Problem Chosen
C

2023
MCM/ICM
Summary Sheet

Team Control Number
2312165

Working on Wordle: Building a Bridge between Correlation and Causation

Summary

Wordle, a puzzle offered daily by the New York Times, has gained immense popularity. The game challenges players to guess a five-letter word in six attempts or fewer and provides feedback after each guess. In this version, every guess must be a valid English word, as unrecognized guesses are not permitted. In this paper, we shall find the inner correlation between the mechanics and report data.

Firstly, we attempt to **predict the number of reports on March 1, 2023**. After analyzing the original data, we **throw out the wrong data points** and find that the tendency of the data appears to fit an exponential function. Thus, we apply several exponential functions, and choose the best model, **Langevin Model**, with $R^2 = 0.9889$. In addition, we apply **Grey Model** to make predictions from the data of last 60 days, which provides similar results as the Langevin Model. We also tried the **ARIMA Model**, however, the confidence in the prediction is poor as the data does not satisfy the the requirement of a smooth time series. In a nutshell, we obtain a precise result with a short confidence interval via the first two models.

Secondly, ANN(artificial neural network) is used to find the association between the lexical information and the distribution of the number of tries. We train two FNN(feedforward neural network) models (the difference between two FNN models is whether or not time information is taken into consideration) and one LSTM(Long Short-Term Memory) Model, which generates **analogous predictions of "EERIE" on March 1, 2023 with low uncertainty**. Furthermore, we find that the data set provided is insufficient to train a large model that is capable enough, and temporal information won't help us to gain a superior model.

Thirdly, to **identify the difficulty of words**, we establish two models using linear and non-linear methods respectively. The former employs **principal component analysis(PCA)** , while the latter utilizes approaches in **survival analysis(SA)**. By considering the first principal component in PCA and choosing a proper statistic in SA, we obtain two quantitative estimators of difficulty , and the ranks of words they provide have a correlation coefficient of 0.99. As for classification, we perform **K-means clustering** using the first two principal components in PCA and the previous statistic in SA to classify words into three levels of difficulty with stability checked. More than 98% percent of words are categorized into the same groups by the two models, and the word "EERIE" is identified as "hard" using results from the previously-constructed ANN models.

Finally, we combine **correlation analysis** and **hypothesis testing** to determine the causal factors that affect the difficulty for solving a specific puzzle. By conducting **paired-sample t-test**, we identify the number of non-repeating letters as an important attribute for classification with $P < 0.01$. Moreover, by exploring further into the occurring frequencies of letters, we discover the underlying structures of words as well as the correlation between letters and the difficulty levels of words.

Keywords: Grey Prediction, ARIMA, FNN, LSTM, PCA, Survival Analysis

Contents

1 Introduction

1.1 Background

Wordle, a delicate game online, has brought up the population of word-filling tricks. Millions of people log in this website to guess daily word. In this game, players had six chances to guess a five-letter word, and each guess gave feedback in the form of a colored card indicating that whether the letters matched or occupied the correct position. The mechanics are almost identical to those of the 1955 pen-and-paper game Jotto and the television game show Lingo. wordle has a single daily solution in which all players attempt to guess the same word.

Now, we are wondering about the correlation between the word itself, the percentage of each attempt and the number of hits on this website. In this article, we will reveal how to assess the difficulty of a word and its impact on people's choices.

1.2 Our work

a) Predict the number of report results

In this problem, we attempt to predict and account for the variation by three different models. At last, we evaluate these models to find the best answer.

- Exponential Function Fitting
- Grey Prediction Model(GM)
- ARIMA Model

We start from the practical meaning of the model and the range of model data selection, combine the model fitting effect, and finally find the optimal model.

b) Determine factors affecting the percentage of hard mode

c) Predict the distribution of attempt times

To achieve this goal, we develop 3 models by pytorch to bridge the information of words and $\{\boldsymbol{p_k}\}(k = 1, 2, \cdots, K)$. After the pre-trained model is obtained, we put the word(and time) into the model to get the prediction of $\boldsymbol{p_k}$. In particular, we get the $\{\boldsymbol{p_k}\}$ corresponding to "EERIE" on March 1, 2023.

- Two FNN Models
- LSTM Model

d) Find Effective Ways to Evaluate Difficulty

we establish two different methods to set up a general criteria to identify the difficulty of words and make comparisons between them:

- Principal component analysis (PCA) + 2-D K-means clustering
- Survival analysis (SA) with adjustments + 1-D K-means clustering

Base on the two models, we evaluate the difficulty of the word ”EERIE”, and employ paired sample t-Test to search for attributes that make a word harder to guess.

e) List interesting points in the data

We conduct correlation analysis to discover the relationship between other factors (e.g. frequencies of letters) and the difficulty levels of words.

1.3 Notation

Symbol	Model	Meaning	Equation
K	(Global)	number of words considered	original data
$\boldsymbol{p_k}$	(Global)	the distribution of the number of tries	original data
$x_k^{(1)}$	Grey Prediction	Original data list	original data
$x_k^{(2)}$	Grey Prediction	Cumulative sum data list	$\mathrm{x}_\mathrm{k}^{(2)} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{x}_\mathrm{i}^{(1)}$
$\mathrm{z_k}$	Grey Prediction	proximity generation series of $\mathrm{x}_\mathrm{k}^{(2)}$	$\mathrm{z_k} = \alpha \mathrm{x}_\mathrm{k}^{(2)} + (1 - \alpha)\mathrm{x}^{(2)}$
$\boldsymbol{p_{kI}}$	Model I(FNN)	$\boldsymbol{p_k}$ predicted by Model I	prediction results
$\boldsymbol{p_{kII}}$	Model II(FNN)	$\boldsymbol{p_k}$ predicted by Model II	prediction results
$\boldsymbol{p_{kIII}}$	LSTM Model	$\boldsymbol{p_k}$ predicted by LSTM Model	prediction results
p	ARIMA	number of Auto Regressive terms	parameter
q	ARIMA	number of sliding average terms	parameter
d	ARIMA	times of differences	parameter
$\boldsymbol{w_k}$	Difficulty Evaluation	weights assigned to p_k	parameter
z_k	Difficulty Evaluation	the estimator of difficulty in PCA	$z_k = \boldsymbol{p_k} \cdot \boldsymbol{w_k}$
h_k	Difficulty Evaluation	the estimator of difficulty in SA	Equation 1

2 Data Cleaning

Contest Number	Column	Problem	Operation
540	word	misspelling naive ⇒ naïve	correct the word
545	word	misspelling probe ⇒ rprobe	correct the word
529	number of report	the number is $\frac{1}{10}$ of the normal data	delete wrong data
281	percent in	the sum of the seven parts is 126 %	delete wrong data
525	word	misspelling clean ⇒ clen	correct the word

3 Prediction of Number of Reports

3.1 Original data

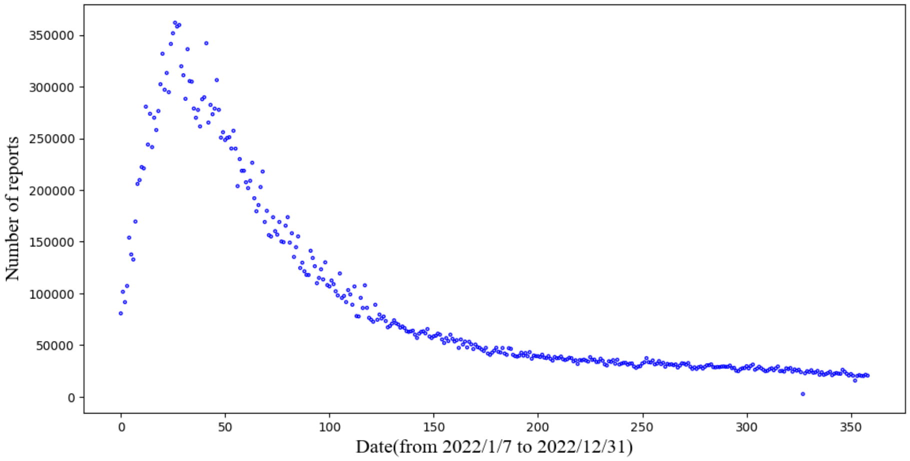


Figure 1: Scatter of Number of Reports

Analyzing the trends in the raw data, we can see a dramatic rise in numbers occurred when the game was first launched, with a peak around day 30.

