Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2303626

Player Data Statistics for the Wordle Game

Summary

Wordle is a very popular game on the Internet recently, which has aroused widespread discussion. The New York Times attempted to analyze and discuss the Wordle game itself, classified words according to the attributes of words, and discussed other features of 359 pieces of data.

As for question 1, we first do data cleaning by removing unnecessary information. Since players increase regularly with **network communication** and **time sequence**, and some seasonal fluctuations are obtained through analysis. Therefore, **Seasonal Autoregressive Integrated Moving Average model with exogenous variables** (ARIMA(1, 1, 2)x(2, 1, [1], 7)) was adopted to calculate the number of predictions reported results on March 1,2023 with the upper bound is 30211.35331 and the lower bound is 9895.00527. We define five word attributes and use difficulty to display the percentage of scores reported. Through correlation analysis, it is found that attributes of the word have less effect on difficulty. We explained from two aspects

As for question 2, this is a prediction problem. Through **correlation test**, the correlation between the five attributes of words and the distribution of the reported results is not strong, so we adopt a nonlinear model. The **Back Propagation Neural Network** is constructed, the hidden layer is 5-3-5-7, and the effect is good. For the word EERIE on March 1, 2023, Our forecast for the percentages of (1, 2, 3, 4, 5, 6, X) is 0.2, 2.4, 13.1, 31.2, 33.4, 16.0, 3.4, which is more conform to the **skewness distribution**. There are several uncertainties in our model, such as **Cognitive preferences** for different Words and **tendency choice** for the first word. Further, we use the **decision tree model** to calculate the **optimal solution**, that is, the optimal solution of the average guess number of all cases.

As for question 3, this is a clustering problem. Since the correlation of all factors is weak, we use the **spectral clustering algorithm** to divide the whole data set into three groups, tagged as *simple*, *general* and *difficult* groups based on the average mathematical expectation of the number of guesses within the group. EEIRE is in the general group.

As for question 4, through the visual analysis and comparison of the obtained data, we found more correlations and trends between these factors and attributes, and obtained some obvious conclusions

Keyword: ARIMAX, BPNN, optional solution, spectral clustering

contents

1 Introduction

1.1 Problem Background

Word guessing game wordle, which became an overnight hit in 2022, caused a lot of discussion. Players are prompted to guess a given five-word word for the day. The rules of the game are as follows. If you submit your word, color of the tiles will change. To be more specific, yellow tile indicates the letter in that tile is in the word, but it is not in the right location. While, green tile shows that the letter in that tile is in the word and is in the correct location. For the daily specific vocabulary, common degree, number of repeated words and the combination may affect the daily answer rate and answer time.

1.2 Restatement of the Problem

The Puzzle Editor of the New York Times are curious about the factors that influence the distribution of the reported results. Based on the analysis of word and existing results, we can predict the results of future answers and the distribution of the reported results for a particular word.

To achieve our goals, specifically, we need to:

- Find how the reported results have changed with time, and predict the number on March 1,2023.
- Discover some attributes of the word affect the percentage of scores.
- Predict the distribution of the reported results according to our model, analysis the uncertainties and predictions. Give our prediction for EERIE.
- Establish a model to sort words given by the degree of difficulty.
- Find out other characteristics of the data.

1.3 Our Approach

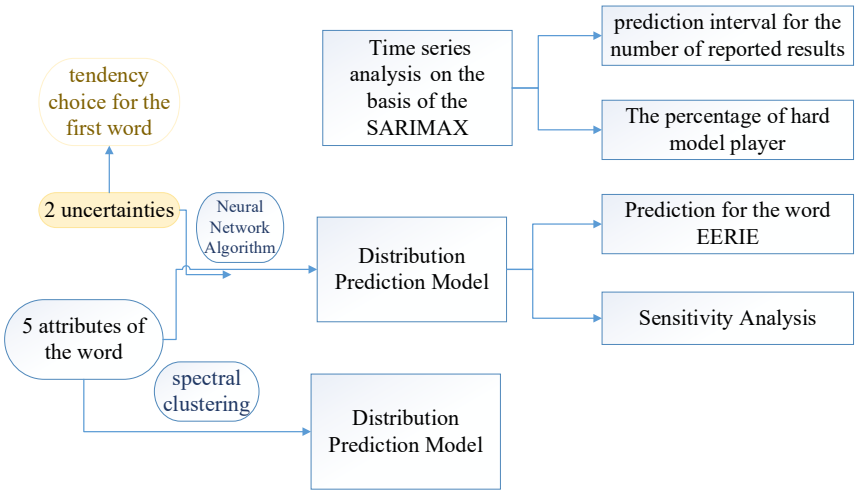


Fig. 1 our approach

2 General Assumptions and Model Overview

To simplify the problem, we make the following basic assumptions, each of which is properly justified.

