Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2315668

The Digital Secret Behind Wordle

Summary

Words play a significant role in language, without vocabulary nothing can be conveyed. The complexity and practicality of the word create its own high value. For more than a year, Wordle has been a concise and entertaining puzzle game with a strong reputation. Players should predict a five-letter phrase in six attempts or less in order to complete the puzzle, and each estimate results in feedback. At the request of the New York Times, we will analyze the data set provided and model **word properties** to make constructive recommendations for Worlde.

Several models are established: Model I: Multiplicative Power Segmented Regression Prediction Model; Model II: Alphabetic Quantization BackPropagation(BP) Neural Network Model; Model III: Word Difficulty Coefficient-Based K-means Clustering Algorithm Model. Before models are established, we **preprocessed** the data to obtain no abnormal data and controlled for rounding percentage errors in the raw data.

For Model I: By visualizing data, we constructed Figure 4 and found that it shows a rapid increase to a peak, then a gradual decrease and a gradual stabilization. Given that partitioning could effectively address this issue, we used the **Multiplicative Power Regression Prediction Model** to account for the temporal variation in stated findings. Based on this fitted model, we obtained the prediction interval **[14053,17858]** for March 1, 2023. The number of repetitions of the letters in the attached words was used as different attributes of the words, and the words were divided into three groups, and the percentages. Thus, we derived the results, which are shown in Figure 8.

For Model II: we prefer to predict percentage of words that are relevant for a given date in the future. Combining the word data provided by Twitter for the last year of 2022, we can predict the relevance percentage of words using the alphabetic quantization into numbers method and the **BP neural network model**. The word 'EEERIE' was brought into Model II to obtain its percentage of successful solution attempts on March 1, 2023 (see Table 3). The model was tested to have correlations of nearly 89% of the predicted results, and we have sufficient confidence in its stability and accuracy (Figure 11 illustrates).

Model III: The model was based on word difficulty rating coefficients for **K-mean clustering analysis**, which resulted in a matrix of central coefficients for the three sets of clusters (see Table 4). The clustering labels corresponding to each group of sample data were then derived by the Euclidean algorithm, and the word classification was completed. Next, the words were analyzed with their attributes to assign difficulty levels to the clusters. Finally, the difficulty rating coefficient matrix of 'EERIE' was brought into Model III to obtain the difficulty of the words (see Table 5). And the accuracy of the model was tested by the Euclidean algorithm and was **94%**.

Lastly, a **sensitivity analysis** of the mathematical expectations demonstrates that our model is not sensitive to changes in the estimate of the number of individuals reporting in the upcoming three months. In other words, our algorithm can be used to predict the number of reports in the near future. In addition to this, we wrote to the New York Times to provide recommendations based on the results of this study.

Keywords: Wordle; Multiplicative Power Regression Prediction; BP Neural Network; K-means Clustering Algorithm; sensitivity analysis

Contents

1 Introduction

1.1 Problem Background

Wordle is an easy-to-play charades game and it has become popular worldwide on social platforms in recent years. The rules are simple, players need to guess an English word consisting of 5 letters in 6 chances, the letter that the player guesses and is in the correct position will be presented with a green background, a yellow background means the word contains this letter but in the wrong position, and a gray background means the letter is not included in the word. It has two modes, normal mode, and hard mode. In hard mode, the player must use the correct letter found for subsequent guesses, as follows:

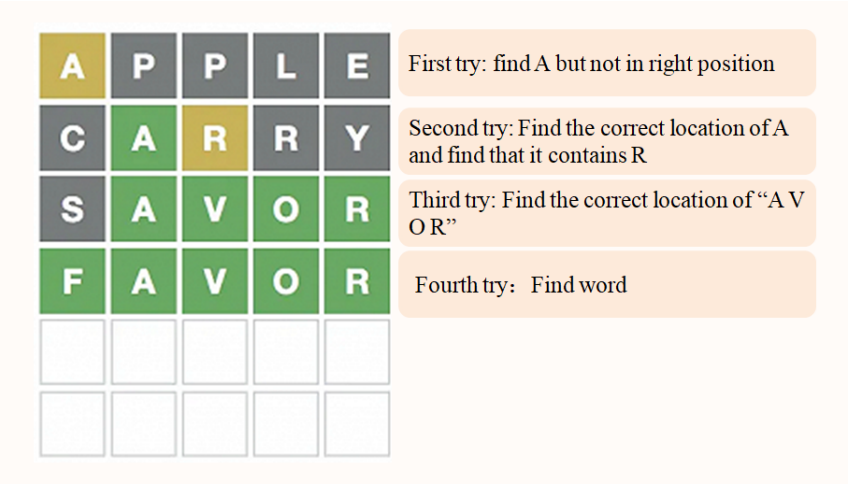


Figure 1: Examples of wordle games

The social sharing feature of wordle is what has really made it famous: a large number of players share their results through Twitter. And since its feature of updating one question a day, it is simple to accumulate the proportion of points for each word over time. At the request of the New York Times, it made sense to analyze existing wordle user performance data. By creating mathematical models based on the data, they might then improve their game designs by making more logical word selections.

1.2 Restatement of the Problem

The game wordle is simple, but the composition of the English alphabet is complex.Through in-depth analysis and research on the background of the problem, combined with the specific constraints given, the restate of the problem can be expressed as follows:

- Build a mathematical model to forecast the overall number that will be logged at a later time (March 1). Likewise, consider how to word characteristics impact the proportion of results.
- Based on the prior work, a mathematical model is built to predict the relevant percentage of game attempts on a future day
- Word difficulty classification for wordle games based on word attributes. And do a robustness test or sensitivity analysis on the model

- Analyze the explicit properties of the dataset based on the established model
- Considering the results obtained above, prepare a one to two pages letter of constructive suggestions about wordle and submit it to the New York Times.

1.3 Literature Review

The primary focus of this inquiry is on developing a model to investigate the connection between words and wordle score data. A traditional area of study in the humanities for a long time has been word characteristics. Particularly, **word nature**, **flexical meaning**, **pronunciation**, and **letter repetition** are the research objectives. The word characteristics that can be used in the model are the main topic of this part.