For such trend, we can initially explain that the game is similar to the growth model of contagion or logical Stiffness at the beginning, and after reaching a certain level of heat, it declines with a trend similar to the exponential function.On the other hand, we find several wrong data in the dataset.(e.g. the report number around day 325 is apparently low)

Following are models we built after eliminating the wrong data.

3.2 Model I : Exponential Function Fitting

The data presented exponential growth and exponential decline in trend, so we fitted the preliminary model using an exponential function regression[1] before and after the peak. So we consider to use common exponential functions and their combinations.

After a series of fitting, we get the optimum model:

	Equation	R^2
Before peak	$y = a - b \cdot c^x$	0.9672
After peak	$y = y_0 + C\left(\coth\left(\frac{x-x_c}{s}\right) - \frac{s}{x-x_c}\right)$	0.9889

	a or x_c	b or s	c or C	y_0
Before peak	4.89×10^5	4.34×10^5	0.9599	
After peak	50.36	-22.21	2.51×10^5	2.54×10^5

Table 1: Exponential Model

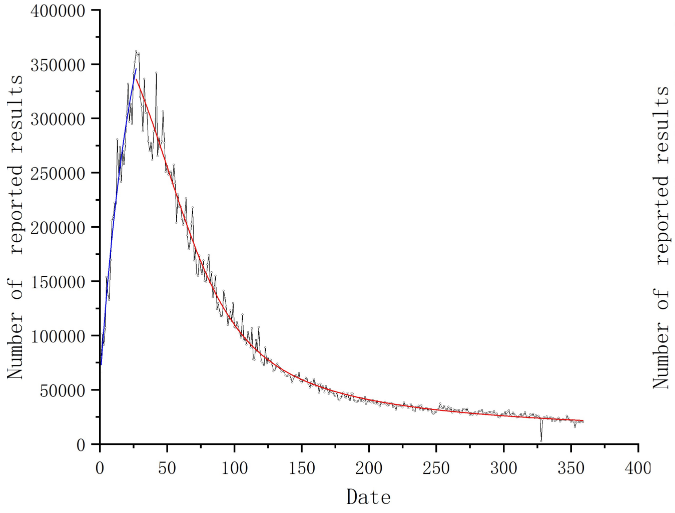


Figure 2: Fitting Curve

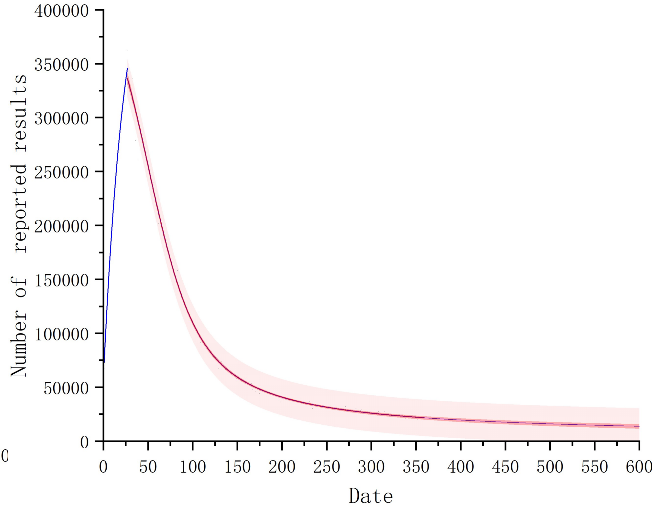


Figure 3: Prediction Curve

Figure 2 shows the fitting consequence of this model. Except for the wrong data(around day 325), the data fits our model well. Figure 3 exhibits the prediction curve and dark red part represents a 95 percent confidence band, light red represents a prediction band.

According to our first model, number of reported results on March 1, 2023 will be **18676**, and the 95 percent confidence interval will be **[16978,20487]**

3.3 Model II : Grey Prediction Model

Grey Model[2] is a model in the form of differential equations using discrete random numbers that have been generated into random numbers with significantly weakened and more regular generations. This facilitates the study and description of the change process. In this section we will use Grey Model to analyze and predict the number of report.

3.3.1 Principle

$x_k^{(1)}$ can be written in this way:

$$\mathrm{x}_\mathrm{k}^{(1)} = \mathrm{x}_\mathrm{k}^{(2)} - \mathrm{x}_\mathrm{k-1}^{(2)}$$

And z_k is defined by following equation, it is proximity generation series of $x_k^{(2)}$:

$$\mathrm{z}_\mathrm{k} = \alpha \mathrm{x}_\mathrm{k}^{(2)} + (1 - \alpha)\mathrm{x}_k^{(2)}$$

And an exponential-like distribution of data should satisfy the following relationship:

$$\mathrm{x}^{(1)} + a \cdot \mathrm{z}_\mathrm{k} = b$$

Where a and b are parameters remains to be determined.

We can estimate a and b by the least squares method.

$$\begin{bmatrix} a \\ b \end{bmatrix} = \left(B^TB\right)^{-1} B^TY, \quad Y = \begin{bmatrix} x_2^{(1)} \\ x_3^{(1)} \\ \vdots \\ x_n^{(1)} \end{bmatrix}, \quad B = \begin{bmatrix} -z_2^{(1)} & 1 \\ -z_3^{(1)} & 1 \\ \vdots & \vdots \\ -z_n^{(1)} & 1 \end{bmatrix}$$

In terms of the data which footstep equals to 1, $x_k^{(1)}$ can stand for the derivative of $x_k^{(2)}$, and this equation can be transformed to an ODE:

$$\frac{dx_t^{(2)}}{dt} + ax_t^{(2)} = b$$

Solve this ODE and we can construct the prediction model by parameter a and b:

$$\widehat{x}_{k+1}^{(2)} = \left(x_1^{(1)} - \frac{b}{a}\right) e^{-ak} + \frac{b}{a}$$

$$\widehat{x}_{k+1}^{(1)} = \widehat{x}_{k+1}^{(2)} - \widehat{x}_k^{(2)}$$

3.3.2 Prediction

Grey Model is more suitable for short- to medium-term forecasts, so we chose the latter 160 days of data.***We have already thrown the wrong data point.***

First, we have to test the suitability of the data. We count the proportion of adjacent data and find that this proportion varies relatively smoothly in our selected data, which indicates that this data set is suitable for the Grey Model.

Second, we use chosen data to predict the number of reported results on March 1, 2023. Get the prediction of following 60 days and put them together with chosen data in following graph.

