Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2309795

Excavate the Wordle with N-gram language model

Summary

Wordle is a popular puzzle game released daily by *The New York Times*. Based on the **Markov assumption** and using an N-Gram language model, we propose a word difficulty index for the game Wordle, and develop a model using exponential fit and least squares regression to predict the number and distribution of reported results each day.

First, we use an n-gram language model and build a word association model to evaluate the difficulty of words in the Wordle game.By using the **perplexity index**, we determined the optimal parameters, n=3, and constructed the **Tri-Gram language model**.Then, we use k-means clustering to classify word difficulty into **10 categories**(figure 5).Finally, we evaluated the difficulty level of ERRIE based on the word **difficulty classification model as 7**, indicating it is a word with significant guessing difficulty in the game.

For predicting the reported results, we adopt a **layered prediction model**. In this model, we divide the time series into two layers, first predicting the overall distribution for each week and then predicting the distribution within each week.Then,comparing the prediction effects of exponential fit, GM(1,1) and ARIMA (0, 2, 0), we chose the **exponential fit** with better model effects.Combining the distribution within each week, we predict the range of reported results on March 1, 2023, with a **95% confidence interval** of **[11637.337, 25646.725]**, with the median value at **18642.031**.

We correlated the distribution of reported results with the attributes of the words and time, and only the difficulty of the words was strongly correlated with the distribution of reported results. And we establish a model for predicting the distribution of reported results based on the **least squares** approach and **skewed distribution**. We obtain the distribution of submission tries for March 1, 2023, and **the result shown in table 7** with RSS = 3300.92

In addition, we conducted a correlation analysis on word attributes such as the number of vowels, word frequency, word difficulty level, and the proportion of difficult patterns. We found a strong positive correlation between word difficulty classification and the proportion of hard mode, with a correlation coefficient of **0.8102**, and no correlation was found for the other attributes. As we continued to explore the data, **we were surprised to discover that**: The proportion of hard mode gradually increased over time;We find that the number of people reported on weekends was significantly lower than week-days;As the difficulty increases, the proportion of hard mode also increases.We explain our interpretation of these findings in the main text.

Finally, we optimize the Markov assumption based on the Tri-gram model and consider **discontinuous situations**. We provide a Wordle strategy guide to help players efficiently pass the game.

Keywords: N-Gram; Markov Chain; K-means; Skewed distribution; Time Series Forecasting;

Contents

1 Introduction

1.1 Problem Background

The prototype of Wordle was born in 2013, inspired by the color-matching game Mastermind. Initially, the game used all 13,000 possible five-letter words in English. Today, the game has been acquired by The New York Times and the word list has been reduced to about 12,000 more easily recognizable words through simple filters.

Nowadays, wordle is popular for its simple rules and one-word feature every day. It requires players to guess an actual five-letter English word in six or fewer guesses and will give feedback in the form of tile color changes after each guess. The hard mode adds a constraint to the regular mode that the previous guessed letter must be used in subsequent guesses, which makes guessing less likely. From the wordle user report, we can summarize interesting features of the game.

Figure 1: Wordle. `https://www.nytimes.com/games/wordle/index.html`

1.2 Restatement of the Problem

The first problem requires us to develop a model to investigate why the reported results fluctuate over time based on player game data, and to predict the range of reported results for March 1, 2023 using the model. Additionally, it is necessary to assess the relationship between word attributes and the percentage of Hard Mode scores.

The second problem necessitates the development of a model to project the distribution of reported results for each day in the future based on the given solution word, as well as to predict the distribution of reported results for the word "EERIE" on March 1, 2023. The uncertainty factors and accuracy of the prediction model must be determined.

The third problem requires us to summarize the difficulty classification model of solution words, determine the attributes of given words in the category, and classify the difficulty of the word "EERIE". The accuracy of the classification model needs to be evaluated.

Finally, we will explore other interesting features of the game's data and present the resulting findings to the crossword game editor at the New York Times.

1.3 Our Approach

This paper mainly establishes an N-gram language model, which first solves task3. Based on this model, the prediction problems of task1 and task2 are solved. Finally, the dataset is analyzed, and surprising parts of the data are identified.

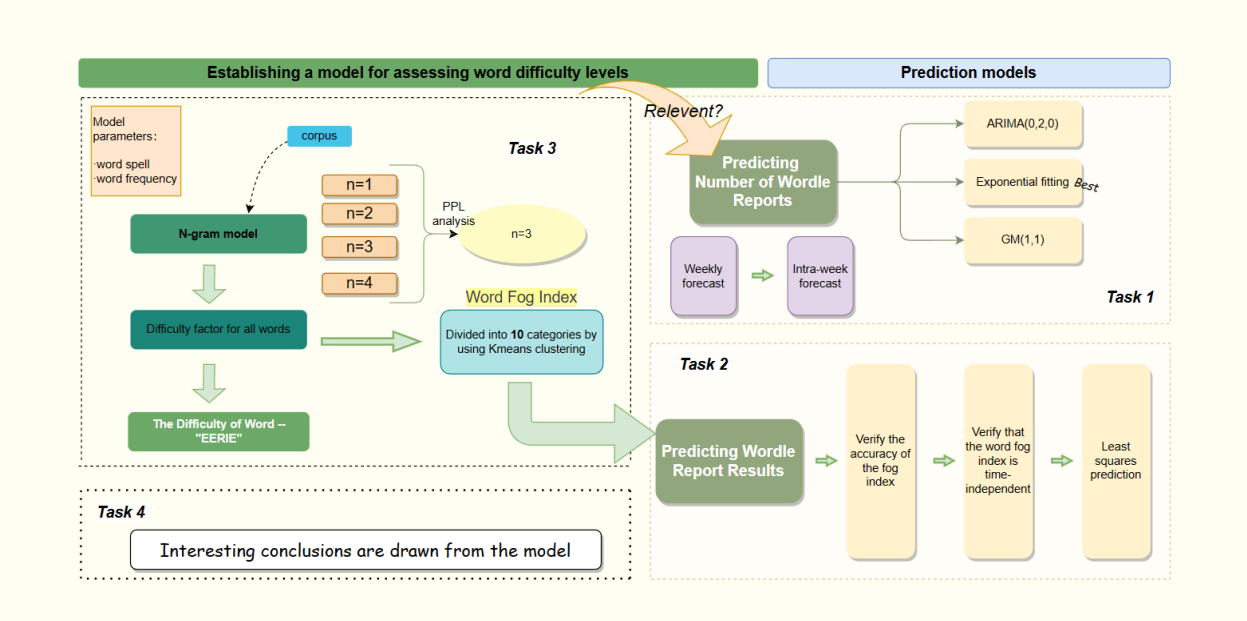

Figure 2: Design Evolution Diagram

2 Assumptions

- Based on the Markov assumption, we posit that each letter or letter sequence within a word is only dependent on the preceding letter or letter sequence.

- We assume that the data we are using is valid and reliable. The frequency calculated from the corpus can accurately reflect the usage level of a word.

- We assume that the popularity of the Wordle will not change violently to ensure the reliability of the prediction results

- We believe that all words with meanings contain at least one vowel letter 'a', 'i', 'u', 'e', or 'o'. For words that do not contain any vowel letters but have the letter "y", we consider it as having one vowel letter.

