Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2300466

Wordle Data Prediction and Evaluation Model via Information Theory and Propagation Dynamics

Wordle, a small game with a simple drawing style, has recently gone viral. So that Wordle players are Posting or tweeting about their reports on various social platforms. In this paper, a variety of word attributes such as word frequency and information entropy are selected for modeling, so as to provide answers to the question and give some suggestions for the setting of the riddle.

For problem one, first, we corrected and eliminated the abnormal data.Considering that the data does not have the condition of time stationarity, we established the **Wordle Information Propagation Model** based on **Complex Network Propagation Dynamics**, and used **Attenuation Model** and other models to predict the future number of reports, and obtained the range of the number of reports on the same day as [17330,20900]. We combine **Information theory** and use **Word frequency**, **Information entropy** and other word attributes to quantify words. **Spearman** correlation analysis shows that the higher the Word frequency, Information Entropy and Popularity of a word, the higher the number of attempts. The higher the degree of repetition of the word letter, the lower the number of attempts.

For problem two, we used **Multiple stepwise regression**.Using word frequency, information entropy, popularity and other information as independent variables, the distribution of the number of attempts is analyzed by regression analysis. However, considering individual player differences, proficiency, Internet hot words and other uncertain factors, the model has a certain deviation. Compared with previous data, the average error of the model is less than $\pm$ 2%. The final predicted EERIE report scores were 0.38, 4.63, 15.86, 29.21, 27.91, 17.45, 5.46.

For problem three, we determined by the elbow rule that when the number of clustering centers is 3, the classification effect is optimal. The sample data is divided into three categories: difficult, medium and easy by **K-means** algorithm. According to the prediction of the word EERIE in question two, it is classified as "difficult". Finally, **BP neural network** classification algorithm is used to verify the classification results of the model, ensuring the accuracy of the classification evaluation model.

For problem four, in studying the above question, it was found that the number of hard mode reports as a percentage of the total number of players increased over time. We also found that the word part of speech had no effect on the difficulty of guessing. And we found that a more interesting phenomenon is that the number of times different letters appear in a puzzle can even affect the difficulty of the puzzle.

Keywords: **Propagation Dynamics; Entropy; K-means; Stepwise Regression**

Contents

1 Introduction

1.1 Problem Background

As a word guessing game, Wordle provides its players with an actual English word as a puzzle every day. Players were welcomed to guess the word in up to six attempts, using the change in tile color after submission as a cue.

Players can choose to play in Normal mode or Hard mode. The hard mode requires that once the player finds a correct letter in a word (the tiles are yellow or green), those words must be used in subsequent guesses. Given the numbers reported by many users on twitter, we did a regression analysis and forecast for the number of reported results, and built a rating model for the game words. According to the relevant characteristics of different words, the "difficulty" is classified, and the classification accuracy is discussed and analyzed.

1.2 Restatement of the Problem

- Build a model to explain the change in the number of reported results and predict the range of reported results by March 1, 2023. Explore and explain how various word attributes affect the number of attempts.
- Given a certain date and daily puzzles, a model is built to predict the percentage of each attempt. And evaluate the advantages and disadvantages of the model.
- Establish a model, classify the difficulty of solving the problem according to the attributes of each word, and judge the difficulty category of EERIE words.
- Further explore other interesting features of the data set.

1.3 Problem Analysis

For problem one:

According to the data in the attachment, the time sequence chart of the Number of reported results since January 7, 2022 is drawn. It can be seen that the Number of reported results shows a trend of first rising and then declining over time. The Number of reported results also reflects the number of people who play the game, and given the nature of the game itself, we consider this phenomenon to be related to the life cycle of the game itself.

Therefore, several prediction models such as Broadbent-Treisman Parameter Model and Wordle information propagation model based on complex network propagation dynamics were established to predict and analyze the previous data. The number of reports is not only determined by time, but also needs to optimize the model by combining the word difficulty and the number of attempts distribution. Finally, the number of reports on March 1, 2023 is predicted to get the forecast range of the number of reports on that day.

In order to explore the influence of Word attributes on the number of attempts, we studied the correlation between Word Frequency, Information Entropy, Popularity and other attributes.

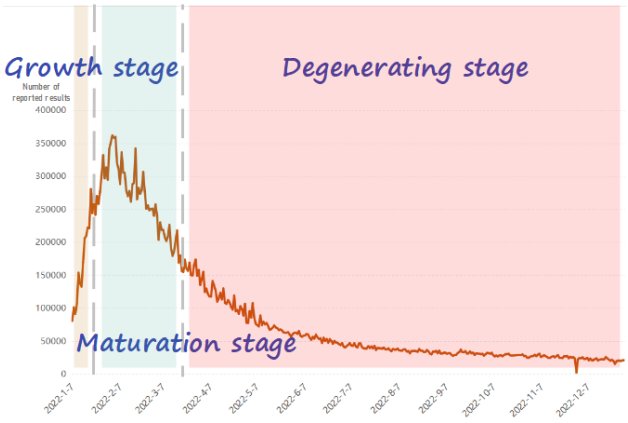


Figure 1: Wordle life cycle

For problem two:

Firstly, an overall analysis of the data should be conducted. Kernel density estimation charts, scatter charts and histograms should be drawn for the percentages of different attempt times and word information entropy, word frequency and popularity. In order to avoid the influence of multicollinearity problem on the prediction model, multiple stepwise regression is selected for modeling. Word frequency, information entropy, popularity, number of letters, classification of repeated letters and other information were taken as independent variables, and the prediction model was obtained by stepwise regression of the percentage of different attempts. At the same time, the influence of uncertain factors on the model should be considered, and the effect of the model should be tested and analyzed to predict the percentage of different attempts of EERIE words.

For problem three:

In order to classify the riddle words according to the difficulty of guessing, we first plan to classify the sample data by K-means algorithm. After the classification of sample data, the difficulty is defined as the evaluation standard of vocabulary to conduct a more detailed quantitative analysis of the classification results. Combined with the prediction of each percentage of EERIE report results in question two, the difficulty is divided. Finally, the accuracy of classification results is verified by dividing training set and test set into sample data.

2 General Assumptions and Notations

2.1 Assumptions

- Assume a positive linear correlation between the number of reports and the number of actual visitors.

- Assume that all twitter users are likely to see the game.
- Assume that there is little difference in individual word cognition.

2.2 Notations

Symbol	Description
N	Twitter user numbers
$S(t)$	Untouched game users
$E(t)$	Temporarily unplayed players
$I(t)$	Players who are playing
$R(t)$	Players who give up
$Rnum(t)$	Number of reports for the day
$Dif(t)$	Average number of attempts for the day
$Try_2(t)$	Percentage of two attempts for the day
β	Probability of seeing a tweet tour
σ	Return rate of old players
γ	Player departure rate
k	Propagation rate of tweets
z	Probability of sending tweets

3 Analysis and prediction of Number of reported results

3.1 Data preprocessing

Data availability must be ensured before data analysis. Firstly, we cleaned the data in the attachment and modified and eliminated individual outliers.

