Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2304682

Secrets of the Wordle

Summary

Wordle is essentially a letter jigsaw puzzle. The game is about using data to find the best guess strategy. In this article, we will analyze the patterns of past data and build a word prediction model to help players move quickly through the game.

In Question 1, we analyzed the trend of the reported results and concluded that the change was due to the number of people playing the game. We proposed a virus invasion model, which consists of two stages: rapid increase and slow decline. The rapid increase is caused by the game's playability and a large number of potential players. The slow decline is due to players losing interest in the game, leading to some players giving up while core players continue to play, resulting in a slow decrease to a stable range. Based on this model, we predicted that the reported results on March 1st would be **21224** and would fluctuate **within a predicted range of 8796**.

In Question 2, we first observed the distribution of the number of attempts to answer the questions over a year and found that it was relatively stable despite fluctuations, leading us to speculate that this distribution only has a direct relationship with the difficulty of the word and other factors are disturbances. With this hypothesis, we used two methods of difficulty classification and extracted corresponding features to determine the distribution of the number of attempts for the target word "EERIE." We matched the word to its corresponding level of difficulty and found that the resulting distribution was similar between the two classification methods, validating each other. We ultimately obtained a distribution of the target word that closely approximated **Table 6**.

In Question 3, we first followed the conclusion of question 1, that is, the change in the number of survey results is caused by the change in the number of people. Through the attribute analysis of words, we established the **Central Gravity Model (CGM)** by using indicators such as letter frequency and word frequency, so as to facilitate the classification of the difficulty of the same words. In this model, we not only consider the influence of the attributes of a word itself, but also consider the influence of the relative relationship between other words in the database and their difficulty. Finally, after quantifying the corresponding difficulty index value of each word, we conduct cluster analysis to establish the difficulty range and realize the division of the difficulty of words. Then put the corresponding word attributes of ERRIE into the model to solve the difficulty score of the word. In terms of accuracy test of the model, we use part of the data given as the database of the model, and the other part as the control group for correlation analysis and fitting degree analysis.

Keywords: Virus Invasion Model, Cluster Analysis, Central Gravity Model

Contents

1 Introduction

1.1 Problem Background

In Wordle, players need to guess a word with five letters in it six times a day at most. After each attempt, the player may receive three kinds of feedback: green means the letters are in the correct position; Yellow means the answer contains the letter but is in the wrong place; Gray indicates the answer does not have the letter. The gameplay is similar to games like Mastermind, but Wordle makes it clear which letters are correctly guessed.

In essence, wordle is a word game in which players have to guess all the letters and their corresponding positions within a limited number of times each time, and the combination of these letters is the formation of words. Of course, there are different ways to play, such as whether the guessed word will be used in the next guess. Different game modes bring different experiences and difficulties to the player.

1.2 Restatement of the Problem

Considering that the data of wordle competition changes from day to day, we need to build a model to solve the following problems：

- Report the results to solve the relationship between the number and date of change, and to make predictions.
- Research word attributes related to whether difficult mode ratio.
- According to the word forecast attempts in the day and the corresponding percentage range.
- To figure out what are the uncertainties of the model.
- In forecasting based on specific word, as well as to the accuracy analysis of the model.
- Classifying words is difficult, and reflect classification basis.
- To quantify the specific word harder, and discuss the accuracy of quantitative models.
- To find some characteristics of data sets.

1.3 Our Work

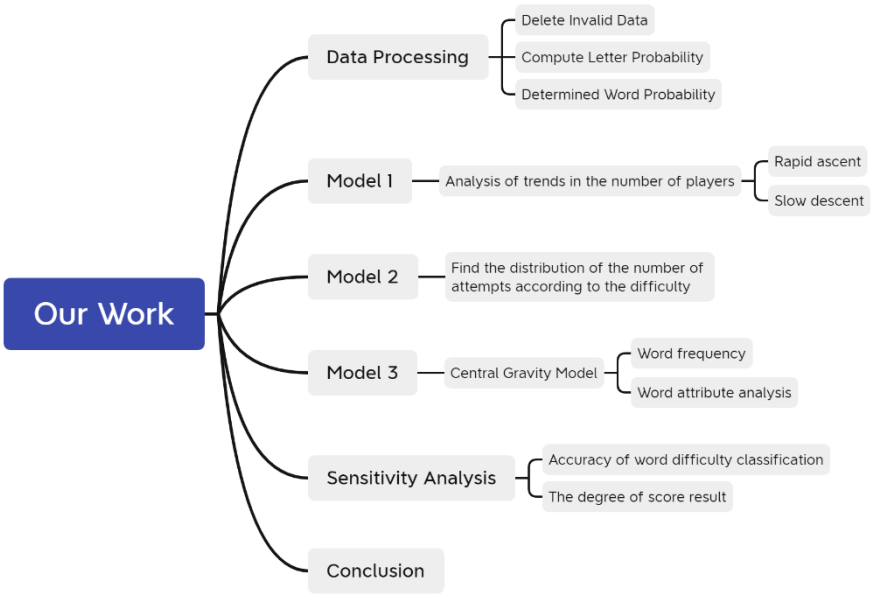

Figure 1: Flow chart of our work

Our work involved several key steps. First, we processed the data by deleting invalid data and calculating the probability of each letter and word. Next, we developed a model to analyze the trend in the number of Wordle players over time, which showed a rapid ascent and slow descent that was positively correlated with the number of survey results. We also developed another model to investigate the distribution of the number of attempts according to the difficulty of the word.

Additionally, we developed a central gravity model to analyze word frequency, classify word difficulty, and identify word attributes. We conducted sensitivity analysis to evaluate the degree of fit for the number of score results and the accuracy of word difficulty classification.

Overall, our work contributes to a better understanding of the Wordle game and sheds light on the factors that drive player engagement and interest in the game. However, there are limitations to our work, including the subjective nature of the parameters and assumptions used in the models and the complexity of the real-world situation.

2 Assumptions and Justifications

Assumption 1: We assume that the change in the number of reported results is caused by the change in the number of participants.

