Prediction of Wordle Results Based on Ridge Regression Model and K-means Clustering

Xiuhan Zheng[1], Xiaoli Jiang[2*], Xiaodong Fan[3*], Xueshu Wu[4]and Yue Zhou[5]

{1929989321@qq.com[1], jxls309@163.com[2*], bhdxfxd@163.com[3*]}

College of Mathematical Sciences, Bohai University, Jinzhou, Liaoning, 121013, China[1,2,4,5],
Faculty of Electrical and Control Engineering, Liaoning Technical University, Huludao, Liaoning, 125105, China[3]

Abstract. Wordle is a word game that became known in January 2022. In order to innovate the game and improve the participation of the game, this paper analyzes a series of data by using time series model, gray prediction model, ridge regression, K-means clustering and other methods. In this paper, we predict the effects of player number intervals and word attributes on the results of Wordle at a certain stage in the future, and classify the words to find out the characteristics of each class of words, and predict the reported results of a certain word based on the word characteristics. At the same time, our experience can be used as classroom examples of innovative word game strategy algorithms and statistics to demonstrate the research of this paper.

Keywords: time series model, ridge regression, K-means clustering, the word game

1 Introduction

Wordle has received a lot of attention since it was released. Players have six opportunities or less to guess a realistic English word with five letters, and players get some feedback [1]. The game was acquired by The New York Times in 2022, and the New York Times website explains that after entering a word, the tile color will change to yellow, green or black. The player determines whether the letter exists based on the color change of the tile, and if so, whether the position is correct [2]. Silva [3] identified the three most suitable starting words to solve the puzzle. Liu [4] learned from using Wordle to design and compare strategies. Wormley et al [5] studied the relationship between cheating and religious belief and culture based on Wordle. Martin [6] dealt with Wordle mathematically. Daniel et al [7] studied Wordle game from the perspective of complexity and prove the np hardness. Bonthron [8] explored the method of converting words into vectors and the role of rank 1 approximation and latent semantic index in Wordle. Although many of the current literature have studied the solution of the Wordle puzzle and other phenomena based on Wordle, there are few data analysis. In this paper, we look for the relevant data [9] of Wordle to predict and analyze the results of the puzzle game.

In this paper, we predict the quantitative interval for a certain day in the future and analyze the influence of the attributes of words on the results [11]. Through the time series model to report the number of results to explain [12], the model is obtained to show a decaying trend. The gray prediction model [10] was used to predict the number of user participation on March 1,2023,

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334479

obtaining an interval of 19121 to 20218 people. Predict the percentage of times a word will be solved (1, 2, 3, 4, 5, 6, X) on a future day and give an example. Ridge regression was used to establish a model [13] to classify the vocabulary in the database [14]. Substituting the data with the number of times the player guessed the puzzle for data analysis can get a prediction of the relevant percentage for a future day. The error analysis was carried out, and the calculation formula of R^2 was used to obtain the values of 1, and the fitting was good [15]. Sorting words of the past year byword's difficulty [14] and determine the attributes of each type of word. K-means clustering model for the percentage associated with the number of times needed to solve the problem successfully [16]. Find the corresponding properties of words in the three clusters, respectively is the part of speech, the frequency of words in various cases [14]. Identify the interesting features of the data set and verify the [17]. A normality test using SPSSPRO yielded the characteristic that the number of daily reported results, the number of participants in the difficulty model, and the percentage of each guess is normally distributed [18].

2 Data collection and processing

Generate daily game's data from January 7, 2022 to December 31, 2022 based on game score reports shared by users on Twitter, which includes the date, the word of the day, the number of people reporting results for that day, the number of players in hard mode, and the percentage of 1 try, 2 tries, 3 tries, 4 tries, 5 tries, 6 tries and 7 or more tries (X) that guessed the word or were unsuccessful.

