Problem Chosen	2023	Team Control Number
C	MCM/ICM	2318006
	Summary Sheet	

A Comprehensive Prediction Model of Wordle Game Based on Linguistic and Statistical Analysis

Summary

In the last few years, a 5-letter puzzle known as Wordle has become popular all over the world. Under the trend of the game, we analyzed the data of the wordle game by developing mathematical models in order to solve the following problems: explain the variation of the number of reported results over time, explain the effect of the feature of words to hard mode rate, predict the distribution of the reported results and classify words by difficulty. Several models are proposed to solve these problems.

Before the models were established, we made some assumptions: Each player sample is independent of each other. The results of the game will not be influenced by the date. Furthermore, we pre-processed the data set to decrease the impact caused by abnormal data.

Model 1: we developed a **regression** model with dates as the independent variable and the number of reported results as dependent variable to explain the change of reported results over time. Then we use this model to predict the number of reported results on March 1, 2023, which is between 13914 and 15288. Through concept and statistical analysis, we conclude that the hard mode rate has no relationship with the words and has no correlation with words' difficulty.

Model 2: we proposed **Back Propagation neural networks** to predict the distribution of the reported results. We utilized the frequency of unigrams and bigrams, the number of vowel letters, the number of repeated letters and the hard mode rate as our inputs. Our output is the distribution of the reported results. We successfully predicted the distribution of the word "EERIE", which is [0 6 25 32 23 12 2].

Model 3: we applied **k-means++ method** to classify all words into 5 different difficulty levels based on the trial distribution. We also use a bigram language model to get a word's probability. It is showed the number of repetitive letters and probability has close relationship with difficulty. Our prediction of difficulty level of word "EERIE" is normal.

Model 4: we utilized **correlation test** to find the correlations between the different days of a week and the number of reported results. We found that the Pearson correlation coefficient is –0.897 and p-value is 0.006 during 2022/7/31-2022/12/31. We used **the Shapiro-Wilk test** to conclude that the statistics satisfy the normal distribution. These calculations indicate that there is a strong negative correlation between the different days of a week and the number of reported results.

The highlight of this paper lies in using normalized models to analyze the data set, which makes the models have reference value.

Keywords: Regression; BP Neural Networks; K-Means++; N-gram; Correlation Test

1. Introduction

1.1. Problem Background

Wordle is an online popular word-puzzle game provided daily by *The New York Times*. The main goal of this game is to solve a five-letter word (sometimes the word has four or six letters) within six tries or less by typing only existing words in English as their guesses.

For every try, the player will receive feedback with different colors. A green letter in one try means that the letter is in the correct location. A yellow letter indicates that the letter is in the wrong location, but it is exactly in the word. A gray letter represents that the letter is not in the word at all. Players also have chances to play this game in a "Hard Mode", which requires players to mandatorily use green and yellow letters from previous tries in subsequent tries. Figure 1 shows a typical example of a hard mode game.

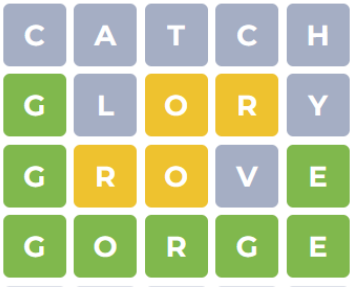

Figure 1 Wordle Game Example

1.2. Brief Overview of the Data Set

We are provided daily results of this game from 2022/1/7 to 2022/12/31, which embraces dates, word of the day, the proportion of players who solved the word in different times of tries or could not solve it, the number of reported results, and the number of results in hard mode.

Firstly, we give a glance at the total number of reported results and the number of players in hard mode over time (Figure 2). We can find that the total number of reported results increases rapidly at first, and then decreases sharply. The number of results in hard mode is only a small proportion of the total results.

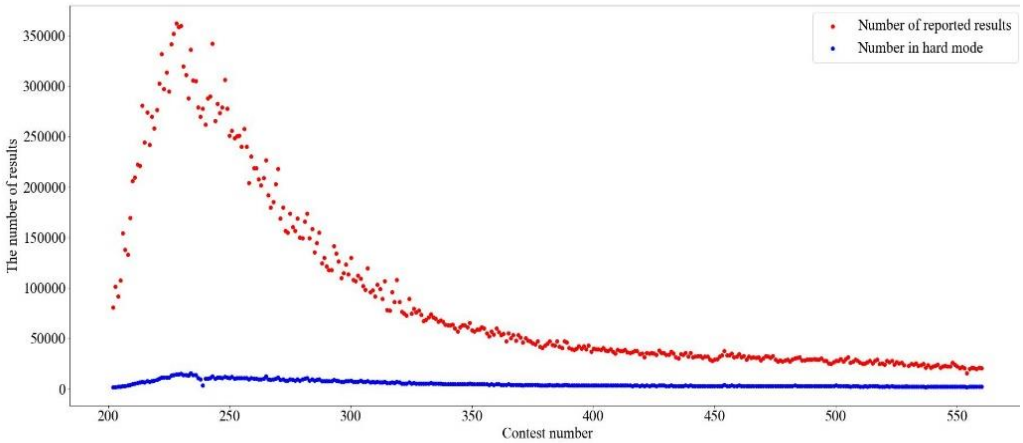

Figure 2 Total number of results and results in hard mode over time

Next, we focus on the proportion of players who solved the word in different tries. We give 5 examples of words in Figure 3, which are words ”foyer”, ”shame”, ”plant”, ”cinch” and ”needy”. We use different shades of color to represent different times of tries. Darker colors represent higher times of tries. We can judge how well the player completed the word by the overall shade of the color. For instance, the word “plant” has lighter colors than the word “foyer” overall, which indicates that the players solved “plant” better than the word “foyer”. To some extent, this also reflects the difficulty of a word. If a word is answered correctly by most of players within lower tries, it means that the word is easy – and the converse is also true. It will be mentioned in subsequent parts.

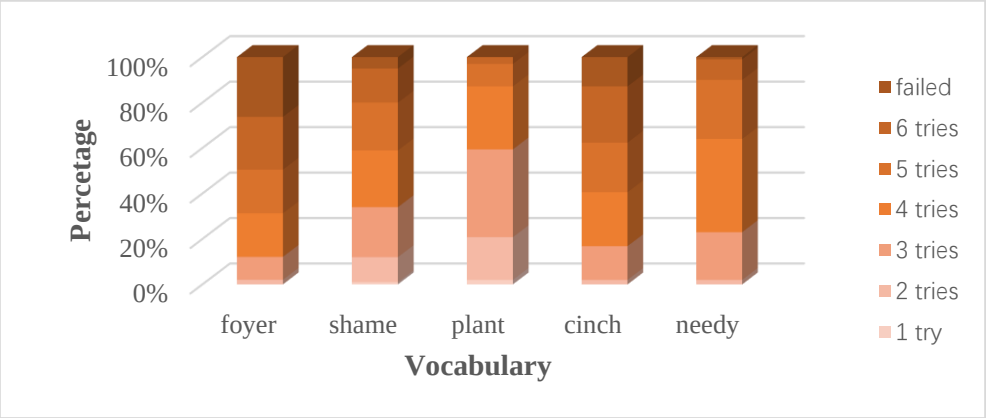

Figure 3 Examples of words about the proportion of different tries

1.3. Restatement of the Problem

Considering the background information mentioned above and restriction identified in the requirement, the following problems are three significant issues that we need to resolve:

- Problem 1: Two sub-questions.
  - Sub-question#1: Develop a model to explain the trend of the number of reported results over time and use this model to predict the number of reported results on March 1, 2023.
  - Sub-question#2: Will attributes of words affect the proportion of “Hard Mode” reports?
- Problem 2: Develop a model to predict the associated percentage of tries that players use to win one game and use this model to predict the word EERIE on March 1, 2023. Evaluate this model.
- Problem 3: Develop a model to illuminate the difficulty of solution words and identify the difficulty of the word EERIE.
- Problem 4: Describe other features of the data set.

1.4. Problem Analysis and Our Work

- **Problem 1:**
  - **Sub-question#1:**

We try to establish a time series analysis model, observe the degree and regularity of data

dispersion, select appropriate curve family to fit and describe the existing data. The accuracy and approximate error of our model are determined and evaluated by the accuracy of the model's description of the existing data. Through this accuracy and the established description curve, we can get the estimated range of the target on March 1, 2022.