- First of all, based on the core properties of words as described in Mark's[5] book, and considering the constraints of the model requirements, words can be classified into: word meaning and alphabetical order
- Secondly, Word meaning is influenced by the alphabetic composition of words, but the presence of synonyms leads to a restriction in the use of word meaning to distinguish between categorized words, while the order of words is not affected
- Finally, The main things that can determine the alphabetical order of words are letter repetition, word nature and pronunciation. The repetition of letters directly determines the order in which words are formed, since it is impossible for certain letters to occur at the same time[1]. The lexical properties determine the letter combinations that must occur in a word[4]. For pronunciation, Abdulrahman [2]contend that each rhymal constituent in English has a maximum of two spaces, whereas an onset has a maximum of three slots.
- The following is a graphic representation of the planning space's assets and weaknesses:

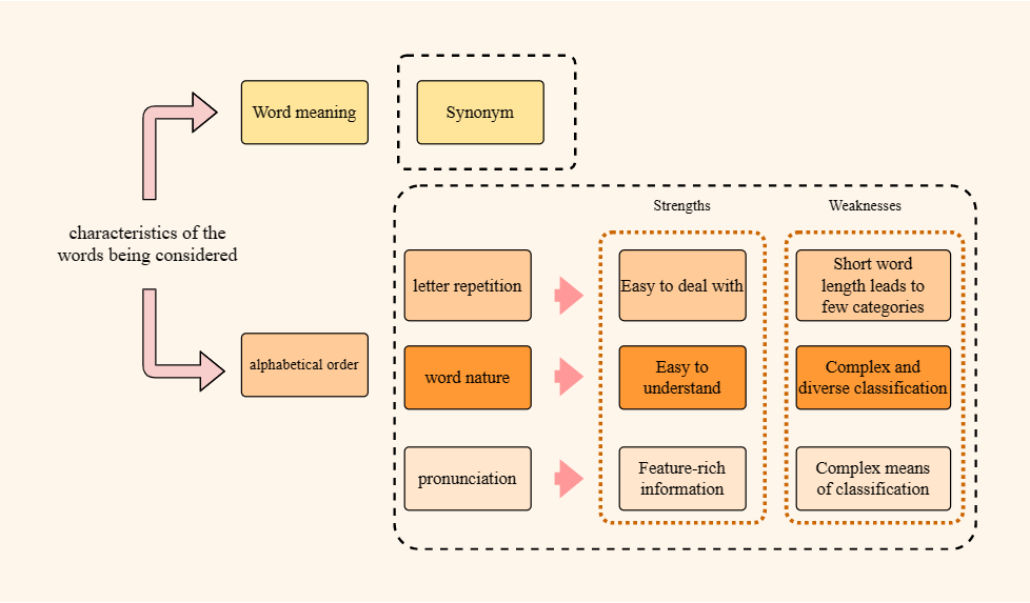


Figure 2: Literature Review Framework

1.4 Our Work

We need to build multiple models and solve known problems based on the given data, after analysis. Our work mainly includes the following:

1. Based on the scatter plot of wordle data, Multiplicative Power Segmented Regression Prediction Model was built.
2. Matrixing the words and building a BP neural network model to obtain predictions of score-related percentages.
3. Create a gradient of word difficulty coefficients and define difficulty for words by a K-means Clustering Algorithm model.

The flow chart is depicted in Figure 3 to prevent a complicated description and to naturally represent our work process:

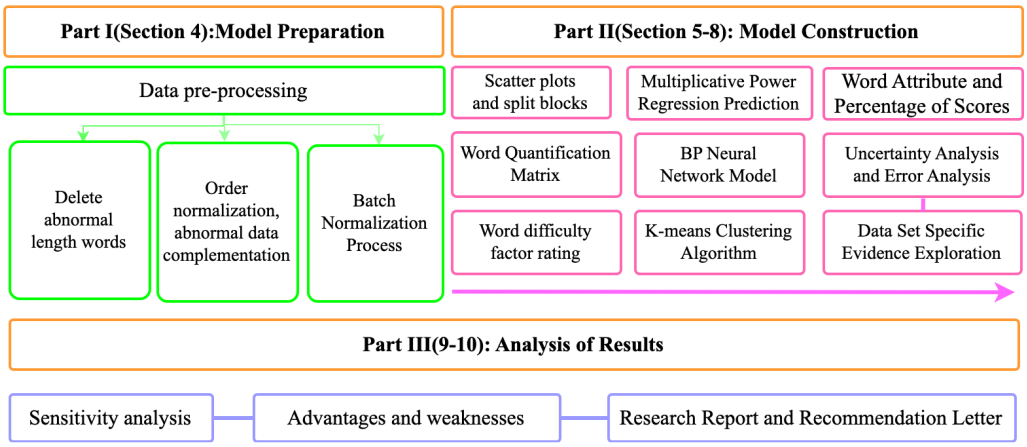


Figure 3: Flow Chart of Our Work

2 Assumptions and Explanations

Since real-world issues almost always involve a variety of intricate variables, we must first make sensible assumptions in order to reduce the complexity of the model. Each premise is followed by an explanation to support it, as follows:

Assumption **1: Ignore the differences in the physiological levels of the participants in the game: intelligence level, memory level.In addition, disregard any outside influences on the user, such as a solo game or a peaceful setting.**
Explanation:Both physiological and external factors may directly affect the results of the questions, which may lead to abnormal bias in the data, so we ignore these complicating factors

Assumption **2: The wordle word bank is large enough to include almost all words**
Explanation:If the vocabulary is too small, it will lead to repeatedly pushing out-of-the-box words like this situation, which indirectly leads to a different probability of each word appearing, so all words can appear by default

Assumption 3: **Assume that all data on Twitter is true and trustworthy**
Explanation:Data collected by WordleStats accounts are for statistical purposes only and are generally not subject to the possibility of falsification.

To make the research for each part easier to understand, additional assumptions are made. The proper places will be where these presumptions are addressed.

3 Notations

Table 1 provides a summary of some significant mathematical notations used in this essay.

Table 1: Notations used in this paper

Symbol	Description
x_c	Score percentage in hard mode
μ_β	Mean of percentage of scores in a group
δ_β^2	Standard deviation of the percentage of scores in a group
x_t	Contest number of time
y_n	Number of reported results
x_n	The corresponding elements in the matrix after word quantization
u_i	Number of neural network layers
y_m	The distribution of correct answer scores corresponding to each variable

*There are some variables that are not listed here and will be discussed in detail in each section.

4 Model Preparation

4.1 Data Cleaning

Examining the data files (2023_Problem_C_DATA.xlsx), both textual and numerical statistics are present in the collection, as we discover. We can quickly identify information that deviates from the game's regulations, such as words like "*rprobe*" and "*clen*" that do not conform to the length of five characters. Due to the fact that its words deviate from the fundamental rationale of data generation—that is, those produced by the wordle game—we chose to remove these two anomalous data.

In addition, we also remove all the blank cells and delete the "*Persent in*" because it only serves as an explanation

4.2 Form Data Pre-processing

We can see that this unprocessed data does not follow traditional visual reasoning and that it includes anomalous data that needs to be processed.

- There is an extreme error here, where the "reported results" in that row of "study" is more than 10 times different from the amount of data above and below, which

we consider as abnormal data. To maintain the sanctity of the data, we chose to alter the abnormal data from "2569" to "25690."

- The overall sequence does not match the visual logic. The top of the table grows older as it descends, so the part that is farthest from the present is at the farthest distance. To conform to the logic of reading from left to right, the appropriate table should have the past as the beginning and the present or future as the farthest conclusion. So we flipped the vertical order of the table
- Considering that the rounding error of the data itself causes the percentages to not sum to 100, we use the difference apportionment method to apportion the error at a distance of 100 to all data to achieve a reasonable effect.