Justification: According to the change trend of the number of reported results over time, it can be seen that the number of reported results increases rapidly at first and then decreases slowly, which corresponds to the novelty hunting mentality that people often have when they meet a new game. At the beginning when the game is launched, many interested people are invested in the game. As time changes, some players' enthusiasm for the game fades, so the number of players gradually decreases. Gradually it becomes stable. This assumption was very important to our model, and from this we came up with the wordle player population change model.

Assumption 2: It is assumed that players who lose interest in Wordle will not start playing the game again.

Justification: Overall, the number of players is declining. In the long run, over time and for players as a whole, the loss of users is far greater than the influx. So don't take into account the few people who get tired of the game and then start playing it again some time later.

3 Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description
S	potential players
I	players
R	players that lost interest in wordle
β	how obsessed the potential player is with the game
c	zoom center
γ	player loyalty
s	magnification

4 Solution of problem 1: Wordle Virus Model

4.1 Data Preprocessing

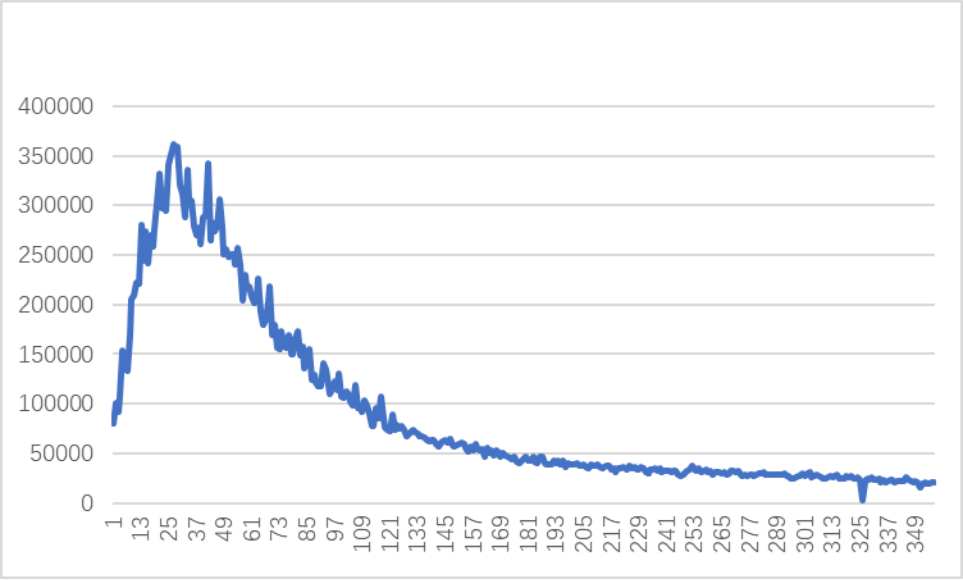

Figure 2: Number of reported results

The dataset given for this problem contains 4 problematic entries, with two having a word length of 4, one having a word length of 6, and one having a significantly different magnitude compared to the other entries. We have removed three of these entries and fixed the remaining one according to the given instructions.

By plotting the trend of the number of reported results over time, we can conclude that the number of results increases rapidly at first and then decreases slowly over time. Through data analysis, we can observe that the expected value of the number of tries by players has not significantly changed. Therefore, we can infer that the change in the number of players is related to the change in the number of reported results. The change in the number of players leads to a change in the number of reported results, and the two are positively correlated. We have developed a player model that varies over time. Next, we will discuss the changes in the number of players.

4.2 The Establishment of Model 1

By observing the change in the number of reported results over time, we found that the trend of the change in the number of reported results is similar to the change in the number of infected people caused by the spread of a virus, which we believe also conforms to the spread

characteristics of the game. This process can be described as follows: when the total number of players is determined, assuming that there is a certain number of initial players, the Wordle game will first spread widely in the potential player population, which will convert potential players into players and cause a sharp increase in the number of reported results. Since the total number of players is limited and players' enthusiasm for the game will gradually diminish, players will gradually become disinterested and lose interest in playing Wordle. Therefore, as time goes on, the total score will gradually decrease.

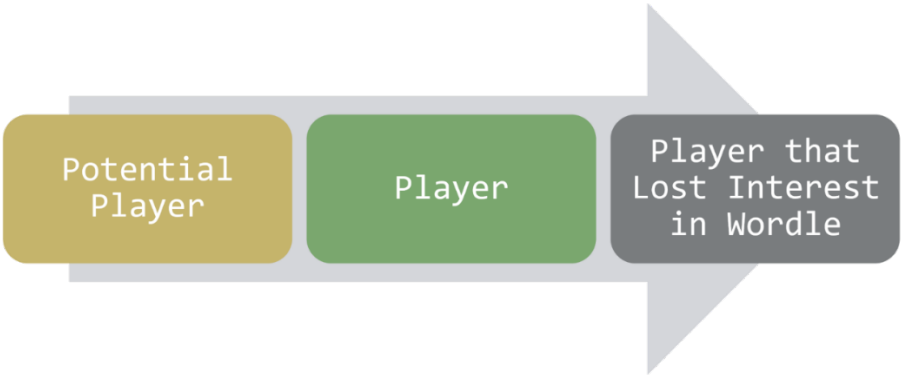

Figure 3: Transformation of Players

We assume that after players lose interest in Wordle, they will not start playing Wordle again. We have constructed a differential equation model to describe this process:

$$\begin{cases} N = S + I \\ \dfrac{dS}{dt} = -\beta \dfrac{S \times I}{N} - \gamma I \cdot \dfrac{1}{1 + e^{(t-c)s}} \\ \dfrac{dR}{dt} = \gamma I \cdot \dfrac{1}{1 + e^{(t-c)s}} \end{cases} \tag{1}$$

To quantify the attractiveness of the game to potential players, i.e. the likelihood that potential players will learn about and start playing the game from current players, we have defined a metric, beta, which is positively correlated with the playability of the game and the number of players. In addition, corresponding to the attractiveness of Wordle to players, we have defined a metric, gamma, which is related to the difficulty of the game and players' personal feelings. Besides these metrics, we have also defined the parameter C as the effective population size. Finally, through multiple simulations using a Monte Carlo model, we obtained the values of these metrics: $\beta = 0.14$, $\gamma = 0.03$, $c = 100$, and $s = 0.012$.

