Word difficulty classification prediction based on ARIMA and DBSCAN

Xilun Li[1, *], Lingyu Sun[2], Jigang Li[1]

[1]School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian, China, 116025

[2]School of Finance, Dongbei University of Finance and Economics, Dalian, China, 116025

*Corresponding author: lxl17604016313@126.com

Abstract. Wordle is an online word-guessing game, where players are given a five-letter word each day and are asked to guess the word within six tries. Since the launch of the game, it has been loved by a wide range of players. To classify the experiment data and point the 'EERIE' at an appropriate label, the article uses unsupervised learning (with DBSCAN and ARIMA), cluster "sum(x)" and "rank(y)" and obtain the values (5.18185152, 8.83924243). Meanwhile, according to the image 'The word positions by DBSCAN', we define the 'EERIE' in moderately easy.

Keywords: Wordle, DBSCAN, ARIMA, Classification.

1. Introduction

The web-based word game Wordle has captured thousands and millions of participants, since its introduction in October of 2021. In January 2022, *The New York Times* bought the puzzle company, pushing the game into an even larger audience Each day, players have six attempts to guess the five-letter word of the day such that previous guesses provide six hints. There is only one word for all players each day.

Players can judge their answer by the color of the squares. Black squares don't appear in the final word. Yellow squares are the letters which appear in the final word but in the wrong space. Green squares are the right letters in the right space. Figure 1 is an example.

Figure 1. An example of a wordle game

Wordle's popularity is owed to multiple factors, like the ease in sharing results and the daily nature of the game, as well as its simplicity. It can be seen to its fullest extent on Twitter, where millions of users share their results every day by pasting their scores via the website's built-in feature. From the daily results, it is common that different people will have different problems: Can we predict the number of participants? We wonder if a player's score will be influenced by the attributes of the word or the mode of the game. Besides, what we are thirsty for is the optimum strategy to win Wordle?

In recent years, many articles have talked about thought-provoking issues behind the game. It can be seen that cheating behavior is negatively related to religiosity and cultural tightness by using data from Google Trends and Twitter to explore correlates of cheating on Wordle. (Alexandra S. Wormley & Adam B. Cohen,2022)[1] There is a study that uses character statistics of five-letter words to

determine the best three starting words. (Nisansa de Silva,2022)[2] Some of the articles talk about the method of winning the game with the least number of tries. Using UPPAAL STRATEGO, the expected number of guesses is reduced from a baseline of 7.67 to 4.40 using 1 million training episodes. (Peter G. Jensen, Kim G. Larsen & Marius Mikučionis, 2022) [3]Using a new reinforcement learning method, Partially Observable Markov Decision Process (POMDP) A good strategy is come up by presenting the resultant Bellman Equation, and using Optimal Classification Trees with Hyperplanes. (D Bertsimas & A Paskov,2022)[4] As for the word difficulty classification, the words are translated into 5-letter Turkish words. After removing the immoral words, through Yandex Wordstat, frequently searched words are grouped as easy words and rarely searched words are grouped as difficult words. (KÜÇÜK DB, ÇOBAN S & ŞENYER N, 2023)[5]

Although there really exists articles to talk about the word difficulty classification, the articles of the area are not sufficient compared with the strategy of winning the game. Based on the situation, we want to classify the words with their difficulty.

The structure of the article is shown as follows:

First, explain the law of the ARIMA and DBSCAN. Then, use the algorithms to cluster the word and forecast the difficulty of the future word “EERIE”. Finally, based on the results, we will give some advice to the creators of the game.

The marginal contribution of this paper lies in: (1) Analyze the game from a new perspective, the classification of the everyday given word. (2) Unsupervised machine learning (clustering algorithms with clear values) makes it easy to identify the category of predicted value "EERIE". (3) Combined with the reality of United States, we will give some advice.