3.1.1 Modification of abnormal words

Through search all data, according to Wordle game rules, the basis of word length should be five. Therefore, we will select the words whose length does not conform to five and replace them with correct words by comparing the data on Wordle's official website with the known data of the title. The replacement result is as follows:

Table 1: Correlation result

Contest number	Abnormalword	Revised
314	tash	trash
525	clen	clean
545	rprobe	probe

3.1.2 Correction of abnormal values

By calculating the percentage of players in hard mode as a percentage of all players with Contest number of 529 words, we clearly find that the percentage of players participating in hard mode as a percentage of all players is significantly outlier than the other sample data.

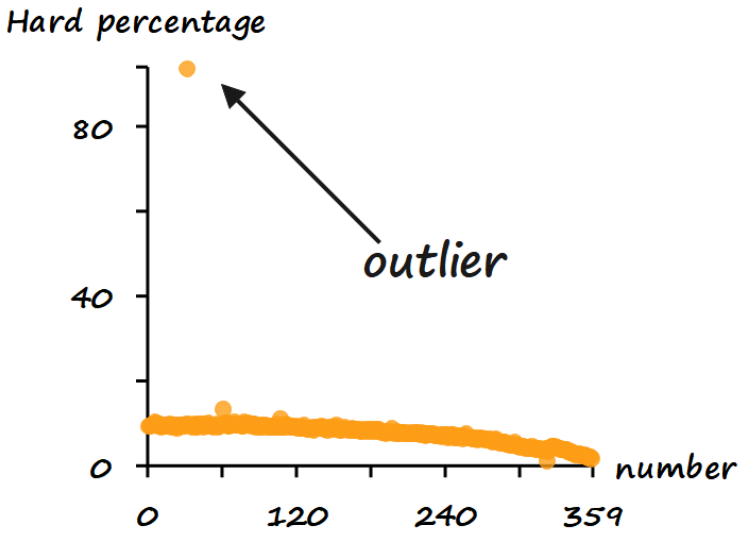


Figure 2: Outlier scheme

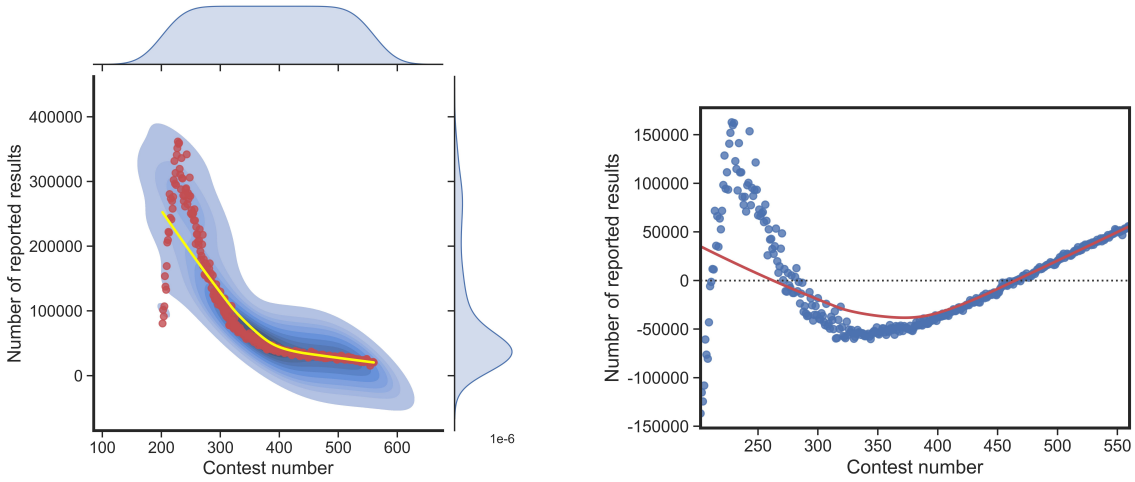
And then we treat that as an outlier. Combined with the data of the first and second days, the interpolation algorithm was used to complete 25569, and the correction was completed.

3.1.3 Delete abnormal data

Given that they're saying that the percentage plus the number of attempts a player makes may differ a little bit from 100% because of some calculation error. After we checked all the data, we found that except for the data with Contest number 281, the number of attempts added up to 126%, the rest of the data was within the range of 100%±2%. Therefore, it is regarded as statistical error and excluded as outlier. At this point, the preprocessing of the data given in this question is over.

3.2 Establishment of model

Contest number is an index of the Wordle puzzles, so Contest number can be regarded as a sequence that has no effect on the data. We plot the number of reported results by Contest number as a scatter plot and plot the overall trend of the data growing with Contest number. It can be seen that the Number of reported results reached a peak in early February. As the number of contests grows, the Number of reported results in the future decreases. Below is the kernel density estimate (KDE) graph, and the linear regression residual graph.



As can be seen from the information in the chart above, the Number of reported result is not simply a set of time-series meaningful data, but something deeper. So we can analyze and predict it using the Broadbent-Treisman parameter Model, complex network propagation dynamics model and other mathematical models.

3.2.1 Broadbent-Treisman Attenuation Model

The traditional filter model is the process of radio wave being blocked by different substances in space propagation, which leads to the blocking of signal. Treisma improved the model and proposed a attenuation model, which believes that the filter does not only allow the information of one channel to pass through, but also the information of other channels can be weakened by just being attenuated.

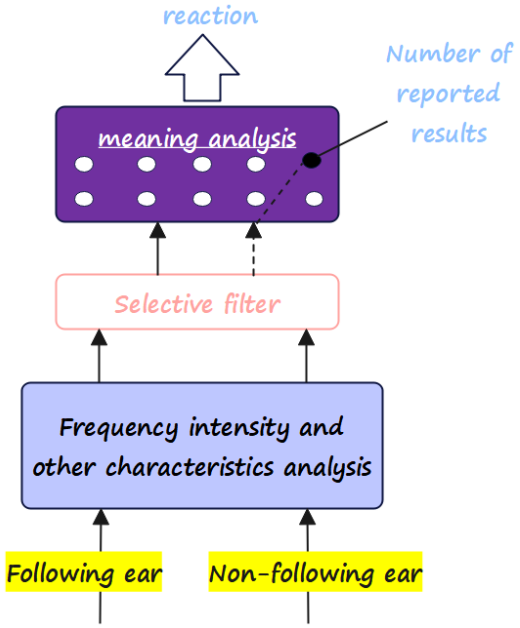


Figure 3: Attenuation model flow

We define Wordle player a to tweet the guess data as event q, and the effect of q on player a is u, discuss the average influence of q on society T_q.

If the probability that a random user b sees the post and becomes interested in it is $r \in (0, 1)$, then the information decays as $f(r) = e^{-\eta r}$ times when propagating from a to b.

Then the impact on person b is:

$$T_a(b) = f(r)u \tag{1}$$

By abstracting player a as a point a^* on the plane, we can assume that society is a unit circle O centered at point a^*, with all people evenly distributed on the surface. Since the probability of player a's tweet being seen by any other user is relatively average, the average influence of each person can be used to represent the influence of event q on society. That is:

$$T_q = \frac{\int_0 uf(r)d_0}{\pi} = u\int_0^1 f(r)dr \tag{2}$$

Furthermore, the above information attenuation model was used to simulate the change of the Number of reported results. The exponential smoothing method was used to predict and fit the attenuation trend of the Number of reported results, and the goodness of fit was 0.91. It shows that the fitting effect of the model is good.

