Research on Predicting Wordle Results Model

Haotian Lin[1,a,*,#], Xiaoyu Wei[2,#], Jiannan Lin[1,#], Wendan Liao[1,#]

[1]School of Ocean Information Engineering, Jimei University, Xiamen, 361021, China
[2]School of Science, Jimei University, Xiamen, 361021, China
[a]linhaotian@jmu.edu.cn
[*]Corresponding author
[#]These authors contributed equally

Abstract: *The reported scores are collected from Twettier only, so the discussion of the prediction of the change of time and the total number of reports is one-sided because it does not consider most users who do not share the game scores. For this problem, the SIRS model can be used to predict the future total number of reports. This model helps combine the process of real-life viral transmission and the changing pattern of the total number of reports curve, resulting in a prediction interval of [17645, 2469z4]. To investigate whether word attributes affect the percentage of reported scores in the complex mode, we use the feature importance scores ranked by the RReliefF algorithm, which builds a Gaussian regression model to rate the influence of word attributes on reported scores in the difficult mode. We found that two word attributes, frequency of letter use and frequency of word use, had the most significant effect on reported scores in the difficult mode, while the type of letters contained in the word had a less significant effect. we use an integrated prediction model and a Gaussian process regression prediction model; first, the data of the percentage of attempts are downscaled into three indicators by principal component analysis, and then the three parameters of word attributes (frequency of word use, cumulative frequency of letters, and the number of letter repetitions) and the three indicators of word difficulty are used as the training set to train the prediction model, Second, we use the prediction model for the example in the question of 2023 The prediction model to predict the word EERIE on March 1, 2023, as the example given in the question. Finally, the three indicators of word difficulty are derived. Because the distribution of the attempted percentage data is normally distributed, after mathematical backpropagation, the final seven broadcast percentages of the word EERIE were obtained: 0, 11, 36, 27, 17, 8, 1.*

Keywords: *SIRS model; Integrated prediction model; Gaussian process regression prediction model*

1. Introduction

1.1 Problem Background

Wordle is the current daily scrabble game offered by the New York Times. The game provides one word per day for people to guess and keeps track of your attempts. Players can try to guess a five-letter word up to six times per game and will receive feedback for each guess. For this version, each guess must be the English word identified by the game. The game has become so popular that versions already support more than 60 languages.

1.2 Our Work

Based on the analysis of the problem, we propose the model framework shown in figure 1 is mainly composed of the following parts:

Data processing: By observing the data, the data that do not conform to the rules of the game and the abnormal data are removed in this paper.

Model Modeling: Analogous to the infectious disease model SIRS, a wordle transmission model was established.

Dimensionality reduction of indicators: To better discover the relationship between the number of attempts and word difficulty, a principal component analysis was performed to downscale the seven attempt count indicators into three difficulty indicators , as shown in Figure 2.

Training Model: Prediction using machine learning models such as integrated prediction models and

Gaussian process regression prediction models.