4.3 Batch Normalization Process(BN)

Considering that the rounding error of the percentage data itself causes the percentages to not sum to 100, we use batch normalization(BN) for percentage data. In addition, the BN algorithm slightly decreases overfitting and lowers the learning rate needed for machine learning.

1. For each set of percentage data, the data percentage x_i is taken as input data and the mean value is calculated μ_β

$$\mu_\beta = \frac{1}{m}\sum_{i=1}^{m} x_c \tag{1}$$

2. Calculate the standard deviation δ_β^2 obtained from the data in the previous step

$$\delta_\beta^2 = \frac{1}{m}\sum_{i=1}^{m} (x_c - \mu_\beta)^2 \tag{2}$$

3. The result of normalization is $\widehat{x_c}$, where ε is a number added near to 0 to prevent the denominator from being 0.

$$\widehat{x_i} = \frac{x_c + \mu_\beta}{\sqrt{\delta_\beta^2 + \varepsilon}} \tag{3}$$

5 Multiplicative Power Segmented Regression Prediction Model

5.1 Data Visualization and Analysis

In order to explore the relationship between the number of reported results and time, and to create a forecast interval for the number of results reported on March 1, it can be visualized as a scatter plot. To normalize the time, we transform the time data into a single standard number, followed by a scatter plot with the **Contest Cumber** representing the time as the independent variable and **Number of Reported Results** as the dependent variable. This is depicted in the following figure:

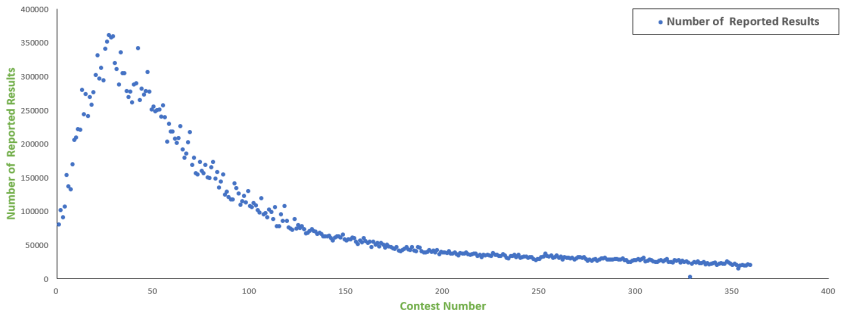


Figure 4: Scatterplot of reported results over time

From the above scatter plot, we observed that the number of results increases rapidly at the beginning. After the results' number reached a certain height, the trend is tapering down. This is probably due to the novelty of the game and the decline in popularity of the game. In order to describe the situation, our group used a piecewise function to divide it into two blocks and quantified separately. And the two blocks are the growth block and the mitigation block.

5.2 Sub-block Processing

By comparison, we split into two blocks with **Contest Cumber** at 30. As a result, the segmented block has a singular tendency, which is more conducive to reducing the interaction of different parts and improving the accuracy of the linear fit.

Growth Block

In the growth block, consider the case where x is an integer between 0 and 30. Through linear fitting, we get the regression equation of the growth block and the correlation coefficient R^2:

$$\begin{aligned} y_n &= 63562x_t^{0.5087} \\ R^2 &= 0.9556 \end{aligned} \tag{4}$$

Its corresponding image is shown in Figure 5

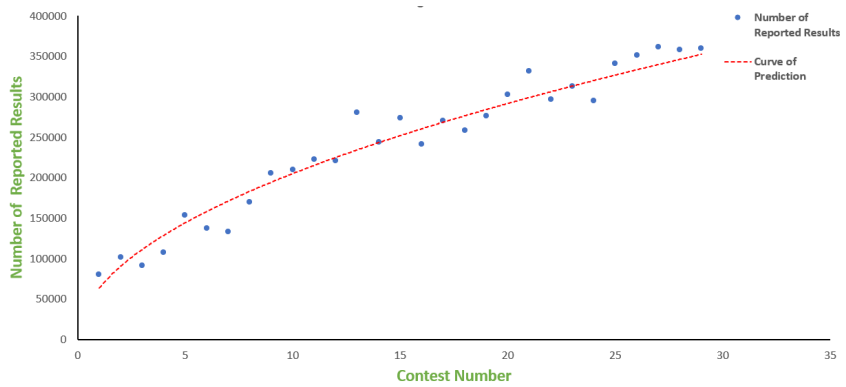


Figure 5: Growth Block

The regression equation is accessible because of its great goodness of fit. According to this graph, the function increases polynomially on the Y-axis and rapidly on the X-axis. Although y and x have a favorable correlation, as x rises, y will rise less quickly.

Mitigation Block

In the mitigation block, We delineate the interval of x as greater than 30. Through linear fitting, we get the regression equation of the mitigation block and the correlation coefficient R^2:

$$\begin{aligned} y_n &= 3 \cdot 10^7 x_t^{-1.252} \\ R^2 &= 0.94 \end{aligned} \tag{5}$$

Figure 6 displays the appropriate picture for it.

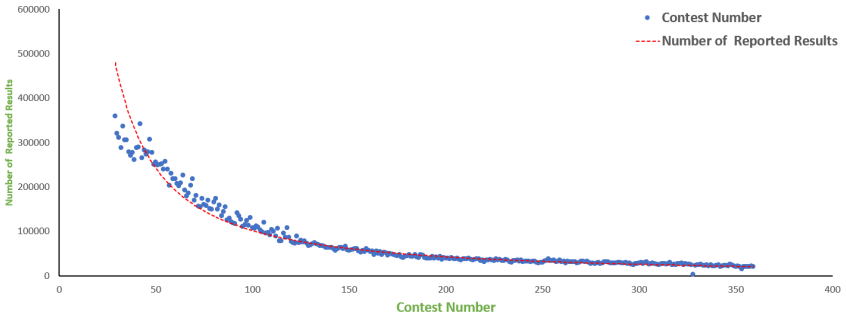


Figure 6: Mitigation Block

It is accessible because the regression solution has a high goodness of fit. This graph shows that the function increases at a constant rate along the Y-axis and an inverse rate along the X-axis. With a negative correlation between y and x, y converges to zero as x near infinity.

In conclusion, we can fit this piecewise function to the quantity of reported findings over time:

$$y_n = \begin{cases} 63562 x_t^{0.5087} & (0, 30] \\ 3 \cdot 10^7 x_t^{-1.252} & (30, \infty) \end{cases} \tag{6}$$

5.3 Multiplicative Power Regression Prediction Model

Create a margin of error for the number of results in Figure 6. Through computer processing, we calculate the average difference between the original data of each sample and the predicted data of the regression function, namely, the average error. Take $\frac{a}{1+w}$ as the lower bound of the predicted value and $\frac{a}{1-w}$ as the upper bound of the predicted value to create an error interval. And then we used that to predict 100 units backwards.