- **Assumption 1:** the players finish the game on your own, not received external help

  Justification: There is no interference with the proportion of attempts made

- **Assumption 2:** basically the same players for the vocabulary

  Justification: The effect of individual cognitive differences on the outcome is eliminated, so that only the player's own subjective judgment and exclusion affect the outcome

3 Model Preparation

3.1 Notations

Important notations used in this paper are lists in Table 1,

Symbol	Definition
syllable	Number of syllables in a word
Freq	Word frequency
Lcount_1~3	letter frequency
Lcount_max	Letter repeat
T1~7	Number of attempts to solve the puzzle

4 Data Processing and analysis

4.1 Data processing

We found that in some days, the number of letters in word were not 5, for example on November 26, 2022, the number of letters in clen was 4. So we deleted those lines of data for processing, and the rest data was verified to be in line with the actual law.

5 Mode I: Time series analysis on the basis of the SARIMAX

5.1 Model Overview

Given that the results of daily reports change, we tend to explore how the results of reports change over time, and explore the inherent regularity of the number of reported results over time scales. We introduce a time series model to study the influence of time factors on the results of daily reports. So we use the SARIMAX to predict the future occurrence pattern based on the past 359 days.

We arranged the data in chronological order, set 1-300 pieces of data as the training set and 301-359 pieces of data as the test set. That is, the forecast is the verification of reported results and actual values after January 3. What's more, in order to find the optimal SARIMAX parameter, we use machine learning GridSearchCV to find the parameter with the highest

precision on the validation set from all the parameters. Finally, we choose ARIMA(1, 1, 2)x(2, 1, [1], 7).

(1) Autoregressive model: An autoregressive model is a model that describes the relationship between current values and historical values. It is a method of predicting itself using the historical event data of the variable itself. is the current value; is the constant term; p is the order; is the autocorrelation coefficient, and is the error value.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-1} + \varepsilon_t \tag{1}$$

(2) Integrated: The most important part of the ARIMA model is the smoothness of the time series data. Stationarity is the requirement that the fitted curve obtained through the sample time series can continue inertia along the existing morphology in the short time in the future, that is, the mean and variance of the data should not theoretically change too much.

(3) Moving average: The moving average model focuses on the accumulation of error terms in the autoregressive model. It can effectively eliminate random fluctuations in predictions.

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2}$$

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-1} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{3}$$

5.2 **Stationarity test**

Only relatively stable data can be used for time series analysis. For the original series, the frequency distribution map can be made, and it can be obtained that p-value=0.063981>0.01, the p-value is too large, which is considered to be unstable. So we conducted a difference on the original data. It can be seen from the figure that except for abnormal fluctuations in the first four months, the rest difference results are all about 0, showing a good stationarity. After testing, it can be seen that p=0.004110<0.01, pass the stationarity test.

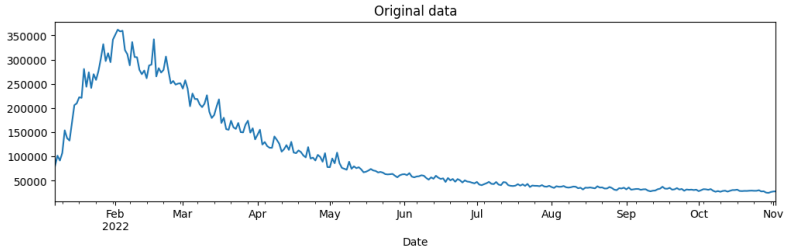


Fig. 2 the distribution of the original data

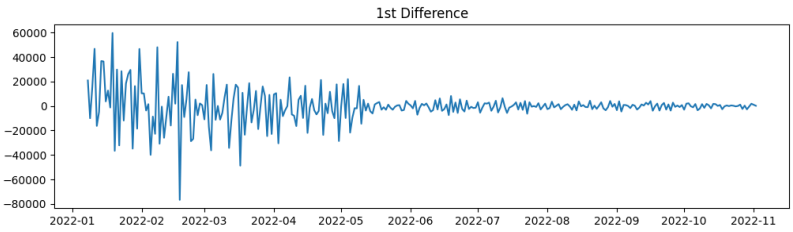


Fig. 3 the distribution of the 1st difference

5.3 Seasonal decomposition

After the preliminary judgment of the graph, it can be found that there is a certain periodic fluctuation, so we carry out the seasonal decomposition. Also, we conducted preliminary seasonal stationarity test, and obtained p-value =0.000003<0.01, passed the test of seasonal stationarity, so we decomposed the seasonality and obtained the cycle period=7. Practically, it is reasonable to take one week as the cycle, because there are fewer players on weekdays and more players on rest days. By separating the periodicity from the original frequency distribution curve, a stable trend line can be obtained.

