ASA Datafest 2022

Team 05 - No pi Charts

Summary Paper

Sayyed Faraz Mohseni,[1] Magnus Magnusson,[1]
Lane Robert Lewis,[1] Akash Satpathy,[1] Gary Bales[1]
[1] *The University of Arizona, Tucson, AZ*

1. Introduction

We used the *Elm City Stories* log data to explore correlations between player demographics, self-perception, and performance in the game. The data were reorganized and relevant information, such as age, gender, and race of the avatar chosen by the user, was extracted using efficient data cleaning methods. A custom sentiment analysis program was used to categorize *aspirational avatar descriptions* into positive and negative personality traits, family-related words, and career-related words. Simple Natural Language Processing (NLP) methods, Pearson Correlation Coefficient analysis, Exploratory Factor analysis, and data visualization were used to examine the newly engineered dataset.

2. Methods

A variety of languages (Java, Python, and R), as well as popular data science packages including Pandas, OpenCV, and Tidyverse, were used to parse the dataset. For sentiment analysis, a dictionary of lemmatized words across the four categories of positive personality adjectives, negative personality adjectives, career-based words, and family-based words was created. A custom program was written to extract the data from the raw log files and categorize them into the four groups mentioned above. Fractions of positive and negative words belonging to each category and demographics data (age, gender, and avatar choice) were also added to the "clean" dataset. Further, the oldest and the newest *S5* scores for a subset of fifty-five subjects, and the change in *S5* scores (i.e., the difference between the newest and oldest score) was also added.

The learning rate of players was examined using a linear regression model with the number of minigame attempts of *People Sense* as the predictor and the scores as the response. Here, the learning rate is defined as the slope of the linear regression for each subject. One subject did not play the game *People Sense* and was dropped from the dataset. The number of minigame attempts and the slopes from the regression was added to our "clean" dataset. To obtain the 3-dimensional latent space on the People Sense minigame, a maximum likelihood exploratory factor analysis on the first 15 measurements of the game for all players with fifteen or greater games was conducted. We selected the three factors that explained the highest variance and transformed the minigame data onto them.

The pairwise correlations between variables on both the datasets were computed (once including the S5 scores, and once without it). A private Github repository was used to share code and the "clean" dataset.

3. Results and Conclusions

- Relatively high correlation between use of job-related words and number of attempts
- Relatively high correlation between gender and use of family-related words
- Relatively high correlation between gender and use of negative personality words
- Relatively high correlation between our learning measure and positive adjectives
- Relatively high correlation between our learning measure and positive feelings towards oneself
- The three largest variance factors of People Sense scores explain most of the data as roughly corresponding to learning, learning then decline, and decline

Additionally, attempts were made to map gameplay progression and study the attention spans of the players. Future studies could log data such that gameplay progression and attention spans could be tracked with ease. Follow-up studies could further aim to evaluate game effectiveness as a preventative tool for risky behavior.