Wordle是由《纽约时报》每天推出的一种受欢迎的益智游戏，我们基于马尔可夫假设使用了n-gram模型，提出了针对游戏Wordle的单词难度指数，并根据指数拟合和最小二乘法开发了一套模型，可以预测每天报告结果的数量和分布。

首先我们利用n- gram模型和构建了一个单词联想度模型，用于评估单词的在wordle游戏中的难度，并利用困惑度指标，进行灵敏度分析，最终选择了Tri-gram模型。然后采用k-means聚类将单词难度分类分为10类。利用该难度指标评估出EERIE的难度分类为7级，属于在游戏中很有猜测难度的单词

对于报告结果的预测，我们采取了分层预测模型。在这个模型中，我们将时间序列分成两层，首先预测每周的整体分布情况，然后预测每周内的分布情况。根据报告结果人数的一阶导数和二阶导数，我们选择了恰当的预测数据训练集。随后，我们利用指数拟合、GM(1,1)和ARIMA（0，2，0）模型，分别预测了第59周的数据。通过对预测结果的分析，最终采用了指数拟合的结果。结合每周内的分布情况，我们预测2023年3月1日在置信区间95%下的报告结果数量的预测区间为【11637.337，25646.725】，取中间值为18642.031

我们根据相关性分析判断时间以及每天的报告结果数量和报告结果分布无关，单词难度分类与报告结果分布呈强相关，我们采用了最小二乘的思想和偏态分布建立了预测报告结果的分布的模型。并得到了2023年3月1日的的不同尝试次数的提交分布。其中，RSS=3300.92，模型准确度有待加强。

此外我们对单词的属性，元音字母数量、词频、单词的难度等级和困难模式比例做相关性分析，得出单词的难度分类和困难模式比例强正相关，相关系数为0.8102，其余无相关性。

随着我们对数据继续挖掘，惊奇的发现：

·随着时间的变化，困难模式通过比例逐渐上升；

·元音字母较多和较少都会导致单词难度更大；

·难度增大，困难模式通过比例上升。

在正文中我们解释了对此的看法。

最后我们基于Tri-gram模型对马尔可夫假设进行优化，考虑不连续情况，我们提供了一个wordle答题策略，指导玩家尽可能地高效通过游戏。

