Assignment 1 Analysis Report

Team A++: Ziyue Chen, Ying Wang, Cheng Shi, Zichao Wang, Ziquan Chen, Wenting Shang

Abstract: This work is implemented with the goal of analyzing the factors that affect the literature fraud rate of a region. After describing and analyzing the Bik dataset, three other datasets are collected to support the analysis. According to the analysis of three features in each of four datasets, university ranking, landslide times, and country economics all have more or less impact on the rate of literature fraud in a region.

1. Introduction

In this essay, we discuss the factors affecting the literature fraud rate by describing the factors, from the Bik Dataset. Also, we extract three or more features from three public datasets, including the university ranking dataset, landslide dataset, and country GDP dataset. In the last part, we use the tika-similarity library to analyze similarities between our data.

2. Bik Dataset

In the Bik dataset, it demonstrates the basic information of the academic literature. To better understand the dataset, we collect other features most by web scraping on the PubMed website, including the university, country, degree area, highest degree, lab size, publication rate of the first author and other journals published in.

By analyzing the dataset, we observe some facts. The year duration in the Bik dataset is 18, from 1996 to 2014. Most papers(91 out of 214) are in “2 - duplications with alteration”. About the first authors, 57 of them come from the USA, and 38 of them come from China; the lab size is from 9 to 24732; the mean number of publications is 84.73, excluding the number larger than 800 as the outlier; the career year is 17.38 on average; the publication rate is 8.39 per year on average, excluding the publication rate number over 100.

3. World University Ranking

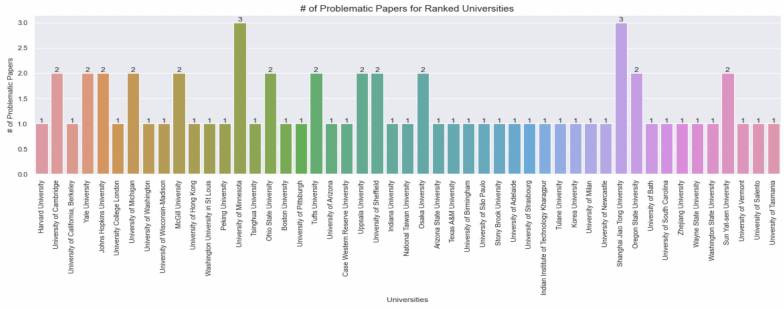
Looking at the “university" feature in Bik, we intuitively ask whether there are connections between the author’s academic behaviors and the quality of his/her affiliation university. By searching on Kaggle, we obtain a CSV file containing 2011-2016 ranking data reported by the Times Higher Education World University Ranking to answer this question.

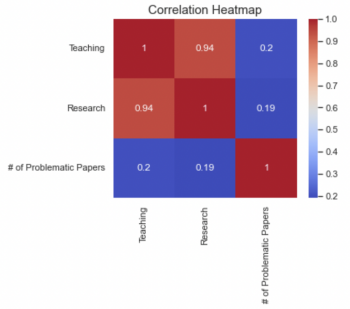
Using Pandas, we first narrow the time period down to 2013 because, in the original Bik dataset, the year that has the greatest number of papers published was 2013. The column “university_name” is necessary as it is the key for merging with the Bik dataset. To assess the overall quality of a university, we select the column “world_rank,” for which a smaller number means a higher ranking and thus higher overall quality. The remaining two features were chosen are “teaching” and “research,” as the former represents a score for the learning environment and the latter provides insights on the volume, income, and reputation of a school’s research. Both scores are on a scale of 100. These four columns are then combined to a new “ranking” data frame that is merged to the Bik datasets using a left join.

As shown in the barplot on the right, among the 214 problematic papers, only 65 of them contain data about ranking/teaching/research, because schools with a ranking after 400 are not reported. This seems reasonable as we expect authors from better universities to possess more serious attitudes towards experiments and research papers. Since unranked universities lack corresponding features, we then focus on ranked ones to find the

correlation between the number of problematic papers and ranking/ teaching/research through barplot and heatmap.

The barplot below has universities ranked from left to right, and there appears to be no trend in the number of problematic papers. In addition, the heatmap also indicates weak correlations between the features, as the correlation coefficients are 0.2 and 0.19. Therefore, we conclude that for ranked universities alone, the ranking/teaching/research features are only slightly correlated with the quality of research papers, but they still tend to result in less problematic papers when compared to unranked ones. We realize if the ranking data is more complete such that it can cover all the universities listed in the Bik dataset, which is hard to accomplish in reality, we may be able to further explore the correlations.