- **Sub-question#2:**

The purpose of this question is to let us explore the relationship between the nature and characteristics of words and the proportion of hard mode selection. According to the mode selection options of the wordle game, players can only choose the difficult mode or not before the game starts, that is, before the first attempt. Therefore, we believe that the proportion of choosing hard mode is more influenced by psychological and behavioral science factors.

Therefore, we decided to test the hypothesis that the nature of words does not affect the proportion of hard mode selection. Because there is no prompt for word formation, we think that the influence of word formation on hard mode selection can be completely ruled out logically.

However, we can reasonably assume that hard mode may affect the overall difficulty. Research on the difficulty of representation plays an important role in the selection of cluster method variables in question three. Here, we simply rely on the distribution proportion of the number of attempts of each word in the data set to establish a rough representation difficulty evaluation coefficient calculation coefficient and analyze whether they are relevant through hypothesis testing.

- **Problem 2:**

The difficulty of this problem is to find the factors associated with the distribution results. What we need to do is to first eliminate the meaningless factors, find out the factors that affect the distribution of results, and eliminate the interference between the factors that affect each other.

According to the basic axiom of mapping and linear transformation, if we can achieve one-to-one mapping, we must make the number of independent factors of input greater than that of output. For this reason, we should find enough independent variables in this problem and reduce the dependent variables describing the distribution results. This requires us to carry out statistical analysis on the result distribution itself.

- **Problem 3:**

Our main idea is to cluster the distribution of the number of answers in the data as an indicator, and then analyze the characteristics of the results of the cluster analysis and the divided classes and establish a complete regression model to explain the clustering difficulty classification.

- **Problem 4:**

The information about time and the number of reported results is significant information that the data set provided for us, so we naturally think that there may be some correlation between the number of players and different days in a week.

In order to demonstrate the relation, we need to write Monday-Sunday to each date and average the number of corresponding reported results. We can primarily discover the correlation by plotting the scatter diagram and then use appropriate model to fit the statistics.

If we cannot find obvious correlation between them, we can shrink the data size to where the number of reported results tends to steadily decline, and then observe the new scatter diagram again. This may reduce the impact caused by the rapid change of the number of reported results over time, since if the number decreased quickly on weekdays and the rate of decline became small on the weekend, it would affect us to find the correct correlation.

1.5. Assumptions and Justifications

Assumption 1: Each player sample is independent of each other.

Justification 1: We assume that the outcome of one player is not influenced by other players. There may be such a situation: one player uploaded the answer of today's game on the Internet, which was discovered by another player, so the player solved the game by using only one try. This kind of situation will impact our statistics, but it is hard for us to rectify it. This assumption is reasonable because these kinds of situations are rare, and it is convenient for us to handle the data set after making this assumption.

Assumption 2: The results of the game will not be influenced by the time.

Justification 2: We exclude the situation that players will get more skills and then solve words better than before as people play the game every day. The assumption is rational, since there are no repeat words in the game and the rule of the game has already imply the skills. As a result, practice will not help you much to play this game. It means that the distribution of different tries to solve a word will not be impacted by time. The distribution may depend on the feature of the word and the proportion of results played in "hard mode". These will be discussed in detail later.

2. Data Pre-Processing

After roughly browsing and evaluating the data, we found the following problems:

- Some data have a strong tendency of outliers. We doubt the correctness of the data of these outliers.
- Most of the words are five in length, and only a few are not five in length, and the word length we need to predict is also five.

Some of the quantities we need to know in the problem sets need to be calculated. According to our experience of the game before modeling, we guess that the number of repeated letters is related to the difficulty of guessing letters, and this amount also needs to be calculated. Furthermore, we guess that different days in a week may have correlation with the number of reported results, so we should add this information.

For the first phenomenon, we decided to conduct data mining for outlier data and verify whether the data is consistent with the official data provided by Wordle game. We retained outlier data consistent with the official data but deleted outlier data inconsistent with the official data. Since the caption of question C clearly mentioned that we can only use the data provided by the official for reasonable mathematical modeling, we decided to respect the requirements of the title and not introduce the external data we have mined for the official website to modify the wrong data, that is to delete the error data directly rather than modify.

Here is a brief demonstration of our outlier data cleaning process:

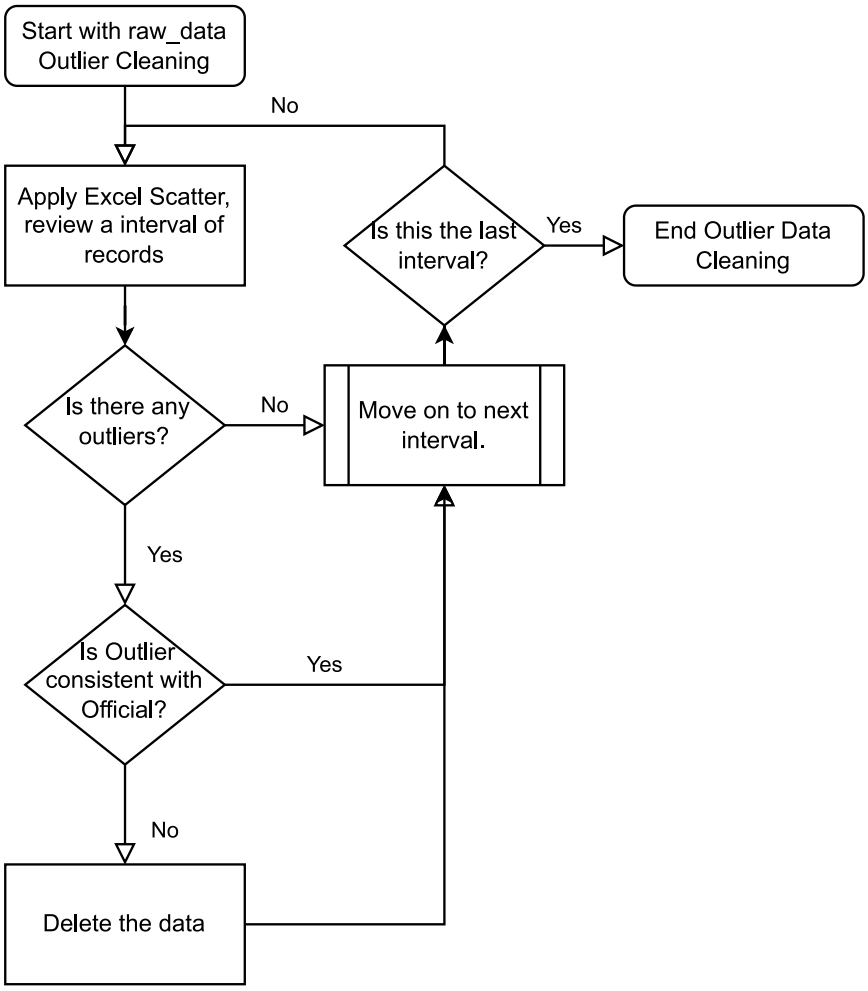

Figure 4 Flowchart Demo for Outlier data pre-processing

For the second and third problem, considering the stability of the prediction results, we decided to keep the data of words with length of 5 only. For the calculation of data, we completed the filtering of five words in length and the calculation of repeated letters through a simple python applet. Another thing that should be mentioned is that we replaced letter ï in the word "naïve" to letter i for convenience.

3. Variation of the Number of Reported Results

3.1. Future Report Quantity Forecast on March 1st.

In this part, we mainly use Excel as a tool for analysis before modeling and Python as a programming language for mathematical modeling and analysis. We established Time Series Analysis and Curve Fitting for the data set, and finally predicted the data of the target date through the properties of the fitting curve and concluded that the quantity interval will be between 13914 and 15288.

3.1.1. Segmental Regression Analysis

In the analyses of the first question, we first used the scatter chart to analyze the trend of the number of reported results and fit the curve.

In general, the number of reported results was rising rapidly and then falling. The rising

is basically smooth and fast, but the decline has experienced form a rapid decline to a steady decrease. As a result, we mainly divided the statistics into three parts: Rising period, rapid decline period and slow decline period.

