Problem Chosen
C

2023
MCM/ICM
Summary Sheet

Team Control Number
2312872

Predictions and Analysis of Wordle Results

Summary

Wordle is a word-guessing crossword game that has always been loved by players because of its fun nature. As this game continues to grow in popularity, some of its deeper patterns have been uncovered, and this has become a research direction. In this paper, by analyzing the data of word attributes and difficulty levels of Wordle game, we classify the difficulty of all puzzles and predict the distribution of players' reports.

For the first part of question 1, at first we preprocess the anomalous data. We use the time series ***ARIMA***(0,1,7) model to build a prediction model about the number of reported results.We predicted that the reported score results on March 1, 2023 were between 18308 and 21737 with 95% confidence interval. The changes in reported scores are also explained using the infectious disease model.

As for the second part of question 1, we extract 4 attributes of words including: information entropy , frequency , number of unfamiliar letters and number of repeated letters.Then we perform ***Spearman correlation analysis*** regarding the attributes of the four dimensions of words and the percentage of reported scores in the difficulty mode. Among them, the information entropy expectation of words, the number of uncommon letter occurrences passed the significance test and had a weak correlation on the percentage.

Regarding question 2, the percentage distribution of the number of user attempts may be related to the word attributes and the percentage of difficulty mode data. We used ***TOPSIS Analysis*** to transform the percentage of user attempts into a series of indicators that quantify the difficulty, and performed a difficulty ranking. The ***GBDT*** regression was performed as the output, and the final distribution of the percentage of attempts for EERIE users from 1 try to 7 or more tries (X) was 0.15%, 3.7%, 18.2%, 34.25%, 28.1%, 13.25%, 2.25%, respectively. in the prediction deviation test, R^2 value is 0.649, indicating that our model is able to predict well.

With regard to question 3, we use ***K-Means Clustering Algorithms*** to quantify the difficulty based on the number of attempts users made to solve the puzzle, which is finally divided into three categories: difficult, medium and easy. By analyzing the word attributes of each category, we find that the words in the difficult category generally have the characteristics of low information entropy, low frequency, high number of unfamiliar letters, and high number of repeated letters. After verification, EEIRE belongs to the category of medium difficulty. Through the distribution of scatter plots, we find that the classification boundaries of the three classes of words are obvious and there are few overlapping parts, which means that our difficulty classification model has a high accuracy rate.

In terms of question 4, we also find some interesting features of the data. For example, fewer people report wordle results on Christmas day than on neighboring dates, possibly due to the holiday causing fewer people to participate in the game on that day. Slate also appeared late, when people have already explored the optimal strategy for solving the problem, resulting in 6% of people guessing correctly in one go.

Ultimately, we write a letter to the editor to summarize our findings and provide some suggestions. We hope this game can be optimized to keep its heat up.

Keywords: ***ARIMA*** ***Spearman Correlation Analysis*** ***TOPSIS Analysis*** ***Gradient Boosting Decision Tree (GBDT)*** ***K-Means Clustering Algorithms***

Contents

1 Introduction

1.1 Problem Background

Wordle is a fun crossword puzzle game. As the popularity of Wordle has grown, players have gone from being unfamiliar to being proficient at first. The game itself has also sparked the curiosity of many players, so many scholars have started to study the deeper rules of the game. By studying the historical answer data of players in the game and the rules of the game's operation, we have had a variety of solutions. Over time, more and more player-reported data flowed out and the game was gradually studied thoroughly.This paper is also based on a large amount of data analysis for the study.

1.2 Restatement of the problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Develop a model to explain the variation of the number of reported results daily and use the model to create a prediction interval for the number of reported results on March 1, 2023. Judge whether any attributes of the word will affect the percentage of scores reported that were played in Hard Modeand then explain how.
- Develop a model that can predict the distribution of the reported results based on the given word. What uncertainty will affect the model and predictions? Use the word EERIE as an example. Explain how confident you have in the model's prediction.
- Develop and summarize a model that can classify solution words according to the difficulty and identify the word attributes associated with each classification.Use the model to determine the difficulty of the word EERIE, discuss the accuracy of the classification model.
- Find some other features from the data then list and description. Finally write a letter to the Puzzle Editor to summarize the research results.

1.3 Our work

2 Assumptions and Notations

2.1 Assumptions

Basic assumptions.

1. Assume that the players who reported Wordle results on Twitter did not cheat
2. Assume that Wordle's answers come from the 2315 words in the source code of The New York Times website
3. Assume that Wordle's daily answers are randomly selected from a thesaurus
4. Assume that players' game strategies remain stable since March
5. Assume that the word frequency reflects the player's familiarity with the word

2.2 Notations

Symbol	Definition
p	Number of autoregressive terms
d	Number of differences performed on the original series
q	Number of moving average terms
Z_{ij}	Standardized index
S_i	Unnormalized score
E_I	Information entropy
$p(x)$	Ratio of words remaining after hints to the total number of words in the thesaurus
k	Total number of categories clustered
n	Number of iterations in clustering process

3 Data Selection and Analysis

Since we are only allowed to use the datasets provided by COMAP official, we need to pre-process the data for both datasets before solving the problem.

In the first part of question 1,we notice an exception in the value of number of reported results on November 30, 2022 and an exception in the value of number in hard mode on February 13, 2022. For both of these anomalies, we use the interpolation method to average the values of the two days before and after as the actual values of that day.

In the question 2 and question 3, we notice that the sum of the percentages of attempts reported by players for the word nymph was 126%, so this data was excluded in the second and third questions.

In the analysis of word attributes, we exclude words with 4 letters and 6 letters.

Based on the above data processing, we performed the problem solving, which greatly improved the accuracy of our model.

4 Task-1

4.1 The Establishment of ARIMA Model

To explore the variation of daily reported scores over time and to give the trend of daily reported scores, the final prediction of reported scores on March 1, 2023 was made. In solving Task 1, first, the original time series with breakpoints in days was used to create a line graph based on the given data, reflecting the trend of daily reported scores in 2022. Then, we selected the Arima model from the time series model to describe the change in daily report-ed scores.

