Wordle word difficulty classification based on K-means

Chen Xiong*, Xinbo Yang#, Jiahui Zhang#

Xian Jiaotong University City College, Xi'an, Shaanxi Province 710018, China

* Corresponding Author Email: 18229046784@163.com

#These authors contributed equally

Abstract. Wordle is a popular word-guessing game, and the analysis of Wordle games will play an important role in updating its iterations. In this paper, a word difficulty classification model based on K-means cluster analysis is developed. Clustering the dataset of the number of times required for word guessing, the words can be classified into three categories according to their difficulty: hard, medium and easy, with the corresponding labels of 3, 2 and 1. The attributes of the words in each category were counted and analyzed to arrive at the following: 1) The more common the word, the less difficult it was. 2) The more repeated letters in the word, the more difficult it was. 3) The experiment substitutes the example word EERIE into the model and the difficulty classification result is medium. The profile coefficient index of the model is 0.372. The cluster analysis method applied in this paper has good training results and is suitable as technical support for Wordle game analysis.

Keywords: Wordle Games, K-means, Difficulty Classification, Word Property.

1. Introduction

Wordle is a word-guessing game offered by the New York Times in which players try to guess a 5-letter word in 6 or fewer attempts. Each guess is met with feedback and each guess must also be an actual English word. The version of the game supports more than 60 languages and is popular with a large number of players.

In recent years, with the continuous progress of machine learning technology, clustering analysis has begun to be applied to a variety of fields such as text categorization, difficulty categorization, and marketing. The K-means algorithm is an extremely classical clustering algorithm, which is widely used because of its easy to understand ideas, easy to operate and excellent clustering results.

Inability to handle various data types is a problem with K-means algorithm, Mohiuddin Ahmed [1] in order to solve this problem by structured K-means and comparing the experimental analysis of various datasets and in-depth study of different K-means clustering algorithms from which their effectiveness is explored. Transformer architecture originated from natural language processing and consists of self-attention and cross-attention, existing converter-based vision models borrowed ideas from NLP and ignored the distinction between language and images, Qihang Yu [2] proposed to formulate cross-attention learning as a clustering process and developed a K-means mask for segmentation tasks, which improves the performance of the technique. Clustering algorithms need to specify the number of clusters K. D T Pham [3] explored the effect of choosing different K on the results and proposed a new method of assisted selection for better clustering performance. K-means. There are many versions of the K-means algorithm, different versions of the K-means algorithm by performing some preprocessing steps or reducing the number of iterations to achieve the effect of reducing the processing time, Ioan-Daniel Borlea [4] proposed the use of supervised learning algorithms on the generated clustering results, which effectively improves the quality of the clusters generated by the K-means algorithm. K-means can be used to construct GIS based mineral prospect maps, in order to classify spatial features into meaningful clusters, Reza Ghezelbash [5] added genetic algorithms to the K-means algorithm in order to automatically select optimized cluster centroids and improve performance accuracy. The distance measure affects the execution time of the K-means algorithm and the number of clusters created, Kamrul Hasan [6] et al. evaluated the K-means

clustering algorithm using three different mathematical metrics, and the results showed that in most of the cases the clustering was best with the implementation of the Manhattan distance measure metric. For distance measure, Mustafa Jahangoshai Rezaee [7] proposed a game-based K-means algorithm, which is to cluster the data using the bargaining game model in K-means algorithm. The clustering centers compete with each other and keep changing their positions to attract more similar targets or entities into their clusters, which is superior and efficient compared to the original algorithm. In order to solve the problem of poor K-means clustering segmentation, HaiXia Zhang [8] proposed an image segmentation algorithm based on particle swarm and K-means, and the new algorithm solves the shortcomings of slow convergence speed of PSO and poor correlation between K-means algorithm and initial clustering center. K-means is highly sensitive to outliers, which can seriously affect the final clustering configuration, due to their removal to obtain high quality results.Zhen Zhang [9] set an objective function on the problem of removing a set of outliers to minimize the cost of K-means clustering of the remaining data points and proposed to utilize a local search algorithm to ensure superior performance of the algorithm. When using the K-means algorithm, the cluster center needs to be initialized, which can lead to unstable cluster performance. Secondly, the performance is poor on non-Gaussian datasets, Han Lu [10] et al. proposed multi-viewpoint k-means based on the adjacency matrix, which maps the affinity matrix to the distance matrix based on the distance of each sample from the domain point, thus making the model insensitive to outliers as well as can be used on non-Gaussian datasets.

