Group: Seize the Data
Course: DSCI 550 – Spring 2022
Assignment 1: Analysis of Media and Semantic Forensics in Scientific Literature

Introduction

Working with the "Bik et al" dataset presented various challenges in each step leading up to our analysis. These obstacles began with the need to clean features in the dataset that was provided to us, such as author names and special characters in DOIs. Additionally, there were issues during the web scraping component, where opening too many websites would result in a security CAPTCHA error. Furthermore, finding additional features from datasets, of different MIME types, that would best help us describe the provided dataset proved to be more difficult than anticipated. Finally, Apache Tika was a completely new toolkit that presented a steep learning curve and required us to adjust our final dataset to be able to run the clustering analysis. In the end we were able to overcome most of these obstacles and do the best possible clustering analysis.

Bik et al Dataset – Adding Author Features

The first step in our program uses the names of the authors and DOIs provided in the "Bik et al" dataset. We created two dictionaries, one with the DOIs as keys and all authors as the value. The second dictionary also used DOIs as keys but only the first author as the value.

The first five features added to our dataset were collected from ResearchGate and Google Scholar using BeautifulSoup to scrape the HTML code of each page. Our program first searched the DOI on ResearchGate, then opened the available author links, which usually allowed for about 4 authors. If there was a "show more" button, the links for those authors were encrypted and unavailable. It was difficult for our program to scrape ResearchGate due to the encryption, which prompted us to also scrape Google Scholar to retrieve more information. The program searched the remaining authors on Google Scholar and selected the first author page from each search, which in most cases was the correct link for the author.

The features collected from both ResearchGate and Google Scholar included Publication Rate, Other Journals, and Lab Size for all authors of each article. Degree Level and Duration of Career were obtained for only the first author of each article. All features were collected from both sites, except for Degree Level which was only found on ResearchGate. Publication Rate was defined as the number of articles published by an author in their entire career. Other Journals consisted of a list of each unique journal that an author was published in at some point in their career. Lab Size was the number of "lab members" as listed in ResearchGate for each author with profiles. However, if the author did not have a profile or was not found on ResearchGate, the number of co-authors was used from Google Scholar to define this variable. The Degree Level included Bachelors, Masters, PhD and MD degrees. This was difficult to obtain due to inconsistent labeling conventions, for example, "PhD" versus "doctor of philosophy." Career Duration was determined as the number of years from the author's first published article until 2022. Not all data was available for all authors: some information was difficult to obtain due to security lockouts and some we were unable to obtain because we could not find the authors in our search. We would have had much more success if we were able to access all author links through ResearchGate which contained the most information in a single place. The obstacle that was the most difficult to overcome for these two sites were the security lockouts. We were unable to run this code on all our data at once because we would get locked out and the final values of the dataset were entered as missing numbers, which were represented by "-1". We had to separate the tsv into 6 smaller sections, run the program on each of those sections, use a VPN, and use pandas to combine the data again afterwards.

Our next step was to find affiliation for the first author of each article. Most articles had this information available, but after closer inspection of each DOI, it was clear that different types of DOIs displayed this information uniquely. Some had it directly under the author's name while others displayed it in a separate tab with many dropdowns; forcing us to create a program that accounted for each of these differences. For this portion we utilized the dictionary containing the first author and DOI to search each DOI webpage, use BeautifulSoup to scrape the HTML, and pull out the affiliated university of each first author. Again, we encountered issues with scraping some of the DOI websites, specifically those with "ijc,"

"bcr," "lungcan," "jaut," and "cyto." These pages either did not give us permission to scrape them or returned the JavaScript of the page rather than the HTML, which did not yield the information we needed. This led to some gaps in our data regarding the affiliations of the first author.

