DSCI 550 HW 1
Instructor: Anna Farzindar
The report of "Bik et al paper" project
Team Member: Wangqi Chen, Erkang Chen, Weijia Fang, Xiang Li, Kaiyi Sun
Submission Date: July 27, 2023

**Abstract**

The dataset "Bik et al 2017 paper" we use as the original data in the project includes 200 biomedical research papers. Our projects can be summarized as four significant tasks: collecting extra data, extracting & incorporating features, applying Tika similarity, and interpreting data.

**Collecting Data and Extracting Features:**

In addition to the original research dataset, our team collects data such as the author's universities/institutions, research paper stats, institutions' locations, and UNDP's (United Nations Development Programme) HDI (Human Development Index ) index.

We utilize the Selenium package to extract data on universities/institutions from the ResearchGate website. This involves parsing the HTML using Xpath to locate the desired information. However, in some instances, the initial Xpath does not contain the required data, necessitating the addition of further Xpaths to retrieve this feature. The extracted data is saved in the text/csv MIME format. Previously, we attempted to use BeautifulSoup for extraction but encountered difficulties resulting in a switch to Selenium for web scraping. Although the code is less complex than Beautiful Soup, it takes approximately two hours to complete the web scraping process.

After getting the feature of "university/institution," we use this feature as the key to web crawl Wikipedia by accessing the Beautiful Soup library to find each location (countries/regions) corresponding to "universities/institutions." It is exported as "2_5.csv", an intermediate dataset waiting for other features to be added. It should be noted that this "2_5.csv" is not included in our final three datasets used to calculate similarities.

However, since there is still a little missing information obtained by purely web crawling, we manually clean the data for the column "Location_list" and create dataset "2_5_cleaned.csv" with an updated column "Location" to ensure a good quality dataset for generating additional predictors for our third dataset.

| SUM | First Authc | URL | Author cou | ResearchG | university/ | Location | Citations | Research Interest Score |
|---|---|---|---|---|---|---|---|---|
| 1 | Inka Regin | https://pul | 3 | https://wv | Ludwig-Ma | Germany | 14 | 9.4 |
| 1 | Jessica M. | https://pul | 7 | https://wv | University | United Sta | 17 | 7.8 |
| 1 | Sreedevi A | https://pul | 6 | https://wv | Stony Broc | United Sta | 104 | 53.7 |

Figure 1. 2_5_cleaned.csv.

The MIME type of data in the feature "Location" is "text/javascript." Besides, we collect the research paper stats by scrawling the "ResearchGate" website using the Selenium library. This includes bypassing some standard web scraping detectors activated by the high website access frequency. After obtaining the data, we

generated our dataset: “research_interest_citation.csv.” The MIME type of it is application/x-javascript. This dataset generated features “research interest” and “citation.”

Then, we download the UNDP’s HDI dataset from its official website, named as ‘HDR21-22_Composite_indices_complete_time_series.csv’.

This dataset calculates the annual HDI from 1990-2021 for 195 countries and regions. The HDI is a geometric mean of normalized indices for each of the following three dimensions: a long and healthy life, being knowledgeable and having a decent standard of living.

Back

Table 1. Human Development Index and its components

| | | | SDG3 | SDG4.3 | SDG4.4 | SDG8.5 | | |
|---|---|---|---|---|---|---|---|---|
| HDI rank | Country | Human Development Index (HDI) Value | Life expectancy at birth (years) | Expected years of schooling (years) | Mean years of schooling (years) | Gross national income (GNI) per capita (2017 PPP $) | GNI per capita rank minus HDI rank | HDI rank |
| | | 2021 | 2021 | 2021 [a] | 2021 [a] | 2021 | 2021 [b] | 2020 |
| | VERY HIGH HUMAN DEVELOPMENT | | | | | | | |
| 1 | Switzerland | 0.962 | 84.0 | 16.5 | 13.9 | 66,933 | 5 | 3 |
| 2 | Norway | 0.961 | 83.2 | 18.2 [c] | 13.0 | 64,660 | 6 | 1 |
| 3 | Iceland | 0.959 | 82.7 | 19.2 [c] | 13.8 | 55,782 | 11 | 2 |
| 4 | Hong Kong, China (SAR) | 0.952 | 85.5 [d] | 17.3 | 12.2 | 62,607 | 6 | 4 |
| 5 | Australia | 0.951 | 84.5 | 21.1 [c] | 12.7 | 49,238 | 18 | 5 |
| 6 | Denmark | 0.948 | 81.4 | 18.7 [c] | 13.0 | 60,365 | 6 | 5 |