This paper focuses on developing a difficulty classification model for Wordle games based on K-means cluster analysis. The more times it takes to guess the word correctly, the more difficult the word is. Word difficulty is categorized by officially providing the number of times a player needs to guess the word correctly each day. The attribute characteristics of words in different classifications are studied and statistically analyzed to derive the key factors affecting the difficulty of Wordle games for the update of Wordle games.

2. Data acquisition and pre-processing

2.1. Data sources

The experimental dataset is from the official Wordle game, and this paper extracts and summarizes the data from the official website.

2.2. Data preprocessing

The purpose of the model is to categorize words in terms of difficulty, which is defined in this paper as being related to two types of attributes: degree of commonness and number of letter repetitions. The experimental data are processed. Some of the word processing results are shown in Figure 1.

2.2.1 Commonness of words and number of letter repetitions

(1) This paper utilizes the Google Books Ngram Viewer online tool, which is a tool that allows you to see how often a word has been used over a period of time, which can be a good indicator of how common the word is.

Step.1: Enter the words from the given data into the tool.

Step.2: Select the time range, this paper uses the duration of 1800 to 2019.

Step.3: The number of occurrences per million words in the selected standardized way is calculated to obtain a number for the degree of commonness of the word. The larger the number, the more common it is, and the smaller it is, the less common it is. By this method, we can better understand the attribute characteristics of the word.

(2) Number of letter repetitions: set 1 for 1 repetition, 0 for no repetition, and so on.

(3) We first processed the data before it could be used in the later model. Since some of the words were very uncommon and not found in the Google Books database, we treated their commonness as

0 by default. In addition, the context of the question indicated that the words of the game had only five letters, but four of the words in the given data appeared with six letters, and after consideration, we did not remove these words.

Word	Number of reported result	Number in hard mode	1 try	2 trie	3 trie	4 trie	5 trie	6 trie	7 or more tries (X	Commonness	ber of repeated	tters
slump	80630	1362	1	3	23	39	24	9	1	77.19	0	
crank	101503	1763	1	5	23	31	24	14	2	103.68	0	
gorge	91477	1913	1	3	13	27	30	22	4	62.91	1	
query	107134	2242	1	4	16	30	30	17	2	22.28	0	
drink	153880	3017	1	9	35	34	16	5	1	324.32	0	
favor	137586	3073	1	4	15	26	29	21	4	384.99	0	
abbey	132726	3345	1	2	13	29	31	20	3	72.15	1	
tangy	169484	3985	1	4	21	30	24	15	5	3.69	0	
panic	205880	4655	1	9	35	34	16	5	1	94.12	0	

Figure 1. Commonness of some words and number of repeated letters

3. K-means based word difficulty classification models

After data preprocessing, the percentage of commonness of the 361 words and the number of repeated letters were obtained.In this paper, we will use the K-means algorithm to cluster analyze the experimental dataset based on the two word attributes, resulting in a difficulty classification of hard, medium, and easy, with corresponding labels of 3, 2, and 1. We will also use this model to test the difficulty classification of the word to be tested, EERIE.

3.1. Introduction to K-means theory

K-means is a division-based algorithm in cluster analysis, and is also an unsupervised learning algorithm. The clustering results are optimized by iterative steps, and the target data set is continuously reallocated to each cluster. The main steps are as follows**.**

- Arbitrary selection of k objects from the sample, each representing the initial mean or center of mass of a cluster.
- For the remaining objects, they are assigned to the nearest cluster based on their Euclidean distance from the mean of each cluster.
- Use the sample means in each cluster as the new center of mass.
- Repeat steps 2 and 3 until the clustering centers no longer change.

For better understanding, we have designed the flowchart as in Figure 2.

