Liu
Shengyu Sun
Zhong Wang
Hang Yang

University of Southern California

13th June 2022

Contents

1 Introduction

In Task4, we added additional features to the data about the author. Then we crawled other features including affiliation university, degree, journals and career duration using BeautifulSoup. We noticed we can use URLs from the dataset to crawl data about the author. As for Task5, with three other datasets we identified, we extracted several features and joined them to the Bik dataset. Finally, in Task6, we calculated and compared Jaccard similarity, edit-distance and cosine similarity. We noticed that the cosine similarity score is higher than the other two similarity metrics. In our opinion, that's because the content in Colleges and Universities.csv was very similar to the Bik dataset.

The report will follow the instructions of homework1 to answer each questions.

- **Question 3.2.1** For each feature added, what type of queries it will allow you to answer and also how to compute the feature?(6pts)
- **Question 3.2.2** Compare the results from Jaccard, Cosine Distance, and Edit Similarity- What similarity metrics produced more(in your opinion)accurate measurements? Why?(6pts)
- **Question 3.2.3** What you noticed about the dataset as you completed the tasks. (6pts)
- **Question 3.2.4** What questions did your new joined datasets allow you to answer the Bik et al papers previously unanswered? (6pts)
- **Question 3.2.5** What did the additional datasets suggest about "unintended consequences related to media forensics data? (6pts)
- **Question 3.2.6** You should also clearly explain which datasets you used to join the problematic papers data and how you extracted the new features from each datasets. (6pts)

Then the report will give a summary and extra credit.

2 Questions' Answer

2.1 Question 3.2.1

For each feature added, what type of queries it will allow you to answer and also how to compute the feature?(6pts)

To get features that Task4 metioned: Lab Size(number of students), Publication Rate, Other Journals Published In, Information about First Author including(Affiliation University, Duration of Career, Highest degree obtained and Degree Area), we go through the source codes of each website. The format of the source codes is html and json. So we can use tools like JsonSQL, JSONPath, TaffyDB, objeq, linq.js, json:select(), beautifulsoup and HTMLParser to query the source code and get the content we need. But in this task we will use beautifulsoup.

When computing the features, different features correspond to different ways, we will explain it in detail. Generally speaking, lab size is counted by getting the length of $''data-js-tooltip'' : ''tooltip-trigger''and'class' :' authors-list-item'$; duration year and publication rate are counted using formula:

$$duration\,career = int(year\,of\,latest\,publication - year\,of\,first\,publication) \tag{1}$$

$$publication\,rate = \frac{total\,journal\,number}{dutation\,career} \tag{2}$$

- **Affiliation university & lab size:**
  - **PLOS papers:**

    In the original data set, the first 48 papers can be viewed on PLOS($https://journals.plos.org/plosone/$), so we observed the introduction of the author information on PLOS. Since it contains the school information of the first author and other staff members of this paper, we scrape this information on PLOS. We found that as long as we input the DOI string of paper in the URL, we can jump to the article page. Therefore, after designing the URL and looking for the label name according to the page content, we used beautiful soup to scrape. Therefore, we wrote a function "$get\_affiliation\_labsize$" to crawl the affiliation University and lab size in PLOS.
  - **Pubmed papers:**

    Also, we notice that other 166 papers' affiliation university and lab size can be found on pubmed($https://pubmed.ncbi.nlm.nih.gov/$). First, we also observe the URL and find that the search result of the article can be obtained by adding the name of the article to the search URL. But there are many results here. We need to find the search result that contains the author information. So we wrote the get_med_link function, used to obtain the article search result link so that we can query the author information from the link.

  Then, by accessing links, we can crawl to the author's relevant information through beautifulsoup. To do this, we wrote the $get\_affiliation\_university\_med$ function.

- **Highest degree** & **Degree area:**

  In order to crawl the highest degree and degree area, we found relevant information on the research gate($https : //www.researchgate.net/$). When we try to scrape by directly accessing the URL, we find that the research gate will hinder our scraper. In order to avoid being monitored by the anti crawler mechanism, we chose to use selenium library to simulate the process of accessing websites using chrome.

  We first define a Driver as an object that simulates a Chrome browser. By inputting the fixed URL + paper's title, we can get the result page of the paper(the image below). On the returned page, Driver will click the first author(in red box) and then it will jump to the author information page.

  From the author information page, we can scrape the highest degree and the degree area by the content in the red box. After getting content from this label, we still need to do some preparation such as dividing the degree and area. After that, we finish scraping the information of the author's highest degree and degree area.