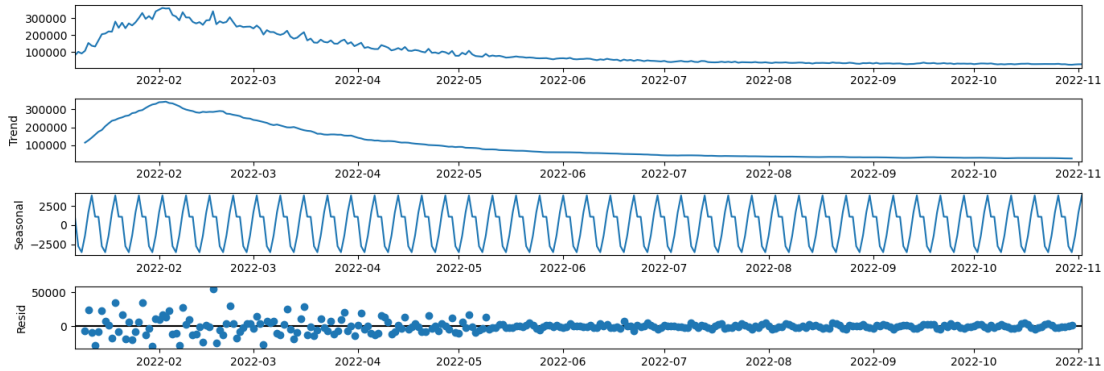


Fig. 4 the distribution of season

5.4 Determination of time series parameters

For the determination of the parameters of the ARIMA p_1, d_1, q_1 and periodic parameters p_2, d_2, q_2, in order to better match the parameters with our results, we calculated the parameters using GridSearchCV, Finally we decide the parameters, ARIMA(1, 1, 2)x(2, 1, [1], 7)

5.5 Prediction on March1,2023

The following figure shows the predicted results of players over time. It can be seen that, except for some outliers, the overall trend is upward to leveling off. Based on the time series results of 1-300 training sets, the forecast of 301-359 data pieces is consistent with the reality. Therefore, we further calculate the forecast interval of March 1, 2023, with the upper bound is 30211.35331 and the lower bound is 9895.00527.

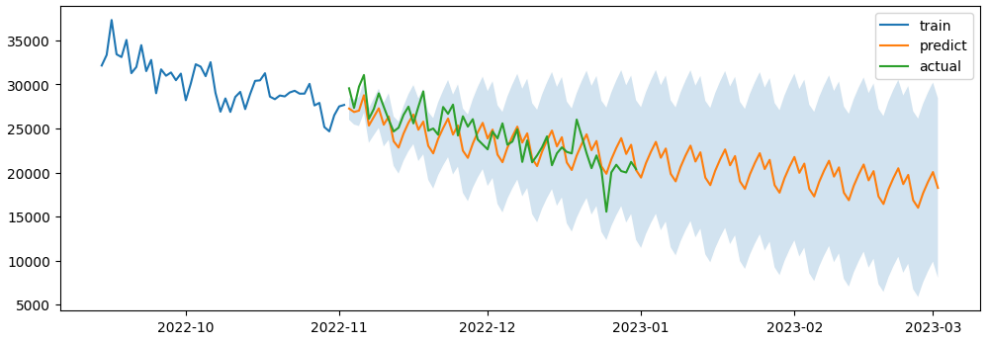


Fig. 5 the prediction on March, 1,2023

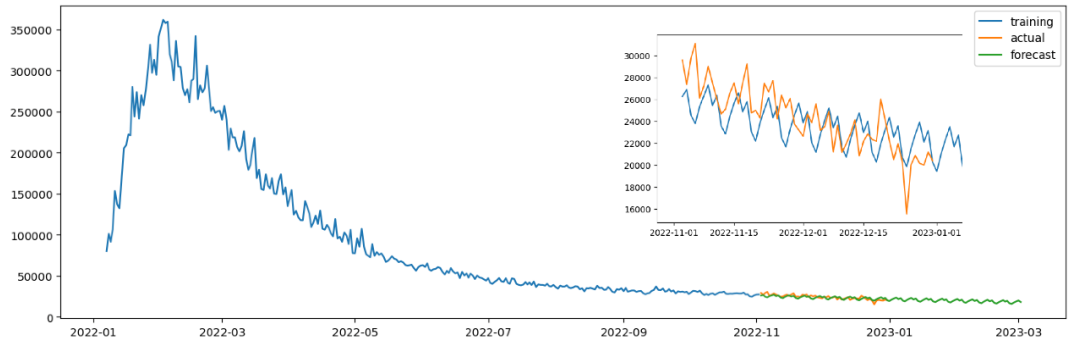


Fig. 6 The result of SARIMA

5.6 The attributes of the word effect on hard mode players

Firstly, we define the attributes of the word that might affect the number of people in hard mode as follows:

Word frequency: indicates the frequency of the word, which can reflect the common degree of the word, that is, the higher the frequency, about easy to think of

Letter repeat: indicates whether a single letter of the word is repeated and how many times it repeats.

