Names: Arya Song, Ashley Yu, Claire Zhou

Strategies to Boost Wikipedia Editing Activities

Introduction

In this project, we aim to explore the key factors contributing to the success of Wikipedia, one of the most well-known and influential crowdsourcing platforms among the world. We want to determine whether Wikipedia's growth is better supported by attracting new editors or by motivating current editors to increase their contributions. To this end, we analyzed data on the number of editors and editing activities, sourced from the Wikimedia Foundation's public statistics portal (https://stats.wikimedia.org/). Our analysis includes various metrics, including edit counts, new page creations, net and absolute byte changes, and page edits, distinguishing between logged-in and anonymous editors who have edited at least once in a given month. Additionally, we considered Wikipedia data across multiple languages—Arabic, Chinese, English, French, Russian, and Spanish—to add another dimension to our analysis. Our objective is to use the growth trends in editing activities and editor numbers to uncover the most effective strategy for Wikipedia to enhance its platform, focusing on either expanding its editor base or boosting activity among existing contributors.

Data Analysis

Our data for different countries have different ranges, so we normalized data to 0-1 range using this formula:

$$\frac{x-min}{max-min}$$

We used Pearson correlation coefficient to measure the relationship between the number of editors and editing activities. Pearson correlation coefficient varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases. The p-value roughly indicates the probability of an uncorrelated system producing datasets with a Pearson correlation at least as extreme as the one computed from these datasets. Both correlation coefficients and p values are visualized as heatmaps in Figure 1 and Figure 2. Here, the "users" means logged-in editors, and "anonymous" means non-logged-in editors.

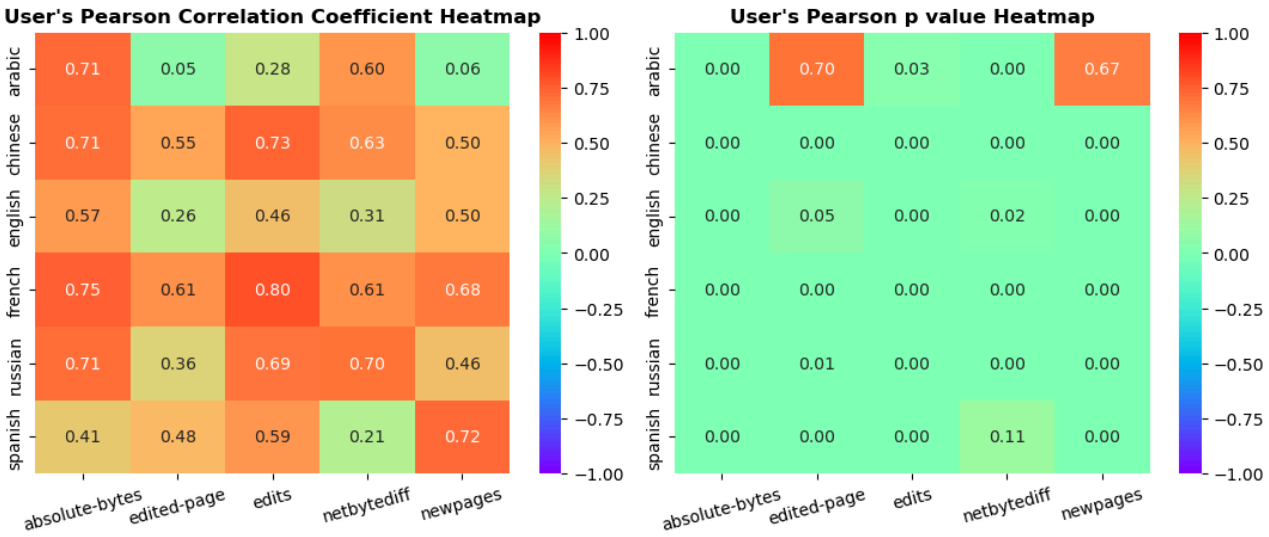

Figure 1: User's Pearson Correlation Coefficient and p Value Heatmaps

For users, all the correlation coefficients are positive, ranging from 0.05 to 0.8. Most relationships have very small p values, meaning it is almost impossible for an uncorrelated system to produce such a dataset. However, for the language Arabic, the relationship between editors and edited pages, and the relationship between editors and new pages have p values around 0.7, which indicates the dataset is nearly random. A potential explanation can be a social

trend that editors using this language focused on a certain group of pages, and were not interested in editing or creating other pages.