In the upper and lower equations above, a is the formula for the piecewise function when x_t in Figure 6. w is the average difference between the actual value and the predicted value of the function for each x, which is

$$w = \frac{\sum_{i=1}^{n} w_i}{n} \quad i \in \mathbb{N}^+$$

w_i is the difference between the true value of the reported quantity and the measured value at each point in time. With this regression prediction model, we obtain Figure 7.

From the above information, we can give a prediction interval for the number of reported results on March 1, 2023, which is **[14053,17858]**. The prediction number is **15729**.

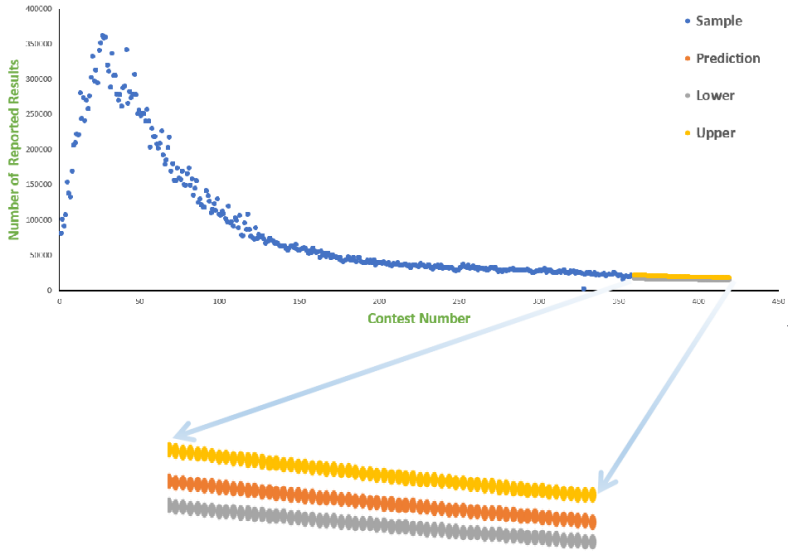
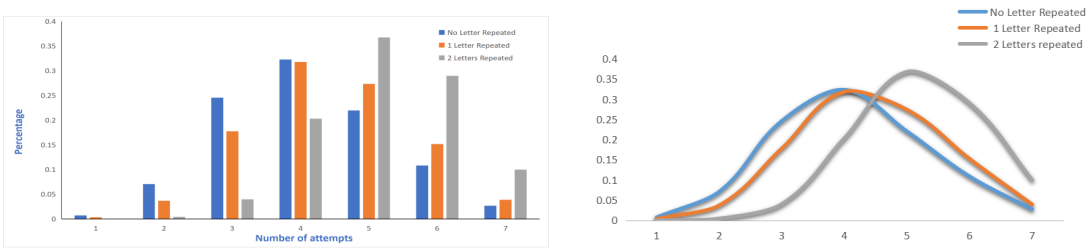


Figure 7: Regression prediction results

5.4 Word Attribute Interference Percentage of Scores

After looking through and evaluating the literature in section1.3, we made the decision to use the number of word letter repetitions as the feature value in order to establish a clear link between words and the percentage of results in the hard mode. To make it easier and more accurate to get the number of letter repetitions for each group of words, we wrote a **Java** program to quickly add feature information to each word. So we can divide the words in the data into three categories:words with no repetition between letters, one repetition between letters and two repetition between letters. After dividing the data into three groups, the number of correct guesses of players in each group was counted, and the percentage of results required by players in each group to guess the correct words was calculated in each group, and presented in the form of a bar chart and line chart in the Figure 8 and Table 2.



(a) Histogram(left) (b) Folding Line Chart(right)

Figure 8: Relationship between the number of repeated letters and the percentage of guesses

Combining these two images, we can observe the curve that repeats the letter two times is a positively skewed distribution. The curve that repeated the letter once did not differ much overall from the curve that did not repeat the letter, but the curve that repeated the letter peaked farther to the right. The farther to the right the spike, the more times it takes to guess a word correctly. Therefore we can infer that the higher the number of repetitions of each letter in a word, the harder the word is to guess.

Table 2: Relationship between the number of repeated letters and the percentage of guesses

Types of word features	1	2	3	4	5	6	7
Repeat 1 Latters	0.0071	0.0070	0.2455	0.3228	0.2198	0.1083	0.0270
Repeat 2 Latters	0.0034	0.0267	0.1774	0.3178	0.2734	0.1518	0.0389
Repeat 3 Latters	0.0000	0.0042	0.0400	0.2030	0.3673	0.2898	0.0998

6 Alphabetic Quantization BackPropagation(BP) Neural Network Model

6.1 Word Quantification Matrix

In order to predict the distribution of the reported results, we need neural network models, so we need new feature attributes that can represent words.

The patent 17/566966 [6]on the USPTO suggests that the conversion of text to audio can be digitized by a matrix, and that the digitized text has faster computer recognition characteristics and a variety of processing modes, and is therefore relatively ubiquitous. Therefore, we considered breaking up words into their component characters, which can be divided into a limited number of English words, and then turning each letter into a different integer so that words could be stored as a matrix.

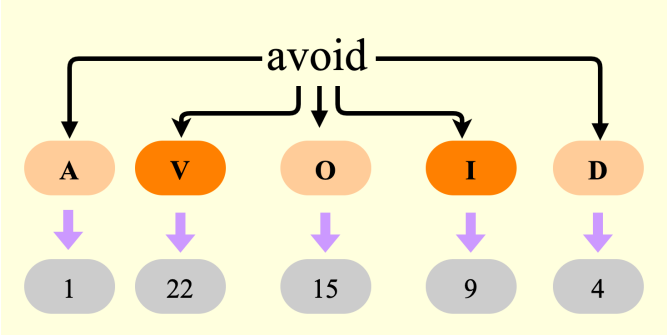


Figure 9: Word splitting into numbers

By such transformation, the BP neural network prediction is made more efficient. Based on the imagined letter quantization BP neural network prediction model, we quantize the letters in each word into numbers 1-26 respectively according to the order of the alphabet, to obtain 359 1×5 matrices. Then it is the work of building BP (Back propagation) neural network in MATLAB.

6.2 Principle of BP Neural Network Model

The standard BP neural network is a multi-layer network that uses a gradient descent algorithm to train the weights of non-linear calculable functions according to W-H learning rules. According to Kolmogorov theorem and BP theorem, for a three-layer BP neural network, as long as the number of hidden layer nodes is enough, it can approximate any complex nonlinear mapping[3]. Because of the 359 sample data

we have, it meets the sufficient amount of sample training of the BP neural network. Therefore, our group believes that setting the number of layers of the neural network at three layers can meet our prediction accuracy. We take the five elements of a 1×5 matrix quantized by letters as independent variables X_1, X_2... X_5 inputs the BP neural network. Similarly, (1,2,3,4,5,6, X) corresponding to the sample data is used as the dependent variable for the 359 neural networks. The principle of neural network is shown in the Figure 9.