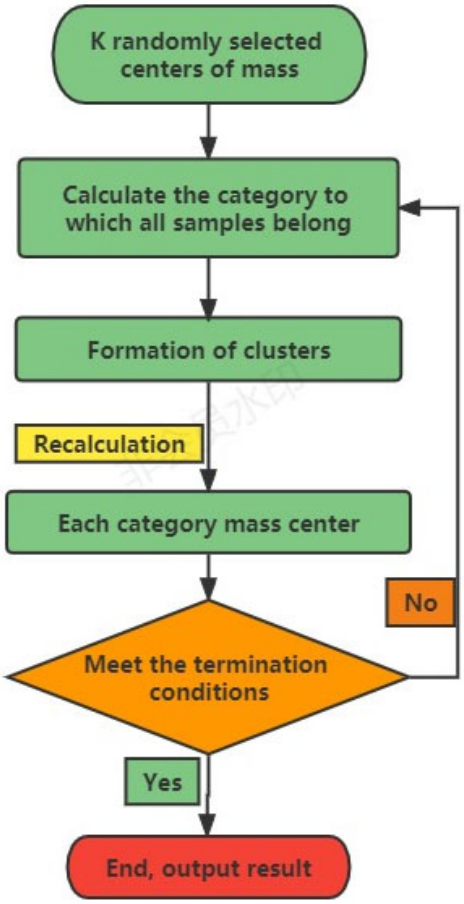

Figure 2. K-means flow chart

where the Euclidean distance between the data objects in space and the cluster centers is given by

$$d(\mathrm{x}, Ci) = \sqrt{\sum_{i=1}^{n} (X\mathrm{j} - Cij)^2} \tag{1}$$

where X is the number of word guesses, Ci is the ith clustering center, n is the dimension of the data object, and Xj, Cij is the jth attribute of x and Ci ; One of the evaluation criteria of K-Means clustering algorithm is the sum of squared errors criterion, which is abbreviated as SSE, and is defined as follows.

$$SSE = \sum_{i=1}^{k} \sum_{p \in Ci} |\, d(p, Ci)\,|^{2} \tag{2}$$

The smaller the SSE value, the closer the data points are to the center of mass, and the better the clustering effect; conversely, if the SSE is larger, the worse the clustering effect is, and the more likely multiple clusters will be considered as one cluster. Therefore, the clusters with larger sum of squared errors need to be divided again in the clustering process.

3.2. Clustering process

The attributes of the words are summarized, the degree of commonness, the number of times players guessed and the number of repeated letters, etc.We first analyze the clustering category variability according to the word attributes, and then analyze the frequency of each clustering category after the clustering summary, and then discriminate the classification of specific sample data according to the clustering annotation; of course, the distance between each sample and the centroid can be derived by observing the coordinates of the cluster center, and finally, an overview of the analysis.