There is one more place where we can not be at least 95% confident to reject the null hypothesis that there are no significant linear relationships. For language Spanish, the p value of the relationship between editors and net byte difference is 0.11, while absolute byte differences and edits had p values of 0. It is possible that some users participated in edit wars, in which many edits were rolled back.

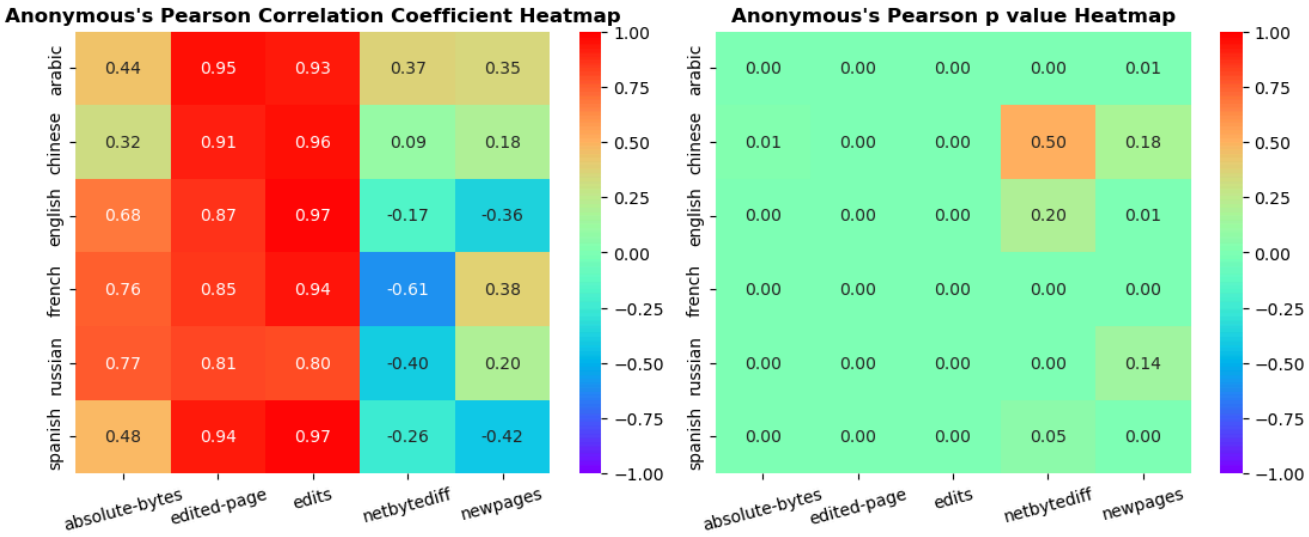

Figure 2: Anonymous's Pearson Correlation Coefficient and p Value Heatmaps

For anonymous editors, the relationships between editors and activities seem more extreme than logged-in users. We can see correlation coefficients from 0.81 to 0.97 in edits and edited pages for all languages we investigated, which do not appear at all for logged-in users. We also see negative coefficients from -0.61 to -0.17 in net byte differences for languages English, French, Spanish and Russian. It does make sense if many anonymous editors are only using Wikipedia when they join an edit war regarding controversial topics they care about.

We also performed Theil analysis to be robust to outliers. Figure 3 shows the slopes for both users and anonymous editors. Comparing the slope values and the correlation coefficient in previous analyses, we can see that they are basically consistent. Negative slopes correspond to negative correlation coefficients, and vice versa. The relativity in numbers also mostly remained.