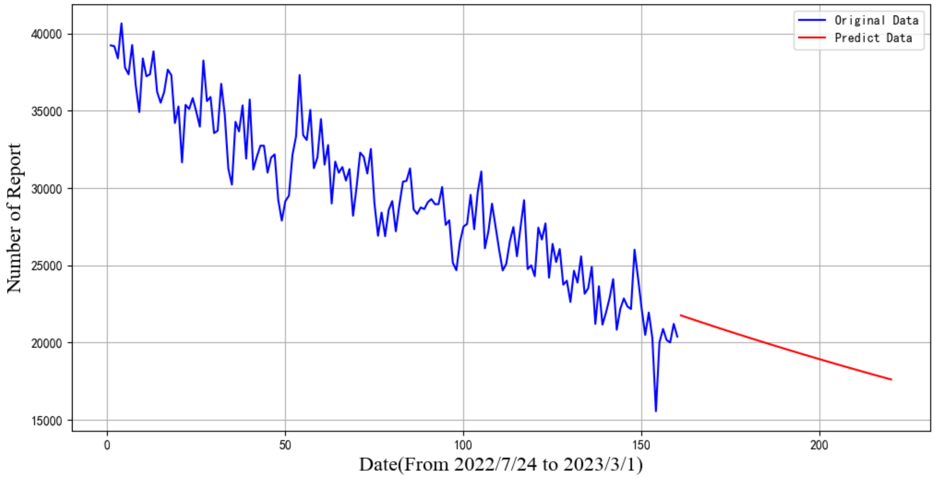


Figure 4: Prediction of Grey Model

According to our second model, number of reported results on March 1, 2023 will be **17608**.

3.4 Model III : ARIMA Model

The third model we tried Auto Regressive Integrated Moving Average model[3]. The following equation tell us the common progress of ARIMA. p is the number of Auto Regressive terms, q is the number of sliding average terms, and d is the number of differences made to make it a smooth series. By using the AIC principlewe choose (p,d,q) to be (1,1,0).

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

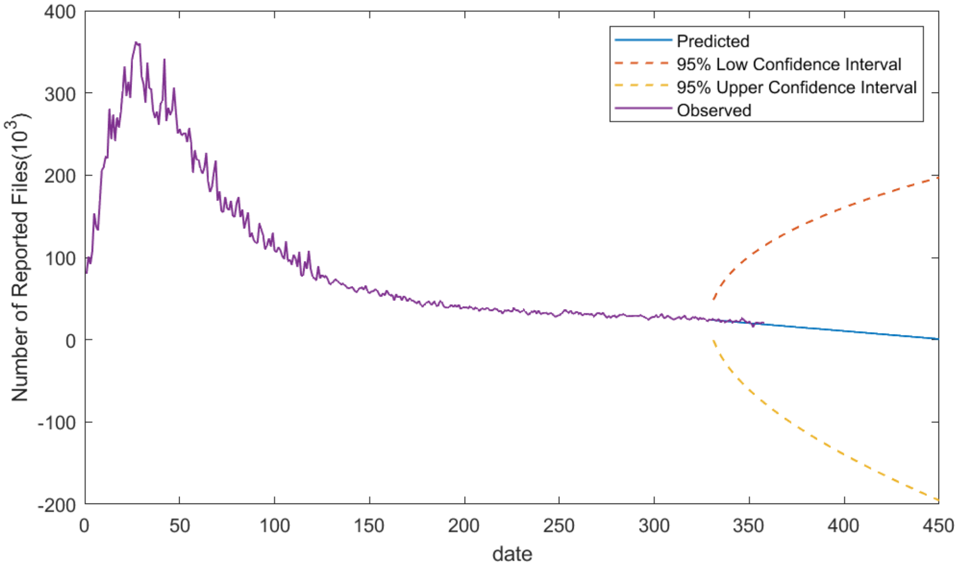


Figure 5: Prediction of ARIMA(1,1,0) Model

According to our third model, number of reported results on March 1, 2023 will be **6685**, and the 95 percent confidence interval will be **[-163447,176817]**.

It is apparently that our third model failed. We consider that this is because the data do not satisfy the requirement of a smooth time series, no matter how many differential has been made.

3.5 Model Evaluation

In the Model I , we divide data into two parts by the peak and apply exponential functions to fit the report data, while in Model II, we use Grey Prediction Model and choose the data of the last 60 days to predict future report data. Our first and second model give the similar results in a short confidence interval. However the third model failed to give a precise result. Because the data set fits the exponential function better.

The first two models predict that the number of reported results on March 1, 2023 will be **17608** or **18676**, with the 95 percent confidence interval of **[16978,20487]**. The relative error is around 9 %, which means the first two models provide satisfying prediction for the required date.

4 Applying Machine Learning to Predict

In this section, we use pytorch to develop 3 models. For the first one, temporal information is not considered. We directly send the information of correct answers into the FNN[4], whose target is the $\boldsymbol{p_k}$ of that day. For the second one, we use migration learning approach to integrate lexical features and temporal information. For the third model, We use LSTM to give a prediction, combing a similar data set made up of five letter words.

We compare the performance of these three models, leading to a conclusion that all of them can prediction with low uncertainty and $\boldsymbol{p_k}$ is not related to the date, and we obtain their predictions of ”EERIE” on March 1, 2023, which are quite similar.

4.1 Two FNN Models

The former, after converting the words into unique hot codes (5*26), we use a FNN model to find the connection between the word information and $\boldsymbol{p_k}$. The latter, we feed the word information into the pre-trained network mentioned above and intercept the middle layer information, then we merge these feature vectors about words with temporal information (converting date to number of days corresponding to the word) as input to train another FNN.

Both models perform well on both the training and test sets (8:2) and the MSE loss, which indicates the uncertainty, on test data are around 0.025 & 0.003, giving us great confidence in our predictions.

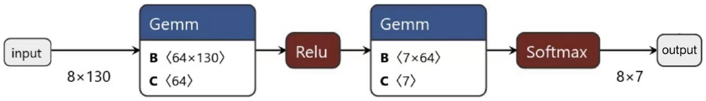


Figure 6: Model I: without temporal information

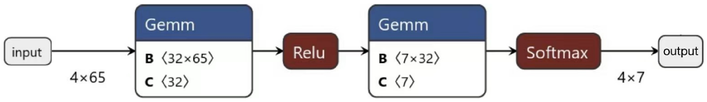


Figure 7: Model II: combining information of words and time

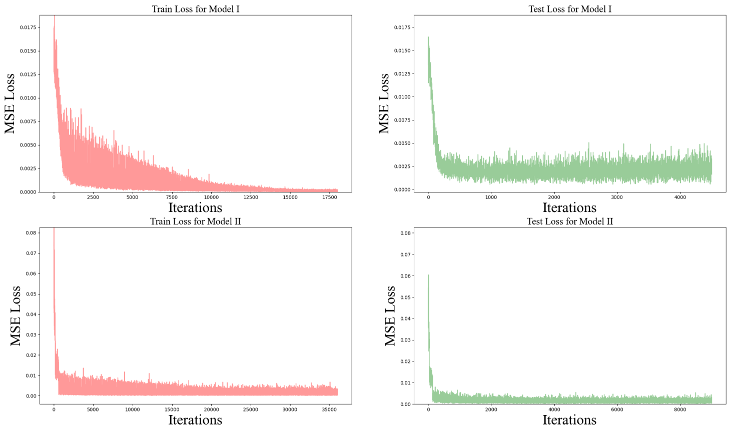


Figure 8: Performance of Model I & II

We put $\{\boldsymbol{p}_k\}$ into the two pre-trained models to obtain $\boldsymbol{p}_{kI}$ & $\boldsymbol{p}_{kII}$. We find high agreement between the two predicted distributions and the actual data, and the consistency between the two sets of predicted data, which increases our confidence in these models.

