WORDLE GUESSING

**Abstract**

本文研究纽约时报如何根据的Wordle谜题游戏中的各项数据，去预测报告结果的数量，并建立出根据solution word预测成绩分布和评价谜题难度的方法。

针对问题一，首先对数据进行清洗，更正或消除异常数据，然后建立Quantity Prediction Model拟合数据集中的的报告结果数量，并预测未来的趋势。

针对问题二，在原数据集中定义了谜题的困难系数直接表现为成绩分布的加权平均，而我们需要做的是挖掘单词本身的潜在特征去预测未来的成绩分布。对各项特征与难度系数的相关性进行分析后，确定与成绩分布相关的几个特征。之后本文以Decision Tree,Naïve Bayes,KNN,Logistic Regression,Linear SVC,Percepton,SGD,SVM作为参考模型，并把困难系数作为标签对单词特征进行训练，并观察这些模型的表现。最终结果表明决策树是其中最好的选择，使用原数据集上80%的数据进行训练，剩下的作为测试集，几乎100%地预测了难度系数。之后以各个成绩区间的占比作为标签去训练，得到能够预测成绩分布的决策树模型。将’EERIE’作为样本进行多次预测，取平均结果作为输出，发现有12%左右的偏移，算是达到了一个不错的预测效果。

针对问题3，(补充一下)

针对问题4，本文尝试在原数据集中挖掘了许多特征，比如报告数量中的困难模式数量占比HR，成绩分布期望即难度系数D，不到j次尝试通过游戏的占比Uj，单词的位置编码P,重复度Mul,元音个数VN,元音位置编码VP，总字母频率系数FS，首字母频率系数FF，双字母组频率系数FB，单词常用频率系数FE。最终保留了有效的特征来训练机器学习模型。

This paper investigates how the New York Times predicts the number of reported results based on data from the Wordle puzzle game, and establishes a method for predicting the distribution of results and evaluating the difficulty of puzzles based on the solution word.

For problem 1, we first clean the data, correct or eliminate abnormal data, and then build a Quantity Prediction Model to fit the number of reported results in the data set and predict the future trend.

For problem 2, the difficulty coefficient of the puzzle is defined in the original dataset as a weighted average of the performance distribution, and what we need to do is to mine the potential features of the words themselves to predict the future performance distribution. After analyzing the correlation between each feature and the difficulty coefficient, several features related to the score distribution were identified. After that, in this paper, Decision Tree,Naïve Bayes,KNN,Logistic Regression,Linear SVC,Percepton,SGD,SVM are used as reference models and the difficulty coefficients are used as labels to train the word features and observe the performance of these models. The final results showed that decision tree was the best choice among them, using 80% of the data on the original dataset for training and the rest as the test set, which predicted almost 100% of the difficulty coefficients. After that, the percentage of each performance interval was used as a label to train and get a decision tree model that could predict the performance distribution. The 'EERIE' was used as a sample for several predictions, and the average result was taken as the output, and it was found to have about 12% bias, which is considered to achieve a good prediction result.

For problem 3,

For problem 4, this paper tries to mine many features in the original dataset, such as the percentage of the number of difficult patterns in the number of reports HR, the expectation of the score distribution i.e. the difficulty coefficient D, the percentage of less than j attempts to pass the game Uj, the position coding of words P,repetition Mul, the number of vowels VN, the vowel position coding VP, the total letter frequency coefficient FS, the first letter frequency coefficient FF, the double letter group frequency The final effective features are retained to train the machine learning model.

Catalogue

[1 Introduction 3](#_Toc127759059)

[1.1 Background 3](#_Toc127759060)

[1.2 Restatement of the Problem 3](#_Toc127759061)

[1.3 Our work 4](#_Toc127759062)

[2 Assumptions and Explanations 4](#_Toc127759063)

[2.1 Condition Assumptions 4](#_Toc127759064)

[2.2 Notation 5](#_Toc127759065)

[3 Preparation 5](#_Toc127759066)

[3.1 Data cleaning 5](#_Toc127759067)

[3.1.1 Fix the Words 5](#_Toc127759068)

[3.1.2 Add new attribute 6](#_Toc127759069)

[3.1.3 Handle abnormal data 6](#_Toc127759070)

[3.2 Feature analysis 7](#_Toc127759071)

[3.2.1 Correlation analysis between features 7](#_Toc127759072)

[3.2.2 Attributes that affect in hard mode 8](#_Toc127759073)

[4 Quantity Prediction Model 10](#_Toc127759074)

[4.1 ARIMA Model 11](#_Toc127759075)

[4.1.1 Principle and implementation 11](#_Toc127759076)

[4.1.2 Forecast Results 12](#_Toc127759077)

[4.2 GM Model 14](#_Toc127759078)

[4.2.1 Principle and implementation 14](#_Toc127759079)

[4.2.2 Forecast Results 15](#_Toc127759080)

[4.3 Other Model 15](#_Toc127759081)

[4.3.1 Polyfit 15](#_Toc127759082)

[4.3.2 LSTM 16](#_Toc127759083)

[5 Difficulty Prediction Model 17](#_Toc127759084)

[5.1 Possible attributes of words 17](#_Toc127759085)

[5.1.1 Letter frequency 17](#_Toc127759086)

[5.1.2 Bigram frequency 18](#_Toc127759087)

[5.1.3 Multiplicity 18](#_Toc127759088)

[5.1.4 Vowel 18](#_Toc127759089)

[5.1.5 Degree of common use 18](#_Toc127759090)

[5.2 Preparatory work 19](#_Toc127759091)

[5.3 Training result and model comparison 19](#_Toc127759092)

[5.4 How difficult ‘EERIE’ is 20](#_Toc127759093)

[6 Model Evaluation 20](#_Toc127759094)

[6.1 Strengths 20](#_Toc127759095)

[6.2 Weaknesses 20](#_Toc127759096)

[7 References 21](#_Toc127759097)

[8 Letter for the Puzzle Editor of the New York Times 21](#_Toc127759098)

[9 About the Authors 22](#_Toc127759099)

[10 Appendix 22](#_Toc127759100)

# Introduction

## Background

（待翻译）纽约时报每天提供一个谜题，玩家在六次或者或者更短时间内猜出这个谜题即算成功。玩家需要在认可的一万三千个单词中进行猜测，每次猜测必须是一个实际的英文单词。每次猜测会提供一个一次反馈，字母的颜色会改变，灰色表示单词中根本没有这个字母，黄色表示单词中有这个字母但位置不对，绿色即表示猜中单词中字母的正确位置。每天有数以万计的人在推特上分享他们的猜测次数，wordle掀起了一股猜词热潮。

The New York Times offers one puzzle per day, and players are considered successful if they guess the puzzle six times or less. Players are asked to guess from a list of 13,000 recognized words, and each guess must be an actual English word. Each guess will provide one feedback, and the color of the letter will change from gray, which means that there is no letter in the word, to yellow, which means that there is a letter in the word but in the wrong position, to green, which means that the correct position of the letter in the word has been guessed. With tens of thousands of people sharing their guesses on Twitter every day, wordle has started a word-guessing craze.

## Restatement of the Problem

• 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?（原题目中问题）

用一个模型解释在推特上分享报告结果的数量变化趋势，并预测在2023年3月1日这天在推特上分享的报告结果数量。猜测是否有其他特征影响在hard模式下人们猜出正确单词的次数。Use a model to explain the trend in the number of reported results and to predict the number of reported results on Twitter on the date of March 1, 2023. Speculate whether there are other features that affect the number of times people guess the correct word in the HARD model.

• 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?(原题中问题)

建立一个模型，当给你后面任意一天的日期，预测出猜词所需次数的分布。预测出猜出ERRIE这个单词的所需次数的分布。描述你的模型中中的不确定性。 Build a model that predicts the distribution of the number of times it takes to guess the word when given any date later in the day. Predict, predict the distribution of the number of times needed to guess the word ERRIE. Describe the uncertainty in your model.

• 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.(原题中问题)

找出可能的属性，对单词按照难度进行分类。对给出的单词eerie的难度进行预测，谈论你分类模型的准确程度。Identify the possible attributes to classify words according to their difficulty. Make predictions about the difficulty of the given word ‘EERIE’ and talk about the accuracy of your classification model.

• List and describe some other interesting features of this data set.（原题中问题）

列出数据集中其他一些有趣的特征

Our work List some other interesting features in the dataset

## Our work

对于问题1，我们使用了多个模型去解释并预测报告结果的数量，包括ARIMA模型、GM 模型和其他模型，其中我们小组认为ARIMA模型最能解释这种变化趋势，而多项式函数并不适合这样的变化趋势。根据多个模型检验，我们认为2023年3月1日的报告结果会在[6000,17500]之间。通过数据可视化分析，我们认为词语的包括字母重复度、词频在内的部分属性会影响困难模式下的分数百分比。

对于问题2，我们首先定义了成绩分布的期望是谜题难度的最直接体现，然后分析了solution word中的潜在属性，将原数据集扩展为能够表示单词中多个特征的数据集，并用一些机器学习模型进行训练。对比各个模型的准确率发现，决策树的预测效果是最好的、其次是KNN模型等，同时也否决了一些明显不适用于该问题的模型。在样本数剩下357的情况下，决策树的预测准确率高达100%，因此我们也使用它作为预测2023年3月1日的谜题’EERIE’的分数占比。我们将这个样本多次投放到模型中，以原数据集中每个成绩占比作为标签进行训练，预测并取平均成绩，最终预测分数占比为(1%,3%,4%,28%,28%,21%,15%)，难度系数为5.02。而采用我们定义的难度系数作为标签取训练，最终预测’EERIE’的难度系数为4.24，根据这个误差，我们有88%的信心认为模型的预测是合理的。

对于问题3，我们沿用问题2的模型，只从单词的本身属性入手，去对solution words的难度进行分类，通过相关性分析，我们发现在Wordle游戏中单词的字母频率、音节频率、元音个数、元音位置、常用程度等属性都与solution words的难度分级相关，但是找不到任何潜在属性能够很大程度上划分solution words的难度，几乎都是这些因素共同作用的结果。对于’EERIE’这个单词，我们将它分为‘较难的’类型，比原数据集90%左右的单词都要难。对这种预测结果，我们的分类模型准确率达到了65%。(这些结果是临时糊上去的，不知道够不够时间做实验了)