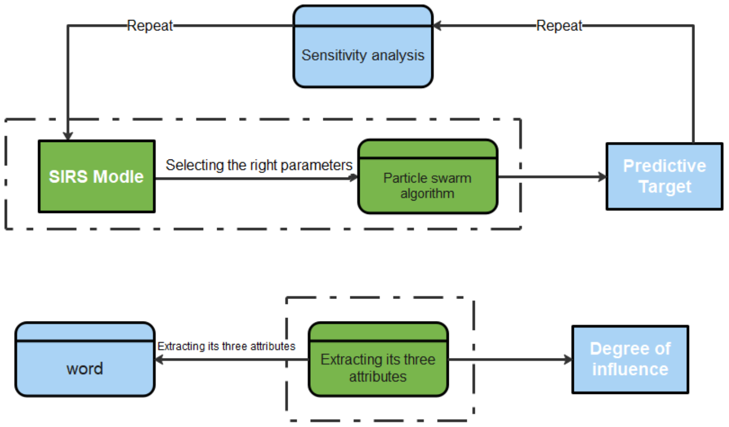

Figure 1: Problem 1

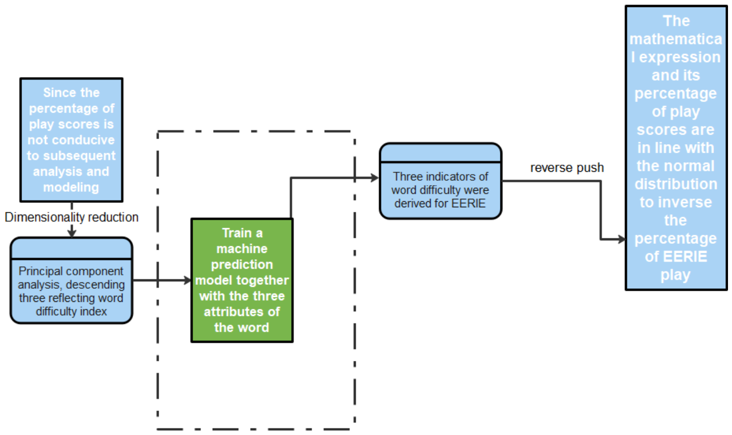
Since the percentage of play scores is not conducive to subsequent analysis and modeling
Dimensionality reduction
Principal component analysis, descending three reflecting word difficulty index
Train a machine prediction model together with the three attributes of the word
Three indicators of word difficulty were derived for EERIE
reverse push
The mathematica l expression and its percentage of play scores are in line with the normal distribution to inverse the percentage of EERIE play

Figure 2: Problem 2

2. Preparation of the Models

2.1 Assumptions and Justifications

Assumption 1: No major international events recently made the wordle topic relatively less hot.

Reason: This assumption is made to ensure that the impact of words on the number of reported results

Assumption 2: The official dataset given is reliable

Reason: this assumption is made to ensure the accuracy of the model solution

2.2 Notations

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

Table 1: Notations used in this paper

Symbol	Description	Unit
α	The transfer rate from recovered to susceptible persons	\
β	Infection rate	\
γ	Recovery rate	\
ω_1	Indicators reflecting the simplicity of words	\
ω_2	Indicators reflecting the commonness of words	\
ω_3	Indicators reflecting the degree of difficulty of the word	\
N	Total number of people	\

3. SIRS prediction model

Since not all users share their scores on Twitter, the data of reported results fluctuates widely from day to day. The relationship between the time of day and the number of reports per day is one-sided, so we use the SIRS model to predict the number of reported results in the future in the context of life.

Assuming an initial total population of N=S+I+R and not taking into account disease-related deaths and natural-born deaths, S refers to the Susceptible, i.e., the healthy, who may become infected at a transfer rate of β. I refer to the Infective, i.e., the sick, who may become infected at a transfer rate of γ. R refers to the Recovered, who may lose antibodies over time. The number of infected persons per unit of time is, therefore, $\beta=\frac{SI}{N'}$, where N' means The effective population in the infectious disease system, i.e., excluding those who leave the system (e.g., those who have acquired antibodies, die, or are effectively quarantined), is N'=S+I. Assuming that the recovery rate from infected to recovered is γ, the corresponding differential equation is Eqs. (1-3), and the state transfer diagram is shown in Fig. 3.

$$\frac{dS}{dt} = -\beta\frac{S \times I}{N} \tag{1}$$

$$\frac{dI}{dt} = \beta\frac{S \times I}{N} - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I - \alpha I \tag{3}$$

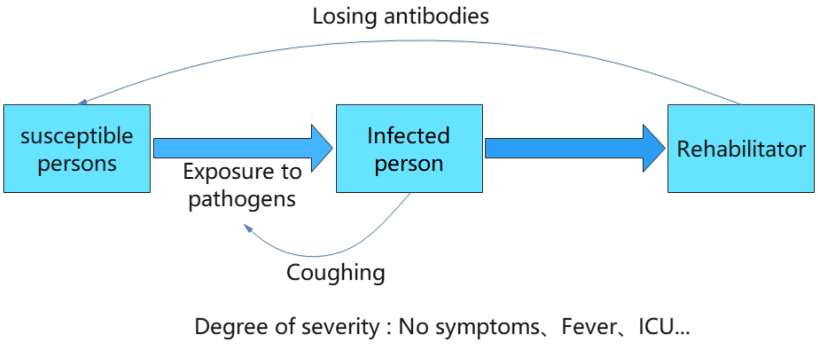

Figure 3: Transfer of infectious disease status

The infected player is the player who completes the Wordle game and shares it on Twitter. The player who shares the score of the Wordle game has a beta probability of attracting new players to play the game on the social platform. Players who play the game alone without sharing their scores do not have access to the system and lose the possibility of attracting new players to play the game.

In contrast, infected recovered persons carry antibodies, and such persons are temporarily immune to viral infections and have an alpha probability of losing antibodies over time and becoming susceptible to infection, thus becoming at risk of infection again. Recovered persons are analogous to users who are temporarily indifferent to Wordle games or temporarily tired of playing them and have a certain probability of regaining their freshness to the game over time.