Wordle requires players to guess a word consisting of five letters to solve the puzzle, from this condition, it can be seen that TASH and CLEN in the data does not meet the requirements. According to the pattern of the number of people reporting scores on that day in the data, it can be seen that the number of people reporting scores on the day of November 30, 2022 is too small and there is an obvious error, so it is deleted. According to the basic laws of mathematics, the sum of the percentage of guessing the word once, twice, three times, five times, five times, six times and failing to solve the puzzle should be 100%, but due to rounding, there is a certain error, due to the error analysis of each number, the sum of the percentage should be 93%~107%, so NYMPH is excluded.

3 Results and analysis

3.1 Results to report the 21 March 2023

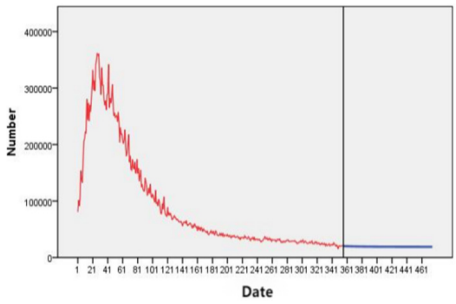

Fig. 1. Gray prediction results

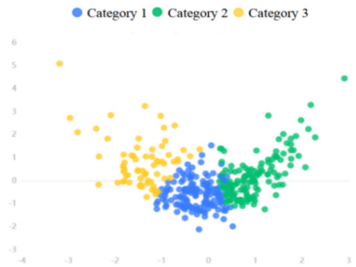

Fig. 2. Clustering scatter plots

The blue curve in Figure 1 is the prediction curve, which predicts that the number of user participants on March 1, 2023 is between 19121 and 20218.

3.2 Correlation percentage for predicting a certain future day based on ridge regression

By using Excel on the data of single word a column to filter it into four categories: Words with no identical letters (coded as 1), words with 2 identical letters (coded as 2), words with 3 identical letters (coded as 3) and words with two pairs of 2 identical letters (coded as 4).

Substitute the above requirements and the given data for data analysis to obtain the predicted values, and take the percentage prediction of X as an example, see Table 1 for details.

Table 1. Forecast results

Variable	Coefficient	Test value
Constant	4.006993006993007	1
Word_2.0	0.9990009990009991	

Variable	Coefficient	Test value
Predicted outcome 7 or more tries (X)		4.007

Table 2. Cluster summary

Clustering categories	Frequency	Percentage
Cluster category _1	154	43.38**%**
Cluster category _2	133	37.465**%**
Cluster category _3	68	19.155**%**
Total	355	100.0**%**

To predict the difficulty of the word EERIE, EERIE contains three same letters, which belongs to the category of number 3, substitute the number into the established model, see Table 3 for details. The sum of the predicted values is approximately 100 %, The error analysis of the "ridge regression" is performed to calculate the goodness-of-fit R^2 of the model, see equation (1), and the R^2 obtained is 1. In general, the closer the R^2 is to 1, the better the fit is [19]. By R^2 is 1, so there is more confident for this model.

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2} \tag{1}$$

Table 3. Prediction results of EERIE

	1 try	2 try	3 try	4 try	5 try	6 try	7 or more tries（X）
Predicted value	0	2.998	0	21.998	40.995	26.004	7.004

3.3 Classification of solution vocabulary based on K-means clustering

The clustering results of the solution vocabulary were divided into three categories [20] based on the number of times needed to solve the problem successfully, with the frequency of cluster category_1 being 154 and the percentage being 43.38%, the frequency of cluster category_2 being 133 and the percentage being 37.465%, and the frequency of cluster category_3 being 68 and the percentage being 19.155%. See Table 2 for details.

The cluster scatter plot only shows the maximum sample size information of 1000. If the sample size is greater than 1000, the random sampling in the whole sample is carried out, and 1000 samples are selected for the scatter plot display. See Figure 2 for details. According to the images and tables of the model, the words can be divided into three categories: simple, moderate and difficult. Cluster 2 is simple, cluster 1 is moderate and cluster 3 is difficult.

3.4 Attributes of words in clustered categories

3.4.1 part of speech

The number of n. and v. is higher in clustering category 1 (moderate), the number of n. and v. is highest in clustering category 2 (simple), and the number of n. and v. is lower in clustering category 3 (difficult). See Figures 3,4,5.