The final author attribute needed was Degree Area, which proved very difficult after we conducted a Google Search on many authors to find a common place where this information may be stored. This information was difficult to find for many reasons. For example, some authors did not have an internet presence aside from the DOI article page. In other cases, authors had very common names and it was impossible for our program to easily determine which was the correct profile. Additionally, even if an author had an Internet presence this information was not always shared or there was, again, inconsistency in labeling. Lastly, a small group of authors were present on a specific site that other groups of authors were not on (Linkedin, ResearchGate, FrontierLoop) which would have required unique scraping programs for each site. We attempted several approaches including scraping Google search results and searching the authors on university websites. However, neither of these approaches were a one-size-fits-all solution and going through each author was going to require its own unique program. We discovered that when pulling the affiliations, the code often pulled the department the author worked in as well. These department names included areas of study. We then defined Degree Area to be the department that the author was working in as we inferred that the author that likely completed their degree in an area similar to what they were working in.We parsed through each affiliation and pulled out all of the scientific areas of study, such as words ending in "ology," "stry," "ics," etc. This approach worked well but was not fruitful for affiliations that were in languages other than English.

U.S. Census Bureau International Database

The first dataset of additional attributes is provided by the U.S. Census Bureau; containing population data about countries around the world, including annual growth rate percentage, total fertility rate, mortality rate, population density, etc. The MIME type of the first dataset is text/csv.

Prior to selecting three features from this downloaded dataset, country names were parsed from the existing dataset in order to join them as many records contained country information at the very end of the "Affiliation University" column. Three features were then added based on the country where the first author of each paper's affiliated university was located: annual growth rate percentage, population density, and life expectancy. No computation was required to obtain these features, although some extra steps had to be taken in finding the country for each paper.

We know that the US and China are two of the top contributors of papers with problematic images. The features added by this dataset allow us to dive deeper into country data and to make queries such as: do authors from countries with a higher life expectancy publish more papers with problematic images? Or maybe knowing that they will lead a long and healthy life, these authors take their time and publish papers with fewer problematic images? Developing nations tend to have higher population growth rates, so do authors from countries with a high growth rate produce more papers with problematic images? China has a relatively high population density (151 people per square km in 2021) while the US's is much lower (36.9 people per square km); so, we can ask if in general, authors from countries with a high population density produce more papers with problematic images.

Environmental, Social, and Governance Data

The second dataset comes from The World Bank and contains annual data that measure how well countries perform environmentally, socially, and how they manage resources. It was saved as a flat application/xlsx file which is a MIME type associated with MS Excel.

With the country name feature from the previous dataset, we were able to search and filter this new dataset by the year articles were published and the country they were associated with. The three features we selected from the new dataset were "Control of Corruption: Estimate," "Government expenditure on education, total (% of government expenditure)," and "Scientific and technical journal articles." We felt it would be very telling if countries who did not control corruption very well could be culprits in the issue of image duplication. Perhaps countries with lower education expenditure provide less resources for academics

so that they might resort to duplicating images. Also, we wanted to see if having a higher number of scientific journals published can be problematic since they will have more reference material to plagiarize.

Air Quality Data

The third dataset was retrieved from World Population Review and consists of air quality information categorized by country. The dataset was obtained via download, with a MIME type of application/JSON. The three added features from the dataset include air quality level, country population, and the percentage of the air quality level compared to the worst air quality level recorded. The third feature was calculated as follows: % Difference = Max Air Quality Level / Current Quality Level x 100
The final step was merging the air quality and ESG datasets using the 'Country' attribute previously described as a relational key, thus resulting in our group's overall dataset output.

In addition to the other two supporting datasets, our group considered how air quality might have contributed to the image duplication issue. Research conducted by IZA Labor Market in 2017 found a direct correlation between air pollution levels and scholastic achievements. During the study, researchers had noticed how prolonged exposure to pollutants resulted in lower oxygen levels, eye irritation, headaches, dizziness, and fatigue-like symptoms. There could be a possibility that air quality levels near the universities might have played a role in either intentional or unintentional research errors and image duplications contributed by contributors experiencing health issues during the time of submission.

Exploring Datasets with Apache Tika

Compare Jaccard similarity, edit-distance and cosine similarity

Since the dataset is small (214 instances), any correlations derived from analysis of the dataset are not likely to be significant, which poses a challenge in assessing the performance of the different similarity metrics. Furthermore, we are calculating the similarity score of papers using combinations of 2 papers at a time while trying to find distinct clusters of more than 2 papers with a high similarity score based on their metadata, and then trying to make conclusions about attributes of the papers which were not run through the similarity algorithms at all (ie. the extracted attributes + attributes from newly joined datasets), all of which further complicates the task of comparing the performance of different similarity metrics.