Infection-prone users are those who share the game and have a certain probability of temporarily losing interest in the game due to various uncertain factors (e.g., work, novelty, environment⋯ ⋯) and becoming recovered users. Thus, a closed loop of game population change is formed, as shown in Figure 4 below.

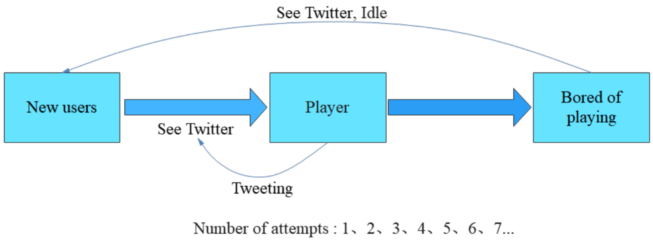

Figure 4: Game propagation state transfer

3.1 Particle swarm algorithm to search for optimal parameters

Since the three parameters of the SIRS function (infection rate β, recovery rate γ, and transfer rate α) are very important to the prediction results of the model, adjusting the variation of the parameters to obtain the best prediction results is a key step for the model to be able to predict accurately. Therefore, we adopt the particle swarm algorithm in a heuristic algorithm to obtain the best parameters efficiently and accurately so that the SSE (Sum of Squares of Error) of the predicted curve and the actual curve is minimized.

The SSE of the predicted curve and the actual curve is taken as the objective function of the particle swarm algorithm, and circular iterations search the smallest SSE, and the SSE is obtained after more than 200 iterations, which is 6.8 in the tenth order, as shown in Figure 5 below.

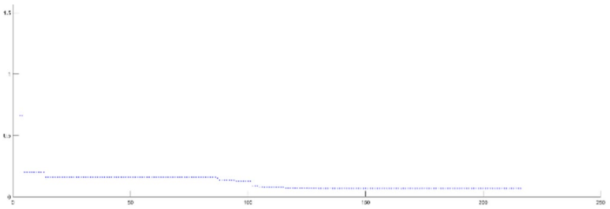

Figure 5: Number of iterations

To better converge to the natural curve and predict more accurate and scientific data, and make the models contain more practical meaning. The curve of the total number of reported results was divided into three segments, the first of which can be called the period of surge in the popularity of the Wordle game, probably due to the simple and interesting content of the Wordle game, which received a lot of retweets and shares from Twitter users and participation from YOUTUBE bloggers. The second period is the period when the game's popularity plummeted, probably due to the monotony of the game's content and the lack of new content updates, which caused users to lose their sense of freshness. The third period is a period of balance after the game has been launched for a long time because the game has lost a lot of players, but there is still a small influx of users, and some loyal players still participate in it.

So according to the combination of the real-life phenomenon, the SIRS model will be a combination of three time periods, and the infection rate in each time period also presents different due to the game heat and the infected population base , as shown in Figure 6.

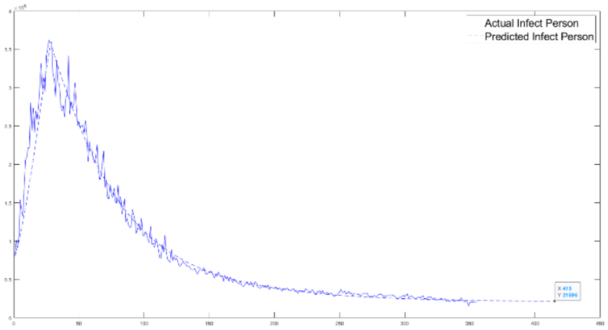

Figure 6: Forecast and actual comparison chart

The model predicts that the total number of results reported on 2023/3/1 is 21696, but the prediction interval is [17645,24694] considering the effect of unknown interference terms and repeating the particle algorithm searching for different SIRS model parameters several times.

If we assume that the audience surface of this game group is 1,000,000 people, β_1=0.1195 in time period one (2022/1/7~2022/2/6), β_2=0.0362 in time period two (2022/2/7~2022/10/14), and β_3=0.8540 in time period three (2022/10/15~) in all time periods α=0.0102 and γ=0.03221.

β_3 is much larger than β_1 and β_2=0.0362 due to the large number base and the fact that the vast majority of the remaining users are loyal users when in period three in the model assumptions, so these users share a higher proportion of the score to Twitter.