As can be seen from the line graph, the daily report score fluctuates less until mid-February, always showing an upward trend and reaching a peak on February 2. From mid-February until May, the daily report scores show a fluctuating downward trend. From May onward, the daily report score tends to level off and decline. The formula is given in equation 2.

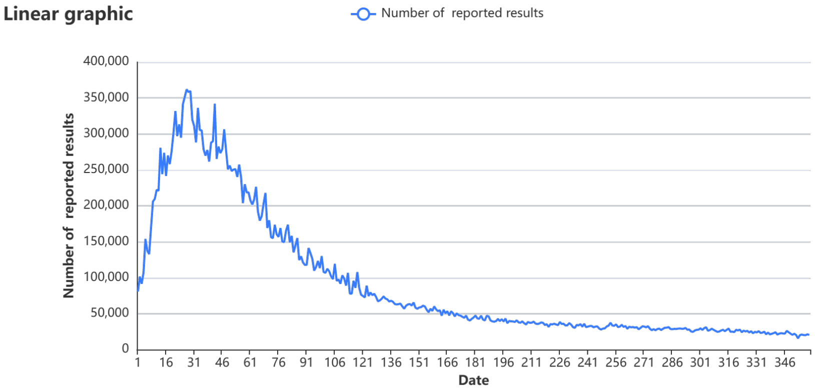

Figure 1: The liner graphic of the number of the reported results

The ARIMA (p, d, q) model is known as the Autoregressive Integrated Moving Average Model, where p is the autoregressive term; d is the number of differences when the time series is stationary; q is the number of moving average items.This model is a combination of autoregressive (AR) and moving average (MA), which can transform a non-stationary time series into a stationary time series, and then regress the lagged values of the dependent variable, the present and lagged values of the random error term to the model established.

Parameter estimation. We estimated the parameters by the least squares method and experimentally fitted them in SPSS software, and finally determined the three parameters in the ARIMA model as $p = 0, d = 1, q = 7$

Model fitness test. By model fitness, we mean that an ARMA model has fully or largely explained the dynamics of the system so that the ε_t in the model is independent. Clearly, the model fitness test is essentially a white noise test of ε_t.

Model	AR(p)	MA(q)	ARMA(q)
ACF	Trailing	Truncation	Trailing
PACF	Truncation	Trailing	Trailing

Table 1: The form of autocorrelation function and partial correlation function of a smooth sequence

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Residual test.This is a plot of the residuals we obtained, with the ACF residuals on the left and the PACF residuals on the right. The autocorrelation coefficient ACF and the partial autocorrelation coefficient PACF are not significantly different from 0 for all lag orders. We conclude that the residuals are white noise series.

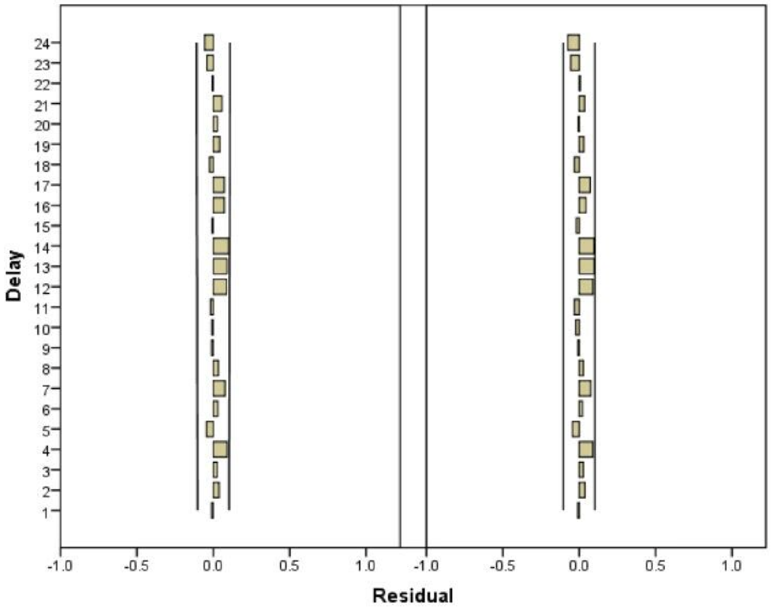

Figure 2: Residual Plot

4.2 The Solution of ARIMA Model

First, we rearranged the table row order by time. We noticed that there were outliers on November 30, which were removed to ensure the validity and accuracy of the data. And then, we integrated all the data and analyzed them, imported them into SPSS, and plotted line graphs to obtain the change in daily reported scores for 2022.

Plotting a line chart allows visualizing the evolution of the daily report scores, allowing to see the trend of the report scores in the time dimension.

The following figure shows the line graph based on the data given for the daily reported scores for 2022.

The prediction results of the ARIMA model are shown in Figure 3. From Figure 3, we can see that the ARIMA model is very accurate in predicting the number of reported results, which is especially obvious in the upward trend of the change. Also, the ARIMA model is still able to capture the trend of the reported result changes well, which means that it can predict well about the linear part of number of reported results.

Based on the forecasts of the ARIMA model above, we believe that the number of reported results in the range of [18308,21737] on March 1, 2023.

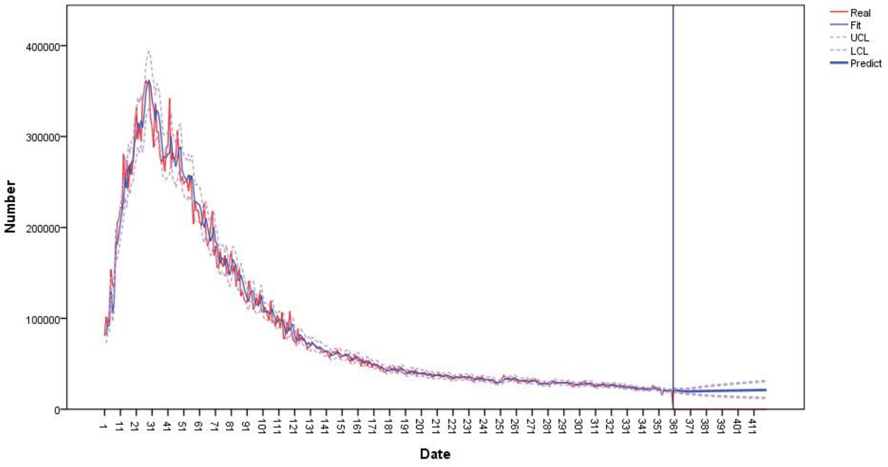

Figure 3: Prediction graph of ARIMA model

4.3 Definition of Word Attributes

Regarding the properties of words, we set the following 4 metrics.