Regarding the observation of the change in the number of reported results over time, it differs from the typical infection curve of a contagious disease in that the decreasing rate of the

number of reported results becomes slower and there is a certain distance from 0. From the overall perspective of players, we speculate that this may be due to the existence of Wordle enthusiasts, who lose interest in Wordle at a much slower pace than regular players. Therefore, we constructed a sigmoid function and applied it to gamma to describe the phenomenon of the slowing rate of players abandoning the game. The modified gamma takes the following form:

$$\gamma(t) = \gamma \cdot \frac{1}{1 + e^{(t-c)s}} \tag{2}$$

4.3 The Solution of Model 1

4.3.1 Model Validation

To test our model, we projected the number of players over the next year and compared it to the number of players already in the data sheet. Qualitatively, we could see if the predicted line matched the actual data to judge the quality of the model.

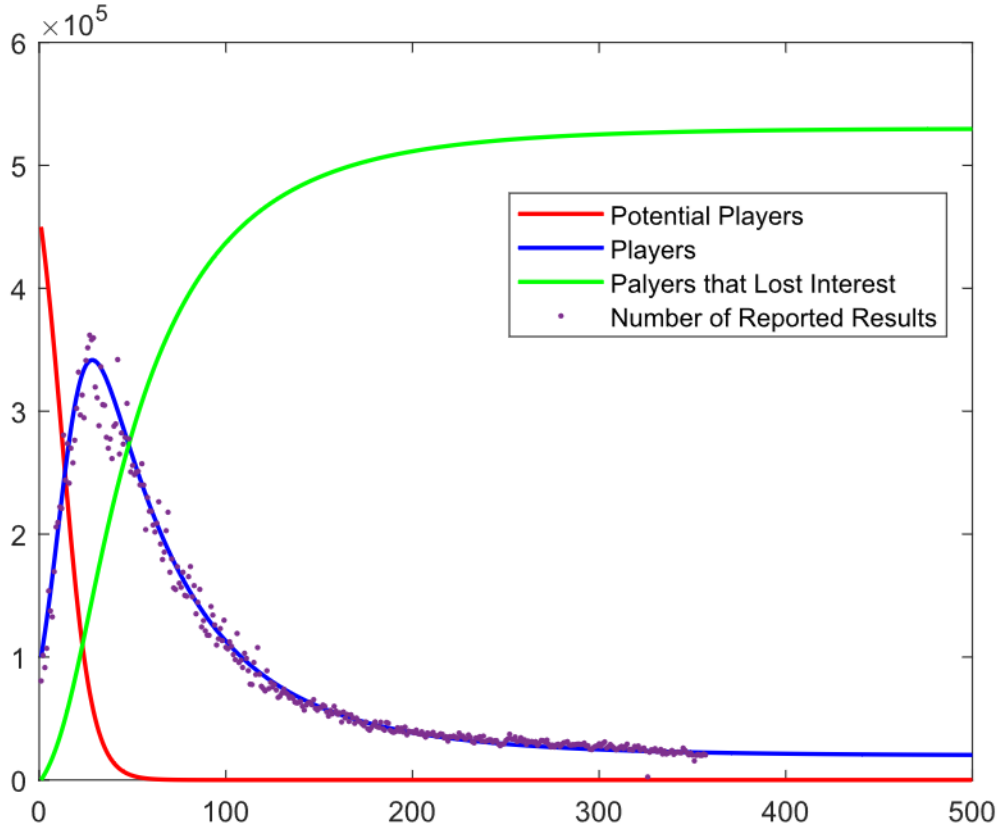

Figure 4: The fitting effect between the actual number of players and the predicted number of players

From this figure, we can see that the change in the number of predicted results is actually caused by the change in the number of players of different categories. When the game was just

launched, a large number of potential players rushed into the new game, leading to a rapid rise in the number of players. After a period of time, potential players have basically invested in the game, while some players in the game left due to the problems existing in the game itself. This will cause the number of players to decrease over time, but there will always be some players who are addicted to the game, which means that the number of players will never go to zero, and the number of players is positively correlated with the number of submissions, so we can know why the number of submissions changes.

According to our model, a reasonable prediction for the report result on March 1, 2023 is 21224. In order to verify the accuracy of this prediction, we have calculated some parameters in the table for verification.

Table 2: Parameter of the prediction of Wordle Virus Model

Parameter	Value	Parameter	Value
SE	4488.150118	R Square	0.927052
SSR	2.84881E+12	Lower bound of prediction interval	12427.22577
SST	3.07298E+12	upper bound of prediction interval	30020.77423

Based on our model, we can predict a reasonable report result of 21224 on March 1, 2023. To verify the accuracy of this prediction, we calculated some parameters in the table. By computing the standard error SE, we obtained the lower and upper bounds of the prediction interval, which are [12427.2, 30020.7]. Furthermore, we calculated the R-squared value, which is 0.927, indicating a good overall fit of the model.

To further explain the fitting of the model, we conducted a residual analysis [4] and obtained the following residual plot. From the plot, we can see that the model error was initially large but gradually decreased as the game developed, indicating a good fit. The reason for the poor fit at the beginning was that as the game started, the user population rapidly increased, and people did not have a scientific understanding of the game's strategy. This resulted in a high level of uncertainty in the initial report results and a poor fit of the model. However, as time went on, the uncertainty gradually decreased, which greatly improved the accuracy of the prediction.