2. The basic fundamental of ARIMA and DBSCAN

2.1. The structure of ARIMA

The Autoregressive Integrated Moving Average model is widely used for time series analysis and forecasting with the advantages for processing linear time series. It fuses two models: autoregressive (AR) and moving average (MA) by regressing the lagged value of dependent variable and its random errors. The intergroup variables have a dependence relationship, including their own change law as well as external factors. With high prediction accuracy, this method is helpful to explain the law of prediction change.

The ARIMA (*p*,*d*,*q*) model can be expressed as:

$$\left(1-\sum_{i=1}^{p}\phi_i B^i\right)\left(1-B\right)^d x_i=\left(1+\sum_{i=1}^{q}\theta_i B^i\right)e_i \tag{1}$$

In this formula, *p* is the order of the autoregressive, *q* is the order of the moving average, and *d* is the number of differences, which makes the time series stationary. In addition, e_i is the random error of the *i* moment, Φ_i is the coefficient of the autoregressive part, θ_i is the coefficient of the moving average part, and *B* is the lagging operator.

2.2. The structure of DBSCAN

Text clustering is one of the core problems in the field of text mining and information retrieval, so the research of text clustering algorithm is a research hotspot in this field. [6]Text clustering is the process of unsupervised division of text into clusters based on the similarity between texts.[7] It has been widely used in many fields, like literature collation, library management and even the identification characters on the clutch flywheel[8] etc.

According to different clustering ideas, clustering algorithms can be roughly divided into four categories: division-based, hierarchical-based, density-based and graph theory-based. And each type of clustering algorithm contains a variety of algorithms, and their derived improvement methods, and

there are many mutual references between various algorithms. At present, in the field of text clustering, the more widely used clustering algorithms include division-based and density-based[9].

DBSCAN (Density based spatial clustering of applications with noise) is a density-based clustering algorithm. The main goal of the algorithm is to require less domain knowledge to determine input parameters than division-based clustering methods and hierarchical clustering methods; Discover clusters of arbitrary shapes; Better efficiency on large-scale databases. The main advantage of the DBSCAN algorithm is that it can divide a sufficiently dense area into clusters and is resistant to noise interference, and can also find arbitrarily shaped clusters in a noisy spatial database[10]. The result of clustering under DBSCAN algorithm is shown in figure2:

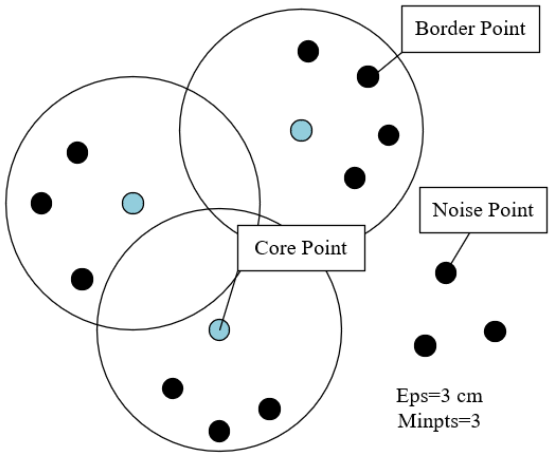


Figure 2. The diagram of the DBSCAN algorithm

Here, to further explain the algorithm, some related concepts are added: Eps is the maximum radius of the neighborhood. Minpts is the minimum number of proximity points within the neighborhood. The core point is a point whose neighborhood with a maximum radius of Eps, is density-reachable with at least Minpts points. The points in the neighborhood are the border points. The others are the noise points. The specific flow of DBSCAN algorithm is shown in figure3:

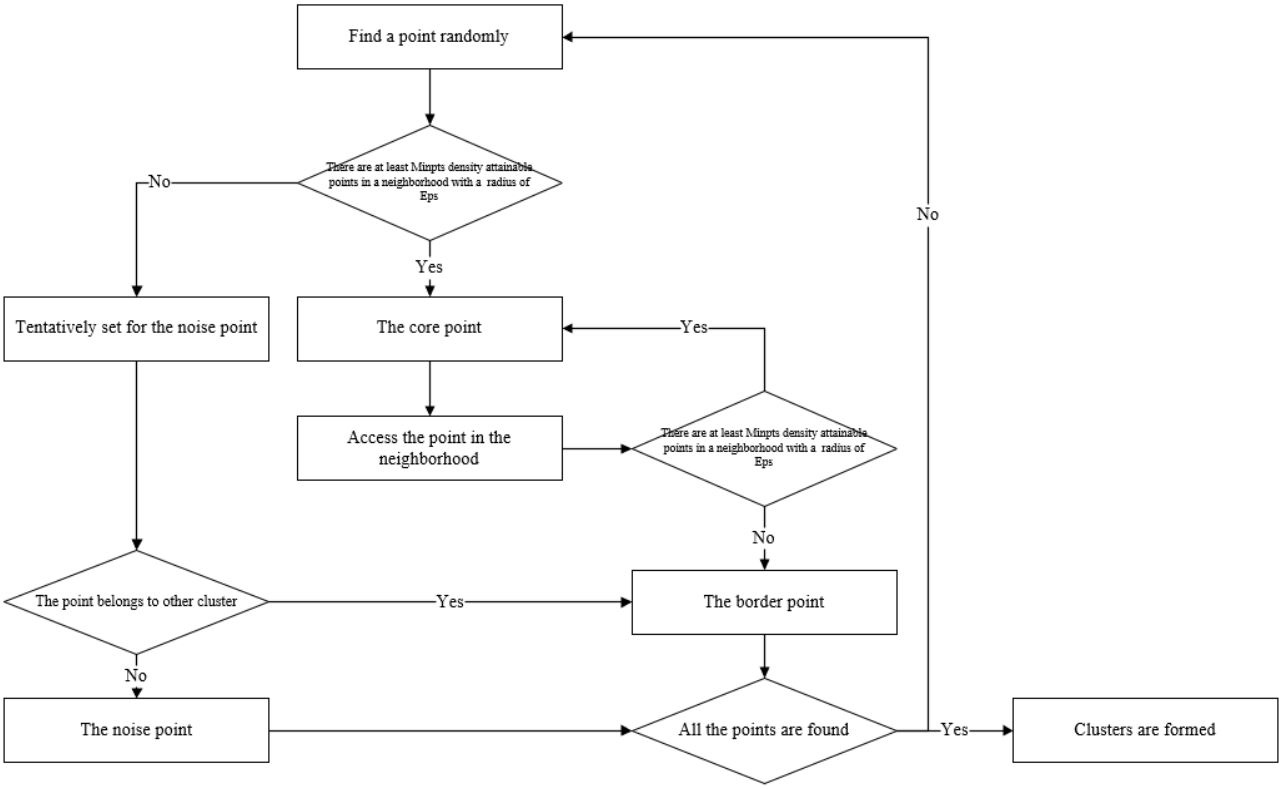
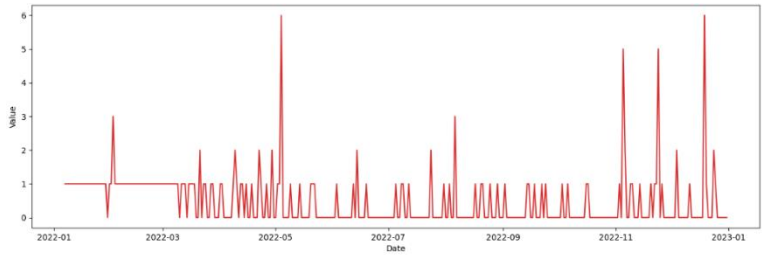