In general Edit-Distance Similarity is more likely to yield accurate results in cases of string comparison, which is the case of this dataset where we are using Tika to generate the metadata of different papers and calculating the similarity score papers based on metadata. This is in contrast to Cosine Similarity which is more accurate when comparing number vectors.

Edit Distance took the longest to run (which makes it less practical to use in scenarios with larger datasets), and resulted in fewer distinct clusters with several papers with a high similarity score than Cosine and Jaccard, and many smaller clusters of papers with lower similarity scores. The maximum similarity score observed using Edit Distance was 1, using Cosine it was 1 and using Jaccard it was the lowest at 0.952 (Generally Jaccard yielded lower similarity Scores than Edit Distance and Cosine).

All 3 methods, Edit Distance, Cosine Similarity and Jaccard gave similar results in terms of which papers have a high similarity score when comparing their metadata. (Appendix A, B and C)

Apache Tika is a powerful tool that was able to extract many of the attributes needed from the metadata of the HTML files and was easy to use with Python due to the Tika Python library.

How the resultant clusters generated highlight the extracted features

Methodology

1. The Tika Similarity Project's Cosine, Jaccard and Edit Distance similarity code were each run using the html files of all papers in the dataset for which an html file was extracted
2. For each similarity metric (Cosine, Jaccard and Edit Distance) the papers' similarity score was generated and visualized.
3. For each similarity metric, the generated scores were used to identify the main cluster of papers which have a high similarity score.

4. Additional synthesized metrics were generated as follows:

- **The Mean, Median and Mode** were calculated for each attribute in each cluster made up of **ONLY the papers with a high similarity score**, for each similarity metric.
- **The Mean, Median and Mode** were calculated for each attribute in the **entire (original) dataset** for all papers for which the html file is available.
- **The Publication Rate** of the 1st Author=1st Author Publications/Duration of Career

5. The Mean, Median and Mode of each attribute in each high similarity cluster from each similarity metric were compared to the Mean, Median and Mode of each attribute in the original dataset.
6. Wherever all 3 similarity metrics (Cosine, Jaccard and Edit Value) agreed that there is significant variance between the Mean, Median and Mode of the high similarity cluster and the Mean, Median and Mode of the original dataset, a hypothesis was formulated.

