**CS 5128/6028 Large Scale Software Engineering**

**Graduate Student Project**

Austin Conn ([connat@mail.uc.edu](mailto:connat@mail.uc.edu))

John Brown ([brow2jp@mail.uc.edu](mailto:brow2jp@mail.uc.edu))

**Overview:**

The graduate student project for Large Scale Software Engineering called for a continuation of the key focuses present throughout the semester. Our group, consisting of Austin Conn and John Brown, chose to focus on the popular chat platform Discord for our main focus for the project. As part of the definition for the project, we compared our findings of the Discord changelog against the datasets previously identified in the course, the WebEx and Zoom datasets. Our analysis of the Discord dataset does extend back to 2017 whereas the datasets from the course are limited to the year of 2022. Our dataset was directly derived from the Discord changelog found on the Discord Developer Portal at <https://discord.com/developers/docs/change-log>.

Our project consisted of two parts: an Exploratory Analysis of our dataset, and a comparison between our dataset and the collective datasets analyzed during the duration of the course, the Zoom and WebEx datasets. During our exploratory analysis, we derived a method of collating and organizing our dataset for analysis, as well as defining several conclusions we arrived at while analyzing the dataset. During our comparison phase of the project, we compare our conclusions from our analysis and compare them to the conclusions found during the course.

**Exploratory Analysis:**

In order to compile a dataset for the Discord changelog, our first step was to scrape the webpage found on the developer portal into a file which we could manually manipulate into a format that could be analyzed. Previous attempts were made at scripting the scraping and manipulation of the webpage. However, it was found that due to a lack of standardized approach Discord took in publishing their changes, the nature in which the webpage data was formatted, and the lack of structured versioning, the level of effort required to implement these irregularities into organizing the data outweighed the level of effort required to manually manipulate the data into an analyzable format.

We do note, as opposed to the datasets from the course, our dataset does extend back to 2017. After compiling the information from the Discord changelog into an analyzable format, we found there were a total of 259 feature entries (Figure 1). Most of these entries were linked together, implying these feature releases were release alongside each other. This implication led to a classification of 85 unique release days over the more than five years spanning from 2017 through 2023. A key data point lacking in the Discord changelog were changes related to bug fixes or issues resolved as a result of user feedback. Due to the lack of granularity in bug fixes being reported, no features were removed or deprecated prior to exploratory analysis. We took the liberty in this decision during further analysis in the hopes that our ability to compare the Discord data with the Zoom and WebEx data.

**Figure 1: Discord Feature Count by Year (2017-2023)**

**Figure 2: Discord Feature Count by Month (2017-2023)**

**Figure 3: Discord Release Day Count by Year (2017 -2023)**

… talk about fourth graph …

**Figure 3: Discord Release Day Count by Month (2017 -2023)**

… talk about fifth graph …

**Figure 5: Discord Feature Count Depreciating Old Features by Year (2017 -2023)**

… talk about sixth graph …

(Just 2022)

**Figure 6: Discord Feature Count by Month (2022)**

… Talk about figure 7 …

**Figure 7: Zoom Feature Count by Month (2022)**

… Talk about eighth picture …

**Figure 8: WebEx Feature Count by Month (2022)**

**Conclusions:**

… discuss conclusions …

\*talk about workflow of each company

\*\*consistent for Webex, zoom and discord have months with nothing

\*talk about release windows for discord/zoom

\*\* talk about quarters (of the year with like a business angle)

**Figure 9: Combined Feature Count by Month (2022)**

TODO:

* Formalize the spreadsheet
  + submit
* Write the report
  + Link to definition: https://en.wikipedia.org/wiki/Exploratory\_data\_analysis
  + Throw as much garbage as possible!
  + Submit
* Effort/peer evaluation
  + Create/submit

“Curate the features” -> organize

Exploratory analysis:

* Total # of features
* Total/average/mean/max min # of features per month
* # of tokens (voc size)
* # of sentences per feature
* Sentence length
* Feature length …

Draw meaningful comparisons with Zoom and WebEx:

* When are released happening (month)
  + Compare by year (2022 for zoom data)
  + “clean boundaries”
* Number of releases per year
* Sentence length