Emma Lynn, Nathan Freestone, & Robiul Islam

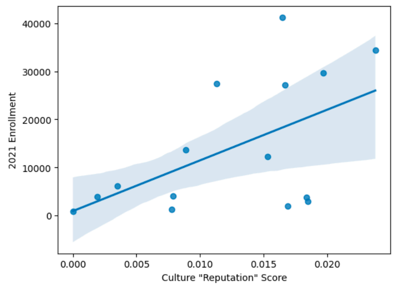
Project 3 Report

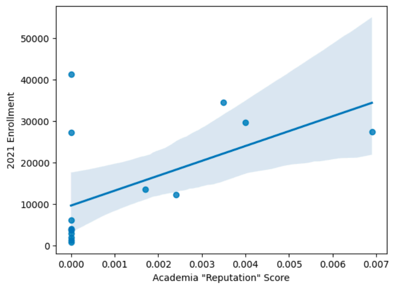
Introduction - Our analysis combines different techniques to analyze the content of Wikipedia articles relating to colleges and universities. Wikipedia is a common source of information for people of many backgrounds, and understanding how your institution is described by Wikipedia can be helpful for understanding many things about its public image. Our analysis focuses in particular on institutions in Utah, and could be helpful to university administration or state lawmakers. Our slides are linked here and GitHub repo is linked here.

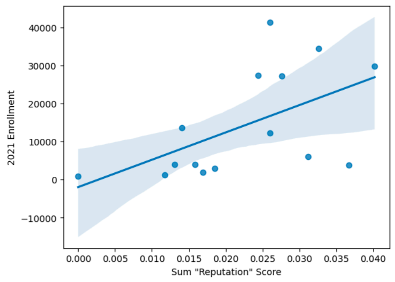
Dataset - For our first analysis, we used schools' Wikipedia pages as samplings of their "reputation". We pulled the content of the Wikipedia pages for 15 public colleges or universities in Utah using the Wikipedia API. We counted the occurrences of certain keywords within the articles. We grouped keywords into "topics". These topics and keywords were Student Culture (campus, university, clubs, events), STEM (science, technology, engineering, math), Athletics (football, basketball, volleyball, sports), and Academia (research, graduate, undergraduate, honors). We also added all of the keyword scores together to come up with a sum score. Then we added data about school enrollment (Dugovic). Finally, we put our data into a dataframe. This dataset is suitable for analysis because it will allow us to compare what information is very easily accessible about a school and enrollment, and determine what factors are potentially related. In the second analysis, BeautifulSoup was used for data scraping from the wikipedia page and for parsing HTML. The major parameters like establishment year, student enrollment, budget, research funding, and faculty numbers were located in specific tags (e.g., table, class). These tags were identified by inspecting the page structure using browser developer tools. For numeric data extraction (such as budget and student numbers), we used Python’s re library to filter out the relevant figures and standardize them for comparison. For the third analysis, the primary data was scraped using BeautifulSoup to access the HTML content of the Wikipedia pages. For numerical data such as student enrollment, research funding, and budgets, regular expressions were used to extract values from the text. In cases where monetary values were inconsistent, the amounts were standardized to millions. Missing values (where data could not be found) were handled by substituting them with "NaN."

Analysis Technique - In our first analysis, we used two techniques: calculating the Pearson correlation coefficient and plotting linear regression. These techniques are suitable for this analysis because they help us determine if there is a relation between two factors over one group of people, which is what our first question was asking about. In the second analysis, we used a combination of metrics to score articles, then looked at the distribution of the scores for different schools (Utah vs Colorado). In the 3rd analysis, Bar plots were used to compare the number of undergraduate and graduate students across universities. This technique is useful for providing a clear visual comparison of categorical data across universities. Line plots were employed to observe trends and relationships between continuous variables. Line plots help in identifying how changes in one parameter (e.g., increased research funding) affect another (e.g., faculty-student ratio).

Results - For our first question, we were wondering what aspects of a school's reputation can be correlated to enrollment at that school. We collected data as described above. Then we ran tests to determine if having a reputation related to any of our specified topics is related to a school's enrollment. While we did not find evidence that a reputation in STEM or Athletics is related to enrollment level, we did find a moderate positive correlation between reputation in Student Culture and enrollment and Academia and enrollment. We also found a moderate positive correlation between the summary reputation score and enrollment (see graphs below).

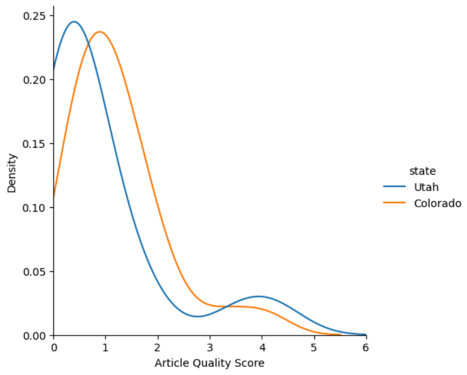




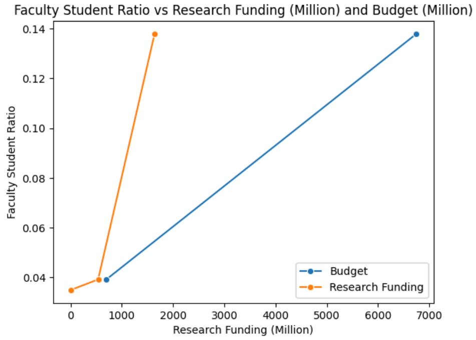


This means that schools who want to increase enrollment could focus their efforts on building a stronger reputation related to Student Culture and Academia.