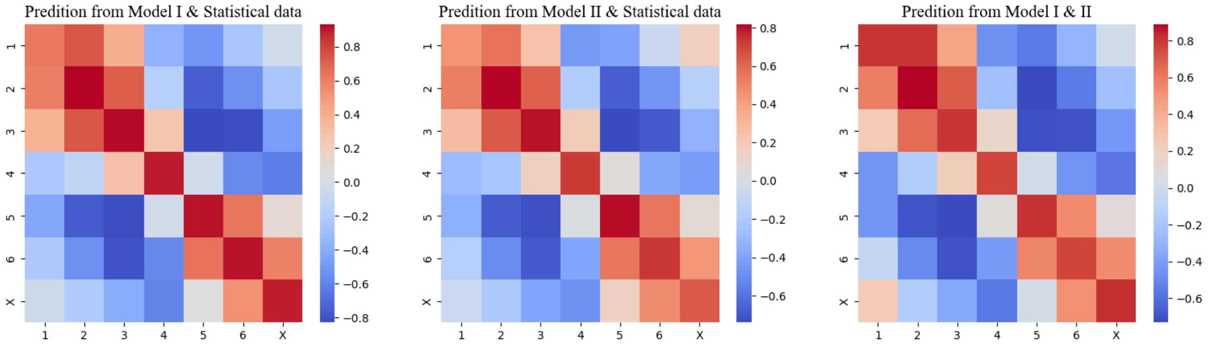


Figure 9: Correlation Coefficient Matrix of the Prediction from Two models & $\{\boldsymbol{p}_k\}$, the Agreement between $\{\boldsymbol{p}_{kI}\}$and $\{\boldsymbol{p}_{kII}\}$

However, as the data set provided is too small for the network to get sufficient information, the R^2 score between their prediction and the original data are 0.756 and 0.508, which are not so satisfying, though the MSE on test data set is good enough.

However, the R^2 value corresponding to the model without integrating temporal information is significantly higher than that of the model considering temporal information, which suggests that the mapping of words to the distribution of trial times may not require temporal information. In other words, people's performance in the game is only determined by the nature of the words themselves, not the specific date and the processing of additional temporal information degrades the performance of the model.

As for distribution of numbers of tries corresponding to ”EERIE” on March 1, 2023, the predictions of two models above is similar, which demonstrates the validity of the two models.

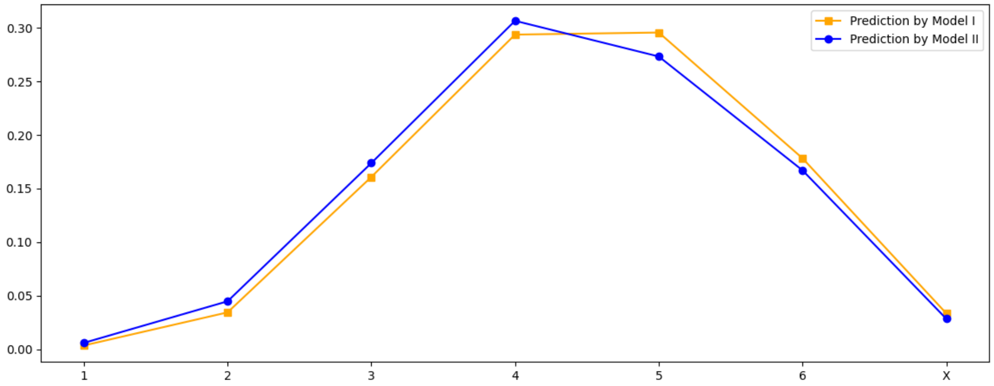


Figure 10: The Prediction of the Distribution of Tries Associated with ”EERIE” on March 1, 2023

4.2 Applying LSTM Model to Predict

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) architecture which has gained significant popularity in the field of deep learning.[5]

LSTMs have memory cells that can store information over time and selectively forget or retain information based on the input, which allows LSTMs to model complex sequential data with varying time intervals between important events, so it can effectively capture long-term dependencies in the data, which is assumed to be suitable to find the law of data fluctuation.

For this attempt, we treat the data as a sequence that fluctuates over time, trying to use LSTMs to capture correlations between the data over time. The model performs well on test data set.