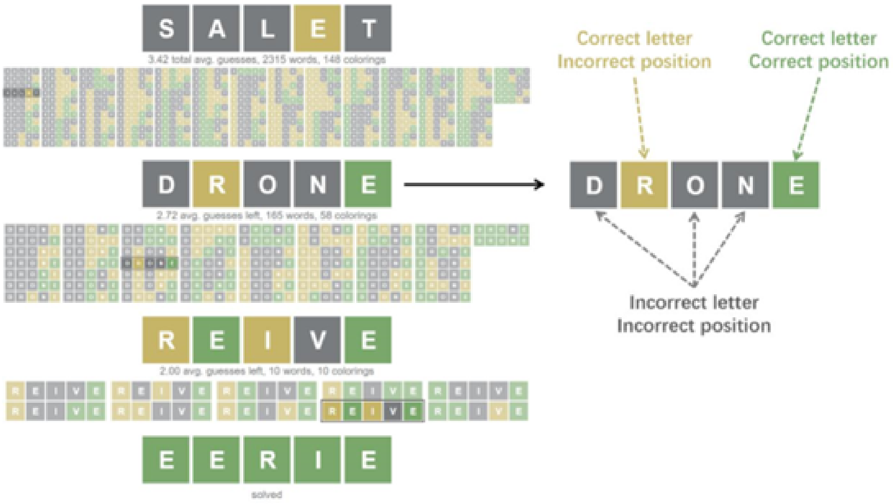
SALET
3.42 total avg. guesses, 2315 words, 148 colorings
DRONE
2.72 avg. guesses left, 165 words, 58 colorings
REIVE
2.00 avg. guesses left, 10 words, 10 colorings
EERIE
solved
Correct letter
Incorrect position
Correct letter
Correct position
DRONE
Incorrect letter
Incorrect position

Figure 4: The liner graphic of the number of the reported results

1.Information entropy:Its calculation formula is as follows.

$$E[I] = \sum_{x} p(x) \cdot \log_2 \frac{1}{p(x)}$$

As shown below, when the first word is entered, Wordle will display the color of different letters according to the answer word. The letters ’s’, ’a’, ’l’, ’t’ are gray in color and the letter ’e’ is yellow, so we will exclude all words containing the letters ’s’,’a’,’l’ ,’t’, without the letter ’e’. The word with the letter ’e’ in the fourth position leaves us with 165 words to choose from. By calculating

$log_2\frac{1}{2315} - log_2\frac{1}{165}$, we get the amount of information that the color result of the word salet gives us. According to the list of all possible answers for all words, the color distribution of salet has a total of 148 cases, each of which provides a different size of information entropy, which we use to get the expected value of slate information entropy.

2.Frequency: Calculated from Google's ebook dataset and referenced to mathematica's word frequency table to quantify the frequency of use for each five-letter word.

3.Number of unfamiliar letters: We counted 2315 frequently used five-letter words and clustered the number of occurrences of 26 of them, and found that certain letters appeared the least frequently. By treating these letters as unfamiliar letters, the number of unfamiliar letters in each word will then be used as a word attribute.

Please refer to the following genealogy chart for the definition of unfamiliar letters: Based on the number of occurrences of each letter in the common five-letter words shown in the genealogy chart, q v are defined as unfamiliar letters from top to bottom in the genealogy chart.

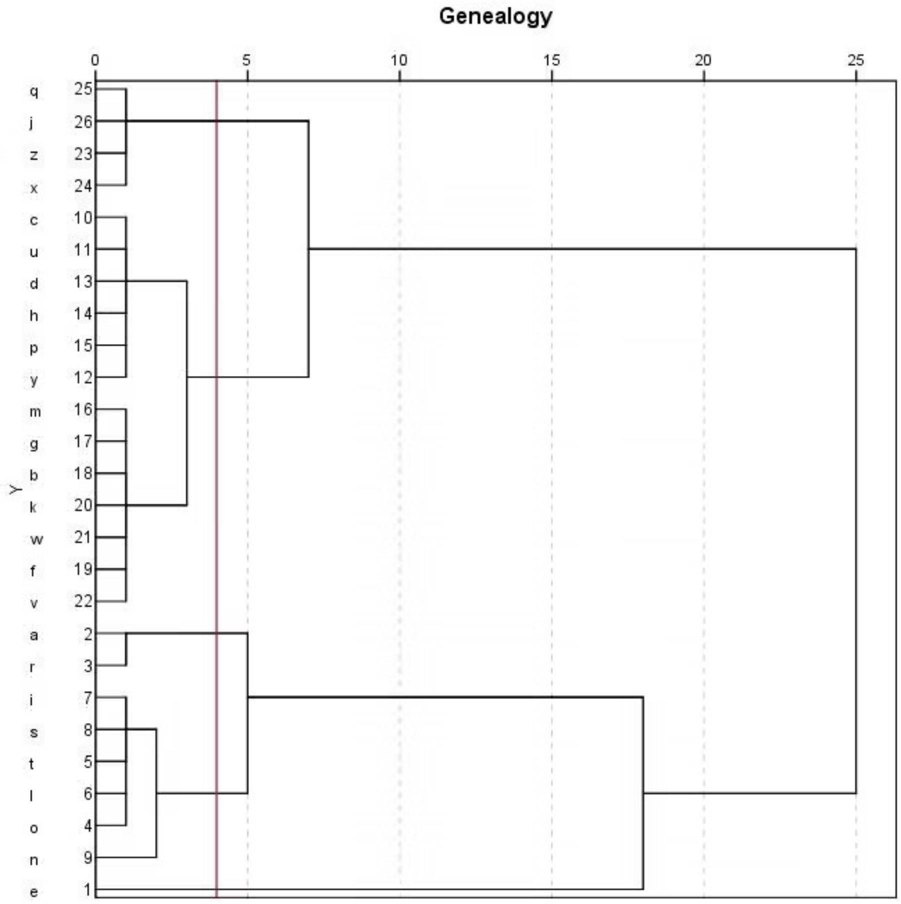

Figure 5: The liner graphic of the number of the reported results

4.Number of repeating letters: We counted the repeating letters of the given words. In particular, if two letters appear twice in a five-letter word, such as cacao, the number of repeated letters is defined as 2. If a letter is repeated three times, the number of repeated letters is defined as 2.5.

