给纽约时报编辑的信：

Letter to the Editor of the New York Times.

我们发现了给定的数据集中有几个异常值，有的字符串只有四个字母，也有的单词的报告数量与前后日期对比差距过多，我们将以上情况视为异常值，首先对其进行了剔除。

We found several outliers in the given dataset, some strings with only four letters and some words with too many gaps in the number of reports compared to the before and after dates, we considered the above as outliers and removed them first.

我们选择了使用指数函数拟合的方法来对未来某天的报告数量进行预测。通过该方法，我们最终得到一个拟合公式，代入要预测的日期，即可得到想要的预测结果及95%置信区间。

We chose to use an exponential function fitting method to make a prediction of the number of reports for a future day. By this method, our final fitting formula was obtained and substituted into the date to be predicted to get the desired prediction result and 95% confidence interval.

另外，我们为单词制定了单词的词性、单词使用频率、元音字母个数、首字母等属性，通过对这些属性与在困难模式下播放的分数的百分比进行可视化分析，我们可以得出哪些属性对在困难模式下播放的分数的百分比是有明显的影响的。通过对结果的分析，我们最终选择了单词的使用频率、元音字母的个数、字母频率以及单词所拥有的字母的种类数作为特征来进行下面的分析。其他的属性由于对结果的影响不大，我们选择将其舍弃。

In addition, we developed attributes for words such as the word's lexicality, frequency of word usage, number of vowel letters, and initial letters. By visualizing these attributes with the percentage of scores played in the difficult mode, we can conclude which attributes have a significant effect on the percentage of scores played in the difficult mode. By analyzing the results, we finally selected the frequency of word usage, the number of vowel letters, the frequency of letters, and the number of types of letters that the words have as features for the following analysis. The other attributes were chosen to be discarded because they did not have a significant impact on the results.

我们建立了BP神经网络，采用Levenberg–Marquardt algorithm和Bayesian Regularization algorithm两种不同的算法分别预测给定解决方案词的百分比。对比选择最优方案我们采取提取到的四种有效属性作为输入，通过多次调整参数和训练，在上百次独立重复实验中可以得到较稳定的eerie百分比分布预测结果大概为【1 7 17 26 27 17 6】。

We built a BP neural network and used two different algorithms, the Levenberg-Marquardt algorithm and the Bayesian Regularization algorithm, to predict the percentage of words for a given solution, respectively. Comparing to choose the optimal solution we take the four extracted valid attributes as input, and by adjusting the parameters and training several times, we can get more stable eerie percentage distribution prediction results roughly [1 7 17 26 27 17 6] in hundreds of independent repetitions of the experiment.

我们建立了一个随机森林，根据属性把各单词进行分类。我们对EERIE这个单词的难度进行了预测，最终得出它属于中等难度，我们计算出它大概平均四次或者五次可以被猜出。通过对已知数据集进行划分并检验，我们发现我们的模型拥有80.4%的准确率。

We built a random forest to classify each word according to the attributes. We predicted the difficulty of the word EERIE and finally concluded that it is of medium difficulty, and we calculated that it can be guessed about four or five times on average. By dividing and testing the known data set, we found that our model has an accuracy of 80.4%.

除此之外，我们还发现了一些该数据集的有趣的属性。我们使用在困难模式下播放的分数的百分比除以总报告数量，得到了一个比值HMR(Hard Mode Rate)。我们将HMR随时间的变化可视化地表示了出来，最终发现困难模式得分占比在缓慢提高，最终稳定在了一个稳定的水平，我们推测这可能与游戏玩法被开发或者游戏玩家的水平提高有很大的关系。

In addition, we found some interesting properties of this dataset. We obtained a ratio HMR (Hard Mode Rate) using the percentage of scores played in hard mode divided by the total number of reports. We visualized the change in HMR over time and eventually found that the percentage of Hard Mode scores was slowly increasing and eventually stabilized at a steady level, which we speculate may have a lot to do with the game play being developed or the level of gamers improving.

在近几天我们每天早上都会打开Wordle的网站一起来玩这个有趣的游戏，在游戏的过程中，我们也总结出了一些浅显的玩法，对此我们感到很开心。尽管有些时候我们并不能猜出来答案，但是它确实使我们认识了一些比较有意思的单词！

In the last few days we have been opening Wordle's website every morning to play this interesting game, and in the process of playing, we have also come up with some obvious ways to play, which we are happy about. Even though we couldn't guess the answers some of the time, it did help us to know some interesting words!