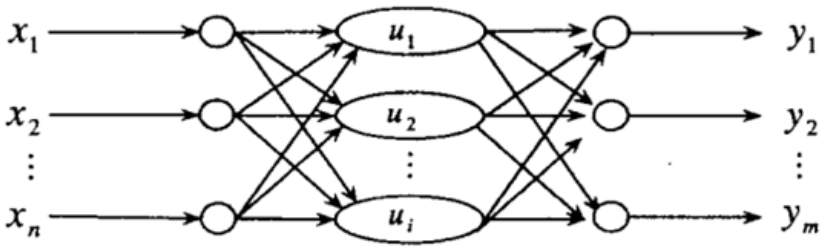


Figure 10: Neural network principle[1][3]

The next step is to normalize the data of all dependent and independent variables using the premix function in MATLAB. In the normalization process, we set the processed value to be in the range [-1,1].

With the data processing done, I started building the neural network model. The model is trained using gradient descent and the results are shown once in 1000 iterations. After many trials and comparing the R (correlation coefficient) obtained from each result, we decided to use 0.01 as our learning rate. To balance the reliability and time cost of training, our maximum training epoch is 50000, and the mean square error is set to $0.65 \cdot 10^{-3}$. After setting the model parameters, the data samples are used as the training set for training, and then the remaining data samples are used as the test set for simulation testing, to continuously improve the BP neural network model.

In the last step, we quantize the letters of the target word 'EERIE' into a row matrix with five elements, then normalize the elements in the row matrix and input them as independent variables into the trained BP neural network. The output of the neural network is normalized and restored to the original order of magnitude.

6.3 Prediction of the Distribution of Unknown Word Results

The resulting (1,2,3,4,5,6, X) distribution for the word 'EERIE' on March 1, 2023 is are shown in Table 3

Table 3: (1,2,3,4,5,6, X) distribution for the word 'EERIE' on March 1, 2023

	1	**2**	**3**	**4**	**5**	**6**	**7**
The resulting distribution	0.4192	7.5915	31.3835	36.5388	17.4892	5.1789	1.0719

[1](where n = 5; i = 3; m = 7)

6.4 Uncertainty Analysis and Error Analysis

However, there are still uncertainties in the model. First,

- First, because there are data whose distribution proportion does not add up to 100% and the rounding process of the data after the results are obtained by the neural network, the predicted data distribution proportion does not add up to 100%. The solution to this uncertainty is to normalize the predicted data.
- Second, after obtaining the distribution percentage of predicted words, if we want to obtain the specific number of predicted words through it, we may get the number of people with decimal numbers, which is not in line with the actual situation.
- The third element of uncertainty is time. Suppose that the number of people playing the game increases dramatically due to the increased popularity of the game due to factors such as increased advertising and social media attention, then the prediction model will not match the actual situation due to these uncontrollable uncertainties.

In the following, the error analysis of the alphabet quantization BP neural network model we constructed will be carried out to illustrate the degree of confidence we have in the model.

After learning by BP neural network, we get the function relation of output on target variable as output$= 0.69 \cdot target + 13$. Data fitting is shown in the figure below:

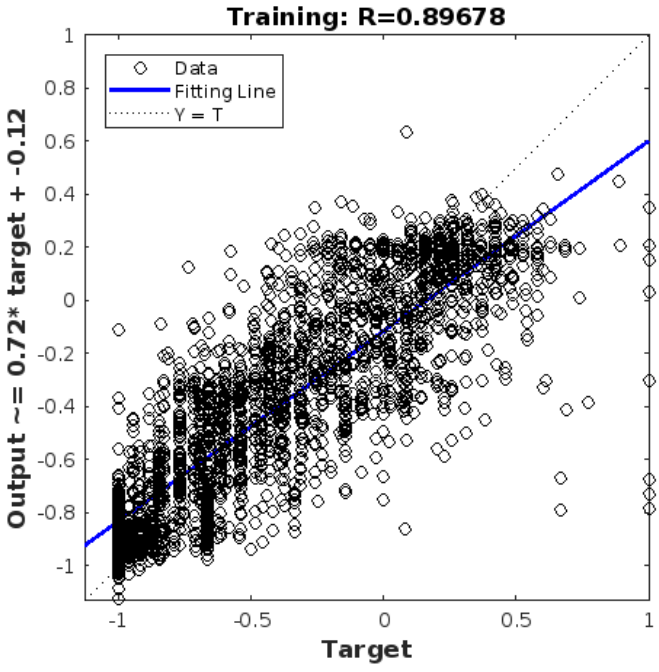


Figure 11: Fitting of quantitative letters to the probability distribution of the number of correct word guesses by neural networks

The value of the determination coefficient R reaches 0.89678, which means that the model can explain nearly 90% of the variation of the target variable. This is a higher degree of fitting, which can be regarded as a better degree of fitting, and a better degree of matching between the model and the data.

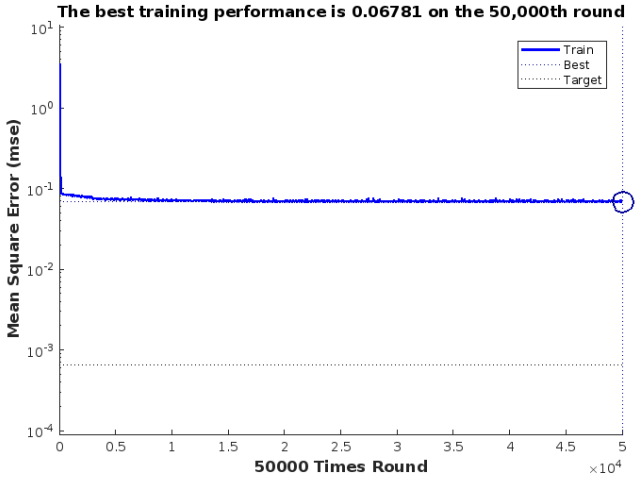


Figure 12: Neural network training mean square error

According to the Figure 12, we can observe that the best performance of the mean square error (MSE) in the 50,000th learning is 0.06781. The mean squared error is the average of the squared difference between the actual and predicted values. The smaller the mean square error value is, the smaller the gap between the predicted value and the actual value of the model is, that is, the better the model fitting effect is. The curve in the image tends to be more and more horizontal, indicating that the mean square error of the model becomes more and more stable as the number of learning increases. It shows that the model obtained by BP neural network learning is relatively accurate and perfect.