4.4 Correlation analysis

In analyzing the percentage of word attributes on the reported scores in the difficulty mode, we used Spearman correlation analysis.

The formula for calculating the Spearman correlation coefficient:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

By analyzing the correlation between the four indicators of word attributes and the percentage of reported scores in the difficult mode, we obtained the following results.

Table 2: Results of Spearman correlation analysis

	Correlation coefficient	Significance	P
Information entropy	-0.124	5%	0.020
Frequency	-0.059	0	0.267
Number of unfamiliar letters	0.153	10%	0.004
Number of repeated letters	0.085	0	0.111

According to the above table, the information entropy and the number of unfamiliar letters passed the significance test, and the correlation coefficients were both greater than 0.1. Therefore, we conclude that the percentage of reported scores in the hard mode showed weak correlation with the information entropy and the number of unfamiliar letters.requency and the number of repeated letters, on the other hand, had no effect on the percentage of scores reported in the hard mode.

5 Task-2

5.1 Basic Analysis

In order to assess the specific difficulty of each word, we consider using the number of attempts per user to represent the difficulty level. In order to evaluate the specific difficulty of each word quantitatively, we decide to use TOPSIS analysis.

5.2 The Establishment and Solution of Topsis Model

For the selection of metrics in the TOPSIS analysis, we decided to use the percentage of users' first four attempts as 4 very large metrics and the percentage of users' last three attempts as 3 very small metrics

In the TOPSIS analysis method, all indices are required to be orthogonalized, which means, transformed into maximum indices. Since the activity of nuclear weapons research is exactly proportional to the two indices, no index transformation is required.The matrix of orthogonalized indices composed of the above data is as follow.

$$x_{ij} = \begin{bmatrix} 1 & 3 & 23 & 29 & 20 & 28 & 47 \\ 1 & 5 & 23 & 31 & 202 & 23 & 46 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 2 & 17 & 37 & 15 & 25 & 46 \end{bmatrix}$$

Next, the standardization of the orthogonalized indices is performed. The purpose of standardization is to eliminate the effect of different scales. The standardization formula is as follows:

$$Z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}$$

where i represents the number of rows and$j$represents the number of columns.

We normalised the normalisation matrix according to the above formula. The results of the standardised matrix are as follows:

$$Z_{ij} = \begin{bmatrix} 0.058421 & 0.022668 & 0.050994 & 0.062185 & 0.050229 & 0.056875 & 0.05504 \\ 0.058421 & 0.037779 & 0.050994 & 0.049429 & 0.050229 & 0.046719 & 0.053869 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0.015112 & 0.037691 & 0.058996 & 0.037672 & 0.050781 & 0.053869 \end{bmatrix}$$

Define the maximum value:

$$Z^{+} = (max\{z_{11}, \ldots z_{17}\}, max\{z_{21}, \ldots z_{27}\}, \ldots, max\{z_{(354,1)}, \ldots z_{(354,7)}\})$$

Define the minimum value:

$$Z^{+} = (min\{z_{11}, \ldots z_{17}\}, min\{z_{21}, \ldots z_{27}\}, \ldots, min\{z_{(354,1)}, \ldots \cdot_{(354,7)}\})$$

Define the distance betweenz_{ij} (i=1,2,... ,354) and the maximum value Z^{+}.

$$D^{+} = \sqrt{(\sum_{j=1}^{7} Z_j^{+} - z_{ij})^2}$$

Define the distance between z_{ij}(i=1,2,... ,354) and the maximum value Z^{-}.

$$D^{-} = \sqrt{(\sum_{j=1}^{7} Z_j^{-} - z_{ij})^2}$$

+ Accordingly, the unnormalized score of each country is calculated as the formula listed below:

$$S_i = \frac{D_i^{+}}{D_i^{+} + D_i^{-}}$$

Obviously, $0 \leqslant S_i \leqslant 1$, which means, the smaller the S_i is, the closer to the maximum value. After calculating the score for each word according to the formula, we normalized all word scores. The final ranking of each word was obtained and the results are as follows:

Table 3: Rank of the Difficulty of Words

Word	Scores	Rank
train	0.91753	1
slate	0.82730	2
dream	0.76447	3
feast	0.73596	4
⋮	⋮	⋮
coyly	0.08423	352
gawky	0.08363	353
fluff	0.8224	354

5.3 GBDT

GBDT stands for Gradient Boosted Decision Trees, a machine learning technique that combines multiple weak models (typically decision trees) into a strong model. It works by fitting new models to the residuals (errors) of previous models, and then adding them together.

We used Information entropy, .Frequency, Number of unfamiliar letters, Number of repeating letters, the percentage of reported scores in the difficult mode as the dependent variables and the data after the topsis treatment of the association percentages as the output. After the experiment, we found that the results obtained by machine learning were better by selecting the data from the last 300 days of the dataset, probably because the wordle fever reached its peak in February 2022, and a large number of wordle solution strategies appeared, which changed the solution strategies of wordle players to some extent. Fourteen days were randomly selected as the prediction set in the dataset, and the prediction results are shown in the following figure. Assuming that the percentage of reported scores in the difficult mode on March 1 is 9.7% (from the December average of this data).