- **Other journals** & **Duration of Career** & **Publication Rate :**

  In order to crawl the papers published by authors and calculate the publishing rate, we crawl the author information on Pubmed. The reason for using Pubmed instead of PLOS is that PubMed has the author information of 214 papers. We can still get the author's search results page by directly accessing the fixed url plus author's name. By querying the labels, we can get the number of papers published by the author.

  Then calculate the author's duration of career and publication rate through the following formulas. Because the website has the problem of inconsistent format, we have also added the classification of different page formats and the handling of abnormal error reports to the function.

2.2 Question 3.2.2

Compare the results from Jaccard, Cosine Distance, and Edit Similarity-What similarity metrics produced more(in your opinion)accurate measurements? Why?(6pts)

For this question, the datasets we use are Bik dataset - papers with endpoint reached.tsv(the original Bik dataset), $updated\_bik\_with\_features.csv$(datasets combined after task4 and task5)and three new datasets with different MIME types: $Colleges\_and\_Universities.csv$, $NSF{-}metadata.xlsx$ and $QSrank.txt$, As a result, we get **cosine similarity metrics** produced more accurate measurements(0.996) comparing to others, details are following.

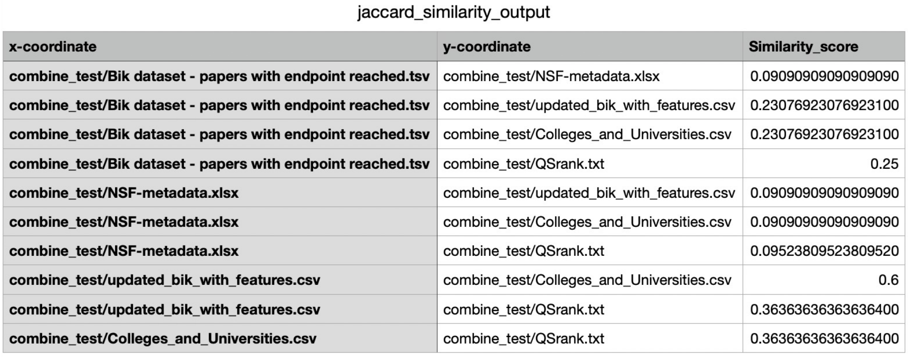
On the above datasets, Jaccard similarity, Cosine similarity, and Edit value similarity are performed.

Similarity: Similarity functions are used to determine how far two vectors, numbers, or pairs are apart. It's a metric for determining how similar two objects are. If the distance between two objects is small, they are said to be similar, and vice versa.

- **Jaccard Similarity:**

Comparing members from two sets to discover which are common and which are unique is done by using the Jaccard similarity index (also known as the Jaccard similarity coefficient). You can see how similar the data is by using a percentage scale that ranges from zero all the way up to one hundred percent. The closer the two groups are, the higher the proportion.

The Jaccard similarity as follow table:



jaccard_similarity_output

x-coordinate	y-coordinate	Similarity_score
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/NSF-metadata.xlsx	0.09090909090909090
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/updated_bik_with_features.csv	0.23076923076923100
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/Colleges_and_Universities.csv	0.23076923076923100
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/QSrank.txt	0.25
combine_test/NSF-metadata.xlsx	combine_test/updated_bik_with_features.csv	0.09090909090909090
combine_test/NSF-metadata.xlsx	combine_test/Colleges_and_Universities.csv	0.09090909090909090
combine_test/NSF-metadata.xlsx	combine_test/QSrank.txt	0.09523809523809520
combine_test/updated_bik_with_features.csv	combine_test/Colleges_and_Universities.csv	0.6
combine_test/updated_bik_with_features.csv	combine_test/QSrank.txt	0.36363636363636400
combine_test/Colleges_and_Universities.csv	combine_test/QSrank.txt	0.36363636363636400

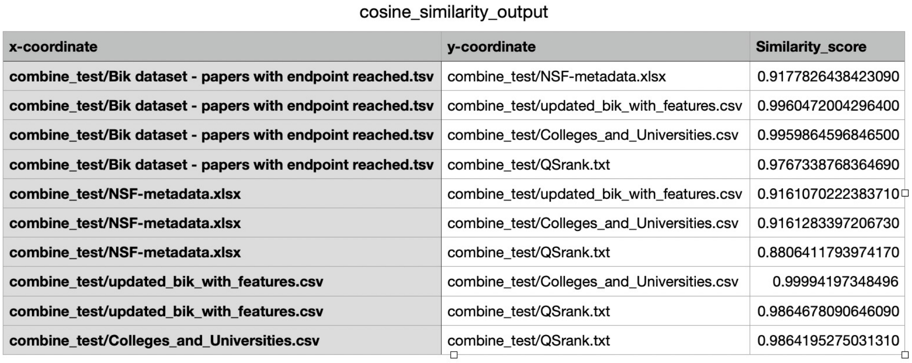
Figure 1: Jaccard Similarity

From the Figure 1, we could find the highest similarity is between file $updated\_bik\_with\_features.csv$ and file $Colleges\_and\_Universities.csv$, which is 0.6

