Wordle Report

Michael Allen, 2/25/2022

Michael Allen
MEGN 200
Brodsky – Section A 2022

Introduction

Wordle is a popular word game that utilizes basic logical conditions to generate a valid 5 letter word, which a player has 6 attempts to guess. By analytically observing patterns found in the game's dictionary – an optimal list of guesses may be made to best solve this pseudo-random puzzle. Given the game's simple nature, a text-based version of Wordle can be coded in many coding languages (including MATLAB) with nominal experience coding.

Background and Problem Statement

An ideal first guess should be taken to best solve a Wordle puzzle. This is a guess that does not have repeated letters and includes letters most likely to be found in any word included in the game's dictionary. Finding the best guesses requires a scoring algorithm to be applied to all possible words that may be guessed, and subsequently giving each word a score. These scores may be sorted into a rank, that will fully define the most informative first guesses - provided the likelihood of the top guesses sharing letters with an unknown word is highest.

Part 1

The first step towards solving this problem is to visualize the distribution of word lengths (should Wordle be extended to non-five letter words). This is done in MATLAB by first scanning a dictionary, then counting the lengths of each word and sorting them accordingly. The results of this process for the Words With Friends (WWF) and Merriam-Webster (MW) dictionary are as follows:

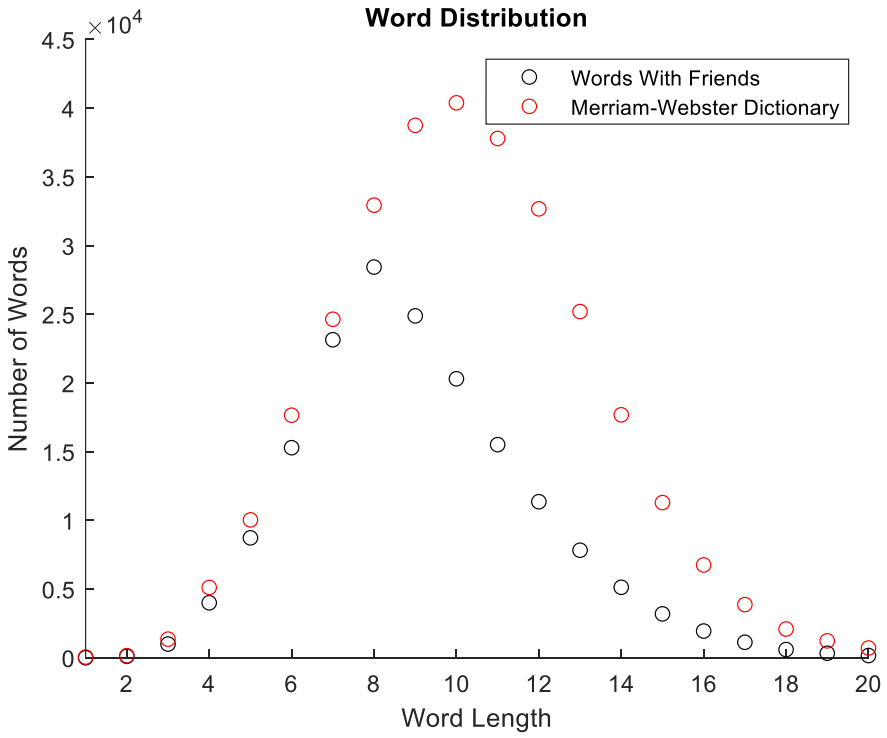


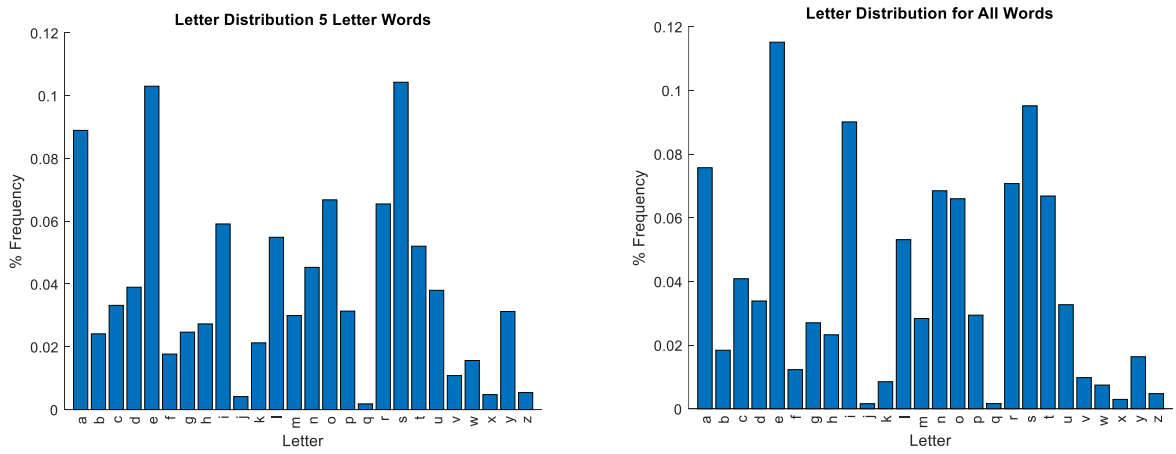
Figure 1: Word length distribution for Words With Friends and Merriam-Webster dictionaries