Initial: indicates the initial letter of the word

Letter frequency: The frequency of the word in the corpus. The higher the frequency, the higher the probability of the word appearing in the five-word word, the more likely the player is to find the prompt message through it.

Syllable:

Difficulty: indicates the difficulty of the word. Specifically, the number of attempts presents a skewed distribution. We use the mean value of the skewed distribution to reflect the difficulty of the word.

However, we represent their correlation coefficients using heat maps and find that percentage is difficulty, word_freq, repeat. The correlation coefficient of is extreme small and can be considered irrelevant. There may be two reasons as follows

- Rules of the game: Choose whether to try the difficult mode first, and then guess the word.

That is to say, as far as people's subjective judgment is concerned, the difficult mode is not related to the attribute of the word, but to the individual behavior pattern

- Timing analysis: We explored the time variation of percentage of scores reported and found that the overall distribution presented a skewed distribution. Interestingly, with the passage of time, the proportion of people choosing the difficult mode increased and the probability gradually stabilized.

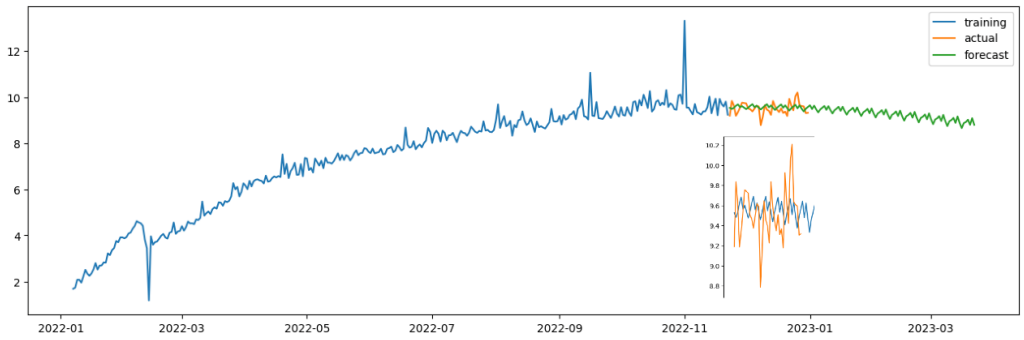


Fig. 7 the result of SARIMAX

6 Model II: Distribution Prediction Model Based on Neural Network Algorithm

6.1 Model overview

The distribution of the reported results is affected by different factors, including the attribute characteristics of the word itself, human decision-making, etc. In order to facilitate quantification, we focus on the impact of word attribute on distribution. So we use BP neural network algorithm to show the relationship.

BP neural network is a multilayer feedforward network trained according to error back propagation algorithm. It consists of forward propagation of information and back propagation of error. The neurons in the input layer are responsible for receiving all kinds of information from the outside world and transmitting the information to the neurons in the intermediate layer. The neurons in the intermediate hidden layer are responsible for processing and transforming the received information and processing the information according to the requirements. In practical application, the intermediate hidden layer can be set as one or multiple hidden layers, and the information can be transmitted to the output layer through the hidden layer of the last layer. This process is the forward propagation process of BP neural network.

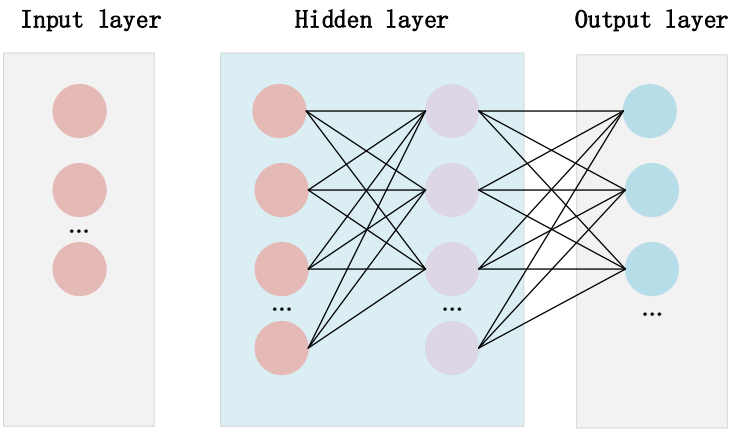


Fig. 8 the Principle interpretation of BPNN

6.2 Model input

For each word, there are a variety of attributes to describe its differences. We construct the following five feature vectors with six dimensions as input layers. More information about the feature vector is shown in Table.

Word frequency: The occurrence frequency of five-character words in the corpus is counted, and the data is compressed within the range of [0,1] through standardized processing and normalization. In addition, some words with the highest frequency are counted, as shown in the figure below.