Figure 3. The flow chart of the DBSCAN algorithm

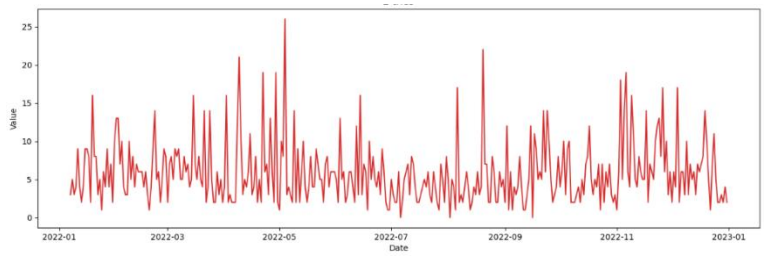
3. Results

3.1. Descriptive Statistical Analysis

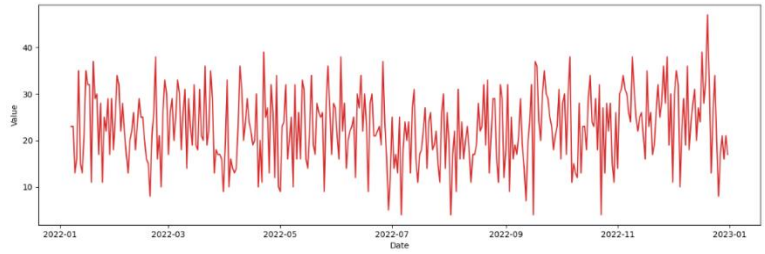
To study the attribution of the attempts every day, we draw the line charts of the percentage of different attempts per day. Figure 4 shows the number of tries, which ranges from 1 to 7 and more.



