Problem Chosen	2023	Team Control Number
C	MCM/ICM Summary Sheet	2308018

Wordle Together :
An Analysis About Player's User Profile

Summary

After all is said and done, 2022 may go down as the year of Wordle. It has become a daily staple for millions around the world. In this paper, we model, analyze, and describe people's enthusiasm, preference, and strategies towards Wordle.

Before modeling, we preprocessed the dataset by using the **Lagrange interpolation** method to replace outliers, **normalized** the percentages that didn't sum to 100. We then conducted **preliminary exploratory** data analysis and discovered some interesting phenomena, such as people's declining interest in Wordle during **holidays**.

In Model I: We utilized **wavelet analysis** to identify the short-term periodic features of the **nonlinear and non-stationary** time series. Based on this, we chose the **SARIMA-ANN** model which has better adaptability. The predicted results almost coincide with the actual results, and the prediction interval for the number of reported results on **March 1, 2023** is $[14554, 15301]$. Our model demonstrated a better fitting performance than others

In Model II: We extracted word features from three aspects: **the number of vowels in a word, letter frequency, and word structure**. We conducted **Pearson correlation** test on these attributes and found that the presence of repeated letters in a word, especially in the first and fourth position, was **positively correlated** with the percentage of right answers. The correlation coefficients were relatively high, and all passed significance tests.

In Model III : We leverage the excellent **Markov property** of the Wordle game rules and use **Markov Decision Process (MDP)** to describe the game process. We design a reward function based on word frequency and information entropy. Subsequently, we improve the **Monte Carlo Tree Search (MCTS)** algorithm based on the idea of random simulation, and its output is the frequency of finding the correct answer at each level after a certain number of games. We convert it into probability and use **BP Neural Network** to train parameters to fit the closest distribution to the reported distribution. The model fits well, and we can predict the distribution of the word "EERIE" as $x = [0.507, 2.535, 18.46433.842, 29.252, 14.497, 0.901]$. Finally, we analyze **the uncertainty of our algorithm**, which may come from the fixed initial word we use because we cannot afford the huge data volume brought by choosing the root node.

In Model IV: As the distribution of reports each day is a **probability vector**, and we used **Dirichlet distribution** to fit this characteristic. First, we defined the peak of the distribution curve as the **difficulty index**, then used **GMM** to cluster the words in the data into three categories, and verified the rationality. According to the expected value, we ranked the difficulty as easy, medium, and difficult. Finally, we can verify the compatibility between a word and one category by the **KL divergence**. We found that **the word "EERIE" is quite difficult**, which is consistent with our previous validation of the characteristics of multi-letter repetition but low word frequency.

As to the very last, we analyze **the strengths and weaknesses** of our model as well as its **sensitivity**, whose results show that our model has high robustness, precision and accuracy. After that, a **memo** is attached.

Keywords: Time Series Correlation analysis Random Simulation Cluster analysis

Contents

1 Introduction

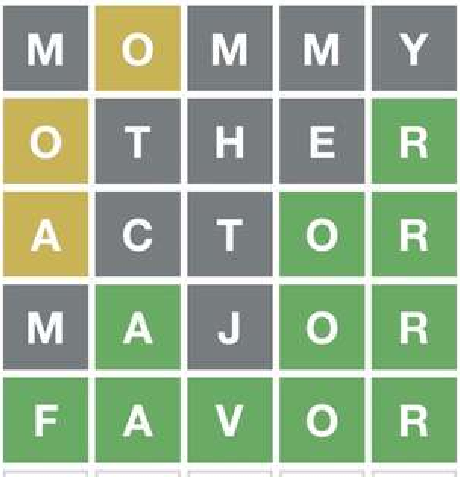
1.1 Problem Background

These days, tweets are popping up in dense clusters on social networks, with only yellow, green and gray squares in the main text, accompanied by a few words of excitement or frustration.[1] Don't worry, this isn't alien writing or a cult code. In fact, this is just a "daily record" of Wordle, a web-based game.

Wordle has become a daily staple for millions around the world. The rules are as follows: Players have six chances to guess a five-letter word, and for each guess, a hint is given based on how well the guessed word matches the correct answer. As is shown in Figure (1), there are three possible hints for each letter: ⬜🟨🟩 . "⬜" indicates that the current letter does not appear in the final answer; "🟨" indicates that the current letter appears in the final answer but is in the wrong place; "🟩" indicates that the current letter appears in the final answer and is in the correct position.

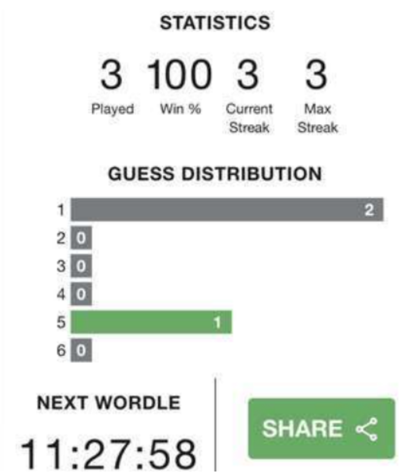
At the end of the game, the player is provided with statistics showing the correct rate of each step, the number of consecutive days played, etc., as shown in Figure (2).

In this paper, we establish a model to deeply analyze and mine data features, simulate players' gaming strategies, and conduct an analysis on the future gaming enthusiasm and the distribution of guessing results.



M O M M Y
O T H E R
A C T O R
M A J O R
F A V O R

Figure 1: Example solution



STATISTICS

3 Played
100 Win %
3 Current Streak
3 Max Streak

GUESS DISTRIBUTION

1 2
2 0
3 0
4 0
5 1
6 0

NEXT WORDLE
11:27:58

Figure 2: Statistical data

1.2 Restatement of The Problem

- Develop a model to account for changes in the number of reported results and predict the number of reported results on March 1, 2023. Determine whether the attributes of the word affect the percentage of scores reported in difficult mode and provide corresponding evidence.

- Develop a model to predict the distribution of the reported results for a given word, analyze the uncertainties in the prediction model, predict the associated percentages of (1, 2, 3, 4, 5, 6, X) for the word EERIE on March 1, 2023, and analyze its prediction accuracy.

- Develop a model for classifying word difficulty, identify the attributes associated with each classification for a given word, then classify the difficulty of EERIE and verify the accuracy of the classification.

- Explore interesting characteristics of the data set and describe them.
- Summarize results in a one to two-page letter to the Puzzle Editor of The New York Times.

1.3 Our Work

The work we have done in this problem is mainly shown in the following Figure(3).

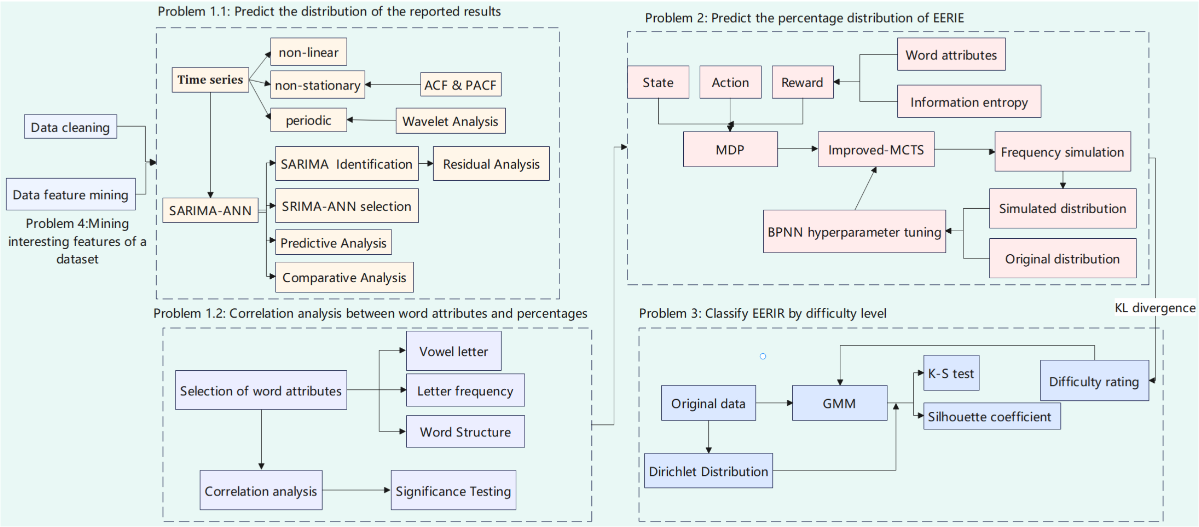


Figure 3: Our Work

2 Assumptions and Justifications

We made the following assumptions to support our model.

Assumption 1: Most of the data comes from trusted sources.

↪ Justification: With the globalization of information, we have reason to believe that the data we have obtained is reliable and meaningful.

Assumption 2: People's guesses were not influenced by others.

↪ Justification: A lot of people shared results when they guessed the right word, and we assume that each player is playing the game on their own, relying only on their own vocabulary and the cues they have been given.

Assumption 3: The correct answer is determined at the beginning of the game and does not change in the middle.

↪ Justification: Markov property requires that states do not change over time, so the external conditions of each game should always be the same.

3 Notations and Definitions

The primary notations used in this paper are listed in Table 1.

Table 1: Parameter Settings

parameter	description
S_t	state at the moment t
P	transition probability
R	reward function
G_t	return function
v_π	value function
$H(x)$	the information entropy of x
$F(x)$	Word Frequency
α	parameter vector
μ	mean of probability distribution
γ	decay factor

4 Data Cleaning

The availability of data must be guaranteed before data analysis, No measures, regardless of its value, can provide accurate assessments if based on unreliable data.