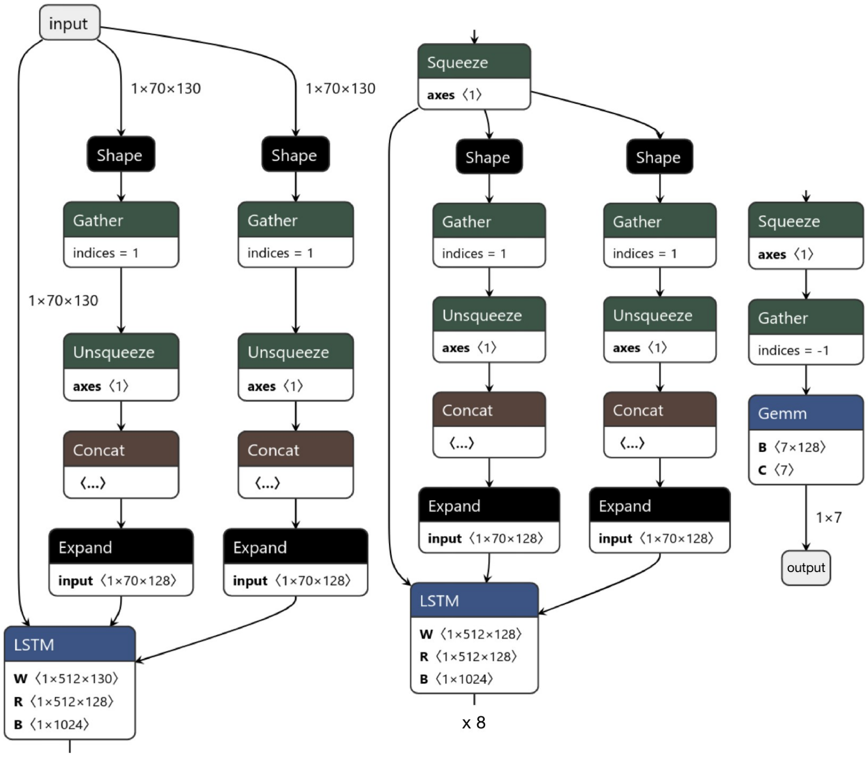


Figure 11: The Structure of LSTM Model We Use

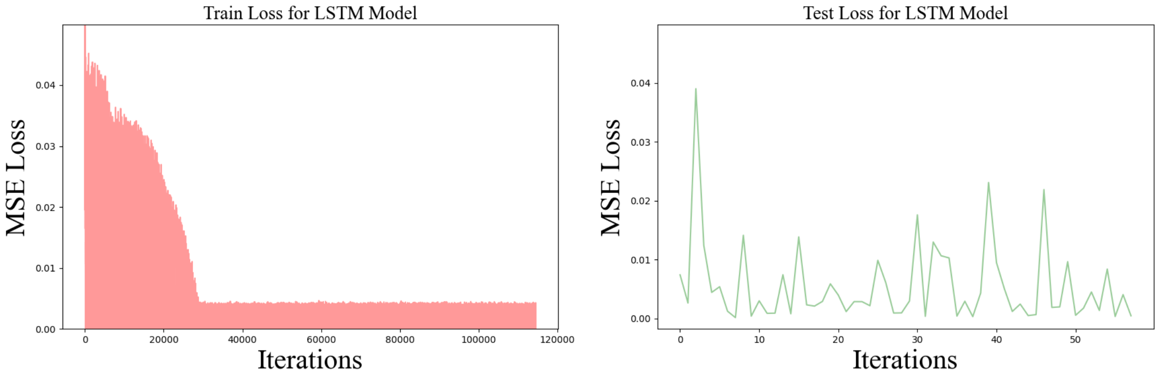


Figure 12: The Train & Test Loss of LSTM Model

	1	2	3	4	5	6	X
LSTM Model	0.021	0.085	0.168	0.256	0.253	0.170	0.046
Model I	0.003	0.034	0.160	0.293	0.295	0.178	0.033
Model II	0.006	0.045	0.174	0.306	0.273	0.167	0.028

Table 2: The Prediction of ”EERIE” from 3 models

We use extra words made of 5 letters to produce a data set to make LSTM have further prediction, setting ’EERIE’ as the input on March 1, 2023. As mentioned above, $\{\boldsymbol{p_k}\}$ is not related to temporal information, so the prediction on ”EERIE” by LSTM Model should not have significant difference from others and actually the results satisfy our expectation. Therefore, the uncertainty of the prediction of ”EERIE” is low.

5 Evaluating the Difficulty Level of Words

In this section, we develop two different models (PCA[6] and survival analysis) to set up a general criteria to evaluate the difficulty of each word based on the distribution of tries. The results of the two models are in great consistency, and show no relevance with respect to other factors (e.g. number of reported results, number in hard mode).

5.1 Model I : Principal Component Analysis (PCA)

From an intuitive perspective, it is natural to regard words that require more tries as harder ones. However, as there are considerable numbers of people who could not solve the puzzle, it is inappropriate to determine the difficulty of each words simply by calculating the average number of tries. We thereby perform correlation analysis and use PCA to extract the most effective features that discriminate the distribution vectors $\{\boldsymbol{p_k}\}$, $\boldsymbol{p} = [p_1, p_2, \cdots, p_7], p_i \in [0, 1], \forall i \in \{1, 2, \cdots, 7\}$), which in turn reflect the difficulty of the words.[7]

5.1.1 Correlation Analysis

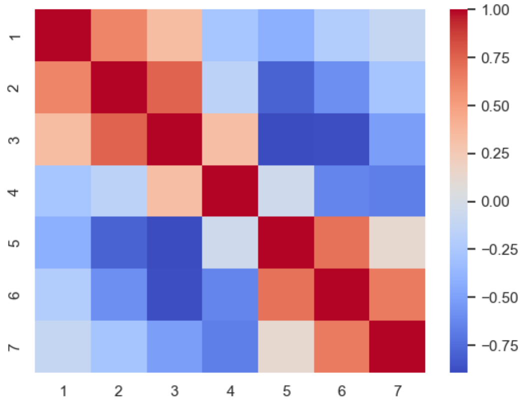


Figure 13: Correlation Matrix of the Percentages of Tries

We calculate the correlation matrix of the percentages of tries ($\{p_i\}$). From the figure **??** we see strong correlation between neighbouring elements p_i and p_{i+1}, despite that p_4 shows little correlation with p_3 and p_5. This result has a strong implication: the threshold attempt times for discriminating the difficulty of a word might be 4, which means that if a man solves a puzzle in 3 or less guesses, he will probably consider it easy ; and if he solves it in 5 or more guesses, it is more likely that he regard this puzzle as a hard one.

5.1.2 Conducting PCA

The fact that the percentages p_i are correlated ensures the rationality of applying dimensional reduction methods. As the sum of percentages approximately equals to 100, the seven-dimensional distribution vectors $\boldsymbol{p_k}$ actually lie in the 6-D hyperplane $p_1 + p_2 \cdots p_7 = 1$. In order to extract features with a better efficiency, we therefore choose to calculate the eigenvectors and eigenvalues of the correlation coefficient matrix of $p_i, i = 2, 3, \cdots, 7$, and the cumulative contribution of variance is shown in figure **??**.

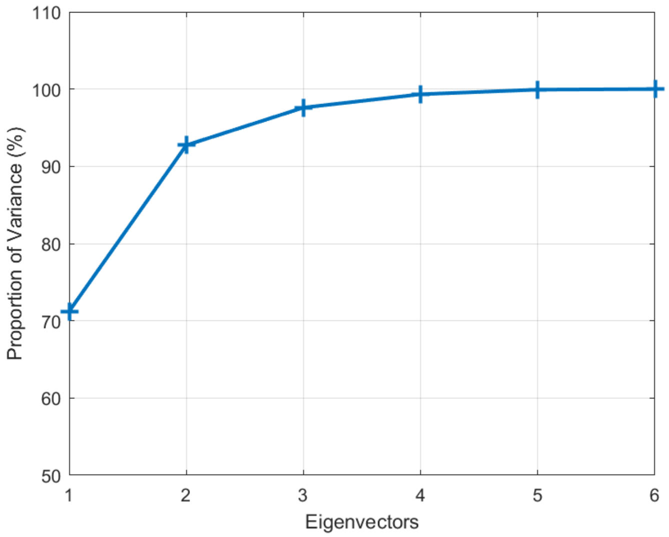


Figure 14: Cumulative Contribution of Variance

We see that the first component, $\boldsymbol{e_1} = [-0.26, -0.65, -0.20, 0.43, 0.49, 0.20]$, is enough to explain up to 72% variance in the data. Thus we can now project our data onto a 1-D plane by calculating $r_k = \boldsymbol{p_k} \cdot \boldsymbol{w}$,in which $w_1 = 0, [w_2, w_3, \cdots, w_7] = \boldsymbol{e_1}$, there by reduce the dimension of data with a minor loss in information.

To check whether r_k provides an appropriate estimation for difficulty, we need to discover the practical meaning of $\boldsymbol{w}$. w_i reflects the contribution of component p_i to the difficulty of a word. The result $w_5, w_6, w_7 > 0$ and $w_2, w_3, w_4 < 0$ indicate that the probability of solving a puzzle in more than 4 times makes positive contributions to difficulty, while solving it with less tries(≤ 4) has the opposite effect. Moreover, the coefficient of variation of r_k is 2×10^3, which indicates that r_k can effectively discriminate between different distributions of attempt times. Therefore, we consider $z_k = \boldsymbol{p_k} \cdot \boldsymbol{w}$ as an effective estimator of difficulty.

5.1.3 K-means Clustering

First, we roll k points to be the initial set of means: $m_1, \cdots, m_k$. Then, we distribute each point to the initial means by following equation, and we define the s_i as the distance between point i and its distribution mean point.