3.2 Determination of word properties

To accurately locate the word attributes of an English word, three aspects will be considered under normal circumstances: spelling, pronunciation, and semantics, where word spelling is related to the use and permutation of the letters that make up the smallest unit, word pronunciation is related to the pronunciation of vowels and consonants, and semantics will trigger the use of words with various scenarios and cause differences in the frequency of use of different words. Based on our preliminary analysis of the game wordle and the psychological thinking of the game audience, we decided to choose the frequency of use of the letters that make up the word, the minimum types of letters that make up the word, and the ranking of the frequency of word use in various scenarios to describe the word properties statistically.

In the process of word attribute dramatization, we visited the authoritative website BNC [1] (British National Corpus) to obtain the ranking of word frequency, and at the same time, we obtained the frequency table of 26 letters by consulting the related websites and conducted the statistics of letter composition types for the daily words in the wordle game given in the annex. The above three indicators were made into a word attribute table corresponding to each word.

Because the three data indicators are not uniform, we have normalized the three data indicators.

$$z_i = \frac{x_i - \bar{x}}{\sigma_x} \tag{4}$$

Where σ_x denotes the variance of the sample; $\overline{\mathrm{x}}$ denotes the mean of the sample.

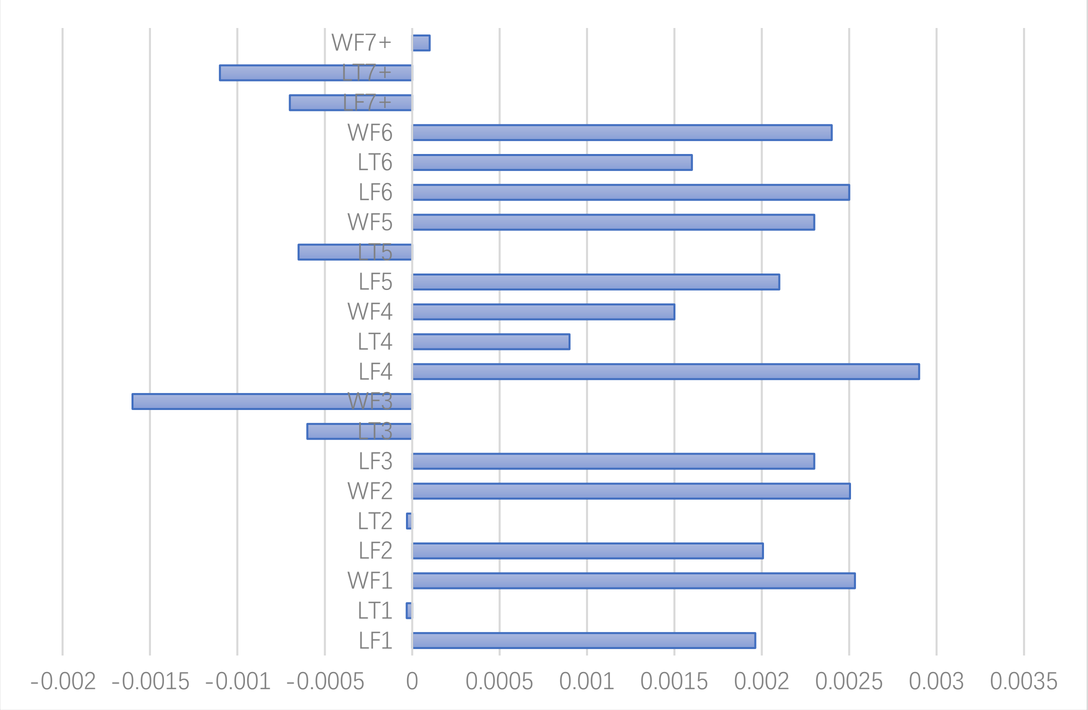

Figure 7: Feature Importance Scores Ranked Using the RreliefF Algorithm

To investigate whether the attributes of words affect the percentage of reported scores in the complex mode, we use the feature importance scores ranked by the RReliefF algorithm to rate the influence of word attributes on reported scores in the difficult mode by building a Gaussian regression model. As can be seen from the graphs, the two-word attributes of letter use frequency and word use frequency have the most obvious influence on the reported scores in the difficult mode, such as LF1, WF1, LF2, WF2, etc. in the graphs, whose feature importance scores are more prominent among the three attributes; while the type of letters contained in words has a less obvious influence on the reported scores in the difficult mode, as in Figure 7.

