INF 550 Homework 3: Building Visual Apps to Explore Fake Scientific People and Literature using Data Science

Project Team 03
Abouelnaga, Raghda - abouelna@usc.edu - 2735798900
Cherry, Carlin - ccherry@usc.edu - 8211265507
Johnson, Andrew - andrewbj@usc.edu - 9950819736
Lee, Matthew - mdlee@usc.edu - 4356300240
May, David - davidmay@usc.edu - 5801939142

Introduction

In this assignment, we explored visualization frameworks like D3 and RShiny to showcase the work we have done over the other assignments in this course. We also tried out open source MEMEX software such as GeoParser and ImageSpace, experimenting with other ways to explore our results.

Visualizations

Force-Directed Graph

We started this series of assignments with basic information about the papers–like the author names and affiliations–and one of the first tasks was to look for connections between them. Our first visualization touches on surface-level connections to find obvious patterns, and to build up to why deeper analysis is necessary.

The **force-directed graph** depicts connections between the 214 authors of the original problematic papers from the Bik dataset. Blue dots represent first authors; orange dots represent co-authors; green dots represent related authors (based on profiles in Microsoft Academic).

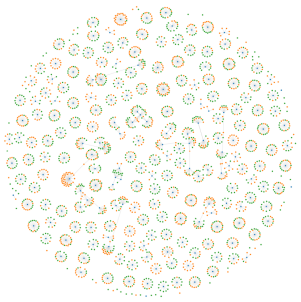

Figure 1: Force-directed graph

We selected this visualization to better understand connections between authors of problematic papers–if there are groups of papers that share co-authors, perhaps from the same circle of researchers or same university, then those connections would be apparent in the graph. Surprisingly, only about 10 clusters of connected webs between authors; for the most part, each author's publication group was an isolated cluster unto themselves.

This type of graph could be useful for someone trying to understand author connections between those who falsify media in their publications. It could also be useful in determining co-authors or "at-risk" populations for media falsification, those who might be candidates for extra review prior to paper publication.

Variable Explorer (scatterplots)

As part of the first assignment, we augmented the Bik dataset with additional features to explore deeper connections between the papers. Our next visualization focuses on exploring patterns and points of interest in the features in our dataset.

The **variable explorer** allows the user to choose two variables of choice for plotting on an XY plane. The variables can be categorical, such as a yes/no to being falsified media, or numeric, such as journal impact factor. The user can color or facet by certain variables for deeper exploration. There are a few options to improve visibility: a sliding scale to select sample size, a toggle to jitter points (forcing them to spread out), and a toggle to draw a smooth line through the average (for noisy numeric plots).

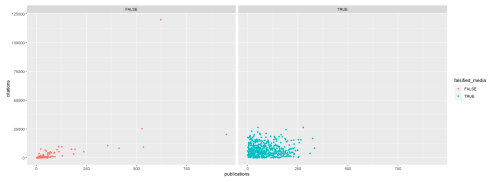

Figure 2: Scatterplots comparing publication by citation distribution for real and fake papers

This can assist with feature selection for clustering, identifying the variables with the most interesting variation. For example, looking at id x publications shows some healthy variation in publication count; id x lab_size_approx is less interesting (as most hit the upper bound at 10). Also, by setting falsified_media as a facet, It also can serve as a sanity check for comparing features between the original and falsified papers. For instance, we can see the distribution of publications/citations is a tad too "clean" among the false papers, as we sampled those numbers from a normal distribution; real papers have some outliers and a less artificial distribution.

This visualization allows for a lot of flexibility on the part of the user. With that, however, the user should know the data well, including its limitations and at least a basic idea of what they would like to visualize, as the number of options could be overwhelming to someone totally unfamiliar with the data.

Cluster Visualization (t-Distributed Stochastic Neighbor Embedding)

Once we had a selection of features in assignment one (determined by visual methods like the variable explorer plots, or by intuition), we dived deeper into analysis by clustering the papers. Now, we revisited our cluster visualization from that assignment, adding user interaction and other improvements.

