INF 550 Homework 1: Analysis of Media and Semantic Forensics in Scientific Literature

Team 03: Matthew Lee, Carlin Cherry, David May

Introduction

In this project, we analyzed features that might make an author more likely to manipulate media in their publications. We developed R scripts to clean Bik et. al.'s dataset for analysis and to add additional features, which we obtained through web scraping, APIs, and manual research. Then, using the Tika-Similarity library in Python and the `cluster` library in R, we explored similarities between the papers.

Featurization

We used R and its `RSelenium` package to scrape Microsoft Academic and LinkedIn for additional information on the **first authors** of the papers. To augment Bik's dataset further, we joined in data from three areas: **Google Trends**, **InCites**, and **US News and World Report**.

First Author Information

Data from Microsoft Academic and LinkedIn provided features requested in the assignment. Microsoft Academic added even more, including: `citations`, the number of times others have cited any of an author's papers, `top_topics`, the fields of study associated with a given author; `top_authors`, other authors associated with a given author.

Design: Microsoft Academic features like `top_xx` were long, delimited strings, so we also added numeric versions of them to make these features more usable. For some, we took the string, separated it into a list by its delimiter, and counted the number of non-NA elements. Other times we created a series of boolean variables that each confirmed whether a specific term was present, like "biology" or "medicine." Microsoft Academic's additional features helped us approximate lab size. We originally planned to scrape university websites for faculty and student numbers, but dealing with the different website structures seemed inefficient with our crawler. Instead, the count of related authors from Microsoft Academic served as a substitute for the scale of an author's lab–an author's lab peers would reasonably be among the names related to that author.

LinkedIn results for degree level/subject varied widely due to free-input text. We also had to use regular expressions to sort the values into standard educational levels. There also were many NA values, so we attempted imputation by converting educational level to a number and filling in NA with the average of the non-NA values. However, this would be assuming all unknown cases are around a bachelor's level, which might be too strong of an assumption. We left these features in the updated dataset but were wary of them when considering features to include in similarity measurements later.

Observations: While each of the papers in this dataset involve biology in some way, only 174/214 mention "biology" in the first three related topics. There might be an important distinction between authors who focus on a sub-discipline versus those who associate with the general field. Looking at degree levels, it was not surprising to find mostly PhD (59 of the 96 with LinkedIn profiles), but the missingness of the LinkedIn data limits its value, with only 44.8% of records being non-NA.

Google Trends

Data available through Google Trends' API provided the following features: `web_interest`, `images_interest`, and `youtube_interest`. Each is a score of search popularity of different content types accessible through Google (web pages, images, and Youtube videos, respectively).

Design: We used the R package `gtrendsR` to extract data from the Google Trends API. The `gtrendsR::gtrends` function takes in user inputs (like keywords, date range, and content types to consider) and outputs a time series for each keyword. For each feature, the query used 4 keywords: 3 taken from the row's `top_topics` (topics relating to the first author from their current Microsoft Academic profile), plus "biology". Google Trends assigns the highest point of interest over time to 100, with all other values being relative to that peak, so always including "biology" provides a baseline of comparison. We averaged across keywords to get a single time series, then filtered down to the month/year of publication to get a single value. While the output is technically numeric text, obtaining this data touches on a variety of MIME types: application/json to encode information packets; application/http to send requests to the API; text/* and image/* data that the Google Trends API sifts through to measure interest.

Observations: These features investigate patterns in an author's research topics. Perhaps an author writes specifically on a topic that rarely sees search results (e.g. matrix metalloproteinase). This might suggest that authors do not collaborate on this topic–or the topic is so obscure that it would be difficult to research. In cases like these, perhaps duplications are accidental, simply due to the lack of search/discussion around the topic. On another note, perhaps the paper was part of a popularized movement around a topic. If something occurred that would create a spike in search interest (say, a virus outbreak or a large grant), then researchers might feel pressure to deliver a high velocity solution at the expense of some veracity.

Interestingly, the 2 highest web interest records have several similarities in their other features. Both published around the same time (2004 ~ 2005) and both relate to "chemistry" and "biochemistry." Moreover, the authors of these records affiliate with the same university (American University of Beirut), and one of them has the other listed in her related authors on Microsoft Academic. This might warrant investigation on the field of biochemistry at this time (were there any initiatives like DARPA's XData for data science?), or on this particular university and its biochemistry department (how does it rank, who funds research?).