- **Cosine Similarity:**

  Cosine similarity is used to quantify the similarity of two vectors in an inner product space. The cosine of the angle between two vectors is used to detect if they are pointing in the same general direction. It's widely used in text analysis to determine the degree to which two documents are similar. Using Cosine similarity, we can see how the datasets compare to each other.

  The Cosine Similarity as follow table:



cosine_similarity_output

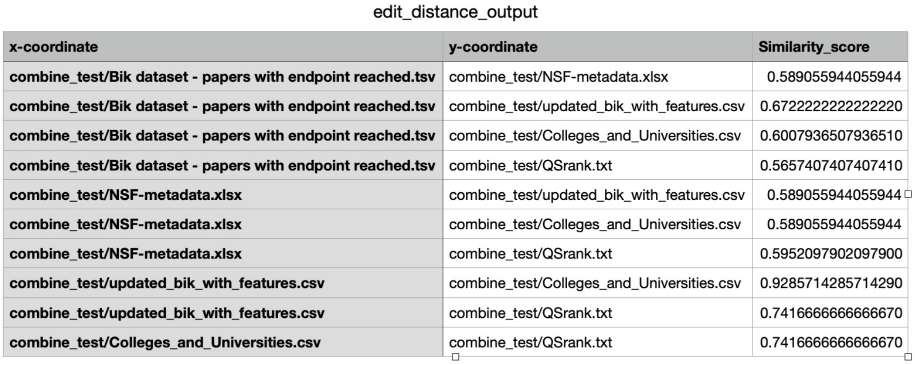
x-coordinate	y-coordinate	Similarity_score
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/NSF-metadata.xlsx	0.9177826438423090
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/updated_bik_with_features.csv	0.9960472004296400
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/Colleges_and_Universities.csv	0.9959864596846500
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/QSrank.txt	0.9767338768364690
combine_test/NSF-metadata.xlsx	combine_test/updated_bik_with_features.csv	0.9161070222383710
combine_test/NSF-metadata.xlsx	combine_test/Colleges_and_Universities.csv	0.9161283397206730
combine_test/NSF-metadata.xlsx	combine_test/QSrank.txt	0.8806411793974170
combine_test/updated_bik_with_features.csv	combine_test/Colleges_and_Universities.csv	0.99994197348496
combine_test/updated_bik_with_features.csv	combine_test/QSrank.txt	0.9864678090646090
combine_test/Colleges_and_Universities.csv	combine_test/QSrank.txt	0.9864195275031310

Figure 2: Cosine Similarity

From the Figure 2, we could find the highest similarity is between file *updated_bik_with_features.csv* and file Bik dataset - papers with endpoint reached.tsv, which is 0.996

- **Edit Value Similarity:**

  The Edit Value Similarity as follow table:



edit_distance_output

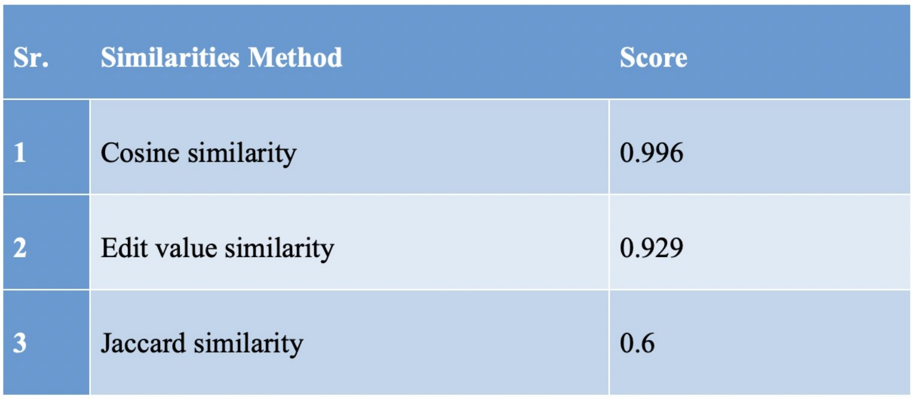
x-coordinate	y-coordinate	Similarity_score
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/NSF-metadata.xlsx	0.589055944055944
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/updated_bik_with_features.csv	0.6722222222222220
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/Colleges_and_Universities.csv	0.6007936507936510
combine_test/Bik dataset - papers with endpoint reached.tsv	combine_test/QSrank.txt	0.5657407407407410
combine_test/NSF-metadata.xlsx	combine_test/updated_bik_with_features.csv	0.589055944055944
combine_test/NSF-metadata.xlsx	combine_test/Colleges_and_Universities.csv	0.589055944055944
combine_test/NSF-metadata.xlsx	combine_test/QSrank.txt	0.5952097902097900
combine_test/updated_bik_with_features.csv	combine_test/Colleges_and_Universities.csv	0.9285714285714290
combine_test/updated_bik_with_features.csv	combine_test/QSrank.txt	0.7416666666666670
combine_test/Colleges_and_Universities.csv	combine_test/QSrank.txt	0.7416666666666670