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \forall j, 1 \leq j \leq k \right\}$$

Then, we calculate the real mean of each clusters and determine the new mean points:

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j$$

Repeat the above process until the change in m_i is small enough each time.[8]

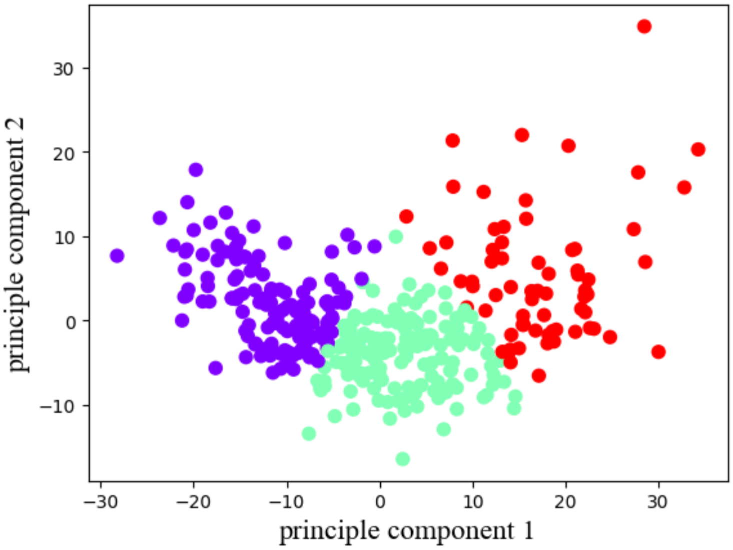


Figure 15: Clusters of PCA

We set three initial mean points and divide the data into three groups on the PCA plane. What's more, we tried a series of random initial points and get the same result, which means that our cluster result is stable. The PCA plane is spanned by $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$.

Eigenvector	1	2	3	4	5	6	X
$\boldsymbol{e_1}$	-0.0454	-0.2563	-0.6471	-0.2002	0.4339	0.4953	0.1996
$\boldsymbol{e_2}$	-0.0396	0.3112	0.1320	-0.7108	-0.4110	0.2469	0.3860

In figure 15, red group stands for hard difficulty, green group stands for medium difficulty and purple group stands for easy difficulty. $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$ represent different evaluation dimensions. $\boldsymbol{e_1}$ mainly weights the fifth and sixth data, while $\boldsymbol{e_2}$ mainly weights the seventh data.

According to the cluster result and PCA, the distribution is more dependent on $\boldsymbol{e_1}$. So the top of the green part will get a little instability. The precise distribution is exhibited in Figure 16(Yellow for Hard; Green for Medium; Blue for Easy).

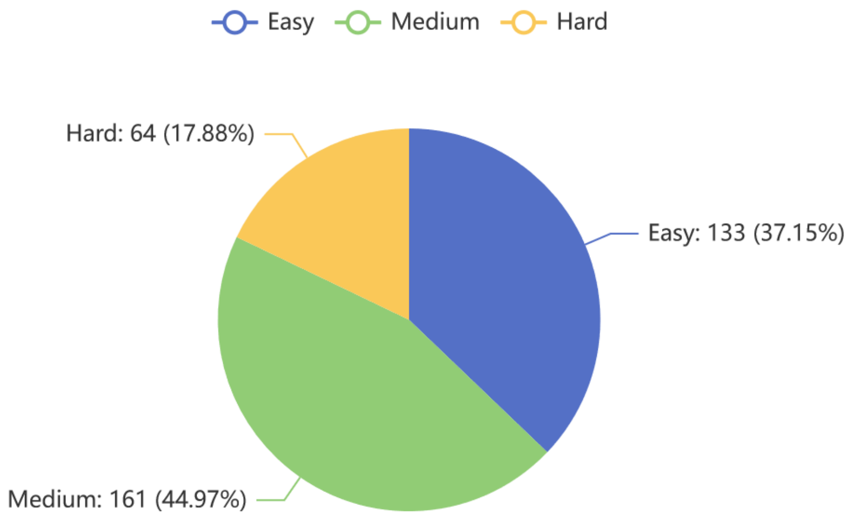


Figure 16: Pie Chart of K-means Cluster

5.2 Model II : Survival Analysis

Apart from assigning different weights to elements of $\boldsymbol{p}$, we set up a different model based on survival analysis method, which in turn considers a non-linear combination of p_k.

5.2.1 Rationality of the Survival Analysis Method

Due to the fact that the number of tries in wordle is limited to 6, special considerations should be taken respect to the 7th element of $\boldsymbol{p}$. The survival analysis method provides a valid way to deal with p_7: the 'survival time' here is the number of guesses made before solving the puzzle, and we can calculate the proportion of remaining subject $S(t) = \sum_{T>t} p_T$, which is also called the survival function. By using $S(t)(0 \le t \le 6)$ instead of $p_1, p_2, \cdots, p_7$, we successfully avoid considering the difference between the meaning of $p_i(i \le 6)$ and p_7.

The survival analysis method is broadly used in clinical trials to assess the effectiveness of a drug. In a typical clinical trial, the better the drug is, the longer the patients live; whereas in our model, the harder a word is, the more the guessing time is expected to be. We can therefore conclude that identifying the difficulty of a word is analogous to testing the effectiveness of a death-preventing drug, which makes survival analysis a reasonable approach to evaluate the difficulty of words.

However, as the number of subjects we study is significantly greater than that in an regular clinical trial, adjustments must been done to satisfy our needs. When comparing the survival distributions among two groups, **Mantel–Haenszel log rank test**[9] is usually conducted as an hypothesis test to determine whether the two distributions are different, by comparing the number of expected and observed events (E_1, E_2 and O_1, O_2) and applying the chi-square test :

$$\chi^2_{\text{MH}} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \rightarrow \chi^2_1$$

$$P - \text{value} = Pr(\chi^2_1 \geq \chi^2_{\text{MH}})$$

In our study, as the number of subjects in each group (the reported results on each day) is large, χ^2_{MH} often reaches a high value. From the above equations we can tell that both the difficulty of words and number of participants affect the χ^2_{MH}, since words with different difficulty levels tend to have a different distribution, and an increase of data size promotes confidence level. Therefore, χ^2_{MH} cannot serve as a unifying statistic to estimate difficulty from hundreds of distributions. To determine the difficulties of words, we need to find another statistic that is independent of the number of reported results.

5.2.2 Model Construction

To set up a general standard for difficulty evaluation, we compare a word's data with the overall data excluding the former word, which correspond to the two groups in survival analysis. The calculations are as follows ($N^{(k)}(0)$ is the number of reported results for word k):

$$\begin{aligned}
N^{(k)}(t) &= N^{(k)}(0) \sum_{T=t+1}^{7} p_T^{(k)} \\
N(t) &= \sum_{k=1}^{K} N^{(k)}(t) \\
O_1^{(k)} &= N^{(k)}(0) - N^{(k)}(6) \\
O_2^{(k)} &= N(0) - N(6) - O_1^{(k)} \\
E_1^{(k)} &= \sum_{t=0}^{5} \frac{N^{(k)}(t)}{N(t)}(N(t) - N(t+1)) \\
E_2^{(k)} &= \sum_{t=0}^{5} \frac{N(t) - N^{(k)}(t)}{N(t)}(N(t) - N(t+1))
\end{aligned}$$

To determine the difficulty of a word regardless of the number of participants, we propose the test statistic $h_k = \frac{E_1^{(k)}}{E_2^{(k)}}$

$$h_k = \frac{E_1(k)N(0)}{N^{(k)}(0)E_2(k)} \tag{1}$$

which eliminate the impact of the number of subjects via dividing $E_1^{(k)}$ by $N^{(k)}(0)$. The greater h_k, the harder the word k is expected to be. If h_k is close to 1, we can tell that the difficulty of the word k is around the average level.