3 Notation

Symbol	Expaination
WAI	The Word Association Index
WDI	The Word Difficulty Index
$P(S)$	The probability of an English word
PPL	The perplexity of a word
w_i	the i_{th} letter in a word
n_i	The frequency of the i^{th} word in the word bank
w_{ij}	The j^{th} letter of the i^{th} word.
q_j	The j^{th} letter of a word.
X_{ij}	The number of reports in the j^{th} day of Week i
f_o	A ractical number
f_e	A theoretical number
H_j	The probability distribution of wordle reported results during the week
z_{ij}	The probability of the j^{th} tries to pass on difficulty i
z_{ijk}	The probability of the k^{th} word of the j^{th} tries to pass on the difficulty i

4 Word Difficulty Evaluation Model

4.1 Description of Data

we obtained all five-letter words from the Kaggle dataset, a total of 12973 words. And frequency data for each word is derived from the online Google Web Trillion Word Corpus.

During data processing, we find there are words beyond corpus, but in the wordle, so we define them as unknown words. Because the unknown word does not exist in the corpus, we assign a minimum frequency to the unknown words. In addition, this paper uniformly divides the frequency by 1 million for the convenience of data presentation.

4.2 N-Gram language Model

we established an evaluation system for the difficulty of words. As a measure of word difficulty, We develop an N-Gram language model to calculate the probability of words . The lower the probability that the word will be associated, the more difficult the word will be.

Using the N-Gram language model, we calculate the occurrence probability of each 5-letter word. Why do we use the N-Gram language model? The reason is that the N-Gram language model is a language model based on statistics, used to calculate the probability of a sentence model and judge the reasonable probability of a sentence. Therefore, it is in line with the conditions of calculating the probability of five-letter words. Besides, the N-Gram language model is faster than RNN and LSTM neural network language models, and owns more practical application significance.

In statistical probability, we calculate the probability of a word by equation(1):

$$P(S) = P(w_1, w_2, w_3, ..., w_n) \tag{1}$$

Where S stands for a word and w_n for a letter. The greater the probability $P(S)$ of an English word, the easier the word is.

Since we can't compute P(S) directly, we combine equation(1) with the conditional probability formula.

$$P(S|A) = \frac{P(A,B)}{P(A)} \tag{2}$$

We can obtain:

$$P(w_1, w_2, w_3, ..., w_n) = P(w_1) * P(w_2|w_1) * P(w_3|w1, w2)...P(w_n|w1, ..., w_{n-1}) \tag{3}$$

The conditional probabilities in equation (3) can be calculated based on statistics over the corpus. However, as the sentence gets longer, the continuous letters we are looking for will also be long. In fact, most continuous letters don't exist in the corpus, which will cause serious sparsity. Hence, this model has no practical significance. To solve the problem, we simplified equation (3) by Markov Assumption in this paper.

The Markov assumption is that the probability of each letter is only related to a few letters before it. For example, the second-order Markov assumption only considers the first two letters, and the corresponding language model is a trigram model. Markov chain is a random process in state space that undergoes a transition from one state to another, which is required to have the property of memoryless, namely the probability distribution of the next state can only be determined by the current state, and the previous events in the time series are irrelevant to it.

Hence, the model in this paper can be greatly simplified by the Markov assumption, which indicates that the current letter is only related to the previous few limited letters, hence, the equation (3) transforms into equation(4):

$$P(w_1, w_2, w_3, ..., w_n) = P(w_i|w_{i-m+1}, ..., w_{i-1}) \tag{4}$$

Among them,w_i stands for the i_{th} letter in a word.n denotes related to the previous n letters.

- When n=1, the 1-Gram model is:

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i) \tag{5}$$

- When n=2, the 2-Gram model is:

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i|w_{i-1}) \tag{6}$$

- When n=3, the 3-Gram model is:

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i|w_{i-2}w_{i-1}) \tag{7}$$

- When n=4, the 4-Gram model is:

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i|w_{i-3}w_{i-2}w_{i-1}) \tag{8}$$

In the above four models, the value of n plays a vital role in model performance: the larger n is, the higher the constraint information on the next letter is; The smaller the n, the more frequently the letter appears in the corpus and has higher reliability. When n=1, each letter is independent, and each letter only depends on itself, which does not accord with the real situation. Hence, we use the three models when n=2,n=3,n=4, respectively to calculate the probability of each word and measure the difficulty of the word, and use the given data to select the model with the best evaluation effect.

4.3 Take n=2 as an example to solve the N-gram Model

4.3.1 Construct the corpus

Hypothesize that n_i(i=1,2,$\cdots$,u) is the frequency of the i^{th} word in the word bank,and w_{ij} (i=(1,2,$\cdots$,u) j=(0,1,2,3,4,5)) stands for the j^{th} letter of the i^{th} word.According to the Markov hypothesis, the i^{th} word can be split into $(w_{i0}, w_{i1}), (w_{i1}, w_{i2}), (w_{i2}, w_{i3}), (w_{i3}, w_{i4}), (w_{i4}, w_{i5})$. where u refers to the total number of words in the dataset and w_{i0}means the empty before the letter at the beginning of the word.

Hence, the relative frequency formula can be denoted as:

$$f_i = \frac{n_i}{\sum_{i=1}^{u} n_i} \tag{9}$$

We use matrices to store information, and our corpus is:

$$\begin{bmatrix} (w_{10}, w_{11}) & \cdots & (w_{13}, w_{14}) & (w_{14}, w_{15}) \\ (w_{20}, w_{21}) & \cdots & (w_{23}, w_{24}) & (w_{24}, w_{25}) \\ \vdots & \ddots & \vdots & \vdots \\ (w_{u0}, w_{u1}) & \cdots & (w_{u3}, w_{u4}) & (w_{u4}, w_{u5}) \end{bmatrix} * (f_1, f_2, \cdots, f_u) \tag{10}$$

4.3.2 2-Gram Solution

We solve the 2-Gram model using conditional probabilities:

$$P(w_i) = P(q_1|q_0) * P(q_2|q_1) * P(q_3|q_2) * P(q_4|q_3) * P(q_5|q_4) = \prod_{i=1}^{5} (q_j|q_{j-1}) \tag{11}$$

Where we denote w_i as the i^{th} word, q_j as the j^{th} letter of the word.

$$P(q_j|q_{j-1}) = \frac{c(q_{j-1}, q_j)}{c(q_{j-1})} \tag{12}$$

Among equation(12), we set the number of occurrences of letter q_{j-1} in the corpus to $c(q_{j-1})$.

Owing to the large frequency of words in the corpus, log calculation of the data will not change the nature and correlation of data, but also make the data more stable and easier for computers to solve. Hence, we calculate the probability by log.

In the 2-Gram model, we take the word EERIE as an example to solve the model, we can obtain the relationship(13).

$$logP(EERIE) = logP(E|0) + logP(E|E) + logP(R|E) + logP(I|R) + logP(E|I) \quad (13)$$

We set $logP(EERIE)$ as the **WAI**(Word Association Index) to measure the difficulty of words in the wordle game.

$P(n|0)$ represents the probability that n is the first letter. The Worlde's rule stipulates that every word has five letters, hence, there's no need to take the probability P(0|e) into account.

As a result, we calculate the probability of 12973 words under the 2-Gram model. Similarly, we establish the 3-Gram model and the 4-Gram model and calculate the probability of 12973 words respectively.

4.4 Model evaluation

We introduced confusion degree as the evaluation index. Take the 2-Gram model for example:

$$PPL(M) = \sqrt[n]{p(q_0, q_1, q_2, q_3, q_4, q_5)} = \sqrt[n]{\prod_{j=1}^{5} P(q_j|q_(j-1))} \quad (14)$$

The perplexity(PPL) of all word data of each N-Gram(n=1,2,3,4) model is analyzed, and finally, and we obtain the geometric mean of all word perplexity. Through this way get the model evaluation score. The specific values are reflected in the following table:

Table 1: Model Evaluation Score

n	n=1	n=2	n=3	n=4
PPL	27.856	16.718	4.127	18.158

The less PPL, the more successful NLP is handled. So we chose the trigram model (n=3) to consider the difficulty rating of words.