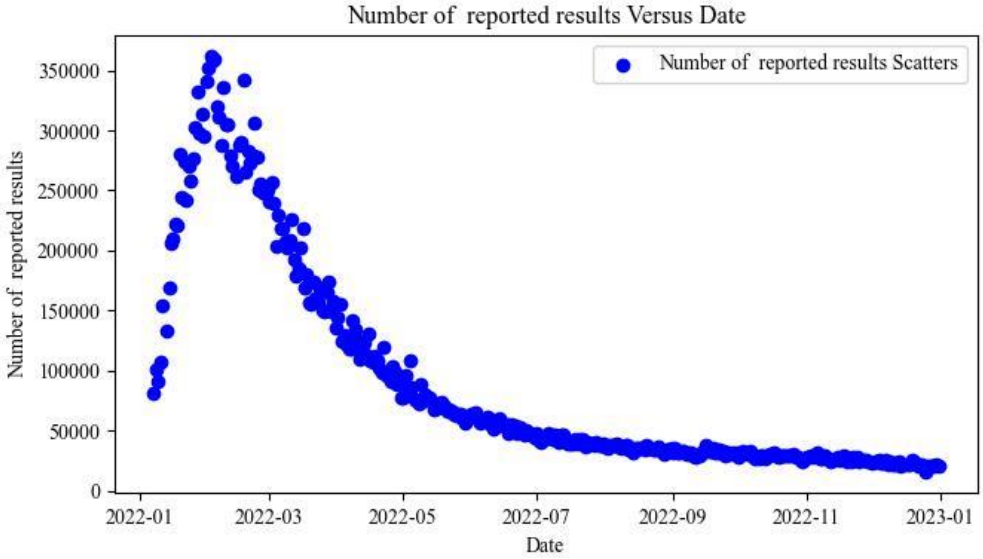

Figure 5 Contest Number Scatter Diagram

The Rising period is from Contest No. 202 to Contest No.232. When the Contest Number $N \in [202,233)$, the function was basically rising, but the rising speed slowed down over time. We found that the quadratic function curve had good matching in this region compared to other curve family provided by Excel. As a result, we chose the quadratic curve to simulate the first segment.

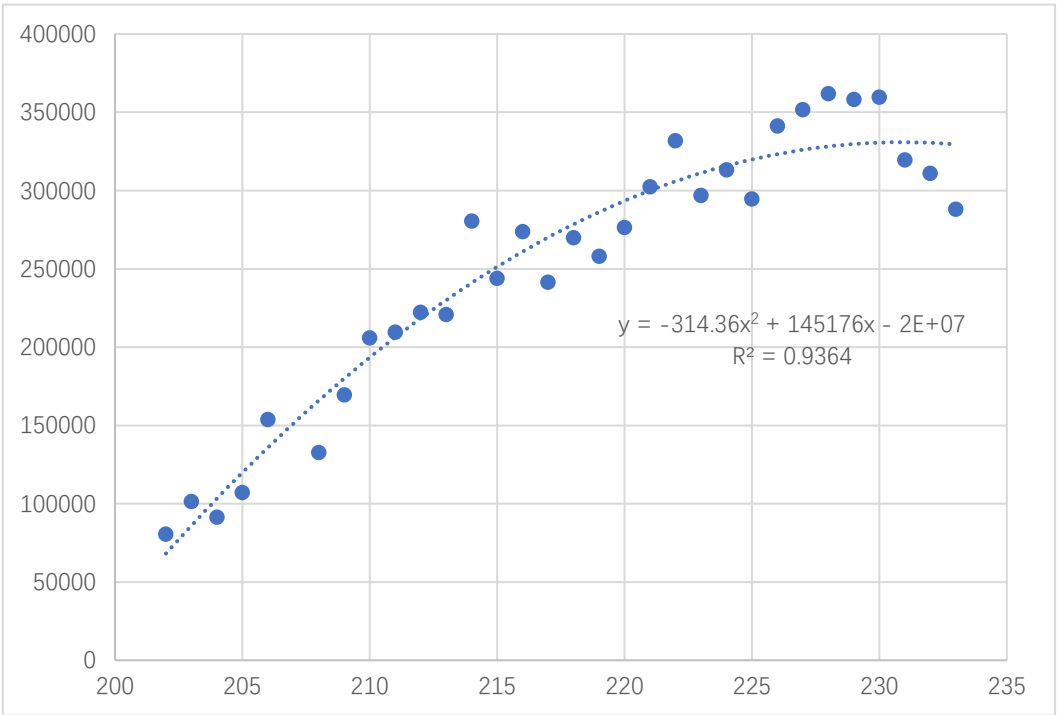

Figure 6 Quadratic function curve generated by Excel

The next two segments fitting curves were generated by similar process. When the Contest

Number $x \in [233,398)$, the curve decreased quickly, and we utilized the cubic curve to approximate it. When the Contest Number $x > 398$, the curve decreased steadily, and we used straight line to fit it. Figure 7 is the result of regression curve diagram with origin scatter data.

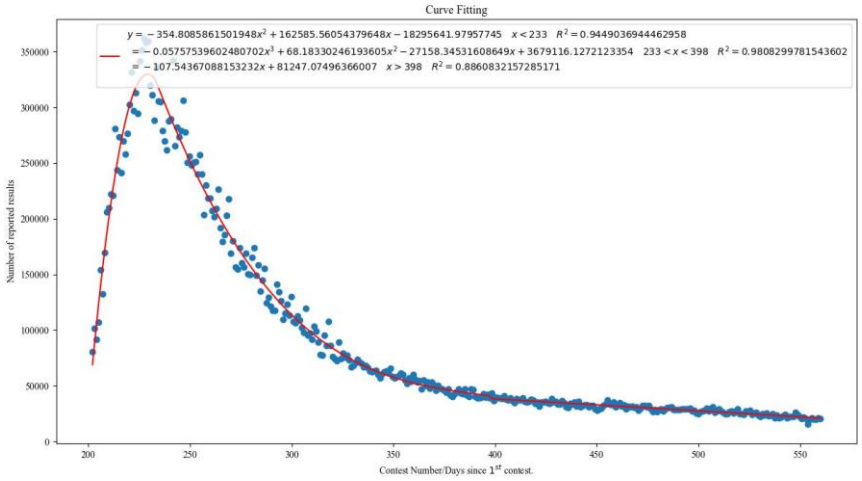

Figure 7 The Regression Plot.

The expression of the total approximation curve is here:

$$y = \begin{cases} -354.8086x^2 + 162585.56054x - 18295641.9796 & 202 < x < 233 \\ -0.057575x^3 + 68.1833x^2 - 27158.345x + 3679116.127 & 233 < x < 398 \\ -107.54367x + 81247.07496 & x > 398 \end{cases}$$

3.1.2. Accuracy and Error Analysis

For the analysis of the accuracy of this model, we analyzed the three segments of the fitting curve separately.

For the first two segments of the fitting curve, we only cared about the overall fitting degree between the current segment data and the current segment fitting curve. For the third segment, we need to pay more attention to the evaluation of the overall fitting degree. Because our main goal is to get the interval of the report number of reported results on March 1, we should pay attention to the relative error between the predicted value and the true value and utilize it to get an interval of the predicted value.

In order to evaluate the overall fitting degree of the three segments, we introduced R-square value, and the definition is as follows:

$$R^2 = 1 - \frac{(y_{\text{predict}} - y_{\text{true}})^2}{(y_{\text{predict}} - y_{\text{average}})^2}$$

This method is a proper way to evaluate the curve fitting degree. The R-square value $R^2 \in (0,1)$, Larger R-square value indicates that the curve has a better performance of fitting.

The overall R-square Values of the three segments are as follows:

$$R^2 = \begin{cases} 0.94490 & 202 < x < 233 \\ 0.98083 & 233 \le x \le 398 \\ 0.88608 & x > 398 \end{cases}$$

For all these segments, the R-square values are very close to 1, which indicate that the fitting curves have a very good performance.

3.1.3. Prediction of the Interval of the Result

For the third segment with contest number $x > 398$, we introduce another parameter δ to evaluate the relative error between the predicted value and the true value.

The value δ is called Mean Absolute Percentage Error, and the mathematical definition is as follows:

$$\delta = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_{\text{predict}_i}}{y_i}\right|$$

Using the following equation, we can get the interval of the predicted data:

$$y_{\text{approximate}_i} \in (\frac{y_{\text{predict}_i}}{1+\delta}, \frac{y_{\text{predict}_i}}{1-\delta})$$

The expression of the third segment is:

$$y = -107.54367088153232x + 81247.07496366007$$

Then we can calculate that the exact prediction data is 14569. The Mean Absolute Percentage Error is 0.0471, and the predicted interval is (13914, 15288).

3.2. Correlation Analysis between “Hard Mode” and Word Properties.

3.2.1. Game Mode and Conception Analysis Before Correlation Analysis

Players can choose to change the game mode to “hard mode” before they play the wordle game. Because the player does not know what the word is at this time, we can judge that the property of the word itself and the choice of hard mode cannot have any conceptual logical relationship.