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{4}$$

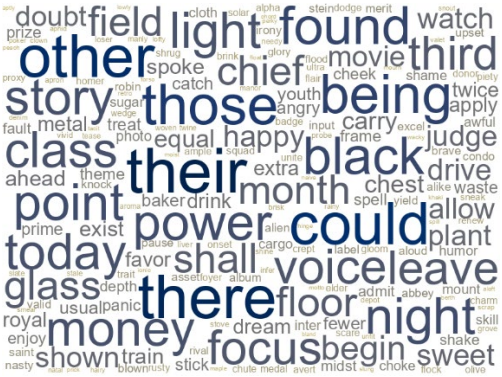


Fig. 9 Word frequency graph of five-character words

letter frequency: We will download the Cambridge dictionary, take out all five-word words. What's more we remove all words whose frequency is less than 10%, because according to the game mechanics, extremely rare words will not be used as the answer to the riddle. The frequency of 26 letters will be sorted. We divide letter frequency into three categories. We assign the value of the five words of the most frequency as 1, and believe that among any input letters, it is more likely to get yellow squares and green squares. The more information we get, the easier it will be to guess, and assign 3 to the five words with the lowest frequency. It is

believed that the probability of the target word containing these words is smaller, and the less easy to guess. The remaining 16 words are assigned a value of 1, somewhere in between.

Table 1 the definition of the categories

letter frequency	Definition
easroiln	1
tdcuhmpyb	2
gkfwvxzjq	3

Letter repeat: Values may occur as 1,2,3,4,5. After analysis, most of them occur between 1 and 2.

Table 2 Feature Definition of Vector

Dimension	Definition	Domain
X_1	Word frequency	[0,1]
X_2	letter frequency	1,2,3
X_3	Letter repeat	1,2,3,4,5
X_4	Initial	[1,26]
X_5	syllable	0,1

All the features above are already defined in the previous sections

6.3 BP Model

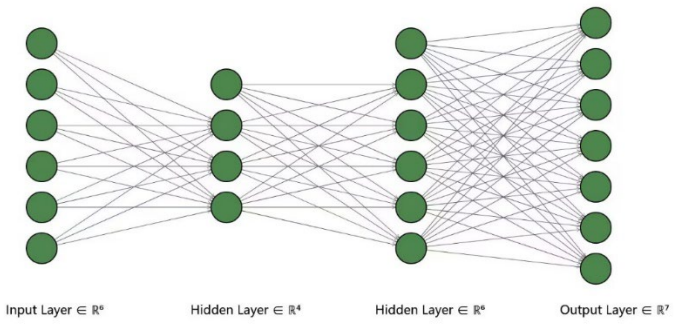


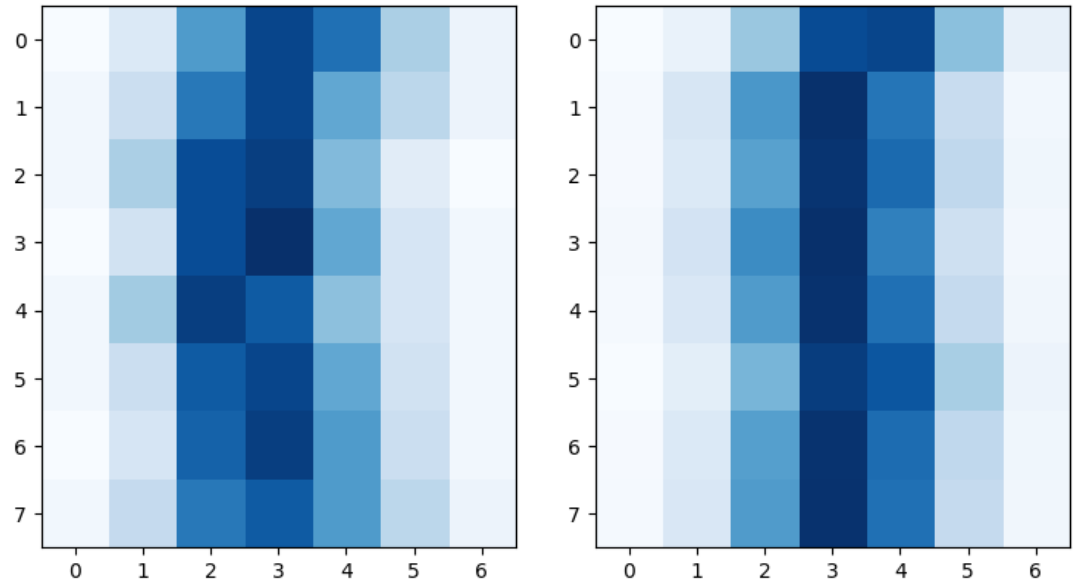
Fig. 10 BP diagram

As the image divides The hidden layer obtained from the input layer of 5 layers into 3 layers, and then into 5 layers, and finally the output layer is eight layers

6.4 Model output

After analysis, we continuously processed the discrete values of attempt times and found that percentages of attempt times were roughly consistent with skewed distribution. percentages of (1, 2, 3, 4, 5, 6, X) were output as 7 targets. If percentages were consistent with the original distribution, the model could be considered reasonable. We randomly selected ten words from

the corpus and checked the prediction results. It was found that the overall distribution was skewed, which was more consistent with the existing results.