Figure 3: Edit Value Similarity

From the Figure 3, we could find the highest similarity is between file *updated_bik_with_features.csv* and file *Colleges_and_Universities.csv*, which is 0.929

- **Comparison of Three Similarity:**

  The Comparison of Three Similarity as follow table:



Sr.	Similarities Method	Score
1	Cosine similarity	0.996
2	Edit value similarity	0.929
3	Jaccard similarity	0.6

Figure 4: Comparison of Three Similarity

From the Figure 4, we could find the highest similarity is Cosine similarity, which is 0.996

Therefore, cosine similarity metrics produced more(in your opinion)accurate measurements(0.996).

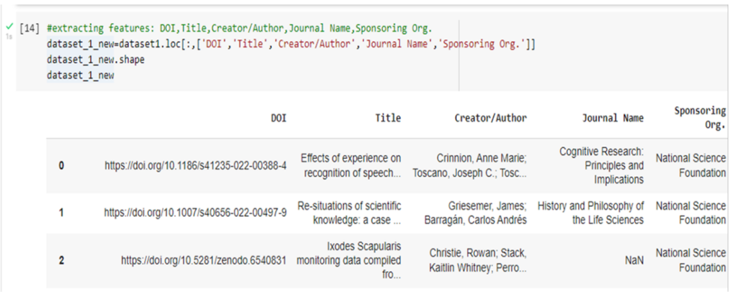
2.3 Question 3.2.3& Question 3.2.6

1. What you noticed about the dataset as you completed the tasks. (6pts) 2. You should also clearly explain which datasets you used to join the problematic papers data and how you extracted the new features. (6pts)

For this question, the datasets we used are Bik dataset - papers with endpoint reached.tsv(the original Bik dataset), $updated\_bik\_with\_features.csv$(datasets combined after task4 and task5)and three new datasets with different MIME types: $Colleges\_and\_Universities.csv$, $NSF-metadata.xlsx$(data link: $https://par.nsf.gov/search/term:PLOS$) and $QSrank.txt$.

- **NSF-metadata.xlsx:**

  We extracted the features of different columns in the NFS metadata, such as DOI, Title, Creator/Author, Journal Name, and Sponsoring Organization. We got about 2000 rows and 5 columns, which is great for getting better results as following Figure 5.



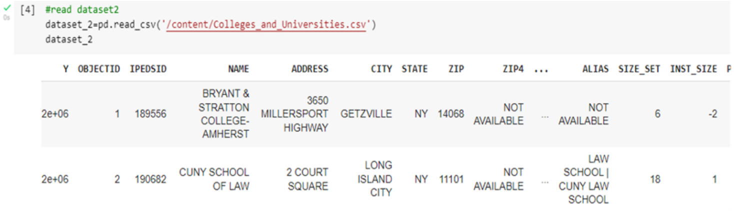
```
[14] #extracting features: DOI,Title,Creator/Author,Journal Name,Sponsoring Org.
     dataset_1_new=dataset1.loc[:,['DOI','Title','Creator/Author','Journal Name','Sponsoring Org.']]
     dataset_1_new.shape
     dataset_1_new
```

	DOI	Title	Creator/Author	Journal Name	Sponsoring Org.
0	https://doi.org/10.1186/s41235-022-00388-4	Effects of experience on recognition of speech...	Crinnion, Anne Marie; Toscano, Joseph C.; Tosc...	Cognitive Research: Principles and Implications	National Science Foundation
1	https://doi.org/10.1007/s40656-022-00497-9	Re-situations of scientific knowledge: a case ...	Griesemer, James; Barragán, Carlos Andrés	History and Philosophy of the Life Sciences	National Science Foundation
2	https://doi.org/10.5281/zenodo.6540831	Ixodes Scapularis monitoring data compiled fro...	Christie, Rowan; Stack, Kaitlin Whitney; Perro...	NaN	National Science Foundation

Figure 5: NSF-metadata.xlsx

- *Colleges_and_Universities.csv*:

  We noticed that there are some categorical data of colleges and universities in the *Colleges_and_Universities.csv* dataset, such as Name, Address, CITY, STATE, ZIP, and so on. We extracted the features of different columns in the Colleges and Universities dataset, such as NAME, ADDRESS, STATE, COUNTRY, COUNTY, SOURCE, WEBSITE. There are 6559 rows and 7 columns in total as following Figure 6.