Figure 2. Human Development Index and its components.

| iso3 | country | hdicode | region | hdi_rank_2 | hdi_1990 | hdi_1991 | hdi_1992 | hdi_1993 | hdi_1994 | hdi_1995 | hdi_1996 | hdi_1997 | hdi_1998 | hdi_1999 | hdi_2000 | hdi_2001 | hdi_2002 | hdi_2003 | hdi_2004 | hdi_20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AFG | Afghanistar | Low | SA | 180 | 0.273 | 0.279 | 0.287 | 0.297 | 0.292 | 0.31 | 0.319 | 0.323 | 0.324 | 0.332 | 0.335 | 0.337 | 0.362 | 0.376 | 0.392 | |
| AGO | Angola | Medium | SSA | 148 | | | | | | | | | | 0.364 | 0.375 | 0.386 | 0.403 | 0.42 | 0.433 | 0 |
| ALB | Albania | High | ECA | 67 | 0.647 | 0.629 | 0.614 | 0.617 | 0.624 | 0.634 | 0.645 | 0.642 | 0.657 | 0.669 | 0.677 | 0.684 | 0.689 | 0.696 | 0.7 | 0 |
| AND | Andorra | Very High | | 40 | | | | | | | | | | | 0.818 | 0.825 | 0.832 | 0.841 | 0.833 | 0 |
| ARE | United Arat | Very High | AS | 26 | 0.728 | 0.739 | 0.742 | 0.748 | 0.755 | 0.762 | 0.767 | 0.773 | 0.779 | 0.787 | 0.796 | 0.8 | 0.804 | 0.814 | 0.818 | 0 |

Figure 3. ‘HDR21-22_Composite_indices_complete_time_series.csv’ sample.

As the UNDP visualizes it as an interactive chart whose top level MIME type would be multipart since it encapsulates ‘image/gif” and ‘text/html’.