5.3 Model Evaluation

5.3.1 Rankings

We sort the words according to z_k from model I and h_k from model II in descending order respectively, and denote the ranks of a word k as $r_2^{(k)}$ and $r_1^{(k)}$. The lower the ranks are, the harder the word tends to be. The below scatter plot demonstrates a strong correlation (correlation coefficient = 0.9904) between the two ranks, which implies that both models are capable of making accurate evaluation of difficulty.

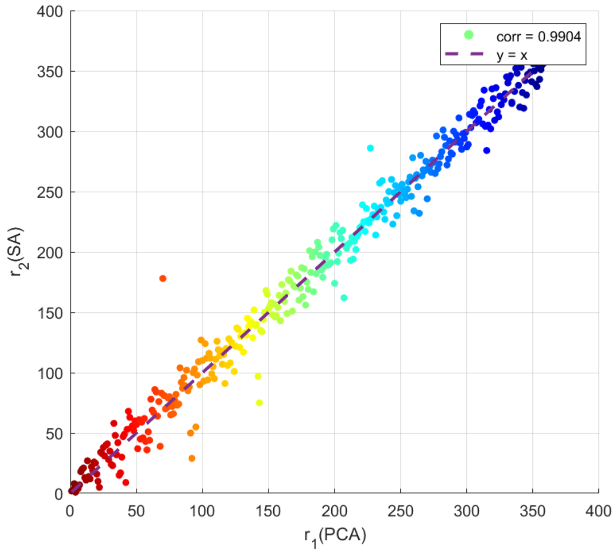


Figure 17: Scatter Plot of (r_1,r_2)

Word	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more	r_1(PCA)	r_2(SA)
parer	0	0	4	11	15	22	48	4	1
mummy	0	1	4	14	27	37	18	1	2
coyly	0	0	4	17	28	35	15	2	3
judge	0	2	8	16	26	33	14	6	4
foyer	0	2	10	19	19	23	26	22	5
inane	0	8	25	30	21	13	3	191	179
equal	0	5	23	35	25	11	2	172	180
shrug	0	4	23	36	26	10	1	171	181
focus	1	4	23	36	24	10	1	186	182
valet	0	4	22	38	25	9	1	182	183
rainy	1	16	38	31	11	3	1	357	355
stair	2	21	36	26	11	4	1	350	356
third	1	10	47	32	9	2	0	359	357
plant	2	19	39	28	10	3	0	358	358
train	6	26	32	22	10	3	0	347	359

Table 3: Part of the Ranking Results

From table3 we can see that although the two methods are generally consistent in determining difficulties, r_1 obtained by PCA tends to underestimate the difficulties of words having greater percentage of trying 7 times or more (e.g. parer and foyer). This is due to the fact that the practical meaning of p_7 is different from that of $p_1, \cdots, p_6$ and only 5% of words have a p_7 greater than 10%,which causes a decrease in the contribution of p_7 to the total variance of $\boldsymbol{p}$. Therefore, we regard the model using survival analysis(SA)[10] as a better way to evaluate difficulties.

5.3.2 Clustering

We apply the 1-D K-means clustering method to h_k and compare the clustering results between PCA and SA with the stability of classification checked. It turns out that most of the words lie in the same category, while four identified as 'Medium' in PCA and 'Hard' in SA, three identified as 'Medium' in PCA and 'Easy' in SA. The distribution of difficulty level in both results are reasonable.

Method	Easy	Medium	Hard
PCA	133	161	64
SA	135	155	68

Table 4: Clustering Results

5.4 The Difficulty Level of EERIE

The z and h of the word "EERIE" according to the distribution predicted by previous models are listed below:

Model Type	z	h	r_1 (PCA)	r_2 (SA)
LSTM Model	8.41	1.10	87	61
Model I	11.49	1.21	68	112
Model II	8.43	1.12	86	96

Table 5: Estimating the Difficulty of EERIE

The results above suggest that the difficulty of 'EERIE' lies in the first third of all words, which means that it is relatively hard to solve. We then project the distribution vector $\boldsymbol{p}$ of this word onto the 2-D plane on which K-means clustering was performed:

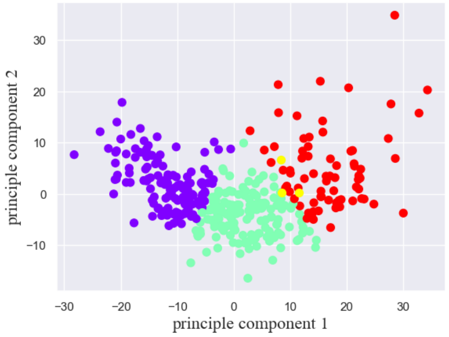


Figure 18: Classification of 'EERIE'

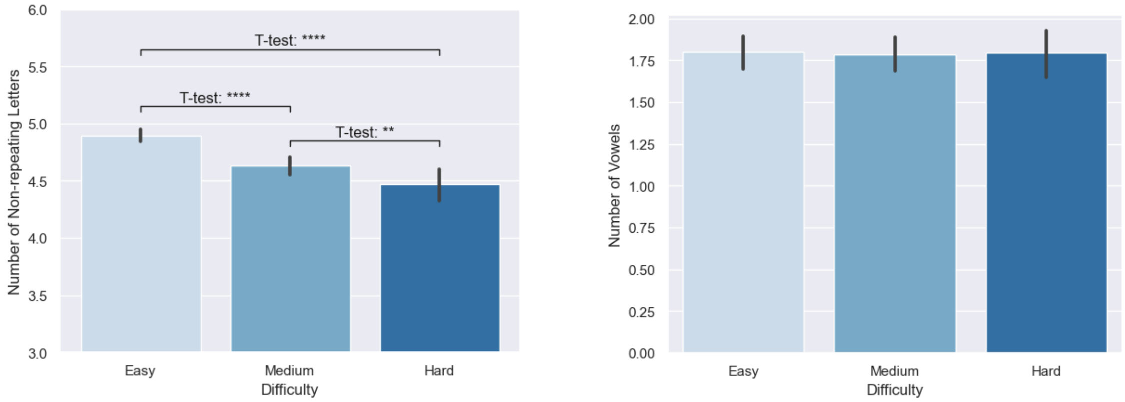
points in yellow represent the three predictions. As these points lie around the border between the 'Medium' group and the 'Hard' group and staying closer to the latter, we argue that 'eerie' is more likely to be a 'Hard' word.

5.5 Attributes of Each Classifications

5.5.1 Number of Non-Repeating Words

Combining the classification results and our personal experience, we propose that the word is harder to guess when it contains repeated letters, as the probability of finding the right letters is lower.

To verify this statement, we calculate the average and variance of the number of non-repeating letters in each category, and perform the Paired-Sample t-Test[11]. The results show strong evidence that the average number of non-repeating letters differ among different categories, and harder words tend to have more repeated letters.



(a) Number of Non-Repeating Letters

(b) Number of Vowels

Figure 19: Illustration of two attributes

5.5.2 Number of Vowels

We test whether the number of vowels differ among the three categories, and find no evidence supporting this statement.

6 Other Interesting Features

6.1 Correlation between Letters

We analyze the correlation between the frequency of occurrence $\{f(X)\}, X = a, b, \cdots, z$ between letters, and find that some letters have a correlation coefficient up to 0.4(e.g. "o" and "s", "c" and "h"). This may results from the innate features of English words: a word can be separated into several syllables consists of 2 or 3 letters, and some syllables are more common than others.