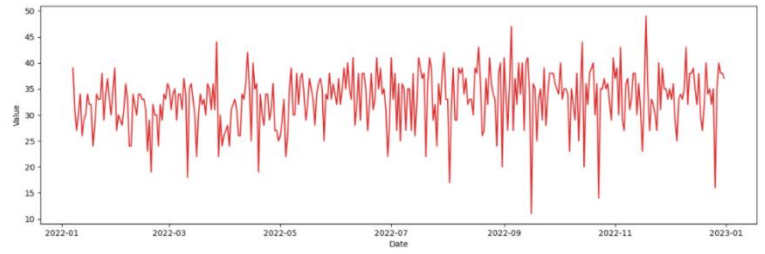
(a) 1 try



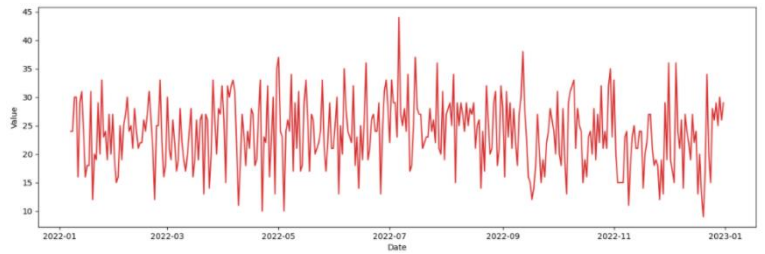
(b) 2 tries



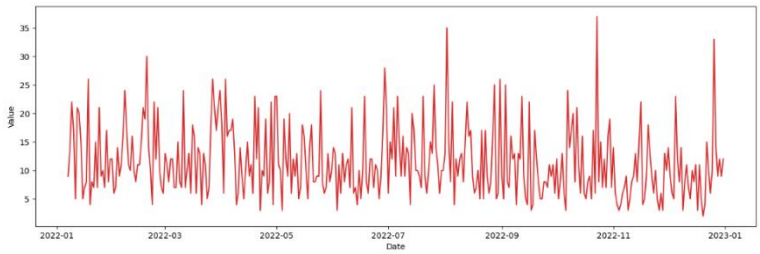
(c) 3 tries



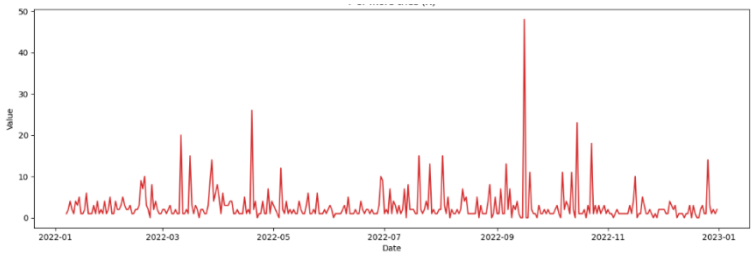
(d) 4 tries



(e) 5 tries



(f) 6 tries



(g) 7 tries and more

Figure 4. The attribution of the number of the everyday attempts

From the above figures, we find that the percentage of players with 3,4 and 5 attempts is higher than other attempts, which also reflects the design of the game to a certain extent, the design of the game is neither immediately guessed by the player, nor will the player be unable to guess all the time, making the game interesting, not too boring to guess easily or too difficult to solve. However, sometimes there also exist some exception. So, we further classify those words.

3.2. Identify the attribute of the word

First, the data is preprocessed to obtain two clustering indicators "sum" and "rank". "Sum" represents the percentage of the user trying to guess the word once and twice. "Rank" represents the percentage of players who participated in hard mode

The points in figure 5 represent rank and sum, respectively, and several clusters can be quickly found, and other points that are not in the circle may be anomalous points.

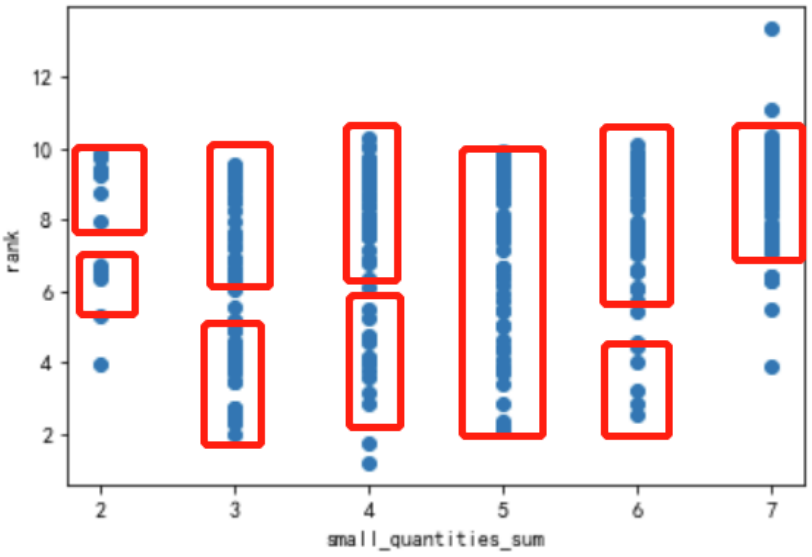


Figure 5. Cluster results

Next, the dataset is validated using density clustering. We save the results under different parameter combinations to find more reasonable clustering results. The optimal combination of parameters is shown in figure 6.

	eps	min_samples	n_clusters	outlines	stats
136	0.851	3	9	4	[76 68 65 64 59 8 6 4 2]

Figure 6. Optimal combination of parameters

As shown in the figure 6, the optimal parameter combination is filtered, and the EPS (The radius of the neighborhood of the core point) is 0.851 and the parameter value is Minpts 3 (because the number of abnormal points under this parameter combination is reasonable). Next, using the parameter combination obtained above, the density clustering model in figure 7 is constructed to realize the clustering of the original dataset. Points of different colors are different clusters, and outliers are anomalies.