In response to this problem, we found that the study of Hard Mode Rate here is very helpful to solve the third question. In this question, except for all linguistic but systematic components that have been excluded by us, there is and only the distribution of the number of word attempts is quantifiable for us to study the relationship between Hard Mode Rate and the word itself, We decided to establish a rough model to determine whether the distribution of word attempts is related to Hard Mode Rate in this part to support the research and answer of the third question.

3.2.2. Representation Difficulty Evaluation Coefficient

The purpose of introducing the concept “representation difficulty” is to consider the influence of hard mode rate on the difficulty of filling the words. In the case that the hard mode rate will affect the distribution of reported results, we can divide our sample into two parts, namely hard mode sample and normal mode sample.

We set up a scoring mechanism for representation difficulty according to the distribution of the number of attempts. We mark one success as 1 point of difficulty, two successes as 2 points, and so on. We mark six failures as 7 points. Their sum is defined as Representation Difficulty Evaluation Coefficient ε, and the mathematical definition is as follows:

$$\begin{cases} \varepsilon = \sum_{i=1}^{7}(i \cdot E_i \times 100) \\ E_i = \text{i tries percentage} \end{cases}$$

Through this method, we have obtained a relatively reasonable evaluation index of representation difficulty.

3.2.3. Correlation Analysis

1. General Idea

In this part, our main goal is to find a suitable model to test the correlation between the hard mode rate and the representation difficulty evaluation coefficient. First, we tried to make a scatter diagram of them and observed whether these two variables had good regression properties. According to the scatter plot (Figure 8), we can see that the distribution between the two variables is very discrete and there may not be strong correlation.

Next, we utilized Hypothesis Testing to prove that there is no correlation between Hard Mode Rate (H.M.R.) and Representation Difficulty Evaluation Coefficient (R.D.C.).

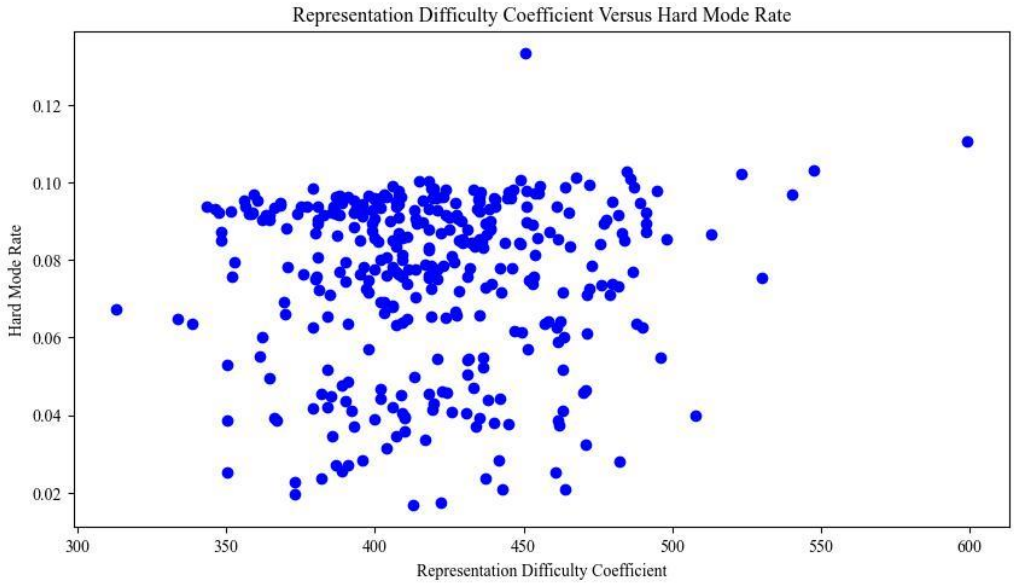

Figure 8 R.D.C. Versus H.M.R., Raw H.M.R. Scatter

2. Hypothesis Testing for Correlation Identity

We find that many samples are concentrated in the middle region, which may lead to the situation that the R.D.C. in the middle region has a particularly large impact on the overall correlation in the analysis. For this reason, we decided to divide the R.D.C. evenly into certain regions, so the R.D.C. sample number of each interval in these discrete regions can be evenly distributed.

The specific method is to use the idea of bucket sorting for reference, divide the difficulty into a certain interval, and assign the average value of H.M.R. in this interval to the middle value of this interval. This can reduce the data fluctuation in a range, eliminate the impact of the difficulty distribution quantity on the overall graph, and make the macro level law in the large interval more obvious. But we cannot get the average interval too large, which will lead to the loss of variability of the data, become too smooth, and produce over-fitting phenomenon, resulting in the reduction of the accuracy of the model.

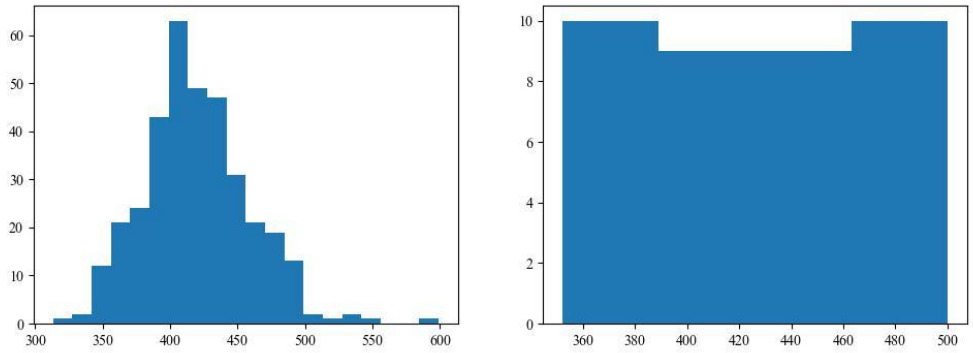

Figure 9 Distribution After Averaged H.M.R.

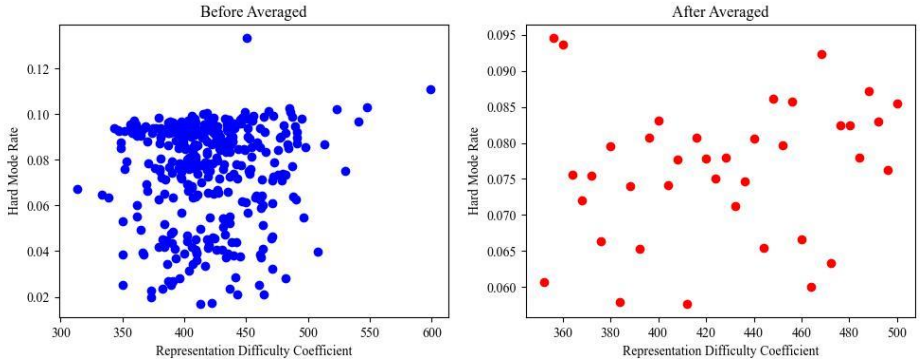

Figure 10 R.D.C. Versus H.M.R., Before and After Average H.M.R.

The specific method is to use the idea of bucket sorting for reference, divide the difficulty into a certain interval, and assign the average value of H.M.R. in this interval to the middle value of this interval. This can reduce the data fluctuation in a range, eliminate the impact of the difficulty distribution quantity on the overall graph and make the macro level law in the large interval more obvious.

After averaging the H.M.R. for the R.D.C. interval, we make a hypothesis test for the linear correlation between the R.D.C. and the averaged H.M.R. through Pearson Correlation Test, which the distribution of H.M.R should satisfy normal distribution. The basic testing idea is here as follows:

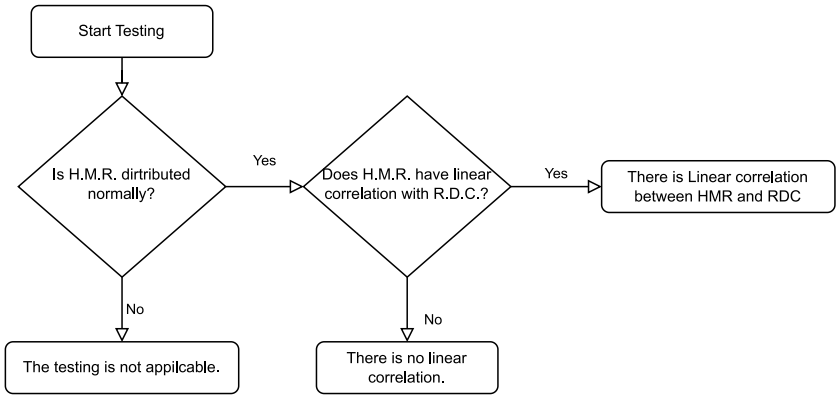

Figure 11 Hypothesis Testing Process

Each rhombus judgment process in the figure above is a hypothesis testing.