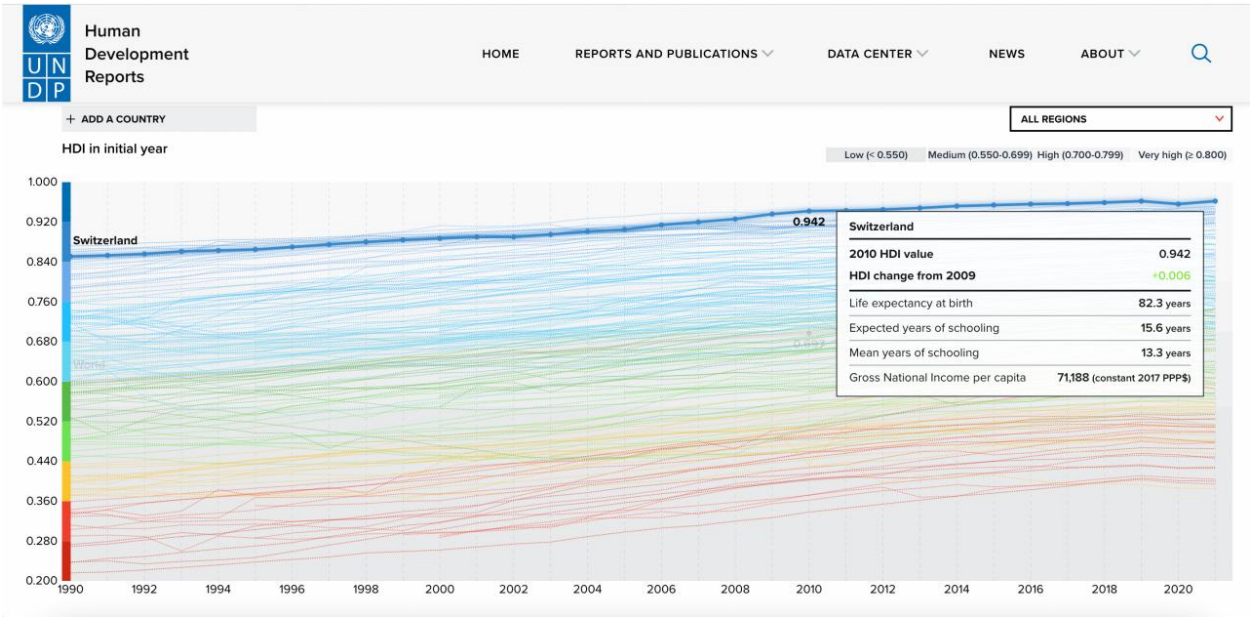



Figure 4. UNDP website screenshot.

This dataset could be downloaded from this official website as ‘HDR21-22_Composite_indices_complete_time_series.csv’

For this dataset, we want to identify the HDI index for each row. That is, we locate the HDI index in the above csv file, according to the "Location" column and "Year" column of the intermediate '2_5_cleaned.csv' file.

So firstly, we first reformat the country and region name in the '2_5_cleaned.csv' to make sure they conform to the UNDP format, so they could be located in the UNDP - HDI dataset. For instance, we reformatted 'Scotland, UK' as 'United Kingdom', and similarly 'Hong Kong' to 'Hong Kong, China (SAR)' and 'South Korea' to 'Korea (Republic of)'.

Then, we locate the HDI index for each row as is discussed above, and export our final dataset "third_dataset.csv".

**Information Interpretation and Tika Similarity:**

To process Tika Similarity, we prepare three datasets: "Bik dataset.csv," "second.csv," and "third_dataset.csv."

Compared to the original dataset, the "second.csv" adds the features "university/institution", "Research Interest Score" and "Citations (the number of citations). Based on the "second.csv," the "third_dataset.csv" updates extra features "Location" and "index."

| SUM | First Authc | URL | Author cou | ResearchG | university/institution |
|---|---|---|---|---|---|
| 1 | Inka Regin | https://pul | 3 | https://wv | Ludwig-Maximilians-University of Munich |
| 1 | Jessica M. | https://pul | 7 | https://wv | University of Colorado Boulder |

Figure 5. updated_bik.csv

| SUM | First Author | URL | Author count | ResearchG | university/ | Research I | Citations |
|---|---|---|---|---|---|---|---|
| 1 | Inka Regine We | https://pubmed.ı | 3 | https://wv | Ludwig Ma | missing | missing |
| 1 | Jessica M. Espa | https://pubmed.ı | 7 | https://wv | University | missing | missing |

Figure 6. Second.csv

| SUM | First Authc | URL | Author cou | ResearchG | university/ | Location | Citations | Research Int | index |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Inka Regin | https://pul | 3 | https://wv | Ludwig-Ma | Germany | 14 | 9.4 | 0.934 |
| 1 | Jessica M. | https://pul | 7 | https://wv | University | United States | 17 | 7.8 | 0.917 |
| 1 | Sreedevi A | https://pul | 6 | https://wv | Stony Broc | United States | 104 | 53.7 | 0.917 |

Figure 7. Third_dataset.csv

The institution/universities feature can be used to retrieve the location of these institutions which in turn will be used to correlate with the UNDP HDI index to answer our research questions.

**RQ1:** Does the impact of papers relate to media manipulation? Are papers with higher research interest scores more likely to manipulate content?

The Research Interest Score is a measured index generated by "unique ResearchGate members, recommendations on ResearchGate, and citations (excl. self-citations)" to evaluate the impact of the research within the scientific community. Because the number of citations is also an important indicator to measure the influence of research papers, we combined both features for its potential of interpreting the significance of the papers and the topics related to scientific fields. In this case, compared to the previous dataset, "second.csv" can answer questions: Does the impact of papers relate to media manipulation? Are papers with higher research interest scores more likely to manipulate content?

By correlation analysis, we set up "Research Interest Score" as the predictor to correlate multiple "manipulation" columns. The computed correlation coefficient corresponding to '0' '1' '2' '3' equals to '-0.0307' '-0.113785' '0.026687' '0.087873.' By this result, the research interest score has the strongest correlation to the '3' manipulation category. The statistics information about the category '3'is: mean= 51.48 and standard deviation= 126.26.