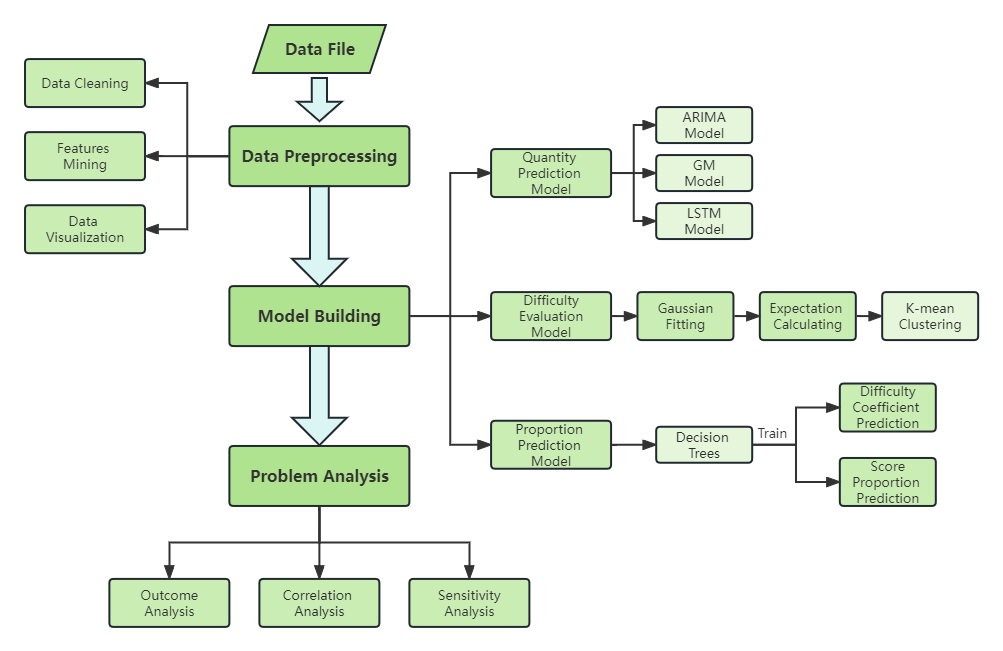
对于问题4，我们小组在原数据集上先是处理异常数据，进行数据清洗，分析数据集中各特征的相关性。在**5.1**节中，我们定义了几个有用的字段作为数据集的挖掘特征，而在**7.1**节中，我们又挖掘了单词本身的潜在属性。这些特征对模型的建立和验证都起到了很大的促进作用。

For question 1, we used multiple models to explain and predict the number of reported outcomes, including the ARIMA model, the GM model, and other models, of which our group believes that the ARIMA model best explains this trend, while the polynomial function does not fit such a trend. Based on multiple model tests, we believe that the reported results on March 1, 2023 will be between [6000,17500]. Through data visualization analysis, we concluded that some attributes of words, including letter repetition and word frequency, affect the percentage of scores in the difficult mode.

For question 2, we first defined the expectation of the score distribution as the most direct representation of the difficulty of the puzzle, and then analyzed the potential attributes in the solution word, expanded the original dataset to one that could represent multiple features in the word, and trained it with a number of machine learning models. Comparing the accuracy of each model, we found that the decision tree was the best predictor, followed by KNN models, etc., while rejecting some models that were clearly not applicable to the problem. With a sample size of 357 left, the decision tree had a prediction accuracy of 100%, so we also used it as a percentage of the score for predicting the puzzle 'EERIE' on March 1, 2023. We drop this sample into the model multiple times, train with each score share in the original dataset as a label, predict and take the average score, and finally predict the score share as (1%,3%,4%,28%,28%,21%,15%) with a difficulty factor of 5.02. And using the difficulty factor we defined as a label to take the training, the final prediction of ' EERIE' has a difficulty coefficient of 4.24, and based on this error, we are 88% confident that the model's prediction is reasonable.

For problem 3, we follow the model of problem 2 and only start from the attributes of the words themselves to classify the difficulty of solution words. Through correlation analysis, we find that the attributes of letter frequency, syllable frequency, number of vowels, vowel position, and commonness of words in the Wordle game are all related to the difficulty grading of solution words, but find We could not find any potential attributes that could largely classify the difficulty of solution words, and almost all of them were the result of the combined effect of these factors. For the word 'EERIE', we classified it as a 'harder' type, which is harder than about 90% of the words in the original dataset. For this prediction result, our classification model achieves an accuracy of 65%.

For problem 4, our group first processed anomalous data on the original dataset, performed data cleaning, and analyzed the relevance of each feature in the dataset. In Section 5.1, we defined several useful fields as mining features for the dataset, while in Section 7.1, we mined the potential attributes of the words themselves. These features contribute significantly to both model building and validation.



# Assumptions and Explanations

## Condition Assumptions

1．谜题的难度只与单词本身的性质有关，并且可以使用用户尝试次数的分布作为标准去衡量，不考虑其他方面的影响；

2．一次猜出答案的数据是允许的，尽管存在许多人在同一天一次性猜对，我们认为是存在玩家掌握其他信息或者高级的游戏策略；

3．数据集给定的百分比对结果的影响影响忽略不计，经过四舍五入后，对训练效果也不会产生太大影响；

4．假设wordle参与人数在一段时间内不会受其他因素影响激增；

5．分析给定数据集所建立的模型足以解释Reported results的变化趋势，不考虑其他因素。

1. the difficulty of the puzzle is related only to the nature of the word itself and can be measured using the distribution of the number of user attempts as a criterion, regardless of other influences.

2. the data on one-time guesses are allowed, although the existence of many people guessing correctly at once on the same day is considered to be the presence of players with other information or advanced game strategies.

3. the percentage of the data set given has a negligible impact on the results and, after rounding, does not have much effect on the training effect.

4. the assumption that the number of wordle participants will not be influenced by other factors to surge over time.

5. the model built by analyzing the given dataset is sufficient to explain the trend of the Reported results, without considering other factors.

## Notation

**Table 1** shows the notations that we use.

Table 1. Notation

|  |  |
| --- | --- |
| Symbol | Description |
|  | 记录i中(1,2,3,4,5,6,X)的比例之和 |
|  | 记录i中Number in hard mode 在 Number of reported results中的占比 |
|  | 记录i中的游戏难度系数 |
|  | 记录i中的使用小于j次通过游戏的占比 |
|  | 单词w的重复度 |
|  | 单词w中的元音字母个数 |
|  | 单词w中所有元音位置的编码 |
|  | 单词w中第j个字母的位置编码 |
|  | 单词w的频率系数FS，由各个位置的字母出现频率决定 |
|  | 单词w的频率系数FF，由首字母的出现频率决定 |
|  | 单词w的频率系数FE，由单词本身的常用程度决定 |
|  | 单词w的频率系数FB，由单词中所有双音节出现频率决定 |
|  |  |
|  |  |

|  |  |
| --- | --- |
| 公式 |  |

**Table 1** shows the notations that we use.

Table 1. Notation

|  |  |
| --- | --- |
| Symbol | Description |
|  | Record the sum of the proportion of (1,2,3,4,5,6,X) in i |
|  | Record the percentage of Number in hard mode in Number of reported results in i |
|  | Record the difficulty factor of the game in i |
|  | The percentage of games with less than j passes in record i |
|  | Repetition of word w |
|  | Number of vowel letters in word w |
|  | Number of all vowel positions in word w |
|  | Position coding of the jth letter in the word w |
|  | Frequency factor FS of the word w, determined by the frequency of occurrence of the letters in each position |
|  | Frequency factor FF of the word w, determined by the frequency of occurrence of the initial letter |
|  | Frequency factor FE of word w, determined by the frequency of the word itself |
|  | Frequency factor FB of the word w, determined by the frequency of all diacritical marks in the word |
|  |  |
|  |  |

# Preparation

## Data cleaning

### Fix the Words

根据游戏的限制，数据集中的Word字段必须是5个小写字母组成的单词，而我们发现数据集中一些记录并不符合这一条件。因此我们通过遍历数据集，发现了5个存在错误的Word，并根据Wordle游戏官网给出的历史谜题[1]更正了一下数据：

• 删除 favor后多出的一个空格(Contest number=207)

• tash 改为 stash (Contest number=314)

• clen 改为 clean (Contest number=525)

• na?ve 改为 naive (Contest number=540)

• rprobe 改为 probe (Contest number=545)

以上修改是可行的，因为它参照了真实的数据，保证了数据质量的同时，避免了删除记录带来的危害，包括样本较小对模型鲁棒性的影响和数据不连续。

According to the restrictions of the game, the Word field in the dataset must be a word consisting of 5 lowercase letters, and we found that some records in the dataset did not meet this condition. So we found 5 Word with errors by traversing the dataset and corrected the data according to the historical puzzle [1] given on the official website of Wordle game.

• Delete one extra space after favor (Contest number=207)

• change tash to stash (Contest number=314)

•clen to clean (Contest number=525)

•na?ve to naive (Contest number=540)

•rprobe to probe (Contest number=545)

The above modification is feasible because it refers to the real data and ensures the data quality while avoiding the hazards of deleting records, including the impact of small samples on the robustness of the model and data discontinuity.

### Add new attribute

将原始数据集” Problem\_C\_Data\_Wordle.xlsx”中的首行删除，并把错误的单词改正后，另存为”data0.csv”，表示初始数据集。为了更好地挖掘数据集中的信息，我们首先给数据集增添以下字段：

• under\_j是表示尝试次数少于j次所占百分比的字段，简称为Uj。特别地，我们称U7为Overall。公式为：

|  |  |
| --- | --- |
|  |  |

• Hard\_rate是表示选择困难模式数量占报道结果的比例，简称为H，公式为：

|  |  |
| --- | --- |
|  |  |

The first row of the original dataset "Problem\_C\_Data\_Wordle.xlsx" is deleted and the wrong words are corrected and saved as "data0.csv", which represents the initial dataset. In order to better explore the information in the dataset, we first add the following fields to the dataset.

• under\_j is the field that indicates the percentage of attempts with less than j attempts, abbreviated as Uj. Specifically, we call U7 Overall. the formula is:

|  |  |
| --- | --- |
|  |  |

•Hard\_rate is the ratio of the number of difficult modes selected to the reported results, abbreviated as H, and the formula is:

|  |  |
| --- | --- |
|  |  |

### Handle abnormal data

通过观察数据集，能够直观清楚地发现两个异常数据，分别为Contest number = 281的记录(它的Overall是126%)，以及Contest number=529的记录(它的Hard Rate高达93%)。我们选择把这些异常的记录删去。

接下来，以Contest Number（或者说是时间）作为x轴，分别以Number of reported results，Number in hard mode，和Hard Rate 作为y轴去散点图图，分别为**Figure X, Figure X**。能够可以看到有个别离群点。 为了检测并纠正异常值，分别做箱线图**Figure X，Figure X**。将异常值用前一个单位的数据代替，分别得到**Figure X，Figure X**，其中绿色表示原来数据组成的线，蓝色表示更正之后的线。

|  |  |
| --- | --- |
| 图表  描述已自动生成 Figure 1. Contest Number与Number of reported results的关系图 | 图表, 散点图  描述已自动生成  Figure 2. Contest Number与Number in hard mode的关系 |

|  |  |
| --- | --- |
|  |  |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 图表  描述已自动生成 | 图表, 散点图  描述已自动生成 |  |  |

Figure 3.