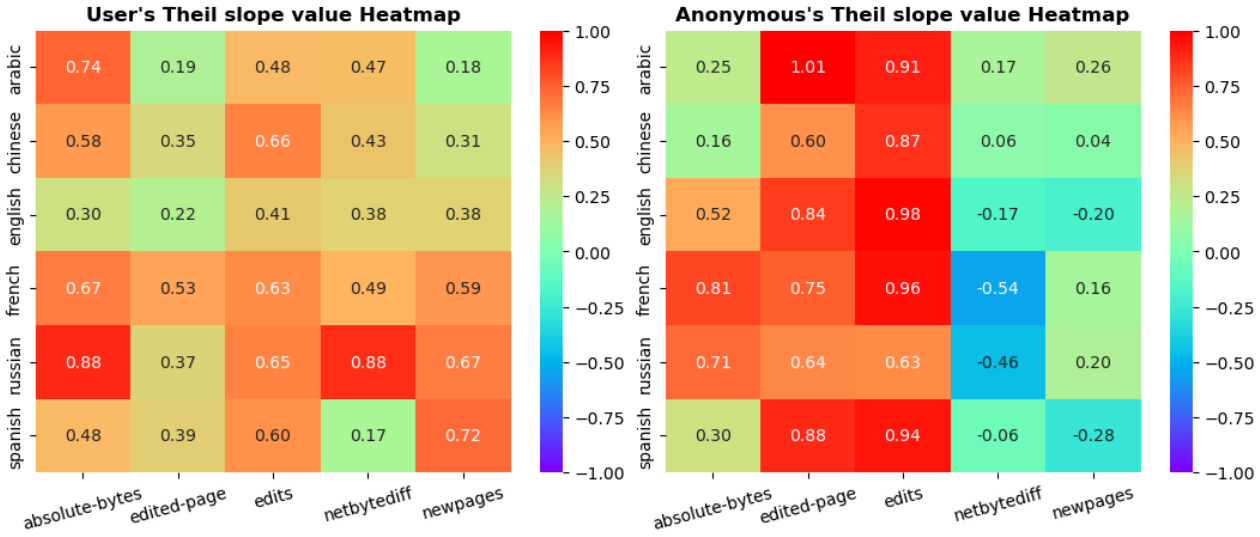

Figure 3: User's and Anonymous's Theil Slope Value Heatmaps

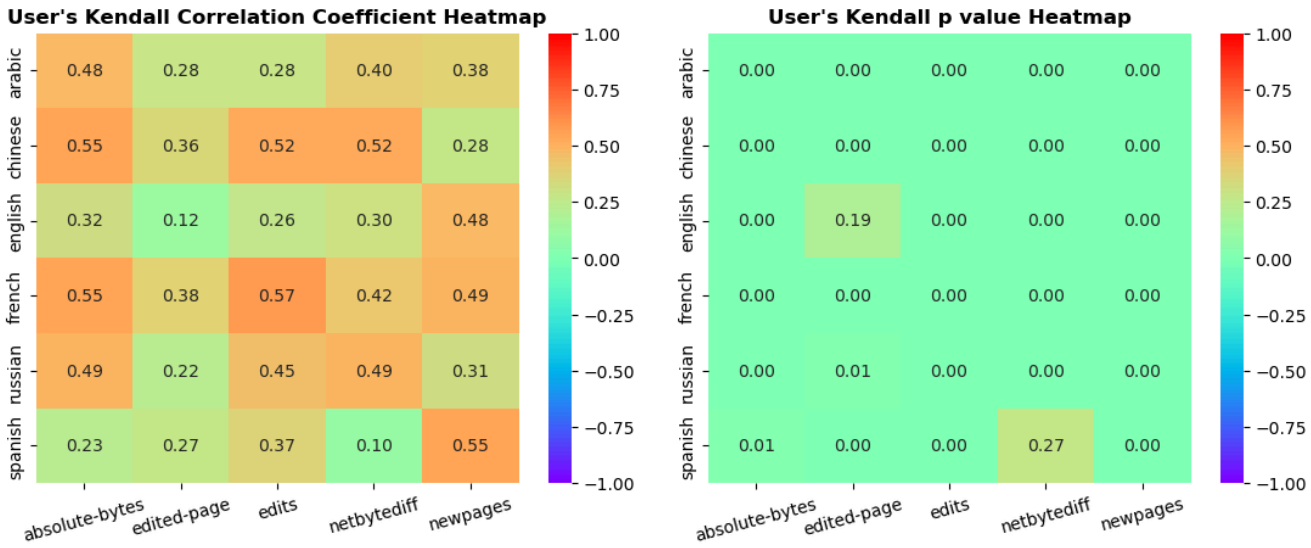

Figure 4: User's Kendall Correlation Coefficient and p Value Heatmaps

However, it might not be enough to only analyze linear relationships that only have one slope. Some metrics may be better modeled by other degrees of the fitting polynomial. Thus, we further calculated Kendall's tau for the monotonic relationship as in Figure 4 and Figure 5.

