Summary

As a word puzzle game, Wordle had become a big hit in 2022. The reported results of it in Twitter even reach 360,000 at the beginning of Feb, 2022. After one year, the people who plays Wordle gradually came down and the reported data seems to be stable. Twitter had collected the data of information of results reported by people for more than one year, using the dataset given by Twitter, we could derive a lot of information, such as predicting the trend of number of people who play the game in the future, getting results of difficulties related to different words in word puzzle, which will be of great significant for the compony to make strategies to popularize their game.

As for the problem in this contest, we developed different models for different questions to better fits the data while getting results. We cleaned the data, fixed some of its errors and make a brief review of the characteristic of the data before we started our work focus on the requirement.

In problem 1(a), we applied ARIMA (4,1,2) model to predict the number of people who reported their results on March,1,2023, and tested it with random forest method. The number predicted by ARIMA is 19698. In order to further verify the reliability of our model, we applied our model to predict number in February,1,2023, results showed that the real data is also in our predicted interval, which was quite good. For problem 1(b), we use hypothesis testing, and found that there’s no high correlation between some basic attributes (word frequency, part-of-speech, sentiment).

In problem 2, in order to predict the number of a particular word in future, we first made 5 attributes that might influence results: …(here is the attribute)… and use Random Forest Regression. Here we use number of people to replace time parameters using the prediction model constructed in Problem 1. The associated percentage of EERIE is [0,3,17,37,32,10,0].

As for problem 3, We use K-means clustering through using the attributes defined above to classify the words, we divide them into 4 classes: Simple, Ordinal, Difficult, Rather difficult. The Standard we use are means and standard Deviation of try times. As for the word ”EERIE”, k-means method classify it to Difficult.

For the last question, we try to restore the data of all users instead of those who shared their results. In order to achieve this, we write an algorithm which simulate the process of human guessing words based on Wordle’s word data and English word frequency, and find a series of interesting information considering the people who report their game results on Twitter.