（a）Actual result **（b）forecast result**

Fig. 11 the prediction of EERIE by BPNN

Table 3 the prediction for the word EERIE

The times of tries	Percentage(%)
1	0.2
2	2.4
3	13.1
4	31.2
5	33.4
6	16.0
x	3.4

6.5 Uncertainties in BPNN Model and optimal solution

Due to the small amount of data, in addition to the objective factors such as the inherent attributes of words, the different thinking of individual players will also affect the results. However, it is difficult to quantify those factors. Viewing that a neural network is described as a black box. There are some hidden factors in the learning process of all neural networks, and we have taken into account the main influencing factors. Cognitive preferences for different Words and tendency choice for the first word are the uncertainties that we take into account.

6.6 One possible way to take tendency choice for the first word in to consideration

Next we will consider the possibilities of various attempts, known as information gain or

information entropy, from an information theoretic perspective, and consider whether there is a word to be tried first that has the least entropy and the most information gain, allowing the player to complete the result with the least number of attempts on average. Intuitively, the 5-word with the lowest entropy needs to get as much information as possible. A word is optimal if the result of a guess allows us to filter out as many words as possible. Therefore, we use the decision tree model to calculate the optimal solution, that is, the optimal solution of the average guess number of all cases.

Decision tree is originally a concept in operational research classification algorithm to describe the process of decision making, but in machine learning it is a model of classification algorithm. In the case of the Wordle guessing game, the goal is to construct a decision tree with as little "depth" as possible by determining which word is guessed at each node.

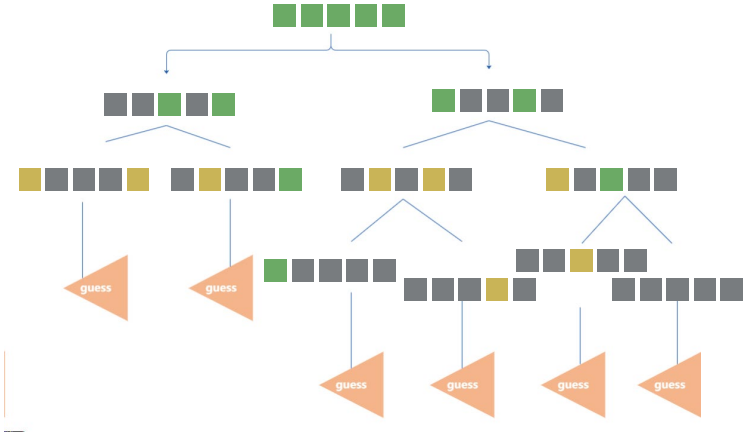


Fig. 12 Decision tree model diagram

Table 4 The smallest total average guess from the decision tree

Dimension	Total average guess
Salet	3.24
Reast	3.42
Crate	3.42
Trace	3.42
Slate	3.43
Crane	3.43
Rance	3.46
Lance	3.46
Ronte	3.47
Alter	3.48

7 Model III: Word Classification System based on spectral clustering

7.1 Correlation analysis between word features and difficulty

As can be seen from the figure, among the five indexes constructed by us, there isn't strong linear correlation among the factors we choose. So simple models like k-means may be not helpful for us. We apply spectral clustering on 359 pieces of data (Since spectral clustering is transductive, we add *EERIE* into the dataset too.) with sub-feature vectors, and use the mean value of the mathematical expectation of the number of guesses players tried to solve the puzzle to identify its difficulty level.

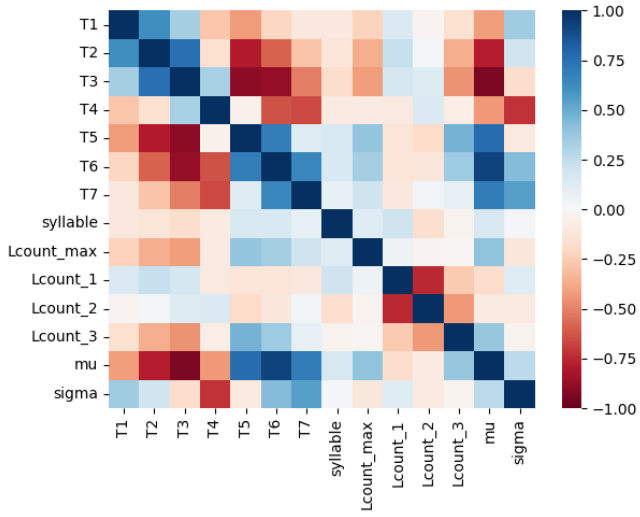


Fig. 13 Correlation analysis between word features and difficulty

7.2 Model overview

1 An undirected graph $G(V,E)$ is constructed according to the data. Each node in the graph corresponds to a data point, and similar points are connected. The weight of edges is used to represent the similarity between data.