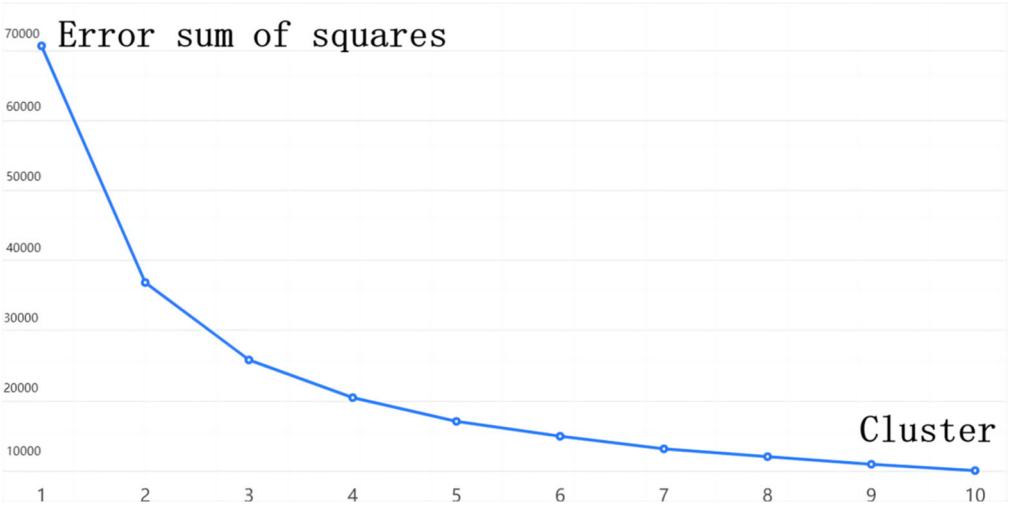

Figure 3. Comparison of the number of clusters

As shown in the figure 3, the horizontal axis is the number of clusters and the vertical axis is the sum of squared errors. It can be seen from the figure: as the number of clusters k increases, the sample division will be finer and the degree of aggregation of each cluster will gradually increase, then the error squared and SSE will naturally decrease gradually. The analysis in the figure shows that SSE tends to level off as the value of k continues to increase, which means that the graph of the relationship between SSE and k is elbow-shaped, and when k=3, it is the elbow, that is, it is really the number of clusters.

Table 1. Clustering Summary

Clustering categories	Frequency	Percentage
Clustering category 1	135	37.604
Clustering category 2	156	43.454
Clustering category 3	68	18.942
Total	359	100

Conclusion from Table 1. The samples were classified using the clustering method, and the final category frequencies were generated using the k-means cluster analysis method as shown in the table above: the clusters yielded 3 categories of groups, and the percentages of these 3 categories were 37.604%, 43.454%, and 18.942%, respectively. Overall, the first two categories are more evenly distributed, and the third category groups are less distributed.

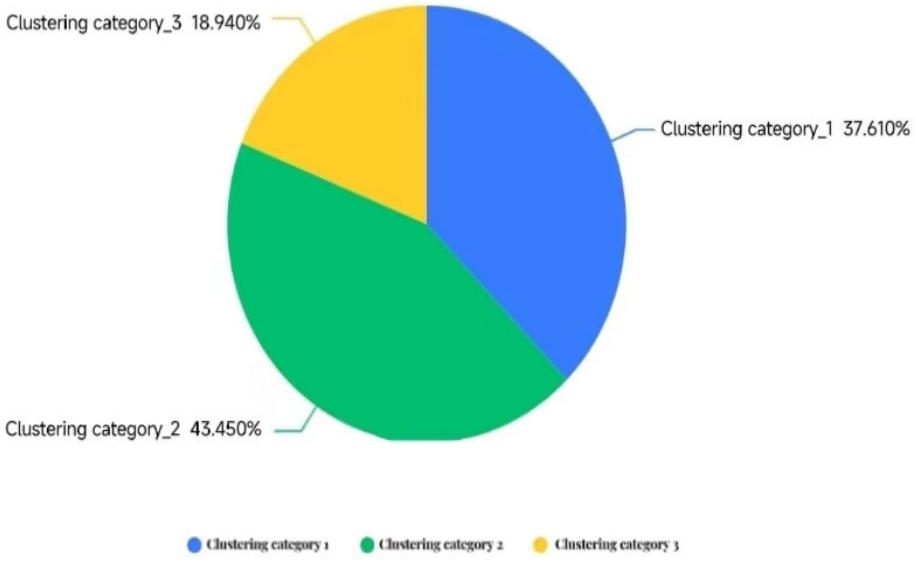

Figure 4. Clustering summary chart

Figure 4 presents the K-means clustering results in a visualized form, and it is possible to visualize the percentage of each category, which is 37.61.%, 43.450%, and 18.940% for the three groups, respectively.

We classify the difficulty as hard, medium and easy, and label them with numbers 3, 2 and 1