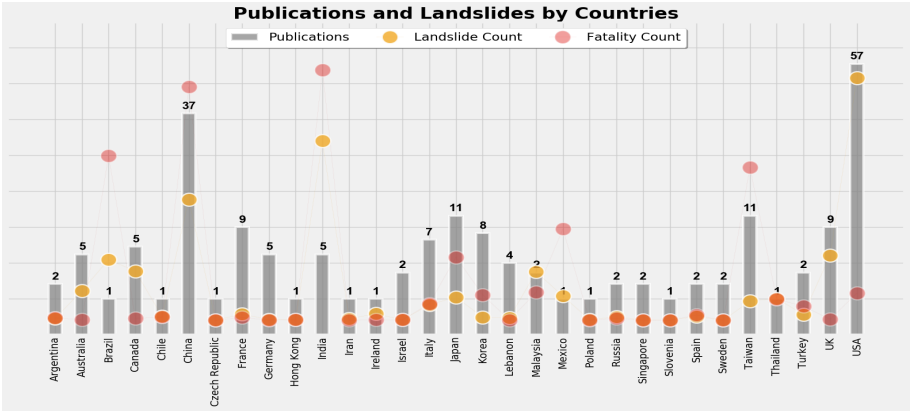


4. Landslide Data

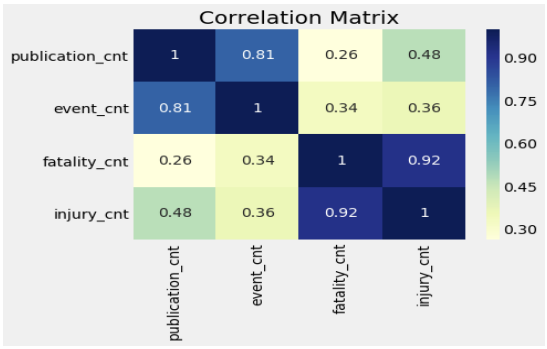
This selected dataset is in JSON and records the landslide relevant data in different countries, identifying rainfall-triggered landslide events around the world. The raw dataset has source information, event details, location related information, landslide category, landslide trigger, landslide size, fatality count, injury count and many other related information.

We calculate how many events happened and how many people died and got hurt in each country during a certain period. By doing so , we have three categorical features: Landslide Category, Landslide Trigger, Landslide Size, and three numerical features: Event Id, Fatality Count, Injury Count. Here we selected the three numerical features that we would analyze with the Bik dataset. Based on the selected features, we want to further explore the integrated data by asking some relevant questions: Does the landslide occurrence/injuries/ fatality somehow relate to the number of problematic publications in certain regions. What's the similarity of landslide data with Bik data?

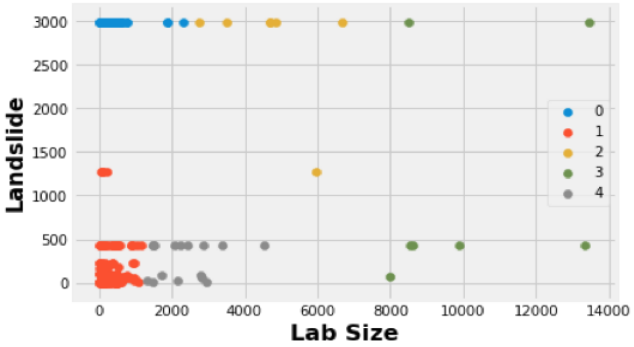
The original landslide dataset contains three categorical features which we will not do any analysis on, so we extracted the three target columns (Event Id, Fatality Count, and Injury Count), and summed them based on regions. We identified some empty data, and we assume that means this certain region did not have a landslide at that time, so we replace the null with 0. After the data cleaning process of the landslide dataset, we apply left join to this dataset and the original Bik dataset by "Region".



Before we applied any statistical analysis, we visualized the landslide data with a number of problematic publications, which makes the analysis more intuitive. The above graph shows that most of the landslide count and fatality count is low, but in countries like China, India, Mexico, and the USA, the occurrence and fatality are significantly higher than in other countries. China and the USA have a much higher number of problematic papers. Then we calculated the correlations of problematic paper count with landslide occurrence/injuries/fatality by implementing the corr() function in python.