For the normal distribution test, we adopted the Shapiro-Wilk normality test. By taking the appropriate R.D.C. interval, we successfully made the Average H.M.R. distribution reach the basic normal and reached the confidence interval of the Shapiro-Wilk normality test under the condition of ensuring that the difficulty distribution is basically uniform. At the same time, we carried out the Pearson test and found that the Pearson test index is not in the confidence interval, so we concluded that, Comprehensive difficulty characteristics have no obvious linear correlation with Hard Mode Rate:

Pearson Result: r = 0.16805, p value = 0.3132 > 0.05

Shapiro Result: p value=0.34547 > 0.05

Combining the averaged H.M.R. scatter diagram with the test results, we can conclude that in the whole sample, Hard Mode Rate has no obvious impact on the distribution of difficulty.

4. Game Trial Distribution Prediction

4.1 Word Attributes Analysis

In order to find the percentage distribution of a future word in a future date, we focus on attributes that may affect the result of a given word. Some attributes of the word w that we think will affect the game reuslt includes the following:

Number of repetition letters: As Wordle only hints letters that appear in the answer word and all repeated letters are not distinguished, we suspect that words with some repetition will be more difficult to guess. For simplicity, we denote a word containing two same letters as 1, a word containing three same letters as 2 and a word containing two different repetitive letters as 2.

Number of vowel letters: Some letters are more frequently used in the composition of a word. And they may be a crucial point for a player to figure out the answer word.

The frequency of unigrams: Unigram of a word is all letters of length 1. For example, the word "extra" has unigram $\{'e',' x',' t',' r',' a'\}$. Unigram information of a letter may influence the result distribution. For example, people would try words beginning with "c" rather than try words beginning with "u", since the letter "c" is more common than "u". In this Wordle word corpus, we collect all counts of 26 unigrams and use the frequency to denote the word. For example, "extra" has the information $\{182,8,128,130,156\}$, where 182,8,128,130,156 denotes the total count of letters 'e', 'x', 't', 'r', 'a' respectively.[4][5]

The frequency of bigrams: Bigram of a word is all letters of length 2. For example, the word "extra" has bigram $\{'\$e','ex','xt','tr','ra','a\$'\}$, where \$ means the pre-start of the word and \$ means the post-end of the word. But as in this discrete word guessing game, we suspect that the pre-start and post-end doesn't have the same importance as other bigrams. We ignore these cases. Bigram also matters because we think that the appearance of frequent letter combinations such as "he", "re" and "in" will make words easier to guess. As we did in the unigram, we count all frequencies of bigrams in the corpus and use the frequency as inputs. [1]

Hard mode rate: As players in hard mode may use a different strategy in solving the problem and there may be different distributions, we put the proportion of hard mode of each word as input.

4.2 Trial Distribution Prediction Model Based on BP Neural Network

Dealing with multi-input and multi-output, we naturally think of back propagation (BP) neural network algorithm, which is one of the most widely applied neural network models.

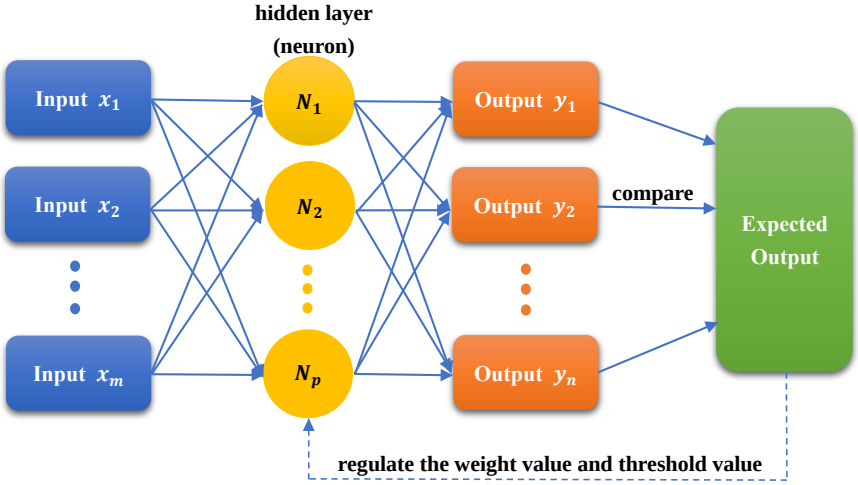

Figure 11 The process of BP neural network algorithm.

Figure 12 describes the process of the algorithm. This algorithm is a multi-layer feedforward network trained according to error back propagation algorithm (Li, J., Cheng, 2012). It's proved that neural networks are a useful model to predict on some data. By training the network using enough data, adjusting parameters in each neuron, using gradient descending to optimize, the network can have a good prediction for an un-trained input.

In our case, we build a simple BP network with one input layer, one hidden layer with 12 neurons and one output layer. As we only have more than 300 datasets, we randomly divide 70% into the train set and 30% into the test set, ignoring validating set for simplicity. The training method we use is gradient descent with momentum and adaptive learning rate back-propagation and the number of epochs is 30000. The performance of the network is evaluated using mean square error (MSE), goodness of fit (R^2) as well as the MSE of test data set.

Further, as we are certain that unigram and bigram have an internal relationship with each other, which may affect the accuracy of the network, we choose to train BP neural network three times using different inputs: only unigram; only bigram and unigram, bigram both. In all cases, the number of repetitive letters, the number of vowel letters and hard mode rate are included in the inputs. For the output, it is the trial percent distribution.

Before inputs data are passed into the network, all data are normalized using the function

$$2\frac{m-min}{max-min}-1$$

to eliminate the effect of dimension.

The results we get are as follows:

Input	*Training Set MSE*	*Testing Set MSE*	R^2
Unigram	**0.031581**	**51.15078**	**0.92729**
Bigram	**0.037942**	**47.88056**	**0.91195**
Unigram + Bigram	**0.019852**	**41.50705**	**0.95494**

Table 1 Results of BP network with different inputs

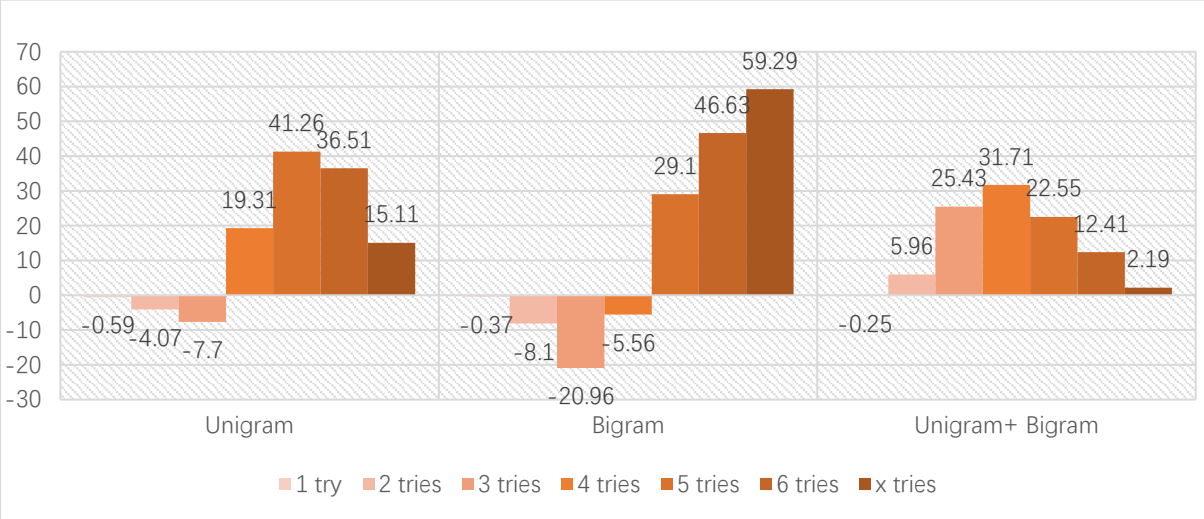

Figure 12 Distribution of "eerie" with different network input

As is shown in the table, both the first and second trial have higher number of MSE during training, MSE of test set and R^2 is higher than the third trail. In addition, both the first and second trial give significant negative number for the word "eerie". This may imply that eerie is an extremely difficult word. But the third trial gives the most reasonable distribution. By training the neural network multiple times, we find that the third trial is most stable (most similar results) and we use this result as our predicted distribution for word "eerie"

As BP neural network is something with great uncertainty itself, there can be lots of factors that will affect the data. Some uncertainty factors include the following:

1. The choice of number of neurons, the activation function as well as the number of hidden layers can be quite indefinite. This is the uncertainty of the neural network itself.
2. The input of neural network may not be complete. Some other attributes include the frequency of a word, the semantic meaning of a word, radical, prefix, suffix of a word. But we don't consider these factors here because of the complicity.
3. The uncertain choice of training data set and not suitable choice of input. The result of the model has great bias toward the training data and it's possible that training data is not sufficient enough. Also, the calculation of repetitive number of letters may not be most suitable.