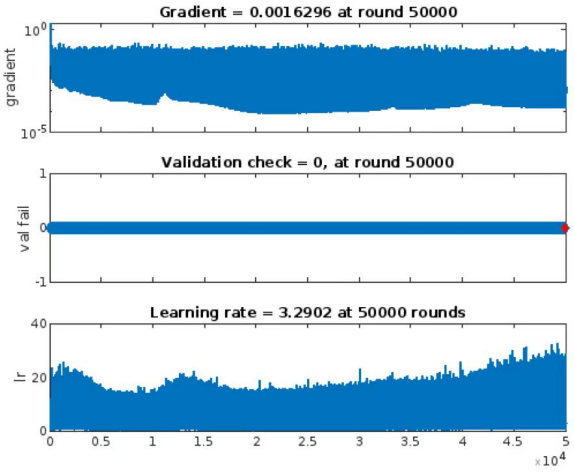


Figure 13: Neural network training state trend

The Figure 13 shows the values of the gradient, the value of validation on the check, and the learning rate at the 50,000th round of forecasting. Gradient=0.0016296, which is relatively small, indicating that the update direction of the current parameter is consistent with the direction of the change rate of the objective function. At this time, the BP neural network takes a large learning value, which is equal to 3.2902. To accelerate the model training, speed up the convergence of the model, and improve the generalization ability of the model, to obtain a better model.

The value of the Validation check is equal to 0, which means the model performs well on the validation dataset. This may be because the model has good generaliza-

tion ability and can adapt well to new data. Through the observation and analysis of our neural network prediction learning data. When a new word is presented, such as "EERIE," our model does a good job of predicting the percentage of times each person needs to guess it correctly.

7 Word Difficulty Coefficient-Based K-means Clustering Algorithm Model

7.1 Word difficulty factor rating

By analyzing the data in the question, we realized that in order to classify the attached words reasonably and accurately, we need to first clarify the difficulty coefficient of the words. This difficulty factor is based on the number of times a player needs to guess each word correctly. The number of times a player needs to guess a word correctly includes the number of people in the easy mode and the number of people in the hard mode. We know that the number of people in hard mode and normal mode has an impact on the percentage of times a player needs to guess a word. In order to objectively describe the difficulty of a word, we introduce the difficulty coefficient.

Let the number of people guessing each word in the normal mode be A and the number of people guessing each word in the hard mode be B.

In simple mode: The number of people required to guess each word correctly is set to $x_i(\mathrm{i} = 1, 2, 3, 4, 5, 6, 7)$

In hard mode: The number of people required to guess each word correctly is set to $y_i(\mathrm{i} = 1, 2, 3, 4, 5, 6, 7)$

With equal probability, players choose two different modes, and the difficulty factor for players to guess correctly on $(1, 2, 3, 4, 5, 6, 7)$ is

$$p_i = \frac{x_i}{A} + \frac{y_i}{B} \quad (\mathrm{i} = 1, 2, 3, 4, 5, 6, 7) \tag{7}$$

In this formula, it is the difficulty factor of each number of times $(1, 2, 3, 4, 5, 6, 7)$ that the player needs to guess each word correctly. The smaller the difficulty factor is, the more difficult it is for the player to guess each word correctly.

Since the data given by the question do not include the actual value of and, the above formula cannot directly obtain the result. In order to quantify the data more accurately, we use the inequality method to deal with the above formula.

$$\begin{aligned} P_i &= \frac{x_i}{A} + \frac{y_i}{B} \\ &= \frac{Bx_i + Ay_i}{AB} \end{aligned} \tag{8}$$

Then there is inequality (A>B):

$$\frac{Ax_i + Ay_i}{AB} > \frac{Bx_i + Ay_i}{AB} > \frac{Bx_i + By_i}{AB} \tag{9}$$

By simplifying the above inequality, we can obtain:

$$\frac{x_i + y_i}{B} > P_i > \frac{x_i + y_i}{A} \tag{10}$$

Then will be approximated by the mean of its interval

$$P_i = \frac{1}{2}\left[\frac{x_i + y_i}{B} + \frac{x_i + y_i}{A}\right] \tag{11}$$

By putting the data in the appendix of the question into the above formula, we can calculate the difficulty factor for each number of times (1,2,3,4,5,6,7) that the player needs to guess each word correctly. Using this difficulty coefficient to classify the words, we adopt the k-means clustering model.

7.2 Model K-means Clustering Algorithm

After successfully assigning a difficulty factor to the distribution of the number of successful guesses for each word, we implemented the K-mean clustering algorithm on MATLAB to classify the words according to their difficulty.

The data was first read into MATLAB, and then we started setting the parameters of the K-mean algorithm. We set the number of clusters to 3 because we wanted to be able to analyze whether the words in the 3 classes obtained by the K-means clustering algorithm were related to the words in the 3 classes classified by word attributes. In order to balance the accuracy of the clustering results and the speed of the algorithm, the number of cycles of clustering was designed to be 5000, followed by setting the distance function and finally calling the K-mean clustering algorithm, which resulted in two results. The first result is a table of the class centers of the 3 clusters, as Table 4.

Table 4: Class Center Distribution

Clustering labels	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Group 1	0.0231	0.3784	1.4692	2.1665	1.5593	0.7523	0.1905
Group 2	0.1169	0.7744	2.8476	3.8913	2.8550	1.5928	0.3261
Group 3	0.2362	1.4441	5.9560	7.5926	5.3072	2.6563	0.4598

The second result is the classification of the sample data according to the difficulty factor of the word. We grouped the samples with the same clusters together and identified the attributes of the given words associated with each classification for the mathematical statistics, and the final data is presented below.

Table 5: Clustering and word attributes correspond to percentage cases

Clustering	**Easy**	**Normal**	**Hard**
	Attribute 1 Percentage	**Attribute 2 Percentage**	**Attribute 3 Percentage**
Group 1	0.8235	0.1764	0
Group 2	0.7117	0.2811	0.0071
Group 3	0.7288	0.2881	0

We base the difficulty classification for the word attribute on Figure 8 made earlier in the article. that legend can be approximated as a sample data curve skewed to a

normal distribution curve, so we look at the data distribution corresponding to these three curves as if they were normally distributed. The coordinates of the highest points of the three normal distribution curves without repeated letters, with one repeated letter, and with two repeated letters are (4, 0.3218), (4, 0.3178), and (5, 0.3673), in that order. Since a more difficult word means a higher number of attempts to make it, we can find by these three coordinates that the word with the attribute of two repeated letters is the most difficult, so in Table 5 we give it the label 'difficult'. Subsequently, in the comparison between (4, 0.3218) and (4, 0.3178), although the highest points of the two curves are close to coinciding, we see that the curve for words with the two repeated letters attribute is left-skewed with respect to the central axis. From this, we can judge that the difficulty of words with the repeated one-letter attribute is slightly greater than that of words without the repeated word attribute. In summary, we assign the label 'Easy' to the word with attribute 1, 'Normal' to the word with attribute 2, and 'Hard' to the word with attribute 3. ' label.