上面三张是Contest Number与Hard rate的关系图，左图表示原始数据集，中间图表示经过一次清洗后的结果，看上去仍有两个点的Hard Rate看上去比较诡异，但数据集整体来说是可用的。我们再限定Hard Rate的范围，经过第二次清洗得到右图。我们把新的数据集保存记作”data1.csv”。到此，数据清理的工作基本完成。

Next, using the Contest Number (or time) as the x-axis and the Number of reported results, Number in hard mode, and Hard Rate as the y-axis to make the relationship graphs, Figure 1, Figure 2, and Figure 3 respectively, we can clearly find two We can clearly find two abnormal data, which are the record with the Contest number = 281 (its Overall is 126%), and the record with the Contest number = 529 (its Hard Rate is 93%). We choose to delete these abnormal records to get Figure 4.

|  |  |
| --- | --- |
| 图表  描述已自动生成 Figure 1. The relationship chart between Contest Number and Number of reported results | 图表, 散点图  描述已自动生成  Figure 2. The relationship chart between Contest Number and Number in hard mode |

|  |  |
| --- | --- |
| 图表  描述已自动生成  Figure 3. The relationship chart between Contest Number and Hard rate before cleaning up abnormal data | 图表, 散点图  描述已自动生成  Figure 4. The relationship chart between Contest Number and Hard rate after cleaning up abnormal data |

(If you have time to add to this place, first make three boxplots to indicate the noise data, and then replace them with interpolation or other processing methods)

Although it seems that the Hard Rate of two points still looks weird, we think it is still manageable. We think the current data set is available overall, and we will save the new data set as "data1.csv". At this point, the data cleaning is basically complete.

## Possible attributes of words

虽然我们通过统计不同尝试次数的占比可以推断出谜题的大致难度，并且得到难度系数D，但归根到底谜题的难度是由单词本身的性质来决定的。比方说，日常生活中人们用的少的单词被猜中的概率更低。另外，带有字母重复的单词更难被猜出来，这是由游戏机制决定的，假设单词中由两个e，一开始你只填了e，即便它填在正确的位置，你也难以推断另外一个e是需要填进去的，更何况填在错误的位置只会返回一个黄色格子。

因此，我们针对单词的各项特征和对应难度系数去建立一个新的数据集”word\_data.csv”，并通过分析来确定这些特征对难度的贡献。

Although we can infer the approximate difficulty of a puzzle by counting the percentage of different attempts and get the difficulty factor D, in the end the difficulty of a puzzle is determined by the nature of the word itself. For example, words that are less used by people in daily life have a lower probability of being guessed. In addition, words with repeated letters are more difficult to guess, which is determined by the game mechanism. Suppose the word consists of two e's, and at first you only fill in the e. Even if it is filled in the correct position, it is difficult to deduce that the other e is needed to be filled in, not to mention that filling in the wrong position will only return a yellow square.

Therefore, we create a new dataset "word\_data.csv" for each feature of the word and the corresponding difficulty factor, and analyze it to determine the contribution of these features to the difficulty.

### Letter frequency

The frequency of occurrence of letters in each position affects the difficulty of the whole word. For example, according to Wikipedia [6] it is stated that the most frequent occurrence of letters in English is 'e', followed by 't', 'a', 'o ' ....... However, the highest frequency of the initial letter is 'a'. Based on this information we can infer that words that start with 'a' and have 'e' or 't' are relatively easy to guess. However, in Wordle this rule is not so applicable, for example, 'tion' as a suffix in many words will greatly affect the frequency of the letter 't', and in wordle, which is a five-letter word, it is almost impossible for 'tion' to appear. It is almost impossible for 'tion' to appear. Therefore, we refer to all datasets available as answers and guessable words in Wordle on the Kaggle website [5] and data mining codes, and obtained a more reliable frequency factor FS by calculating the following formula.

|  |  |
| --- | --- |
|  |  |

where count(a,b) denotes the frequency of occurrence of letter a in dictionary b, and sum(b) denotes the frequency of occurrence of all letters in dictionary b.

Among the letters in each position, the initial letter frequency is the feature that we need to pay extra attention to. We obtain the frequency coefficient FF of initial letters in a similar way, calculated as follows.

|  |  |
| --- | --- |
|  |  |

The word0 represents the initial letter of the word word.

### Bigram frequency

It is obviously not reasonable to define the frequency coefficient only in terms of each letter, let's say that 'st' occurs more frequently than a syllable like 'ee', which consists of two high-frequency letters, so we also have to consider the frequency of syllables that occur in the word. To simplify the problem, we consider only diphthongs and define all combinations of adjacent letters in a word as diphthongs (although some cannot be called syllables) and calculate the sum of the frequencies of all combinations in the word, FB, with the following formula.

|  |  |
| --- | --- |
|  |  |
| where count(a,b) indicates the frequency of occurrence of adjacent letter combinations a in dictionary b, and sum(b) indicates the frequency of occurrence of all adjacent letter combinations in dictionary b. |  |

### Multiplicity

As mentioned earlier, if there are repeated letters in the word, the difficulty of the puzzle will be greatly increased. We define the repetition Mul as the sum of the product of the number of species of each repeated letter in the word and the number of occurrences of the letter, calculated as follows.

|  |  |
| --- | --- |
|  |  |

### Vowel

元音是单词中不可缺少的部分，因此我们元音的位置和个数都是值得考虑的因素，另外记录元音个数的同时也挖掘了单词中辅音个数的影响。我们定义VN为单词中的元音个数，VP为单词中所有元音位置的二进制编码，比如’EEIRE’中元音的位置编码为[1,1,1,0,1]。

Vowels are an indispensable part of words, so the position and number of our vowels are factors worth considering. In addition, recording the number of vowels also taps into the effect of the number of consonants in a word. We define VN as the number of vowels in the word and VP as the binary code of all vowel positions in the word, for example, the position of vowels in 'EIRE' is coded as [1,1,1,0,1].

### Degree of common use

我们在前面根据单词本身的性质，挖掘了许多影响单词难度的潜在因素。从日常生活的角度出发，单词的常见程度可以非常直观地体现单词的难度，因为人们往往猜测常用的单词。我们定义单词常见程度系数FE，这是通过统计大文本中单词的出现频数得到的，这些数据可以在Kaggle上面找到[4]。

We have previously uncovered many potential factors that influence word difficulty based on the nature of the words themselves. From the perspective of daily life, the commonness of words can be a very intuitive representation of word difficulty, as people tend to guess the commonly used words. We define the word commonness factor FE, which is obtained by counting the frequency of word occurrences in large texts, and these data can be found on top of Kaggle [4].

## Feature analysis

### Correlation analysis between features

在python中，seaborn库的heatmap是特征相关性可视化的一个便捷工具，它以冷色表示正相关的程度，暖色表示负相关的程度，从而让我们更好地发现特征之间的相关性。我们尝试对data0使用heatmap，结果如**Figure 5**所示。

图表, 树状图

描述已自动生成针对由单词本身挖掘出的的属性，我们使用热图观察特征之间的相关性，尤其是与难度系数的相关性，在此之前，我们把一些难以处理的频率系数转化为了信息熵。结果如**Figure X**所示

图表, 树状图

描述已自动生成

Figure X.

Figure 5. data0中出去Word和Date字段后得到的heatmap

In python, the heatmap of seaborn library is a convenient tool for feature correlation visualization, which indicates the degree of positive correlation with cool colors and the degree of negative correlation with warm colors, thus allowing us to better discover the correlation between features. We try to use heatmap for data0, and the result is shown in Figure 5

### Attributes that affect in hard mode

**属性一：**单词重复次数：我们小组在进行wordle游戏时发现若目标单词中出现重复的字母，那么解题成功的次数或者解题时间会增长，换言之，由于个人的知识储备中大部分都是没有字母重复的单词，因此我们在做题是很少会将字母重复的单词作为目标单词。由于目标单词长度都为5个，因此我们按照字母重复次数将单词划分为每个字母出现一次,一个字母出现两次,一个字母出现三次,两个字母各出现两次 等情况。

根据常识判断，排除不存在的情况，包括但不限于一个字母出现4或5次、单词为两种字母的组合。我们将单词按照字母重复次数划分为以下4类：

W1：每个字母出现一次（无单词重复）

W2：一个字母出现两次

W3：一个字母出现三次

W4：两个字母各出现两次。

并对数据集中的单词按照以上四种分类进行划分，对每种类别求其完成次数的平均百分比，并将四种分类绘制称折线图，如图所示。

在假设参与人的单词储备并无较大差异的情况下，我们将四种类型单词分类的占比看作参与人人脑中储备四种单词量的平均占比。占比越大则猜想倒目标单词的概率越大。那么我们便可以得出四种分类的难度排序

DW1<DW2<DW4<DW3.

这个假设我们从折线图中也可以近似分析出，我们认为4次及之前答出的人数越多，5次及以后答出的人数越少则说明单词越容易，那么我们按照从左到右排序同样可以得出上述的难度排序。

接下来我们使用平均得分进行量化，我们将七个尝试次数的加权平均作为难度系数，

得出其具体的难度排序，最终排序相同，那么我们便可以得出单词中字母的重复率与得分占比有较强的相关性，即单词中字母的重复率对结果有较大影响。

Attribute 1: Number of word repetitions: Our group found that if there are repeated letters in the target word, the number of successful solutions or solution time will increase. In other words, since most of the individual's knowledge base are words without letter repetitions, we will rarely use words with repeated letters as target words in our questions. Since the target words are all 5 in length, we divide the words according to the number of letter repetitions into cases where each letter appears once, one letter appears twice, one letter appears three times, two letters appear twice each, and so on.

Based on common sense, we exclude cases that do not exist, including, but not limited to, where a letter appears 4 or 5 times and where the word is a combination of two letters. We divide words into the following four categories according to the number of letter repetitions.

