DATA301 Project Progress Report

Phoebe Zhang 62244228

What have you done so far on your project?

In my project so far, I have completed the data cleaning and preprocessing phase. I have loaded in both the steam_games.json and steam_reviews.json datasets into Dask, normalized tag formats and converted release_date to datetime. I filtered the games and reviews datasets to only include the data needed to answer the research question. I subsequently joined the two datasets on their respective matching id and product_id columns to create a single merged dataset with the relevant columns (product_id, app_name, tags, hours, release_date). From there, I aggregated the data by game to compute the number of reviews and total playtime recorded by users who left reviews. To define a threshold for high-engagement dependent on the total number of hours played and number of reviews, I used the top 75th percentile of total hours and number of reviews. I filtered the dataset to only contain games with total hours and number of reviews surpassing the defined thresholds. To identify recent games, I defined them to be those released in the past 2 years (relative to the latest release date in the dataset) and a separate subset was formed for recent games.

What are you doing next? What is your plan to complete the required project components on time?

The next step is to apply the A-Priori algorithm to identify frequent combinations of user-defined tags in the high-engagement games subset. Afterwards, I will use the Jaccard similarity to compute how much recent games resemble the tag combinations found in high-engagement games to help predict their potential success. To complete the project on time, I plan to implement this code within the next week while simultaneously drafting the final report.

Do you have any road blocks?

In the project so far, my main road block has been handling the large JSON datasets, which contained inconsistent schema and posed memory challenges. Reading in the review data required multiple attempts to deal with inconsistent column orders and missing columns like compensation which lead to differences in expected and actual metadata. Another challenge I faced was defining recent games. I initially had defined recent games to be those released in the last 6 months (relative to the latest date in the dataset) but I found that the data was sparse for that range. To confront this problem, I tested multiple different time frames and concluded that a sizable subset of the data could be formed by defining recent games as those released in the last 2 years from the latest release date in the dataset. The Dask documentation has been extremely helpful so far for coding and resolving these issues.