3.2.2 Wordle Information Propagation Model based on Complex Network Propagation Dynamics

The problem calls for a more reasonable explanation of the daily variation in the number of reported results. We built a network information propagation model for Wordle games to explain the changes in the reported quantity in the existing data, so as to predict the future quantity changes.

The spread of a game or piece of information goes through periods of growth, maturity, decline, and stability. The reported number is the number of Wordle players who played and tweeted that day. We believe that the number of tweets is mainly affected by the number of players, and that more tweets attract more players. Therefore, we establish Wordle information propagation model based on complex network propagation dynamics.

When the average Twitter user sees a tweet about the game, there is a $\beta\%$ chance that they will choose to become a new Wordle player. When a player plays a game, he will send a tweet report $z\%$ of the time and become a regular player. Potential players have a $\sigma\%$ chance of returning to play, or a $\gamma\%$ chance of leaving after multiple plays. The flow chart of information dissemination is as follows:

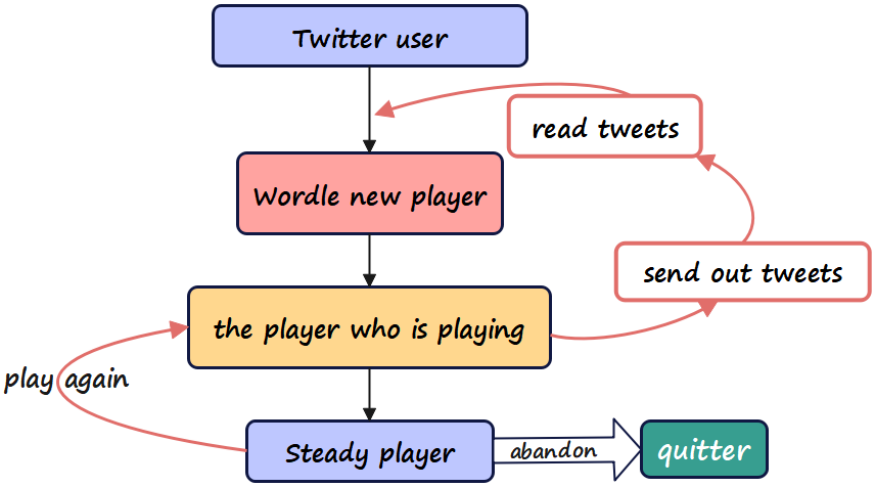


Figure 4: Wordle information propagation dissemination process

The parameter relationship of each variable is as follows:

$$\begin{aligned} N &= S + E + I + R \\ \frac{dS(t)}{dt} &= -\frac{\beta S(t) I(t)}{N} \\ \frac{dE(t)}{dt} &= \frac{\beta S(t) I(t)}{N} - \sigma E(t) \\ \frac{dI(t)}{dt} &= \sigma E(t) - \gamma I(t) \\ \frac{dR(t)}{dt} &= \gamma I(t) \end{aligned} \tag{3}$$

The number of daily reports can be obtained through calculation:

$$Rnum(t) = zI(t) \tag{4}$$

By fitting the original data, the original model is obtained through multiple fitting.See appendix for its initial parameters:

After visualizing the number of reports, we can intuitively find that there are significant fluctuations in the number of daily reports. It fluctuates for reasons that cannot be explained by time t. We further analyzed the data and found that the larger the percentage of the first three attempts on the day, the more significant the fluctuation in the number of reports on the day. So we assume that the number of daily reports fluctuates in relation to word difficulty.

To further verify the conjecture, $Dif(t)$ is calculated, and correlation analysis is made between it and $Rnum(t-1) * Rnum(t)$. They were found to be significantly correlated, and $Try_2(t)$ was the most correlated.

From this analysis, we can conclude that the fewer attempts a player takes to complete a game, the more likely they are to send a tweet.

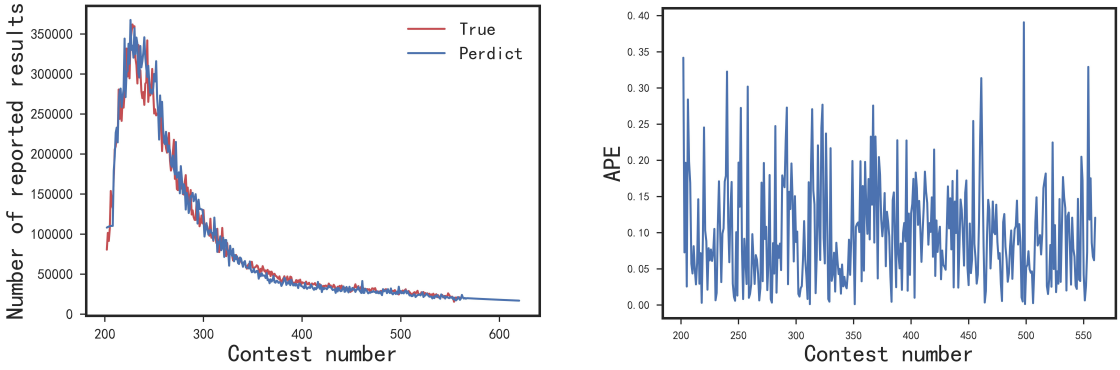
In the Wordle information propagation model, the number of attempts and the probability z of sending tweets have an effect. Therefore, we establish the probability evaluation equation with $Dif(t)$ as the basis and $Try_2(t)$ as the disturbance term:

$$z = \frac{\text{Dif}(t) + \text{sigmoid}\left(Try_2(t) - 0.5\right)}{10} \tag{5}$$

Combined formula4:

$$Rnum(t) = \frac{\text{Dif}(t) + \text{sigmoid}\left(Try_2(t) - 0.5\right)}{10} I(t) \tag{6}$$

According to the Wordle information propagation model, the Fitting Graph and the APE Graph are obtained. It can be seen that the fitting degree is excellent, and MAPE is 9.892.



3.2.3 Model summary and interval prediction

After the prediction of the above model, we used ARIMA model, LSTM algorithm and XGB algorithm to predict the number of reported results in order to make the analysis more perfect. By comparing the confidence intervals of the various models, it is obtained that the number interval of reported results on March 1, 2023 is [17330,20900] at the 95 percent confidence level. Below is the forecast result graph of the five different models.

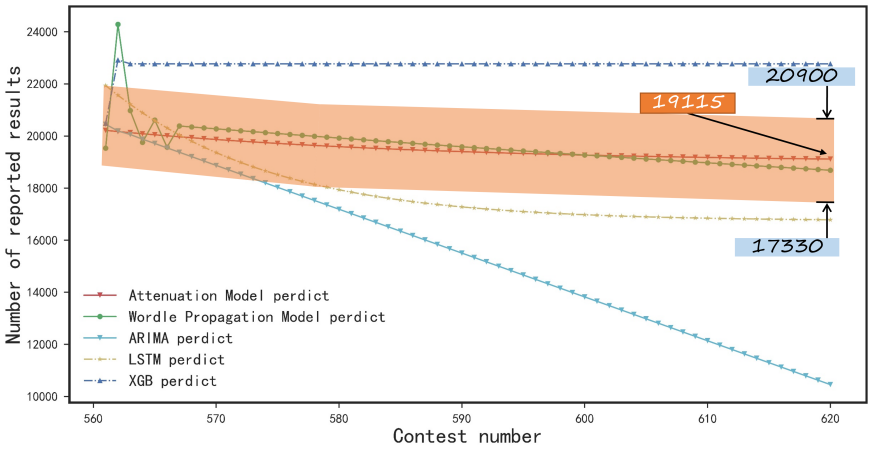


Figure 5: The models' predictions for the future

By observing the figure, it is found that the original data does not conform to a stable time series, its mean does not have a stable amplitude on the time axis, and its variance does not tend to a stable value. Therefore, there is a big difference between the prediction results of ARIMA and other prediction models of deep learning. However, the Wordle information propagation model based on the complex network propagation dynamics tends to be stable and accords with the expected results.