W1: One occurrence of each letter (no word repetition)

W2: One letter appears twice

W3: One letter appears three times

W4: Two occurrences of each of the two letters.

The words in the data set were also divided according to the above four categories, and the average percentage of their completion times was found for each category, and the four categories were plotted as line graphs, as shown in the figure.

Under the assumption that the participants' word reserves do not differ greatly, we regard the percentage of the four types of word categories as the average percentage of the amount of the four types of words in the participants' human brain reserves. The larger the percentage, the higher the probability of guessing the target word. Then we can derive the difficulty ranking of the four categories

DW1<DW2<DW4<DW3.

This hypothesis can also be approximated from the line graph, where we assume that the higher the number of answers from 4 times and earlier, and the lower the number of answers from 5 times and later, the easier the words are, so we can also derive the above difficulty ranking by sorting from left to right.

Next, we quantify the difficulty using the average score, which is a weighted average of the seven attempts.

The final ranking is the same, so we can conclude that there is a strong correlation between the repetition rate of letters in a word and the score percentage, i.e., the repetition rate of letters in a word has a strong influence on the result.

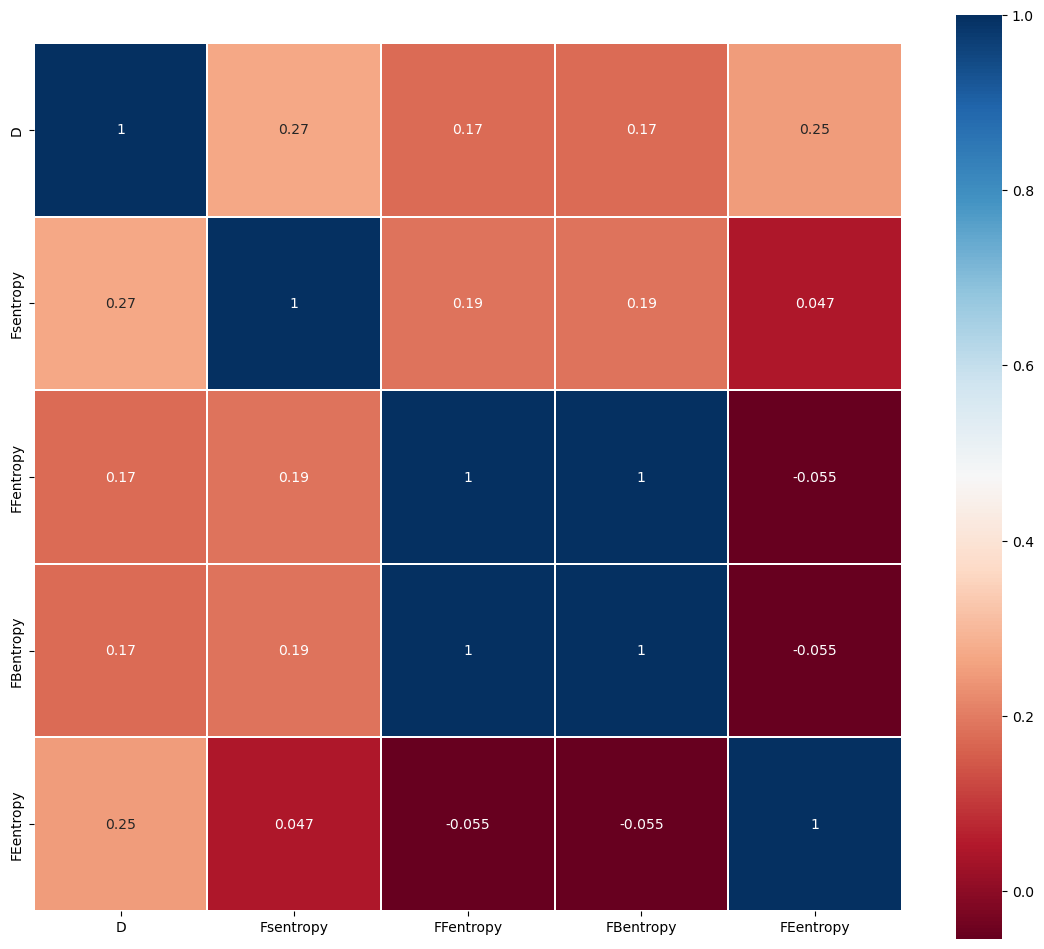
属性2：之后，我们通过查阅相关资料确定了另外四个属性，FS各个字母出现的频率；FF首字母出现的频率；FE单词的熟悉度；FB双单词中双音节频率。之后我们利用信息量计算公式来计算出每一个属性下单词的信息量，之后再计算每个属性分别与难度的Pearson相关系数。所得出的结果统计显著，总体呈现为低度相关性。计算信息量的公式如下：

|  |  |
| --- | --- |
| *;* |  |

其中为频率

我们分析由于单词的可选属性较多，并且尽不能较好地独立表征单词性质，但可以证明所选各属性与结果分布具有低度相关。所以我们认为将上述提到的四个属性进行加权得到的属性与结果分布具有较强相关。

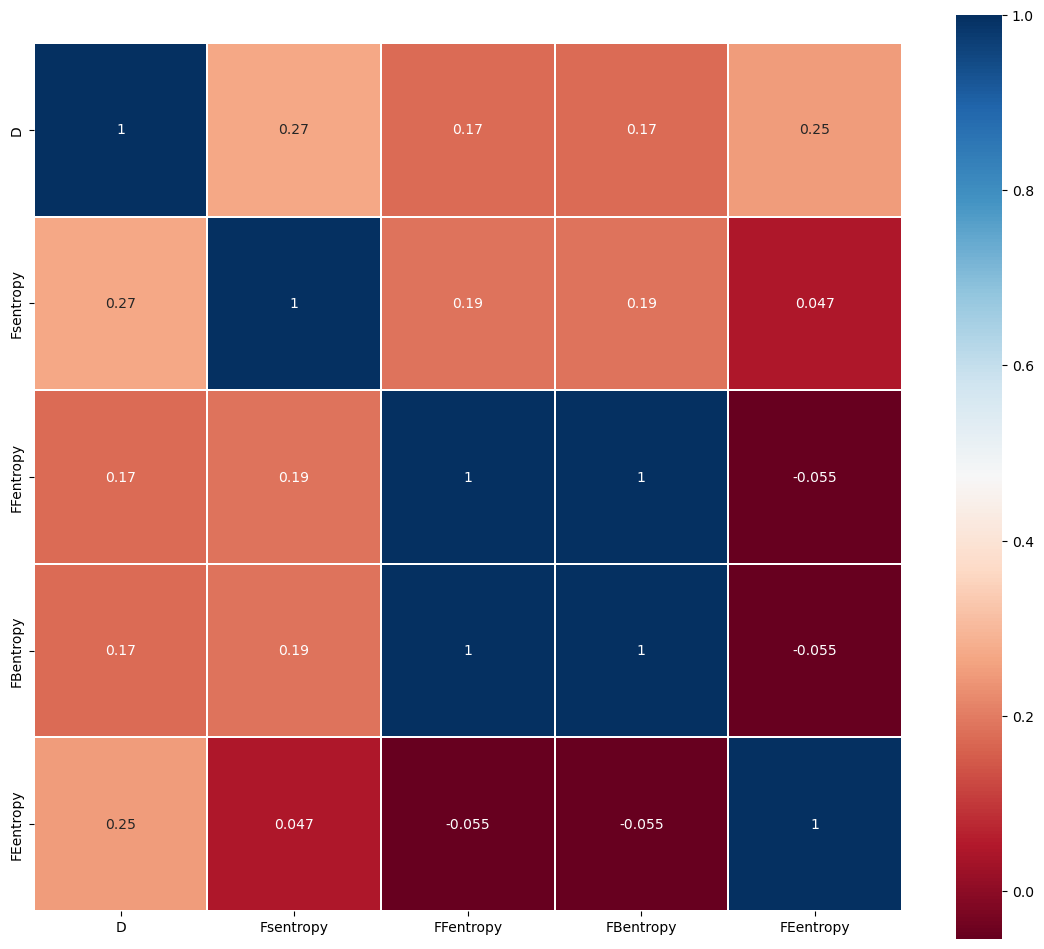
|  |  |  |
| --- | --- | --- |
| attribute | Pearson相关系数 | 显著性检验 |
| FS  FB  FF  FE | 0.268  0.173  0.173  0.249 | 显著  显著  显著  显著 |



Attribute 2: After that, we determined four other attributes by reviewing relevant information, the frequency of occurrence of each letter of FS; the frequency of occurrence of the initial letter of FF; the familiarity of FE words; and the frequency of diphthongs in FB double words. We then used the information content formula to calculate the information content of the words under each attribute, after which we calculated the Pearson correlation coefficient of each attribute with difficulty separately. The results obtained were statistically significant and showed a low correlation overall. The formula for calculating the information content is as follows.

|  |  |
| --- | --- |
| *;* |  |
| We analyze that due to the large number of optional attributes of the words and as much as they cannot characterize the word properties well independently, it can be demonstrated that each of the selected attributes has a low correlation with the outcome distribution. Therefore, we consider that the attributes obtained by weighting the four attributes mentioned above have a strong correlation with the outcome distribution. |  |

|  |  |  |
| --- | --- | --- |
| attribute | Pearson correlation coefficient | Significance testing |
| FS  FB  FF  FE | 0.268  0.173  0.173  0.249 | Significant  Significant  Significant  Significant |



# Quantity Prediction Model

为了解决问题1，我们尝试使用ARIMA、LSTM模型、GM模型来预测未来的报告结果数量，通过我们的实验发现，ARIMA模型是相对而言更加适合的预测模型，并且能够给出一个可靠的预测结果。经过我们小组的实验分析，2023年3月1日的报告结果数量预测在[6200,17200]这个区间内。

To solve problem 1, we tried to use ARIMA, LSTM model, and GM model to predict the number of reported outcomes in the future, and through our experiments, we found that ARIMA model is relatively more suitable prediction model and can give a reliable prediction result. After our group's experimental analysis, the number of reported outcomes on March 1, 2023 is predicted to be in the interval [6200,17200].