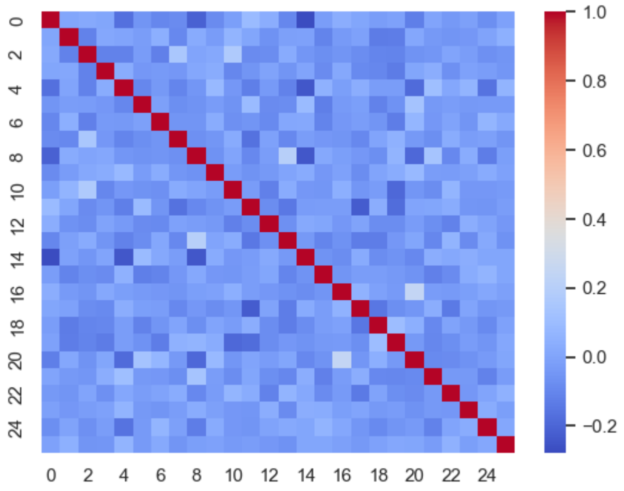


Figure 20: Correlation between letters

6.2 Correlation between Letters and Difficulty

We propose that a word containing more frequently used letters may be easier to guess. To test this statement, we calculate the correlation $\{\rho(X)\}, X = a, b, \cdots, z$ between the estimator of difficulty h_k and the occurrence of 26 letters respectively, and take a step further to test the relevance between the former correlation $\{\rho(X)\}$ and the frequency of occurrence $\{f(X)\}$. A correlation coefficient of 0.5 is obtained, which implies that there might be a weak correlation between letters and difficulty, with words containing more frequently-used letters easier to guess.

6.3 The Impact of Word Familiarity

As non-native English speakers, we observe that we recognize most of the words in the ’easy’ group while encounter dozens of words we’ve never seen before in the ’hard’ group. However, as the data provide no information about word frequencies, we cannot make quantitative analysis to assess the impact of familiarity to words on their difficulty levels.

Letter

Dear Puzzle Editor of the New York Times:

It is our honor to help you analyze the data on the *Wordle* website. Actually, having received your data of *Wordle*, we consider to make account for the inner correlation of the numbers from following aspects:

- The number of reports vary by date.
- There is association between word and the distribution of number of tries.
- How can the distribution above reflect the difficulty of words.
- Some problems in data need to be updated.

1) Variation of the number of reports

The tendency of the number of reports turn out to be different exponential functions before and after peak. So we apply common exponential function and Langevin Model to fit the data curve and both get excellent R^2. To make account for the model, we prefer to assume that the dissemination of *Wordle* expanded as pestilence at the beginning. Month later, the temperature of *Wordle* population slowly declined. On the other hand, we cast the data of last 60 days and put them into Grey Prediction Model, which assess to similar consequence. In our prediction of the two models, the number of reports on March 1, 2023 will decrease to **17608** or **18676**.

2) Possibility of inferring distribution of number of ties from words

Training 3 different ANN models, we find all of them perform well on finding the mapping relationship of words and the statistical data, and we are confident with their prediction results as they are highly similar to the published data (even for some words it has never met during its training process) and they have remarkable similarity with each other.

Their predictions of ”EERIE” on March 1,2023 show high consistency, indicating that the uncertainty level is low and these three models have found the precise relation, though they can hardly be understood by human being.

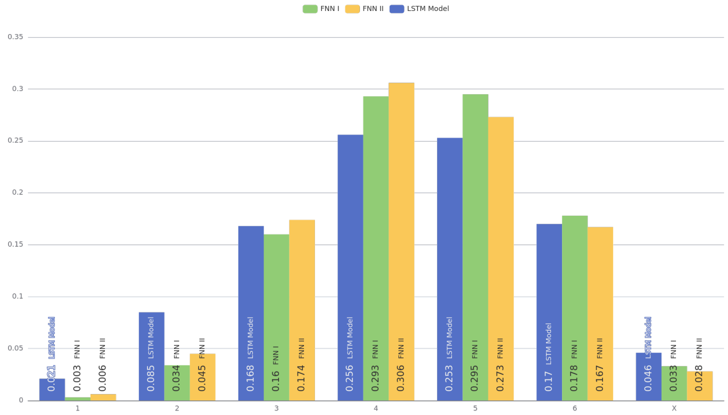


Figure 21: Predictions from 3 models on ”EERIE” March 1, 2023

3) From distribution to difficulty evaluation

By analyzing the distribution of tries using linear and nonlinear methods, we develop two different models (PCA and survival analysis) to evaluate the difficulty of each word. We categorized the words into three difficulty levels, and are able to make finer discrimination between each categories using both models. The figure below illustrates the distribution of difficulty considering words from January 7, 2022 to December 31, 2022, which is a quite reasonable.

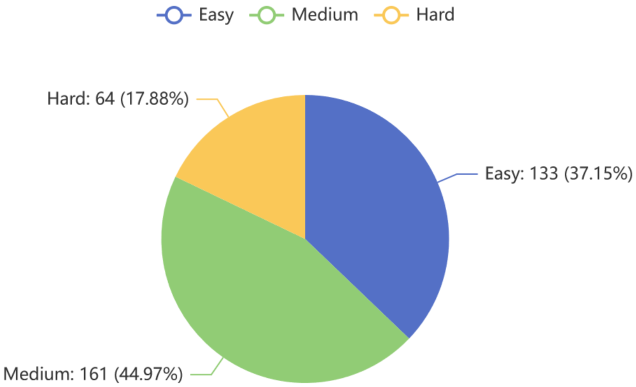


Figure 22: The Distribution of Difficulty

4) Data problem

The data you have provided us appears to have some wrong data point, several words are misspelled and some number of report are apparently lower then the commom level. What is more, sum of the ’percent in’ column turn out to be larger then 100 %. It is suggested that you can update these wrong data and the prediction may be more precise.

Thus, we tend to provide you some suggestions according to your data of the website:

Firstly, you could broaden your game mode, provide 4-letter or 6-letter words. It will not only abstract more visitors, but can also cater to fresh players.

Secondly, the distribution can be well predicted by our model, which provide you a chance to adjust the word reasonably. As a result, it may improve the user stickiness of your website.

At last, you can combine our prediction model of the distribution with our analysis of the attribute of the words, designing a series of questions from easy to hard. It can improve the player’s online time and the game experience.

These are all of our suggestions. Hope they will be useful to you.

Best wishes,
Team 2312165

Reference

[1] Bradley Efron. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81(395):709–721, 1986.

[2] Deng Julong et al. Introduction to grey system theory. *The Journal of grey system*, 1(1):1–24, 1989.

[3] George EP Box and David A Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332):1509–1526, 1970.

[4] Murat H Sazli. A brief review of feed-forward neural networks. *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, 50(01), 2006.

[5] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

[6] Andreas Daffertshofer, Claudine JC Lamoth, Onno G Meijer, and Peter J Beek. Pca in studying coordination and variability: a tutorial. *Clinical biomechanics*, 19(4):415–428, 2004.

[7] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006.

[8] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[9] Nathan Mantel. Propriety of the mantel-haenszel variance for the log rank test. *Biometrika*, 72(2):471–472, 1985.

[10] Stephen P Jenkins. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, 42:54–56, 2005.

[11] Robert W Mee and Tin Chiu Chua. Regression toward the mean and the paired sample t test. *The American Statistician*, 45(1):39–42, 1991.