Figure 7. Two-iteration density clustering model

Through two iterations, it can be seen that there are two outliers with very large errors: "eject", "study", delete the anomaly "eject", "study", re-experiment, obtain clustering results, and find that the remaining two anomalies should be "piney" and "wrung". The coordinates of the new word "EERIE" predicted by the ARIMA model in this figure are: (5.18185152, 8.83924243), and the position in figure 8:

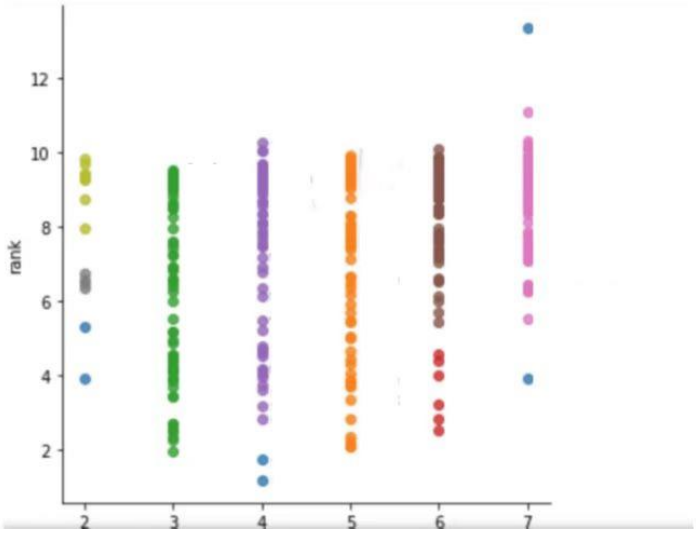


Figure 8. The word positions by DBSCAN

This clustering effect predicts that the words are moderately easy, which is very consistent with the previous conclusion of supervised machine learning (supervised machine learning prediction is "B")

4. Conclusions

One possibility of the logic of the propositional person is to rely on an electronic dictionary with all five words randomly selected one a day as an unexpected rule such as the word guessing game of the day. Then everyone can only guess arbitrarily in this situation.

It is recommended to reveal some rules and logic of the questions in the next few competitions to make it easier for everyone to predict words. It can be difficult for a while, simple for a while. The common words in different places are slightly different, and some words commonly used by everyone are suggested.

References

[1] Alexandra S Wormley, Adam B Cohen. C-H-E-A-T: Wordle Cheating Is Related to Religiosity and Cultural Tightness [J]. Perspectives on Psychological Science, 2022: 17456916221113759.

[2] De Silva N. Selecting seed words for wordle using character statistics[J]. arXiv preprint arXiv:2202.03457, 2022.

[3] Jensen P G, Larsen K G, Mikučionis M. Playing Wordle with Uppaal Stratego[M]//A Journey from Process Algebra via Timed Automata to Model Learning: Essays Dedicated to Frits Vaandrager on the Occasion of His 60th Birthday. Cham: Springer Nature Switzerland, 2022: 283-305.

[4] Bertsimas D, Paskov A. An Exact and Interpretable Solution to Wordle[J]. Available at URL. (Accessed: 14 November 2022)

[5] KÜÇÜK DB, ÇOBAN S, ŞENYER N. Leveled Wordle Game[J]. International Journal of Advanced Natural Science. 2023, 7(2): 11-15.

[6] Liu Yingying, Liu Peiyu, Wang Zhihao, Li Qingqing, Zhu Zhenfang. A text clustering algorithm based on density peak discovery[J]. Journal of Shandong University(Science Edition),2016,51(01):65-70.

[7] Wu Jinchi, Yu Weijie. A study on text clustering based on knowledge base semantics[J]. Journal of Intelligence,2021,40(05):156-164.

[8] Chen Suxin, Liu Wei, Wan Shouxiang. Multi-line character detection and recognition of clutch flywheel based on machine vision[J].Combined Machine Tool and Automatic Processing Technology,2022(07):127-129+133.

[9] Sun Mingxi & Liu Chunqi. Research on the discovery of hot topics based on DBSCAN algorithm and intersentence relationship. Library and information work[J]. 2017,61(12)

[10] Tang Miaojia and Zhang Yong. Research on tobacco Internet illegal data based on DBSCAN algorithm. Modern computers [J]. 2022,28(22),52-55