By observing Table 5 above, the change from cluster 1 to cluster 3, attribute 1 accounts for the overall decreasing trend of the cluster ratio, while attribute 2 accounts for the increasing trend of the cluster ratio, and finally attribute 3 accounts for the cluster ratio is basically the same as 0. Therefore, the comparison between clusters 1 and 2 shows that the proportion of attribute 1 in the cluster is decreasing, while the proportion of attribute 2 is increasing. Therefore we conclude that cluster 1 should be classified as 'Easy' difficulty and cluster 2 as 'Hard' difficulty. In contrast, during the comparison of cluster 2 and cluster 3, the percentage of each attribute in the two clusters is almost equal to the cluster ratio, so clusters 2 and 3 should be classified as equal 'Hard' difficulty.

Let the difficulty matrix for ratings 1 to 7 of the words to be tested be W. Then.

$$w = \left[\begin{array}{cccc} p_1 & p_2 & p_3 & \cdots p_7 \end{array} \right] \tag{12}$$

The category centers obtained from the above clustering are Hi, respectively, where the sequence numbers denoted by

$$H_i = [P_{1i}, P_{2i}, P_{3i}, \cdots P_{ii}] \tag{13}$$

Then there exists.

$$\min \cdot d_i = \|W - u_i\|^2 \quad (i = 1, 2 \cdots n) \tag{14}$$

When di takes the minimum value, we find its i value, i.e., the category to which the word belongs is I, and the overall difficulty is Hi. The difficulty of the predicted word can be found by bringing the data in the above table.

After that, we calculated the difficulty coefficient for the word "EERIE" and used the Euclidean algorithm to calculate its difficulty coefficient matrix using the class centers obtained from the first result, respectively. We discovered that "EERIE" is in the second cluster, which was given the highest difficulty score, and that it is part of cluster 2. Table 6 illustrates.

In the process of testing the accuracy of the model, we still use the Euclidean algorithm to verify it. Consistent with the principle and process of classifying 'EERIE' described above, the clustering labels of each group of data under the Euclidean algorithm are derived, and the clustering labels obtained with the Euclidean algorithm are subsequently compared with the clustering labels obtained with the K-mean algorithm. The identical rate of the labels obtained by the two algorithms is taken as the accuracy rate of the model. The result obtained is an accuracy rate of 94.258

Table 6: Correspondence between the clustering labels of 'EERIE' and the difficulty factor ratings

Clustering labels	Hard						
	1	2	3	4	5	6	7
Group 2	0.0223	0.4039	1.669	1.9441	0.9305	0.2755	0.0570

8 Data Set Specific Evidence Exploration

In Model 3, we used the K-means clustering algorithm and were surprised to find that the second and third classes were intelligently assigned the same difficulty in the process of assigning difficulty to the three clustered classes. Therefore, we conjecture that the actual difficulty of the data samples from the first inflection point of the overall trend at the given time (2022/2/5) to the end date (2022/12/31) may be similar. This is because as can be seen from the figure below only the figures for cluster 2 and cluster 3 appear after February 5.

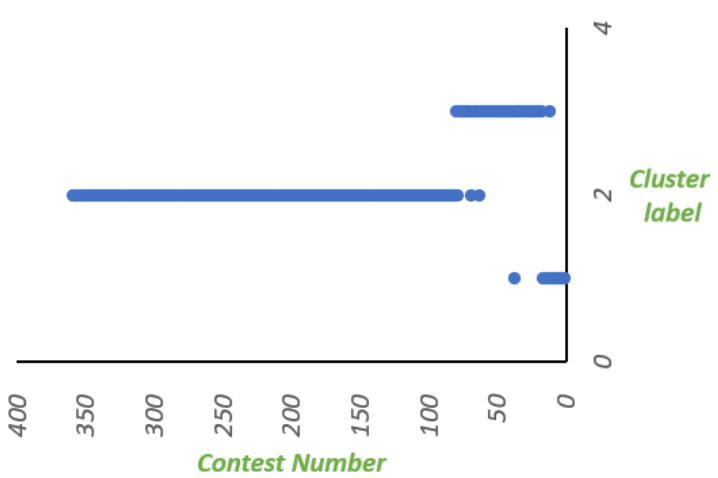
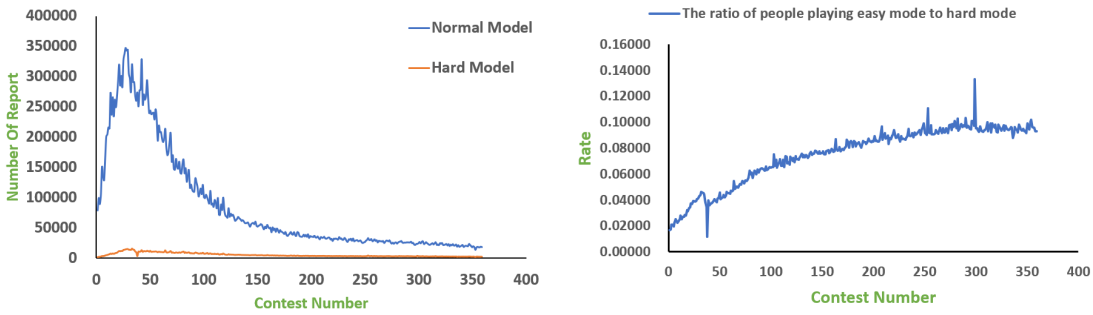


Figure 14: Time variation of clustering labels

And in the case of similar game difficulty, will more people switch from normal mode to hard mode, so that the proportion of hard mode players in the overall player will be increased? With this question we drew two graphs, one when the change in the number of players playing in easy and hard mode compared to the other is the percentage of players in hard mode.

After February 5, the number of people playing Easy mode is decreasing, but the number of people playing Hard mode is increasing (see Figure 15, left). The rate of the number of people playing hard mode is increasing (see Figure 15, right).

By the above can be guessed, may be with the passage of time, the game difficulty reached an equilibrium point, thus leading to more and more people from playing easy mode to play hard mode.



(a) Number of players in easy and hard modes(left)

(b) The ratio of people playing easy mode to hard mode(right)

Figure 15

9 Sensitivity Analysis

In order to ensure the generalizability of model I, we decided to use sensitivity analysis to test the model, mainly for the second segmentation function of model I (data samples after 2022/2/5). Because the trend of the second segmentation function has a strong correlation with the future prediction results, at this time the game's in the change of time, the heat is decreasing, so the total number of people playing the report is also decreasing, while the first half is increasing with the heat of time, and the number of people playing the game surges to the peak. The expression of the second segmentation function is:

$$3 \cdot 10^7 x_t^{-1.252}(30, \infty) \tag{15}$$

This is the multiplicative power function, so that 3 is the factor R. In the process of sensitivity analysis, we make R belong to the interval [2.7,3.3] and finally calculate the error ratio between March 1, 2023 and R=3. It was found that the coefficient factor R expanded or reduced by up to 10

Table 7: Model 1 sensitivity test

Sensitivity factor R	2.7	2.85	3	3.15	3.3
Sensitivity factor change	10%	5%	0%	5%	10%
Number	13029	15033	15635	16615	17198
Influence	−16.652%	−3.85%	0%	6%	10.132%

10 Evaluation of Strengths and Weaknesses

10.1 Strengths

We have the following advantages with our model:

1. Data visualization methods are applied to rationalize the interpretation of the raw data. Specifically, the scatter diagram visually expresses the trend of the reported

quantity data and the distribution characteristics. Further, it is processed in partitioned blocks.

