A Word Difficulty Classification Research Based on K-Means Method

Chengxi Wei [1, *], Di Kuang [2], Keran Xu [1]

[1] School of Future Science and Engineering, Soochow University, Suzhou, China, 215200

[2] School of Mechanical and Electrical Engineering, Soochow University, Suzhou, China, 215100

* Corresponding Author Email: chengxiweiclaire@gmail.com

Abstract. Wordle is a globally popular game and many researchers have conducted research on its game mechanics. However, few studies have explored the influence of attributes on the difficulty of the word. Therefore, this paper uses K-means algorithm to classify word difficulty based on certain word attributes. In the paper, the main attributes character repeat times, the presence of "th" or "er", the initial letter (s/c/a/t), and the final letter (e/y/r/t) that affect word difficulty are selected, and a comparison is made regarding the number of difficulty categories, and the most appropriate number of categories is three. Finally, it has been validated by Spearman that this method possesses strong reliability.

Keywords: Wordle, Word Attributes, K-means, Spearman.

1. Introduction

Wordle, a web-based word-puzzle challenge, is now in vogue on social media sites like twitter. To solve the puzzle, player need to guess the correct five-letter word within six tries and each try will provide player with some information hinting the correct answer. To be specific, each letter of the word is marked with a different color: grey means “not in the answer at all”; yellow means “in the answer but wrong position”; green means “in the answer and correct position” [1]. Moreover, Wordle offers “Hard Mode” to those who want to challenge themselves in which once a correct letter is revealed, it must be used in the next guess. Figure 1 is a screenshot of our attempt on February 17, 2023.

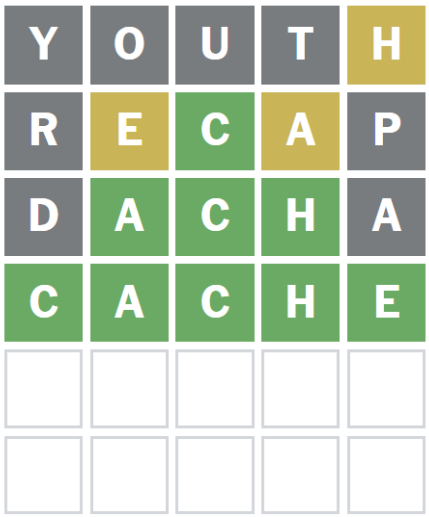

Figure 1. Guessing the word cache from February 17, 2023

A well-designed puzzle can let people dig deep into its underlying logic and summarize the game experience. Since Wordle went viral around January 2022, this word-puzzle challenge has drawn numerous people’s attention to the word-formation which is an intriguing and profound study field.

Researchers hypothesize that there must be underlying rules governing this word puzzle [2]. Research like Sidhu [3] attempted to find the best starting word from a linguistic perspective, while other researchers, such as Anderson and Meyer [4], prefer using mathematical methods like machine

learning to determine the optimal human strategy for solving wordle. There are also puzzle lovers sharing their own experiences. However, there is little work researching the word attributes that contribute to word difficulty. This paper aims to fill this gap.

The paper attempts to utilize machine learning to assign a difficulty level to each word based on its attributes. It explains the connection between the attributes of a word and its difficulty level, which will provide a dependable reference point for Wordle operators and technical support for future research endeavors.

2. Methods

2.1. Dataset and data preprocessing

The game: Wordle studied in this paper is the official version developed by Josh Wardle and published by the New York Times Company. Clones and variations of the game are not included.

A Twitter account (@WordleStats) collects all the results shared on Twitter and posts a bar chart illustrating the score distribution and the number of hard mode players daily [5]. This article utilizes data provided by this Twitter account from January 7, 2022 to December 31, 2022, which contains the date, word of the day, the number of people reporting scores that day, the number of players on hard mode, and the percentage of people that guessed the word in one try, two tries, three tries, four tries, five tries, six tries, or could not solve the puzzle (indicated by X) to facilitate the research.

2.2. Word attribute selection

It is widely accepted that meaning, part of speech, spelling, and pronunciation are all common attributes of a word. These attributes all together affect people's results when playing the game Wordle. But taking every last one of them into consideration can be very difficult. What's more, it is neither necessary nor scientific to do so because not every attribute distributes much to the final prediction result.

Based on the research findings of predecessors, repeated characters, letter frequency [6], the number of vowels [7], and the use of two-letter sequence have the most intimate relationship with difficulty level of words. In this paper, Spearman correlation [8] is used to simplify the problem and reduce disturbances from the low relevant attributes. After analyzing the word attributes with the percentage of scores reported, five attributes that possess strong correlation are selected.