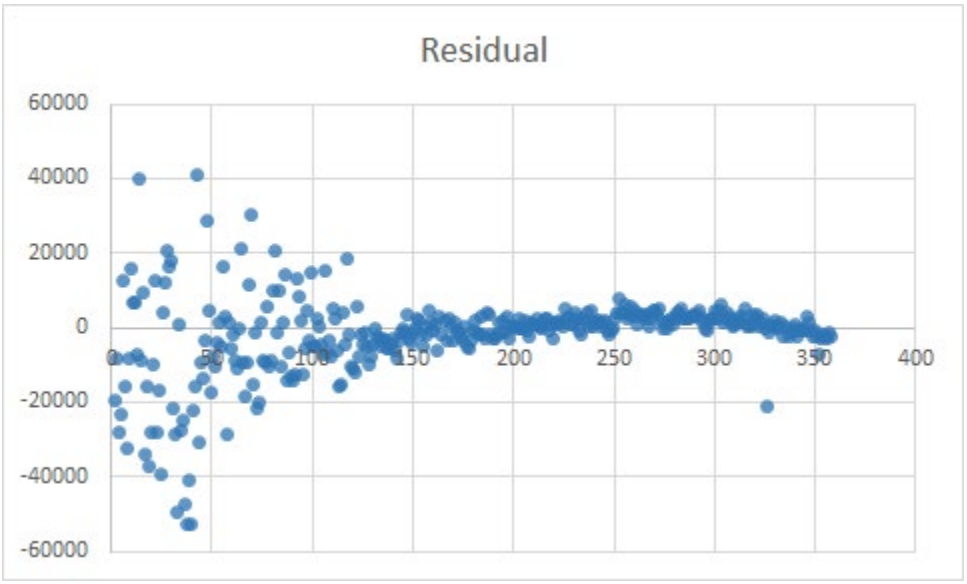

Figure 5: Residual of prediction of Wordle Virus Model

4.3.2 Relationship between word attributes and difficulty patterns

In order to test whether the word attributes affect the percentage of the report in hard mode, each data in the database, namely the probability of letters contained in the word, is extracted, weighted average processing is performed on it, and correlation analysis is performed on the percentage of hard mode of its corresponding date. The significance is found to be <0.01, which proves that it has a correlation relationship, and the correlation of Pearson coefficient is -0.143.From this we can see that the higher the frequency of letters in a word, the lower the proportion of difficult patterns.

5 Solution of Problem 3: Central Gravity Model

The idea for this part of the model is inspired by the article "Adaptive Propagation Clustering Algorithm Based on Universal Gravity" by He Wang Zhi[7], which introduces the concept of universal gravity and proposes a "Central Gravity Model" (CGM) to measure the difficulty level of each word. The basic idea of the Affinity Propagation (AP) clustering algorithm is to consider all data points as potential cluster centers, and then construct a network by connecting each pair of data points with edges representing the similarity matrix. Through message passing over the network edges, which includes reliability and validity, the clustering center for each sample can be computed. However, the traditional AP model, which uses the Euclidean distance as the measure of similarity between data points, cannot reflect the global consistency

of sample features. Therefore, in the improved algorithm proposed by Wang Zhi et al., the concept of universal gravity is introduced to perform global searching over samples, measuring the similarity between objects.

Universal gravity reflects the attractive force between objects based on their mass and distance, with objects having greater mass and closer distance exerting stronger attraction. Inspired by this property of universal gravity, the CGM takes into account various word features such as word frequency, letter frequency, and letter matching degree at corresponding positions, to determine the strength of "universal gravity" between different words. In Wordle, when a user inputs a word that is close to the target word of the day, it is more likely that the user will guess the target word quickly. Here, "close" means that the answered word and the target word have similar properties as described above. However, in the game corpus, there are also many words that deviate greatly from the properties of the target word. When users select these words to try, it will lead to more attempts or even inability to complete the game.

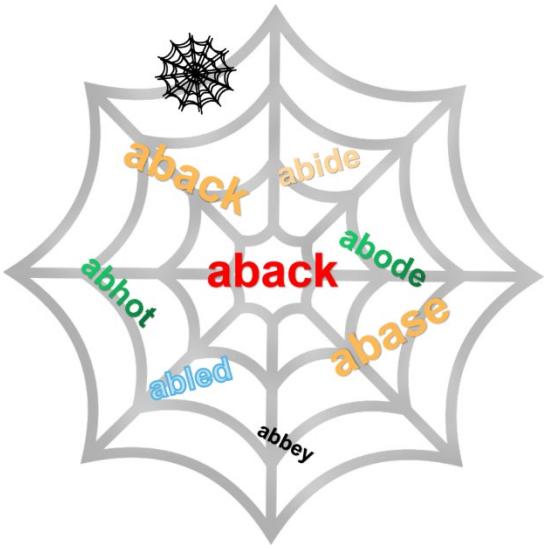

Figure 6: Central Gravity Model

Based on the game background and the requirements of the task, we have developed the following model to measure the difficulty of each word based on its characteristics and interactions with other words, and ultimately categorize the words into different difficulty levels.

$$d_A = \sum_{i=1}^{N} \frac{\log(f_i) \cdot \log(f_A)}{r_{Ai}^2} \tag{3}$$

$$r_{Ai} = 1 - P(l_i) \tag{4}$$

Here, d_A represents the difficulty of the current word A; f_A represents the frequency of word A in daily life, which is obtained by importing the "word_frequency" function from the "wordfreq" library in Python. Frequency f_i represents the frequency situation of all words in the word library except word A, obtained in the same way as above. Taking the logarithm of the frequency is to reduce the impact caused by the uncertainty in frequency statistics. N represents the number of all words in the corpus.

The original meaning of r_{Ai} was the distance between word A and word i. In this model, we abstract distance as the probability of the event l_i occurring, denoted as $P(l)$. The event l_i refers to the probability of the user selecting word A given the same letter combination and arrangement pattern in the word library. The explanation is as follows:

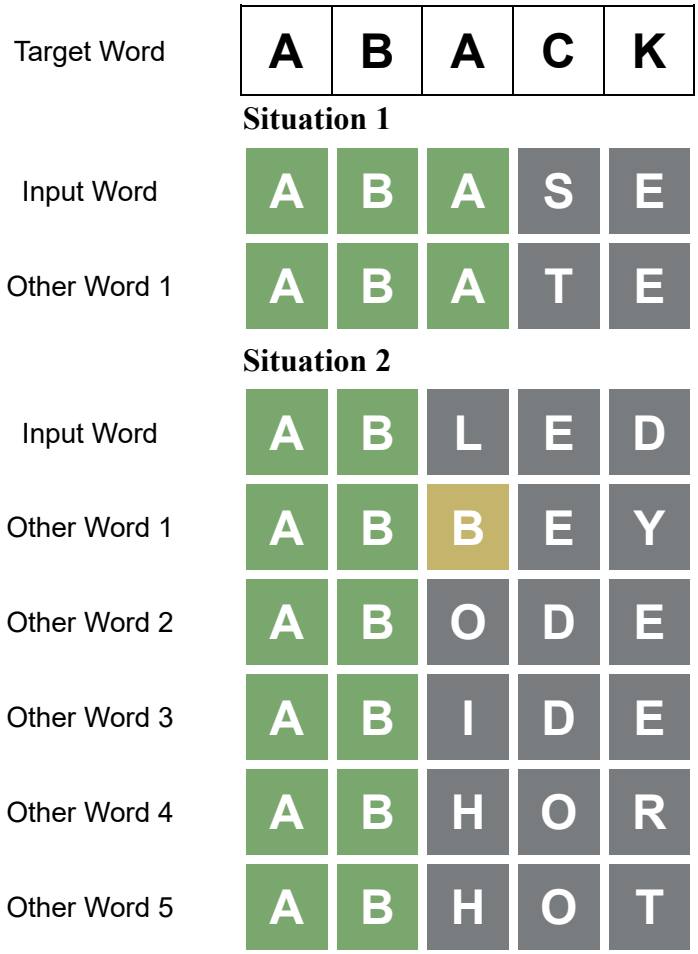

Figure 7: Explanation of Algorithm for calculating probabilities

In the case where the target word A is "aback", in situation one, when the input word i is "abase", there are still two words "aback" and "abate" left in the corpus, but the letter "e" in

"abase" is grayed out, so "abate" can be excluded. Therefore, in this situation, the probability of the event l_i happening is 1. In situation two, when the input word i is "abled", after excluding the letters "l", "e", and "d", there are three words left: "aback", "abhor", and "abhot", so the probability of the event happening, i.e., the user selecting the target word, is $1/3$.

Based on the above explanation, we redefine the meaning of "distance" as follows: after selecting an input word i, this word helps the user to bring the "distance" between the input word i and the target word A closer. Because if the relationship between the input word i and the target word A is closer, then the value of $P(l_i)$ will be larger, which leads to a smaller $1 - P(l_i)$, and the final result is a closer "distance".

5.1 The Solution of Model 3

During the solving process, we calculated the difficulty of each word in the "Problem C Data Wordle.xlsx" file using the above model, and then performed K-means clustering [6] based on their difficulty levels. Finally, we obtained three categories of difficulty levels. The word ERRIE belongs to the most difficult category, with a difficulty level of 80116.0633.

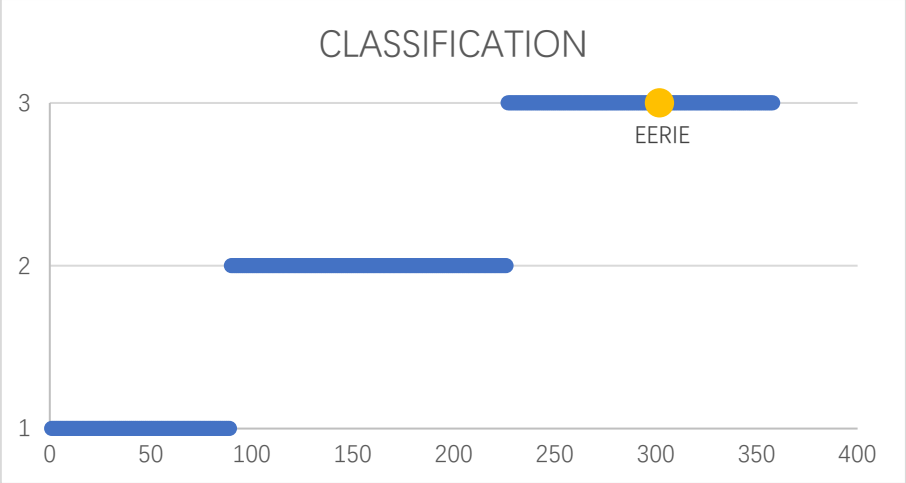

Figure 8: Difficulty classification of words

The position of the red triangle in the above figure indicates the difficulty level of the target word EERIE in the third category. It can be observed that the word is located in the upper-middle part of the third difficulty level, which is due to the repetition of the letter "E" in the word and its relatively low frequency.

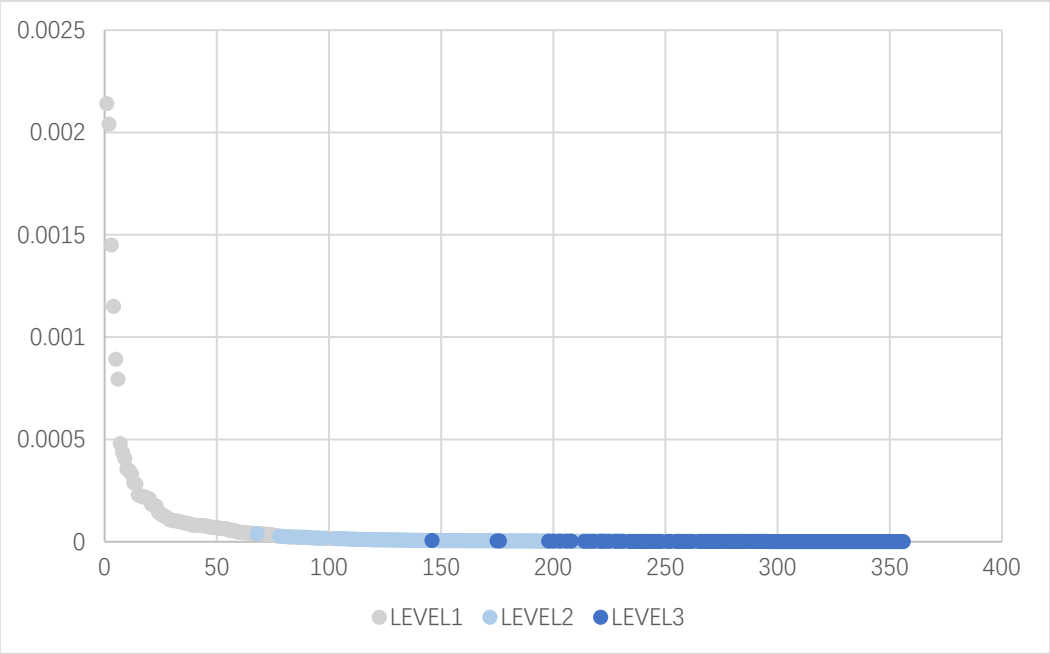

Figure 9: Word frequency expectation

Table 3: Different grades correspond to different attribute expectations

Difficulty Level	Word frequency expectation	Letter frequency expectation	Difficulty interval
First class	1.93E-04	0.3960	[35477,61383)
Second class	7.03E-06	0.3951	[61383,75245)
Third class	1.40E-06	0.3942	[75245,97460)