Note that the most common words are 8 letters long (WWF) and 10 letters (MW) respectively. Due to formatting in the text-based MW dictionary, entries for 1 and 2 letter words appear to be excessively large – ie. 56 entries in the 1 letter word category, despite only one ("I") being valid. Despite this, the overall distribution was not affected by this error in a significant manner.

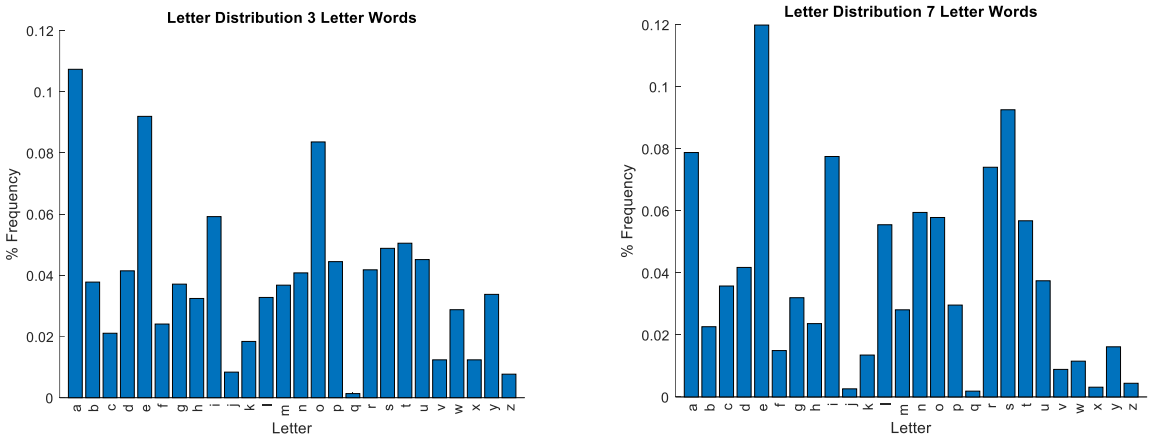
Analyzing the two distributions, it is apparent that not only is the average word larger in the MW dictionary, but there are also substantially more words overall. Observing the relative peaks of each distribution, MW dictionary had a peak of ~40.4-thousand words, whereas WWF had only ~28.4-thousand. This gap in dictionary size may be attributed to selective word choice for the Words With Friends game (as some words may be too obscure, illegitimate, or invalid for the syntax of the game – ie. hyphenated words). Cutting out hyphenated words and game syntax (max length of playable words) may also be to blame for the shorter average words in the WWF dictionary, as hyphens were counted as 'letters' in the word length algorithm.

Part 2

After defining possible lengths for each word, letter frequency was analyzed (as this will become the standard for scoring in the next section). Given Wordle is a 5 letter word game, the scoring standard is based on the distribution for 5 letter words. However, for speculative purposes, letter distribution was also analyzed for 3 letter words, 7 letter words, and all words in the WWF dictionary. Below you may see letter distribution for 5 letter words [left] in comparison to all words in the WWF dictionary [right]



Figures 2 + 3: Letter Distribution for 5 letter words and All words respectively (both WWF Dictionary)



Figures 4 + 5: Letter Distribution for 3 letter words and 7 letter words respectively (both WWF Dictionary)

Observing the distribution of these words, the following table may be constructed which displays the top 3 most common characters for each length category and their relative % Frequency.