Pearson correlation is not appropriate for this study because it requires the variables to meet the following requirements: Both variables are continuous variables； Both variables come from populations that are normally distributed or close to normal unimodal symmetric distributions； The variables should be paired； Both variables are linearly related.

In this study, the two variables, attempts number and word attributes, clearly do not meet these requirements, so Pearson's correlation coefficient cannot be used for correlation calculation. Compared to Pearson correlation, Spearman correlation can be used in non-linear relationships. Additionally, Spearman correlation has a wider applicability and produces high accurate results.

Spearman correlation coefficient is often used in measuring the correlation between ordinal variables. In this study, the number of attempts is obvious an ordinal variable. As long as the importance order of word attributes is determined, word attributes can also be treated as an ordinal variable. Therefore, this study uses Spearman correlation to analyze the correlation between word attributes and the attempts number.

Since the impact of word's attributes on word's difficulty is unknown and for the sake of simplicity, it is reasonable to assume that each attribute contributes equally. That is to say, each attribute possesses the same importance. During calculation, it can be found that the results of the Spearman correlation analysis do not differ significantly when their order switch, which verifies the assumption.

Therefore, Spearman correlation is chosen to investigate the correlation between the word's difficulty and attributes in this study.

The principle of Spearman correlation is as follows:

$$r_s = 1 - \frac{6\sum(x_i - y_i)^2}{n(n^2-1)}, \tag{1}$$

Where x_i and y_i are the corresponding sizes of the two variables respectively, n is the sample size, n>20 can be tested with the t statistic:

$$nt = r_s\sqrt{\frac{n-2}{1-r_s^2}} \tag{2}$$

After analyzing the word attributes with the percentage of scores reported, four attributes that possess strong correlation are selected as shown in Table.1.

Table.1. The selected attribute

repeat	Character repeat times
th	Whether it contains the two-letter sequence “th”
er	Whether it contains the two-letter sequence “er”
first	Whether the first letter is s/c/a/t
last	Whether the last letter is e/y/r/t

2.3. K-means Clustering

K-means is commonly used for datasets whose dimensions and values are small and continuous, such as grouping similar items from a random distributed set. The schematic diagram of K-means method is shown in Figure 2.

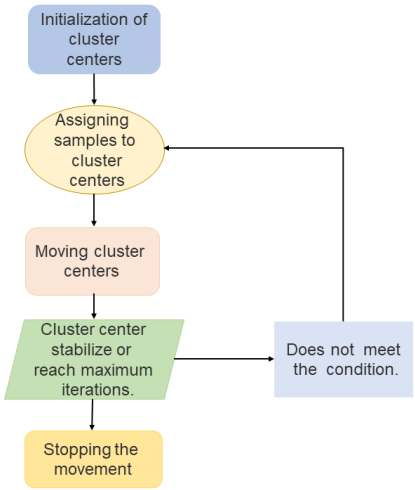

Figure 2. The schematic diagram of K-means

The main steps for the K-means clustering algorithm are as follows:

Step 1: Initialization of cluster centers

Randomly select K data points as the initial cluster centers, which can come from any K points in the dataset or be selected through other methods.

Step 2: Assigning samples to cluster centers.

Assign each data point to the category where its nearest cluster center is located. Specifically, for each data point, calculate the distance between it and each cluster center, and then assign it to the category where its nearest cluster center is located.

Step 3: Moving cluster centers.

Recalculate the cluster center for each category, that is, use the average of the coordinates of all data points in that category as the new cluster center.

Step 4: Stopping the movement.

Repeat steps 2 and 3 until the cluster centers no longer change significantly or the predetermined number of iterations is reached.

K-means clustering model guarantees convergence and can warm-start the positions of centroids. Thus, this algorithm is perfectly suitable for this paper's study. Given the aforementioned reasons, this paper finally decides to use the distribution of the reported results as classification criteria to construct a K-means clustering model [9]. It should be noted that the K-means algorithm uses an iterative method to obtain a local optimal solution, so multiple runs are needed in practical applications to select the optimal result.

3. Model Building and Problem Solving

3.1. Word Difficulty Assessment Model Based on K-means Clustering

Here is how K-means is implemented [10]:

Step 1: Set the sequence of all word's score results as $\{x^{(1)}, \ldots, x^{(m)}\}$.. Here, $x^{(i)} \in R^7$..

Step 2: Initialize cluster centroids randomly (denoted by $\mu_1, \mu_2, \ldots, \mu_k \in R^7$).

