Problem Chosen	2023	Team Control Number
A	MCM/ICM	2318982
	Summary Sheet	

世界:一个字母就能改变世界

摘要：自 2022 年初推出以来，世界码在社交媒体上引发了一波分享黄色、绿色和灰色方块的热潮。世界有简单但具有挑战性的规则，只需要很短的注意力持续时间。基于世界数据集，我们深入挖掘隐藏在数字和报告结果百分比背后的信息。

首先，我们关注随时间变化的报告结果数量。我们尝试建立一个 ARIMA 模型，为 2023 年 3 月 1 日报告的结果数量提供预测区间。这表明世界在发布一年后仍然保持着很高的热情。然后，我们探讨了影响硬模百分比的因素。通过拟合多元线性回归模型，结果表明，重复字母的数量和单词出现的频率与游戏难度相关。玩家事先从社区获得的难度信息可能会影响他们对游戏模式的选择。

接下来，我们好奇的是，报告结果的分布在未来会发生怎样的变化。为了简化模型，我们将玩家的游戏状态概括为他们已知的每种颜色的方格数。然后可以将世界建模为马尔可夫链，并将问题转化为求解它的初到分布。这需要了解初始分布和依赖于玩家选择的策略的转移概率。此外，转移概率被假设依赖于状态之间信息量的差异。因此，我们提出了一种测量状态中当前信息量的方法。在此基础上，建立了整个马尔可夫链模型，求解了不同策略下的第一到达时间分布。

为了使模型更加合理，我们假设选择上述两种策略的人的比例随时间而变化。据此，提出了一种基于历史数据的方法来估计这一比例。最后，我们将估计的比例与高斯过程回归模型相结合，预测未来玩家策略选择的比例。然后将其与马尔可夫链模型相结合，以预测未来报告结果的分布。最后得到 EERIE 的分布为(0.00,0.15,11.05,28.44,35.46,21.16,3.76)。

最后，我们要根据单词的难易程度对其进行分类。由于单词的难度只与单词本身有关，因此我们认为，根据单词属性进行聚类可以反映单词的难度水平。针对这一思路，进行了 k - prototype 聚类，并设置了合理的单词难度指数。然后，我们提取每个类别的难度信息，然后绘制密度函数并计算 Kullback-Leibler 散度。两种结果都表明，不同属性的词具有不同的难度等级。这证明了我们的思路是合理的，分类模型是准确的。进一步，我们根据 EERIE 的属性将其划分为“硬”类，这与上面得到的百分比分布是一致的。此外，我们还讨论了关于数据集的其他信息，如难词、易词和意外词。最后，对模型的敏感性分析表明，我们的模型具有良好的鲁棒性。

关键词:ARIMA;多元线性回归;马尔可夫链;k - prototype 聚类

目录

世界:一个字母就能改变世界 1
1 介绍 4
1.1 背景 4
1.2 问题重述(Restatement of The Problem 4
1.3 我们的工作 4
2 模型假设和注释 5
2.1 假设与证明 5
2.2 符号 5
2.3 数据预处理 6
3 任务 1:数字预测和单词属性 6
4.模型拟合 9
4.1 残留分析 9
4.2 词属性的影响 10
4.2.1 The Word 的属性 10
4.2.2 多元线性回归 11
5 任务 2:基于马尔可夫链模型的分布 14
5.1 状态空间 14
5.2 初始分布 15
5.3 转移概率 16
5.4 报告结果的分发 17
5.5 使用两种策略的比例 17
5.6 预测未来报告结果的分布 18
6 任务 3:解词分类 18
6.1 难度评分 18
6.2 k - prototype 聚类 18
6.2.1 求解步骤 18
6.2.2 结果 19
6.3 解词难度分类 19
6.4 EERIE 这个词的难度 19
7 任务 4:其他有趣的特征 19
8 敏感性分析 21
9 模型评估和进一步讨论 21

9.1 优势 21
9.2 缺点 21
10 给拼图编辑的一封信 22
References 23

1 介绍

1.1 背景

Wordle 是《纽约时报》每天提供的一款很受欢迎的五字母解谜游戏，玩家尝试在 6 次或更少的时间内猜出正确的单词，每次猜出都会得到反馈。它有 60 多种语言版本，有两个级别:普通模式和困难模式。在困难模式中，正确猜测的字母必须在随后的猜测中使用。猜中后，瓷砖会变色:黄色=字母放错位置，绿色=字母放对位置，灰色=未包含字母。

1.2 问题重述(Restatement of The Problem

考虑到题目中给出的背景资料和相关条件，我们需要解决以下问题:

开发一个模型来解释报告结果的每日变化，并使用它来创建 2023 年 3 月 1 日结果数量的预测区间。困难模式分数的百分比是否受到单词属性的影响?如果是，如何影响?如果没有，为什么不呢?

