Trexquant Interview Project (The Hangman Game)

Algorithm Author: Jason Chen
Overall Success Rate on Test Set: 61.1%

Before the algorithm:

Following the analysis, the subsequent findings were deduced:

- Almost every word contains at least one vowel.
- Among these vowels, the most frequently occurring are in the order of e, i, a, o, and u.
- As demonstrated in the plot, the majority of words comprise 20% to 55% vowels relative to their length.

These findings suggest the following strategies:

- Predict vowels at the beginning.
- Predict vowels according to their frequency order: e, i, a, o, u.
- When provided with a word, continue predicting vowels until the proportion surpasses 20%.
- Once the proportion of vowels exceeds 50% in a given word, refrain from predicting additional vowels.

Algorithm

1. At the beginning, guess vowels until the proportion surpasses 20%.
2. Proceed to guess letters based on words that match the clean_word pattern across the entire length. Retrieve qualified words, count the occurrence of each letter in every word from the all_word_dictionary, identify the most frequently occurring letter, and then guess it.
3. Repeat stage 2, but replace "entire length" with "half of its length".
4. Repeat stage 2, but replace "entire length" with "one-third of its length".
5. If no word matches in the training dictionary, default back to the ordering of the full_dictionary.

Notes:

- The all_word_dictionary is a dictionary where each key stores different lengths, and its corresponding value stores the parsed words from every word in the full_dictionary.
- The full_dictionary is a list of words provided by the host.