3.3 Word feature

In Wordle games, the factors that affect the guessing times of players include the complexity of word structure, the frequency of word use, the part of speech of the word and other related factors. In order to distinguish the effect of different words on the difficulty of guessing games, we quantified words by analyzing the following attributes.

3.3.1 Intuitive features

- **Letter combinations** According to word formation, there are some common roots and affixes in English.We split the word and count the frequency of each letter combination, considering that the player may substitute his own vocabulary in the process of guessing the word,It is found that there will be some degree of influence on the percentage of scores reported that were played in Hard Mode.
- **Repeat letters** According to the rules of the game, the program doesn't tell the player how many letters they guessed during each guess. So when there are duplicate letters, the difficulty increases. At the same time, this paper defines Classification variable, which can be divided into 1, 2 and 3, among which:

  1: There are no repeated letters in the word.

  2: There is only one recurring letter in the word.

  3: The word has two or more repeated letters.

  Below is a violin diagram of the difficulty for different Classification. You can see that the difficulty increases as the degree of letter repetition increases.

3.3.2 Hidden Features

- **Word frequency** The frequency with which a letter appears in text. Wordle's mechanic is to guess and eliminate letters step by step to determine the final answer, so how often letters are used affects the game's problem-solving process. We collected all 5-word words and calculated the frequency with which each letter was used, and found that the higher the frequency of the word, the greater the percentage of scores reported that were played in Hard Mode.
- **Word popularity** The total number of 5-word words is about 13,000, compared to just over 2,000 common 5-word words. When playing games, the average player is more likely to guess common words. We obtained the popularity of each word through the Mathmetica program interface and normalized the data using the sigmoid function. Found that low popularity of the answer to the riddle increased the difficulty of the game.

3.3.3 Information entropy

Charades are about eliminating letters to get to the final answer. And because the letters don't have the same frequency, they don't carry the same amount of information. So we introduce the concept of information entropy to quantize the information of each word.

Take the letter 'e' for example, its word frequency is the highest, so the number of words formed by 'e' is the largest, the uncertainty is the highest, and the information content of 'e' is the largest.

The following is the flow chart of information entropy evaluation:

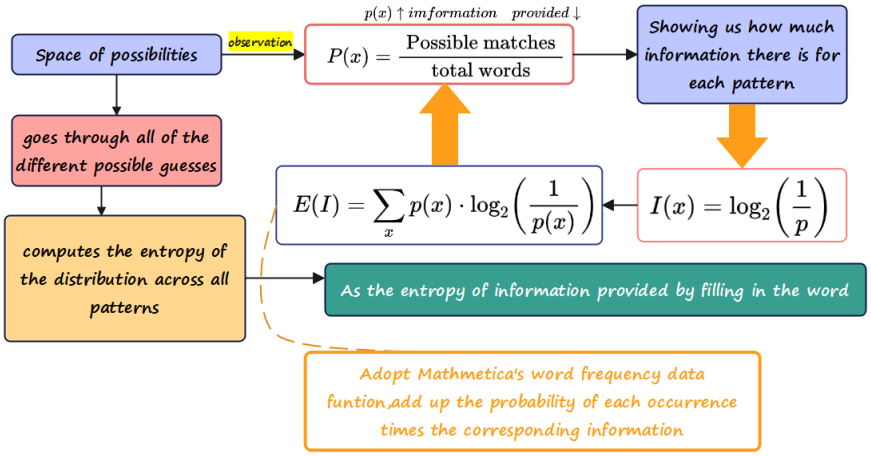


Figure 6: Word information entropy calculation process

Because of a meaningful word, there are combinations of letters in a particular order. For example, combinations of 'es' and 'er' are more common than combinations of 'rc' and 'fy'. Therefore in combination with context, the appearance of a letter is not independent, the letter S in the possibility of $P(S) = P(x_n|x_1, x_2, ...x_n - 1)$. Since there are too many possibilities, we follow the Markov hypothesis - the probability of any word w_i is only related to the word x_{i-1} before it, and the probability of S is $\mathrm{P(S)} = \mathrm{P(x1)P(x2 \mid x1)P(x3 \mid x2)} \ldots \mathrm{P(xi \mid xi - 1)}$

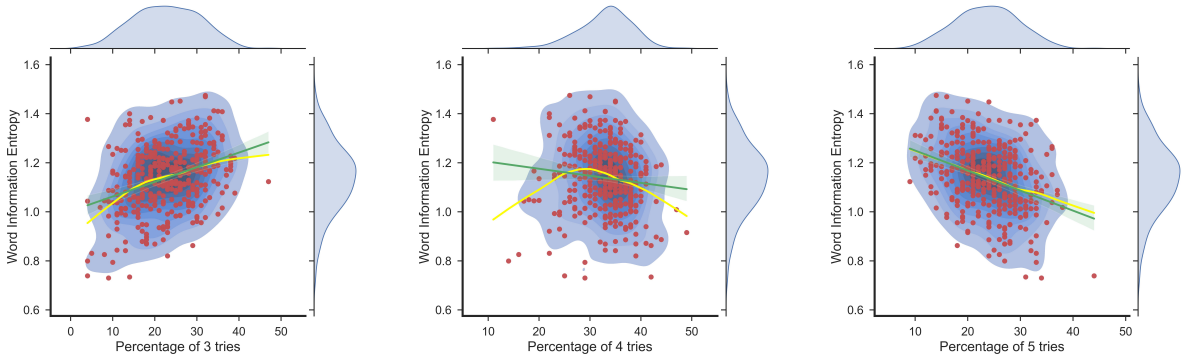
According to Markov hypothesis, we build a dual language model:

$$E(I) = -\sum_x p\left(x_n \mid x_{n-1}\right) \cdot \log_2 p\left(x_n \mid x_{n-1}\right) \tag{7}$$

The binary model takes into account the context and word formation, which is more in line with human spelling habits.

Linear regression was performed using information entropy and percentage of attempts. It is obvious that the fourth time is the watershed between information entropy and the percentage of attempts. When the number of attempts is less than 4, the information entropy is proportional to the percentage. This means that the greater the entropy

of information, the greater the number of words contained, the greater the proportion of correct attempts in the first three attempts. When the number of attempts is greater than 4, the information entropy is inversely proportional to the percentage. This indicates that the scope of determined words is narrowed, and the smaller the information entropy is, the more correct the word can be determined.