- ***Step 1:* Missing value processing**

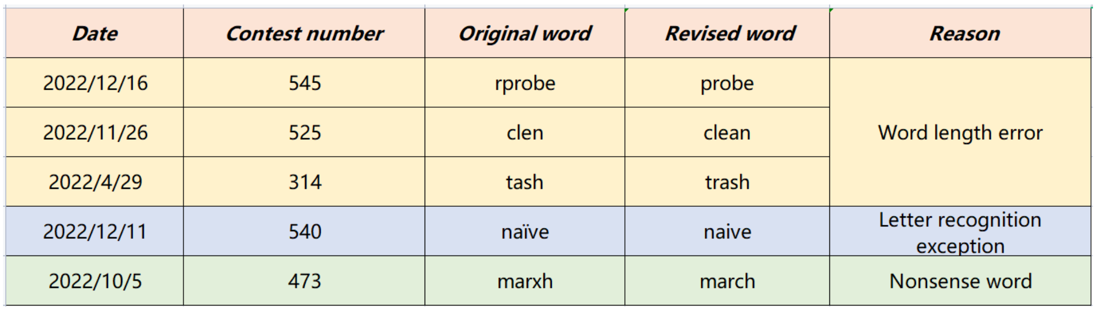
We choose python due to its wonderful ecosystem of data-centric packages. We roughly check *Problem_C_Data_Wordle.xlsx* with the describe() and info() function, which displays the type of data and whether there are missing values in the data set.

- ***Step 2:* Outlier handling**

In the process of data cleaning, we filter out two abnormal cases, and we provide corresponding solutions for each case.

Abnormal word :The solution of Wordle puzzle is a five-letter word and it must be an actual word in English. The following table (4) lists the abnormal words, the reasons and the revised words.

Figure 4: Abnormal word



Date	Contest number	Original word	Revised word	Reason
2022/12/16	545	rprobe	probe	Word length error
2022/11/26	525	clen	clean	
2022/4/29	314	tash	trash	
2022/12/11	540	naïve	naive	Letter recognition exception
2022/10/5	473	marxh	march	Nonsense word

Note : The revised word is made without changing the order of the existing letters, taking into account the frequency of the word and other factors.

Abnormal data : By visualizing the number of reported results and number in hard mode to number of reported results in the raw data set, we find an abnormal data in the number of people reported for Wordle 529 on November 30, 2022, as shown in Figure (5) below. We compared various methods and finally use Lagrange Interpolation Polynomial to replace it with 22462.The effect of eliminating outliers is shown in Figure (6) below.

In addition, we also find that the sum of percentages may not always be 100% due to rounding (In particular, on March 26, 2022, the percentage was 126 %) . In order to

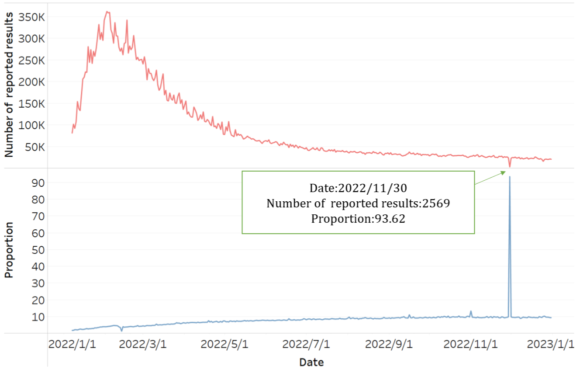


Figure 5: Primary data

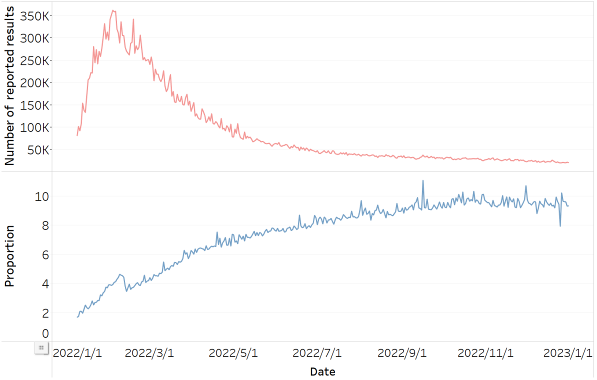


Figure 6: Data after excluding outliers

ensure the accuracy of the data, we normalize the percentages.

Note : Given the continuity of the time series, we choose to replace the outliers rather than delete them outright.

5 Interesting findings in the data set

After data cleaning, We perform data visualization and analysis in the following steps.

- ***Step 1:* Trend analysis**

Firstly, the number of reported results says a lot about how Wordle's popularity has changed. Figure (6) above shows that the number of reported results rose rapidly since January 2022, reaching a peak on February 2, after which the number of reported results gradually declined and leveled off.

To be more precise, we find that the number of reported results experiences sudden changes (either a sharp increase or decrease) at specific time points. Interestingly, we discover that these time points, or the days immediately before or after them, are usually holidays. For example, on December 25, 2022, the number of reported results suddenly decreased. We speculate that on this day, people place more importance on family reunions and the warmth of being with loved ones may alleviate feelings of loneliness and boredom.

In addition, as time flowing, the number in hard mode to the number of reported results has gradually increased and leveled off. We suspect that this is because more and more players are getting the hang of Wordle and are willing to play in Hard Mode.

- ***Step 2:* Distribution analysis**

Then we calculate the average proportion of attempts, as shown in Figure (7) below. We found that about one-third of people were able to figure it out in four tries, about a quarter guessed the word on the third and fifth try, and very few people got it right on the first try or fail after six tries

- ***Step 3:* Periodicity analysis**

Finally, Figure (8) shows that the number of reported results in April 2022, so we guess that the number of reported results may fluctuate periodically in a smaller time scale.

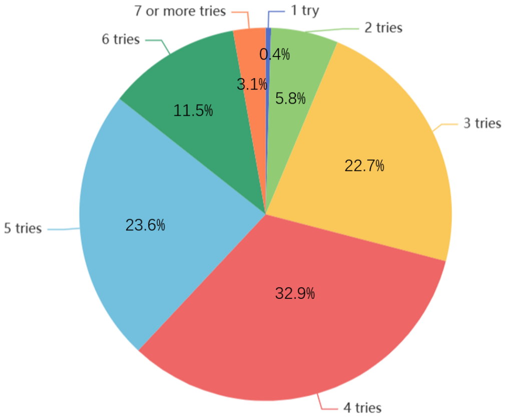


Figure 7: The average proportion of attempts

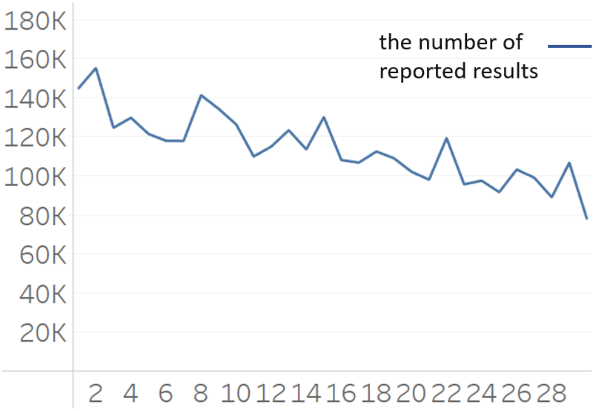


Figure 8: the number of reported results for April 2022

6 Model I: Series Forecasting Model Based on SARIMA-ANN.

The number of reported results vary daily. We use the SARIMA-ANN model based on Wavelet Analysis to explain this variation , predicting the number of reported results on March 1, 2023.

6.1 Feature mining of time series

- ***Non-linearity***

From Figure (6), it can be clearly seen that the number of reported results first increased sharply and then slowed down over time, indicating that the time series is non-linear.

- ***Non-stationarity***

In time series analysis, ensuring the stationarity of data is a necessary prerequisite for subsequent time series analysis steps. For non-stationary time series, statistical analysis will be subject to the limitations of methods and theories. Due to the instability of time series, statistical properties obtained from historical data are not meaningful to the future.

We use SPSS to test the stationality of the data, ACF test and PACF test are conducted, and the results are shown in Figure (9) and (10) below.

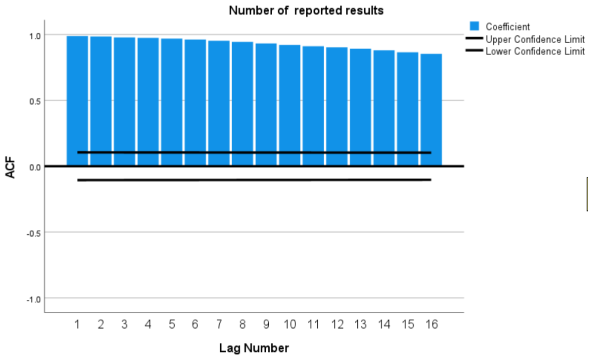


Figure 9: ACF Test

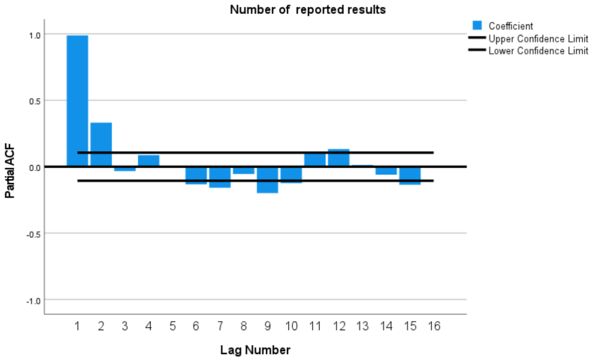


Figure 10: PACF Test

The original ACF is similar to the tail, but the subsequent data is still outside the confidence interval and does not show fluctuations. However, the coefficients in PACF fluctuate up and down around the zero axis, so the original time series is determined to be non-stationary.