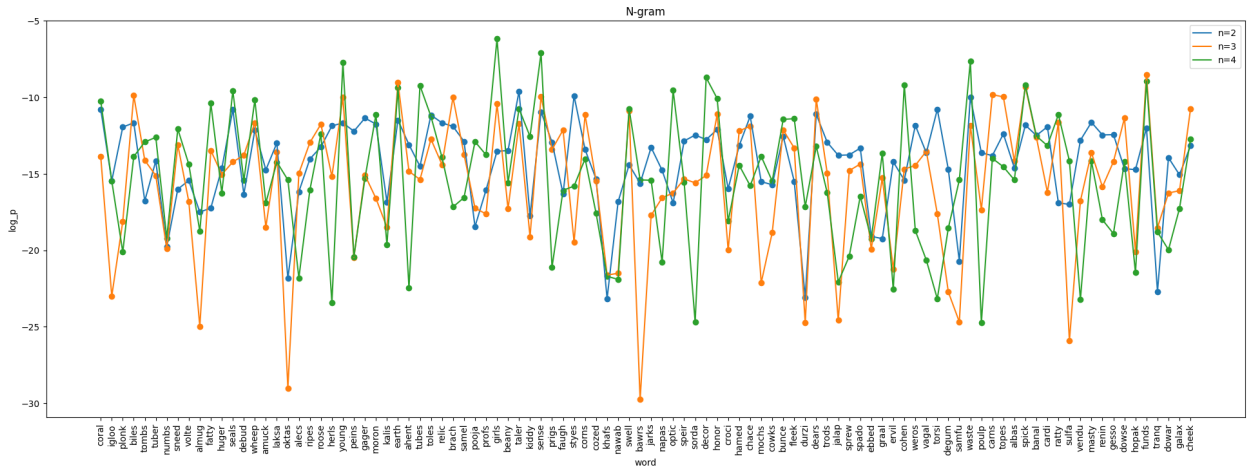

Figure 3: When n = 1,2,3,4, the change in the logarithm of the probability of being associated with a random part of a word

We randomly selected 100 words and entered them into the model of n=1, 2, 3, 4, and after many experiments, the value of the logarithm of the associative index of each word fluctuated, and we can see that when n=3, the fluctuation is more violent. This verifies the PPL test above, and also shows that the clustering effect at this time is also better.

4.5 Obtain the Difficulty level by K-means

Firstly, based on the word association model constructed in the previous section, we calculate the probability of word association for each 5-letter word in the dataset. Due to the small value of the association index, we take the logarithm for easier observation of fluctuations. Figure 4 depicts the relationship between the association index and word frequency.

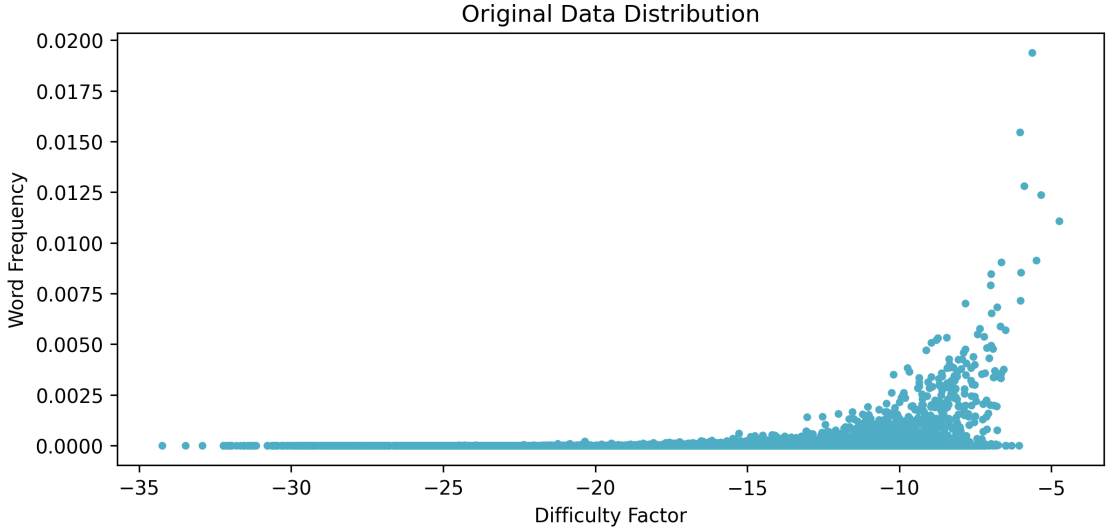

Figure 4: Origin data

We use the K-means algorithm to cluster the above datasets, and through the CH coefficient analysis, we think that clustering into 10 classes is better. The clustering results are shown in the figure, and the greater the difficulty level, the less likely its words are to be associated, and the greater the difficulty in the wordle game.

The word "EERIE" is substituted into the Tri-gram model to calculate its association:

$$WAI(EERIE) = logP(EERIE) = -22.0632$$

After K-means clustering, the difficulty level of the word 'EERIE' was rated 7, indicating that it is a very unfamiliar word, not easy to be associated with, and it will be very difficult to appear in the game Wordle.

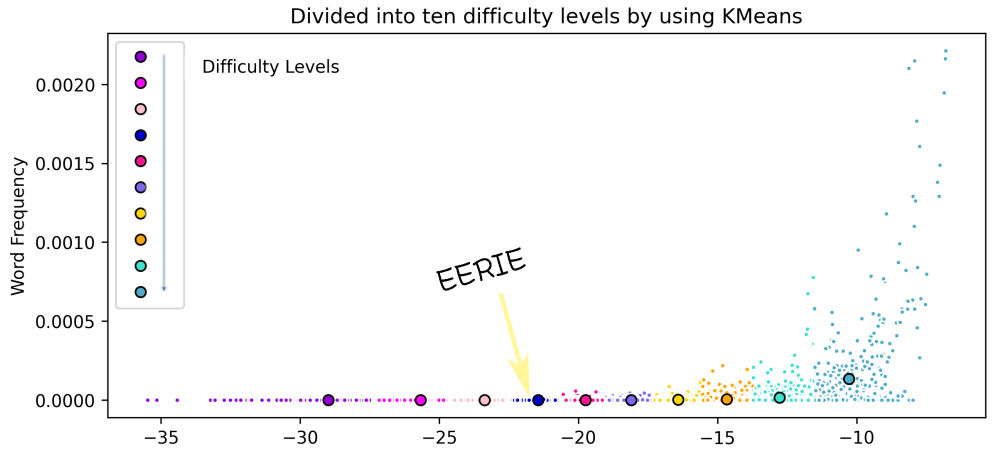

Figure 5: Difficulty levels and the position of the word "EERIE"

5 Predicting the Number of Wordle reports

5.1 Data Processing

Firstly, we process the outliers in the data. cf.Table2 and Table3.

Table 2: Before

Date	Word	Number of Reported Results
2022/12/16	rprobe	22853
2022/12/11	naïve	21947
2022/11/30	study	2569
2022/11/26	clen	26381
2022/10/05	marxh	30935

Table 3: After

Date	Word	Number of Reported Results
2022/12/16	probe	22853
2022/12/11	naive	21947
2022/11/30	study	24419
2022/11/26	clean	26381
2022/10/05	marsh	30935

- **Justification**
  - On 2022/11/30, the number of reported results is only 2569, while the Number in hard mode is 2405. Hence, We replace the value of that day with the mean of the Number of reported results of that week, which is 24419.
  - Since the answers in the game all have meaning and are 5-letter, we modified the meaningless and misspelled words to common words, and added them to the data set.