Clusters revealed

1. Highlights from Extracted Features

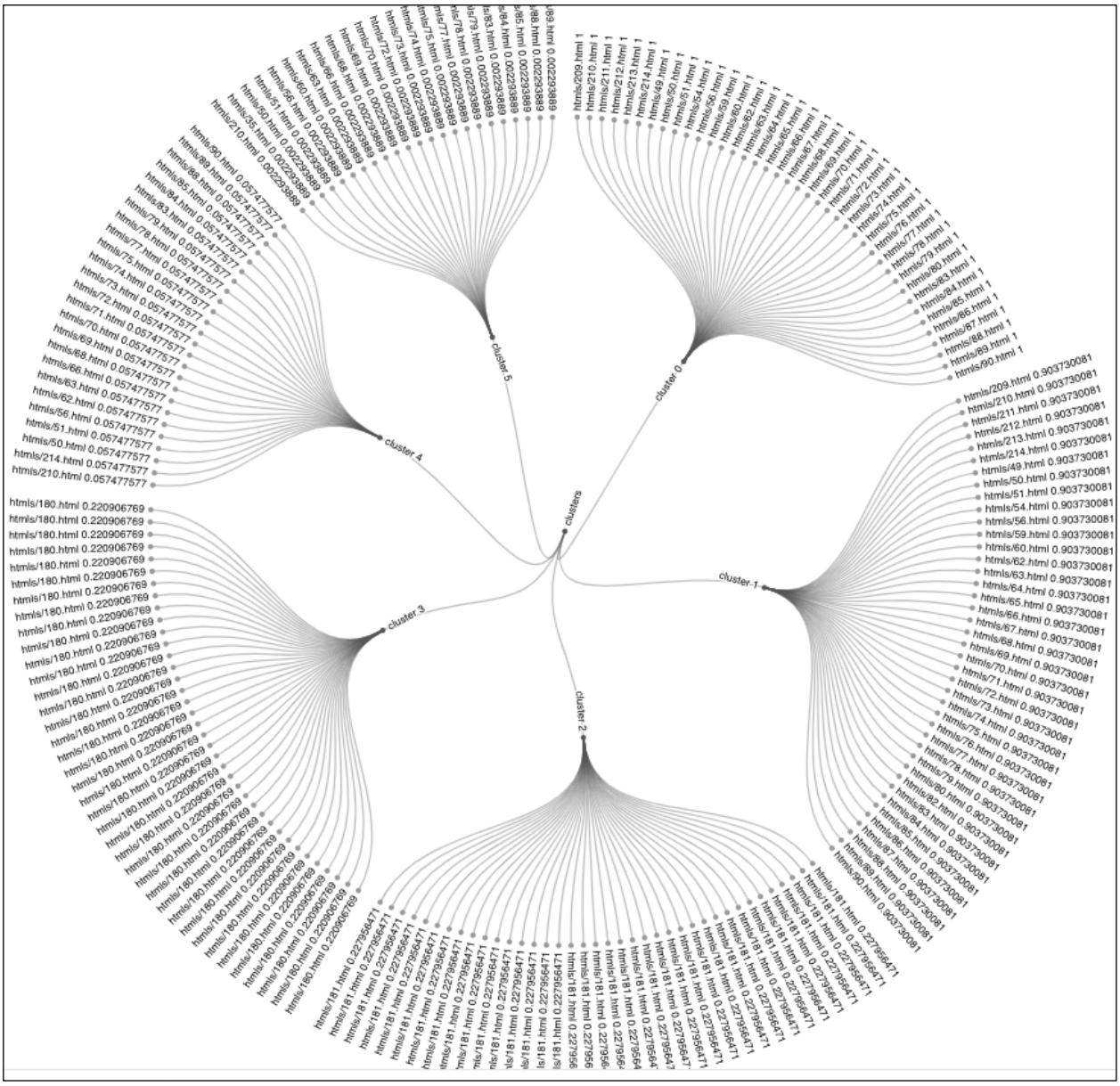
- Using an unweighted "Ensemble" voting approach across the similarity metrics (simply by examining the results), the Publication Rate of the 1st Author is consistently lower in the high similarity clusters than in the full dataset (all papers in sample), while the duration of career is mostly higher. The exception is the Mean and Mode Publication Rate using Edit Distance which are slightly higher than in the full dataset, and the Mode of Career Duration using Edit Distance which is slightly lower than its equivalent in the full dataset.
- **Hypothesis:** It is possible that authors with 11-17 years of experience and a lower-than-average Publication Rate are trying to get tenure and want to increase their Publication Rate, therefore they are not thoroughly reviewing the data included in their papers before submitting them for review and publication.
- No correlation could be established between lab size and likelihood of data falsification as lab size range remained between 5 and 13 members in both the high similarity clusters and the full dataset, across all similarity metrics. (Appendix D)