3.4 Correlation analysis between word attribute and number of attempts

For the above word attributes, Spearman correlation analysis was conducted between them and the percentage of attempts. The heat map is as follows:

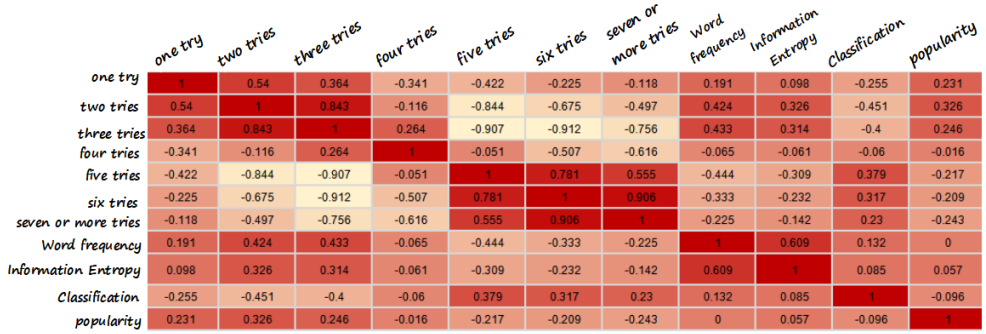


Figure 7: Heat map

Word frequency is the harmonic average of the frequency of each letter in the word. It can be found that Word frequency, Information Entropy and Popularity are significantly positively correlated with the first three attempts, and negatively correlated with the latter several attempts. In conclusion, when the Word frequency, Information Entropy and Popularity of a word increase, players are more likely to guess the word correctly in the first three guesses, and the word is simpler. Classification is negatively correlated with the first three attempts and positively correlated with the next few attempts. This suggests that the more letters in a word are repeated, the more guesses users need to make, and the more difficult the word is.

4 Regression prediction on the number of guesses

4.1 Stepwise regression model

For problem two, we classified the percentage of different attempts according to Classification, and carried out visual analysis on the feature data of words.It can be found that the percentage of attempts is related to word frequency, word information entropy, classification of repeated letter number, word popularity, letter number and other factors.

In order to avoid multicollinearity problems, multiple stepwise regression method was used for prediction, and mathematical models were established for different percentages of attempt times, such as word frequency, word information entropy, classification of repeated letter number, word popularity, letter number and other factors. After introducing all the independent variables into the model, the independent variables that contributed little to the sum of the squares of the residual parameters and the squares of other parameters were eliminated one by one. After several iterations, no independent variables met the exclusion conditions. In this way, a more accurate multiple linear regression equation is established.

Dependent variable

- Percentage of each guess (1 try, 2 tries,.... ,7 or more tries (X))

Argument

- Word frequency (F) : The word frequency of the word
- Word information entropy (E) : The information entropy of the word
- Repeated letter classification (C) : The word is categorized by the number of repeated letters
- Word popularity (P) : The popularity of the word
- Letter count (L_i) : The number of letters appearing in the word, where $i = a, b, ..., y, z$

$$\text{Percentage }_i = \alpha_i + \beta_i F_i + \gamma_i E_i + \delta_i C_i + \varepsilon_i p_i + \sum \lambda_i L_n + \mu_i \tag{8}$$

According to the formula, the percentage of each attempt times was substituted into the data, and finally 7 corresponding multiple regression equations were obtained.All regression equations passed the F-test (joint hypothesis test) and were valid equations.

By constructing stepwise regression model, the prediction equation corresponding to each number of guesses is obtained.See the table below:

Table 2: Stepwise regression equation result table

No. of attempts	Regression equation	F-test
Try_1	$y = 0.031 + 15.436F - 0.358C + 0.384P$	F=20.527 P=0.000***
Try_2	$y = 2.444 + 123.637F - 2.942C + 2.258P + 1.272L_t - 1.245L_y + 0.858L_a - 0.665L_l$	F=43.161 P=0.000***
Try_3	$y = 17.592 - 6.586C + 183.392F + 3.32P + 2.051L_t + 2.357L_p + 2.22L_h - 2.913L_y$ $-1.756L_e - 2.998L_w + 10.036E + 1.666L_d + 1.36L_i - 2.482L_v - 1.548L_g$	F=32.081 P=0.000***
Try_4	$y = 34.213 + 2.537L_p - 7.611L_z - 1.297L_r - 1.236L_e - 1.151L_a$	F=8.485 P=0.000***
Try_5	$y = 31.75 - 198.567F + 5.047C - 2.58L_h - 2.475P - 1.982L_t - 2.08L_p$ $-1.741L_c - 1.142L_i + 1.077L_e - 4.448L_q - 1.493L_d - 7.01E$	F=38.512 P=0.000***
Try_6	$y = 12.054 - 115.629F + 4.28C - 2.471P - 2.081L_p + 2.698L_y$ $+1.008L_e - 1.573L_t + 7.533L_z + 2.687L_v + 2.141L_w$	F=22.326 P=0.000***
Try_7	$y = 1.724 + 1.477C - 1.509P$	F=10.566 P=0.000***

According to the obtained regression analysis table, the regression model was established based on the retained independent variables, and the significance p-values of F-test were all less than 0.01, showing the level of significance. The original hypothesis that all regression coefficients were 0 was rejected, indicating that the model was well constructed.

4.2 Uncertain factors

In the prediction process of the model, we take into account the following possible uncertainties:

1.Since the sample data is collected from players voluntarily Posting on Twitter, considering the differences in individual personalities of each player, some players may be inclined to post better results on the social platform, and it is also possible that the word as the answer to the riddle on the same day is more difficult, resulting in more players willing to show the results they are proud of. As a result, the collected data may have some subjective bias, which cannot guarantee the absolute randomness of the sample data.

2.During the prediction process, no relevant metrics were found to distinguish the categories of players. As the game progresses over a longer period of time and the number of older players increases, it is possible that this increase in proficiency will lead to an increase in the average player's ability to guess words, taking into account the growth of the players themselves.

3.At the same time, due to the unpredictability of the word as the answer to the riddle, the word itself may be fermented by some events and appear as a hot word on the Internet, which cannot guarantee the absolute perfection of the model's judgment on the characteristics of the word, which may lead to certain deviation between the predicted results and the actual situation.

4.3 Prediction of model

Here are EERIE's indicators and predictions:

The Indicators of EERIE						
E	P	F	C	L_e	L_i	L_r
0.558886	0.463777	0.080604	3	3	1	1

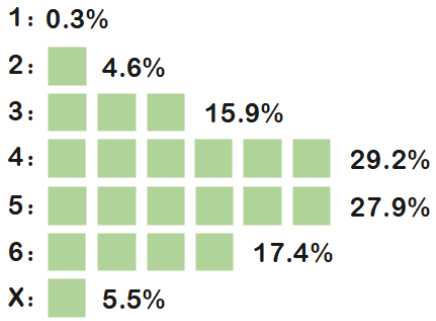


Figure 8: The predicted results for EERIE

4.4 Test of the model

For the test of the model, we calculated the average error of the actual value and the predicted value of the percentage of each attempt (see the figure below). It can be seen that the error is between $\pm 2\%$, indicating that the model has high accuracy in the percentage of each attempt in the prediction of difficult mode.