5.2 Data Analysis

First, we investigate whether there is a difference in the number of wordle report results from Monday to Sunday and whether the number of results is uniformly distributed.

We use the fit test in the chi-square test to determine whether the distribution is uniform. n.b.f_o is practical number,f_e is theoretical number.

$$\frac{\sum_1^7 (f_o - f_e)^2}{f_e} \sim \chi^2(6) \tag{15}$$

We denote that X_{ij} is the number of reports on the j^{th} day of Week i,and f_{oj} is the f_o on j^{th} day of Week, we can get:

$$f_{oj} = \sum_{i=1}^{50} x_{ij} \tag{16}$$

$$f_e = \sum_{j=1}^{7} \frac{f_{oj}}{7} \tag{17}$$

Table 4: Difference Judgment

	f_o	f_e	$\frac{(f_o-f_e)^2}{f_e}$
Monday	4586312.06	4609354.517	115.1907125
Tuesday	4709580.1	4609354.517	2179.300252
Wednesday	4787753.82	4609354.517	6904.721939
Thursday	4659214.83	4609354.517	539.3490106
Friday	4647933.88	4609354.517	322.9014459
Saturday	4462440.12	4609354.517	4682.616624
Sunday	4412246.81	4609354.517	8428.826221
	32265481.62		23172.9062

We find:

$$\frac{\sum_1^7 (f_0 - f_e)^2}{f_e} = 23172.9062 > \chi^2_{0.05}(6) = 12.5916 \tag{18}$$

Hence, rejecting the null hypothesis, the number of reported results has a difference from Monday to Sunday. Specifically, the number of reported results on weekdays from Monday to Friday was significantly higher than that on weekends, while the trend increased from Sunday to Wednesday and then decreased to Sunday.

5.3 Development of Prediction Model

According to figure 6, the number of reported results initially increased rapidly to the peak, and then rapidly decreased with a decline rate that was initially fast and subsequently slow.

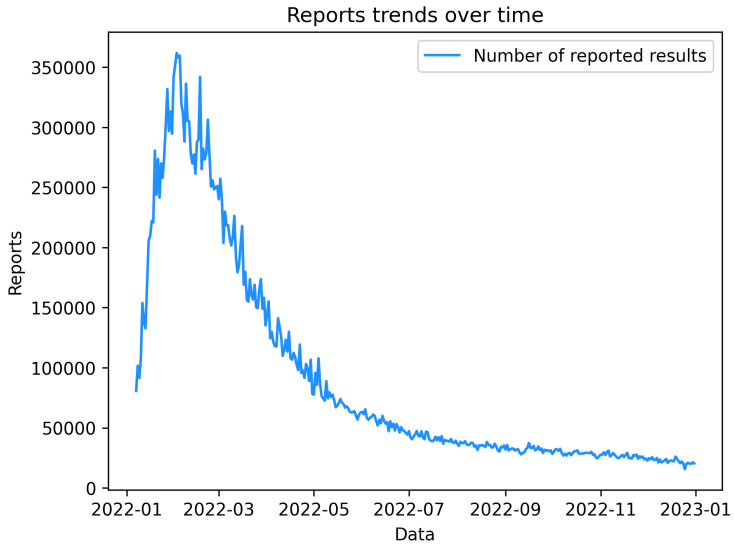

Figure 6: Line Chart of History Reports

Since intra-week differences exist, we build a hierarchical time forecasting model.We divide the time series into two layers, first predicting the overall distribution of each week, and then predicting the distribution within each week. We set 2022/1/10-2022/1/16 as the first week, so 2023/3/1 as the 59th week.

First, we calculated the difference change in the number of daily reports within a week and constructed a weekly wordle hot distribution. Where $H_j(j = 1, 2 \cdots, 7,)$ denotes the probability distribution of wordle reported results during the week.

$$H_j = \frac{\sum_{i=1}^{50} \frac{x_{ij}}{\sum_{j=1}^{7} x_{ij}}}{50} \tag{19}$$

After calculation, we get the weight vector for each day of the week:

$$\begin{aligned} H &= [H_1, H_2 ... H_7]^T \\ &= [0.145125, 0.146178, 0.144752, 0.143599, 0.145098, 0.138321, 0.137607]^T \end{aligned} \tag{20}$$

In order to explore the rate of change in the number of reported results, we solve for the second derivative, see Figure 7, and we find that the absolute value of the second derivative fluctuates

around 0 from 20 weeks, and the decline in the number of reported results is slow from 20 weeks. Therefore, in order to reduce the noise interference, we use the data from week 20 to week 50 to predict the number of results reported on Wednesday of week 59, which is March 1, 2023.

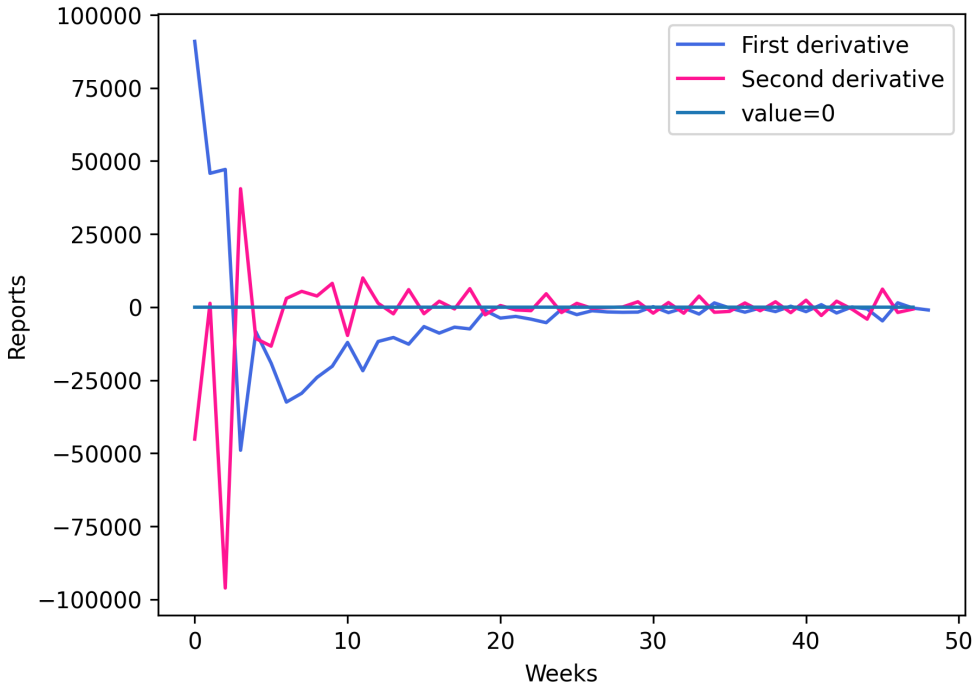

Figure 7: The Second Derivative of the Function

Then we forecast the inter-week differences, using the weekly average of 50 weeks to predict the average of the reported results in Week 59. After that, we forecast the number of reported results on Wednesday, March 1, 2023, according to the intra-week differences and the different distribution from Monday to Sunday.