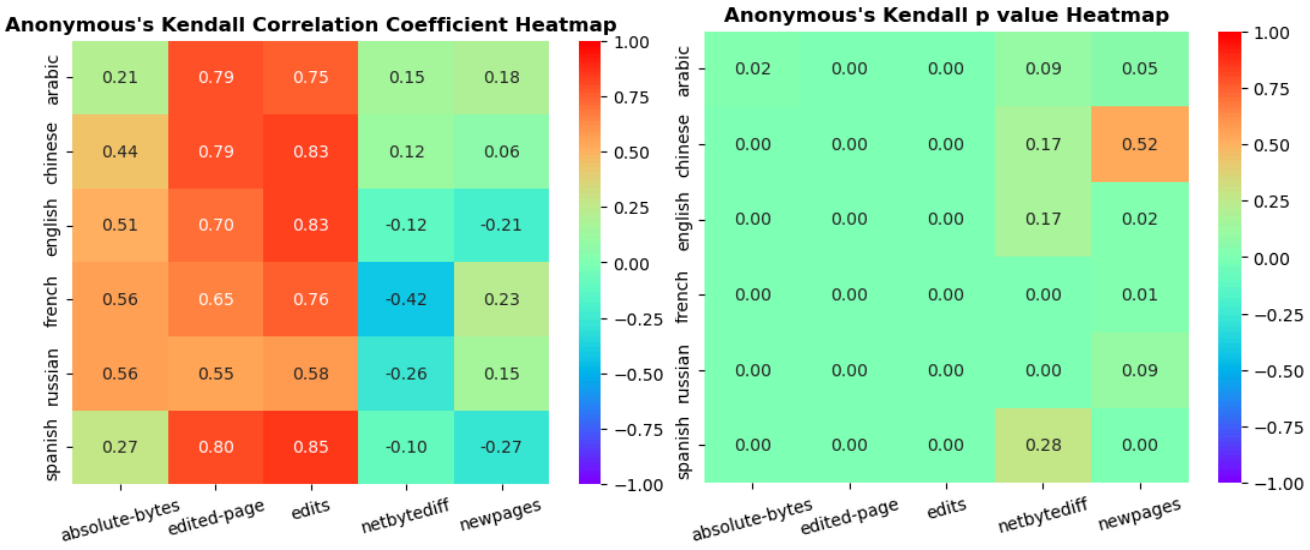

Figure 5: Anonymous's Kendall Correlation Coefficient and p Value Heatmaps

The coefficients and p values of Kendall's are mostly consistent with the results of Pearson's. It is notable that some relationships with high Pearson p values got lower p values in Kendall's, such as the new pages and edited pages by Arabic logged-in users. Some, on the other hand, increased a little.

The difference between the two results may be due to the limited number of users in some languages. For example, Arabic users are the fewest among the six language groups. However, to determine which result better fits the reality, we decided to first fit the scattered data into polynomial functions using regression, and then visualize them. We employed degrees of 1 to 5, then chose the degree with the lowest sum of absolute residuals. As we expected, many metrics are better fitted with the highest degree, and only some few are not. The visualizations are shown in Figure 6.