## ARIMA Model

之前我们已经分析了报告人数随时间的变化趋势，为了分析报告人数未来的变化趋势，我们使用ARIMA模型对未来60天内的报告人数进行预测，并按照题目要求得出3月1号的报告人数区间。ARIMA模型的一种基本方法为差分方法，即对时间序列进行差分，消除其趋势性季节性等特性，使得变换后的序列是平稳时间序列。此时便可将变换后的序列假设为ARMA序列进一步研究。

We have already analyzed the trend of the number of reports over time, and in order to analyze the future trend of the number of reports, we use the ARIMA model to predict the number of reports in the next 60 days, and to derive the interval of the number of reports on March 1 as required by the question. a basic method of the ARIMA model is the difference method, that is, the time series is differenced to eliminate its trend seasonality and other characteristics The transformed series is a smooth time series. At this point, the transformed series can be assumed to be an ARMA series for further study.

### Principle and implementation

For a zero-mean smooth sequence {X\_t,t=0,1,2,---}, if it can be expressed as a weighted sum of the first p terms and the sum of zero-mean smooth white noise, as follows:

|  |  |
| --- | --- |
|  |  |

and introduce the backward shift operator:

|  |  |
| --- | --- |
|  |  |

and the arithmetic polynomial:

|  |  |
| --- | --- |
|  |  |

where is a smooth white noise with zero mean variance of . Then is said to be an autoregressive series of order p, denoted as an AR(p) series. The model can be rewritten as:

|  |  |
| --- | --- |
|  |  |

If satisfies

|  |  |
| --- | --- |
|  |  |

Then call a sliding average series of order q, denoted as MA(q) series. As above, the model with the introduction of the backward shift operator can be rewritten as

|  |  |
| --- | --- |
|  |  |

If satisfies

|  |  |
| --- | --- |
|  |  |

Then call an autoregressive sliding average series of order p, q, denoted as ARMA (p, q) series, and the model can be rewritten after introducing the same backward shift operator as

|  |  |
| --- | --- |
|  |  |

The smoothness condition of ARMA model is φ(B) X\_t=0 all roots fall outside the unit circle, and the reversibility condition is α(B) = 0 all roots are outside the unit circle. The two properties are very important in theoretical and practical problems

Therefore, the time series we want to predict is differenced to eliminate its trend seasonality and other characteristics, and then it can be considered as a smooth time series, after which the ARMA model is used to fit the prediction. the specific steps of the ARIMA prediction model are.

Step 1: Calculate the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series we want to predict, and determine whether it obeys the ARMA model by judging whether it is truncated or trailing, and if at least one of them is not, the original series is non-stationary, then it is differenced to the first order, and the ACF and PACF are discussed until the differenced series is a stationary series

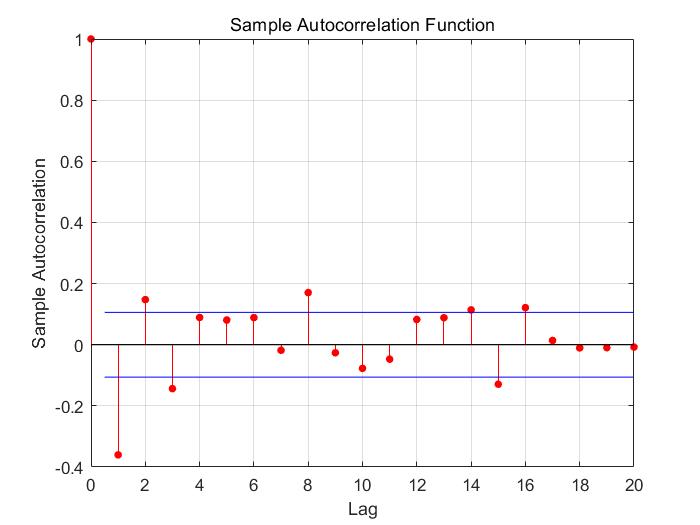
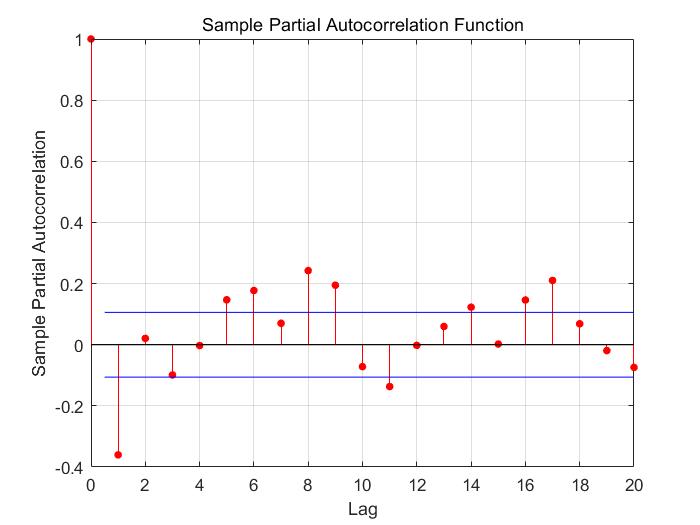
Step 2: Determine the parameters in ARIMA (p, 1, q): use the AIC and BIC criteria to fix the order and select the optimal parameters p, q

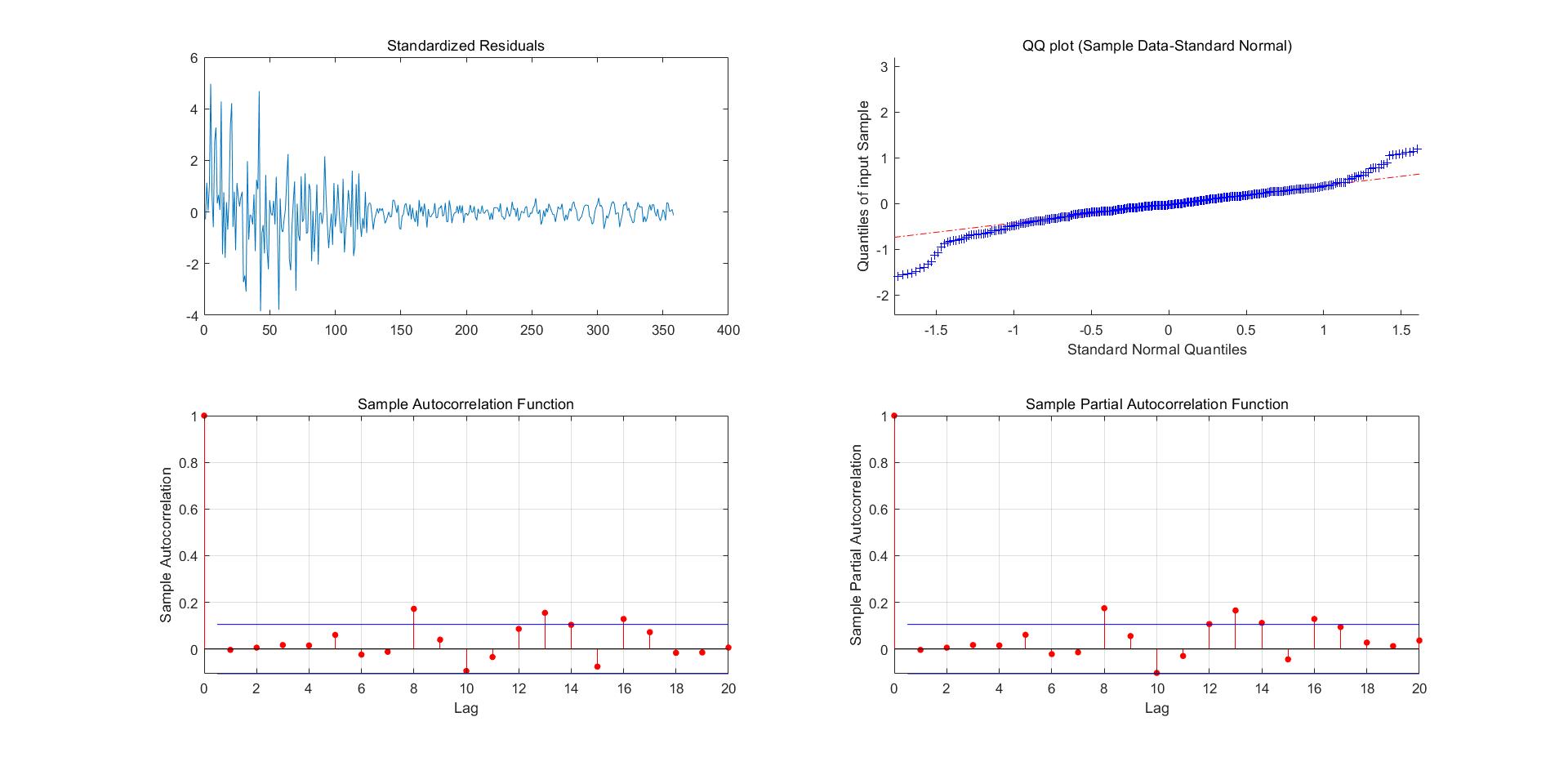
Step 3: Validation of the model: the residual test is performed to judge the reasonableness of the model, after which the model is used for prediction.

### Forecast Results

首先，我们计算报告人数的ACF和PACF，并对其进行平滑性检验，并未通过，在进行完一阶差分以后，再次进行平滑性检验，得到的逻辑输出结果如下表：

|  |  |
| --- | --- |
| Adf | 1 |
| Kpss | 0 |

一阶差分后的时间序列的ACF和PACF如下图**Figure X**所示  


之后我们进行参数定阶确定模型为ARIMA（3，1，5）时间序列预测模型，并对该模型进行残差检验，得到其标准化残差折线图，直方图、ACF图、PACF图以及QQ图。通过标准化残差图可以看出，残差在0附近随机分布，分析QQ图可以看出，绝大多数的点落在红线上。对所得到的误差进行Durbin-Watson检验（简称D-W检验）得到的DW统计量DW0=2.0066，与2极其接近，得出残差无自相关性。  