2 Calculate the adjacency matrix A and degree matrix D of the graph, and solve the corresponding Laplacian matrix $L_{N\times N} = D - A$. In constructing the adjacency matrix A(Affinity matrix), we use the Radial Basis Function(RBF) in the full-join method, which can project the original feature to infinite dimensions. It is useful in cases where data may need to be classified in a non-linear way.

3 The former of L k from small to large order of eigenvalue $\{\lambda\}_{i=1}^{k}$, and the corresponding eigenvectors $\{v\}_{i=1}^{k}$.

4 The k feature vectors are arranged together to form a matrix $N\times k$, each row is regarded as a vector in the dimensional space, and the K-Means algorithm is used for clustering, and the final clustering category is obtained.

As for the number of attempts, the number of attempts is generally in the normal distribution, and there are slight differences for words of different difficulty. The vast majority

of people can finish all the words in the first four times, no matter the difficulty of the words. Under normal circumstances, if a player uses the first two attempts wisely, he/she can get enough information and solve the puzzle in the third or fourth time.

7.3 Result

The spectral clustering algorithm divides the dataset into three clusters as follow:

Table 5 the result of cluster

Cluster	count	Average number of guesses	Tag
0	135	4.33	Difficult
1	154	4.09	Simply
2	69	4.16	general

The kernel density estimate is used to calculate the probability density of the mathematical expectation of the number of guesses players tried. The following picture shows the difficulties of the three groups we divided.

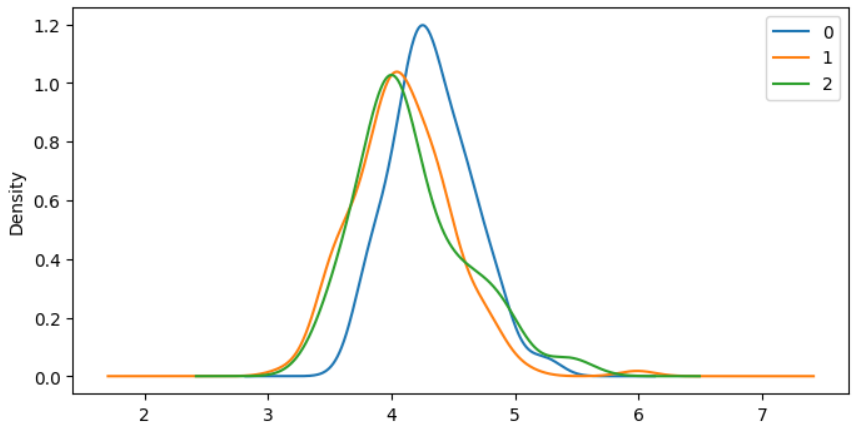


Fig. 14 the result of groups

8 other interesting features of this data set

As for the number of player, after exponential growth, the total number of people gradually decreases within a certain period of time and gradually approaches a stable equilibrium in the near future. This may be because most people do not maintain lasting interest after playing the game after its popularity. Such a complex neural network dynamic model reaches a stable and equilibrium point after this period of time, and the remaining people who can keep playing the game become the main players of the game. At the same time, based on temporal variations and predictive models, the percentage of people playing the game who choose the hard mode increases, which can be interpreted as players being able to take on more difficult difficulties as they become proficient. And given the difficulty of the word, over a period of time, the difficulty of the word has a slight effect on the player's choice of difficulty mode later, based on the complex thinking of individuals and networks, the reasons taken into account should be more complex and unstable.

As for words, the frequency of each word in daily life obviously has a significant impact on the difficulty of words. For people's guess, the part of speech of a word also has a great impact on the guess. Due to the reasons of English grammar and language itself, people tend to have stronger reactions and thoughts for verbs, adjectives and nouns, while for prepositions, auxiliary words, articles and adverbs, etc. At the same time, because there are fewer syllables combined with five letters, the word range will be greatly reduced after knowing that a syllable is guessed. Then, according to the degree of common use, people tend to guess more commonly used words more easily. Therefore, fixed letter combinations and syllables greatly reduce the number of people's guesses. A higher frequency of the first letter and multiple repeated letters were also strongly linked to difficulty, but whether a word had multiple meanings was slightly less important.

As for the number of attempts, the number of attempts is generally consistent with the correct distribution, and there are slight differences for words of different difficulty. The vast majority of people can complete all the words in the first four times, no matter the difficulty of the words. Under normal circumstances, if they need the first two trials and errors, they can get more information and guess the words in the third and fourth times.

9 Test the Model

9.1 Sensitivity Analysis

In real life, statistics are often inaccurate and there may be some biases in the inputs to our models. These biases may affect the results of our model. To test the robustness of our model, we will analyze the sensitivity of our decision model in the analysis.