InCites

Data from the Clarivate Analytics InCites Journal Citation Reports provided the following features: total citations of that particular journal by year of publication; an index called Impact Factor that is used as a proxy for the relative importance of a journal within its field, with a higher score representing a more prestigious and/or important journal; and a similar rating called Eigenfactor of the total importance of the journal encompassing additional factors to the pure number of citations it received in a given year.

Design: We first attempted to web scrape this data, using the `keyring` R package to pass USC login credentials to a crawler. However, we ultimately were concerned about the security of doing so with our real credentials. After attempts at a workaround or finding other non-password-protected sites with this information, we ended up manually searching the database for each journal. While we saved the output as a text/* MIME type, trying to obtain this data using a crawler utilized a variety of MIME types.

Observations: These features added to Bik's dataset describe the journal of publication's importance. We hypothesized that the importance of a journal as described by these features in combination with other indicators might help differentiate the intentional image duplications from the non-intentional. Bik. et al estimate that about half of the images they examined were intentionally and wrongfully manipulated prior to journal submission, and of that subset, we predicted that an author might be more likely to intentionally falsify their results when the stakes are higher. On the other hand, the data might reveal that more prestigious journals have a more thorough review process and thus are less likely to accept duplicated images, which would translate to journals with high impact factor having fewer of these types of images. Ultimately, because the dataset we examined is a curated list of images we already know to be duplications, Impact Factor data features might prove to be more valuable in a mixed dataset of known and unknown images so that it can be used as a predictive indicator rather than as an explanatory one.

US News and World Report

Data from US News and World Report and other research provided the following features: `academic_reputation`, an index based on research reputation, publications, citations, and international collaboration; and `affiliation_funding`, whether a university is publicly or privately funded.

Design: We faced a larger than anticipated challenge finding comprehensive data on industry funding. There is no governing body for how research should be funded; in fact, we discovered specific industry funding in academic research is often kept secret. In order to sidestep this roadblock, we pursued data on yearly research budgets by university. This also proved to be a challenge since comprehensive yearly research budgets for each of the universities are housed in entirely different places on the web; however, we were at least able to determine the public/private funding type of universities from manual searches. In the search to illuminate the reputability of the research from our list of academic institutions, we discovered US News and World Report's global university ranking system. We scraped academic reputation scores for each university in the dataset from this site.

Observations: If we had information on each paper's top sources of funding, we could determine if there is any financial link between these papers or if certain funding streams are associated with problematic image duplications. We could also examine research budgets; perhaps institutions with lower research budgets during the time when the papers were published could have felt additional pressure to inappropriately duplicate images in their research to save money/resources.

Analysis

For our analysis, we used the cosine, Jaccard, and edit-value similarities from Tika-Similarity. We also experimented with the Gower distance in R (a composite index that supports numeric and categorical data). We used Partitioning Around Medoids (PAM) in R to make clusters, and t-SNE plots to visualize them.

Similarity Measures

To improve accuracy, we only used a subset of our features in calculating similarity scores. For Tika-Similarity, we excluded identification columns like `doi` (no insights from knowing every row has a distinct ID) and features with uniform values like `reported` (no variation to draw insights from). For R, we made similar exclusions, plus a few more due to nuances with Gower distance. Gower distance performs one-hot encoding on categorical variables, but we already had done something similar by making numeric features that capture the information of the categorical ones. We opted to use our numeric features to narrow down on specific terms rather than having Gower distance explode the dimensions of our dataset. We also excluded features with high missingness, like `month` and `highest_degree`.

With Tika-Similarity, we obtained the cosine, Jaccard, and edit-value similarity scores between all pairs of rows. Some of the .py scripts hinted at support for file input–iterating through all pairs of rows in a .csv instead of through all files in a directory–so we built on that code to get the three desired similarity measures. With R, we used the `cluster::daisy` function to obtain a dissimilarity matrix with Gower distance. We imported the Tika-Similarity outputs and copied them into dissimilarity matrices for R to cluster on them. We also copied the Gower distance matrix into similarity scores for easier comparison to the other measures.