The "unintended consequence" is that the publication count has a strong correlation with the landslide occurrence of 0.81 coefficient. We also tried to cluster the lab size with landslide occurrence. The data is very scattered, by clustering them into 5 groups as above, we could not identify any patterns. And the correlation between these two features is 0.05, showing a very weak correlation.

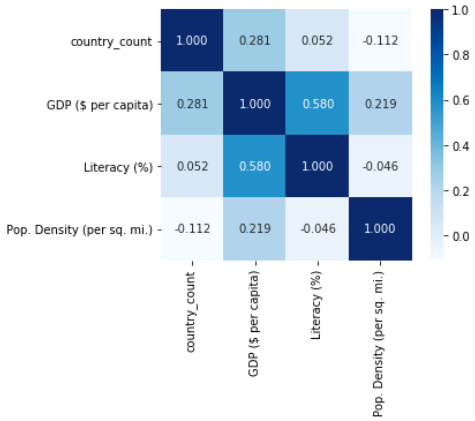


By doing the above analysis, we concluded that the number of problematic papers has a quite strong relation with landslide occurrence count, but a very weak relation with non-regional data, such as lab size.

5. World Countries Data

In order to have a comprehensive understanding of the correlation between scholars' academic behaviors and their countries, we decided to add some features indicating a country's overall living quality and educational level to the Bik dataset. After searching on Kaggle, we found a CSV file of each country's GDP and other related features that perfectly met our needs. And then we performed data cleansing, feature selection, and merged it with the Bik dataset by using pandas.

The original country dataset contains 19 features, many of them such as coastline or arable, is very likely not relevant to our analysis regarding scholars' academic behaviors in each country, so we filtered them out and picked the most three relevant features: GDP, literacy, and population density. We also noticed some problems with the original dataset, for example, it used commas instead of dots to separate integer digits and decimal digits. In addition, it has extra space that comes after the values that may affect matching. After handling those issues, we merged the country dataset with the Bik dataset by using left join and set "Region" as the key.



By doing a count of the occurrences of problematic paper groups by country in the Bik dataset, we assume this is a random distribution and it reflects how different country's scholars behave. Then we calculated the correlations of countries' count, GDP, literacy, and population density by using the corr() function.

The above figure shows that there aren't any strong correlations between the attributes. Only the GDP has a weak correlation with country occurrences count, meaning that the GDP of each country may be slightly related to the scholar's academic behaviors in this country.

6. TIKA Similarity

In general, Jaccard similarity measures the similarity of two sets (A, B) by Jaccard distance, which is calculated by the intersection of AB divided by the union of AB. The edit-distance similarity is usually used on two strings, by deleting, inserting, etc. operations to make one string the same as the other. The cosine similarity is used on two sets, counting the cosine similarity between two vectors. The formulas is $K(A, B) = \langle A, B \rangle / (||A||*||B||)$. In our opinion, the edit-distance similarity is the most accurate measurement to compare two strings, because it records the exact steps that transformed from one string to another, which is logical and measurable.

We apply two modes for the cosine similarity function, inputFile and inputDir mode. The inputFile function directly processes the rows in the dataset, but in the inputDir mode, we split the rows into 200+ files for comparison. The mean similarity score from inputDir mode(0.9998) is much higher than the inputFile's (0.9025). The reason for the difference may be that the inputFile mode measures the content of rows, and the inputDir processes the metadata of every file to compute similarity scores. However, the scores of both modes are above 0.9, which is caused by natural characteristics of a table. Because the content in rows share the same columns and fixed traits, it possibly leads to high similarity scores.

We use the python file "compare_in_cosine_similarity.py" to compare the descriptive statistics generated by the cosine similarity function.