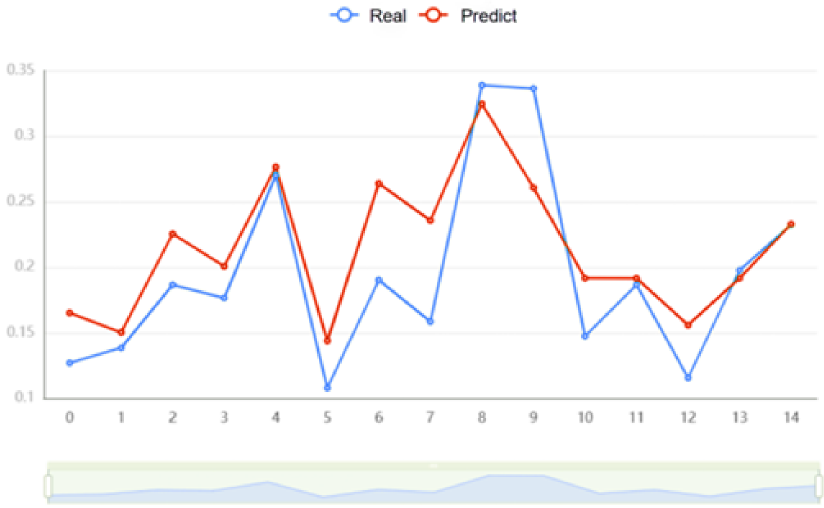

Figure 6: The liner graphic of the number of the reported results

We get the predicted result of EERIE as 0.186440691, after sorting we take the distribution of the association percentage of the 20 words adjacent to EERIE and average them to get the distribution of the reported results of EERIE.

Table 4: The Prediction for the Distribution of the Word EERIE

1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)	total
0.15	3.7	18.2	34.25	28.1	13.25	2.25	99.9

In terms of prediction accuracy, we use five common indicators to evaluate the prediction results of GBDT to analyze the deviation between the predicted value of dependent variable and the difficulty of the prediction. The MSE, RMSE, MAE, MAPE and R2-score results calculated by python are shown in the Table below.

Table 5: Evaluation of the GBDT prediction results

Predict value	MSE	RMSE	MAE	MAPE	R^2
Training set	0.001	0.024	0.019	8.898	0.947
Test set	0.002	0.043	0.033	15.634	0.649

The R^2 value for our test set is 0.649, indicating that our model can better respond to the relationship between the five dependent variables we extracted and the difficulty.

6 Task-3

6.1 Basic Analysis

We will make the classification of different word attributes based on difficulty, and judge EERIE the word belongs to a class. Further,in a bid to cater to the reasonableness of the classification model and sensitivity analysis, we established the word type K-means system clustering model .

6.2 The Establishment and Solution of K-means Model

The K-means algorithm is one of the division methods in the clustering algorithm The Euclidean distance is used as the evaluation index of similarity, that is, the closer the dis-tance between 2 The closer the distance between two samples, the greater the similarity. The form of Euclidean distance is as follows.

$$D(\mathbf{X},\mathbf{Y}) = \{\Sigma_i |x_i - y_i|^2\}^{\frac{1}{2}}$$

k points are randomly selected as the initial clustering centers; then the samples are grouped into categories according to the distance from each samples into classes based on the distance from each sample to each cluster center; recalculate the center of each class (i.e., the average of all points in the class, i.e., the geomet-ric center) of each class, and again group the samples according to the distance from the cluster center, as This cycle is iterated until the squared error criterion function is stabilized at the minimum value.

$$J_c = \sum_{1}^{k} \sum_{i}^{n_j} ||x_i - m_j||^2$$

In this equation,k is the number of clusters to be formed; n_j is the number of samples in the jth class m_j is the mean value of the samples in class j, representing the centre of the data set of this type. of this type of data set.

This is the flow of the K-means clustering algorithm, and we strictly follow this flowchart for clustering.

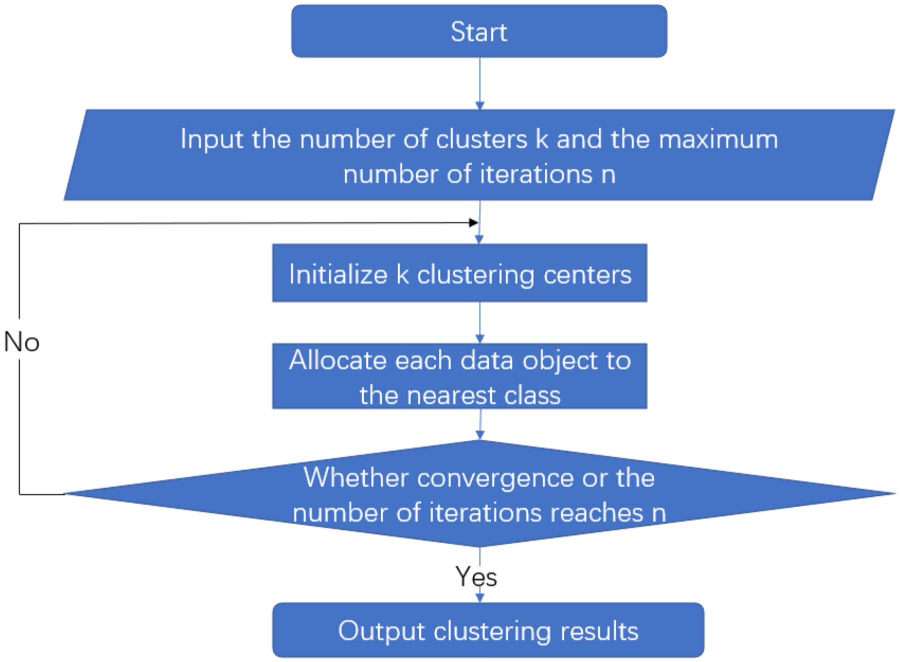
Start
Input the number of clusters k and the maximum number of iterations n
Initialize k clustering centers
Allocate each data object to the nearest class
Whether convergence or the number of iterations reaches n
No
Yes
Output clustering results

Figure 7: Clustering algorithm process

This is the folding graph of aggregation coefficient. According to this line graph, it can be seen that the degree of distortion of the aggregation coefficient is greatest when the value of the horizontal coordinate k is in the range of 1 to 3. When the value of k exceeds 5, the degree of aggregation coefficient variation decreases significantly. Therefore, when the number of categories is set to $k = 3$, the decline of discounting tends to slow down, so the number of categories can be set to $k = 3$.