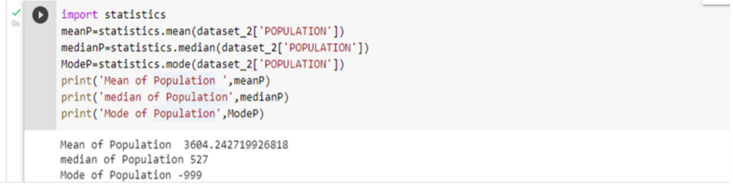


```
[4] #read dataset2
    dataset_2=pd.read_csv('/content/Colleges_and_Universities.csv')
    dataset_2
```

Y	OBJECTID	IPEDSID	NAME	ADDRESS	CITY	STATE	ZIP	ZIP4	...	ALIAS	SIZE_SET	INST_SIZE	P
2e+06	1	189556	BRYANT & STRATTON COLLEGE-AMHERST	3650 MILLERSPORT HIGHWAY	GETZVILLE	NY	14068	NOT AVAILABLE	...	NOT AVAILABLE	6	-2	
2e+06	2	190682	CUNY SCHOOL OF LAW	2 COURT SQUARE	LONG ISLAND CITY	NY	11101	NOT AVAILABLE	...	LAW SCHOOL | CUNY LAW SCHOOL	18	1	

Figure 6: *Colleges_and_Universities.csv*

  We conducted a statistical analysis of the population of colleges and universities as following Figure 7.



```
import statistics
meanP=statistics.mean(dataset_2['POPULATION'])
medianP=statistics.median(dataset_2['POPULATION'])
ModeP=statistics.mode(dataset_2['POPULATION'])
print('Mean of Population ',meanP)
print('median of Population',medianP)
print('Mode of Population',ModeP)
```

```
Mean of Population  3604.242719926818
median of Population 527
Mode of Population -999
```

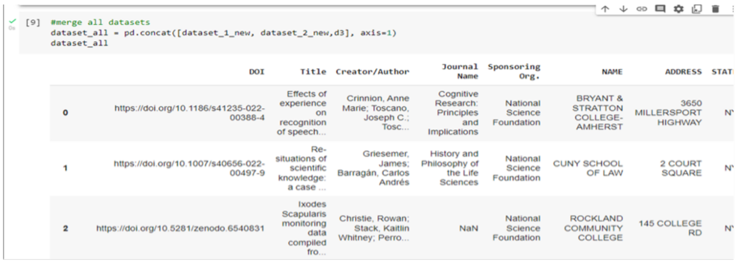
Figure 7: statistical analysis

- **QSrank.txt:**

  We noticed that different university datasets are located in Asia, Europe, North America, Oceania, and Latin America in this dataset.

  We noticed that the data was arranged and split because the dataset lacked proper columns. We've also extracted the data from a text file and also, we have assigned the column names UNI_RANKS, UNI_NAMES, and UNI_AREAS. We found 2000 records in 3 columns across 2000 rows.

After we've completed our analysis of the three datasets, we'll merge them all into a single csv file as following Figure 8.



```
[9] #merge all datasets
    dataset_all = pd.concat([dataset_1_new, dataset_2_new,d3], axis=1)
    dataset_all
```

	DOI	Title	Creator/Author	Journal Name	Sponsoring Org.	NAME	ADDRESS	STAT
0	https://doi.org/10.1186/s41235-022-00388-4	Effects of experience on recognition of speech...	Crinnion, Anne Marie; Toscano, Joseph C.; Tosc...	Cognitive Research: Principles and Implications	National Science Foundation	BRYANT & STRATTON COLLEGE-AMHERST	3650 MILLERSPORT HIGHWAY	N
1	https://doi.org/10.1007/s40656-022-00497-9	Re-situations of scientific knowledge: a case ...	Griesemer, James; Barragán, Carlos Andrés	History and Philosophy of the Life Sciences	National Science Foundation	CUNY SCHOOL OF LAW	2 COURT SQUARE	N
2	https://doi.org/10.5281/zenodo.6540831	Ixodes Scapularis monitoring data compiled fro...	Christie, Rowan; Stack, Kaitlin Whitney; Perro...	NaN	National Science Foundation	ROCKLAND COMMUNITY COLLEGE	145 COLLEGE RD	N