- ***Periodicity***

In the previous discussion, we found that the non-linear and non-stationary time series may also have a certain short periodicity, because the time span is large, we use multi-scale wavelet analysis to dig the periodic characteristics of time series data deeply.

Wavelet analysis can be used to test the periodicity in time series. Usually, the time-frequency structure of the signal can be obtained by analyzing the coefficients after wavelet decomposition, and then periodic analysis can be conducted according to the obtained structure.

We choose Symlet-5 wavelet base. Then, we perform J wavelet decomposition on the original time series, where J is the scale of decomposition. So J =4. The next steps are as follows:

1. First, the original time series is assigned to the approximation coefficient a_J.

2. For each level $j = j-1, \ldots, 0$:

a. Perform low-pass filtering and under-sampling on $a_{j+1}$to obtain a_j.

b.Perform high-pass filtering and down-sampling on $a_{j+1}$to obtain d_j.

3. Finally, the wavelet coefficients $c_1, \ldots, c_J$and the approximation coefficients a_J are obtained.

As can be seen from the figure (11), as the number of decomposition layers decreases gradually, as shown when J=1, the periodic component with a period of 5 days appears.

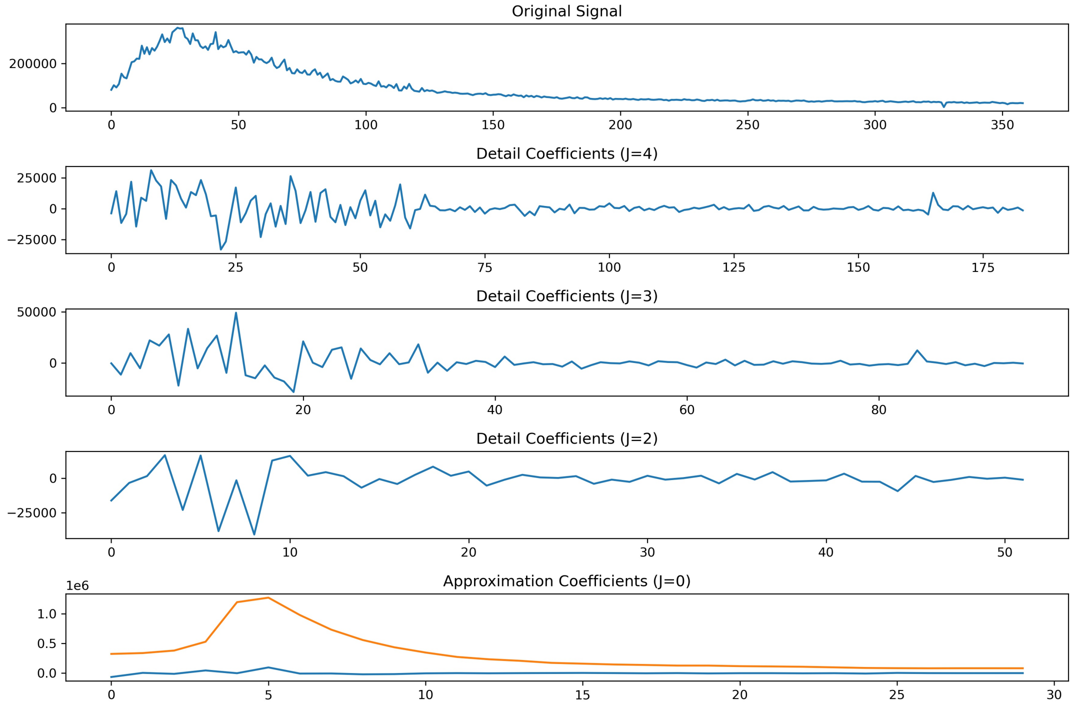


Figure 11: Wavelet Analysis

To sum up, time series has the characteristics of nonlinear, non-stationarity and periodicity, which is an important premise for us to choose the time series prediction model.

6.2 Prediction of time series

The traditional time series prediction model ARIMA is suitable for some stable or weak stable time series, it can capture the long-term trend and periodicity of time series. Due to the nonlinear, non-stationary and five-day period of time series, the prediction based on ARIMA model has some defects.

We use the SARIMA-ANN model to fit and predict the number of reported results, and capture the linear characteristics and seasonal characteristics of time series through the SARIMA model. It also makes full use of the strong curve fitting ability of ANN model to capture the nonlinear features in the time series, so as to give consideration to both the seasonal characteristics and the nonlinear characteristics.

- ***Step 1:*** **Model recognition and test for SARIMA**

Time series data contains trend and periodicity, we adopt SARIMA $(p,d,q)(P,D,Q)_S$, perform first-order difference on the time series x_t and five-step first-order seasonal difference to get the difference sequence:

$$Y_t = \nabla_5 \nabla X_t = (1-\beta)\left(1-\beta^5\right) X_t \tag{1}$$

Then,we use Grid Search to select the appropriate parameters. Grid search is a violent enumeration method, which traverses all possible parameter combinations and selects the optimal parameter combination through cross-validation and other methods. Finally, we find SARIMA $(2,1,2)(1,1,1)_5$ is the best model.

After selecting the final model, white noise test is also needed for residual terms. If the residuals are autocorrelated, consider adding an autoregressive or moving average interpretation, re-modeling and model evaluation, then white noise testing of the new model residuals, and so on until the residuals are determined to be white noise.

White noise can still be tested by autocorrelation graph, and the results are shown in Figure (12).

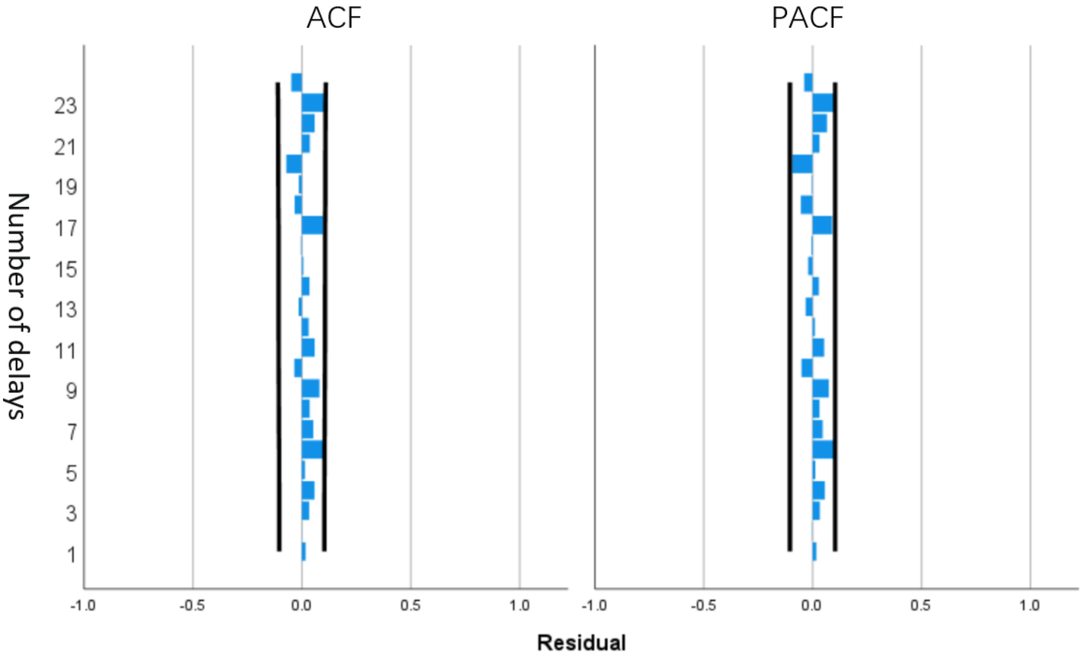


Figure 12: Residual white noise tes

It can be seen that all the autocorrelation coefficients fall within the confidence interval, and the trend gradually approaches 0. Similarly, the residual PACF coefficient is approximately 0. Meanwhile, through Ljung-Box test, it is found that most sig values are greater than 0.05, that is, the lag term is not significant, and the null hypothesis that each residual term is not correlated cannot be rejected. The effective information of timing data has basically been extracted by the model.

- ***Step 2:*** **Construction and selection of the SRIMA-ANN model**

The traditional SARIMA-ANN model[2] distinguishes the linear part of time series from the nonlinear part:

$$Y_t = L_t + N_t, \tag{2}$$

where Y_t is the observed value of time series, L_t represents the linear factor of time series, and N_t is the nonlinear factor of time series.

First of all, the fitting value $\hat{L}_t$ is established through the linear part of the SARIMA model fitting time series. Then, the residual difference e_t of the time series is obtained by comparing the fitting value $\hat{L}_t$ and the actual value Y_t, namely $e_t = Y_t - \hat{L}_t$.