最终得到的模型：

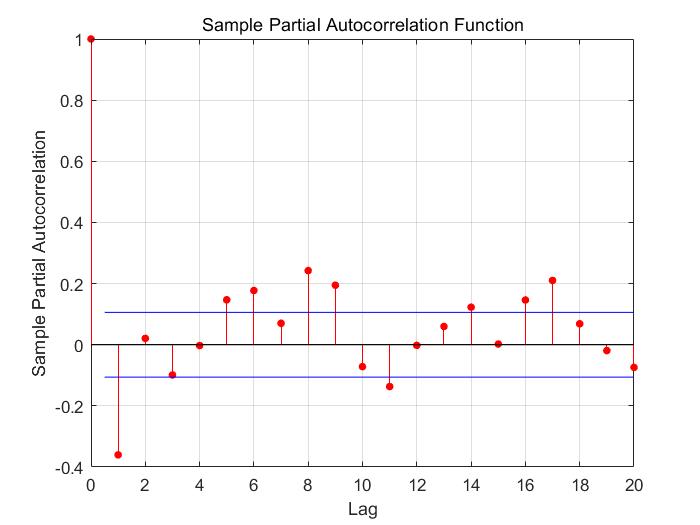
|  |  |
| --- | --- |
|  |  |

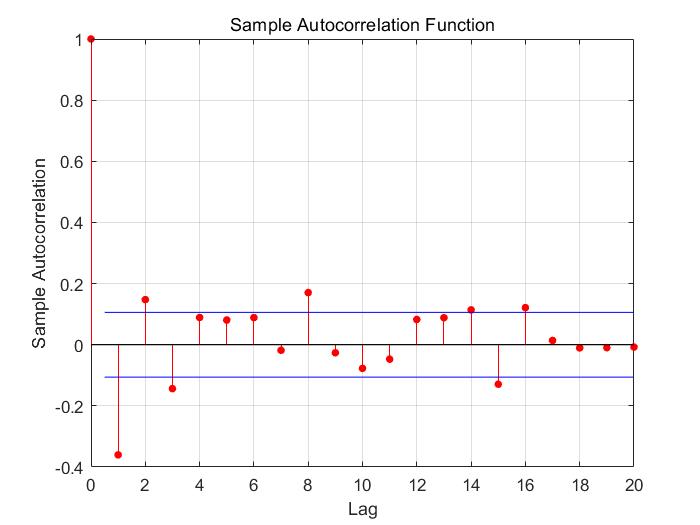
最后，利用该模型对2022年1月31号向后预测60天，得到2023年3月1号的数据.=17079，即在当天的报告人数约为17079。

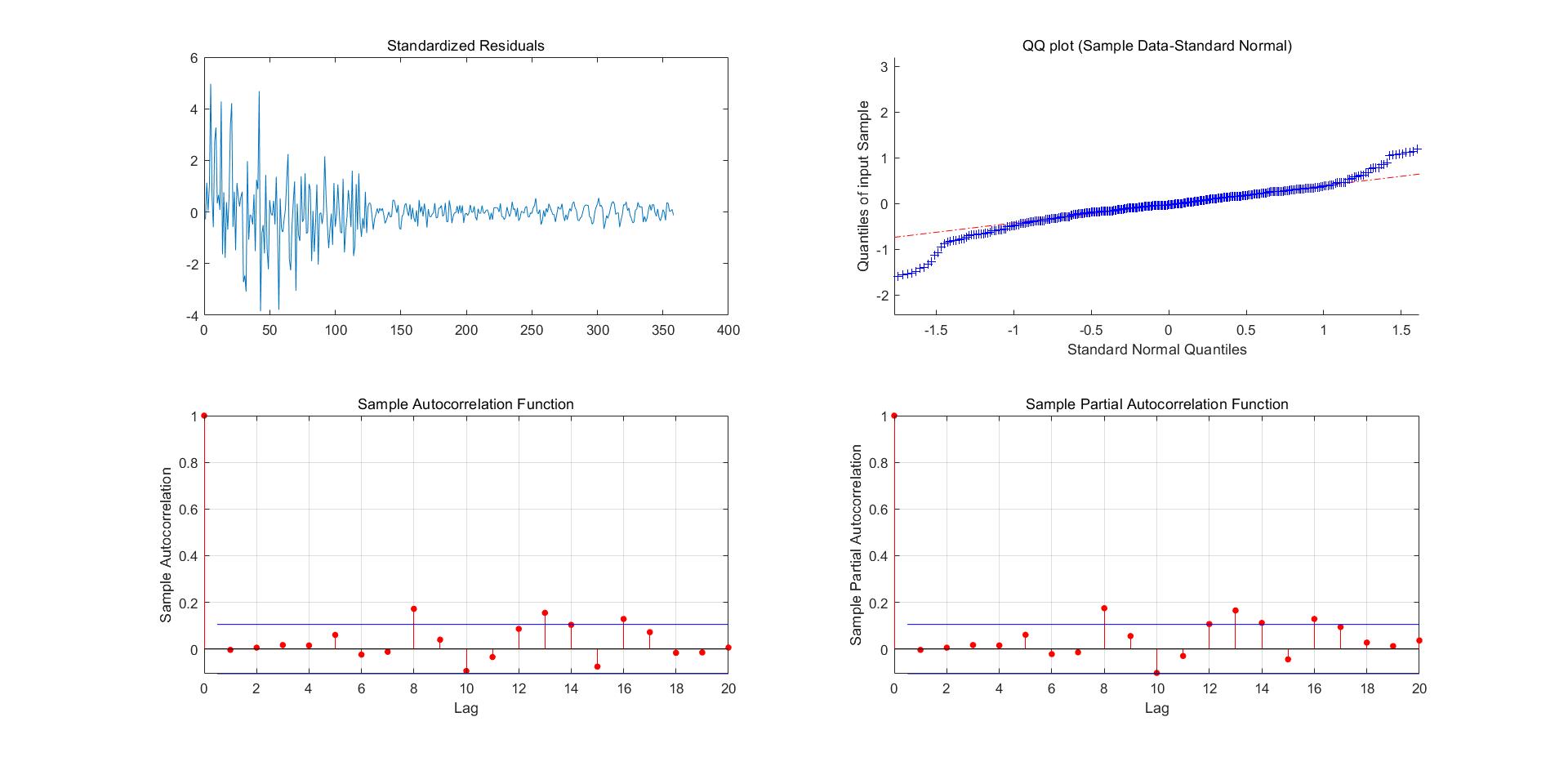
First, we calculated the ACF and PACF for the reported number of people and performed the smoothing test, which did not pass. After performing the first-order difference, the smoothing test was performed again and the logical output was obtained as shown in the following table.

|  |  |
| --- | --- |
| Adf | 1 |
| Kpss | 0 |

The ACF and PACF of the time series after first-order differencing are shown in Figure X below





Afterwards, we performed parameter sizing to determine the model as ARIMA (3, 1, 5) time series forecasting model and performed residual tests on the model to obtain its standardized residuals line plot, histogram, ACF plot, PACF plot, and QQ plot. The standardized residual plot shows that the residuals are randomly distributed around 0. Analysis of the QQ plot shows that the majority of the points fall on the red line. The Durbin-Watson test (D-W test for short) of the obtained errors yields the DW statistic DW0 = 2.0066, which is extremely close to 2 and yields no autocorrelation of the residuals.

The final model obtained:

|  |  |
| --- | --- |
| Finally, using the model to project backward 60 days for January 31, 2022, we obtain the data .=17079 for March 1, 2023, which means that the number of people reported on that day is approximately 17079. |  |

## LSTM Model

LSTM(Long-Short Term Memory)是一种循环神经网络的特殊变体，具有“门”结构。LSTM可以通过门单元的逻辑控制决定数据是否更新或是选择丢弃，这样预测的数据可以充分感知前面的数据，从而能够有效提高预测精度。LSTM 拥有三个门， 分别为遗忘门、输入门、输出门，以此决定每一时刻信息记忆与遗忘，输入门决定有多少新的信息加入到细胞当中，遗忘门控制每一时刻信息是否会被遗忘，输出门决定每一时刻是否有信息输出。我们小组尝试使用LSTM对Reported Results数据进行时间序列预测，得到结果如**Figure X**所示。

LSTM (Long-Short Term Memory) is a special variant of recurrent neural network with a "gate" structure. The LSTM has three gates, namely the forgetting gate, the input gate and the output gate, which determine whether information is remembered or forgotten at each moment, the input gate determines how much new information is added to the cell, the forgetting gate controls whether information is forgotten at each moment, and the output gate determines whether information is output at each moment. Our group tried to use LSTM to predict the time series of Reported Results data, and the results are shown in **Figure X**.

图表, 直方图

描述已自动生成

从结果图不难发现，使用LSTM模型非常近似地拟合了Reported\_results的变化趋势，并在预测任务上的表现也相当合理。

From the result graph, it is easy to see that the trend of Reported\_results is fitted very approximately using the LSTM model and performs reasonably well on the prediction task.

## GM Model

灰色预测是一种预测灰色系统的预测方法。灰色系统介于白色（完全已知）和黑色（完全未知）之间。灰色预测通过鉴别系统因素之间发展趋势的相异程度，即进行关联分析，并对原始数据进行生成处理来寻找系统变动的规律，生成有较强规律性的数据序列，然后建立相应的微分方程模型，从而预测事物未来发展趋势的状况

Gray forecasting is a forecasting method that predicts gray systems. The gray system is between white (completely known) and black (completely unknown). Gray forecasting is done by identifying the degree of dissimilarity of development trends among system factors, i.e., correlation analysis, and generating the original data to find the pattern of system changes, generating a data series with strong regularity, and then establishing the corresponding differential equation model to predict the future development trend of things

### Principle and implementation

Here we use the GM (1, 1) grayscale model, the GM (1, 1) bit first-order differential equation model containing only one variable. If the number of original samples is known to be n nn,here first define the tolerable coverage interval Θ as follows:

|  |  |
| --- | --- |
|  |  |

Define the column order ratio λ(k) as follows:

|  |  |
| --- | --- |
|  |  |

If the following conditions are met:

|  |  |
| --- | --- |
|  |  |

Then we say that x ( 0 ) x(0) x(0) is predictable by grayscale as GM(1,1) data.

We write the predicted value as x 0 ( k ) :

|  |  |
| --- | --- |
|  |  |

At the end of modeling, residuals and level difference tests are also required, and the residuals ρ are defined as

|  |  |
| --- | --- |
|  |  |

Define the level difference ξ as :

|  |  |
| --- | --- |
| ) |  |

### Forecast Results

图表, 直方图

描述已自动生成

从图中不难发现，灰度预测模型的拟合效果并没有ARIMA模型那么准确。因为灰度预测模型适合预测较短的时间序列，在该数据中，灰度预测模型的结果有偏小的趋势。

It is easy to see from the figure that the grayscale prediction model is not as accurate a fit as the ARIMA model. Because the grayscale prediction model is suitable for predicting shorter time series, the results of the grayscale prediction model have a tendency to be small in this data.

## Other Model

### Polyfit

我们尝试使用简单的多项式来拟合这种变化趋势，并且做出了1到6次多项式对数据的拟合结果如**Figure X**所示。

We tried to fit this trend using a simple polynomial and made a polynomial of order 1 to 6 to fit the data as shown in Figure X.