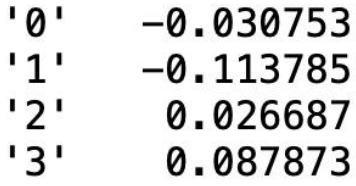
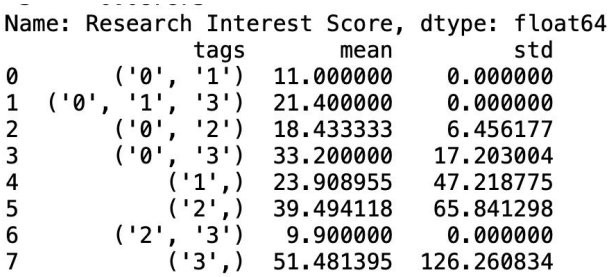



Figure 8. correlation coefficient between research interest score and media manipulation



Name: Research Interest Score, dtype: float64

| | tags | mean | std |
|---|---|---|---|
| 0 | ('0', '1') | 11.000000 | 0.000000 |
| 1 | ('0', '1', '3') | 21.400000 | 0.000000 |
| 2 | ('0', '2') | 18.433333 | 6.456177 |
| 3 | ('0', '3') | 33.200000 | 17.203004 |
| 4 | ('1',) | 23.908955 | 47.218775 |
| 5 | ('2',) | 39.494118 | 65.841298 |
| 6 | ('2', '3') | 9.900000 | 0.000000 |
| 7 | ('3',) | 51.481395 | 126.260834 |

Figure 9. Statistics Information

**RQ2**: Is there any correlation between the citation number and the HDI index?
When looking into this final dataset, we propose a second question: is there any correlation between the citation number and the HDI index?

To be more specific, the number of citations of a paper can reflect the quality and validity of the research and hence indicates the academic reputation. So, do countries with higher HDI index possess more papers with higher citation numbers? Or, to take one step further, do researches from countries with higher HDI index enjoy better academic reputation?

After running a correlation test between citation numbers under the "Citations" column and the index under the "index" column, we get the correlation coefficient of 0.0064186141875820675, which is too small to support our hypothesis.

**Apach Tika Visulization:**

***Edit distance****:*

After calculating the edit distance among the three datasets, we find that the second and third datasets are the most similar, whose similarity score is 0.814815. The similarity score between first and third datasets is 0.796296, that between the first and second datasets is 0.777778.


Figure 10. Edit distance circle packing visualization

***Cosine distance:***

After calculating the cosine distance among the three datasets, we find that the results are highly alike. The similarity between the second and third datasets is still the highest, reaching 1.0, indicating that they are entirely the same. The similarity between the first and second datasets, and between the first and third datasets remain the same and soar high as well, reaching 0.999967.

And we get a similar visualization in the combination cluster viz and circle packing viz.