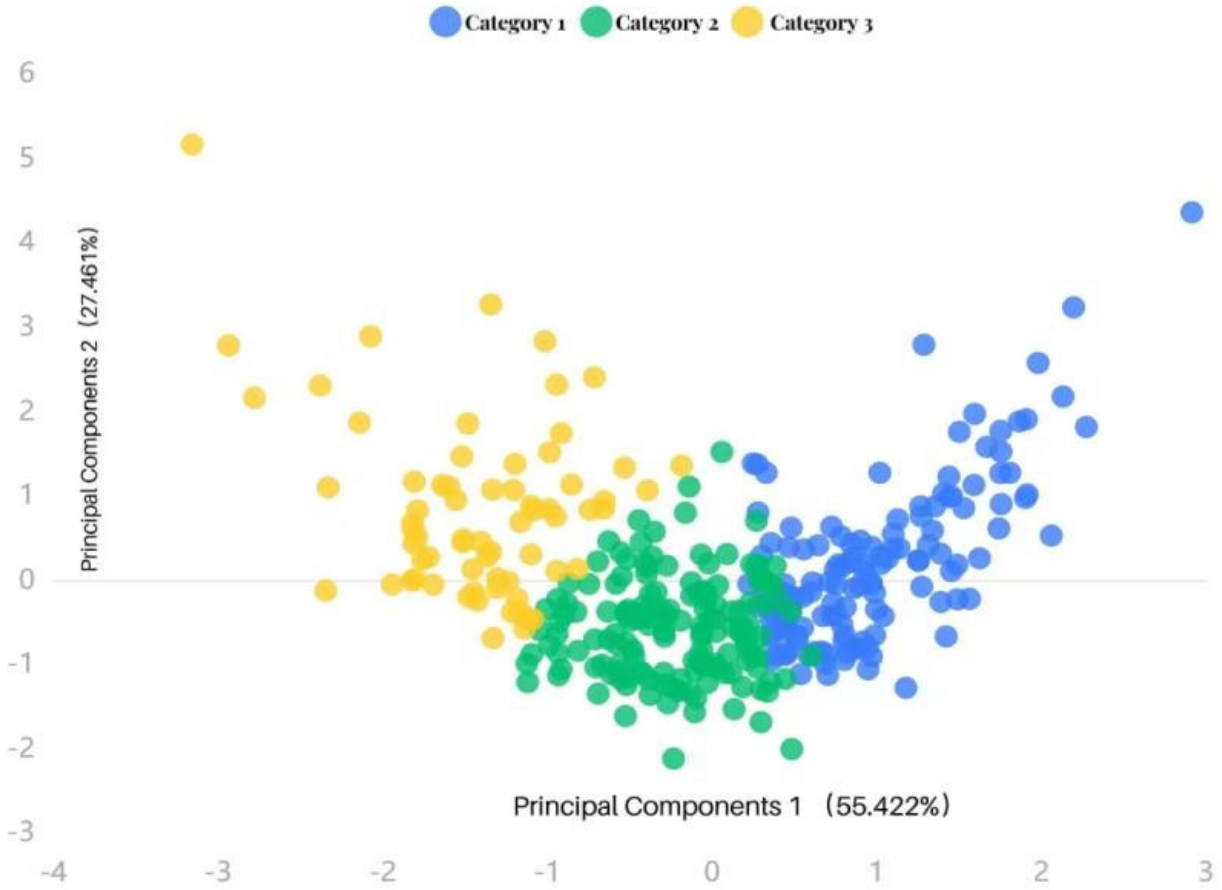

Figure 5. Cluster scatter plot

Since the number of variables is greater than 2, figure 5 illustrates the clustering effect very well. Observing the scatter plot, it is easy to see that the clustering result is the same as the previous paper, which shows that the feature screening and dimensionality reduction processing of the original data do not affect the accuracy of the clustering result, while greatly reducing the complexity of the original data, reflecting the superiority of the algorithm

Table 2. Evaluation indicators

Contour factor	DBI	CH
0.372	0.916	309.326

A total of three metrics are listed in the table above, namely, profile coefficient, DBI and CH(Table 2).The clustering results show the feasibility of the optimization algorithm, for the contour coefficient of 0.375 between the values, DBI of 0.916 indicates that the lower the indicator the better the clustering effect, CH indicator is obtained from the ratio of separation to tightness, the larger the CH indicates the better the clustering effect.

3.3. Sorting statistics on word attributes

The statistical analysis of the word attributes in the difficult, medium, and easy difficulty classifications was performed and the results are shown in Figure 6.