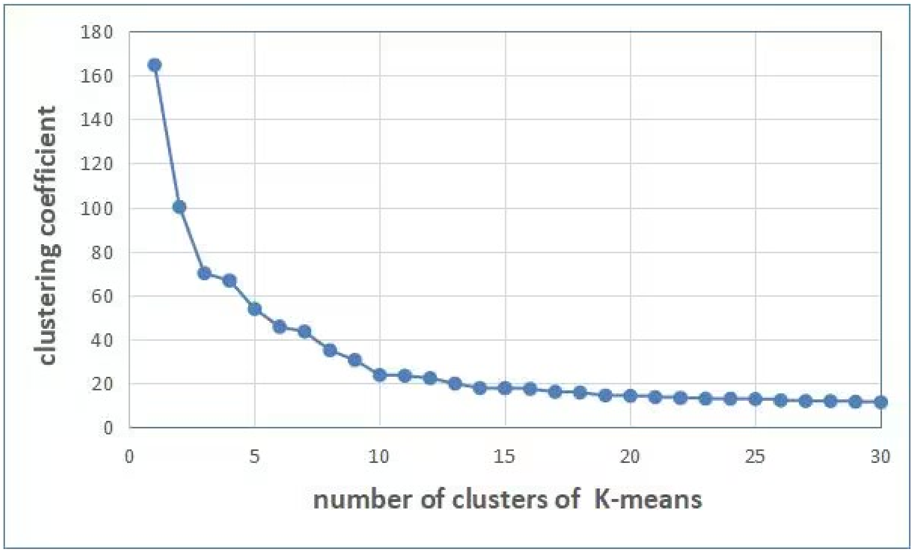

Figure 8: Aggregation coefficient line graph

In the clustering process, we first selected the initial number of clustering centers as $n = 3$.The initial clustering centers are as follows:

Table 6: Initial Clustering Center

	1	2	3
1 try	6	0	0
2 tries	26	0	0
3 tries	32	4	4
4 tries	22	21	25
5 tries	10	15	44
6tries	3	2	23
7 or more tries (X)	0	48	4

We set the number of iterations $n = 10$, and after the completion of the iteration, the final clustering center is generated.Here is the final clustering center.

Table 7: Final Clustering Center

	1	2	3
1 try	1	0	0
2 tries	9	3	4
3 tries	30	12	19
4 tries	34	25	35
5 tries	18	29	27
6tries	7	22	12
7 or more tries (X)	1	9	2

After the clustering centers were determined, all words were clustered based on the number of attempts, and they were classified into 3 categories of words: difficult, medium, and easy. The clustering results are as follows.

Table 8: Final Clustering Results

Category	Number
1	152
2	60
3	142

6.3 The Attributes Associated with Each Classification

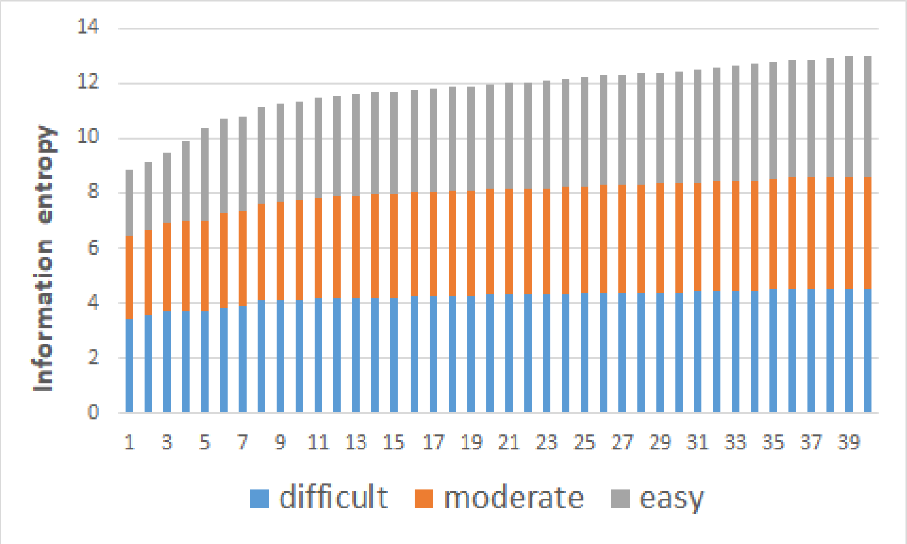

Figure 9: Aggregation coefficient line graph

From the above figure, we can see that the information entropy of the difficult class of words is generally low, and the information entropy of the easy class of words is generally high, which indicates that difficulty is negatively correlated with information entropy, which is a property of the words we selected.

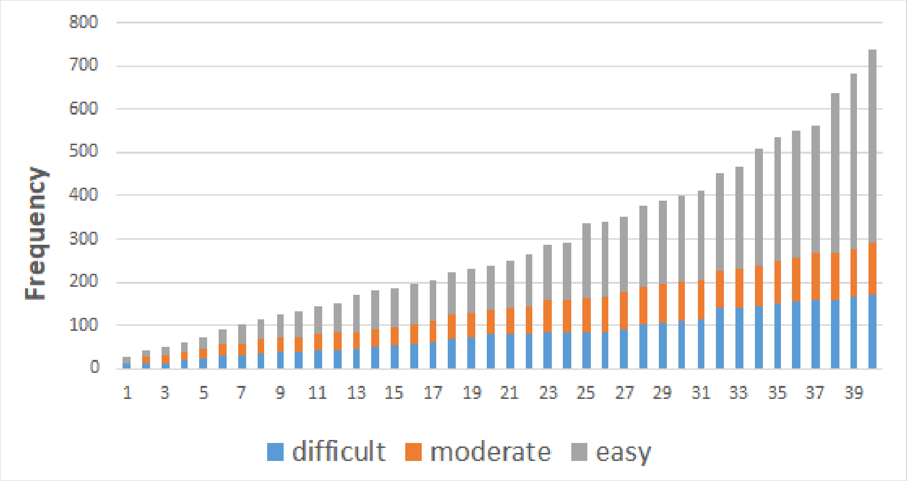

Figure 10: Aggregation coefficient line graph

As we can see from the frequency graph, the classification of difficult words generally occurs less frequently and the classification of easy words generally occurs more frequently. This indicates that the difficulty of words is negatively correlated with the frequency of occurrence.