The **cluster visualization** allows users to explore clusters in our final Bik dataset. The user can select a distance metric to cluster with, variable to color by, number of clusters, and size of points. There is also a toggle to include or exclude fake papers, which has an additional slider to alter the point size for fake papers (to make them more apparent). The technique generating the plots (t-SNE) has a randomness component, so there is a toggle to unlock the seed and reroll another visualization. The graph is also interactive, allowing the user to hover over a point for more info about the data point it represents as well as zoom in and out on specific clusters.

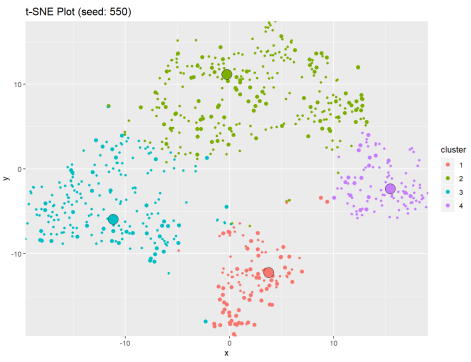

Figure 3: Cluster visualization; larger points are medoids; smaller points are generated papers

We selected this graph to build on the cluster analysis that we performed on our data in assignment one. This graph is useful to a user who wants to find similar characteristics between subsets of data–for example, checking whether all papers involving medicine congregated toward a particular cluster. With the addition of falsified media points, we also want to verify how well the faked metadata obscures the falsification; glancing over a few different seeds, it seems fake papers tend to stick around existing clusters rather than making their own (as expected with how we generated their fake metadata, which was mostly by resampling from the original papers).

Our cluster analysis in assignment one lacked some of the user-friendly features we've updated in this visualization. Particularly, we like the flexibility of allowing the user to hover over each point to get more information about the particular paper that data point represents, in addition to the other more user-friendly aspects of the graph.

Word Cloud

The second assignment explored the concept of media generation and detection. Our next visualization aims at ways to quickly compare original and generated media.

The **word cloud** shows the frequency of words in a given problematic paper. We included sliders to alter the tuning parameters for the word cloud–"minimum frequency" filters out words with frequency below

the specified value; "maximum words" affects the number of total terms that can appear. There is also a checkbox to generate another word cloud; the secondary word cloud shows the frequency of words in the falsified version of the selected paper.

Figure 4: Word clouds of an actual paper (left) and its corresponding generation (right)

Here, we explore the difference in content between the original problematic papers and the falsified papers that we generated in the second assignment. Back in assignment two, we assessed the content by reading through a small sample of papers, looking at logical flow and visual structure; this visualization is more of a quick comparison of content. Notably, each paper and its fake seem to share similar vocabulary among the most frequent words, which makes sense given how Grover treats its contextual input.

This kind of visualization is often for sentiment analysis, but we use it to confirm that Grover excels at emulating a corpus. However, a corpus is just words–without proper training on syntax and structure, the resulting generations lack authenticity, as we know from the second assignment.

Author Map

One aspect of this final assignment is looking for geographical insights; in addition to using GeoParser, we attempted to generate a geographic visualization with our own calculations.

The **author map** shows geographic incidence of papers by their affiliated university. We include a toggle to display/hide fake papers, a drop-down menu to let the user determine the variable for the legend (publications or citations), and drag bars to filter by number of author publications and number of author citations. The bubbles on the map are color coded by the quantity of variable of choice (number of publications or number of citations).

We can use this map to visualize concentrations of fake papers by geography; the "hot spots" for falsified papers become even more evident. For example, there are a large number of papers associated with the east coast of the United States, in central Europe, and in east Asia. Among the authors of problematic papers, there appear to be more smaller authors in the eastern hemisphere compared to the west. Interestingly, when toggling between publications and citations, the largest authors by publication count are not necessarily those with high citation counts.

We distributed the location of fake papers proportionally with the locations of the real papers, so the fake papers do not "add" much to the visualization in terms of novel information. We do give the user to toggle between adding and removing the fake papers from Assignment 2. The graph's feature of selecting the variable of representation gives added dimensionality to the graph as the user can choose their variable of interest.