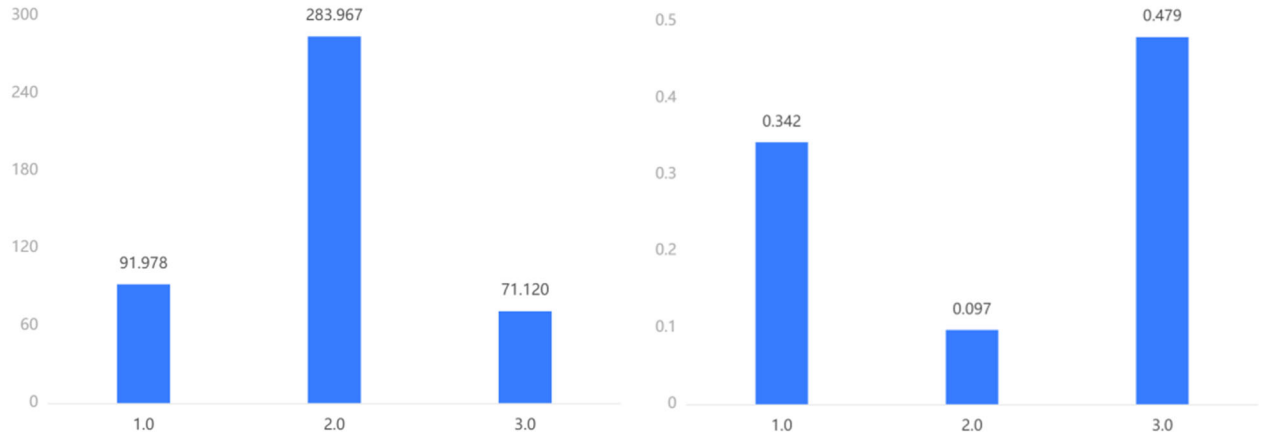

Figure 6. Commonness(left)and Number of repeated letters(right)

Table 3. Category Summary

Clustering type	Commonness	Number of repeated letters
1	91.978	0.342
2	283.967	0.097
3	71.12	0.479

In summary, the data obtained from the classification summary can verify that our previous k-means clustering model analysis is correct, the more common, the lower the difficulty; conversely, the higher the number of repeated letters, the higher the difficulty(Table 3).

4. Conclusions

In this paper, we are studying the difficulty of Wordle game, we collect the number of guesses required to pass the game per day from different players, crawl the Internet to find out how common each word is, and count the repetitive letters of the words are processed as an experimental dataset. The properties of the words were utilized as a criterion to define the difficulty of the game, i.e., the more repetitive letters and the less common they are, the harder this guessing game is. We imported the dataset into the K-means clustering algorithm for clustering experiments, and determined the appropriate K-means according to the auxiliary algorithm [3] in order to improve the clustering effect, and the accuracy of the CH indicator 309.326 model is good. The clustering results are divided into three categories according to the difficulty, labeled 3, 2, and 1, corresponding to difficult, medium, and easy. In this paper, the test word EERIE is selected for model evaluation, and after algorithmic experiments, the result is medium difficulty. This experiment not only helps the subsequent iterations of the Wordle game's version, but also broadens the application area of the K-means clustering algorithm.

Reference

[1] Ahmed M, Seraj R, Islam S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.

[2] Yu Q, Wang H, Qiao S, et al. k-means Mask Transformer[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 288-307.

[3] Pham D T, Dimov S S, Nguyen C D. Selection of K in K-means clustering[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2005, 219(1): 103-119.

[4] Borlea I D, Precup R E, Borlea A B. Improvement of K-means cluster quality by post processing resulted clusters[J]. Procedia Computer Science, 2022, 199: 63-70.

[5] Ghezelbash R, Maghsoudi A, Shamekhi M, et al. Genetic algorithm to optimize the SVM and K-means algorithms for mapping of mineral prospectivity[J]. Neural Computing and Applications, 2023, 35(1): 719-733.

[6] Ghazal T M. Performances of K-means clustering algorithm with different distance metrics[J]. Intelligent Automation & Soft Computing, 2021, 30(2): 735-742.

[7] Rezaee M J, Eshkevari M, Saberi M, et al. GBK-means clustering algorithm: An improvement to the K-means algorithm based on the bargaining game[J]. Knowledge-Based Systems, 2021, 213: 106672.

[8] Zhang H, Peng Q. PSO and K-means-based semantic segmentation toward agricultural products[J]. Future Generation Computer Systems, 2022, 126: 82-87.

[9] Zhang Z, Feng Q, Huang J, et al. A local search algorithm for k-means with outliers[J]. Neurocomputing, 2021, 450: 230-241.

[10] Lu H, Gao Q, Wang Q, et al. Centerless multi-view K-means based on the adjacency matrix[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(7): 8949-8956.