Then using ANN model to identify the nonlinear relationship between residual:

$$\widehat{N_{t+nh}} = \hat{e}_{t+nh} = f\left(e_t, e_{t-nh}, e_{t-n\cdot nh}\right) + \varepsilon_t \tag{3}$$

where ε_t is the random error term, and $\hat{e}_t$ is the deviation between the fitted value and the actual value.The fitted value :

$$Y_{t+nh} = \hat{L}_{t+nh} + \widehat{N}_{t+nh} \tag{4}$$

Thus model 1 is established :

$$\text{SARIMA -ANN 1: } y_{t+nh} = \hat{L}_{t+nh} + f\left(\hat{e}_{t+nh}\right) + \varepsilon_{t+nh} \tag{5}$$

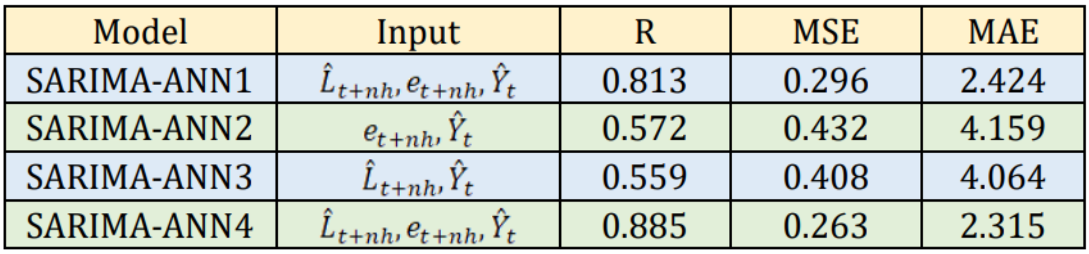
The fitting value L_t and residual e_t obtained through the SARIMA model are applied to the ANN model at the same time, so as to obtain another SARIMA-ANN model :

$$\begin{aligned} &\text{SARIMA -ANN 2: } y_{t+nh} = f\left(e_{t+nh}, \hat{Y}_t\right) + \varepsilon_{t+nh} \\ &\text{SARIMA -ANN 3: } y_{t+nh} = f\left(\hat{L}_{t+nh}, \hat{Y}_t\right) + \varepsilon_{t+nh} \\ &\text{SARIMA -ANN 4: } y_{t+nh} = f\left(e_{t+nh}, \hat{L}_{t+nh}, \hat{Y}_t\right) + \varepsilon_{t+nh} \end{aligned} \tag{6}$$

With the fitting value of SARIMA model as input data and the actual value as learning sample, ANN is trained to realize the establishment of SARIMA-ANN model. Respectively for $(2,1,2)(1,1,1)_5$ model fitting value $hatL_{t+nh}, e_{t+nh}$ and the actual value Y_t for the input data, The network training function is the adjusted conjugate gradient, the initial Lambda is 0.0000005, the initial Sigma is 0.00005, the hyperbolic tangent is the hidden layer activation function, and the identity is the output layer activation function. The time series is fitted, so as to establish the SARIMA-ANN model.

To determine the optimal SARIMA-ANN model, we measure SARIMA-ANN1,SARIMA-ANN2,SARIMA-ANN3 and SARIMA-ANN4 by using R (the standard correlation coefficient),MSE (the Mean Square Error) and MAE (the Mean Absolute Error).

Figure 13: Choice of SARIMA-ANN model



Model	Input	R	MSE	MAE
SARIMA-ANN1	$\hat{L}_{t+nh}, e_{t+nh}, \hat{Y}_t$	0.813	0.296	2.424
SARIMA-ANN2	$e_{t+nh}, \hat{Y}_t$	0.572	0.432	4.159
SARIMA-ANN3	$\hat{L}_{t+nh}, \hat{Y}_t$	0.559	0.408	4.064
SARIMA-ANN4	$\hat{L}_{t+nh}, e_{t+nh}, \hat{Y}_t$	0.885	0.263	2.315

As shown in the following table (13), SARIMA-ANN4 with maximum R and minimum MSE and MAE is finally selected as the optimal model.

6.3 Result analysis and comparison

In the coupling model of SARIMA and ANN, SARIMA-ANN 4 is adopted:

$$y_{t+nh} = f\left(e_{t+nh}, L_{t+nh}, Y_t\right) + \varepsilon_{t+nh} \tag{7}$$

where the linear and nonlinear factors of time series are regarded as a unified whole. The predicted value $\hat{L}_{t+nh}$and the predicted residual $\hat{e}_{t+nh}$are regarded as input variables to establish a unified ANN model.The trend of the number of reported results was fitted and predicted, as shown in Figure (14).

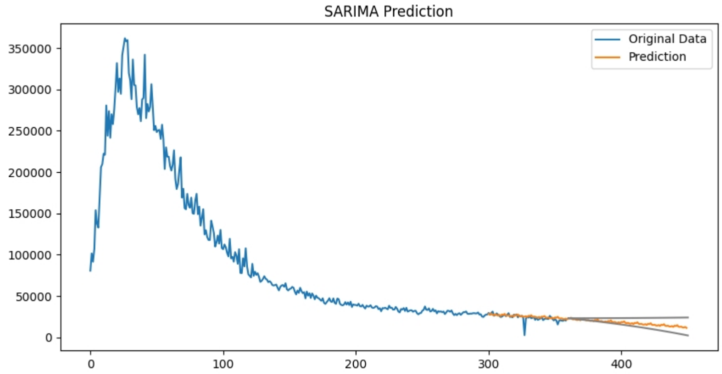
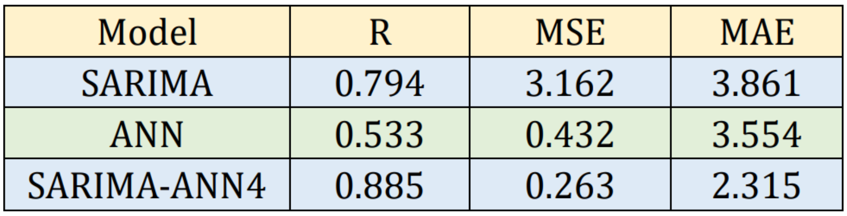


Figure 14: SRIMA-ANN Prediction

Our fitting result is very good, as evidenced by the near-perfect match between predicted and actual values. By setting a 95% confidence interval, we can obtain a prediction interval for the number of reported results on March 1, 2023, which is [14554, 15301].

We compare the fitting results of SARIMA model, ANN model and SARIMA-ANN4 model, as shown in table (15) below.

Figure 15: Comparison of fitting results



Model	R	MSE	MAE
SARIMA	0.794	3.162	3.861
ANN	0.533	0.432	3.554
SARIMA-ANN4	0.885	0.263	2.315

We can find that MSE and MAE of SARIMA-ANN4 are obviously smaller than SARIMA model and ANN model, while R is obviously larger SARIMA model and ANN model, indicating that the correlation between the fitting value and the actual value of the observed data in operation of SARIMA-ANN4 model is better. Therefore, SARIMA-ANN4 model is better than SARIMA model and ANN model in predicting the number of reported results.

7 Model II: Quantification of Word Attributes and Correlation Analysis

7.1 The choice of word attributes

For the attributes of words, we consider as following aspects:

1. ***Vowel letter***. In English, a word usually has a vowel (a,e,i,o,u), but there are some words that don't. We suggest that the number of times each vowel appears in a word n_a, n_e, n_i, n_o, n_u may influence the percentage of scores reported in Hard Mode.

2.***Letter frequency***. We think that the frequency of the letters may affect how people try to navigate difficult patterns. We define the sum of probabilities of words formed by each letter in daily life and data set as f_1 and f_2, and the probability of each letter of words in data set appearing in daily life as $c_i(i = 1, 2, ...5)$.

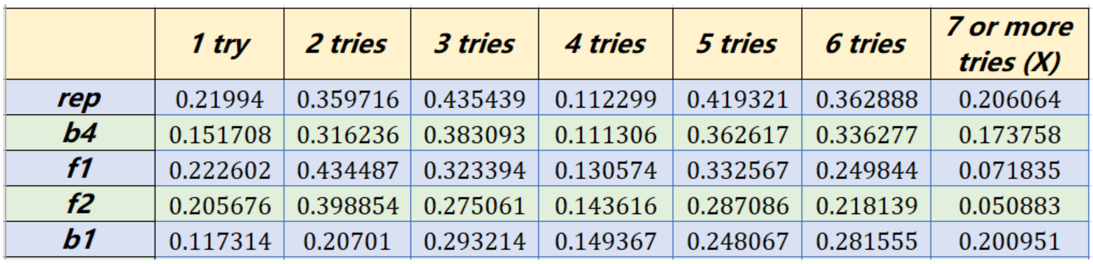
3. ***Word structure***.We think that word structure may also be a factor that people consider when filling in words. We consider from the aspects of whether the word has a root(ro), affix(a), repeated(rep) letter and whether the word has a repeated letter in the i-th bit($b_i(i = 1, 2, ...5)$). These values are 0-1 variables, which are 1 if the conditions are met, and 0 if not.

7.2 Correlation analysis

Above all, the above word attribute variables are classified, which can be divided into continuous variables (such as f_1) and binary variables (such as s), while percentage variables are continuous variables. According to the attributes of variables, we use Pearson correlation coefficient and dot diallel correlation coefficient to discuss the correlation between word attributes and percentages.

The correlation analysis results are shown in the figure (16) below. Because we selected a large number of word attributes, only the word attributes with high correlation with percentage are listed here.