For the accuracy of the model, we give a rough estimation. Let $X_i^{(j)}, Y_i^{(j)}$ be i[th] testing data item's j[th] trial percentage. The inaccuracy rate is estimated as

$$Avg\left(\frac{\left|X_i^{(j)} - Y_i^{(j)}\right|}{Y_i^{(j)}}\right), for\ all\ i,j\ testing\ data\ items.$$

And accuracy is $1 - inaccuray\ rate$. Our value for accuracy(confidence) is 65.79%, which indicates that our model has a good predication accuracy, but the confidence level is still not enough. So, there can be more improvements to our model.

5. Game Word Difficulty Classification

5.1 K-means++ Word Classification Model

There are different criteria to divide language ability or word difficulty to different levels, but for this problem, we only focus on the difficulty of figuring out the answer. A direct reflection of the difficulty is the trial distribution of a word. Intuitively, if the percentage of trial number 7 or more is high and the percent of trial 1 or 2 is low, the word is considered as difficult. To solve this problem, we built a k-means++ model to solve the classification problem.

k-means++ is a common and effective method to classify different objects. By initially selecting k cluster center that are far away from each other, put all objects to the cluster that the object is closest to. Then update the center of each cluster based on the object cluster to reduce the summation distance of center to all objects in this cluster. Keep iterating this step until the cluster centers converge. Overall, we need to minimize within-cluster sum of squared errors (SSE), where

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} (x - \frac{1}{|C_i|} \sum_{x \in C_i} x), where\ C_i\ is\ i^{th}\ center$$

For the output of the model, we use 5 different difficulty levels L_1, L_2, L_3, L_4, L_5, with L_1 as the easiest level and L_5 as the most difficult level. We will explain the meaning of each level later. Before we apply k-means++, we find the word "parer" with 48 % failure rate and it will make L_5 class contains only one word, so we delete this extremely hard word before we start our k-means++ analysis.

Further, we assume that the hard mode rate has a possibility to affect the trial distribution of a word, which will in turn affect the result of the classification model. In order to take the hard mode rate into consideration, we try our k-means++ model again using modified trial distribution. For this modified distribution, we do the following analysis.

Variable	*Description*
N	Normal mode number of each word
H	Hard mode number of each word
α_i, β_i	i^{th} trial percentage of normal mode and hard mode respectively
γ_i	i^{th} trial percentage of given distribution

Table 2 Hard mode & trial distribution variables

Clearly, we have

$$N\alpha_i + H\beta = (N + H)\gamma_i$$

Take α_i dependent variable, γ_i as independent variable and β_i as disturbance term. Further, simply assume $\alpha_i = k_i\beta_i$, we can get the following:

$$\alpha_i = \frac{\frac{H}{N} + 1}{\frac{H}{N} k_i + 1} \gamma_i$$

Then key point here is the choice of k_i. To reduce the degree of freedom (DOF) of k_i, we set k_1 to 1, meaning that hard mode won't affect the percentage of the first trial and $k_i = k_{9-i}^{-1} \; for \; i = 2,3,4$. If calculated α_i doesn't sum to 1, we normalize them. Then we choose different value of k_2, k_3, k_4 and use α_i as input for the k-means ++ classification model.

Ultimate Cluster Center

	Clustering				
	L1	**L2**	**L3**	**L4**	**L5**
1_try	**1.15191**	**0.32260**	**0.15196**	**0.19445**	**0.68119**
2_tries	**11.48727**	**5.67937**	**3.15678**	**1.77979**	**5.53500**
3_tries	**33.13507**	**25.51242**	**17.23635**	**9.98903**	**17.96171**
4_tries	**32.37546**	**36.19436**	**34.68948**	**25.54527**	**25.28776**
5_tries	**15.73683**	**22.33891**	**28.99210**	**32.05304**	**22.62336**
6_tries	**5.28311**	**8.59549**	**13.47691**	**23.87223**	**18.49143**
more_tries	**0.83035**	**1.35685**	**2.29642**	**6.56618**	**9.41954**

Table 3

With 5 cluster centers and after enough times of iteration, we successfully get good final cluster center with sparse distance with each other. Based on the distribution of these centers, when rank them by estimating the deviation coefficient and comparing them. For model with modified input, we get similar inputs. This corresponds with our result in the first model that hard mode rate has no obvious correlation has difficulty. We only present the result of the original model here. The following are the results of classification along with some representing words that are closet to cluster center.

Difficulty Level	**Word Number**	**Example**
L1	**72**	**Stein, Poise, Charm, Heist**
L2	**124**	**Shown, Inter, Field, Hinge**
L3	**92**	**Bough, Condo, Undue, Rupee**
L4	**36**	**Vouch, Lowly, Gully, Fewer**
L5	**29**	**Shake, Liver, Homer, Baker**

Table 4 Difficulty class overview

5.2 Statistical Analysis with Bigram Language Model

In our BP neural network model, we use unigram and bigram to represent the morphological attributes of a word. It's possible to analyze these attributes of each classification but there will be too much data. Hence, we apply bigram language model to calculate the probability of a word in the Wordle corpus using unigram and bigram.

Denote a 5-length word w as $l_1 l_2 l_3 l_4 l_5$, in which l_i means i^{th} letter. Denote the probability of sub-letters in a word as $P(l_i \dots l_j)$, then $P(l_2|l_1)$ can be computed as $\frac{P(l_1 l_2)}{P(l_1)}$. Using chain of probability, we get the following:

$$P(l_1 \dots l_5) = P(l_1)P(l_2|l_1) \dots P(l_5|l_1 \dots l_4) = \prod_{i=1}^{5} P(l_i|l_1 \dots l_{i-1})$$

This probability can be time-consuming to compute so we can estimate it using Markov assumption. Markov assumption says that current state only depends on last state. By applying this assumption, the upper equation becomes:

$$P(l_1 \dots l_5) = \prod_{i=1}^{5} P(l_i|l_{i-1})$$

When can estimate this probability by using maximum likelihood estimation using the count of the bigram $C(l_{i-1}|l_i)$. We calculate by divide the count of a bigram with sum of all bigrams sharing the same head letter:

$$P(l_i|l_{i-1}) = \frac{C(l_{i-1}l_i)}{\sum_i C(l_{i-1}l)} = \frac{C(l_{i-1}l_i)}{C(l_{i-1})}$$

As we can see, the probability of a word can be calculated by multiplying all probability of bigrams of this word. The probability can be simply calculated by counting all the number of bigrams and unigrams. We take this probability as an attribute of the word and then identify its possible relationship with difficulty level.[2][3]

5.3 Attributes Identification of Different Classes

In addition to the feature of the data set mentioned above, we also discovered some other interesting features which are worth mentioning. For each classification, to identify the attributes of each classification, we select the top 20% words with closet distance to their respective cluster center from each class (most representative words from each class). Then use the average value of their attributes and do a statistical analysis.