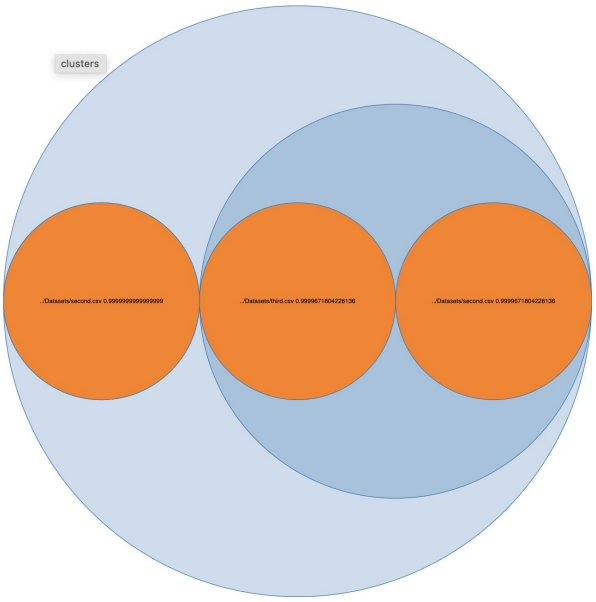



Figure 11. Cosine distance circle packing visualization

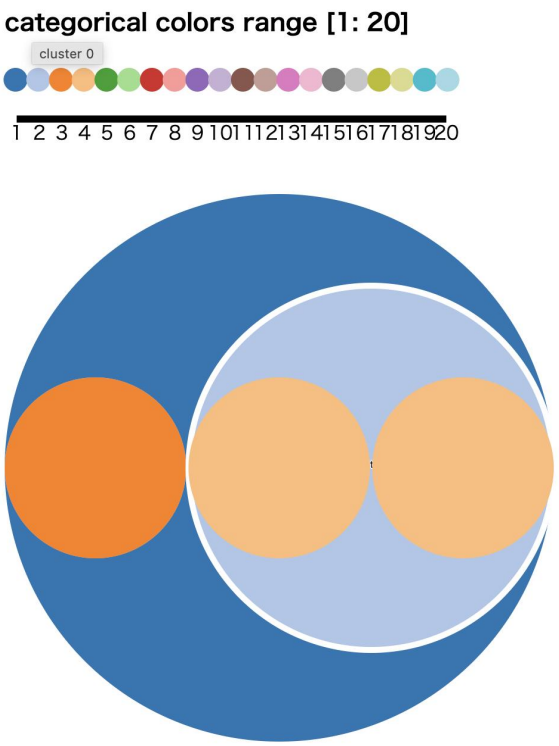



Figure 12. cluster viz and circle packing composed visualization

***Jaccard distance:***

After calculating the jaccard distance among the three datasets, we find that the results are identical. The results are 0.538462.

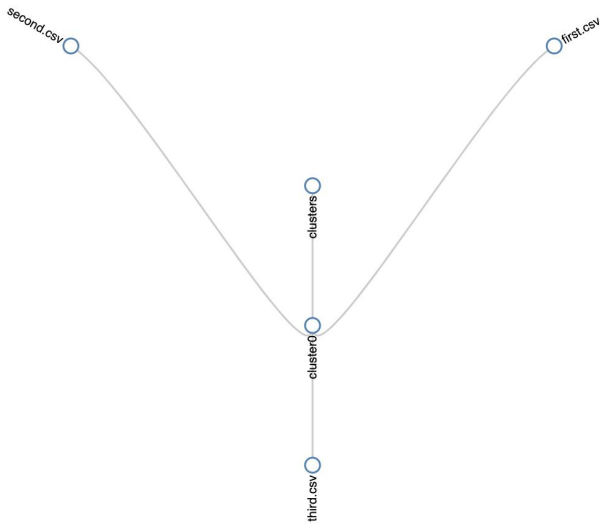



Figure 13. Jaccard distance cluster visualization

**Similarity Measurement Comparison:**

We believe that the cosine similarity is the most accurate one. The cosine similarity measures the 214 rows of our final document, so it is reasonable that the similarity scores are all very high. And since the modification on the second dataset is smaller than that on the first one, we can understand that the similarity between the second and third datasets is bigger than the other two scores.

The edit distance calculation operates on sequences of characters and measures the minimum number of operations required to transform one string into another. So, it does not make that much sense that the similarity between the first and the third datasets is greater than that of the first and second datasets. So are the results of the jaccard similarity, which remain all the same.

**Thoughts on Apache Tika**

Apache Tika is a power tool that supports a wide range of file formats, such as csv and json. However, in this project, when calculating the similarity between files, Sklearn is faster and easier to implement. For instance, when using Apache Tika, we have to jump back and forth between files to find the specific .py file for certain similarity measures.

**Summary**

In this project,we incorporate multiple datasets to enrich our analysis, including university information, citation counts, and research interests index from ResearchGate, country data via Wikipedia, and Human Development Index from UNDP dataset. Additionally, we conduct a comprehensive analysis of academic papers using various distance methods, including Jaccard Similarity, Edit Distance, and Cosine Similarity, to assess their metadata similarities using Tika-similarity package without delving into their content. We successfully visualized the clustering results using hierarchical circle packing, providing insights into the relationships among the papers. The correlation coefficient between the media manipulation and research interest shed light on unexplored aspects of media forensics data and enabled us to explore how the influence of the research paper interact with media manipulation.