Word Length	3 Letter Words			5 Letter Words			7 Letter Words			All Words		
Top 3 Characters	a	e	o	s	e	a	e	s	a	e	s	i
% Frequency (descending)	10.73	9.19	8.36	10.43	10.30	8.89	11.98	9.25	7.88	11.50	9.51	9.01

*Table 1: Top Letters for each Word Length *(WWF Dictionary)*

From the table, it may be observed that the most common letter is “e” – with all words consisting of 11.5% “e”. The second most common letter is “s”, which appears to make up 9.51% of all words. This is congruent with 7 letter words, as “e” and “s” do not deviate in percentage by more than .26% each. An interesting note is that for five-letter words, these two characters are very close with “s” appearing .13% more often. Strangely, 3 letter words do not contain “s” as a most-common letter (although a close runner up as 5th). The third most common letter for all words was “i” at 9.01%, which was not shared with any of the other three categories (whose third letter was “a”). “a” was only most common in three-letter words and was the third most common in 5 and 7 letter words. Provided this unique distribution, the dynamic of optimal guesses would change depending on the word length defined for Wordle. Provided the game is played with 5 letter words, it can be assumed that the optimal guesses will include letters “e”, “s”, and “a”.

Part 3

Multiplying a word's character content by respective scores was a straightforward and efficient way to score words. Sorting the scored words in descending order results in the word list to the right [Table 2]. All the top words on the list include "e", "a", and "s." This is congruent with the prediction given in Part 2 of the report. Looking for optimal words on CNN and WIRED articles, it may be seen that these words (or words like them) are consistently listed as good starting words – including the top result 'arose' being cited as a good starting point alongside 'soare' (Asmelash). An important note may be that the location of these letters was not considered in the MATLAB simulation. This detail was highlighted by both publications – "There are more solutions where the letters in *soare* are already in the correct positions compared to *arose*, [a fan told CNN]" (Asmelash). These sites also mentioned the importance of including letters 'E, A, R, I, O, T, N, [and] S' (Guinness) as they were most common for Wordle. It is possible that discrepancies in letter occurrence caused the variation in optimal guesses, as WIRED listed 'notes, resin, tares, [and] senor' (Guinness) as the best starting words. While there are some similarities in character composition, there is a clearer emphasis on the letter "r". Furthermore, these words may have been selected given the optimal following guesses, as these second guesses will "tick off any remaining letters in the top 10" (Guinness). A final note may be that the Wordle dictionary is not the same as the one used for this report, meaning that variations in the most common letters is possible.

Top 20 Starting Wordle Words
1. 'arose'
2. 'arise'
3. 'raise'
4. 'serai'
5. 'aloes'
6. 'arles'
7. 'earls'
8. 'lares'
9. 'laser'
10. 'lears'
11. 'rales'
12. 'reals'
13. 'seral'
14. 'stoae'
15. 'aster'
16. 'rates'
17. 'stare'
18. 'tares'
19. 'tears'
20. 'aisle'

Table 2: Optimal Starting Words, Wordle

Part 4

After defining the best guesses for a word, programming the text-based game became the primary focus. To accomplish this task, the following programming logic [Fig. 6, page 5] was implemented – this logic is representative of a way to accomplish a basic (and most likely non-optimal) version of the game. One of the greatest challenges was properly defining correct guesses, without incorrectly categorizing a word. This was primarily tested using words with two or more of the same letters, as these scenarios saw the most incorrect categorization of guesses (as the logic would properly categorize the first letter, then incorrectly categorize the second). The primary cause of this problem may be attributed to an incorrect ordering of conditional statements when categorizing characters in a word. For example, checking if a letter is in the correct spot first makes the loop vulnerable to skipping conditional statements that would check for duplicate letters in the word (both correctly and incorrectly guessed).

To solve this issue, the best method was to first determine if a letter was in the word; If this case wasn't met, the letter would be immediately appended to the incorrect character vector. By processing the guess in this manner, incorrect categorizations were omitted in the final product.