Where LF: is the frequency of letter use, LT is the type of letters contained in the word, and WF is the frequency of word use. Also, some outliers in the chart are difficult to interpret, such as the negative feature importance score of WF3 and the significant decrease of the three indicators when the game is played for the seventh time, which is an interesting phenomenon, probably because the information queried on the website does not fully represent the people in their daily word usage, which may be due to the fact that the sources of information on the frequency of word usage included in the UK national website are paper-based materials such as books, journals, papers, etc., and there is not much crossover in the collection of data sources.

4. Predictive modeling of the distribution of reporting results

4.1 Data Description

Table 2: Variance of various attempts

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
Variance	0.611	16.57	60.38	28.59	35.27	38.42	16.94

According to the rules of the wordle game, the number of attempts used by players who share their game results on Twitter to experience the game falls in the range of seven or more attempts at a time, so combined with the percentage of player game attempts in the difficult mode given in the annex , as shown in Table 2. We can divide the difficulty of the set words each day according to the number of attempts made by players. First, by analyzing the data, we can conclude that the number of attempts of players who finished the game and uploaded tweets were mostly 3, 4, and 5, so this makes the percentage number of attempts 3, 4, and 5 relatively serious and with a large variance, i.e., the data fluctuates strongly. If the data were used directly without data processing, these more volatile data would dominate the neural network regression prediction model. In contrast, in the actual game, the best indicator of word difficulty is the percentage of people who complete the game in 7 attempts and in 1 and 2 attempts. Therefore, using principal component analysis to reduce the dimensionality of the data as in Equation (5-6)

$$\omega_1 = \sum_{i=1}^{3} \frac{F_i - F_{i\_mean}}{F_i + F_{i\_mean}} \tag{5}$$

$$\omega_2 = \sum_{i=3}^{5} \frac{F_i - F_{i\_mean}}{F_i + F_{i\_mean}} \tag{6}$$

$$\omega_3 = \sum_{i=5}^{6} \frac{F_i - F_{i\_mean}}{F_i + F_{i\_mean}} + 2\,\frac{F_{7+} - F_{i\_mean}}{F_{7+} + F_{i\_mean}} \tag{7}$$

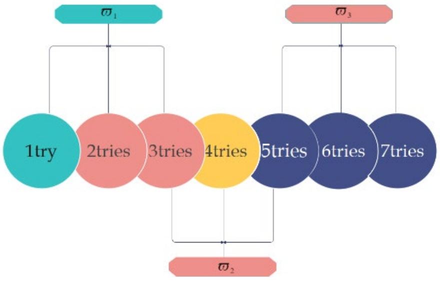

Figure 8: Data dimensionality reduction processing

From Figure 8, ω_1 is an indicator reflecting the simplicity of the word, and the greater the difficulty of the word, the greater its value. ω_2 is an indicator reflecting the commonness of the word, and the greater its value, the more common the difficulty of the word? ω_3 is an indicator reflecting the difficulty of the word, and the greater its value, the more difficult the word?

$F_{i_mean}$is the average number of people who pass the game after I attempt. ω_1, ω_2, ω_3 are the indicators after dimensionality reduction, as shown in the figure.

4.2 Integrated prediction model with Gaussian process regression prediction model[2]

4.2.1 Integrated prediction model

The results of a single model are somewhat one-sided, and combining the results of multiple models with some strategy can usually achieve better generalization performance than a single model. This combination of multiple models to obtain better generalization performance is called an integrated model, and a present model is an integrated approach using the bagging method [3] by combining multiple models, such as decision trees, gradient ascent trees, etc., in parallel to make them as independent of each other as possible. Used in classification, a strong classifier is obtained, which can effectively reduce the variance, and used in regression to reduce the mean square error. As shown in Figure 9.

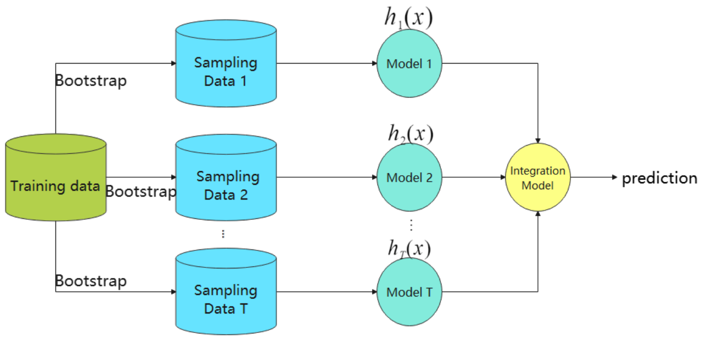

Figure 9: Integration model based on the bagging method

4.2.2 Gaussian process regression prediction model (GPR) [4,5]

Gaussian process regression GPR is a nonparametric and kernel-based probabilistic model. A GPR model introduces latent variables to explain the response through a Gaussian process (GP). GPR introduces a latent variable f(x_i) that obeys a Gaussian distribution for each sample x, and all $f(x_i)$ together are a set of random variables that together have a joint Gaussian distribution. Thus a GP has a mean function m(x) (in GPR, m(x) is generally equal to 0) and a covariance function k(x, x'). The Gaussian process regression is modeled as follows.