According to the requirements of Task 3, we identified the word frequency, letter frequency, and difficulty interval characteristics of different difficulty levels in the table above.

It can be observed from the table that as the word difficulty level increases, the expected word frequency and the sum of letter frequencies at each position show a decreasing trend, which is consistent with the fact that people are more likely to answer the target word correctly when they are more familiar with the word or letter.

For the accuracy of this classification model, the following discussions are made:

- First, we need to explain why our model can be used to measure the difficulty of a word d_A.
    1. We use the universal law of gravitation to measure the degree of influence between words, which is scientifically reliable.

2. We incorporate word frequency as the "mass" of the word itself into the formula, and taking the logarithm of the frequency itself has been scientifically validated.
3. In considering the "distance" r_{Ai}, we take into account the frequency (commonality) of letters, and reasonably define distance based on the probability of events occurring. As explained above, the greater the correlation between the input letter and the target letter, the greater the probability of the corresponding event li occurring, leading to a decrease in the "distance" r_{Ai}, i.e., when the input word is closer to the target word, the difficulty of guessing the target word will also decrease accordingly. Therefore, in the final summation process, the difficulty of the input word i for the target word i is increased less, which is consistent with cognitive laws.

- The difficulty level of a word is negatively correlated with its frequency and the commonality of the letters it contains, as shown in **Table 3**, and the classification results are generally accurate.
- It can be seen that the target word EERIE contains many letter E's. According to research, the letter E has the highest frequency among all letters, accounting for up to 10%. However, due to the high repetition of the letter E in the Wordle game, words with high repetition rates like EERIE often require more attempts to succeed, which matches our classification of EERIE as the most difficult level.

6 Solution of Problem 2

We used MATLAB to dynamically simulate the distribution of daily attempts for each word from January 7th, 2022 to December 31st, 2022. From the animation, we found that the peak attempt frequency for the majority of dates was around four attempts, accounting for about 72% of the total attempts. Therefore, we speculate that the distribution of attempts for different words on different dates may exhibit a "quasi-normal" distribution with a peak around four attempts. Deviations in peak occurrence on some dates may be caused by uncertain factors.

Although the distribution of attempts for different words may follow a "quasi-normal"

distribution with a peak at four attempts, the difficulty level of different words may have a certain impact on the distribution. Therefore, dividing the words into different difficulty levels is the key to solving this problem.