开发一个模型来预测特定未来单词的解(1,2,3,4,5,6,X)分布。讨论与预测相关的不确定性。提供 2023 年 3 月 1 日 EERIE 的预测示例，以及模型的置信度。

创建一个分类模型，根据单词的难度对它们进行分类，并描述每个单词的特定属性。根据模型，EERIE 这个词有多难?评估模型的准确性。

最后，描述数据集中其他有趣的特征。

1.3 我们的工作

结合背景和问题，我们的工作主要包括以下几个方面:

我们假设可以通过构建具有最优参数的 ARIMA 模型来预测 2023 年 3 月 1 日的报告结果数量。为了进一步深入了解单词属性，我们运行了一个多元线性回归来检验单词属性对难度模型中报告的分数百分比的影响。

我们将玩世界游戏的过程建模为离散状态马尔可夫链，并基于导出的信息推导出两种游戏策略。然后，我们使用信息熵和马尔可夫链性质等理论工具估计了两种策略报告结果的分布。随后，将获得的结果结合起来，对未来日期报告结果的分布进行预测。

此外，任何给定单词的难度是由其属性决定的。因此，根据属性对单词进行聚类可以为每个类别的难度提供有价值的见解。

最后，在对数据集进行仔细分析后，我们观察到几个值得注意的特征。

为了避免复杂的描述，直观地反映我们的工作过程，流程图如图 1 所示。

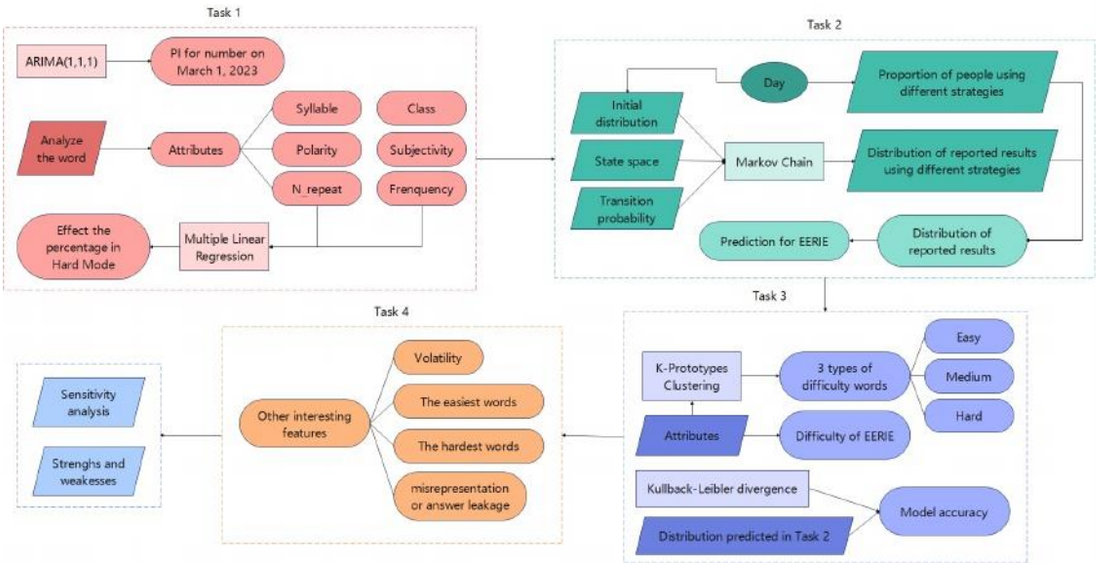

图 1:我们的工作流程图

2 模型假设和注释

2.1 假设与证明

为了简化问题并方便我们对世界游戏的建模，我们做出了以下基本假设，每个假设都有适当的理由。

(1)在玩世界游戏时，玩家总是基于他们当前所知道的信息做出局部最优选择。

(2)大多数时候，世界游戏的正确答案来自于更常见的单词。

(3)玩家在玩世界游戏时是理性的，一般不会轻易抛弃已知信息。

(4)玩家通过从大致相同的单词库中选择填充词来玩世界游戏，并且他们对一个单词是否是正确答案的主观概率是相同的。

(5)一个事件的信息增益越高，其出现的概率越低

(6)在不同大小的同义词典上测量一个词的信息含量，得到的结果大致相同。

(7)使用每种策略的玩家比例在一天中没有明显变化。

2.2 符号

表 1:标注

Symbols	Description
A_i	The set of states that are reachable in one step of state i.
S	The state space of the Markov chain.
W	All the words a player may fill in.
p_x	The subjective probability that word x is the correct answer.
$freq_x$	The word frequency of word x.
I_x	The amount of information obtained by filling in the word x at the opening.
$x_{true}^{(r)}$	The correct word of the r th day.
G_i	The set of words that the player has guessed when he is in state i
$p_k^{(r)}(i,j)$	The transfer probability from state i to j in Markov chain on day r.