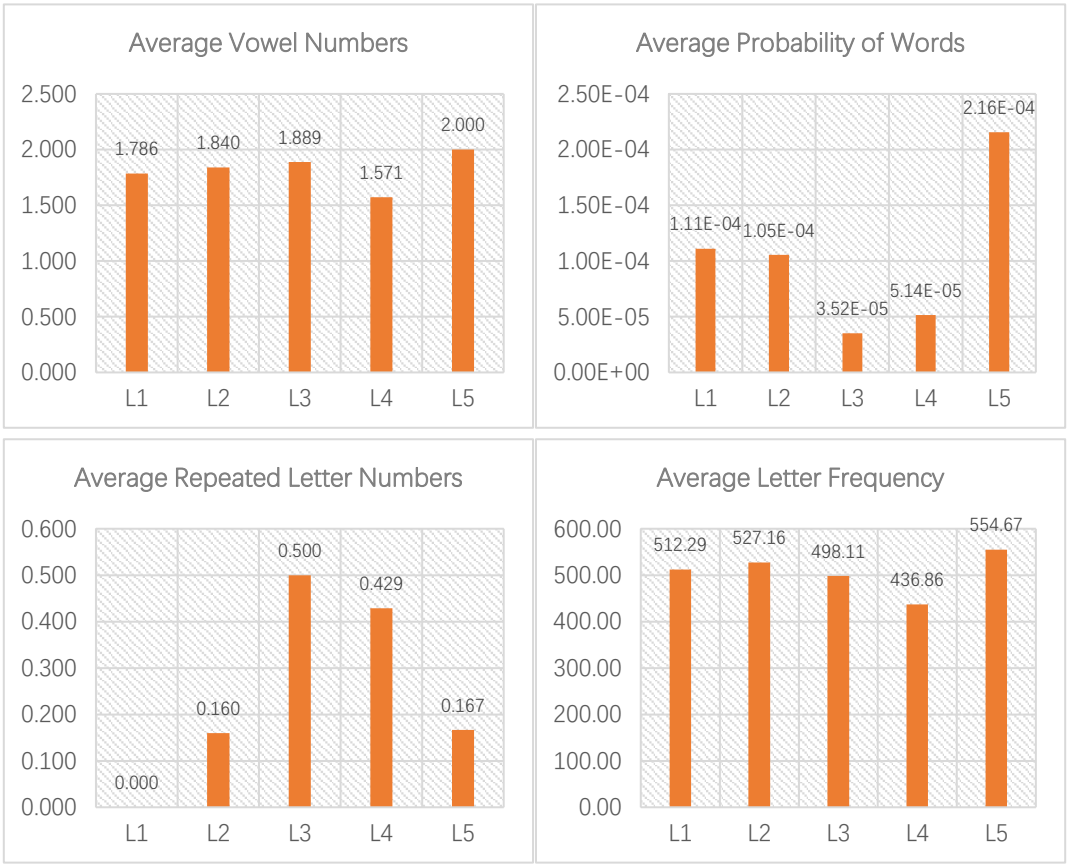

Figure 13 Word Attribute overview of each difficulty level

The graph above shows each class's average repeated number of letters, vowel letter number, probability and letter frequency. Based on the result, we conclude that vowel letter number and letter frequency of a word has no relationship with its difficulty. The simplest word usually has no repetitive letters and difficult words usually have more repetitive letters. This corresponds with our assumption that repetitive letters increase the difficulty of a word. Some most difficult words may be common with high probability, as L_5 in probability graph suggests. A good example is the word "shake", which is common in daily life, but players tend to try more attempts to figure it out. At the same time, most difficult or fair words (L_3, L_4) usually have less probability than other words.

Based on the analysis above, we can now give a good definition of each difficult level.

Difficulty Level	**Definition**
L1	Easy, words that are usually commonly used and has no repetition, usually can be figured out within 3 trials
L2	Normal, words with some repetitive letters, most frequently appear and can be figured out with normal effort
L3	Tricky, words that don't appear often, especially with repetitive letters but can still be solved with some effort.

L4	Difficult, words that are hard to guess, usually has repetitive letters and not common and even can't be figured out
L5	Challenging, words that are extremely difficulty to guess, even some common words are included and often can't be figured out

Table 5 Definition of each difficulty level

For the given word "eerie", our classification model indicates that it's in L_2 class, which means the difficulty is normal. To get the accuracy of our model, we label each word using representation difficulty coefficient from last model to different five levels and compare the two results. Then accuracy is calculated using F(F-measure), P (precision) and R (recall),

$$F = \frac{2PR}{P+R}, P = \frac{|M \cap N|}{|N|}, R = \frac{|M \cap N|}{|M|}$$

Where M means our manually labelled clusters and N means model's clusters, our value for F is 70.53%. So overall our model has great accuracy in classification.

6 Additional Feature

In addition to the feature of the data set mentioned above, we also discovered some other interesting features which are worth mentioning.

6.1 Different weekdays

People may have different rest time arrangements on different days of a week, which may impact the number of players, so we predict that there may be some correlation to a certain extent between the number of reported results and different days of a week. We average the number of players on corresponding days and display them in one scatter diagram. Figure x shows the trend of the number of reported results changing with weekdays during 2022/1/7-2022/12/31.

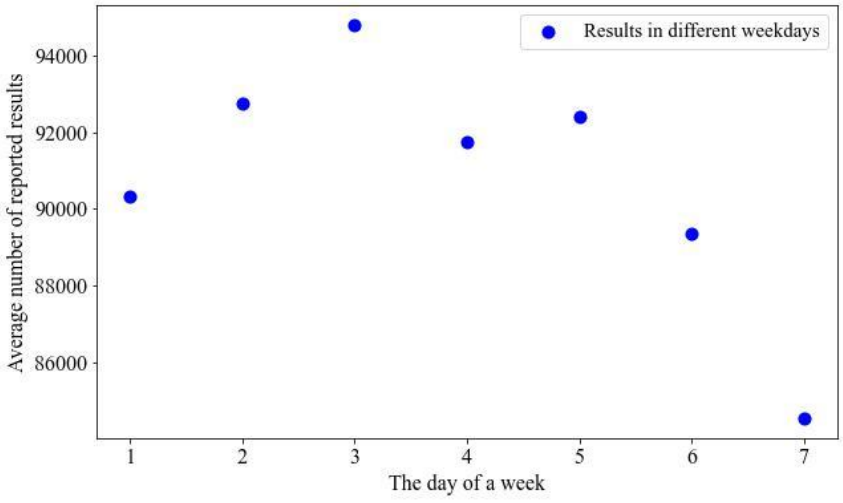

Figure 14 The number of reported results changing with weekday.

In statistics, we use **Correlation Test** to indicate the strength of the association between these two variables by calculating Pearson correlation coefficient, which refers to:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \cdot \sum(y - m_y)^2}}$$

where r is the Pearson correlation coefficient, m_x and m_y corresponds to the means of x and y, respectively. In this example, we calculate $r \approx -0.622$. We cannot conclude that there is a strong negative correlation between the number of reported results and different weekdays until we do the **Hypothesis Testing** to the Pearson correlation coefficient.

Before we do it, we must check whether the data is normal distribution. We use Shapiro-Wilk test and obtain p-value $p \approx 0.460 > 0.05$, which rejects the null hypothesis that there is no significant difference between the distribution of sample data and normal distribution. It means that the data is normal distribution. Then we can do the Hypothesis Testing to the Pearson correlation coefficient. Assume:

Null hypothesis: H_0: $r = 0$

Alternative hypothesis: H_1: $r \neq 0$

We obtain the p-value by using python, which is $p \approx 0.136 > 0.1$. It indicates that we failed to reject the null hypothesis at a 90% significance level. As a result, we are not able to conclude that there is a strong negative correlation between the number of reported results and different days of a week in the entire data set.

As we have already known, the number of reported results increased rapidly in the early days of the website, and after the trend of the game faded, it decreased sharply and then declined slowly and steadily. We tried to check our hypothesis by only choosing data in the period of data slowly decreasing. Figure x shows the trend of the number of reported results changing with different days in a week during 2022/7/31-2022/12/31 (22 weeks in total).

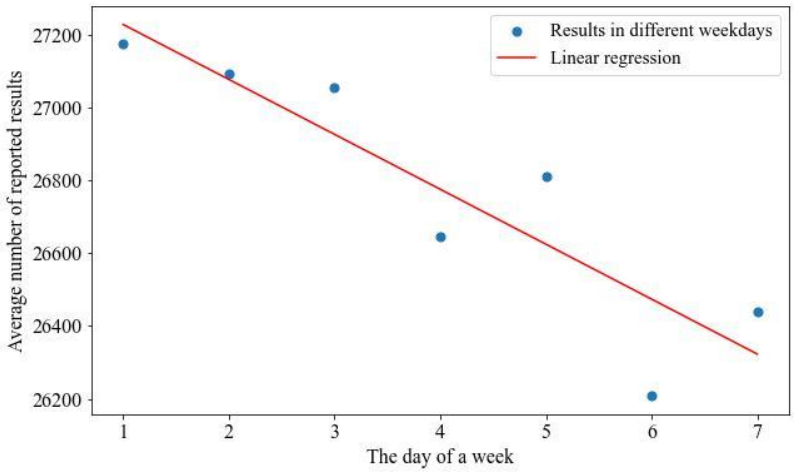

Figure 15 The number of reported results changing with weekday.