$$h(x)^T\beta + f(x) \tag{8}$$

h(x) is a set of basis functions, which convert the d-dimensional original eigenvector into a new p-dimensional eigenvector to be consistent with the dimensionality of the right-hand side $f(x)$, and β is the reference vector of p*1

The method of model training is maximum likelihood estimation, during which three main coefficients need to be estimated: β, σ^2 and θ (hyperparameters of the kernel) viz.

$$\hat{\beta},\hat{\theta},\hat{\sigma}^2 = arg_{\beta\theta\sigma^2}maxlogP(y|X,\beta,\theta,\sigma^2) \tag{9}$$

When applying the model for prediction, again, according to the Bayesian formula, a new sample of f* is included in f. This aggregate conforms to the following joint probability distribution.

$$p(y_{new}|y,X,x_{new}) = \frac{p(y_{new},y|X,x_{new})}{p(y|X,x_{new})} \tag{10}$$

$$\begin{bmatrix} \vec{f} \\ f^* \end{bmatrix} \sim N\left[\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{11}\, K_{12}\, K_{13}\, K_{14} \\ K_{21}\, K_{22}\, K_{23}\, K_{24} \\ K_{31}\, K_{32}\, K_{33}\, K_{34} \\ K_{41}\, K_{42}\, K_{43}\, K_{44} \end{bmatrix}\right] \tag{11}$$

Where, $\vec{f} \sim N(0, k)$, $\vec{f} = [f_1, f_2, f_3]^T$, $f^* \in R$

Finally, it can be obtained that:

$$f^* \sim N(u^*, \sigma^*) \tag{12}$$

$$u^* = k_*^T k^{-1} \vec{f} \tag{13}$$

$$\sigma^* = -k_*^T k^{-1} k_* + k_{**} \tag{14}$$

4.3 Prediction results

The integrated prediction model and Gaussian process regression prediction model were used to first downscale the seven words attempted data into three indicators by principal component analysis, and then the three parameters of word attributes (frequency of word use, letter accumulation frequency words, and the number of letter repetitions) and the three indicators of word difficulty were used as the training set to train the prediction model, followed by the prediction of the word EERIE on March 1, 2023, as exemplified in the question The three-word attributes of EERIE are:

(1)Weighted sum of the frequency of letter usage: 0.5216

(2)Type of letters contained in the word: 3

(3)Frequency ranking of words used: 15351

The three-word attributes of EERIE were predicted by feeding them into a Gaussian process regression prediction model, which finally yielded its three indicators of word difficulty as: -0.18154769, -0.31976692, and -1.950707116. Because the indicators obeyed a normal distribution, the seven broadcast percentages of the word EERIE were finally derived by the backward process of equations (5), (6), and (7): 0, 11, 36, 27, 17, 8, 1.

References

[1] http://www.wordcount.org/main.php

[2] C. Guo, J. Zhang and M. Yang, "Weighted Gaussian Process Regression for Single Image Super-resolution Based on Randomized Sample Clustering and Augmentation," 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2021, pp. 744-750.

[3] M. Ward, K. Malmsten, H. Salamy and C. H. Min, "Data Balanced Bagging Ensemble of Convolutional- LSTM Neural Networks for Time Series Data Classification with an Imbalanced Dataset," 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Korea, 2021, pp. 1-5.

[4] Jiang Y , Yang Y , Wu Q ,et al. Research on Predicting the Short-term Output of Photovoltaic (PV) Based on Extreme Learning Machine Model and Improved Similar Day[C]//2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia).IEEE, 2019.DOI:10.1109/ISGT-Asia.2019.8881620.

[5] Chen Xiaokang, Tu Xuan, Xu Weidong, Xie Runzhong. Tool Life Prediction Based on Bagging Integrated Gaussian Process Regression Model [J]. Manufacturing Technology and Machine Tool, 2020 (12): 110-115+121.