$T_j^{(r)}$	The number of steps to first reach state j from state i on the Markov chain at day r.
$C_k^{(r)}$	The set of absorbing states of Markov chains on day r.
$T_{absorbed}^{(r)}$	Number of steps before falling into an absorbing state on Markov chain at day r.
$q_k(r)$	The proportion of all players using strategy k on day r.

其中我们定义了主要参数，而这些参数的具体值将在后面给出。

2.3 数据预处理

由于 COMAP 官方只允许我们使用“Problem_C_Data_Wordle.csv”数据集，所以我们需要在解决问题之前对数据集进行预处理。对数据集的初步检查显示，存在一些异常值和缺失值。

在单词列中，我们发现一些单词的长度不等于 5，例如“rprobe”、“clen”和“tash”。正如 COMAP 官方提到的，在第 18 行，对于 545 号比赛，所列的单词是“rprobe”，而应该是“probe”。通过查阅世界发布的今日解词，我们也得到了“clen”应该是“干净的”，“tash”应该是“垃圾”。

此外，在第 34 行，对于 529 号比赛，列出的报告结果数为“2569”，而

正确的数字应该是“25569”。

3 任务 1:数字预测和单词属性

在本节中，我们通过构建 ARIMA 模型并选择最优参数来预测 2023 年 3 月 1 日报道的结果数量。然后我们总结单词属性，然后通过构建多元线性回归来探索单词属性对难度模型中报告分数百分比的影响。

自回归综合移动平均，即 ARIMA，是一种利用时间序列数据预测未来趋势的统计分析模型。ARIMA 的基本思想是将预测随着时间的推移而形成的数据序列看作是一个随机序列和一个模型可以用来近似描述这个序列。一旦确定了这个序列，模型就可以从时间序列的过去和现在的值来预测未来的值。ARIMA 模型包括自回归(AR)模型和移动平均(MA)模型。AR 模型描述当前值与滞后值之间的关系，并利用历史数据预测未来值。MA 模型利用过去残差项的线性组合来观察未来残差。ARIMA 预测模型可写成如下公式:

$$\hat{p}^{\{t\}} = p_0 + \sum_{j=1}^{p} \gamma_j p^{\{t-j\}} + \sum_{j=1}^{q} \theta_j \varepsilon^{\{t-j\}}, \tag{1}$$

式中，p 为自回归模型(AR)阶数，q 为移动平均模型(AM)阶数，ε{t} 为时间 t 至 t - 1 之间的误差项，γj 和 j θ为拟合系数，p0 为常数项。

根据本例的相关情况，综合考虑各种时间序列分析模型后，我们选择 ARIMA 模型对报告的结果数量进行分析，并给出 2023 年 3 月 1 日的预测区间。

我们的模型构建和求解步骤如下:

1.时间序列的预处理

(1)划分训练集和测试集

为了更合理地评价模型，将数据分为训练集和测试集。

(2)平稳性和白噪声检验

从观察到的记录对随机过程的结构进行统计推断时，通常需要对其做出一定的简化(近似合理)假设。这些假设中最重要的是平稳性。平稳性的基本思想是，决定过程性质的统计规律不随时间而改变。

