Math 142 - Project Proposal

Group 3 Members: Tong Xie, Yuheng Ding, Carol Kang, Sunil Trivedi

Problem

Wordle is a game where players try to solve the puzzle by guessing a five-letter word in six tries or less, receiving feedback with every guess. In this project, we work with data of players' reported scores on Twitter (percentage of users solving the problem in [0, 1, 2, …, 6, X] tries). In particular, we analyze the daily number of reported results and their distributions.

Main Question

The main questions that this project is aiming to answer is showing the correlation between the rarity of a word and the distribution of the number of tries it takes for people to answer the Wordle problem and forecasting the number of reported results on a future date.

- **Difficulty & Distribution of Results**: How does the difficulty of the solution word relate with the number of tries it takes to solve the wordle?
- **Number of Reported Results**: How many people will be participating in reporting their results in a future date?
- **Prediction**: We use our model to predict the number of reported results and score distribution on March 1, 2023 with the quest “eerie.”

Mathematical Approach

- For Number of Reported Results, we are planning on using discrete time models (reported results on day T).
- For now, we model the number of participants and distribution of scores/attempts independently.
- For the number of participants, we will use a linear recurrence or a SIR model.
- For the difficulty of the puzzle, use the discrete time subpopulation distribution where the categories are numbers of attempts, with consideration of a word's rarity.

Simplifying assumptions

- The less frequently a word is used in English, the more difficult it is to guess in Wordle.
- The distribution of guesses on a certain day does not vary based on the number of participants on that day.
- Treat every participant as the same (i.e. everyone has the same skill level)
- Reported Results on day T are independent of results on day K where K ≠ T.

Potential Challenges

- There are typos in certain solution words in the given dataset (words with length != 5), which may require some data cleaning.

- The letters in the word might impact the difficulty as well. For instance, if we have “z” or “x”, there are limited options, whereas simple words like “mummy” are more difficult to guess. (Could acknowledge this in conclusion)
    - Number of repeated characters might also play a role
    - Could gather data on “rarer” characters or character combinations to support this acknowledgement

Work Distribution

We divided the work based on stages and individual problems, where two people work on each problem and we work on the Introduction and Conclusion parts as a group.

- Introduction - Together
    - Background
    - Related works / papers
    - Early Data Exploration (summary statistics, attributes of solution words, etc)
- Difficulty & Results Distribution - Yuheng, Tong
    - Model Construction - Yuheng, Tong
    - Data / Equilibrium Analysis - Yuheng
    - Example Generation + Summary of Results - Tong
- Number of Reported Results - Carol, Sunil
    - Model Construction - Both
    - Data + Equilibrium Point Analysis - Carol
    - Example Generation + Summary of Results - Sunil
- Conclusion / Summary - Together
    - Report
    - Presentation

Resources

We will use the python library wordfreq from https://github.com/rspeer/wordfreq to determine the rarity/frequency of each word.