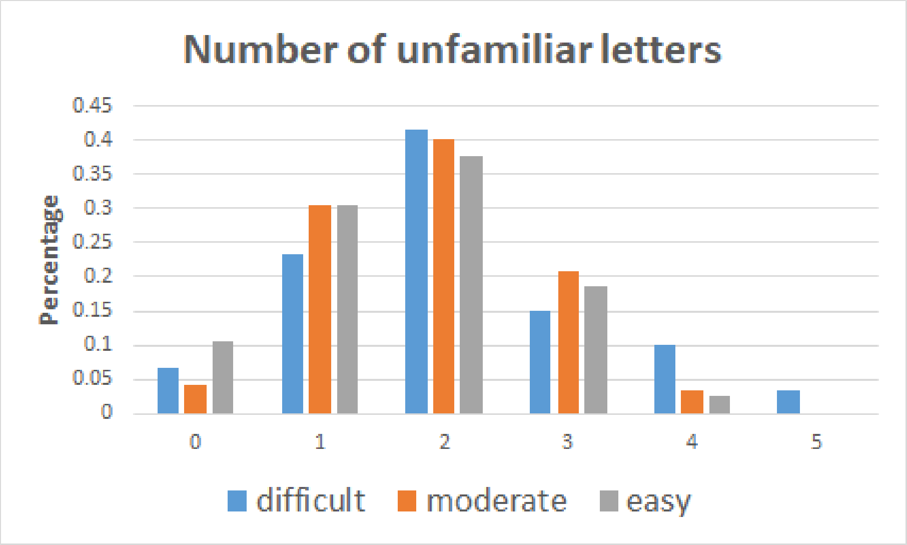

Figure 11: Aggregation coefficient line graph

From the above graph, we can find that the simple category of words usually has a small number of unfamiliar letters between 0 and 2, while the words with 5 unfamiliar letters are all difficult category words. We conclude that the difficulty is positively interrelated with the number of unfamiliar letters.

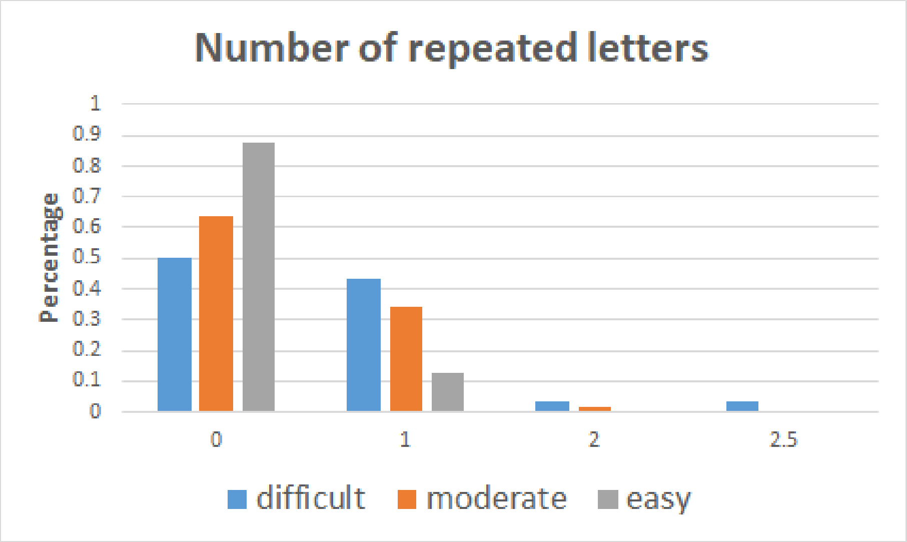

Figure 12: Aggregation coefficient line graph

Observing the above graph, we notice that the words in the easy category all have 0 or 1 repetition letters, while the words with 2.5 repetition letters are all in the difficult category.