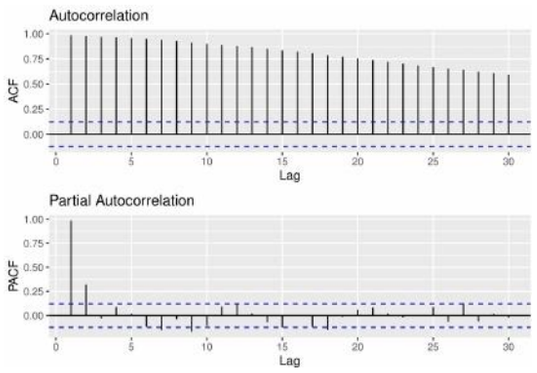

图 2:原始时间序列相关函数图

在进行平稳性检验之前，首先通过简单的 ACF 图来识别序列的季节性因素，时间序列没有明显的季节性趋势。ACF 图、PACF 图和时间序列图如图 2 所示。

通过绘制报告结果数量的时间序列，我们发现该序列看起来并不平稳。对于非平稳过程，使用差分方法进行转换。从下面的图 3 和图 4 可以观察到，差分后的序列总是围绕某一值随机波动，没有明显的趋势。

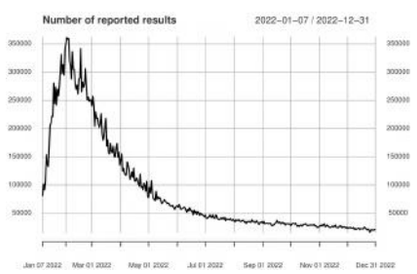

图 3:原始时间序列

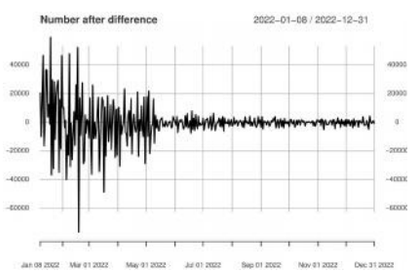

图 4:一阶微分时间序列

采用增强 Dickey-Fuller 检验检验平稳性。发现原始序列在α= 0.05 的水平上没有拒绝序列不平稳的原始假设。一阶差分之后的时间序列则拒绝原始假设，并通过平稳性检验。差分前后的统计值如下表 2 所示。

表 2:增强 Dickey-Fuller 检验

	Dickey-Full statistics	p value	lag
before difference	-2.8951	0.1989	6
after difference	-5.0648	0.01	6

进一步确定该序列是否为白噪声序列。在白噪声序列中，任意两个时间点的随机变量是不相关的，序列中没有可以利用的动态规律。因此，历史数据不能用来对未来进行预测和推断。从数据来看，Box-Ljung 白噪声检验的统计值为 253.36,p = 0.000。原始假设序列是白噪声序列，在α= 0.05 的水平上被拒绝。这表明该序列可以通过适当的时间序列模型进行预测。

2.参数 d 的确定

根据第 1 步的分析，时间序列经过一阶差分后趋于稳定。因此，ARIMA 模型的参数 d 可以确定为步骤 1。

3.参数 p, q 的确定

表 3:p 值和 q 值的确认方法

Model	ACF	PACF
$AR(p)$	attenuation tends to 0	truncation after p-order
$MA(q)$	truncation after q-order	attenuation tends to 0
$ARMA(p,q)$	attenuation tends to 0 after q-order	attenuation tends to 0 after p-order

我们尝试使用自相关函数(ACF)和部分自相关函数(PACF)来确定 p 和 q 的值，确认方法如表 3 所示。

进行一阶差分后，绘制 ACF 图和 PACF 图。

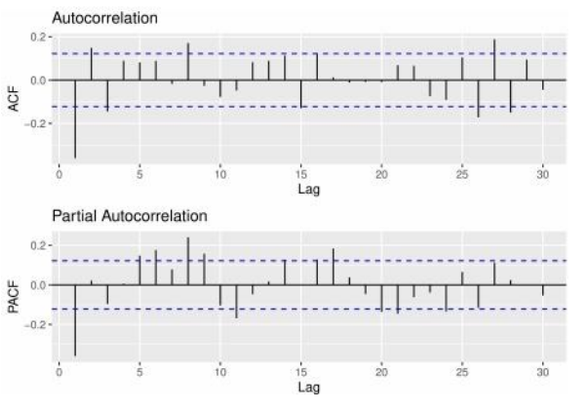

图 5:一阶微分序列相关函数图

对于这张图，有三种解释:

ACF PACF 是尾随的，可以考虑 ARIMA(1,1,1)模型。

ACF 尾随，PACF 一阶截断，考虑 ARIMA(1,1,0)。

ACF 三阶截断，PACF 尾随，考虑 ARIMA(0,1,3)。

对 ACF 和 PACF 图的解释是困难和主观的。仅凭这一点来判断 ARIMA 的顺序是不够的。拟合 ARIMA 的最佳方法是从低阶模型开始，然后尝试每次添加一个参数，看看结果是否发生变化。

所以，我们通过计算机编程实现了这一点。根据 Hyndman-Khandakar 算法的一种变体(Hyndman & Khandakar, 2008)，通过最小化评价标准 AICc 获得 ARIMA 模型。

在步骤 2 中已经确定了阶数 d。然后通过将数据差 d 次后最小化 AICc 来选择 p 和 q 的值。该算法不是考虑 p 和 q 的每一个可能组合，而是使用逐步搜索来遍历模型空间:

(a)拟合四个初始模型:ARIMA(0,d,0)、ARIMA(2,d,2)、ARIMA(1,d,0)、ARIMA(0,d,1)。除非 d = 2，否则包含常数。如果 d≤1，则还拟合一个附加模型:不带常数的 ARIMA(0,d,0)。

(b)将(a)中拟合的最佳模型(AICc 值最小)设为“当前模型”。(c)考虑当前模型的变化:

p 和/或 q 与当前模型相差±1;

从当前模型中包括/排除 c。

到目前为止考虑的最好的模型(要么是当前模型，要么是这些变体中的一个)成为新的当前模型。

(d)重复(c)直到找不到较低的 AICc 为止。

根据历史数据，我们执行上述步骤，求出使 AICc 最小的阶数 p 和 q，即模型的最优排序。选取 ARIMA 的参数为(1,1,1)。

4.模型拟合

确定最优顺序后，我们进行参数估计。对模型系数进行了显著性检验。模型拟合曲线如图 6 所示。

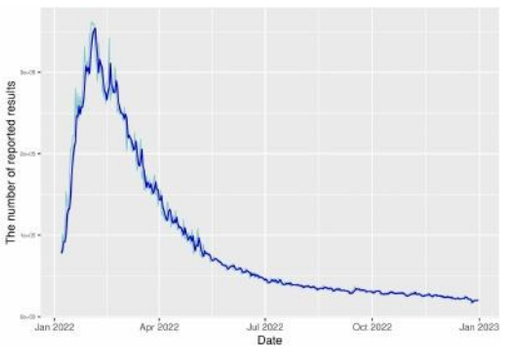

图 6:拟合曲线

4.1 残留分析

标准化残差的 p 值、残差的 ACF 和残差的 Ljung-Box 白噪声检验如图 7 所示。可以看出，残差的自相关系数在一阶差分后均落在置信区间内，趋势逐渐收敛于 0。对多个滞后值进行白噪声检验计算。水平虚线(0.05 线)上方的 p 值表明残差序列已经是白噪声，表明 ARIMA 模型已经从数据中充分提取了有用的信息。

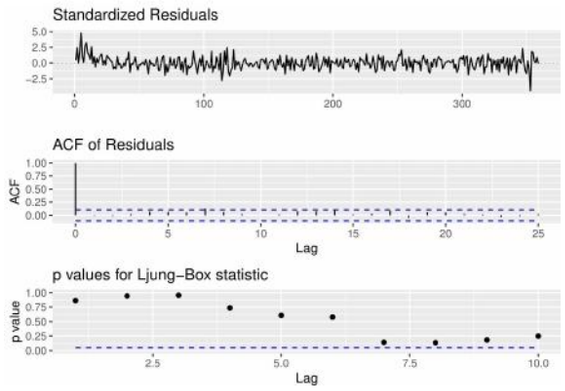

图 7:残差

同样，通过绘制 Q-Q 图，模型残差基本服从均值为零的正态分布。结合自相关检验的结果，确定拟合最优模型后的残差为白噪声序列，无需进一步建模。

6.预测

基于 ARIMA(1,1,1)模型，使用 2022 年 1 月 07 日至 2022 年 12 月 31 日的数据，可以预测此后 60 天内报告的结果数量，如图 8 所示。进一步，我们在 2023 年 3 月 1 日获得了一个具有 90%置信度的预测区间[10517,27007]，预期预测值为 16529。