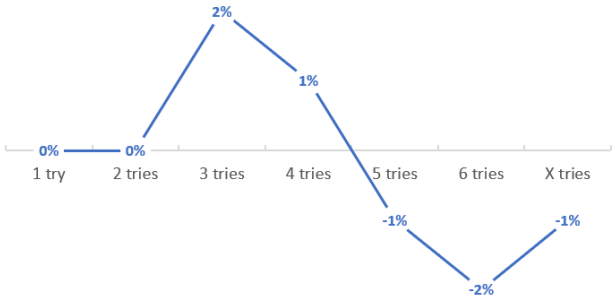


Figure 9: The average error

5 Classification and evaluation of word difficulty

5.1 K-means construction of classification model

To solve this problem, we build a classification model by K-means algorithm. Words were classified by difficulty by analyzing the number of guesses in the data. Then the

classification results are used to define "difficulty" as the attribute of the word, and the difficulty of guessing is evaluated quantitatively. Finally, the difficulty level of the word EERIE is calculated based on the prediction results of the problem two model, so as to make a specific division of its guess difficulty.

K-means algorithm first needs to randomly specify the initial cluster number k and the corresponding initial cluster center C from the given data object. The distance from the initial clustering center to the remaining data objects is then calculated. Euclidean distance is selected in this paper. The Euclidean distance formula from the clustering center to other data objects in the space is:

$$d\left(x, C_i\right) = \sqrt{\sum_{j=1}^{m}\left(x_j - C_{ij}\right)^2} \tag{9}$$

x is the data object, C_i is the ith clustering center, m is the dimension of the data object, x_j, C_{ij} is the attribute value of the JTH dimension of the data object x and clustering center C_i.

According to Euclidean distance, the similarity is measured, and the target data with the highest similarity to the clustering center is allocated to the C_i cluster. After the allocation, the data objects in the k clusters are averaged to form a new round of clustering center, so as to reduce SSE of the data set. The calculation formula is as follows:

$$SSE = \sum_{i=1}^{k}\sum_{x \in C_i}\left|d\left(x, C_i\right)\right|^2 \tag{10}$$

SSE value is used to measure the quality of clustering results. When it no longer changes or converges, the iteration is stopped and the final result is obtained. The flow chart is as follows:

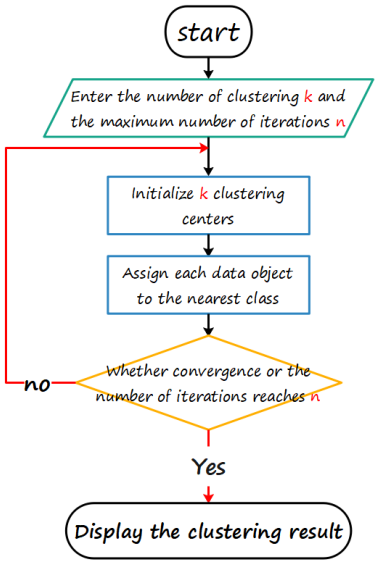


Figure 10: K-means flow chart

According to the number of player attempts as the clustering basis, the bungalow distance error and the degree of distortion between the particle of each cluster and the sample points within the cluster are called. Therefore, for a cluster, the lower the degree of distortion, the closer the members in the cluster, the higher the degree of distortion, the looser the structure in the cluster, and the degree of distortion will decrease with the increase of the category. However, for data with a certain degree of differentiation, the degree of distortion will be greatly improved when it reaches a certain near point, and then slowly decreases. This critical point can be considered as a point with better clustering performance.

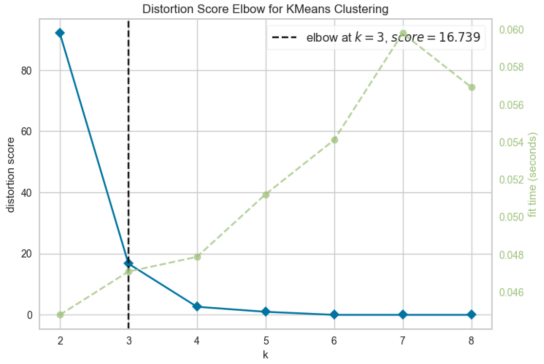


Figure 11: K-means elbow chart

	Easy	Middle	Diffcult
1 try	1	0	0
2 tries	10	2	0
3 tries	47	10	4
4 tries	32	25	11
5 tries	9	36	15
6 tries	2	23	22
7 tries	0	4	48

Table 3: Initial cluster center

From the figure above, it can be seen from the elbow rule that the clustering effect is optimal when there are three clustering centers. Therefore, we divide the words in the attachment into three categories: easy, medium and difficult. By K-means algorithm, we selected third, woken and parer as the initial clustering center of the three difficulties respectively. The number of attempts is shown as follows:

5.2 Construction of difficulty evaluation model

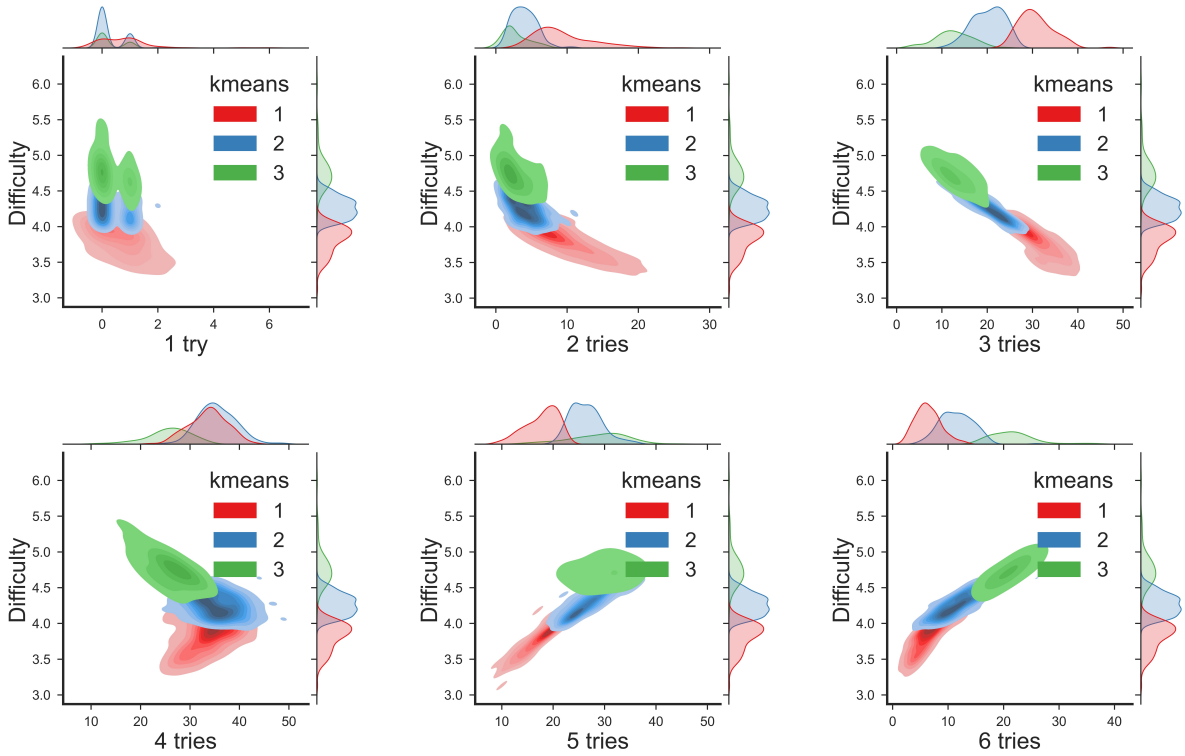
In order to provide a quantitative assessment standard for the difficulty of riddle words, we defined the word "difficulty" F as the measurement standard.

$$F = \frac{\sum_{x=1}^{6} x \cdot N(x \text{ tries })}{\sum_{x=1}^{6} N(x \text{ tries })} \tag{11}$$

Where x is the number of times it took the player to guess the answer to the riddle.