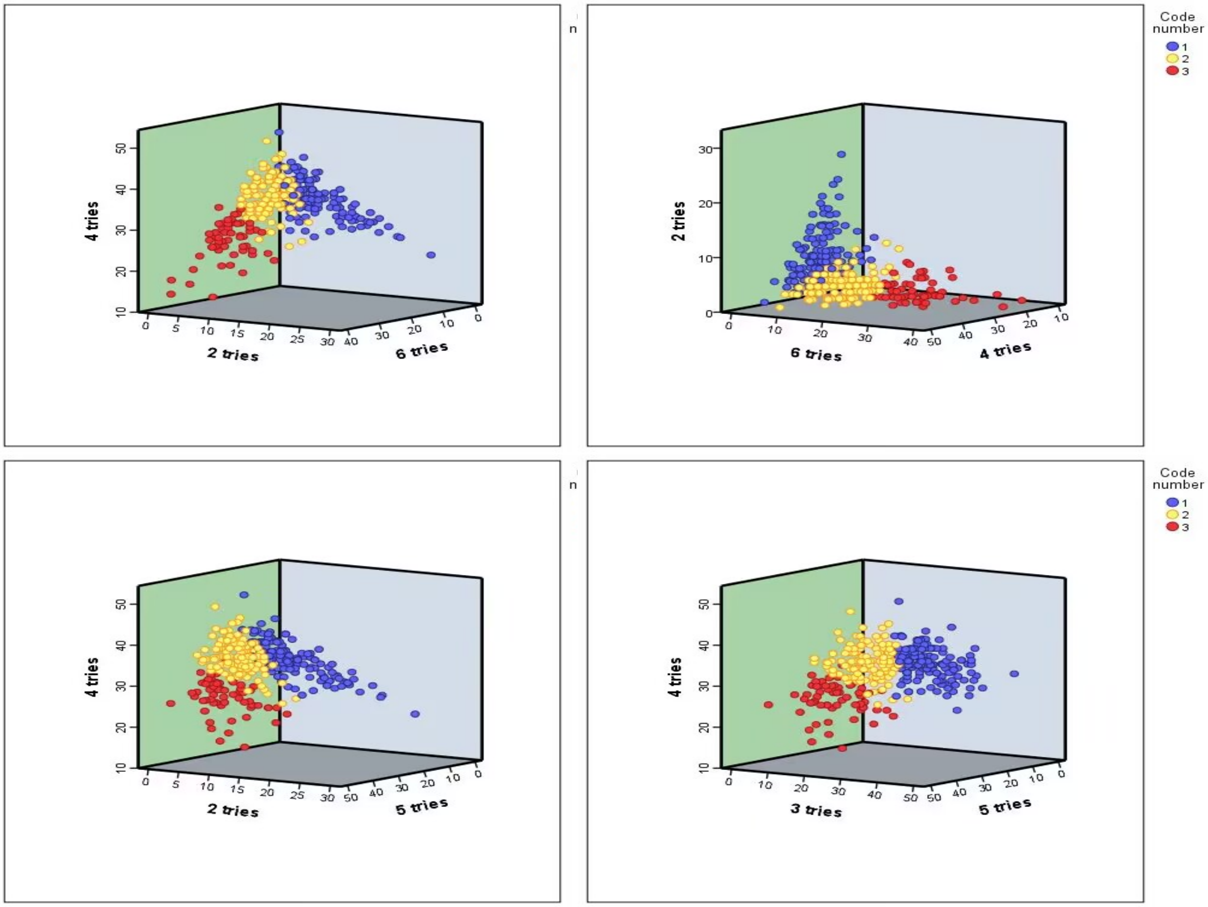

Figure 13: Scatter plot reflecting the accuracy of the classification model

For the above four scatter plots, we set different number of attempts to reflect the difficulty level of each category of words, which is also our indicator for classification. The above four scatter plots show that there is almost no overlap or coverage between the scatter points, and the distinction between each class is obvious. And it can be found that the difficult category corresponds to a high number of attempts and the easy category corresponds to a low number of attempts. Therefore, our classification model is very accurate.

According to the above difficulty classification model, we consider the word EERIE will be in the moderate classification.

7 Task-4

We found that on Christmas Day, December 25, the number of participating players reported a significant decrease, with a huge difference from the two days before and after. Our speculation on this is that because Christmas is the most important holiday of the year in the United States, people enjoy the holiday and do not have time to participate in Wordle games.

We found that the word slate appeared on December 19. The number of players who guessed it correctly on the first attempt was around 6%, which is out of the ordinary. Our speculation is that after a long time of playing the game, players have figured out a set of solution strategies in which slate is known as the universal beginning word. Therefore, many players follow the optimal strategy to solve the problem and will guess the word slate first.

8 A letter to the Puzzle Editor of the New York Times

Dear Editor,

Nowadays, Wordle is a popular puzzle currently offered daily by the New York Times.The version of the game now available in over 60 languages, which illustrates the popularity of the game. This game not only adds fun to everyone's life, but also examines everyone's vocabulary ability and reaction capacity, because we must guess a real English word, and every guess we will receive feedback, and according to the feedback to make the next guess. This game is so fun and our group was very interested in this game, and we analyzed the data of the Worldle game on Twitter over the past 354 days(outliers were removed), and found some interesting things.

First of all, we found that in the early days of the release of the game, the number of players increased exponentially. By February 2022, the popularity of the game reached its peak, and then the popularity of the game began to decline over time, to about July 2022, the number of players tended to level, with only a small decay trend. In addition, we also found that with the release of the game, the percentage of scores reported that were played in Hard Mode increased from 1.69To predict the distribution of the word-reported results for a given future solution word on a future date, we developed a model that can predict the distribution of report results more accurately for just a given word answer on any day in the future. In other words, the relevant percentage of the predicted future date (1,2,3,4,5,6, x). To illustrate our model, such as the word EERIE on March 1, 2023, we predict that the distribution of report results is (0.15,3.7,18.2,34.25,28,13.25,2.25).

What's more, we used K-means clustering algorithm to classify words, and we use the elbow method to find the best categories,which is shown that these words are more suitable to fall into 3 categories, and after 10 iterations, we get the final result:three categories were easy, moderate and difficult. Later, we verified the accuracy of the model for the three types of data, and found that the accuracy of the classification model was very high. In this model, EERIE is of moderate difficulty. Later, we analyzed the attributes of words and found that the attributes of the three categories were greatly different. The results were: words with easy categories are common words and are more often used in our daily lives; words in the difficult category had the same letter, and more low-frequency letters were used; and the attributes of the moderate category were between the easy words and difficult words.

Finally, we found some interesting data, such as Christmas Day, the number of players was small, showing that some large festivals affect the number of participants. In addition, if you want to improve the difficulty of the game, you can change the selection of words, such as choosing words with low frequency, the words can contain more letters like q, j, z, x and other rare characters, at the same time, the words can contain the same letter, which can also increase the difficulty of the game. If you want to reduce the difficulty of the game, you can choose some common words, which can contain more characters like e, a, r, i and other common letters, and the words try not to contain the same letter.

Above is the result of our research on Worldle game data, I wish Worldle games better and better.

References

[1] Wolfram.Word frequency source URL:https://reference.wolfram.com/language/ref/WordFrequencyData.html

[2] The best strategies for Wordle. URL:https://sonorouschocolate.com/notes/index.php?title=The_best_strategies_for_Wordle

[3] Chen Fuji,Hu Gaili. On the Spread Pattern of Network Public Opinion Hot Topics.IN:Institute of Public Management 2014,33(01):97-101+156

[4] Barrysmyth.Three Million Wordle Tweets Later.URL:https://towardsdatascience.com/three-million-wordle-tweets-later-3d3af23bd5c7

[5] Alcolea Adrián, Resano Javier.FPGA Accelerator for Gradient Boosting Decision Trees.IN:Electronics 2021, 10(3), 314