Step 3: Repeat until convergence: {

For every i, set.

$$c^{(i)} = \arg\min_j \| x^{(i)} - \mu_j \|^2 \tag{3}$$

For each j, set.

$$\mu_j = \frac{\sum_{i=1}^{m} 1c^{(i)} = jx^{(i)}}{\sum_{i=1}^{m} 1c^{(i)} = j} \tag{4}$$

}

To determine the appropriate number of levels for classification, SSE (Sum of Squared Errors) is utilized.

$$SSE = \sum \left(y_i - \overline{y}^2\right), \tag{5}$$

here, y_i represents the actual value of the i-th observation, $\overline{y}$ represents the predicted value of the i-th observation.

The ideal number of classifications can be identified by locating an inflection point on the graph of the sum of squared errors. By observing the inflection point in Figure 3, it is judged that the inflection point should be at 2.

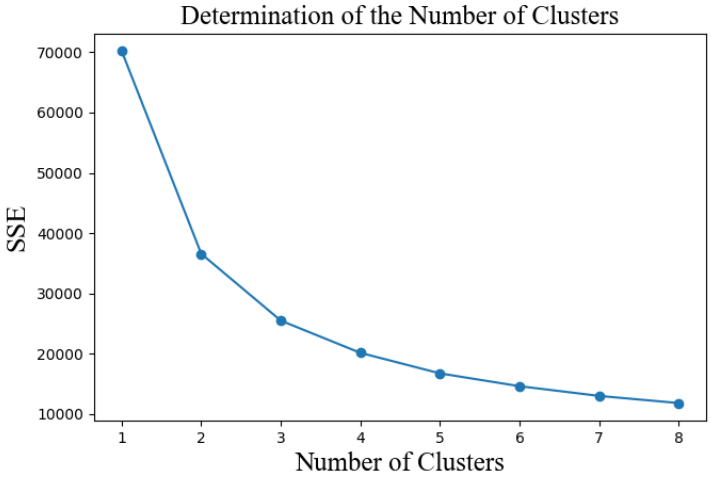

Figure 3. Determination to the number of clusters

Based on preliminary judgment, if the words are only divided into two categories, it is too broad to become a good classification method. However, in order to make the inference process more rigorous and complete, the cluster centers through K-means clustering when k equals 2 is still provided as follows.

Table.2. The cluster center when k=2

	Clustering	
	1	2
1 try	0.6	0.3
2 tries	7.9	3.2
3 tries	28.3	15.5
4 tries	34.7	30.6
5 tries	19.9	28.5
6tries	7.3	16.9
7tries	1.2	4.9

The Table.2 illustrates that in the first four attempts, the cluster centers of Class 1 are more delayed, while in the last three attempts, the cluster centers of Class 2 are more delayed. The distribution of cluster centers does not have a universal regularity at the same number of attempts, which verifies the inference that difficulty patterns cannot be summarized when k=2.

In addition, if there is only little difference between the distribution of attempts number and the mean square error of the two cluster centers when classifying specific words, the judgment result can be extreme. In other words, the judgment of word classification is either/or, thus the probability of misjudgment is very high.

The above evidence is sufficient to indicate that word difficulty should be classified into more than two categories. So, k = 3 is more reasonable. Thus, the value of k should be 3, which means the difficulty levels should be divided into three.

Step 4: The clustering centers are shown in Figure 4.

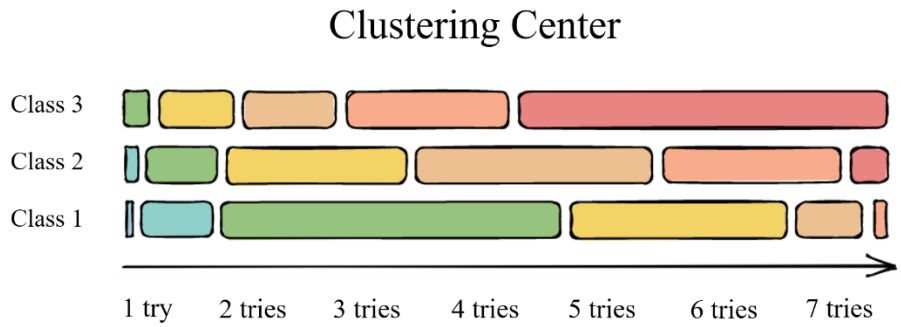

Figure 4. Final clustering centers

As the number of answers increases, the frequency of attempts number is gradually increasing, making the third category hardest, the second category moderate and the first category easiest. So the third category is the most difficult, the second is the second, and the first is the easiest. The levels are arranged the in the order of difficult to easy, as A, B, C.