Figure 16: Correlation analysis



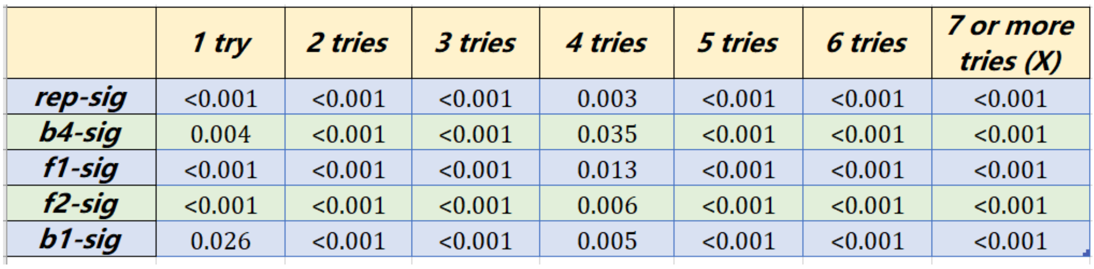
	1 try	*2 tries*	*3 tries*	*4 tries*	*5 tries*	*6 tries*	*7 or more tries (X)*
rep	0.21994	0.359716	0.435439	0.112299	0.419321	0.362888	0.206064
b4	0.151708	0.316236	0.383093	0.111306	0.362617	0.336277	0.173758
f1	0.222602	0.434487	0.323394	0.130574	0.332567	0.249844	0.071835
f2	0.205676	0.398854	0.275061	0.143616	0.287086	0.218139	0.050883
b1	0.117314	0.20701	0.293214	0.149367	0.248067	0.281555	0.200951

From the figure above, we can find that whether there are repeated letters in a word (especially when the repeated letter is in the first and fourth place) is positively correlated with the percentage and the correlation coefficient is high. Secondly, the frequency of using each letter in the word in daily life and in the data set will also affect the distribution of the percentage. Therefore, we believe that the letter with higher frequency is more likely to be remembered by people first. Other word attributes such as the number of vowels in a word and the presence or absence of a root have little effect on the percentage distribution.

7.3 Significance Test

Due to the sample correlation coefficient has a certain randomness, whether it can explain the correlation degree of the population is also related to the capacity. When the sample size is relatively small, the calculated correlation coefficient can not necessarily reflect the real correlation of the population, which will result in false correlation phenomenon. In order to judge the representativeness of the sample correlation coefficient to the overall correlation degree,significance test of the correlation coefficient is needed.

Figure 17: Significance Test



	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
rep-sig	<0.001	<0.001	<0.001	0.003	<0.001	<0.001	<0.001
b4-sig	0.004	<0.001	<0.001	0.035	<0.001	<0.001	<0.001
f1-sig	<0.001	<0.001	<0.001	0.013	<0.001	<0.001	<0.001
f2-sig	<0.001	<0.001	<0.001	0.006	<0.001	<0.001	<0.001
b1-sig	0.026	<0.001	<0.001	0.005	<0.001	<0.001	<0.001

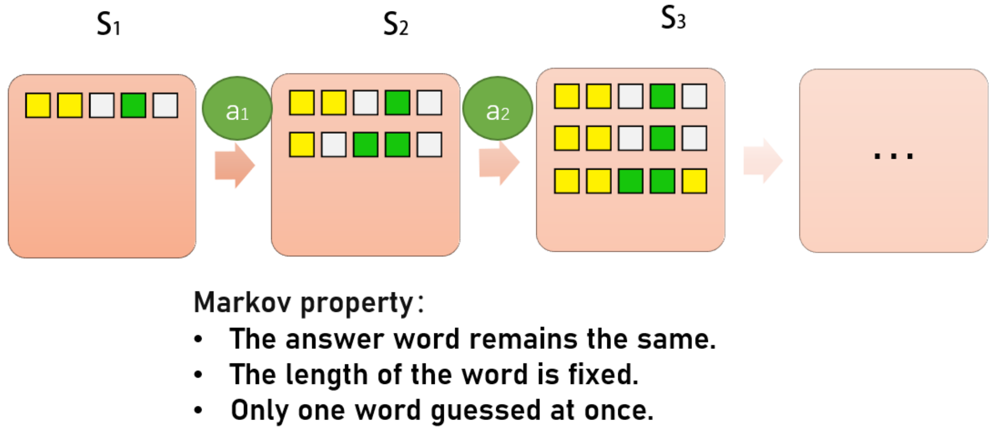
From the figure (17) , we can find that all attribute variables and percentages of words are significantly correlated at the 0.05 level, and most of them are significantly correlated at the 0.01 level. The sample correlation coefficient can represent the overall correlation degree.

8 Model III: Stochastic Simulation Search Model Based on MDP

8.1 Markov Decision Process

In Wordle, after each guess of the word, the result (the number of guessed letters, and the position of the guessed letters) is only related to the current guess word ,and the answer word does not depend on any previous guess process, so we think that Wordle has a typical Markov property.[4]

Markov property is a concept in Probability Theory. When a stochastic process is given the present state and all past states, the conditional probability distribution of its future state only depends on the current state, that is, the future is independent of the past given the present, then the random process has a Markov property. MDP(Markov Decision Processes) is a mathematical model that simulates the stochastic policy and reward of agents in the environment, and the state of the environment has the Markov property.



Markov property:

- The answer word remains the same.
- The length of the word is fixed.
- Only one word guessed at once.

Figure 18: State transition process

In order to represent Wordle as an MDP, we need to define some important components. MDP is often expressed as A quintuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, which includes:

$\mathcal{S}$ is the state space, which represents the set of all possible states. $s_t \in \mathcal{S}$ represents the state at time t, which is the arrangement of the 5×7 grid on the game board. $\mathcal{A}$ is the action space, where $a_t \in \mathcal{A}$ represents the letter to be filled in. $\mathcal{P}$ is the transition probability, and $P(s'|s,a)$ represents the probability of transitioning from state s to state s' under the action a. It can be obtained from the following reward function R and the discount factor γ.

$\mathcal{R}$ is the reward function, where $R(s_t, a)$ represents the immediate reward obtained by taking action a in state s_t. Its value can be fixed or changed according to the requirements.

$$R(s_t) = E\left[R_{t+1} \mid s_t\right] \tag{8}$$

γ indicates the current value of the future reward, which is between 0 and 1. A smaller γ indicates a "shortsighted" focus on the present, while a larger γ indicates a "far-sighted" decision maker is more focused on the future. In predicting the distribution, it represented the group behavior of users.

In addition, there is the termination condition $Done$.[7] In Wordle, we believe that the game is over when the player guesses the correct word, that is, when $S = Done$, the game is over.π represents a set of policies, where $\pi(a|s)$ represents the probability of taking action a in state s, i.e., $\pi(a \mid s) = P(a \mid s)$.According to the above definition, we can obtain the transition matrix:

$$P^{\pi}\left(s' \mid s\right) = \sum_{a \in A} \pi(a \mid s) P\left(s' \mid s, a\right) \tag{9}$$

The objective of MDP is to choose the optimal action a_t at each time step t to maximize the cumulative reward. Here, "optimal" refers to finding a policy $\pi(a_t|s_t)$ that maximizes the expected cumulative reward for selecting action a_t in state s_t.

We define the return of an action as the sum of all possible future rewards, controlled by a discount factor γ to account for diminishing returns over time. The value function $v^{\pi}(s)$ is used to represent the total reward that can be obtained by following a policy π.

$$G_t = \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k}) \tag{10}$$

$$\begin{aligned} v^{\pi}(s) &= \sum_{a \subset \wedge} \pi(a \mid s) \cdot \left(G_t \cdot v^{\pi}(s') + R(s,a)\right) \\ &= \sum_{a \subset \wedge} \pi(a \mid s) \cdot \left(R(s,a) + \gamma \sum_{s' \in S} P_{(s'|s,a)} v^{\pi}(s')\right) \end{aligned} \tag{11}$$

Therefore, the optimal policy for a Markov decision process is to find the policy π^* that maximizes the value function $v^{\pi}(s)$ for each state s.

$$\pi * = \arg\max v^{\pi}(s) \tag{12}$$

8.2 Reward function

Due to the fact that decisions depend on the value return of the policy, and the value function is closely related to the reward function, an appropriate reward function is crucial. Giving the correct answer a higher reward value means that it can be easily guessed. Therefore, we define the reward function based on the characteristics of how people use words:

1.Word frequency: As we have seen in previous discussions, it is a property that affects people's guesses. Specifically, for most people, the higher the frequency of a word, the more familiar they are with it, which means it may be selected as a guess first. This means that we can give it a higher reward value.

2.Information entropy[8]: Information entropy is used to describe the degree of uncertainty in the information and can be expressed in probabilities. The meaning of entropy is the expected value of the information generated by an event. The larger the entropy, the greater the information generated by the event, which means it can provide us with effective reference value for our decision-making.

$$H(I) = \sum_x p(x) \cdot \log_2(\frac{1}{p(x)}) \tag{13}$$

The utilization of information entropy in playing the game Worldle highlights the rational aspect of human behavior. When guessing a word, people tend to consider more appropriate options based on the information provided by previously guessed words and block colors, rather than making random guesses. In the game, the impact of information entropy can be explained in two ways.

First, we analyze the information provided by the block colors, ignoring the intrinsic differences in each word.

• When a letter corresponds to a green block, it provides significant information, including the precise letter type and position. By identifying the letters in the same position in the candidate word, a large number of words can be eliminated.

• When a letter corresponds to a yellow block, it also provides a considerable amount of information, allowing the position to be determined and incorrect positions to be excluded.

• When a letter corresponds to a white block, it only excludes one incorrect letter and provides relatively little information.