The 4 similarity measures differed slightly, but they agreed on the straightforward cases; the pairs with the highest similarity described papers with the same author. This made sense, as many of the features that went into our analysis describe the author. Papers that shared similar topics also seemed to group together, perhaps due to the proliferation of one-hot encoding variables that we added when breaking down top topics. Revisiting that pair from the Google Trends observations that felt "close" across many features, the distance measures disagree. With Jaccard, the similarity is rather low (0.29); with edit-value (0.61) and Gower (0.69), it is higher but still somewhat low; with cosine (0.99), it is appropriately high.

From the aforementioned observations and other checks, we have some notion that cosine similarity performed the best with our data, and Jaccard lagged behind. It is possible that the cosine distance excelled due to the many numeric variables that we created, and perhaps Jaccard similarity suffered because it is more suited for comparing across sets of feature names rather than arrays of values. Now that we have a grasp of the similarity between our records and a slight favor for cosine similarity, we will look at how those measures fare in clustering the data.

Clusters

Surprisingly, the clustering visualizations in Tika-Similarity placed all observations in a single cluster. Upon further investigation (and helpful office hours), we found that the clustering looks for parity across *feature sets* rather than *values*. Since all rows have the same column names, they all seem identical when clustered. To address our clustering needs, we turned to R.

We used R's `cluster::pam` function to perform clustering with partitioning around medoids, which requires a predetermined number of clusters k. To find this, we looked at the silhouette criterion for the various distance measures. The silhouette plots were lackluster, with most distance measures returning average silhouettes $<$ 0.25 for all tested k values (usually hinting that the structure is superficial and requires additional analysis). The exception was cosine distance, which returned reasonable values in the 0.5 ~ 0.6 range. The silhouette plot suggested that k = 2 was the optimal number of clusters. However, the silhouette for k = 4 was not much worse, and a 4 cluster split seemed more reasonable for further analysis.

Proceeding with cosine distance and PAM clustering with k = 4 gave us the cluster visualization below. Each dot represents a paper from the dataset, with the larger dots representing the cluster medoids.

The clusters seem to divide up the data nicely. To explore the clusters, we first looked broadly at summary statistics. On average, cluster 1 (red) has fewer publications, fewer citations, shorter career durations, and smaller variety in associated authors/journals. In contrast, cluster 2 (green) is the hightest in all those areas. Clusters 3 (blue) and 4 (purple) are in between. We also looked at the distribution of categorical variables across the clusters, but only found statistical significance with 2 of them : whether an author's research topics involve "biology" and whether they involve "medicine." There is some notion that our clusters have distributed the biology/medicine researchers in a meaningful way; cluster 2 and 4 have higher concentrations of biology/medicine; cluster 1 has a lower percent in biology.

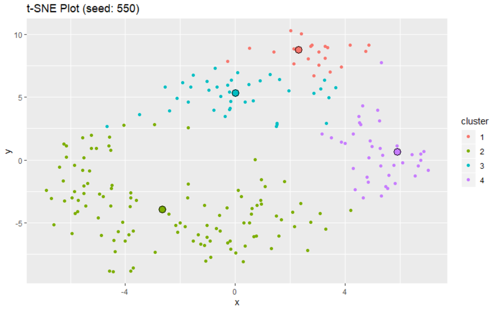

Figure 1: Clustering results

Taking that further, cluster 1 seems to be researchers with smaller profiles in the field of biology, perhaps due to inexperience or a focus on a specific sub-discipline. Larger names who have accumulated publications and citations over time (to the point where their names are associated with the general field) land in clusters 2 and 4, with 4 being slightly more concentraed in medicine. Interestingly, the t-SNE plot corroborates this, with clusters 2 and 4 being apart from 1, and cluster 3 being somewhere in between (however, it should be mentioned that t-SNE will bring similar objects closer, but distance on a t-SNE plot is not necessarily measurable).

It would have been more interesting if the categorical variables for the various media manipulations had been statistically significant–this would hint that records with certain manipulations were not randomly distributed across the clusters.

Conclusion

We found a few clusters with reasonable structure within this Bik et. al. dataset. However, we question the value of our results in regards to identifying patterns in media manipulations, as our clusters did not differentiate the various manipulations in a meaningful way. Future analysis might seek out better features or extend to a general dataset of papers where we seek to understand characteristics that might indicate media manipulation versus a "clean" journal submission. This being said, through this assignment we developed a conceptual and technical understanding of R, web scraping, Tika, and cluster analysis. As mostly coding novices, we felt we faced and overcame significant knowledge gaps during this assignment, and we look forward to implementing what we learned about coding, statistics, and Tika in the future.