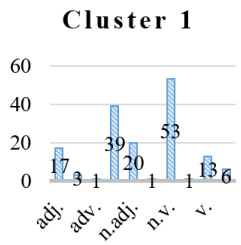

Fig. 3. Clustering category 1

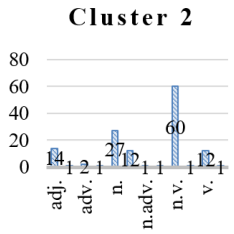

Fig. 4. Clustering category 2

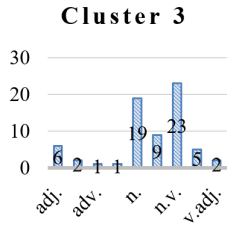

Fig. 5. Clustering category 3

3.4.2 The frequency with which the words appear

The frequency of each category representing words in various situations was obtained by the function in Corpus of Contemporary American English Chart. Due to the large amount of data, this paper identifies one representative word in each category and concludes that in general, difficult words appear less frequently and easy words appear more frequently. Take the frequency of cluster category 1 words as an example, as detailed in Figure 6.

CLICK TO SEE CONTEXT

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	13242	1428	1558	2829	968	1811	1590	2778	280	1393	1561	1808	1761	1489	2244
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	121.1	125.2	124.6	123.1	123.3	122.8
PER MIL	13.34	11.10	12.54	22.09	7.67	15.31	12.61	22.82	2.34	11.50	12.47	14.51	14.31	12.07	18.28
SEE ALL SUB-SECTIONS AT ONCE															

Fig. 6. SQUAD

In summary, from the perspective of speech, the number of n. and v. in each clustering category is the largest, cluster 2 (simple) is the most obvious, and cluster 3 (difficult) is less obvious. From the frequency of words, in general, the frequency of words in cluster 3 (difficult) is small, and the frequency of words in clustering category 2 (simple) is large.

3.5 Other characteristics of solving words

The normality test was based on the S-W test or the K-S test to obtain the results [21]. Variable analysis terms: number of players reporting results on the same day, number of players in difficult mode, 1 try, 2 tries, 3 tries, 4 tries, 5 tries, 6 tries, 7 or more tries (X), and the data are subjected to Shapiro-Wilk (small data sample, general sample size of 5000 or less) or Kolmogorov-Smirnov (large data sample, general sample size of 5000 or more) test to see the significance. See Figure 7,8. Analyze the histogram of the normality test, if the normality plot basically shows a bell shape (high in the middle and low at the ends), it means that the data,

although not absolutely normal, are basically accepted as normally distributed. All the data are normally distributed.

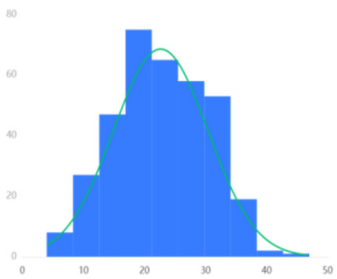

Fig. 7. 3 tries

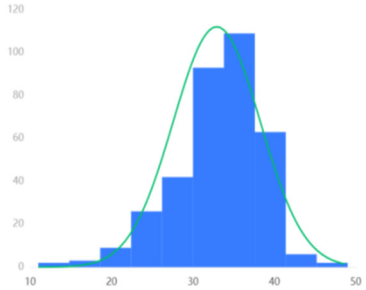

Fig. 8. 4 tries

4 Conclusions

Firstly, the collected data is pre-processed to predict the number of user participation on March 1, 2023 using a gray prediction model, and the interval of the reported results is obtained to be roughly 19121~20218. Secondly, using ridge regression model, according to whether the word has repeated letters and the number of repeated letters of the database is divided into four categories, and code them as 1,2,3,4, and then the table data and the player guess the puzzle into data analysis, can get the future day related percentage prediction, predict the difficulty of the word EERIE. Using the calculation formula of R^2, the value is 1, that is, the model is well fitted. Then, the K-mean clustering model is used to process the relevant percentage of the successful problem solving (1,2,3,4,5,6, X), and obtain three types of simple, moderate and difficult results. Find out the corresponding properties of words in the three clusters, that is, the parts of speech and the frequency of words in various cases. Finally, other characteristics of the words are listed, and the characteristic that the number of daily reported results, the number of participants in the difficult mode, and the percentages of each guess are normally distributed was obtained by analyzing the histogram of normality tests.