图表, 折线图

描述已自动生成

Figure X. 使用多项式函数拟合的结果图

从图中不难看出，多项式函数不适合拟合这样的变化趋势。当次数小时，没办法尽可能地算出在散点附近的结果；当随着次数增大，函数虽然更加贴近三点附近的位置，但会出现过拟合现象，即拟合的样本中表现良好而在预测的样本中有明显偏差。比方说图中的五次方程的结尾有急剧上升的趋势，而六次方程的结尾有急剧下降的趋势，这些都不是我们想要的。

From the figure, it is easy to see that the polynomial function is not suitable for fitting such a changing trend. When the number of times is small, there is no way to calculate the results as close as possible to the scatter point; when the number of times increases, the function, although closer to the location of the triple point, will appear overfitting phenomenon, that is, the fitted sample performs well while in the predicted sample there is a significant deviation. Let's say that the graph has a sharp upward trend at the end of the fifth equation and a sharp downward trend at the end of the sixth equation, which are not what we want.

# Difficulty Prediction Model

## Preparatory work

为了得到一个准确预测单词难度的模型，将6.1节挖掘的所有特征放入数据集中存为”word\_data.csv”，预览图如**Figure X**所示。将难度系数D标准化后的结果作为标签Y，其他的字段作为输入特征X，用于训练通过这些特征来得出难度系数的机器学习模型。在训练之前，我们将Word字段去掉，因为它属于数值类型。取而代之的是用位置编码posi来代替，单词每一位上的字母的编码，对应着字母频率的排名。这样就可以在保留Word信息的同时处理训练中不合法的数据。

表格

描述已自动生成

Figure X. word\_data数据集的预览图

For the above mined attributes, we use heat map to observe the correlation between features, especially with the difficulty coefficients, before we transform some difficult frequency coefficients into information entropy. The results are shown in Figure X图表, 树状图

描述已自动生成

Figure X.

To get an accurate model for predicting word difficulty, all the features mined in Section 6.1 are put into the dataset and saved as "word\_data.csv", and the preview is shown in Figure X. The normalized difficulty factor D is used as the label Y, and the other fields are used as the input features X. We use these features to train the machine learning model to derive the difficulty factor. Before training, we remove the Word field because it belongs to the numeric type. Instead, it is replaced by the positional encoding posi, the encoding of the letters on each bit of the word that corresponds to the ranking of the letter frequencies. This allows to handle the data that are not legal in the training while preserving the Word information.表格

描述已自动生成

Figure X. Preview image of the dataset word\_data

## Training result and model comparison

在划分好训练集和测试集后，我们选用了常见的几个模型训练并以验证的准确率作为标准来比较模型的优劣，这些模型包括Decision Tree,Naïve Bayes,Logistic Regression,KNN,Linear SVC,Perceptron,SGD,SVM，结果如**Figure X**所示。

|  |  |
| --- | --- |
|  | 图表, 条形图  描述已自动生成 |
| Figure X. | Figure X. |

Figure X.各个模型训练效果一览

从表中我们看出，在像Perceptron,SGD,SVM这类模型是明显不适合处理该类问题的(当然我不否认也有我们的模型没有得到很好调整的原因，但实际上所有模型都是从sklearn库中直接调用的)；而决策树在该数据集上有着非常惊人的表现，要知道，尽管样本数较少，用80%的样本去训练，从而100%通过剩下的20%的样本也是相当惊人的成绩。因此我们决定采用决策树来解决接下来的问题。

After dividing the training and testing sets, we chose several common models to train and use the accuracy of validation as a criterion to compare the models, including Decision Tree, Naïve Bayes, Logistic Regression, KNN, Linear SVC, Perceptron, SGD, SVM, and the results are shown in Figure X.

|  |  |
| --- | --- |
|  | 图表, 条形图  描述已自动生成 |
| Figure X. | Figure X. |

Figure X.各个模型训练效果一览

From the table we see that in models like Perceptron, SGD, SVM are clearly not suitable for this type of problem (of course I do not deny that there are reasons why our models are not well tuned, but in fact all models are called directly from the sklearn library); while decision trees have a very impressive performance on this dataset, bearing in mind that despite the small number of samples Using 80% of the samples to train and thus passing 100% of the remaining 20% is also quite an impressive performance. Therefore, we decided to use decision trees to solve the next problem.

## How difficult ‘EERIE’ is

我们的按照6.1的步骤获得’EERIE’这个单词的其他特征，并放入模型中进行预测，结果得到的预测标签值为107，标准化反转后的结果为难度系数4.32，这个难度放在’data1.csv’里意味着在358个单词中它比241个单词更难。而比它更难的基本上都是生僻词，尽管如此，我们仍认为’EERIE’的实际难度是更加大的，主要因为它非常罕见而且字母重复度很高。

我们认为造成这种误差的原因在于样本较小，使得训练出来的模型并不能很好的顾及到这一类单词，另外需要说明原数据集中并没有出现过一种元音字母重复三次的单词，在机器学习中，这会很大程度上阻碍模型的判断。总而言之，对于常用的或者与样本空间距离较小的单词，我们认为该模型有90%以上的准确度可以区分难度；而在样本中采集不到主要特征的单词，我们的预测率只保证在70%左右。

Our following the steps in 6.1 to obtain other features of the word 'EERIE' and put it into the model for prediction resulted in a predicted label value of 107 and a difficulty factor of 4.32 after normalized inversion, this difficulty put in 'data1.csv ' means that it is more difficult than 241 words out of 358 words. The words that are more difficult than it are basically rare words, but despite this, we believe that the actual difficulty of 'EERIE' is greater, mainly because it is very rare and has a high letter repetition.

We believe that this error is due to the small sample size, which makes the trained model not take into account this type of words very well, and the fact that there is no word in the original dataset that has a vowel letter repeated three times, which can greatly hinder the model's judgment in machine learning. All in all, for words that are commonly used or have a small spatial distance from the sample, we believe that the model has more than 90% accuracy to distinguish the difficulty; while for words that do not capture the main features in the sample, our prediction rate is only guaranteed to be around 70%.

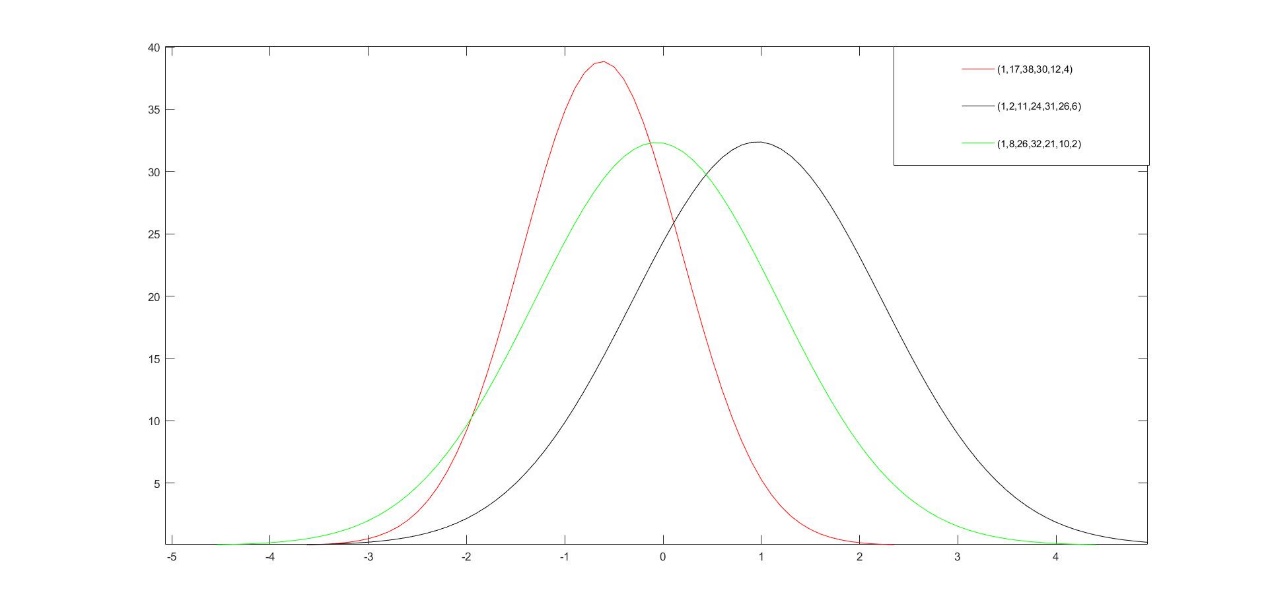
# Difficulty Evaluation Model

在第二问中，我们已经通过机器学习训练数据并预测得出了eerie的结果占比。为了实现对单词按照困难程度进行划分，我们首先需要找到作为划分依据的指标。

## 划分依据的指标确定

分析题目所提供的数据，我们得出能够最直观的体现单词难度的数据是完成次数(1, 2, 3, 4, 5, 6, X)各自的百分比。但若直接将百分比作为划分依据指标则较为复杂，因为数据为7维，并且这7维数据特性有差异。对于（1，2，3）的占比，我们可以大致认为占比越低单词难度越高，但细分下来，我们还需要确定1，2，3分别的对难度的贡献程度；对于（4，5，6，X）的占比，我们可以大致认为占比越高，单词难度越高，同样我们还要计算其各自的贡献程度。如此分析，若直接作为划分依据指标，需要确定的参数较多且复杂。对此，我们通过查阅资料与分析数据，提出了一种简化划分依据指标的方法：将每个单词的(1, 2, 3, 4, 5, 6, X)分布拟合为高斯函数，取高斯函数的期望作为聚类依据的指标

通过观察(1, 2, 3, 4, 5, 6, X)占比数据的分布规律，我们得出几乎每一个单词的(1, 2, 3, 4, 5, 6, X)占比都可以近似为高斯分布，并且所拟合出的高斯函数的期望与单词的难度有一定的关系。为了探究这种关系，我们首先从已有单词中挑选出3个(1, 2, 3, 4, 5, 6, X)占比分布差异较大的单词，并对其结果分布进行高斯拟合，并将三个拟合函数绘制在同一个图中，由高斯函数的性质：高斯函数的期望为其函数图像的对称轴，方差表现其图像的变化趋势。我们分析得出，三个拟合函数方差差异较小，即图像的0变0化趋势差异较小；主要差异表现在拟合函数的期望差异较大，并且（1，2，3）占比较高的单词对应的拟合函数期望值越小。