10 Conclusion

10.1 Strength

- The ARIMA time series model can make better predictions and is more stable than others.
- BP neural network system has the characteristics of non-linear and intelligent. Qualitative description and quantitative calculations, precise logical analysis and non-deterministic reasoning are well considered
- Compared with the traditional clustering algorithm, the spectral algorithm we use can cluster on any shape of sample space and converge to the global optimal solution, and when the number of clustered categories is small, the spectral clustering effect will be very good.

10.2 Weakness

Due to the small amount of data, the predicted data will be overfitting.

The spectral clustering effect depends on the similarity matrix, and the final clustering effect obtained by different similarity matrices may be very different.

10.3 Improvement

For the time series model, after seasonal adjustment is used to remove the influence, the empirical mode decomposition (EMD) can be used for denoising. We can also consider the influence and change of the accumulation of a period of time on the subsequent results, which is more accurate. For the complex network dynamic system, the influence of time delay function on the equation can be considered, so as to solve the stability more accurately.

For BPNN Algorithm, we are supposed to optimize the uncertainty construction method and pruning method of network layer number, node number of each layer and node connection mode in the topology structure of the network structure. The network structure can be transformed with the sample space to simplify the network structure. Meanwhile, other algorithms can be mixed to increase the stability.

For the improvement of clustering method, we can improve the initial partition and apply a more suitable method to find k value, judge the difficulty, and use information entropy to increase entropy for classification.

11 Our Letter

Dear Sir/madam

We would like to thank you for your trust in our team. We have now completed our report about the popular puzzle—Wordle. After our analysis, we got the change rule of game players, and divided the difficulty level of the words in the corpus of Times New York. I hope the results obtained by our data analysis can be helpful and enlightening to you.

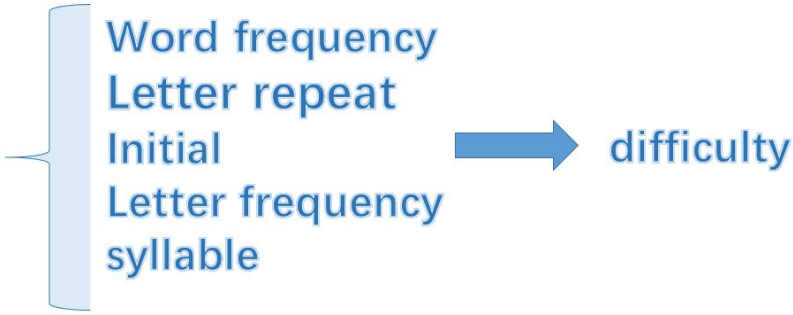
The followings are our suggestions:

Increase user engagement： The number of players grows rapidly in the early period, decreases afterwards, and then levels off. The rapid growth occurred when wordle sparked discussion on the Internet about the internal laws of Wordle, suggesting that new gameplay could be developed to increase the sense of experience and freshness

Increase the difficulty: and the five-word corpus was updated. In view of the increasing proportion of people in difficult mode, the in-depth discussion on the best solution of wordle made players more inclined to challenge difficult answers to riddles

constructed the difficulty evaluation system: predict and analyze the problem solving situation of the specified words, judge its rationality, and then adjust and change.

For the classification of difficulty levels and prediction of response distribution, we mainly built models and predicted results based on the attributes of words themselves. The associated attributes of words are as follows：



If you want to know more details, please refer to our thesis. We will be glad to discuss with you on our solution details.

Sincerely yours

MCM 2023 Team

References

[1] Xiaocai Zhang, Lining Zhao, Haibao Wang, Hai jiang Li. Classification Method ofNavigational Aids in Inland waters Based on Big Data [C]// Proceedings of the 9thInternational Conference on Frontier of Computer Sci ence and Technology (FCST 2015),2015: 204-208.

[2] Fukun Xing, Research on English Legibility Based on Information Computing and Development of IRMS Application System, Master's Thesis. Xue Wei, Statistical Analysis and SPSS Applications, Beijing: Chinese Minmin University Press, 2008

[3] Wang Jianxin, Construction and Application of Computer Corpus, 2005.

Zhou Mingqiang, Exploration of Language Cognition and Language Application, Beijing: China Social Sciences Press.

[4]Yuan Ye, Li Dan, Experiential and Operational Language Comprehension, Journal of Sichuan University of Foreign Languages, 2007, 5,. Yulin Yuan, Cognitive Research and Computational Analysis of Language, Language and Writing Applications, 1996, Yulin [5]Yuan, Language Research in the Context of Cognitive Science, Foreign Linguistics, 1996, 2: 1-12.

[6]Yu Guoliang, Research and Application of Corpus Linguistics, Chengdu: Sichuan University Press, 2009.

[7]Juecheng Zhang, Discussion of Some Basic Issues in Natural Language Understanding Research, Journal of Ningxia University.