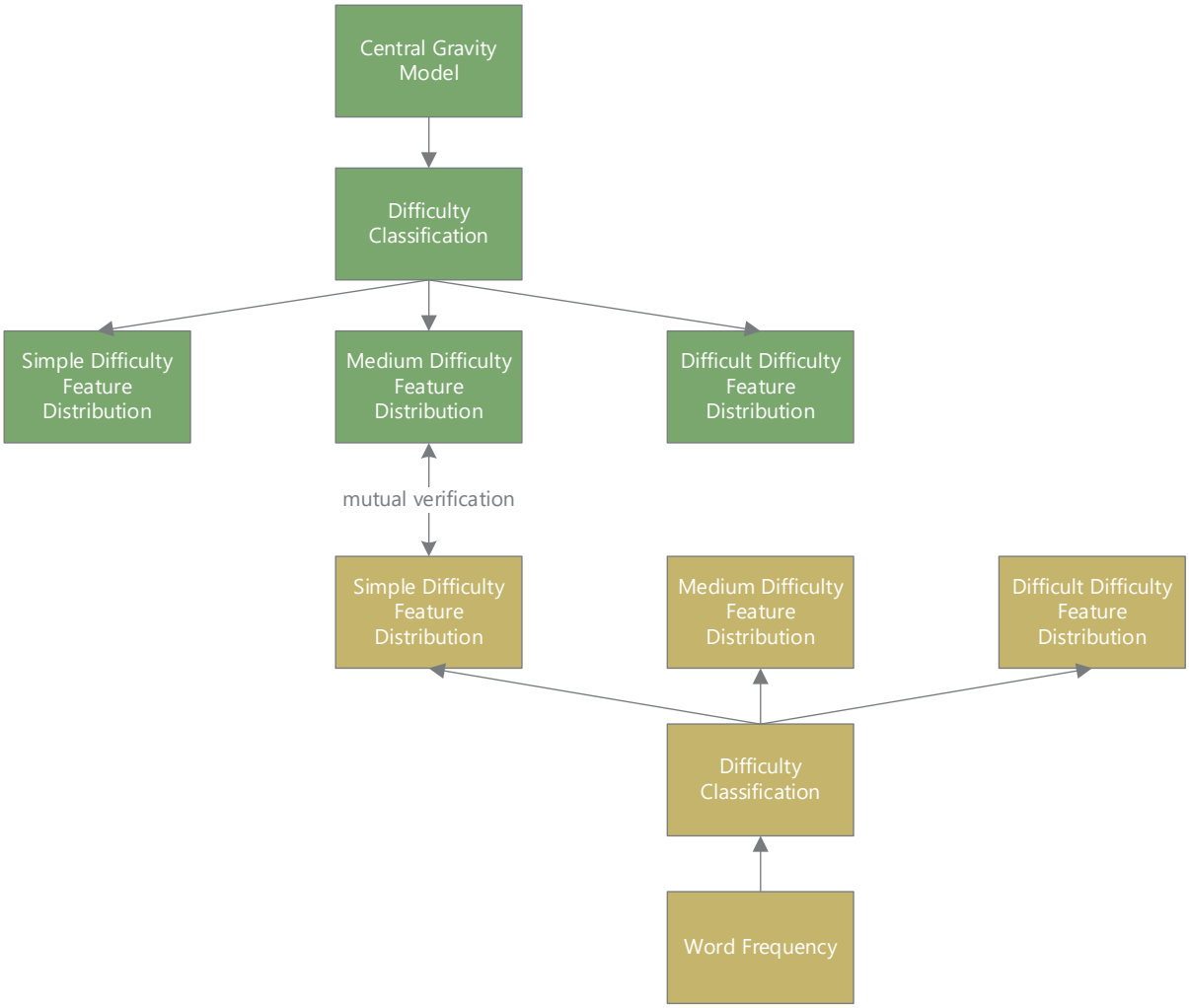

Figure 10: Flow chart of question2

According to the central gravity model, we have obtained a difficulty classification model that divides words into three levels. In order to obtain the distribution of the number of attempts for the target word "EERIE", we need to summarize and extract the distribution of the number of attempts for each difficulty level in the model, and then assign the difficulty of the target word to the corresponding difficulty level. Therefore, the distribution of the number of attempts for the target word can be determined by the characteristic distribution of attempts for the corresponding difficulty level.

Table 4: Distribution Table 1

LEVEL	1	2	3	4	5	6	7 or more tries (X)
1	0.758	7.540	25.770	32.632	21.103	9.7126	2.321
2	0.2517	4.4217	20.061	32.945	25.639	13.258	3.496
3	0.512	6.146	23.560	33.178	23.219	10.967	2.357

We know that the target word is in the third difficulty level, so the distribution of the number of attempts for the word is shown in the last row of the **Table 4**.

To verify the reliability of the above results, we also divided the words into three categories based on word frequency, and extracted the corresponding distribution of attempts for each difficulty level, as shown in the **Table 5**.

Table 5: Distribution Table 2

LEVEL	1	2	3	4	5	6	7 or more tries (X)
1	0.704	7.796	24.713	31.065	21.519	11.037	3.130
2	0.429	5.186	22.311	33.366	24.267	11.696	2.634
3	0.239	4.375	20.773	34.500	25.341	12.136	2.773

The frequency of the target word "EERIE", and found that it belongs to the second difficulty level.

We have obtained the trial distribution of the target word in two ways. We compare the two distributions by plotting the curves, and we can see that the two curves overlap to a high degree, which verifies the reliability of our results.

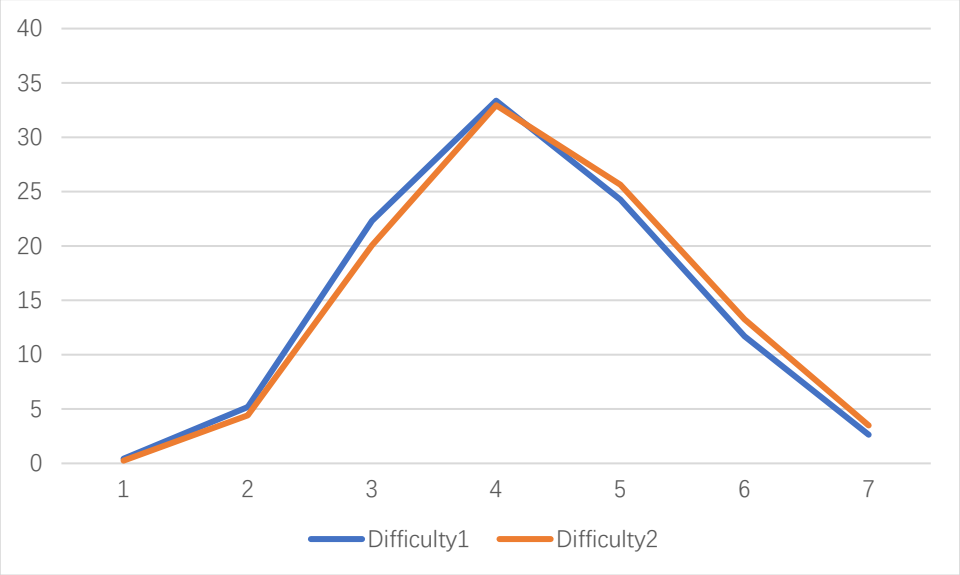

Figure 8: Trials distribution comparison

Finally, we obtain the following distribution for the target word by averaging the results of the two methods:

Table 6: EERIE final trial distribution

Attempts	1	2	3	4	5	6	7 or more tries (X)
EERIE	0.340	4.804	21.186	33.156	24.953	12.477	0.340

Although we have predicted the distribution of the number of attempts for the target word, there are still many uncertainties in our observations. For example, the number of participants in the game on a given day can greatly affect the stability of the distribution, and only when there are enough participants can we maximize the possibility of reproducing the distribution of the number of attempts. In addition, if a special holiday or an unexpected event occurs on that day, which directly affects the mindset of the participants, it will greatly affect the final answering situation.

7 Solution of Problem 4

Interest 1: Some simple words may also require people to try several times before answering correctly.

The average expected number of attempts for all words in the Excel can be calculated as 4.195 times. The table below lists the words with an average number of attempts above 5 and their frequency.

Table 7: Word feature ranking

Word	Tries Average	Word Frequency
Paper	5.99	3.55E-04
Nymph	5.84	7.41E-07
Mummy	5.53	3.72E-06
Coyly	5.35	1.20E-07
Foyer	5.25	0.00000123
Judge	5.18	8.91E-05
Gawky	5.08	8.13E-08
Swill	5.08	2.09E-07

It can be observed that some common words, such as "paper" and "judge", require more attempts for players to answer correctly despite their frequent occurrence in daily life.

Interest 2: The scoring situation has a relatively small expected impact on the average number of attempts.