By quantifying the difficulty, we calculated the difficulty of all the words in the sample data. The difficulty level of the three clustering centers selected during classification is used as the standard for quantitative evaluation of word difficulty classification. See appendix for the difficulty table:

The following figure is the kernel density estimation chart of the percentage of different attempts divided according to clustering. It can be intuitively seen that the data is obviously divided into three categories, indicating that the clustering summary model of K-means is excellent.



Then we took the percentage of each attempt as the independent variable and the clustering result as the dependent variable. The first 90% of the sample data in the attachment is taken as the training set, and the last 10% is taken as the test set. BP neural network classification is carried out on the data.

5.3 Verification of classification evaluation model

In order to verify the accuracy of classification results, we cross-validate the classification results of bp neural network. When the accuracy rate is 99%, the recall rate is also 99%, indicating that the model is very excellent. Even in the case of incomplete sample size, it still has a high classification effect, and we believe that the difficulty evaluation classification model built by us has a good effect. The specific results are as follows:

Table 4: BP model evaluation

BP	Accuracy rate	Recall rate	F1
Training set	1	1	1
Test set	0.992	0.989	0.991

5.4 EERIE difficulty prediction

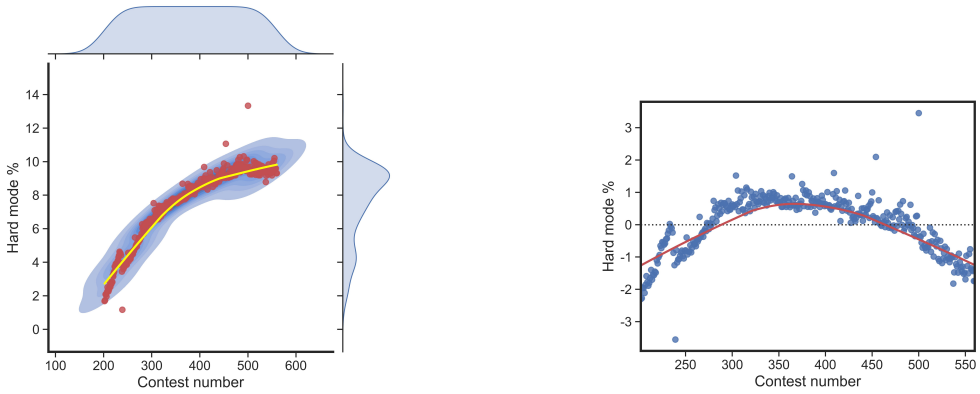
Through the model of problem two, we have completed the guessing situation of the players in Wordle when the word EERIE is used as the answer to the riddle. According

to the difficulty evaluation system we constructed, the difficulty level of this word is calculated as $F(EERIE) = 5.03$, thus indicating that the difficulty level of this word is difficult.

6 Some interesting discovery

6.1 Percentage increase in hard mode

Interestingly, although the Number of reported result decreased over time, the number of players participating in difficult mode increased steadily. The following is the trend plot and linear regression residual plot of the number of players participating in difficult mode, the percentage of total players and the Contest number:



6.2 Part of speech has no effect on difficulty

We analyze the parts of speech of the given words. Part of speech symbols and frequency of occurrence are listed in appendix.

By counting the occurrence times of these parts of speech, the parts of speech and difficulty are analyzed, and the boxplot is drawn. It can be found that the parts of speech and difficulty are not necessarily related.

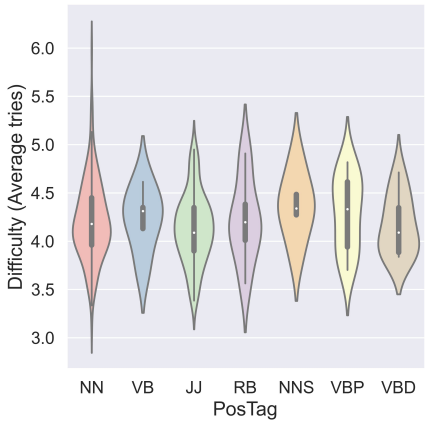
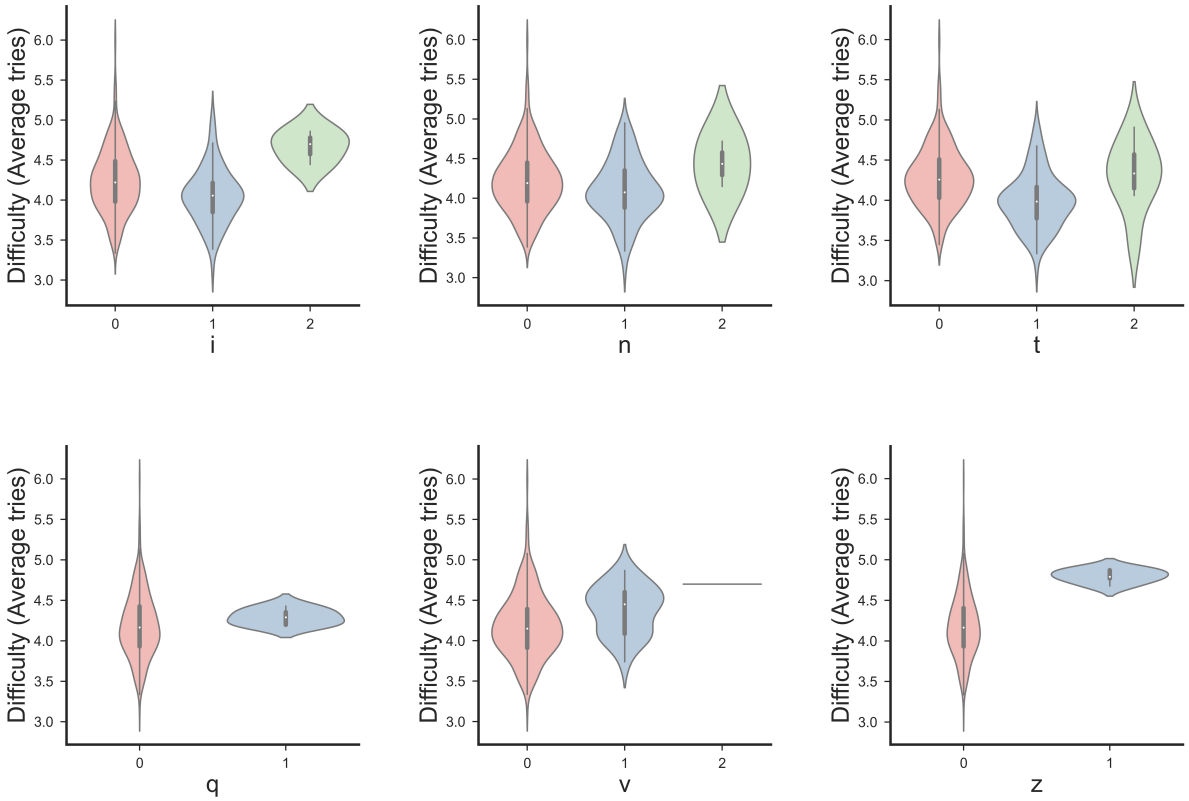


Figure 12: Comparison of difficulty of PosTag

6.3 The number of letters in the puzzle will affect the whole difficulty

By counting the number of letters, when the letters 'a,h,i,n,p,r,s,t' appear in the riddle word, the overall difficulty will become simpler.The difficulty becomes harder when 'b,g,j,k,q,v,w,x,y,z' occur in words.