The more unreasonable words we exclude, the fewer reasonable words are left, and the provided information entropy is greater.We use n to represent the number of possible numbers left after selecting the word w, and m to represent the excluded word, then its information probability is:

$$p(w) = \frac{n}{n+m} \tag{14}$$

Then we consider the impact of some word attributes. From the discussion above, we can see that the presence of repeated letters in a word (especially when the repeated letter is in the first and fourth position) is positively correlated with the percentage and the correlation coefficient is high. Additionally, the frequency of each letter used in the word in daily life and in the data set also affects the percentage distribution.

Finally, we integrate all these pieces of information. In order to represent a trade-off between players' strategy and blindness, we normalize them all and then map them to the range of (0,1) using the sigmoid function $\sigma(x)$. Finally, we introduce a beta value to control the relative weight between these two factors

$$R_s = H'(w)^\beta \cdot \sigma(F'(w)) \tag{15}$$

Where H' and F' represent the normalized values of information entropy and word frequency. The normalization ensures that $R_s \in (0, 1)$, which conforms to the range of probability values. β is a control factor that represents the ratio between personal attributes and player skills, and by adjusting β, we can better describe player behavior patterns.

8.3 State space tree search algorithm

The State Space tree is the solution space tree of the problem, and the process of solving the problem in Wordle is the process of state transition. Players calculate the probability of each guess in the guess space according to the prompt information, and update the State Transition Matrix according to the result of the guess, so as to make predictions and decisions. There are many State Space Tree Search Algorithms that can be used to optimize guessing strategies, such as Depth-first Search(DFS),Breadth-first Search(BFS), etc. Monte Carlo Tree Search (MCTS) [9]is a game tree algorithm, which has achieved good results in the man-machine battle of Go. In Wordle, we can use MCTS to simulate and fit the guessing strategy of players. However, not all states require a long search time to determine the best operation that the agent can find. For some common words, people may get the correct result on the first try, so we improve MCTS:

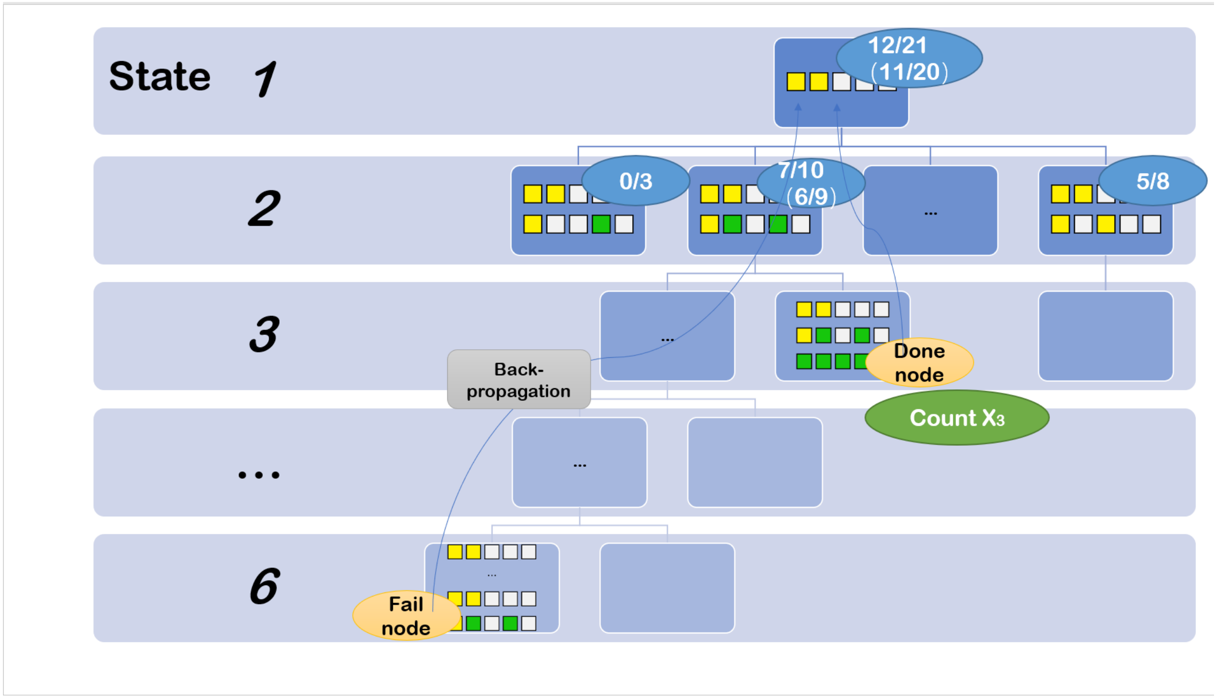


Figure 19: Improved Monte Carlo Tree Search

- ***Step 1:*** **Initialization**

In Wordle, the choice of the initial word can have a significant impact on the search process. A good initial word (i.e., a word with a higher number of correct letters in the answer) can greatly reduce the number of branching paths in the search, thus accelerating the search process. The selection of the "best initial word" has been widely discussed on the Internet. In our study, we adopted the well-recognized "SALET" algorithm for selecting the initial word.

- ***Step 2:*** **Search**

The search process is the key to this algorithm. Generally, the search step is performed according to the following four steps:

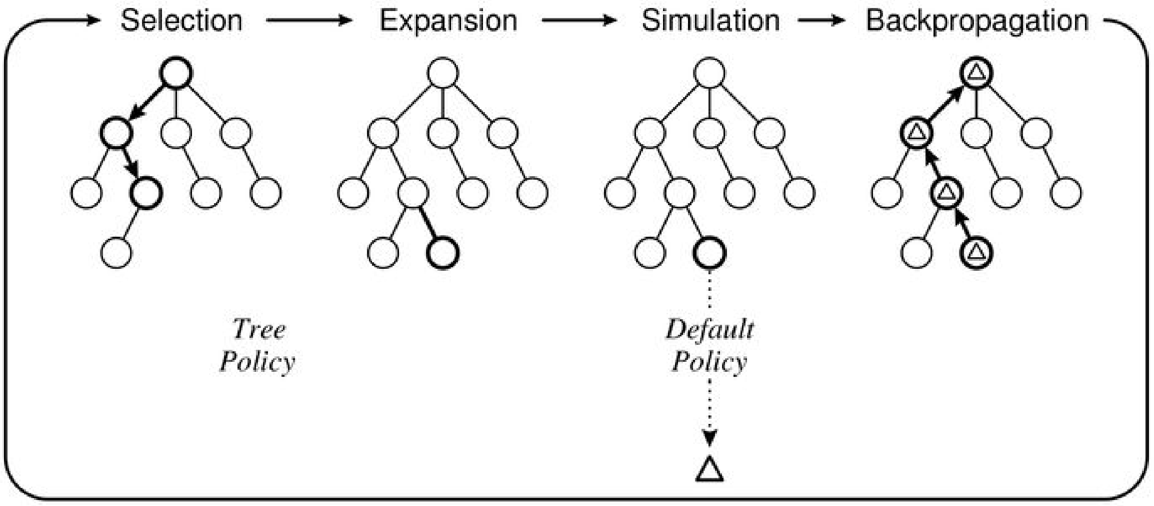


Fig. 2. One iteration of the general MCTS approach.

Figure 20: Repeated searching[5]

1.Selection: The search tree starts with the current state as the root node, and recursively applies the child selection policy to traverse the tree until an unexpanded node is reached. The selected child node represents the optimal choice in the current state. In this game, each selected child node represents the state S after a player has filled in a word.

2.Expansion: Nodes are added to the search tree by expanding it according to the available actions, and are recorded as (0/0), indicating that they have not yet been visited. However, we made the following improvement: if ”Done” is selected, we consider that we have made a good guess and should skip the expansion step and backtrack to the root node, thus completing the simulation of the game process.

3.Simulation: During simulation, the game is played according to the rules starting from the newly expanded node. If the result is ”win”, the node is recorded as (n+1/m+1), where n is the number of times it has been won and m is the number of times the node has been visited. If the result is not ”win”, the node is only recorded as (n/m+1), representing the number of times the node has been visited.

4.Back propagation : During the simulation phase, the game results generated are propagated back through the selected leaf node using reverse propagation, and the scores of all nodes on this branch are fed back to all parent nodes.

During the search process, the above four steps will be repeated over and over again.

- ***Step 3:*** **Termination**

In general, MCTS does not automatically exit. It returns the node with the highest visit rate as the ”optimal strategy” when it reaches the maximum number of iterations or stops based on a timer. This simulates the process of repeatedly considering multiple options and selecting the best one. However, in this problem, our goal is not to determine the optimal strategy, but to infer the players’ strategies based on their performance. Therefore, we need to modify the value returned to a 1×7 array that stores the probabilities.