Figure 8: QSrank.txt

2.4 Question 3.2.4

What questions did your new joined datasets allow you to answer the Bik et al papers previously unanswered? (6pts)

The 3 additional datasets could allow us to answer these questions that papers from original Bik dataset could not be answered:

First dataset NSF-metadata.xlsx shows : Since the first 48 papers from PLOS in the Bik et al papers, are the papers from NSF-metadata.xlsx different from the papers from Bik et al papers? Compared with the original Bik, this dataset has more characters of PLOS articles, such as sponsor organization, journal name, publication date, editor, these characters may be more supportive and intricate to information of PLOS articles. And from the type of the dataset, this dataset is from a website with a partnership with the Technical Department of Energy, Office of Scientific and Information. It includes different contents and angles from those of Bik dataset.

Second dataset $Colleges\_and\_Universities.csv$ which illustrates the geographical location of thousands of universities could be corresponding to these questions: The dataset is obtained through Homeland Infrastructure Foundation-Level Data (HIFLD), and it is in the education field showing the colleges and universities, which is specific and inclusive. As to the answer of the question, are the names of universities from this dataset almost the same as the authors' affiliation universities which we scraped based on the original articles from the original Bik dataset? Are the geographical factors correlated to the authors' affiliation universities, such as x,y coordinates, addresses, latitude, longitude? Meanwhile, the 'SOURCE' from this dataset shows the website about basic information of each university or each college beyond the dataset: enrollment and financial situation, therefore, is this information helpful to describe how the university or college shape the authors to conduct the scientific papers?

Third dataset QSrank.txt shows the universities from QS rankings. It is a difficult, political, and controversial practice. There are hundreds of different national and international university ranking systems, many of which disagree with each other. This dataset contains three global university rankings from very different places. It covers more dimensions than the original Bik dataset. The dataset could respond to these questions: Are the universities from QS rankings almost the same as the authors' affiliation universities which we scraped based on the original articles from the original Bik dataset? If almost the same, do the papers have higher reputations from these authors from universities with high QS rankings around the world, if not, will the credibility of these papers be questioned or be doubted around the world? Does the area of universities from QS rankings influence the authors to conduct the scientific research which can be later written as research papers?

2.5 Question 3.2.5

What did the additional datasets suggest about "uninten- ded consequences" related to media forensics data? (6pts)

NSF-metadata.xlsx shows the ranking of universities around the world, therefore, there are several "unintended consequences" related to media forensics data:

1. The access to these articles might be closed for users who do not register in the websites.

2. If the journal or the sponsor organization is closed, there will be the possibility that the articles related to PLOS cannot be accessible.

3. Obviously, it is hard to find the information about authors related to the PLOS articles, which means that we could not make sure the papers' credibility.

Colleges_and_Universities.csv depicts basic information of each university or each college, especially geographic information, therefore, there are some"unintended consequences" related to media forensics data:

1. The location of each university or each college might be changed as time goes by.

2. X,Y coordinates might be different based on the different standard of geographic measurements.

3. The labeled ID, such as OBJECTID, IPEDSID, can be varied based on the data scientists' preferences.

QSrank.txt mainly shows the ranking of universities around the world, therefore, there are several"unintended consequences" related to media forensics data:

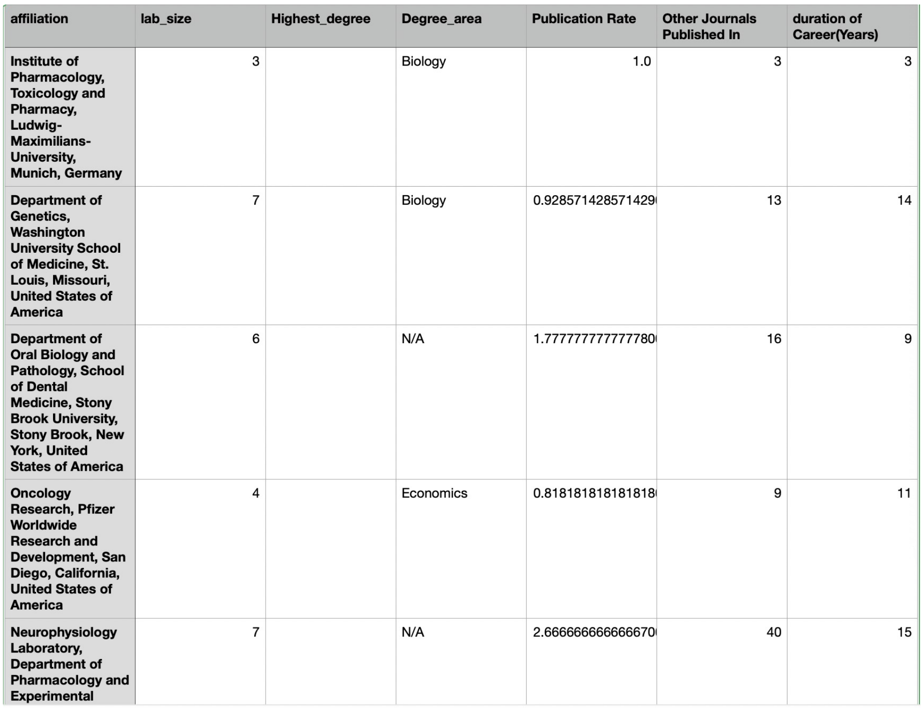
1. Ranking of each university can be changed and varied every time