2. Highlights from New -Joined Datasets:

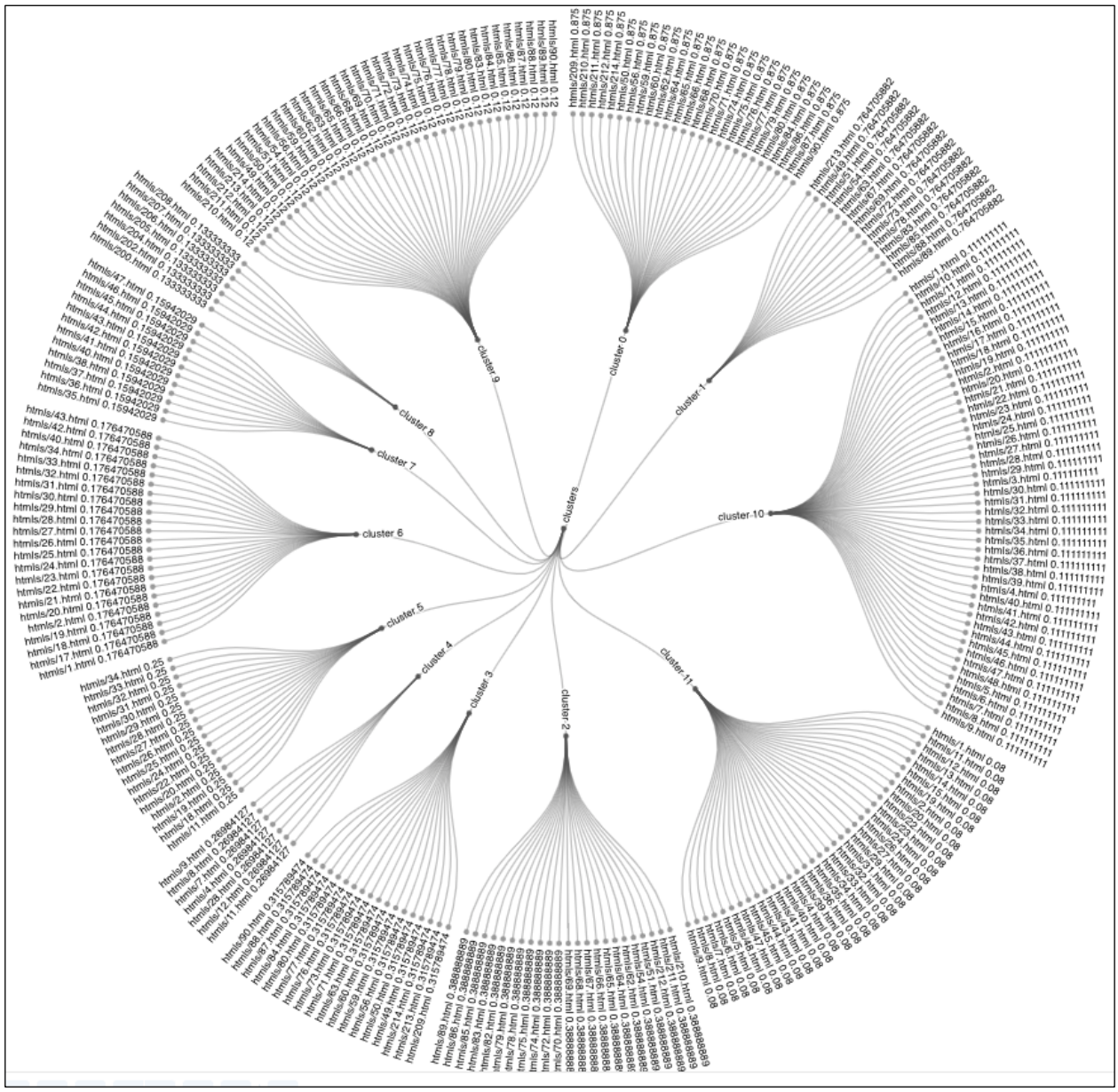
- **Hypothesis:** It is possible that there is a correlation between higher-than-average Annual Growth and lower than Average Population Density of the Author's country of origin, and increased likelihood that the author intentionally or unintentionally falsifies data. This could potentially be due to the Author coming from an ambitious and competitive culture and being pressured to increase his/her publication rate therefore not committing to a thorough review of data before publication.
- No correlation could be established between Corruption and Air Quality Attributes of the first Author's country and likelihood to falsify data. (Appendix E)
- Since the additional datasets are country specific, some of the "unintended consequences" related to media forensics data suggested by additional datasets are that if a correlation is established between a specific nationality/country and a negative outcome (for example, falsifying data), this could fuel pre-existing racial and cultural biases between people and promote an anti-diverse culture in the scientific community.