By training the parameters of the BPNN to map to the source data, we obtain a fitted prediction model. Based on the residual analysis of the training data, we can conclude that the model can simulate the distribution of the reports obtained after a given word is presented. For the word ”EEIRE,” the predicted result is:

$$[0.507, 2.535, 18.464, 33.842, 29.252, 14.497, 0.901].$$

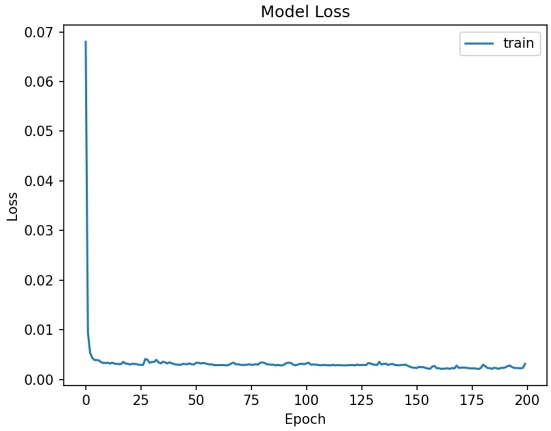


Figure 21: model loss

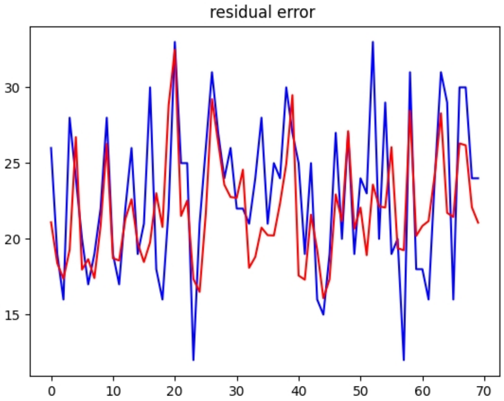


Figure 22: residual error

8.4 Uncertainty of Prediction

The selection of the initial word at the beginning of the search tree initialization is crucial because, in fact, there is a completely empty state S_0 hidden above the root state S_1. Once the root state of the search tree is evaluated, more than half of the future states can be selected, which will directly affect the entire Monte Carlo tree search process. Our improved Monte Carlo tree search algorithm is actually only limited to a certain tree in the forest formed by the state tree.

Due to the limitation of search computing power, we cannot afford the huge amount of data brought by the selection of the root node, so we cannot simulate the impact of different people’s choice of the first guessing strategy. Therefore, there is uncertainty in our prediction.

9 Model IV: Clustering Model Based on the Dirichlet Distribution

9.1 Dirichlet Distribution

Dirichlet Distribution is a multidimensional probability distribution, and its samples have the following characteristics:

1. The range is limited within the interval (0, 1), and the sum of all probability values is 1, indicating probability normalization;

2. The values of each sample are correlated in all dimensions, because their sum must be 1;

3. The number of samples is usually small, and normalization processing is usually not required.

For the daily reported percentage distribution, it is a vector with a fixed dimension of 7, each dimension representing the player's share in the guess, and their sum is 100%. Therefore, we can assume that the daily distribution of reported data may belong to some Dirichlet distribution[10].

The Probability density function of the Dirichlet distribution (PDF) is as follows:

$$f(\mathbf{p}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{k} p_i^{\alpha_i - 1} \tag{16}$$

Where, $\mathbf{p} = [p_1, p_2 ... p_7]$ is a probability vector with dimension k, k represents the number of tries from 1 to 7;

$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_7]$ is the parameter vector of the Dirichlet distribution, and it can represent the degree of preference for each event; For example, if $\alpha_1 < \alpha_2$, that means the number of people who guessed the word the second time was higher than the first time.$\boldsymbol{\alpha} = [1, 1, ...1]$ means that the probabilities of each attempt are about the same.

$B(\boldsymbol{\alpha})$ is normalized coefficient, makes $f(\mathbf{p}|\boldsymbol{\alpha})$ meet the definition of probability density function.

The parameter vector $\boldsymbol{\alpha}$ determines the shape of the distribution. The K dimensions enable it to have flexible shape changes, which can be well adapted to different data distributions. Therefore, we can use the Dirichlet distribution to describe the difficulty of different words.

9.2 Difficulty of words

In formula (16) above, each α_i of the parameter vector of the Dirichlet distribution represents the weight of the i-th dimension in the distribution, that is, α determines its shape, which includes the peak position and the degree of variation of the distribution.

When $\alpha_i > 1$, x_i has a higher probability density; When $\alpha_i < 1$, x_i has a lower probability density. If a dimension has a high value of α_i, the distribution will be more inclined to take a high value on that dimension. When each element α_i is equal, its shape is an uniform distribution, meaning that all possible values have equal weight.

For example, when we use a Dirichlet distribution to represent the probability of people guessing a word correctly, if α_i is larger, more people would have to guess n times to get the right word, and vice versa. If $\alpha_1 = \alpha_2 = \cdots = \alpha_K = \alpha_0$, then each word will have an equal probability of being guessed correctly.

The expectation of Dirichlet distribution is a vector where each component represents the expectation of a random variable , and all components sum to 1. To be specific, If the parameters of the dirichlet distribution vector is $\boldsymbol{\alpha}$, that means the expectation can be expressed as $\boldsymbol{\mu} = (\mathbb{E}(x_1), \mathbb{E}(x_2), \ldots, \mathbb{E}[x_K])$, where

$$\mathbb{E}[x_i] = \frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j} \tag{17}$$

From the distribution image, the peak position of the Dirichlet distribution is determined by the expectation vector $\boldsymbol{\mu}$, and the value of the curve on each component x_i is equal to its expectation $\mathbb{E}[x_i]$. To describe these positions more easily, we map the expected vector of dimension 7 to the interval 0,1, and obtain a new vector $\boldsymbol{\mu}'$. Thus, we can convert the new vector into a normal distribution defined continuously at 0,1, whose

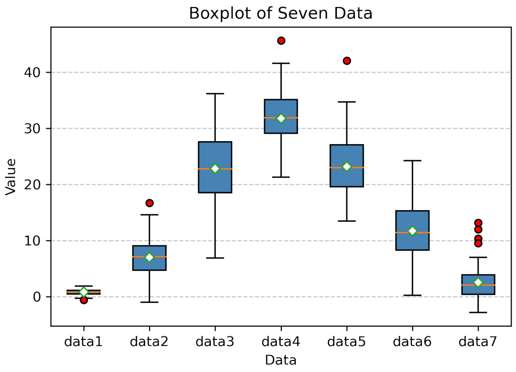


Figure 23: boxplot of distribution

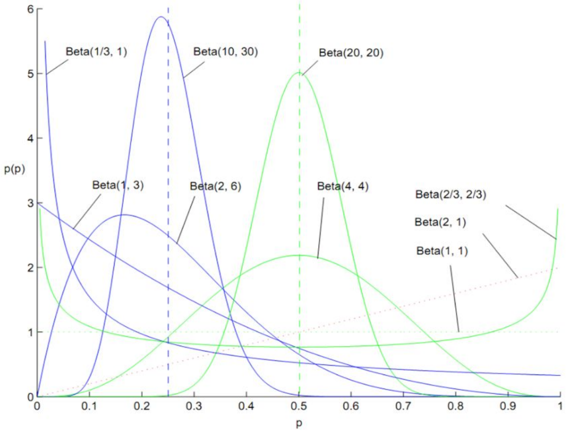


Figure 24: Dirichlet distribution[6]

peak value can be represented by the expected value of the new distribution. Finally, we take the left and right position of the peak, that is, the expected value of the new distribution, as the difficulty of the word:

$$Difficulty = \sum_{i=1}^{7} (\mathbb{E}[x_i] \cdot \frac{i}{7}) \tag{18}$$

With $Difficulty$, we can cluster the words first, and then rank their difficulty according to the parameters of each cluster to get a ranking from easy to difficult.

9.3 GMM clustering and analysis

The Gaussian Mixture Model (GMM) assumes that the samples in each cluster follow the same probability distribution, and takes into account the mean and covariance of the data distribution, which conforms to the characteristics of many data distributions. The commonly used algorithm for solving GMM is the Expectation-Maximization (EM) algorithm, which achieves this through the iterative Expectation step (calculating the probability that each sample point belongs to each cluster) and Maximization step (updating the distribution parameters of each cluster).

The advantages of GMM clustering are that it can fit clusters of any shape and output the probability that each sample point belongs to each cluster. The disadvantage is that it has a high computational complexity and is susceptible to the influence of initial parameters. To speed up the calculation, we can first use the k-means algorithm to divide the data into several clusters, and then use the parameters as the initialization parameters of GMM.

Then, we use parametric tests and silhouette coefficient tests to analyze the effectiveness of the clustering.

- Parametric tests

The K-S test is a non-parametric statistical method used to determine whether a sample distribution fits a certain theoretical distribution. The method calculates the maximum difference between the cumulative distribution functions of the two distributions, referred to as the K-S statistic. The smaller the K-S statistic, the more likely the sample data comes from the target distribution. Based on the K-S statistic and degrees of freedom, a p-value can be calculated to represent the probability that the sample data

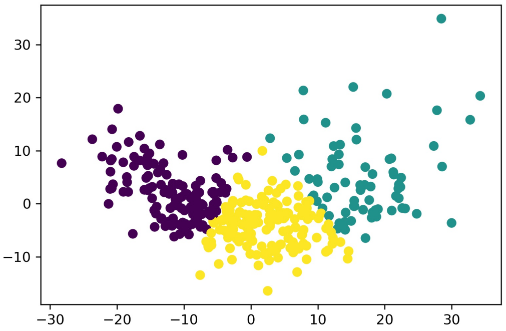


Figure 25: initialize with k-means

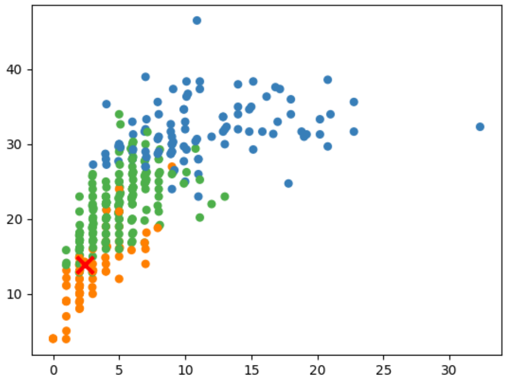


Figure 26: GMM clustering

fits the target distribution. If the p-value is smaller than a given significance level (usually 0.05 or 0.01), the null hypothesis can be rejected and it can be concluded that the sample data does not fit the target distribution. In classification tasks, we can perform parameter estimation and K-S test for each class, and determine the validity of the classification results based on the significance of the p-values. In this model, our p-values are [0.92, 0.87, 0.74], indicating the probability that the sample data fits the target distribution.