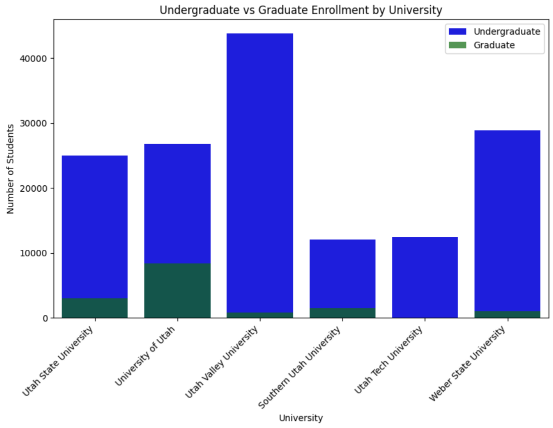
Our second analysis focuses on comparing the quality of information that exists about Utah universities to that of Colorado's. Colorado was chosen as a comparison point, because it is a state in the same region with a similar population and number of institutions. The Wikipedia articles were scored on a number of metrics including length of article, number of citations, and number of images. We then performed an analysis on the distribution of those scores between the two states.



As seen in this figure, Colorado had a slightly higher mean score. However, the difference between the two means is not statistically significant (p = 0.25), meaning we can't conclude that Utah schools are being underrepresented on Wikipedia compared to Colorado.



The third analysis provides a comparative analysis of six major universities in Utah, based on parameters including the year of establishment, the number of undergraduate and graduate students, research funding, annual budget, number of faculties, and faculty-student ratios. Undergraduate enrollment varies significantly, with Utah Valley University leading. Utah Tech University only offers undergraduate programs. The University of Utah leads in graduate enrollment due to its focus on research, followed by Utah State. Universities with higher research funding tend to have stronger research output and larger budgets. Smaller schools, like Southern Utah and Weber State, focus more on teaching with lower budgets. Higher funding typically leads to better faculty-student ratios, as seen at the University of Utah and Utah State, allowing for smaller class sizes and better academic support.



Technical - For our first question (school reputation and enrollment), we first divided the number of occurrences of each key word by the word count of the article to normalize for articles of different lengths. We used two analysis techniques: calculating the Pearson correlation coefficient and plotting linear regression. These techniques are suitable for this analysis because they show us if there is a correlation between two variables in one population, which is the format of our question. For the Pearson test, we set our significance level to .05. After running the tests, we got these values:

	Culture	STEM	Athletics	Academia	Sum
Test stat	.54	-.106	.211	.539	.553

p-value	.037	.707	.451	.038	.033

We have statistically significant results for the tests run on the Culture, Academia, and Sum categories, since their p-values are less than our significance levels. Since our test statistics were .54, .539, and .553, respectively, we can conclude a school's reputation for Student Culture, Academia, and combined "reputation" for all of our topics have moderate positive correlations to its enrollment.

The second analysis involved creating a score for each wikipedia article that represents the quality of the article. The score was based on 5 metrics. (Section Count: articles with larger numbers of sections are likely better organized and/or have greater detail, Reference Count: articles with more references are considered better, Word Count: longer articles can have more information in them, Image Count: images can help readers engage more with the subject, or provide additional context, and Internal Links: articles with a large number of links to other wikipedia pages are easier to navigate and have more use.) Each of these metrics was then MinMax normalized, and summed (evenly weighted) to arrive at the final score. We believe this score to be a reasonable metric of article quality without requiring advanced language processing. The primary limitation of this metric is that it has a heavy bias towards longer articles, and does not consider the density of information in the articles. It also fails to consider other more subjective things like writing style and voice.

In our third analysis, preprocessing was necessary to convert text into usable numerical formats. Monetary values were found in different formats (thousands, millions, etc.). To standardize the data, all figures were converted to millions using appropriate scaling. Some data points, such as research funding or faculty numbers for smaller universities, were not available. In such cases, the missing values were filled with nan (Not a Number), allowing for seamless analysis without causing calculation errors. To compute the faculty-student ratio, we combined the undergraduate and graduate student populations and divided them by the number of faculty members for each university. With the cleaned data, two statistical plots were generated to visualize relationships between the key variables. Bar plot provided insights into how undergraduate and graduate student enrollments differ across the universities whereas line plot was used to demonstrate the relationship between a university's financial resources (budget and research funding) and its faculty-student ratio.

References

Dugovic, Trisha. “Enrollment at Utah’s Public Colleges and Universities Grows Overall.” *Utah System of Higher Education*, 25 Oct. 2021, https://ushe.edu/2021-fall-enrollment/.