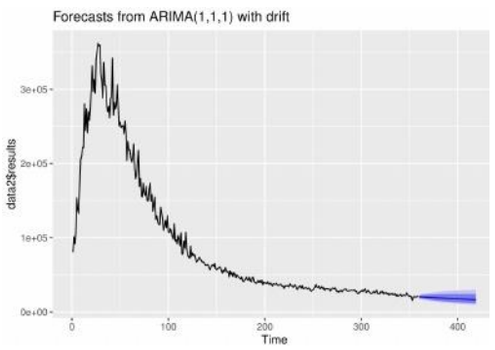

图 8:预测结果

4.2 词属性的影响

为了探究单词属性对在困难模式下打出的分数所占报告分数百分比的影响，我们首先总结了单词的属性。然后，以属性为自变量，百分比为因变量，通过建立多元线性回归模型，进一步确定两者之间的关系。

4.2.1 Word 的属性

通过回顾与词属性相关的文献，我们总结出以下属性。

1.字类

根据 Python 中最流行的英语自然语言处理库 NLTK，我们用“NLTK”标注了每个单词的类。pos_tag”功能。由于其类的结果比较详细，有 34 个[x]类别，结合“Problem_C_Data_Wordle.csv”数据集中词类的实际情况，我们最终将词类分为以下四类。

NN:名词。

RB:副词 very。

JJ:形容词。

Other:除上述三个词类以外的词类。2.频率

显然，一个词的常见程度会影响一个人使用它的次数。根据 Kaggle 上的“英语词频”数据集，其中包含了 333333 个单词的计数

英语网络上最常用的单个单词，来源于谷歌网络万亿词语料库。

我们获得每个解词的计数，并将每个词自身的计数除以总计数，得到其出现频率。

3.音节数

每个单词都有一定数量的音节，五个字母的单词通常有一个或两个音节。我们使用 Python 中 textstat 库的“text- stat.音节 le_count”函数来计算每个解词的音节数。

4.字母重复的次数

虽然 Wordle 的解词是一个 5 个字母的单词，但很可能其中有两个甚至更多相同字母的重复，所以我们计算每个单词中字母的重复次数。比如出现两个 a，就记录为一次，出现三个 a 就记录为两次，出现两个 a 和两个 b 也记录为两次。

5.极性

通常一个词会包含积极或消极的情绪。我们用“blob”。Python 中 TextBlob 库的 sentiment”函数来获取单词的极性，其值介于(−1,1)之间，该值越高意味着单词所代表的情感越积极。