5.4 Model Solving

We used curve fitting in Matlab for exponential with a 95% confidence interval, we assume that the equation of function is:

$$ax^b + c \tag{21}$$

We figured out:

$$a = (1.43 \times 10^6, 2.183 \times 10^6) \tag{22}$$
$$b = (-1.189, -1.065) \tag{23}$$

We take the average and finally determine:

$$a = 1.807 \times 10^6 \tag{24}$$
$$b = -1.127 \tag{25}$$
$$c = 0 \tag{26}$$

Table 5: Model Evaluation Index

SSE	R-square	RMSE
8.147×10^7	0.9794	1675

According to Table5, the fitting effect of this problem is good.

Meanwhile, we adopted the grey prediction model and ARIMA (0,2,0) time series prediction model to predict the 59^{th} week.

Table 6: Prediction Results

week	Arima prediction	GM(1,1) prediction	Exponential fitting
51	20555.0197	18086.856	21504
52	19602.48768	17136.429	21039
53	18655.11823	16197.232	20592
54	17712.91133	15269.131	20163
55	16775.867	14351.996	19750
56	15843.98522	13445.698	19353
57	14917.26601	12550.107	18971
58	13995.70936	11665.098	18603
59	13079.31527	10790.546	18248
R^2	0.684	0.864	0.9794
Mean Relative Error	/	6.86%	/

We can clearly see from the figure 8 that the exponential fit works well, with data points near the fitted curve.At 95% confidence, the confidence interval for the 59th week is [11485, 25311].

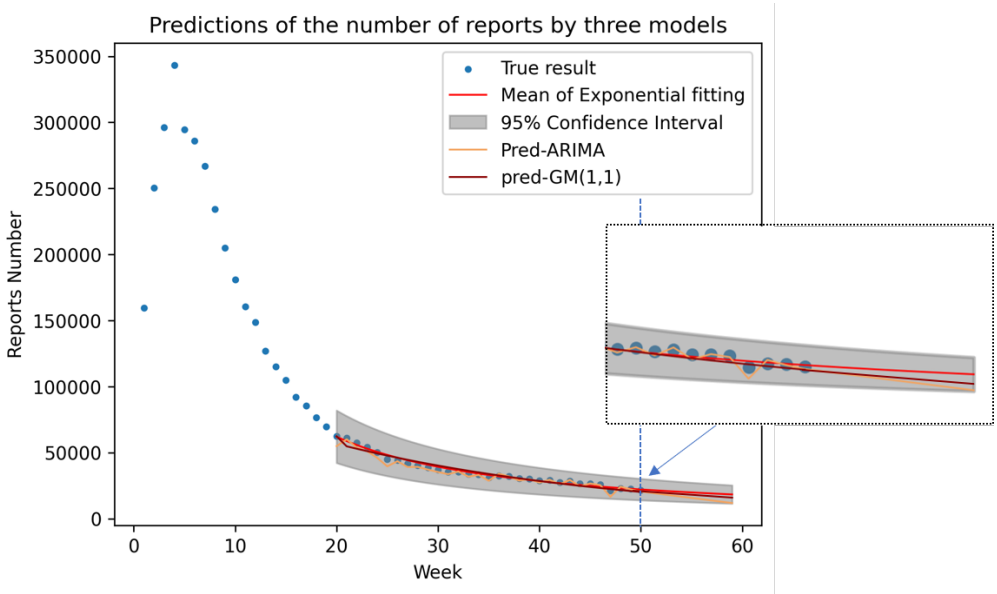

Figure 8: Predictions of the Number of Reports by Three Models

According to wordle hot distribution and week 59 mean of 18248, we can get:

$$X_{ij} = \text{Weekly Average} * 7 * H_j/vec$$

$$X_{593} = 18490$$

Based on the predictions of the three models, we decided to use exponential fitting to predict the 59^{th} week data.These are our justification:

1. The fitting effect and model accuracy of exponential fitting are significantly better than those of the other two.
2. Since the answers in the game all have meaning and are 5-letter, we modified the meaningless and misspelled words to common words, and added them to the data set.
3. Combined with relevant research and papers, we found that after a game's popularity is gradually decreased, the players of the game will gradually decline and the decline speed will slow down and finally tend to a minimum value. Under the GM(1,1) and ARIMA model predictions, it will eventually be negative.

5.5 Attributes Affecting the Percentage of Hard Mode

5.5.1 Effect of Word Relative Frequency and vowel letters' Number

We first investigate the effect of word frequency and vowel letters' Number on the proportion of hard mode. We perform Pearson correlation analysis on them.

It can be seen from Figure 10 that there is no correlation between word frequency and the proportion of hard mode.Besides there was no correlation between the number of vowels and the proportion of hard mode, while the correlation coefficient was 0.09.

5.5.2 Effect of Word Difficulty

In order to eliminate the influence of time and individual difference of words, we calculated the average value of difficult mode proportion under different word difficulty, to explore the correlation between difficulty and hard mode proportion.

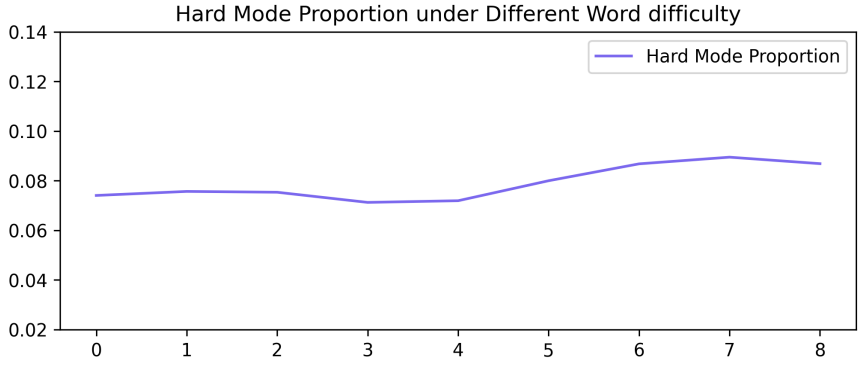

Figure 9: Hard Mode Proportion under Different Word difficulty

We calculated the correlation between the difficulty level and the proportion of commits of the Hardmode using the Spearman rank correlation coefficient:

$$r_{sperman}(WDI, HardMode\%) = 0.810151762 \tag{27}$$

We do correlation analysis on the number of vowels, hard mode percentage, and word Frequency. The correlation coefficients of Hardmode percentage, Number of vowels, and Word Frequency are 0.017 and -0.11, respectively, without correlation.

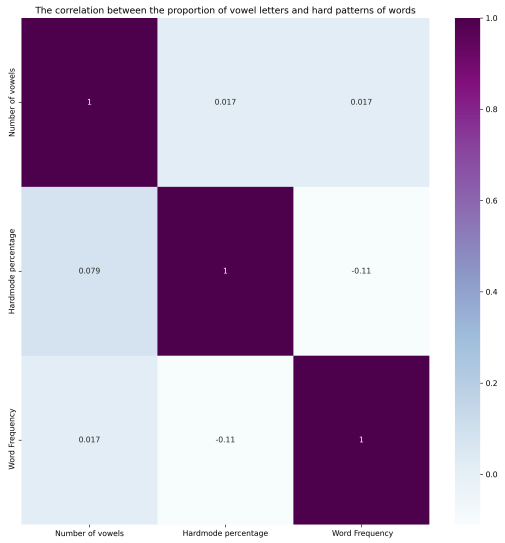

Figure 10: The Correlation between the proportion of vowel letters and hard patterns of word

6 Predicting the Distribution of Tries

6.1 Correlation analysis between difficulty index and times of try.

We first performed data processing. To eliminate individual word and time differences, we averaged the distribution probability of the times of try under the difficulty index.