- Silhouette coefficient tests

The silhouette coefficient is a metric used to evaluate the quality of clustering, with values ranging between -1 and 1. It combines the compactness within clusters and separation between clusters to measure the similarity between each sample and its assigned cluster. The closer the value is to 1, the better the clustering result is as it indicates larger distances between the samples. The closer the value is to -1, the worse the clustering result is as it indicates smaller distances between the samples. In the current clustering model, the silhouette coefficient is 0.308, indicating that the distances between samples are relatively large and the clustering result is good.

9.4 Difficulty classification of words

We can use the Kullback-Leibler divergence (KL divergence) to measure the matching degree between a new word and a certain class, as we have already obtained the distribution parameters of the three classes of difficulty. KL divergence is a measure of distance or similarity between probability distributions, measuring the distance or difference between two probability distributions. The smaller the value of KL divergence, the more similar the two distributions are.

If $p(x)$ and $q(x)$ are two probability distributions, their KL divergence is defined as:

$$D_{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{19}$$

Suppose we add a new word to the sample, denoted as $x = (x_1, x_2, ..., x_7)$, and want to calculate its similarity with a certain parameter vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_7)$.

If the value of $D_{KL}(x||\alpha)$ is smaller, it indicates that the sample x is more likely to come from the Dirichlet distribution of parameter α.

For the three pre-classified categories $\mathbf{ff^1}, \mathbf{ff^2}, \mathbf{ff^3}$, we can calculate the KL divergence for each and classify x to the category whose KL divergence is the smallest.

$$\min_{i=1,2,3} D_{KL}(x||\alpha^i) \tag{20}$$

Computing the divergence of the word list with difficulty ranging from intermediate to difficult, we found it to be in the range of [-0.251, 0.353, 0.381]. We noticed that it is very close to the intermediate difficulty, but ultimately it was categorized as difficult. Based on the analysis of the words in Model 2, the following characteristics are very obvious: 1. containing multiple repeated letters, 2. low word frequency, and it is not a common word, and 3. it does not contain common letter combinations. These attributes add to its difficulty, and therefore this classification result is consistent with our understanding and expectation.

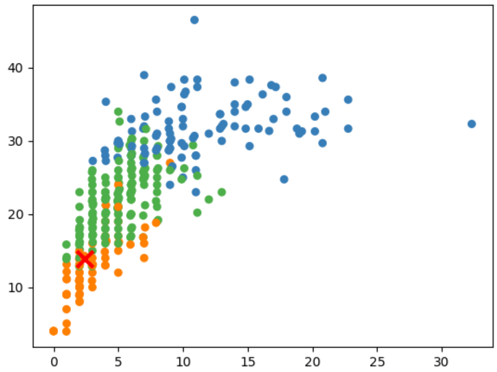


Figure 27: the position of word "eeire"

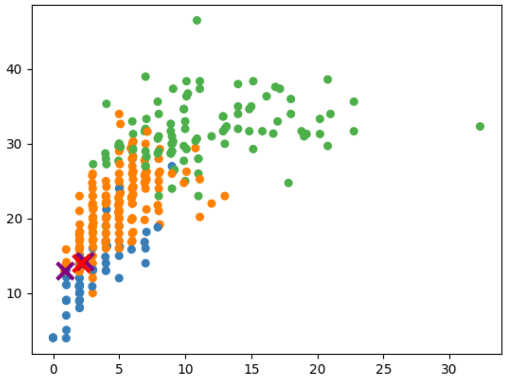


Figure 28: sensitive analysis

10 Sensitive Analysis

We fit the parameters of MCTS using BP neural networks to simulate the distribution of results and predict words in Model IV. To test the sensitivity of the model, we can adjust one parameter of the MCTS random simulation algorithm and compare the deviation of the results before and after adjustment. For example, in Figure (28), we adjusted the discount factor γ by $\gamma' = \gamma \times (1 \pm 5\%)$, and found that the predicted result for the word "EEIRE" was very close to the original prediction, which indicates that our model is not sensitive to changes in parameters.

11 Strengths and Weaknesses

11.1 Strengths

• We utilized wavelet analysis to extract the periodicity of the time series, then used the SARIMA model to capture the seasonal, trend, and cyclic features of the time series, and the ANN model was able to capture non-linear relationships and complex interactions, resulting in more accurate time series predictions.

- Since the sum of the normalized percentage of attempts is equal to 1, we innovatively used the Dirichlet distribution to describe the distribution characteristics of the percentage.

- We conducted a comprehensive data analysis on the data set, including data distribution analysis, clustering analysis, correlation analysis, and time series analysis, providing a comprehensive understanding of the features of the data set.

- We meticulously processed the data set and corrected the outliers in the case of high model accuracy.

11.2 Weaknesses

- Due to the limited availability of datasets we searched for, although our model achieved good predictive performance on existing datasets, its adaptability to large datasets may be limited.

- Due to computer performance issues, the scale of the Monte Carlo random seed generation cannot be too high, which may lead to some loss of precision.

References

[1] SHEN Xueyan. Design and implementation of backgammon algorithm based on Monte Carlo tree and neural network[D]. Shenyang University of Chemical Technology, 2021.

[2] PENG Lijun, MA Yueru. Seasonal problem of employee turnover and application of SARIMA-ANN model[J].Operations Research and Management,2021,30(02):139-145.)

[3] Peng Mingyi. Review clustering based on Dirichlet process and multi-item distribution mixing model[D].Shenyang University of Science and Technology,2021.DOI:10.27323/d.cnki.gsgyc.2021.000042.)

[4] Moeeni H, Bonakdari H, Ebtehaj I. Monthly reservoir inflow forecasting using a new hybrid SARIMA genetic programming approach [J].Journal of Earth System Science, 2017, 126(2)1-13.

[5] Get started with the Monte Carlo Tree Search MCTS[EB/OL].[2023.02.21]. https://zhuanlan.zhihu.com/p/26335999

[6] Codefmeister.Dirichlet distribution [EB/OL].[2023.02.21]. https://zhuanlan.zhihu.com/p/425388698

[7] Brown K A. MODEL, GUESS, CHECK: Wordle as a primer on active learning for materials research[J]. npj Computational Materials, 2022, 8(1): 97.

[8] Ruiz-Aguilar J J, Turias I J, Jiménez-Come M J. Hybrid approaches based on SARIMA and artificial neural networks for inspection time series forecasting[J]. Transportation Research Part E: Logistics and Transportation Review, 2014, 67: 1-13.

[9] Anderson B J, Meyer J G. Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning[J]. arXiv preprint arXiv:2202.00557, 2022.

[10] de Silva N. Selecting seed words for wordle using character statistics[J]. arXiv preprint arXiv:2202.03457, 2022.

MEMO

To: Puzzle Editor
From: Team 2308018
Date: February 20th 2023
Subject: Wordle Together : An Analysis About Player's User Profile

Dear editor:

I'm glad to show you some research findings of our team on Wordle game. For the prediction of the reported number problem, missing value query and outlier detection are quite important. For outlier data, we use Lagrange method for interpolation processing. After data cleaning, we conducted a simple visual exploration of the data, and found that it has an obvious trend in the long span and a periodic fluctuation in the small scale. According to this characteristic, we use wavelet analysis to decompose it and get the characteristic wave with periodicity of 5. Then we use seasonal ARIMA to search for appropriate parameters and predict the prediction interval of March 1(14554.8,15301.2)

In order to obtain the distribution of the number of attempts, we first put forward word attributes such as word frequency, vowel, whether it contains repeated letters, initial letters, does it contain roots and affixes, and select appropriate quantitative methods to calculate the correlation coefficients between them and the seven distributions. We found that the correlation coefficients of word frequency, duplicate letter and letter frequency are high, which shows a great impact on the distribution.

Based on the very good Markov nature of the wordle game rules, we adopted the Markov Decision Process (MDP) to describe the game process and designed a reward function based on word frequency and information entropy. Then we improved the Monte Carlo Tree Search (MCTS) based on the idea of random simulation. Its output is the frequency of finding the correct answer at each level after a certain number of games. We converted it into probability, BP neural network training parameters are used to fit the closest report distribution. The fitting degree of the model is very good, and the distribution of the word "EEIRE" is predicted to be [0.16 1.97 13.99 30.55 29.98 18.26 5.09].Finally, we analyzed that the uncertainty of our algorithm may come from the use of fixed initial words, because we can't afford the huge amount of data brought by selecting the root node

The distribution reported every day is a probability vector with a sum of 1, and according to this characteristic, the Dirichlet distribution is used to fit. Firstly, we define the peak position of the distribution curve as an index of word difficulty, and then use GMM to cluster the words in the data into three categories, and verify the rationality of the clustering results with the help of K-S test and contour coefficient. According to the expected size of the distribution vector, we rank the difficulty in simple, medium and difficult order. Finally, by calculating the KL divergence, we can verify the matching of a word with this class and classify it as the most suitable classification. We found that the difficulty of the word "EEIRE" is very difficult, which is also consistent with the multi-letter repetition but low word frequency feature we have previously verified.