Score comparison of changing features in the Bik dataset

By comparing between the original Bik dataset(bik) and the Bik dataset with different columns, including Bik with affiliation(bik_aff), Bik with affiliation and labsize(bik_aff_labsize), Bik with affiliation, labsize and other journals published in(bik_aff_labsize_otherpub), and Bik with affiliation, labsize, other journals published in and publication rate(bik_aff_labsize_otherpub_pub), we can see that adding additional feature makes our Bik data less similar gradually. The "other journals published in" feature contributes the most, lowering the average score by about 0.5%. In total, the mean similarity score decreases from 0.9174 to 0.9030, which means that the Bik dataset's mean similarity score is sensitive to the additional features we add.

	bik	bik_aff	bik_aff_labsize	bik_aff_labsize_otherpub
mean	0.917397	0.908414	0.908410	0.903017
std	0.076395	0.074449	0.074434	0.072545
min	0.406944	0.412170	0.412220	0.416913
25%	0.887602	0.877910	0.877915	0.872420
50%	0.941888	0.930126	0.930111	0.922698
75%	0.971671	0.961764	0.961765	0.954879
max	0.999798	0.998942	0.998883	0.998891

	bik_aff_labsize_otherpub_pub
mean	0.902969
std	0.072518
min	0.417090
25%	0.872377
50%	0.922675
75%	0.954784
max	0.998740

Score comparison of changing features from Bik, Ranking, Landslide, GDP datasets

	bik_aff_labsize_otherpub_pub	bik_aff_labsize_otherpub_pub_landslide
mean	0.902969	0.902896
std	0.072518	0.072488
min	0.417090	0.417208
25%	0.872377	0.872360
50%	0.922675	0.922582
75%	0.954784	0.954693
max	0.998740	0.998689

	bik_aff_labsize_otherpub_pub_landslide_ranking	all_features
mean	0.902569	0.902503
std	0.072458	0.072379
min	0.417208	0.417451
25%	0.872075	0.872057
50%	0.922258	0.922175
75%	0.954391	0.954230
max	0.998690	0.998569

By comparing similarity scores between the bik_aff_labsize_otherpub_pub dataset (original Bik with affiliation, labsize, other journals published in and publication rate features), and the datasets with different features from the other three datasets one by one, it shows that the cosine similarity score of our

dataset decreases from 0.9030 of bik_aff_labsize_otherpub_pub dataset to 0.9025 of the dataset adding additional features from other three datasets. From the descriptive data, the dataset of world university ranking contributes most among those datasets, and our additional features play a significant role in distinguishing our data.

Thoughts about TIKA

TIKA is a good tool for comparing the similarity between two MIME file types, and it includes different methods for calculating similarity scores. However, it's not easy to use for new users because of the limit of file types. Thus, users have to check the source code package and transform their files to specific file types so as to run TIKA. Also,

7. Conclusion

According to the analysis above, here come the conclusions:

1. **The GDP of each country may be slightly related to the scholar's academic behaviors.**
   According to the analysis of country GDP data, the GDP of a region may have a weak correlation with country occurrences count, so it may have something to do with the literature fraud rate.
2. **The University ranking level in a region has an impact on the literature fraud rate.**
   The ranking/teaching/research features are only slightly correlated with the quality of research papers, but they still tend to result in less problematic papers when compared to unranked ones.
3. **The landslide occurrence times have strong correlations with the literature fraud rate.**
   The "unintended consequence" is that the number of problematic papers has quite a strong relation with landslide occurrence count, but a weak relation with non-regional data, such as lab size.
4. **The similarity between articles is high but can be lowered by adding additional features.**
   Two of the three similarity methods have high similarity scores on average. Besides, adding additional features, such as the other journals published in and the university ranking, can lower the score to some extent.

Other mandatory questions:

What did you notice about the dataset as you completed the tasks?
We noticed that the original dataset does not have much information for us to do further analysis, but after adding demographic data, we were able to get more insights. Statistics sometimes reveal results that are opposite to our common sense, this is where we can dive into the data and apply more scientific methods.
What questions did your new joined datasets allow you to answer about the Bik et al papers previously unanswered?
After we joined the datasets together, we were able to analyze the authors' personal information in detail and how they were related to geographical, educational and financial features.
What similarity metrics produced more (in your opinion) accurate measurements? Why?
In our opinion, the edit-distance similarity is the most accurate measurement to compare two strings, because it records the exact steps that transformed from one string to another, which is measurable.
What did the additional datasets suggest about "unintended consequences" related to media forensics data?
The number of problematic papers has quite a strong relation with landslide occurrence count, but a relatively slight relation with school ranking.
What insights do the "demographic" features of the authors tell us about the data?
By adding the demographic features, we found out that most authors are from China and the US. Also, merging Bik dataset with landslide, GDP and ranking datasets using demographic features, we found more situations like we discussed above.