2. Some universities with 800-1000 around the world can sometimes be ignored in the data analysis.

3. Meanwhile, we could not ignore the possibility that some universities do not want to be mentioned in the QS rankings and some universities from QS rankings might be closed next year.

3 Summary

First, we add additional features to the data about the Lab Size(number of students), Publication Rate, Other Journals Published In, Information about First Author including(Affiliation University, Duration of Career, Highest degree obtained and Degree Area) to update the original Bik dataset as Figure 9.

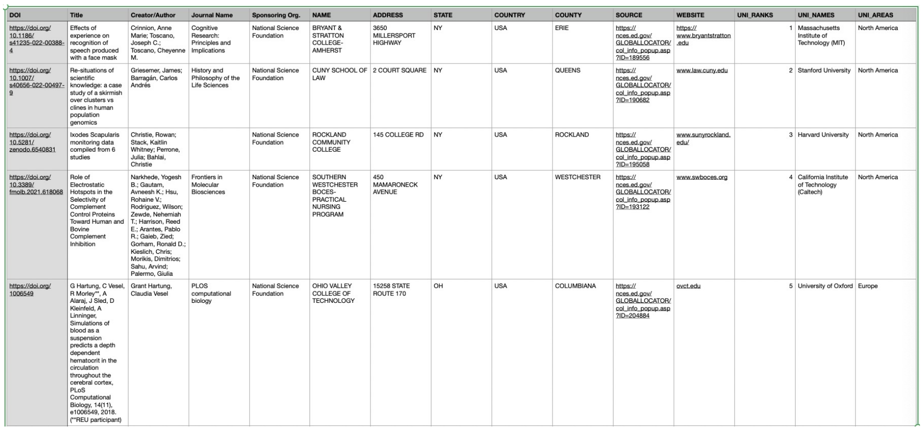


affiliation	lab_size	Highest_degree	Degree_area	Publication Rate	Other Journals Published In	duration of Career(Years)
Institute of Pharmacology, Toxicology and Pharmacy, Ludwig-Maximilians-University, Munich, Germany	3		Biology	1.0	3	3
Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, United States of America	7		Biology	0.928571428571429	13	14
Department of Oral Biology and Pathology, School of Dental Medicine, Stony Brook University, Stony Brook, New York, United States of America	6		N/A	1.77777777777778	16	9
Oncology Research, Pfizer Worldwide Research and Development, San Diego, California, United States of America	4		Economics	0.818181818181818	9	11
Neurophysiology Laboratory, Department of Pharmacology and Experimental	7		N/A	2.66666666666667	40	15

Figure 9: *updated_bik_with_features.csv*

Figure 9 shows an example that the first five rows of additional features that would be add into the bik dataset and become the *updated_bik_with_features.csv*.

Second, We selected new features from each three datasets *Colleges_and_Universities.csv*, *NSF − metadata.xlsx* and *QSrank.txt* to combine the updated Bik dataset to become a new updated Bik dataset as Figure 10.