Appendix

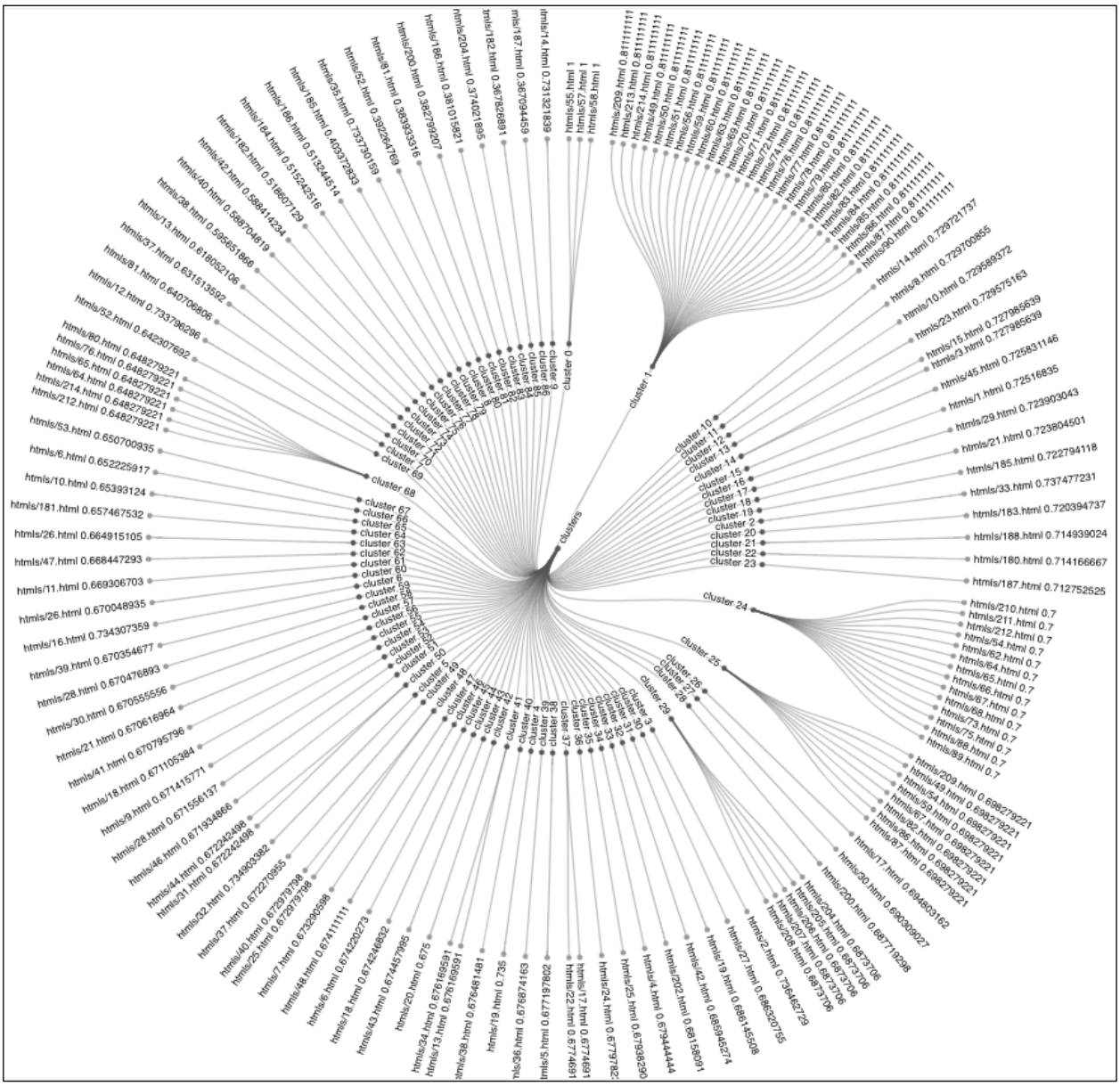
Appendix A: Cosine Similarity



Appendix B: Jaccard Similarity



Appendix C: Edit Distance Similarity



Appendix D: Comparison of Similar Papers vs. All Papers for Extracted Features

Measure	Duration of Career	Lab size	1st Author Publications	Publication Rate
Metrics of All Papers				
Mean	14.73	10.06	47.70	3.24
Median	12.00	10.50	24.00	2.00
Mode	12.00	5.00	100.00	8.33
Metrics of Similar Papers based on Cosine				
Mean	16.40	10.00	43.78	2.67
Median	13.50	11.00	23.00	1.70
Mode	11.00	13.00	13.00	1.18
Metrics of Similar Papers based on Jaccard				
Mean	17.75	9.82	53.25	3.00
Median	15.50	11.00	26.00	1.68
Mode	#N/A	13.00	9.00	#N/A
Metrics of Similar Papers based on Edit Distance				
Mean	16.46	8.63	53.48	3.25
Median	14.00	7.00	25.00	1.79
Mode	11.00	13.00	100.00	9.09

Appendix E: Comparison of Similar Papers vs. All Papers for Extracted Features

Measure	Annual Growth	Population Density	Life Expectancy	Control of Corruption	Education Expenditure	Scientific Journal	AirQuality	Country Population	Compared to Worst AQ
Metrics of all papers									
Mean	0.63	295.87	78.62	0.77	12.80	232,823.07	21.49	568,932.98	(0.74)
Median	0.67	146.10	80.61	1.29	13.32	108,995.98	11.36	334,805.27	(0.86)
Mode	0.73	34.80	76.17	1.38	13.32	433,192.28	9.04	334,805.27	(0.89)
Metrics of similar papers based on Cosine									
Mean	0.88	157.29	79.30	1.12	12.86	200,598.54	16.79	320,157.04	(0.80)
Median	0.78	63.80	81.48	1.31	13.70	97,905.28	9.39	201,651.59	(0.89)
Mode	0.69	34.50	#N/A	1.31	#N/A	429,570.05	9.04	334,805.27	(0.89)
Metrics of similar papers based on Jaccard: NA as highest similarity score was 0.344									