Conduct correlation analysis on the data:

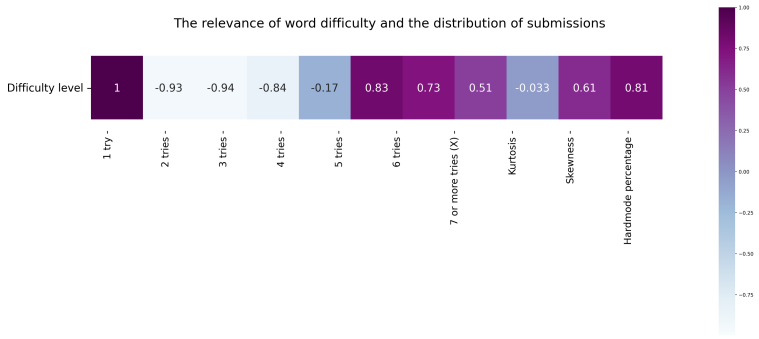

Figure 11: Effect of Time to the proportion of Hard Mode

We find that difficulty is strongly negatively correlated with 1try, 2tries, and 3tries, and the correlation coefficients are -0.934, -0.940, and -0.843 respectively, number the more difficult a word is, the smaller the probability of try times is. There is little correlation between difficulty and 4tries, and the correlation coefficient is -0.17 because the probability of 4tries is stable, and a stable value is imposed.

There is a strong positive correlate action between difficulty and 5tries,6tries 7tries, and the correlation coefficients are 0.826, 0.729, and 0.510 respectively. In other words, the more difficult a word is, the higher the probability of clearance times will be, which is It's in our perception that the harder it is, the more try it takes to complete the game.**This further proves the accuracy of the word difficulty model**

6.2 Correlation analysis between time and times of try

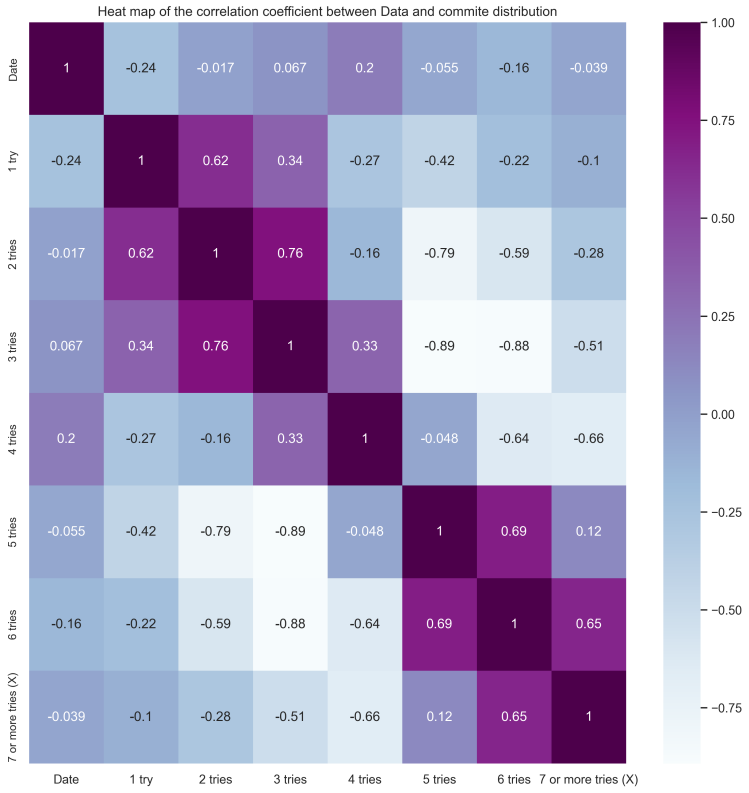

Figure 12: Correlation analysis between time and times of try

Since there is no correlation between time and probability distribution of try times, the probability distribution of try times is only related to word difficulty.

We counted the probability distribution under each difficulty. For the difficulty classification problem, to increase the model accuracy of the predicted value, we used the idea of the ordinary least squares method:

$$\min Z = \sum (z_{ij} - zijk)^2 \tag{28}$$

Where z_{ij} is the probability of the j^{th} tries to pass on difficulty i, And z_{ijk} represents the probability of the k^{th} word of the j^{th} tries to pass on the difficulty i.

We find that when between the minimum, z_{ij} is the prediction of the probability distribution under this difficulty index, and the minimum z is also the minimum sum of error squares.

Finally, we predict the distribution probability according to the difficulty index of the given words.

Based on this, the word 'EERIE', difficulty level is rated as level 7. Therefore, when the wordle puzzle of the day is "EERIE", the distribution of player attempts should be consistent with the distribution in the WDI=7 image shown in the figure13. By performing the above calculation, we obtained its distribution as:

Table 7: The distribution of the tries for "EERIE"

	1 Try	2 Tries	3 Tries	4 Tries	5 Tries	6 Tries	7 or more Tries(X)
Distribution(%)	0	3	18.9	35.3	27.2	12.8	2.6

The RSS of the model is 3300.92, and the model effect is relatively good.

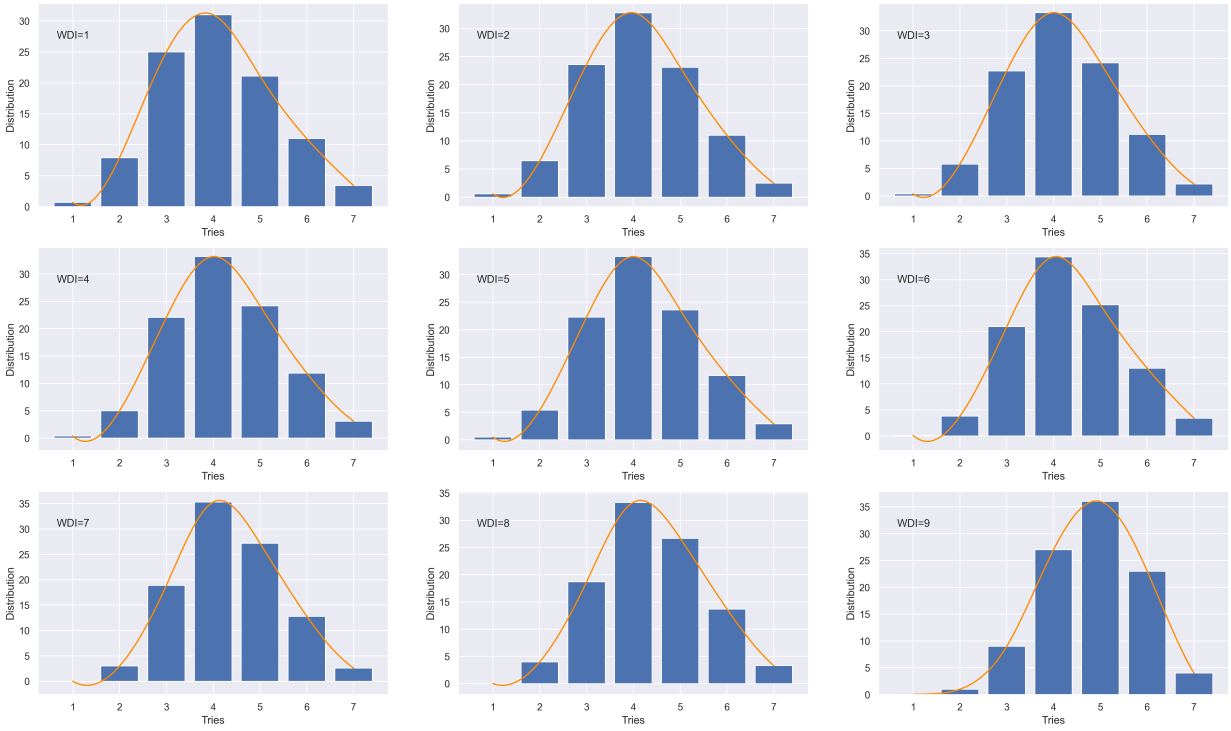

Figure 13: Pearson Correlation on Word Relative Frequency and Hard Mode Proportion