6.主体性

一个词通常也包含一种能力或客观的情感。和 5 一样，我们可以得到这个词的主观性，它的值在(0,1)之间，值越大，这个词所代表的情感就越主观。

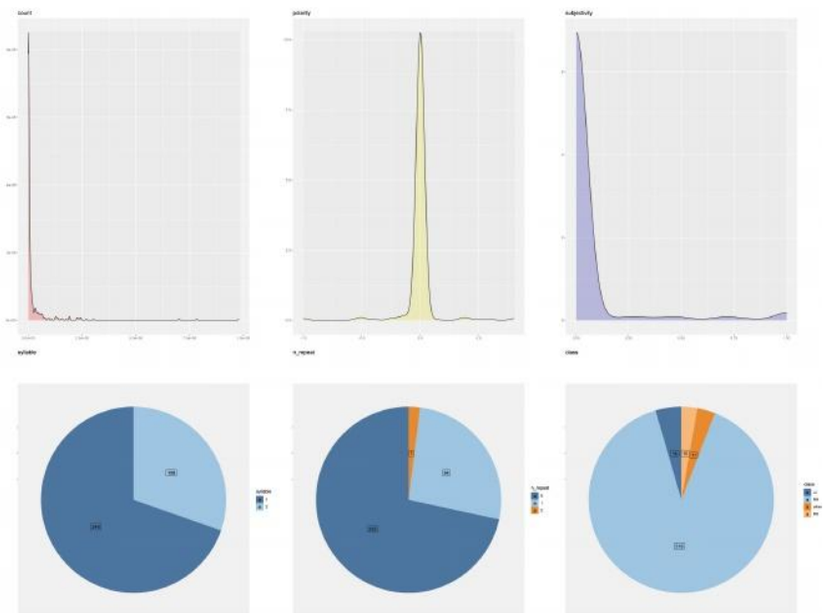

图 9:词的属性

4.2.2 多元线性回归

为了探索单词的每个属性与在 Hard Mode 下播放的报告分数百分比之间的关系，我们选择建立一个多元线性回归模型。

1.模型设计

为了考察变量之间的关系，我们计算了比例与其他自变量的 Pearson 相关系数，并进行显著性检验，结果如图 10 所示。可以看出，自变量之间存在一定的相关性，与比例相关性较强的因素是“音节”、“n_repeat”、“极性”、“主观性”和“频率”，可以通过建立回归模型来进一步研究。

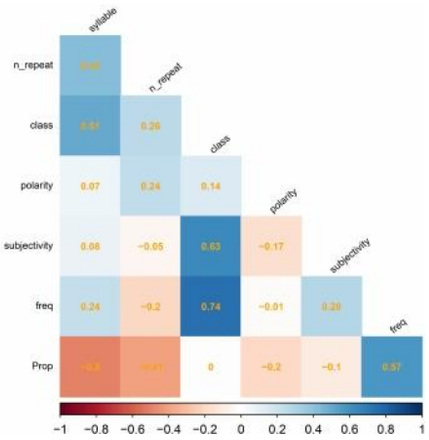

图 10:Pearson 相关系数

2.模型建立

完整的模型是通过前面的研究开发出来的，具体如下:

$$\boldsymbol{Y} = 0.0692 - 0.0045\boldsymbol{X}_1 - 0.0026\boldsymbol{X}_2 - 0.0064\boldsymbol{X}_3 - 0.0033\boldsymbol{X}_4 + 0.2823\boldsymbol{X}_5, \qquad (2)$$

式中 1 ，x1 表示音节数，x2 表示字母重复数，x3 表示极性，x4 表示主观性，x5 表示频率。

唯一发现的显著变量是“n_repeat”和“frequency”。然后，用 LASSO 回归法对这两个变量进行过滤，建立最终的回归模型如下:

$$\boldsymbol{Y} = 0.0750 - 0.0035\boldsymbol{X}_1 + 0.2663\boldsymbol{X}_2, \qquad (3)$$