Mean	0.16	234.50	80.58	0.46	10.72	178,620.53	22.52	544,773.00	(0.73)
Median	0.16	208.90	82.26	0.08	10.72	108,995.98	17.09	125,584.84	(0.79)
Mode	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Metrics of similar papers based on Edit Distance									
Mean	0.89	217.87	77.94	0.07	13.72	243,314.54	33.02	968,621.90	(0.60)
Median	0.46	146.10	76.17	(0.36)	13.32	359,274.07	39.12	1,448,471.40	(0.53)
Mode	0.46	146.10	76.17	(0.36)	13.32	359,274.07	39.12	1,448,471.40	(0.53)

Resources

Homeland Infrastructure Foundation-Level Data. 28 June 2019. Updated March 1, 2022. Retrieved from: https://hifld-geoplatform.opendata.arcgis.com/datasets/geoplatform::colleges-and-universities-campuses/about

U.S. Census Bureau. 2022. *International Database*. Retrieved from: https://www.census.gov/data-tools/demo/idb/#/table?COUNTRY_YEAR=2022&COUNTRY_YR_ANIM=2022&menu=tableViz&POP_YEARS=2022&TABLE_YEARS=1990,1991,1992,1993,1994,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2022&TABLE_RANGE=1990,2014&TABLE_USE_RANGE=Y&TABLE_USE_YEARS=N&TABLE_STEP=1&quickReports=CUSTOM&CUSTOM_COLS=POP,GR,RNI,POP_DENS,TFR,CBR,E0,IMR,CDR,NMR

IZA World Labor. 23 August 2017. *New Report: Exposure to air pollution adversely affects educational outcomes.* Retrieved from: https://wol.iza.org/press-releases/air-pollution-educational-achievements-and-human-capital-formation

The World Bank. 26 July 2019. *Environmental, Social and Governance Data*. Retrieved from: https://datacatalog.worldbank.org/search/dataset/0037651

World Population Review. N.A. *Most Polluted Countries 2022.* Retrieved from: https://worldpopulationreview.com/country-rankings/most-polluted-countries