We utilize the **Correlation Test** again to indicate the strength of the association between these two variables, and we calculate the Pearson correlation coefficient $r \approx -0.897$. Next, we utilize the Shapiro-Wilk test and obtain p-value $p \approx 0.565 > 0.05$ and conclude it is a normal distribution. Then we can do the Hypothesis Testing and obtain the p-value $p \approx$

$0.006 < 0.01$, which means that we can reject the null hypothesis $r = 0$ at a 99% significance level. In conclusion, the number of reported results has a strong negative correlation with different days of a week. The red line in the above diagram is the result of linear regression by using **Ordinary Least Square regression**.

7 Sensitivity Analysis

7.1 Average Representation Difficulty Coefficient Method.

For the second sub-question solution in the first question, when taking different uniform intervals for the representation diversity coefficient to do the arithmetic mean of the Hard Mode Rate, the results of the model for the Hypothesis Testing will change. If the interval is wide enough, there may be linear correlation results, but we think this is unacceptable. Although the wide interval makes it possible to describe the two parameters in the linear regression relationship, However, too wide an interval will lead to poor fitting of the linear regression model to the original data. The following is an example of the of different length of interval values to the original data:

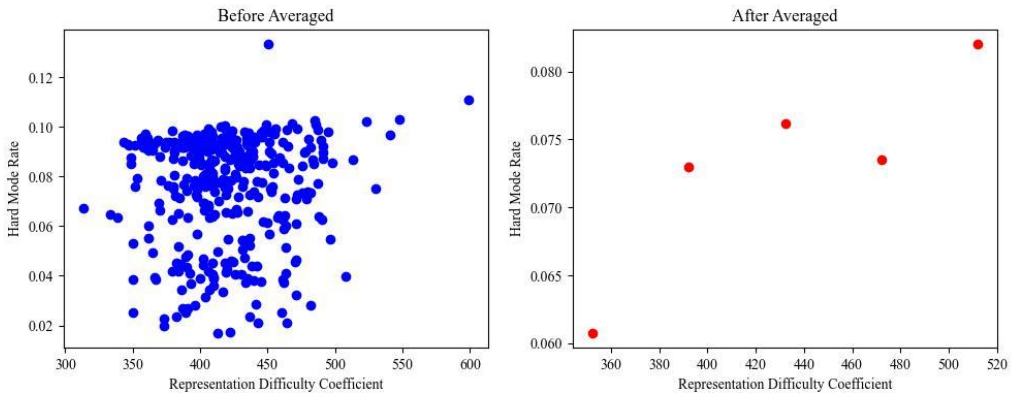

Figure 16: Averaged Comparison when interval length = 40

As we can see, the averaged data is surely to have a good linear correlation, but the corresponding fitting curve obtains a very bad fitting performance with original data. The nature of our model for sub-question is to sacrificing local accuracy for global relatively good linearity.

7.2 Network Improvement

There can be all kinds of choices for the number of neurons, number of hidden layers and so on. We try to analyze whether the performance of the network can be improved by adjusting these parameters. In all cases, we add both unigram and bigram as inputs. See the table below.

As we can say, by adjusting these parameters, there is no significant change in the metrics. The reasons can be 1. The number of epochs is high enough. 2. The training data set is quite limited. 3. The network is overall simple and doesn't contain very complicated layers.

The main factor that can be affected by changing the number of neurons or hidden layers is the training time, so the choice of the network work is not the main concern for this problem, which further indicates that our model has great universal applicability.

Number of neurons in hidden layers	Number of hidden layers	Training Set MSE	Testing Set MSE	Accuracy
10	**1**	**0.021654**	**50.345**	**60.32%**
12	**1**	**0.019852**	**41.50705**	**65.79%**
14	**1**	**0.019363**	**42.5342**	**63.59%**
10	**2**	**0.018432**	**36.6321**	**66.43%**
12	**2**	**0.031243**	**44.4231**	**62.74%**
14	**2**	**0.023214**	**32.4125**	**70.34%**

Table 6 Network Metrics with different parameters.

8 Model Evaluation and Further Discussion

8.1 Strengths

1. Our model has a great achievement of completeness. All kinds of attributes of a word into consideration, with focus on the composition of a word. We also try to use different model input to eliminate some unrelated factors.
2. Data visualization is diverse and sufficient, and each model has corresponding intuitive graphical representation.
3.

8.2 Weaknesses

1. The accuracy of our model is not enough. The result of BP neural network model is passed to the classification model, resulting in accumulative error of the whole outcome. The final prediction may not be persuasive enough.
2. Our model contains lots of uncertainty, mainly caused by the insufficiency of data. Large amount of repeated testing should be implemented to verify the result of our model.

8.3 Further Discussion

1. In terms of the composition of a word, it's reasonable to take 3-gram of a word into consideration. There can also be a better evaluation of repetitive number of letters instead of just counting. This can make a better sense of attributes of a word
2. In our model we only classify words into 5 levels. This may not be the optimal choice for the optimization. One can try more classification number and also a different classification method.

9 Letter to the Editor

From: Team #2318006
To: Puzzle Editor of the New York Times
Date: February 20, 2023

Dear Puzzle Editor of the New York Times,

Thank you for hiring our team as your consultants of the world game project! Our team built up mathematical models to research statistics of the wordle game and finally got some achievements. Here are the details of our findings.

Firstly, the total number of reported results varies daily, and we promoted a model to simulate the change. We have found that the number of reported results surged in the first month of the data set. After this trend of playing wordle, it decreased sharply in the following 5 months. Finally, the rate of decline went steady in the remaining days. For each segment, we used different functions to approximate it, and finally we utilized the third segment to predict the number of reported results on March 1, 2023, which is between 13914 and 15288. Additionally, we did not find any obvious evidence to indicate that attributes of words would influence the percentage of reports in "hard mode" by using Correlation Test.

Next, we developed a model to predict the distribution of the results for a certain word on a future date by using Back Propagation neural networks. We utilized this model because there was no need to get the mathematical equation in advance. In accordance with our analysis, we found that the following features of word would influence the distribution: the frequency of unigrams and bigrams, the number of vowel letters, the number of repeated letters and the hard mode rate. We put these data as inputs of the neural networks, and the network would regulate itself automatically by using error back propagation algorithm and produce applicable outputs. Once the network was built up, we applied the word "EERIE" to the neural network and obtained the distribution [0 6 25 32 23 12 2] with 65.79% accuracy rate.

What's more, we summarized a model to classify words by difficulty. We used K-means++ method and classified all words into 5 categories. Through our classification, the word "EERIE"'s difficulty can be classified as normal.

In addition to what is required for our team, we also find some other interesting features. We found that there is a strong negative correlation between the number of reported results and different days of a week. We made this conclusion by averaging the number of players on corresponding days and using Correlation Test to indicate the strength of the association.

That's the summary of our research about the wordle game. We sincerely hope that our work will help you deal with problems. If you want to know more information, please contact our team.

Yours sincerely,
Team #2318006

10 References

1. Zhang, S., Jia, Q., Shen, L., Zhao, Y. (2020). Automatic Classification and Comparison of Words by Difficulty. In: Yang, H., Pasupa, K., Leung, A.CS., Kwok, J.T., Chan, J.H., King, I. (eds) Neural Information Processing. ICONIP 2020. Communications in Computer and Information Science, vol 1332. Springer, Cham. https://doi.org/10.1007/978-3-030-63820-7_72
2. Daniel Jurafsky and James H. Martin. 2023. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3st. ed.).
3. Cluster Effect Evaluation- Fmeasure and Accuary and its implementation in Matlab. May, 2015. https://www.cnblogs.com/zhangduo/p/4504879.html
4. Aman Anand. What is a Bigram Language Model? Educative. https://www.educative.io/answers/what-is-a-bigram-language-model
5. "English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU". norvig.com. Retrieved 2019-10-28. https://norvig.com/mayzner.html
6. Li, J., Cheng, Jh., Shi, Jy., Huang, F. (2012). Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement. In: Jin, D., Lin, S. (eds) Advances in Computer Science and Information Engineering. Advances in Intelligent and Soft Computing, vol 169. Springer, Berlin, Heidelberg.