2. Using BP neural network prediction, thanks to the mapping function from input to output of BP neural network, According to mathematical theory, any nonlinear continuous function can be approximated by a three-layer neural network with unlimited precision. In addition, it has the ability to learn and has a high degree of generalizability.
3. Sensitivity analysis was performed on model 1, and it was concluded that model 1 has high generalizability and can be used for similar wordle data and can produce correct results

10.2 Weaknesses

The following are our model's shortcomings and suggested fixes:

1. The K-means clustering model used can be time and memory consuming with large amounts of data. Therefore, we believe that it can be optimized by reducing the number of clusters K as well as reducing the feature dimensionality of the samples.
2. The sample data is insufficient, and BP neural network analysis will produce more compelling findings if we have more score data.

Our Letter

From: Team #2315668

To: The New York Times

Date: February, 20, 2023

Dear Puzzle Editor,

It's our honor to write to share several results regarding the daily Wordle puzzle. To explore wordle more deeply, we establish three customized models to investigate game participation is affected by time and how well players work on words with different attributes and predict the distribution of the number of times players need to guess a specific word right. Additionally, after doing some research and analysis, we have noticed a few areas where the puzzle could be improved to enhance the player's experience. So we will make some suggestions for improvement based on our investigation.

Firstly, we use the "Multiplicative Power Segmented Regression Prediction Model" to fit the number of reported results over time and predict intervals gradually decreased, which we believe is related to the decline of the game's popularity, so it is necessary to innovate the game on the b for the number of reported results on March 1, 2023. It is worth noting that since February 7, 2022, the number of reported results chassis of the original to attract players back. At the same time, we divided the attributes of the word into no repetition of the letter, one repetition of the letter, and two repetitions of the letter. We counted the correlation between the number of times the player needed to guess the word correctly and the number of repetitions of the letter in the word, and plotted the following graph to show As a word repeats more letters, it becomes more difficult for the player to get the word right.

Secondly, we quantized letters into numbers 1-26, thus constructing a letter quantization BP neural network prediction model to predict the percentage of relevance for a particular word "EERIE".So when you need to name a particularly difficult word, our model can help you find that word easily.

Building upon previous research, we used the k-means clustering algorithm to group the words by difficulty and split them into three clusters. Cluster one is defined as easy, while clusters two and three are defined as difficult. By comparing with the cluster center, we classify "EERIE" as the "difficult" type. When you need to know whether your word is difficult or easy, our model can help you classify the word.

Lastly, we found that people generally needed more time to guess a word, especially when there were letters repeated in the word. To increase the playability of the game, we suggest that If there is a repetition of five letters in a word, the game system can tell the player that a letter is repeated in the word when the player correctly types the repeated letter, that is when the letter turns orange or green. But only once. For example, for the word "EERIE ", when we first guessed "BEAST", the system would tell us that there was a word repetition, and then I would go on to guess "TEETH", although the E was repeated three times in "EERIE ", the game system would no longer tell us that there was a letter repetition. This slightly reduces the difficulty of the game, but at the same time increases the playability, and at the same time can better improve the interest of players, which is beneficial to the promotion of the Wordle puzzle.

Overall, the daily Wordle puzzle offered by the New York Times is an excellent way to promote mental agility and creative thinking. It is a fun and challenging game that is accessible to people of all ages and backgrounds. Thank you for offering this engaging puzzle, we hope that some of the suggestions for improvement will help you improve and upgrade this anagram game, and we look forward to continuing to play it in the future.

References

[1] Annelie Ädel and Britt Erman. "Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach." In: *English for Specific Purposes* 31.2 (2012), pp. 81–92. ISSN: 0889-4906.

[2] Abdulrahman Alwadea. "What do disyllabic words tell us about syllable structure, vowel quality, and stress in English?." In: (2021).

[3] Xu Haijing Liu Ying Jiao Shuhua Xia Bing. "MATLAB implementation of BP neural network prediction". In: *Journal of Harbin Higher Institute of Finance* 55-56 (2009). ISSN: 1004-9487.

[4] Emilia Kerr, Jonathan Mirault, and Jonathan Grainger. "On non-adjacent letter repetition and orthographic processing: Lexical decisions to nonwords created by repeating or inserting letters in words." In: *Psychonomic Bulletin Review* 28.2 (2021), pp. 596–609. ISSN: 1069-9384.

[5] Peter Mark Roget. *Thesaurus of English words and phrases : classified and arranged so as to facilitate the expression of ideas and assist in literary composition.* Cambridge library collection. Linguistics. 2014. ISBN: 9781108074179.

[6] "SYSTEM AND METHOD FOR SPEECH TO TEXT CONVERSION." In: (2022).

Appendices

```
>>[x_n,min_x,max_x,y_n,min_y,max_y]=premnmx(x,y); % x:input; y:output
>>u=ones(n,1);
>>dx=[-1*u,1*u];              % After normalization, the minimum value is -1 and the maximum value
is 1

>>net=newff(dx,[x_row,7,y_column],{'tansig','tansig','purelin'},'traingdx'); % Build the model and train
it with gradient descent
>>net.trainParam.show=1000;          %The results are displayed once in 1000 cycles
>>net.trainParam.Lr=0.01;            % The learning rate is 0.01
>>net.trainParam.epochs=50000;       % The maximum training epoch is 50,000
>>net.trainParam.goal=0.65*10^(-3);  % Mean square error
>>net=train(net,x,y);        % Start training, where x, y are the input and output samples respectively

% The BP network is simulated using the original data
>>an=sim(net, x);        % Simulation was performed with the trained model
>>a=postmnmx(an,min_y,max_y); % The data obtained by simulation is restored to the original order
of magnitude

>>predict=[5;5;18;9;5];
% Input the parameters of the independent variables, each row represents an independent variable, and
the number of columns represents the number of predictions
>>predict_=tramnmx(predict,min_x,max_x); % The new data is normalized using the normalization
parameter of the original input data
>>anewn=sim(net, predict_);          % The simulation was carried out using the normalized data
>>final_result=postmnmx(anewn,min_y,max_y) ; % The data obtained by simulation is restored to the
original order of magnitude
>>error=abs(t- final_result)./y ;                % Relative error
>>final_result;
```