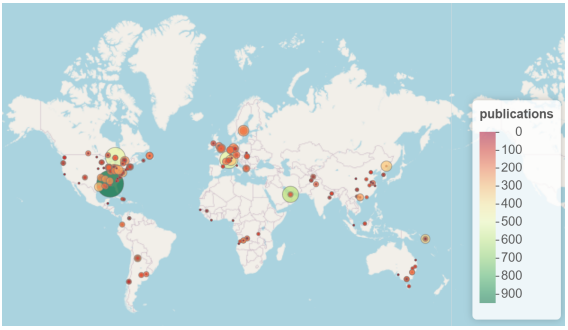

Figure 5: Geographic locations of authors, with size/color of point indicating that author's publication count

Deeper Insights

Location Data (GeoParser)

Another source of geographical insights is the MEMEX GeoParser application.

Comparing the two mappings–the one from the previous section with the location of affiliated universities, and the one from GeoParser with mentions of locations in content–we still see main groupings in eastern United States, central Europe, and eastern Asia. However, there are some differences. GeoParser excludes

points in some places like China and Australia; it is interesting that problematic papers originate from these places, but the content of articles seldom mention locations there. GeoParser also captures points that are not in the author-affiliation map, like a few points in Canada and Russia.

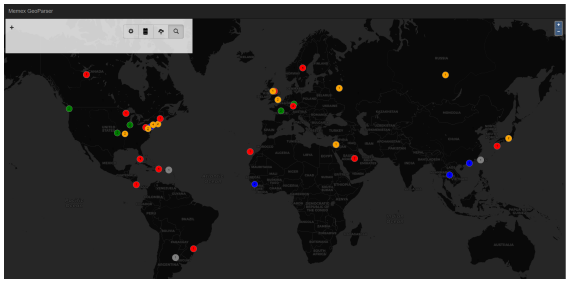

Figure 6: GeoParser results of the original 214 problematic papers

Since GeoParser looks at content, it might excel at finding a given paper's connections–such as plotting the locations of all related papers. However, that would rely on a paper referring to another by location, when often only names/times appear in citations. The author map is not infallible either; due to how Microsoft Academic counts publications, it is possible that a common name interferes with results by pulling the wrong person or inflating the count. Moreover, Microsoft Academic counts all publications involving an author, even if that author played a minimal role (based on position in the author names) in the research. To supplement the geographical insights from the data, one might look at another metric besides the raw publication count. For instance, the Bik et. al. paper used the ratio of published papers to problematic papers from each country to find the probability that a paper from a specific location would contain problematic images. The maps do better at representing information in an interactive way, while Bik's graph of problematic images by country is a better representation of which countries are producing a higher percentage of problematic papers. The ratio of contributed problematic papers is something that could be incorporated into the author-affiliation map moving forward.

Both maps add insights that we did not have in assignment one; back then, we had considered some features based on country of origin, but joining in those features would be quite expensive. To generate the author map, we ultimately did go with such a calculation at the expense of a few hours. One strength of GeoParser is definitely the speed of its insights; configuration difficulties aside, GeoParser provides a nice, quick option for when speed is important.

Similarity (Image Space)

We ingested all images from the previous assignments into ImageSpace. While we were unable to successfully run SMQTK similarity searches, we explored the SMQTK documentation and a basic installation of ImageSpace enough to obtain a general grasp of the concepts.

In theory, we would expect the similarity search to do well in differentiating the kinds of fake images we used (blots, tissues, fluorescent stain, etc). Each of those types has a distinctive color scheme, and the attributes in ImageSpace seem to factor in color in some fashion. From the second assignment, the image generation software, DCGAN, seemed to generate perpendicular lines due to tiled images; it would be interesting to see whether ImageSpace can pick up on images that consist of a panel of smaller images. Perhaps most of all, it would be interesting to see how ImageSpace relates the real images and the fakes. A lot of our fakes look the same, so those probably are close to one another in a nearest-neighbor index.

Overall, configuring ImageSpace was not easy. There was often a hiccup in running the scripts, though these issues were, to some extent, due to our own limitations in hardware and experience.

Conclusion

This final assignment was a fun way to revisit our past work, augmenting old visualizations and discussing ideas anew. This assignment challenged us to think about how to communicate data and how to think about it critically. With that, we feel as though we have taken our first steps to think like a true data scientist, and we hope to carry this thinking with us well into the future.