According to the above figure, it can be found that the average number of attempts per day by players is not greatly affected by the changing of the report score. In Model 1, we also linked the report score with the number of daily participants in the game. Therefore, we can further conclude that the number of participants in the game has a relatively small expected impact on the average number of attempts. This expectation can also be understood as the overall level of the players participating in the game. Thus, the conclusion can be further generalized as the number of participants in the game has little impact on the overall level of players participating in the game.

8 Model Evaluation and Further Discussion

8.1 Strengths

- The model takes into account the communication nature of the game and reasonably describes the changing trend of the number of Wordle players over time.
- The use of differential equations makes the model more accurate in representing the actual situation.
- The model considers the situation where players' interest in the game gradually wanes and provides a reasonable explanation for it.
- The Central Gravity Model incorporates the idea of the physical gravity model, which is simple, easy to understand, and considered a comprehensive range of factors.

8.2 Weaknesses

- The model uses the Sigmoid function to describe trends in the rate at which players quit the game, but in reality, this can be more complex and may not fully capture all factors affecting players' decision to quit playing the game.
- The model's parameters and indicator values are based on assumptions and guesswork,

without actual data validation, thus it has some subjectivity.

- In Central Gravity Model, the calculation method for the interaction between words in the model is relatively simple, which may result in certain inaccuracies.

9 Conclusion

This paper analyzes in detail the relationship between time, word attributes, the number of survey results and the number of attempts. Before introducing our model, we have done a lot of data preprocessing work. We then analyzed trends over time and data and came to the conclusion that the number of players was changing.

Next, we built a game viral transmission model, fitted the trend of the number of survey results over time, and determined that the change of survey results was caused by the change of the relative number of different groups of players. Then, through the correlation analysis of a variety of word attributes and the proportion of difficult patterns, we finally determine that there is a correlation between letter frequency and the proportion of difficult patterns.

In the second question, we weighted the difficulty of words by classification, analyzed them based on word frequency and word difficulty score, and then calculated the corresponding distribution by classification. The correctness analysis is carried out by taking part of the data as sample and the other part as reference. For the third question, we use the method of central gravity model to quantify the scores of different words, and then through cluster analysis and classification, finally get the specific classification of words.

10 References

[1] Wang Binggang, Qu Bo, Guo Haiqiang et al. Research on mathematical model of infectious disease prediction [J]. China Health Statistics,2007(05):536-540.
[2] Pan Difu, Liu Hui, Li Yanfei. Wind farm Wind speed prediction Optimization Model based on Time Series analysis and Kalman filter algorithm [J]. Power Grid Technology,2008,No.276(07):82-86.
[3] Huang Yi, Duan Xiusheng, Sun Shiyu et al. Research on Deep Neural Network Training Algorithm Based on Improved sigmoid Activation Function [J]. Computer measurement and control, 2017, 25 (02) : 126-129. The DOI: 10.16526 / j.carol carroll nki. 11-4762 / tp. 2017.02.035.
[4] Zhang Lei, Sun Changqing. High Order Residual Modified GM(1,1) interval prediction model and its application [J]. Journal of Ordnance Equipment Engineering,2017,38(02):177-181.
[5] Wang Zhihe, Chang Xiaoqing, Du Hui. Adaptive Nearest neighbor Propagation clustering algorithm based on Gravitation [J]. Journal of Computer Applications,2021,41(05):1337-1342.
[6] Sun Jigui, Liu Jie, Zhao Lianyu. Research on clustering Algorithm [J]. Journal of Software,2008(01):48-61. (in Chinese)
[7] Wang Zhihe, Chang Xiaoqing, Du Hui. Adaptive Neighbor Propagation Clustering Algorithm Based on Gravity [J]. Computer Applications, 2021, 41(05): 1337-1342.

Dear Puzzle Editor,

I hope this letter finds you well. I am writing to share with you the latest findings from our research on the word puzzle game that you are interested in.

We are a group of young people who love word guessing games. When we first heard about Wordle, we were immediately intrigued and had infinite ideas about how to guess the daily word faster. We have thought deeply about various questions, such as whether there is any impact between words and how many attempts it takes to guess some easy words. Now, we would like to respond to your letter.

First, let us answer your question about the changes in the score curve. We understand that such changes can have a huge impact on the game's future development. We believe that the essence of the game can be understood as a virus. When a game is first launched, almost everyone is part of the target audience, and the "virus" will spread rapidly among the population. However, over time, if a game does not have enough attractions to retain players, the population will gradually decrease at a certain rate, resulting in the effect shown in the title. Therefore, we believe that changing the rules of the guessing game can make the game shine again.

In the process of studying the scoring curve, we also found an interesting point. As time goes by, the ratio of scores between the difficult mode and the simple mode almost stabilizes at 1:9. In reality, we expect that the average daily active participation in the game will stabilize at 21224 after March 1st, 2023. This indicates that after that time, the players who participate in the game will be stable, and can be regarded as old players. However, these old players almost play every day and are familiar with the rules and strategies of the game, but they cannot improve their score ratio in difficult mode, or there are no more old players to try the difficult mode, which is strange.

Above is our answer to your doubts. Next, we will tell you about our findings. We are very curious about how to measure the difficulty of a word on a certain day, and why most people cannot answer it or need more attempts. So we sat under an apple tree and thought about it. Suddenly, an apple fell from the tree, and we suddenly realized that if the earth can attract the apple to fall, then we can guess that the word can also be used to attract the answer word

each day. Inspired by the law of universal gravitation, we established a central gravity model to measure the mutual influence between these words based on their frequency, letter frequency, and other factors, and then consider the difficulty of the word that needs to be answered on that day. After extensive calculations and verification, we classified the words in the corpus into three levels of difficulty, and the target word "EERIE" was classified into the most difficult category of words.

After determining the word difficulty, what attracted us even more was how many attempts an average person needs to answer a word correctly each day. To answer this question, we carefully observed the distribution of attempt numbers over time and found that although the peak values fluctuated left and right, the overall distribution remained stable at around four attempts. Using feature matching, we first divided the words into different levels and then extracted the feature distribution of different levels. Finally, we matched the target word to its corresponding level and obtained an approximate distribution of the number of attempts for the word.

After several days of research, we have gained a deeper understanding of this word puzzle game, and we are grateful for the opportunity to share our findings with you.

Sincerely
MCM