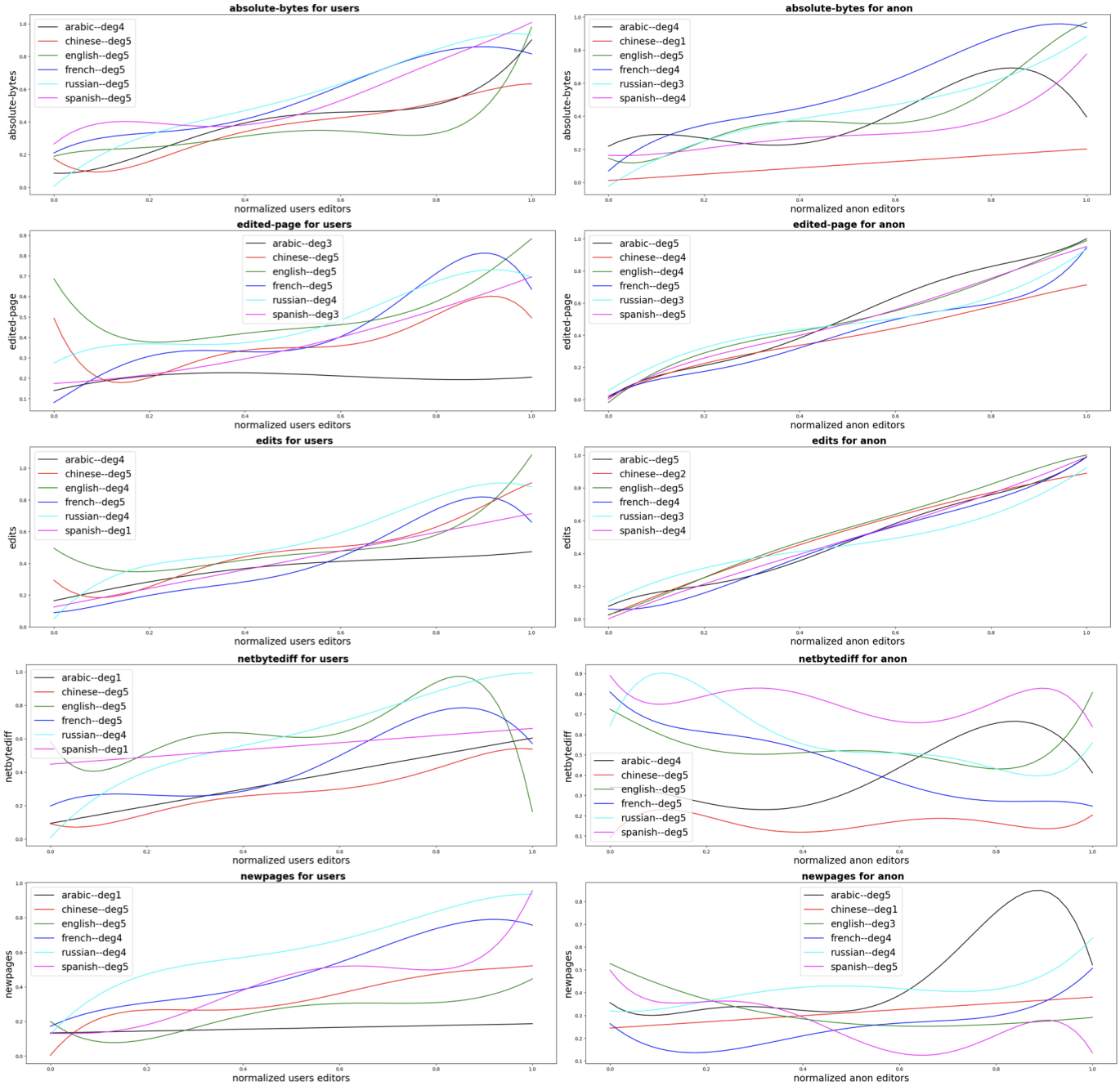

Figure 6: User's and Anonymous's Regression graphs

According to the graphs, the new pages and edited pages by Arabic logged-in users are relatively static, compared with other metrics. More specifically, the number of new pages has a linear relationship with a small slope, and the number of edited pages is more like a tiny wave.

Results

Among the five metrics we chose, it is clear that net byte differences can be significantly and negatively influenced by edit wars. Anonymous editors are more likely to join such wars. Although anonymous editors are more linearly correlated with edits than logged-in users, they do not always contribute to Wikipedia. For a healthier growth of the community. It is more important to attract people who are willing to join as a user, than to attract people who only utilize the platform for their own interests.

Increasing editors or encouraging existing editors to contribute more, which is more effective to boost meaningful editing activities? Generally, we conclude that there is a significant monotonic relationship between the number of editors and editing activities. Thus, attracting more editors would be a better strategy to vitalize Wikipedia. The effectiveness, though, would highly depend on the language one is using, and which metric matters most in one's opinion.

Future Work / Limitation

In our proposal, we assumed that there exists a causal relationship between the number of editors and the number of editing activities. However, we are only safe to conclude that there exists a correlation between the two variables in our work. If we want to go a step further and establish causality, we need to demonstrate that changes in one variable directly caused changes in the other, while controlling for potential confounding variables. In our future work, we can investigate other factors that might have an impact on the editing activities beyond the number of editors. We can perform our regression analysis controlling for these variables and assess whether the coefficients of the number of editors are still statistically significant. If so, we can proceed with our conclusion of the causal relationship between the two variables.