7 Something Interesting features

7.1 Effect of Time

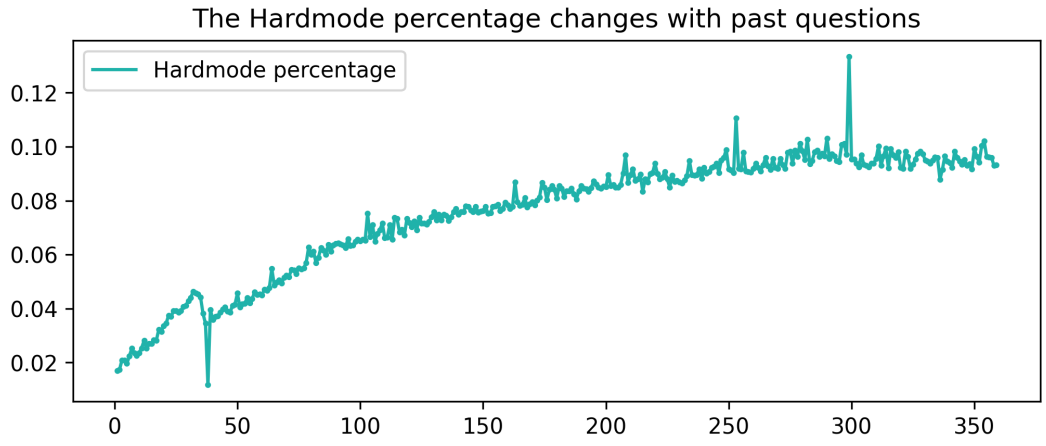

Figure 14: Effect of Time to the proportion of Hard Mode

In Figure14, the horizontal axis represents the time of the question. As people have a deeper understanding of wordle gradually, the proportion of hard mode increases.

- **Explaination**
  - Everyone has a deeper understanding of the game and is more proficient in word-guessing skills.
  - Due to the competitive and social nature of games, people are more likely to complete difficult modes and get satisfaction.
  - Online guess strategy gradually perfect and universal.

7.2 Other interesting feature

- We explain that people have a heart of comparison. When people around them complain that today's puzzle is too difficult, people tend to choose the hard mode rather than the ordinary mode and may pass the difficult mode with the help of external for85ces.

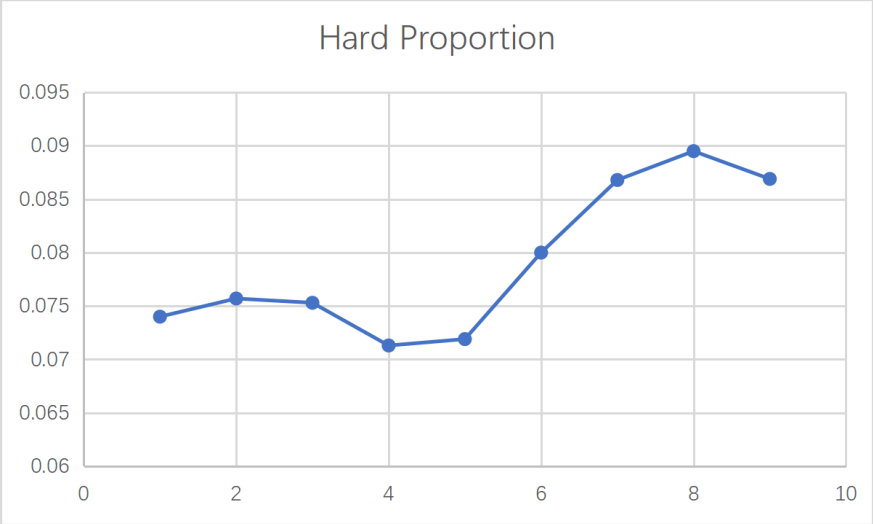

Figure 15: Hard Mode Proportion

- Minimum and maximum vowels correspond to a higher difficulty and hard mode proportion because common words have 1 to 2 vowel letters, hard mode proportion.

Table 8: Number of Vowel Letter, Word Difficulty and Hard Mode Proportion

Number of Vowel Letter	Word Difficulty	Hard Mode Proportion
0	5.333333333	0.081444075
1	3.085714286	0.073193772
2	3.217592593	0.075303226
3	4.142857143	0.081082017

- We find that the number of people reported on weekends was significantly lower than weekdays. We boldly guessed that people spent less time reading the New York Times on weekends and had other entertainment activities, so they could play wordle on weekdays to relieve work pressure.Or people tend to share wordle scores in real life instead of twitter in weekends.

8 Model Extension

8.1 Tri-Gram Model based on Discontinuous Markov Hypothesis

For the word Association Index based on the Tri-Gram model, the hypothesis in this paper is the continuous Markov hypothesis. However, discontinuity often occurs in hard mode, so we use the discontinuity Markov hypothesis to further optimize the model and give a wordle game strategy.

Based on the Tri-Gram model of the discontinuous Markov hypothesis, we built a corpus. And

we divided all the words in the word bank into:

$$\begin{bmatrix} (w_{10}, w_{11}) & (w_{10}, w_{11}, w_{12}) & \cdots & (w_{12}, w_{13}, w_{14}) & (w_{13}, w_{14}, w_{15}) \\ (w_{20}, w_{21}) & (w_{20}, w_{21}, w_{22}) & \cdots & (w_{22}, w_{23}, w_{24}) & (w_{23}, w_{24}, w_{25}) \\ \vdots & \ddots & \vdots & \vdots & \\ (w_{u0}, w_{u1}) & (w_{u0}, w_{u1}, w_{u2}) & \cdots & (w_{u2}, w_{u3}, w_{u4}) & (w_{u3}, w_{u4}, w_{u5}) \end{bmatrix} * (f_1, f_2, \cdots, f_u) \quad (29)$$

Assume that the correct word is $(r1, r2, r3, r4, r5)$.Input the initial value $(s1, s2, s3, s4, s5)$. When we determine a letter position, assuming it is $(01, 02, 03, r4, 05)$, 0_jindicates that the j^{th} position is unknown, we find all the words in the form of $(01, 02, 03, r4, 05)$from the corpus, calculate the probate ability of each word, and substitute the word with the maximum probability. If the second letter is not determined, select the word with the second highest probability, input it, and operate in turn until the second letter appears to determine the position $(01, r2, 03, r4, 05)$. We find all the words in the form of $(01, r2, 03, r4, 05)$from the corpus, calculate the probability of each distribution of the reported results' highest probability, and r repeat the operation until all 5 words were determined completely.

When the initial value of $(s1, s2, s3, s4, s5)$ is input and $2^{th}, 3^{th}, 4^{th}$ letters are determined, operate this until optimal.

• **Game strategy**

Since most words contain the vowels a, e, i, o, u, we assume that the strategy is to guess four vowels' word $Aeidu$ first, determine a correct letter position, and if it's all gray, guess $tryst$ without vowels. If we get 1 to 4 green positions, we can adopt the established Tri-Gram model based on the discontinuous Markov hypothesis.