Mean squared error is a metric that quantifies the difference between specified or predicted value and actual value. The smaller mean squared error is, the closer the specified or predicted value is to the actual value. In order to ascertain the difficulty level of a given word in this task, the paper computes the mean squared error between the reported percentage and cluster centers of the three difficulty levels and assigns the word to the difficulty level with the lowest mean squared error. Solution words are shown in the Figure 5.

(a) Level-A (b) Level-B (c) Level-C

Figure 5. Solution words in different difficulty levels

3.2. Associate Word Attribute with Difficulty Level

This paper analyzes the word attributes with the percentage of scores in all three levels using Spearman correlation. Take level-C as an example. Following is a heat map (Figure 6) of its correlation coefficient.

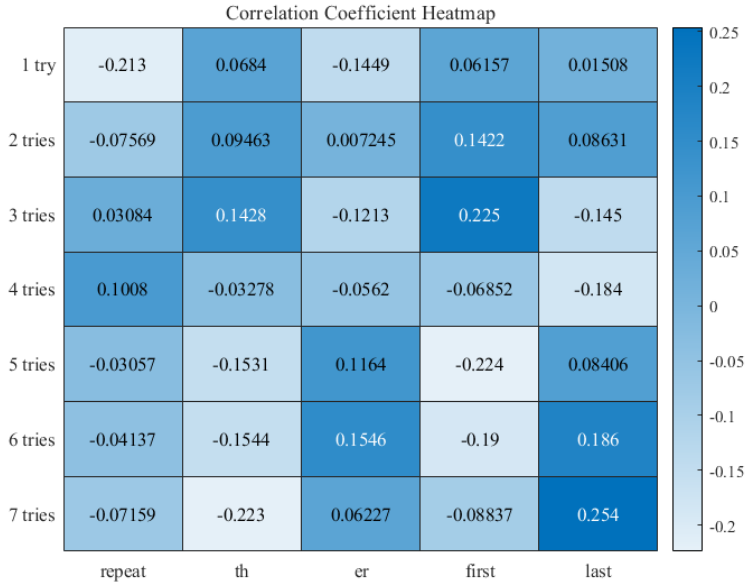

Figure 6. Spearman's rank correlations between attributions and score percentage under level-C

The same possess is repeated in every difficulty level, and the negative and positive correlation results as shown in Figure 7.

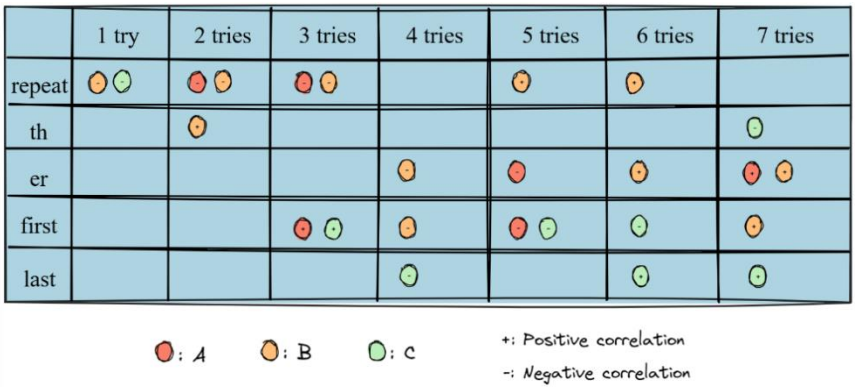

Figure 7. Correlation results between word attributes and percentage of each attempt under different difficulty levels

It can be seen that the attribute “repeat letter” always exhibits negative correlation with hypo-tries and positive correlation with multi-tries. So “repeat letter” can be preliminarily classified to high difficulty level (either A or B). Further looking into the result, this paper discovers that the "repeat letter" is more strongly associated with multi-tries in the level-B as opposed to the level-A.

Performing the same analysis to each attribute, the result is shown as Figure 8.

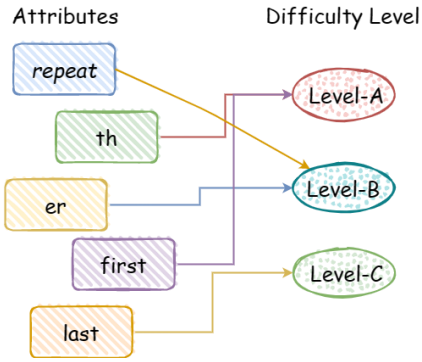

Figure 8. Attribute analysis

3.3. Model verification

In order to verify the accuracy of the difficulty classification model, another four words along with their score distribution is picked. If the difficulty level drawn from the score distribution corresponds with the level drawn from the attribution, the accuracy can be considered as good.