7 Conclusion

7.1 Strengths

- Because of the volatility of the data, the traditional time series model cannot be used for analysis. Therefore, based on the complex network communication dynamics model, this paper explains the data volatility by using Twitter information transmission rate, user publication rate and other factors. Accurately predict the range of the number of reported results in the future.

- With various word features as independent variables, multiple stepwise regression was used to model the percentage of attempts. Successfully predicted the release of EERIE scores.

- Combining with information theory, word features are divided into objective conditions such as information entropy, popularity and letter combination to make the model more specific and comprehensive.

- Using K-means clustering algorithm, cluster analysis was conducted on the percentage of attempts. Define the difficulty type according to the classification result. BP neural network was used to construct a difficulty evaluation model with the number of attempts as independent variable and difficulty as dependent variable.

7.2 Possible Improvements

Our model has the following limitations and related improvements

- When calculating the attributes of word information entropy, word usage frequency can be added to optimize the calculation mode of information entropy, so that the result is more matching with the reality.
- When establishing the information propagation model, some optimization algorithms such as neural network can be combined to further optimize the parameters.
- If we had more complete data, we could add variables of player characteristics to the prediction of percentage of attempts, which would be more accurate.

Undeniably, in order to ensure the accuracy of prediction results as much as possible, this paper uses a variety of data about the attributes of words that can be collected, among which there must be some influencing factors that we have not taken into account that may cause deviations between model results and future actual values.

References

[1] Huan Liu; Qing Bao; Hongjun Qiu; Ming Xu; Benyun Shi. 2021. Source Identification of Asymptomatic Spread on Networks.

[2] Zhongdong Yin; Junye Cheng. 2023. Distribution transformer winding material identification algorithm based on Kmeans clustering.

[3] Imran, H.; Al-Abdaly, N.M.; Shamsa, M.H.; Shatnawi, A.; Ibrahim, M.; Ostrowski, K.A. 2022. Development of Prediction Model to Predict the Compressive Strength of Eco-Friendly Concrete Using Multivariate Polynomial Regression Combined with Stepwise Method.

[4] Koponen I; Södervik I. 2022. Lexicons of Key Terms in Scholarly Texts and Their Disciplinary Differences: From Quantum Semantics Construction to Relative-Entropy-Based Comparisons.

MEMORANDUM

To: Puzzle Editor of the New York Times

From: Team #2300466

Subject: Research suggestions for Wordle answer words

Date: February 20, 2023

Dear Puzzle Editor of The New York Times,

Wordle, a free word-guessing game, has taken the world by storm, with more than 2 million players worldwide. In our opinion, as a small game with simple style, the five-character words preset by the system every day are not only the only puzzle of the day, but also the breakthrough of Wordle's viral spread. Players only need to spend a few minutes to complete the game every day, which effectively strengthens the "sense of scarcity" of the game experience.

After our research, we are not worried about the total number of words used as puzzles at first. The total number of five-character words is more than 12,000, excluding rare words, the range of game answers can still reach about 2,500, and such a vocabulary bank can last for nearly seven years. So our research is more about distinguishing the specific properties of words for games. Combined with word frequency, information entropy and other data, we established a unique evaluation model, which can effectively evaluate the difficulty of guessing words as the answer to riddles quantitatively. It can also predict roughly how well the player will answer the riddle.

We analyzed the mystery of 2022 for an entire year. Based on feedback from players, the most difficult word is parer on September 16, while the least difficult word is train on May 4. We also combined the complex network communication theory to construct a prediction model with a high degree of fit for the number of reports, trying to explain the small fluctuations in the number of reports over the past year. To some extent, the number of reports also reflects the number of game players, which can provide some reference value for the development of the game.

Based on the above research, we have the following suggestions for the running of games and the selection of riddle words:

1. Select words with large differences in difficulty coefficients on adjacent days

Developers have a certain amount of information guidance, and can use the data of player feedback on various social platforms to describe a certain amount of player experience. In turn, changes can be made to the player experience. When the majority of players said the answer to the riddle was easy, the more difficult word was chosen as the answer, and vice versa. Maximize the player experience and thus increase the engagement of game players.

2. Develop multiple game modes to give players more options based on word attributes

As the current game tips are more in the process of player speculation, for the player's different choices to give hints. Before the game started, there were only easy and hard modes, and the difference between the existing modes was that the difficulty was set more in the guessing process, rather than giving different players more freedom to accumulate words in different directions. Therefore, we want to create more game modes to provide some bias to the puzzle according to the player's preference. For example, some hints are given about the properties of the riddle word, the noun pattern, the verb pattern, or the fact that the riddle word is an animal, etc., so that players with different subjective preferences can have a better game experience.

3. Give your game a longer life cycle with a little more hype

Combined with the data, we found that the number of players for wordle has skyrocketed since its launch. Compared to other games of the same type, the achievement is very proud. However, by analyzing player data over the past year, it is clear that the total number of players has declined in the second half of the year. We think it probably has something to do with the fact that most people are starting to go back to work. Therefore, we suggest that while optimizing the word of the riddle, we should also increase the publicity of the game to ensure that more new players continue to flow in.

In the end, we hope that our research and related suggestions can help you in solving puzzles, and we sincerely hope that wordle will get better and better.

Your Sincerely,
Team #2300466

Appendices

Table 5: Model initial parameter

Parameter	Initial Value
N	4500000
β	0.2
σ	0.1
γ	0.1
k	3.3
z	0.25

Table 6: Classification of difficulty

Degree of difficulty		
Easy	Middle	Diffcult
$0 \sim 3.43$	$3.43 \sim 4.8$	$4.8 \sim 5.99$

Table 7: Part-of-speech statistics

PosTag	Description	Number
NN	noun singular	205
JJ	adjective	87
RB	adverb	13
VBP	The present tense of the non-third person singular	12
VBD	past tense	10
VB	verb prototype	5
NNS	noun plural	5
VBN	past participle of verb	4
IN	a preposition or subordinate conjunction	3
VBZ	third person singular present tense	3
VBG	gerund or present participle	3
MD	modal verb	2
CC	joint conjunction	2
PRP$	possessive pronoun	1
RBR	comparative adverb	1
JJS	superlative adjective	1
DT	qualifier	1
JJR	comparative adjective	1