在此基础上，我们对所有单词对应的结果分布进行高斯拟合，得到第i个单词对应的高斯函数的期望值：和方差：

使用的拟合函数为：

于是我们得到了257个单词对应的结果分布的高斯拟合的期望和方差部分数据如下表

|  |  |  |
| --- | --- | --- |
| word | 期望 | 方差 |
| slump  crank  gorge  query  drink  favor  abbey  ··· | 0.06892  0.1769  0.7404  0.5015  -0.3959  0.6747  0.67  ··· | 1.452  1.829  1.805  1.738  1.451  1.898  1.7  ··· |

之后我们对和难度系数D进行相关性检验，得出其相关系数高达0.965，并且通过了显著性检验。于是我们得出了单词结果分布的高斯拟合的期望与其难度系数具有很强的正相关性。

之后我们使用K-means算法对257个期望值进行聚类，不同的类别对应不同的难度等级。K-means聚类算法是一种无监督分类算法。其将划分出的簇的均值当作该类簇的中心点，可以在不确定划分前提下，通过对数据集不断迭代对数据集划分，自动计算并更新每个簇的中心点。

算法的主要步骤为：

* 步骤一：选定聚类类别数k，并选择k个初始中心点
* 步骤二：针对样本点，找距离最近的中心点，并将其划为中心点代表的簇
* 步骤三：计算簇的均值得到新的中心点
* 步骤四：重新计算对象与中心点距离，重新划分。若聚类结果发生变化，则返回步 骤3；若无变化，则返回聚类结果。

我们对257个期望进行聚类。选定聚类类别数5，并选择5个初始中心点，将划分出来的5个簇按期望由小到大进行排列，划分为5个难度等级，分别为：easy、normal、medium、hard、ultimate.最终的得到的划分结果如下表：

|  |  |  |
| --- | --- | --- |
| Level | Expectation range | Number |
| Easy  Normal  Medium  Hard  Ultimate | （-1.016，-0.2587）  （-0.2782，0.1166）  （0.1201，0.5524）  （0.5599，1.262）  （1.513，2.942） | 54  127  113  57  6 |

之后，使用我们在第二问中建立模型求解所得到的单词eerie的结果分布。由于每次得出的结果分布有差异，并且部分数据结果分布之和与100相差较大，于是我们利用我们的决策树模型对eerie进行1000次的结果分布预测，并取删除和在（98.5，102.5）之外的数据，最终得到166条有效数据，然后我们再对166个结果分布求平均，得出了eerie较为可靠的结果分布

然后对该数据进行高斯拟合，得到的拟合函数的期望与方差如下(括号内为95%置信区间)：

比较得出eerie的难度为hard水平。可视化结果如下图，图中不同颜色的点代表不同的类，由上到下难度依次为：ultimate、hard、medium、normal、easy。红色方框表示的点代表eerie单词结果分布的高斯拟合所得期望所在位置

图表, 散点图

描述已自动生成

模型准确性讨论：模型的准确性主要与数据集以及单词eerie在高斯拟合时产生的误差有关。之前提到我们将数据集高斯拟合所得到的期望与我们对（1，2，3，4，5，6，X）的占比的加权平均（初步定义的难度系数）进行了相关性检测，得到高达R=0.965的显著相关性。所以我们将R=0.965作为我们模型的评价系数之一。经过验证，由于分类的区间较大，高斯拟合所得期望产生的误差在计算对分类的准确性影响时可忽略不计。又由于我们使用第二问的模型预测的结果作为我们的输入，那么我们再考虑该模型的准确性时需要将第二问模型的准确性考虑进来。最终我们得到分类模型的准确性：A2=R\*A1；得到结果A2=84.92%

因此我们认为我们的模型的准确率为84.92%

# Sensivity Analysis

**对于难度评估模型的灵敏度检验：**

假设一些影响因素使得（2）占比发生了100%左右的向上浮动，其增加的百分比在（3，4，5，6）上平均减少。在此种情况下我们得到期望值为0.7724，分类仍为hard。同理，若（3）占比发生100%的上浮，其增加的百分比在（4，5，6）上平均减少，此种情况下我们得到的期望值为0.6781，分类仍为hard.。若（4）占比发生了20%的向上浮动，其增加的占比在（2，3）上平均减少，得到的期望值为0.683，分类仍为hard。我们得出在任意一个占比上的合理情况下的扰动会对期望值产生影响，但对分类结果并无影响。因此，我们得出我们的模型用于分类较为稳定。

# Model Evaluation

## Strengths

1. 在Quantity Prediction Model中使用多个模型做预测，在比较拟合、预测效果的同时，也保证了预测区间的准确度。
2. 引入了困难系数的概念，它直接表现为原数据集中成绩分布的期望。而事实上，谜题的难度基本上由单词本身的属性决定的，因此在训练任务中，它作为预测难度的标签，使得我们能够对单词的特征进行监督训练。
3. 充分考虑了单词本身的属性，不仅挖掘了原数据集中能够推断出来的属性，如元音的个数和位置，字母的重复次数等，还通过在网上查阅单词、字母的使用频率等数据获得更多特征，使得训练过程更精确。
4. 决策树模型易于理解和解释，适用于类似题目给定的小数据集。

(1) Multiple models are used to make predictions in Quantity Prediction Model, which ensures the accuracy of the prediction interval while comparing the fitting and prediction effects.

(2) The concept of difficulty coefficient is introduced, which is directly expressed as the expectation of the grade distribution in the original data set. And in fact, the difficulty of the puzzle is basically determined by the properties of the words themselves, so it serves as a label for predicting the difficulty in the training task, enabling us to perform supervised training on the features of the words.

(3) The attributes of the words themselves are fully considered, not only mining the attributes that can be inferred from the original dataset, such as the number and position of vowels and the number of repetitions of letters, but also obtaining more features by consulting data on the Internet on the frequency of words and letters used, making the training process more accurate.

(4) The decision tree model is easy to understand and interpret, and is suitable for small data sets given by similar topics.

## Weaknesses

1. 对Reported Result数量的预测仅从以往数量本身出发，并没有考虑其他外部因素的影响。
2. Difficulty Prediction Model需要更大的数据集来支撑。因为样本数少意味着很多单词的潜在属性都没办法被训练到，一旦遇到生僻词，模型将容易做出误判。
3. 能够挖掘到的特征相当多，但基本不对难度起决定性因素，这使模型冗余复杂且训练成本较高。

(1) The prediction of the number of Reported Results starts only from the previous number itself, and does not consider the influence of other external factors.

(2) Difficulty Prediction Model needs a larger dataset to support. Because the small sample size means that many potential attributes of words are not trained, the model will be prone to make misjudgment once it encounters a rare word.

(3) The number of features that can be mined is quite large, but basically does not play a decisive role in the difficulty, which makes the model redundant and complex and expensive to train.

# References

1. “Wordle-The New York Times.” The New York Times, 2022. Accessed December 13, 2022 at <https://www.nytimes.com/games/wordle/index.html.>
2. “Wordle-The New York Times.” The New York Times, July 21, 2022.
3. “Wordle Stats.” Twitter, July 20, 2022.
4. Dataset of English Word Frequency at <https://www.kaggle.com/datasets/rtatman/english-word-frequency>
5. Dataset of Wordle Valid Words at <https://www.kaggle.com/datasets/bcruise/wordle-valid-words>
6. Letter frequencies as recorded in Wikipedia at <https://zh.wikipedia.org/wiki/%E5%AD%97%E6%AF%8D%E9%A2%91%E7%8E%87>

# Letter for the Puzzle Editor of the New York Times

Dear Puzzle Editor of the New York Times:

最近我们总是在社交媒体上被大批莫名奇妙的格子图猜谜游戏wordle令我们十分着迷，我们都想着是否能用最少的次数解出谜题。每天都有数以万计的人们在推特上分享他们的猜测结果，形成一股猜词风潮。语言学家、统计学家、计算机学者也纷纷开始研究wordle，运用他们领域的知识来分析谜题，求解单词。

我们参考了2022年1月7日到12月31日的数据，发现每天在推特上分享他们成果的人都在逐渐变化，我们建立一个模型解释这种变化，并希望能给出未来某一天在推特上分享他们wordle结果的人数区间，我们使用了Aria模型、GM模型、LSM三个在机器学习领域常见的模型来进行时间序列预测。时间序列预测是通过对按照一系列时间间隔排列的一组数据的分析，来进行预测，发现现象和规律，比如这里的每天在推特上分享他们wordle答案的人数。通过三种模型，我们得出预测区间大概（）

（词语的属性影响报告在困难模式的百分比）

（预测各猜中次数区间）

同样的，在实际做题中，我们总会感慨今天的词好难，今天的词比较容易。我们知道（As we all know）wordle词库中只有12937个，我们是否能过通过大量数据真的对每个单词给进行难度分类，判断今天遇到的单词是否真的是颇具难度的。我们同样地采用2022年1月7日至12月31日的数据来进行预测。在这里我们考虑单词本身的性质，选取了字母频率、双音节频率、字母重复度、含有元辅音个数、单词常见度、单词每一位上的字母的编码个可以通过分析单词能得出的特征。在这里需要特别注意的是单词常见度这一特性，生活中越常见单词就越好猜，人们也往往会猜测自己较为常用的单词，我们通过大文本统计了wordle词库中较为常见的单词，这些数据可以在kangol中查到。

我们将挖掘到的所有特征存入数据集,将难度系数D标准化的结果作为标签Y,其他的字段作为输入特征X。我们选用了大量的机器学习模型来进行训练，这些模型包括Decision Tree,Naïve Bayes,Logistic Regression,KNN,Linear SVC,Perceptron,SGD,SVM。在划分好训练集和测试集后，我们选用了常见的几个模型训练并以验证的准确率作为标准来比较模型的优劣。我们惊人的发现，决策树训练效果甚至达到了百分之一百，而perception、SGD、SVM等几个模型，则只有百分之几的准确率。

我们选取了eerie这个具体的单词，希望能获得关于他的难度分类。向上面一样我们获取这个单词的特征，并放入难度模型中进行预测，我们得到的预测标签值为107，标签化反转后的结果为难度系数4.32，这个难度放在’data1.csv’里意味着在358个单词中它比241个单词更难。而比它更难的基本上都是生僻词，尽管如此，我们仍认为’EERIE’的实际难度是更加大的，主要因为它非常罕见而且字母重复度很高。

I am writing to express my appreciation for your puzzles. They are always challenging and engaging, and I look forward to seeing a new puzzle every day. I think the puzzles are a great way to exercise my mind and learn something new every day.Thank you for your hard work in producing these puzzles each day. I look forward to seeing what new challenges lie ahead(最后写点客套话)

Sincerely,three college students

# About the Authors

# Appendix