DOI	Title	Creator/Author	Journal Name	Sponsoring Org.	NAME	ADDRESS	STATE	COUNTRY	COUNTY	SOURCE	WEBSITE	UNI_RANKS	UNI_NAMES	UNI_AREAS
https://doi.org/10.1186/s41235-022-00388-4	Effects of experience on recognition of speech produced with a face mask	Crinnion, Anne Marie; Toscano, Joseph C.; Toscano, Cheyenne M.	Cognitive Research: Principles and Implications	National Science Foundation	BRYANT & STRATTON COLLEGE-AMHERST	3650 MILLERSPORT HIGHWAY	NY	USA	ERIE	https://nces.ed.gov/GLOBALLOCATOR/col_info_popup.asp?ID=189556	https://www.bryantstratton.edu	1	Massachusetts Institute of Technology (MIT)	North America
https://doi.org/10.1007/s40656-022-00497-9	Re-situations of scientific knowledge: a case study of a skirmish over clusters vs clines in human population genomics	Griesemer, James; Barragán, Carlos Andrés	History and Philosophy of the Life Sciences	National Science Foundation	CUNY SCHOOL OF LAW	2 COURT SQUARE	NY	USA	QUEENS	https://nces.ed.gov/GLOBALLOCATOR/col_info_popup.asp?ID=190682	www.law.cuny.edu	2	Stanford University	North America
https://doi.org/10.5281/zenodo.6540831	Ixodes Scapularis monitoring data compiled from 6 studies	Christie, Rowan; Stack, Kaitlin Whitney; Perrone, Julia; Bahlai, Christie		National Science Foundation	ROCKLAND COMMUNITY COLLEGE	145 COLLEGE RD	NY	USA	ROCKLAND	https://nces.ed.gov/GLOBALLOCATOR/col_info_popup.asp?ID=195058	www.sunyrockland.edu/	3	Harvard University	North America
https://doi.org/10.3389/fmolb.2021.618068	Role of Electrostatic Hotspots in the Selectivity of Complement Control Proteins Toward Human and Bovine Complement Inhibition	Narkhede, Yogesh B.; Gautam, Avneesh K.; Hsu, Rohaine V.; Rodriguez, Wilson; Zewde, Nehemiah T.; Harrison, Reed E.; Arantes, Pablo R.; Gaieb, Zied; Gorham, Ronald D.; Kieslich, Chris; Morikis, Dimitrios; Sahu, Arvind; Palermo, Giulia	Frontiers in Molecular Biosciences	National Science Foundation	SOUTHERN WESTCHESTER BOCES-PRACTICAL NURSING PROGRAM	450 MAMARONECK AVENUE	NY	USA	WESTCHESTER	https://nces.ed.gov/GLOBALLOCATOR/col_info_popup.asp?ID=193122	www.swboces.org	4	California Institute of Technology (Caltech)	North America
https://doi.org/1006549	G Hartung, C Vesel, R Morley**, A Alaraj, J Sled, D Kleinfeld, A Linninger, Simulations of blood as a suspension predicts a depth dependent hematocrit in the circulation throughout the cerebral cortex, PLoS Computational Biology, 14(11), e1006549, 2018. (**REU participant)	Grant Hartung, Claudia Vesel	PLOS computational biology	National Science Foundation	OHIO VALLEY COLLEGE OF TECHNOLOGY	15258 STATE ROUTE 170	OH	USA	COLUMBIANA	https://nces.ed.gov/GLOBALLOCATOR/col_info_popup.asp?ID=204884	ovct.edu	5	University of Oxford	Europe

Figure 10: New features added from New datasets

Figure 10 shows an example that the first five rows of new features that would be added into the bik dataset and become the final updated datasets. New features are: The DOI, Title, Creator/Author, Journal Name, and Sponsoring Organization from $NSF-metadata.xlsx$, NAME, ADDRESS, STATE, COUNTRY, COUNTY, SOURCE, WEBSITE from $Colleges\_and\_Universities.csv$ and UNI_RANKS, UNI_NAMES, and UNI_AREAS from $QSrank.txt$.

Third, we collected Bik dataset - papers with endpoint reached.tsv(the original Bik dataset), $updated\_bik\_with\_features.csv$(datasets combined after task4 and task5)and three new datasets with different MIME types: $Colleges\_and\_Universities.csv$, $NSF-metadata.xlsx$ and $QSrank.txt$ to calucate three different similarity metrics Jaccard, Cosine Distance, and Edit Similarity, As a result, we get **cosine similarity metrics** produced more accurate measurements(0.996) comparing to others from the table we mentioned in question 3.2.2.

4 Extra Credit

First, we selected more than three features from each new datasets, new features are: The DOI, Title, Creator/Author, Journal Name, and Sponsoring Organization from $NSF - metadata.xlsx$, NAME, ADDRESS, STATE, COUNTRY, COUNTY, SOURCE, WEBSITE from $Colleges\_and\_Universities.csv$ and UNI_RANKS, UNI_NAMES, and UNI_AREAS from $QSrank.txt$.

Second, we used BeautifulSoup python package to collect this data ($https : //www.crummy.com/software/BeautifulSoup/bs4/doc/$), which is the crawler of USC data science.