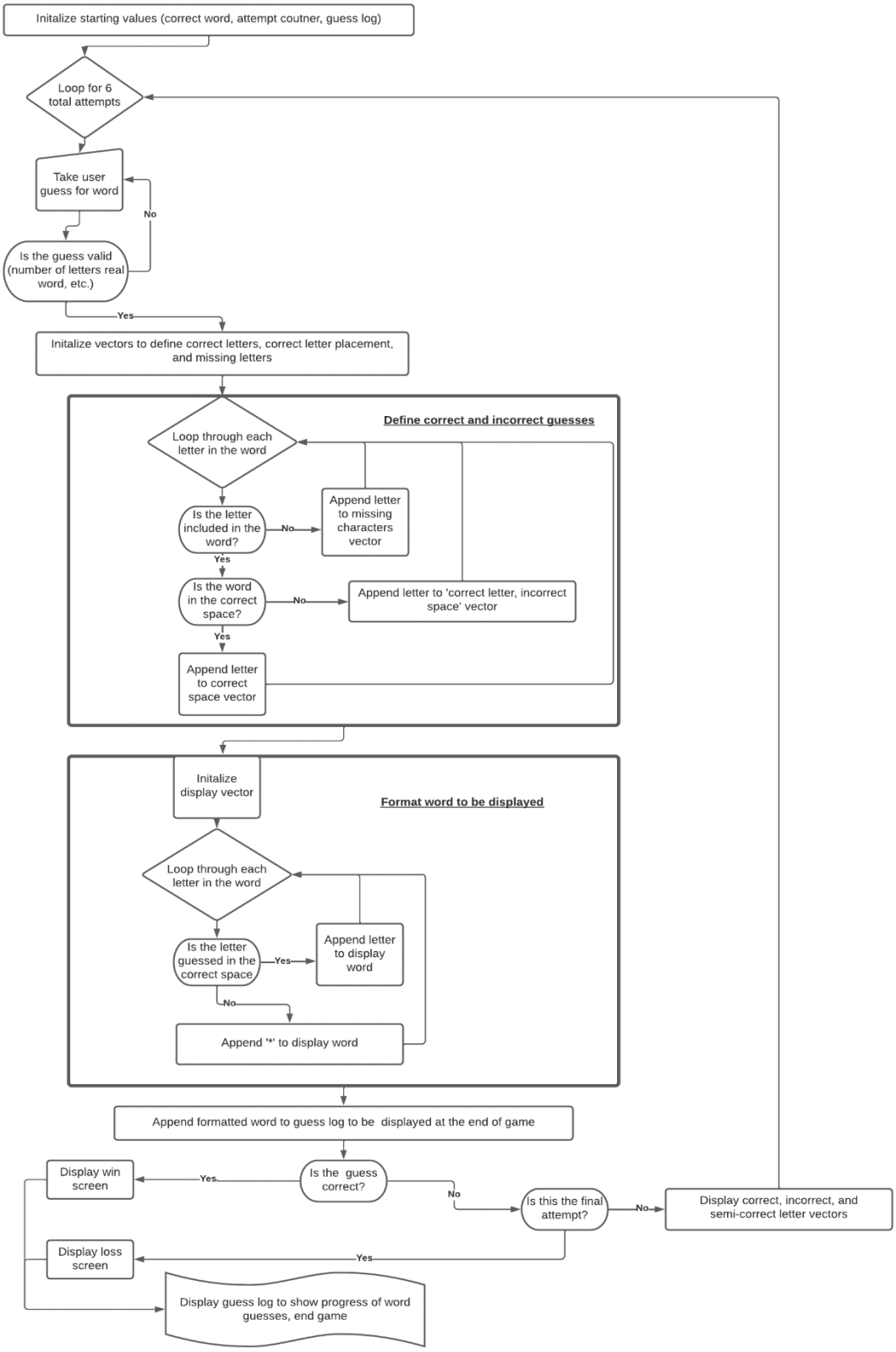


Figure 6: Wordle Logic Flowchart

References

Asmelash, Leah. “Want to Win at Wordle? Use These Words.” *CNN*, Cable News Network, 1 Feb. 2022, https://www.cnn.com/2022/02/01/us/wordle-top-strategies-winning-words-cec/index.html.

Guinness, Harry. “The Best Starting Words to Win at Wordle.” *Wired*, Conde Nast, 13 Jan. 2022, https://www.wired.com/story/best-wordle-tips/.