Footnotes

Xiaoli Jiang[2*] and Xiaodong Fan[3*] are the corresponding authors of this article.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61976027, 11871117), Department of Education of Liaoning Province (LJKZ1026, LJKZ1030, LJ2019011), Liaoning Natural Foundation Guidance Plan (2019-ZD-0502), Liaoning Revitalization Talents Program (XLYC2008002).

References

[1] Wordle Unlimited emerges as a word-guessing game designed to guess and create new words.[J]. M2 Presswire,2023.

[2] "Wordle-TheNew York Times."TheNewYork Times,2022.Accessed December 13,2022at https://www.nytimes.com/games/wordle/index.html.

[3] Silva N D . Selecting Seed Words for Wordle using Character Statistics[J]. 2022.

[4] Chao-Lin Liu. Using Wordle for learning to design and compare strategies, proceedings of the 2022 IEEE Conference on Games (IEEE CoG 2022). Beijing, China, 21-24 August 2022. (virtual)

[5] Wormley Alexandra S,Cohen Adam B. C-H-E-A-T: Wordle Cheating Is Related to Religiosity and Cultural Tightness.[J]. Perspectives on psychological science : a journal of the Association for Psychological Science,2022.

[6] Short Martin B. Winning Wordle Wisely—or How to Ruin a Fun Little Internet Game with Math[J]. The Mathematical Intelligencer,2022,44(3).

[7] Locshtanov D, B.Wordle is NP-hard [J]. 2022.

[8] Bonthron M. Rank One Approximation as a Strategy for Wordle[J]. arXiv e-prints, 2022.

[9] Bhambri S,Bhattacharjee A,BertsekasD. Reinforcement Learning Methods for Wordle: A POMDP/Adaptive Control Approach[J]. arXiv e-prints, 2022.

[10] Jintao Y ,Xican L ,Shuang C , et al. Grey fuzzy prediction model of soil organic matter content using hyper-spectral data[J]. Grey Systems: Theory and Application,2023,13(2).

[11] Eric M. Extending a word property for twisted Coxeter systems[J]. Advances in Applied Mathematics,2023,145.

[12] Yucheng S ,Huaiyi C ,Xiaomeng S , et al. STG-Net: A COVID-19 prediction network based on multivariate spatio-temporal information[J]. Biomedical Signal Processing and Control,2023,84.

[13] NanYang .The unique role of ridge regression analysis in solving the problem of multicollinearity [J]. Statistics and Decision-making, 2004 (03): 14-15.

[14] PanDu. Application of classified memory strategies in junior High school English vocabulary teaching [D]. Yan'an University, 2021.DOI:10.27438/d.cnki.gyadu. 2021.000426.

[15] Li Y ,Hoang-Le M ,S. K , et al. [J]. Engineering Structures,2023.

[16] Shibaprasad B ,Kanak K ,Robert Č , et al. [J]. Materials,2021.

[17] XinyiLi,Yuzhuo Wu,JijuYang, etc. Analysis of Zedoary oil injection based on real world data [J]. China Journal of New Drugs, 2023,32 (05): 547-552.

[18] GuanghuiCAI,Zhimin Wu. Efficacy and AUC study of several classes of normality tests under the two-stage sequencing set sampling of perfect and non-perfect [J]. Systems Science and Mathematics, 2023,43 (01): 227-243.

[19] Chao Liu , Regression Analysis- -Application of Methods, Data and R, Higher Education Press, 2019.

[20] Saroj,Kavita.Review:study on simple k mean and modified K mean clustering technique[J].International Journal of Computer Science Engineering and Technology,2016,6(7)：279-28

[21] Shuping Zong , Yulan Yao . The statistical distribution of the data was quickly tested using the Q-Q plots and the P-P plots [J]. Statistics and Decision-making, 2010 (20): 2.