Question:

1. If the percentage is hard mode or entire: entire
2. Can we use additional data(up-to-date data and check dictionary): maybe check dictionary

Task:

You have been asked by the New York Times to do an analysis of the results in this file to answer several questions.

1. The number of reported results vary daily. Develop a model to explain this variation and use your model to create a **prediction interval** for the number of reported results on March 1, 2023. Do any attributes of the word affect the percentage of scores reported that were played in Hard Mode? If so, how? If not, why not?

* 1. This seems to be a Math problem-check prediction interval, the idea is probably simply to use the normal distribution, we should also consider the time series to see if there is a trend with the number; Time series
  2. The ratio of hard mode and overall:

NLP

2. For a given future solution word on a future date, develop a model that allows you to

predict the distribution of the reported results. In other words, to predict the associated

percentages of (1, 2, 3, 4, 5, 6, X) for a future date. What uncertainties are associated with your model and predictions? Give a specific example of your prediction for the word

EERIE on March 1, 2023. How **confident** are you in your model’s prediction?

* 1. This seems to be a Math problem-check prediction interval, the idea is probably simply use the normal distribution, we should also consider the time series to see if there is a trend with the number
  2. Ratio of hard mode and overall

3. Develop and summarize a model to classify solution words by difficulty. Identify the

attributes of a given word that are associated with each classification. Using your model,

how difficult is the word EERIE? Discuss the accuracy of your classification model.

4. List and describe some other interesting features of this data set.

Finally, summarize your results in a one- to two-page letter to the Puzzle Editor of the New York Times.

* SVM
* LSTM+时间序列

Idea:

* 1. A prediction task, use ML, DL, NLP and time series
  2. The concept of uncertainties and confident should be investigated
  3. What does difficulty mean here: the distribution
  4. If it is based on the distribution, we can use a clustering method such as K-means to classify difficulty; train an additional model to predict EERIE’s distribution, and classify based on K-means
  5. Can also try to play the game and find an “optimal method” for it - RL might be good for this
  6. Investigate the word(word distribution, how far each words is to each other, their characteristic features…)
  7. Time series with features
  8. Features itself

Basic knowledge:

Time Series + NLP

Times Series:

1. It takes information unrelated to words, only related to the sequence of date, and many tricks to play here. A classic way to do this is by doing characteristic engineering and applying XGboost; since the data is relatively simple, linear regression might be a good way as well.

Reference: <https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting>

To do:

1. **Graph related to time**
2. **Feature engineering and modeling**
3. **Question 1 part 1**

NLP:

1. Use pre-trained model to map words to high dimensions and cancat to train a NN
2. Words itself contains many information, such as counting the character, and check the dictionary for the frequency or hardness of the word, we can add additional features related to word

To do:

1. **Understand the game thoroughly -> word features**
2. **Pre-trained model relation and word frequency in the dictionary**
3. **Modeling**

Math:

1. Stats and Probability, especially the idea of prediction interval and confidence, how we use them to interpret machine learning model

To do:

1. **Stats analysis**
2. **preprocessing**

Data Analysis:

1. Visualize, analyze, and clean data

RL

Timeline:

2-16: Question settle and Work division

2-17,18: Task finished

2-19: Task combined and paper construction started

2-20: Paper construction and submission