In pursuit of heightened accuracy in the validation process, this paper embarks upon an exhaustive search of the Internet for 100 new words. In order to render the verification process more straightforward and to introduce an element of chance in the selection of words, the 100 words are numbered and four of them are selected by generating random numbers. As a result, the words tough, cider, above, and ruddy are chosen.

Take the word RUDDY as an example. The score percentages of (1, 2, 3, 4, 5, 6, X) for RUDDY is (0, 2, 23, 33, 24, 13, 4).

Before calculation, it is necessary to introduce mean squared error (MSE) first. Mean squared error (MSE) is a commonly used measure of the average squared differences between the predicted and actual values in a regression analysis. It measures the average of the squared differences between the predicted and actual values of the dependent variable. The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero indicate a better fit.

The formula for calculating the MSE is:

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2, \tag{6}$$

Where m is the number of observations, y_i is the actual value of the dependent variable for observation i, and $\hat{y}_i$ is the predicted value of the dependent variable for observation i.

The MSE is a useful metric for evaluating the performance of a regression model, as it provides a measure of how well the model is able to predict the values of the dependent variable. It is commonly used in machine learning and statistics to compare different models and select the best one based on the lowest MSE value.

To determine the word’s difficulty level, this study calculate the mean squared error (MSE) between the set of word resolution times and the final clustering centers of level A, B, C respectively. The results are as follows: 143.43 for level-A, 68.14 for level-B and 421 for level-C. The MSE with level-B is the smallest, which states the difficulty level of RUDDY is B.

In order to explore the accuracy of the assessment model, the correlation between attributes and difficulty level is constructed to make a second judgment on the difficulty of the word RUDDY.

RUDDY has one attribute: two repeated d, which preliminarily rank RUDDY ‘s difficulty level as B. Table.3 and Table.4 show the specific process of classification.

Table.3. Mean square error.

Word	Mean Square Error			Difficulty Level
	with A	with B	with C	
tough	45.14	207.86	601.57	A
cider	148	60.43	366.43	B
above	129.71	79.57	478.14	B
ruddy	143.43	68.14	421	B

Table.4. Difficulty classification

Word	Related Attribute	Category	
		Preliminary Inference	Final Result
tough	first	A	
cider	er, first, last	B, A, C	B
above	first, last	A, C	B
ruddy	repeat	B	

4. Conclusion

This study utilizes K-means clustering to classify word difficulty. The model shows high accuracy and suitability after thorough testing. However, the paper lacks quantified correlations and visualized data. Recommending listing formulas based on correlation coefficients to calculate a final correlation score for better data interpretation and model accuracy. Future research can explore additional techniques and visualization methods for comprehensive representation, improving the understanding of word difficulty and model performance.

References

[1] Wordle. (2023). Retrieved February 17, 2023, from https://en.wikipedia.org/wiki/ Wordle.

[2] de Silva, Nisansa.” Selecting seed words for wordle using character statistics.” arXiv preprint arXiv:2202.03457 (2022).

[3] Wordle – the best word to start the game, according to a language researcher. (2023). Retrieved February 17, 2023, from https://theconversation.com.

[4] Anderson, Benton J., and Jesse G. Meyer.” Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning.” arXiv preprint arXiv:2202.00557 (2022).

[5] Twitter. Available at: https://twitter.com/WordleStats?s=20 (Accessed: April 9, 2023).

[6] N. De Silva,” Selecting Optimum Seed Words for Wordle using Character Statistics,” 2022 Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 2022, pp. 1-6, doi: 10.1109/MERCon55799.2022.9906176.

[7] Tankersley, Karen. The threads of reading: Strategies for literacy development. ASCD, 2003.

[8] Croux, C., Dehon, C. Influence functions of the Spearman and Kendall correlation measures. Stat Methods Appl 19, 497–515 (2010). https://doi.org/10.1007/s10260-010-0142-z

[9] Hartigan, J. A., and M. A. Wong. “Algorithm AS 136: A K-Means Clustering Algorithm.” Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, 1979, pp. 100–08. JSTOR, https://doi.org/10.2307/2346830. Accessed 20 Feb. 2023.

[10] A. Chadha, Distilled Notes for Stanford CS229: Machine Learning, https://www.aman.ai, 2020, Accessed: Aug 1, 2020.