8.2 Skewed Distribution to Predict Report Result Distribution

For the prediction of the distribution of reported results, we conduct a normal distribution test on the distribution of 359 words, which can be the approximately normal distribution, but there are differences in skewness and kurtosis.

We operate correlation analysis on the reported results distribution, skewness, and kurtosis.

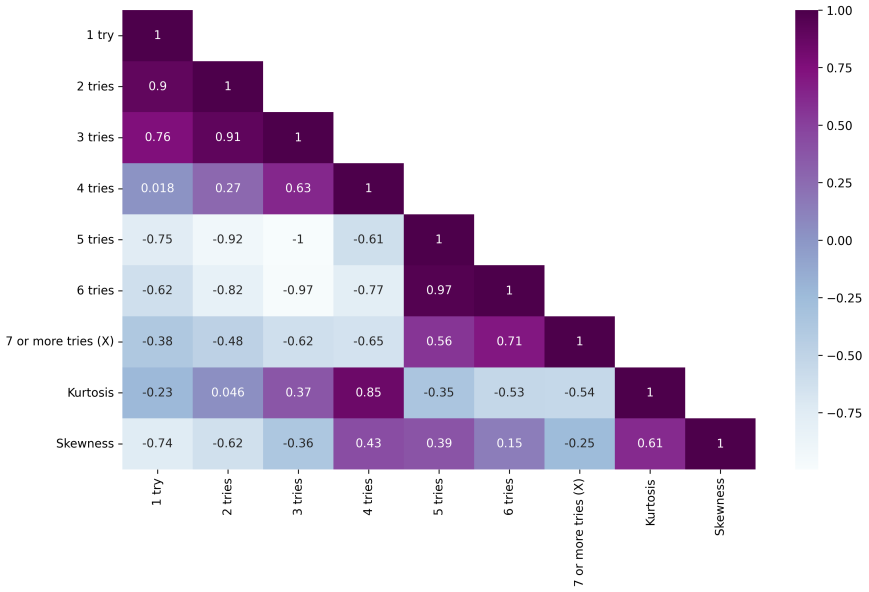

Figure 16: correlation analysis on the reported results distribution, skewness, and kurtosis

We found that skewness and kurtosis were correlated with the distribution of the reported results to varying degrees.Hence, we introduce a skewed distribution to predict report result distribution

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{30}$$

$$\Phi(x) = \int_{-\infty}^{x} \phi(x)dt \tag{31}$$

Y obeys the skewed distribution of SN($\mu, \sigma\lambda$), and similar probability density functions are defined as follows:

$$f_Y(Y) = \frac{2}{\sigma}\phi(\frac{y-\mu}{\mu})\Phi(\lambda\frac{y-\mu}{\sigma}) \tag{32}$$

Then we use parameter estimation to estimate parameter to get different probability densities of skewness distribution, which helps us to predict the distribution of reported results, and the model is more accurate than above.

9 Strength and Weakness

9.1 Strength

1. The analysis of our data mainly relies on the establishment of an N-gram language model, and the data of the training set and the experimental set are highly consistent, and the data is real and effective, which is a model with practical significance.

2. For the prediction of the report results, we considered the second derivative and adopted exponential fitting, which better predicted that the declining trend of the number of players would gradually slow down.

3. The classification and rating of the difficulty of words in this paper are designed specifically for wordle games, taking into account the association degree of letters and the frequency of words in daily life of players, with high accuracy of the model.

4. Our paper adopts a hierarchical time prediction model, which can better deal with multiple scales and periodicity in time series and improve the prediction accuracy.

9.2 Weakness

1. The attributes of words are rarely considered, and the emphasis is placed on the association degree and difficulty classification of words, while the attributes of words are not studied deeply enough.

2. The prediction model of reported result distribution is not accurate enough, so a better prediction method is proposed in the model derivation.

3. The study on the occurrence of yellow is not deep enough and the consideration is few.

10 A Letter to the New York Times

Dear Puzzle Editor:

Thank you very much for running Wordle, a seemingly simple game that not only provides physical and mental pleasure, but also gives us a taste of the interesting phenomena behind the data.

We studied the change rule of the number of woldle report results shared on Twitter over time, the relationship between word attributes and the proportion of difficult patterns, predicted the distribution of attempts on future dates, and established a word difficulty classification mechanism.We use mathematical methods to deduce phenomena and combine the knowledge of linguistics, psychology and sociology to explain these phenomena.And we are honored to present our results to you.

First, we use Tri-Gram model with the highest sensitivity to quantify the difficulty of words in Wordle game, and then use k-means clustering to divide the difficulty of words into 10 categories. Using this difficulty index, the difficulty of all English words can be judged. Therefore, the Puzzle editorial department was able to optimize the word selection mechanism according to our classification method to select a difficulty suitable for the public.

Second, we build a hierarchical forecasting model, which can predict the overall distribution of each week and the distribution within each week in the future. We also combined a variety of prediction models, such as curve fitting, GM(1,1) and ARIMA(0,2,0) models, to predict the distribution of the number of attempts in the 59^{th} week respectively, and selected the curve fitting with the highest accuracy, which proved to be of high reliability. Combined with the distribution for each week, we get the submission attempt distribution for March 1, 2023. These models allow Puzzle to increase the popularity of wordle by changing the word Settings on days with fewer weekly submissions, depending on the number of weekly submissions.

Thirdly, through correlation analysis, we found that the reported result count and distribution are not related to time or the number of reports each day. Word difficulty classification was strongly correlated with the distribution of reported results. Based on least squares method and skew

distribution, we built a model to predict the distribution of report results. Using this model, Puzzle editorial department was able to affect the distribution of report results by setting the difficulty of words.

Furthermore,a strong positive correlation between word difficulty classification and the proportion of hard mode was founded, and no correlation was found for the other attributes. Which suggested that the puzzle editorial department could focus on changing the difficulty of words when controlling the proportion of hard modes.As we continued to explore the data, we found something interesting: The proportion of hard mode gradually increased over time;Words with more or fewer vowels tend to be more difficult.The reasons may lie in the growth mechanism of game psychology, the distribution of vowel words in English linguistics, and players' psychology of challenge or cheating.

Finally, we optimize the Markov hypothesis based on the triplet model to provide a Wordle strategy that guess four vowels' word $Aeidu$ first, determine a correct letter position, and if it's all gray, guess $tryst$ without vowels.This helps the player complete the game more efficiently.What's more,it is a starting point for Puzzle editorial department to change the game more fun.

Thank you again for considering our models, and hope that our model will be benefit to future update and operation of woldle.

Yours sincerely,
Team#2309795

References

[1] Qiong Wang,Wenzhen Kuang, Li Xu. Text error checking algorithm based on improved N-gram model and knowledge base[J].Computer Applications and Software,2021,38(10):310-315+320.)

[2] I. L, M. M, M. L, E. A, et al. Word n-gram attention models for sentence similarity and inference[J], Expert Systems with Applications, 2019, 132(): 1-11.

[3] Adam Pauls, Dan Klein. Faster and smaller N-gram language models[C], Annual Meeting of the Association for Computational Linguistics, 2011, P11-1(): 258-267.

[4] Martin B. Short. Winning Wordle Wisely, arXiv preprint arXiv, 2022, 2202(02148)

[5] Wordle. https://www.nytimes.com/games/wordle/index.html.

[6] Kaggle. https://www.kaggle.com/datasets/rtatman/english-word-frequency.

[7] Wordle Estimation Tool. https://public.tableau.com/app/profile/wrenathan/viz/WordleEstimationTool/WordleEstimationTool.

[8] vision. https://visionon.cn/clipboard.