其中，x 1 表示字母重复的次数，x2 表示频率。R 2 adj是0.782。回归方程和回归系数均通过显著性检验。

3.适当的测试

对于线性回归模型，假设是:误差独立同分布并服从正态分布，之间不存在线性关系

解释变量，且响应变量与解释变量呈线性相关。这些假设将在下一节进行检验。对 R 中的“lm()”函数使用“plot()”函数，得到的回归诊断结果如图 11 所示。

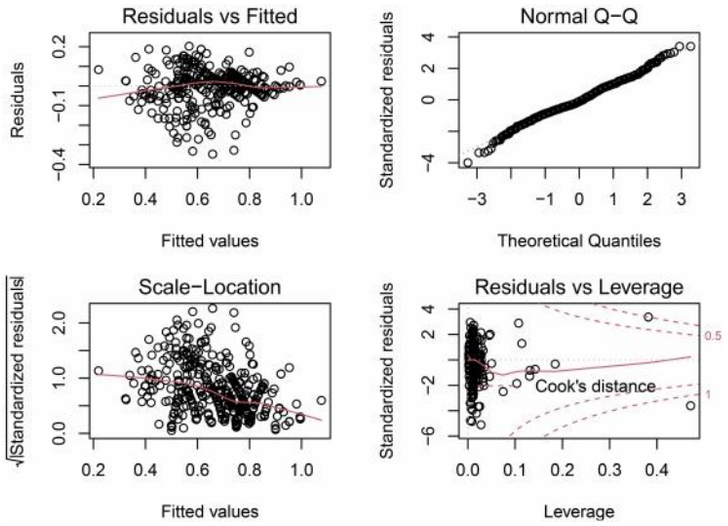

图 11:充分性测试

残差正态性检验

回归诊断图 11 右上方的面板显示，所有点都在直线周围，表明没有理由怀疑正态性假设。

残差的独立同方差检验

图 11 左下角的面板显示，残差没有明显的趋势，因此认为误差是独立同分布的。

模型线性检验

在图 11 的左上角面板中没有明显的曲线关系，因此认为线性是有效的。

模型多重共线性检验

Table 4: VIF

	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$
VIF	1.166984	1.341964

检验多重共线性的一种方法是计算方差膨胀因子(V IF)，如果单个变量的方差膨胀因子大于 10，则认为存在多重共线性。模型(x)两个变量的 V IF 计算结果如表 4 所示，表明模型中不存在显著的多重共线性。

综上所述，认为模型的上述假设是有效的。

4.结果

经过一系列的模型测试，最终确定建立的多元线性回归方程为:

$$\boldsymbol{Y} = 0.0750 - 0.0035\boldsymbol{X}_1 + 0.2663\boldsymbol{X}_2, \tag{4}$$

其中，x1 表示字母重复的次数，x2 表示频率。

这个等式表明，影响在困难模式下的得分百分比的重要变量是字母重复的次数和频率。频率与比例正相关，而 n_repeat 与比例负相关。

我们知道困难模式是在游戏开始前选择的，并且应该与单词的属性没有关系。然而，对于字母重复的次数，它的增加将使更少的玩家选择困难模式。这可能是因为当一个单词有多个重复的字母时，人们一般不会认为隐藏在绿色方块中的解词中还有字母，所以不容易快速猜出正确答案，导致

无法解出的概率要高得多。这个时候，人们也会从 Twitter 上知道，今天的解词是困难的，所以他们不会选择困难模式。

同样，对于频率，它与比值正相关，可能是因为当频率高时，人们更有可能快速得到正确的答案并将结果分享到 Twitter 上，这样其他玩家在发现今天的问题更容易时就会选择困难模式来挑战自己。当频率较低时，玩家不会选择困难模式，以获得更高的猜对答案的机会。

5 任务 2:基于马尔可夫链模型的分布

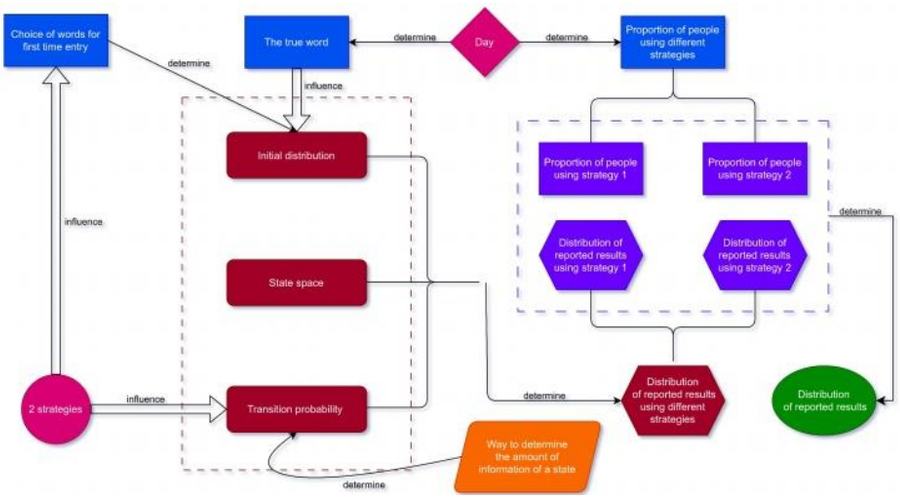

图 12:马尔可夫链流程图

根据假设(1)，玩家是理性的(他们不会放弃他们已经拥有的信息)和贪婪的(他们只是试图根据手头的情况给出局部最优解)，所以游戏过程可以建模为马尔可夫链。我们将玩家在世界游戏中的状态概括为已获得的每种颜色的方块数，并将玩家在 6 步中可能达到的所有状态作为这个马尔可夫链的状态空间。我们还为玩家设想了两种策略。

概率第一原理。根据假设(2)，常用词更有可能是答案，所以可以测量一个词的使用频率，得到它是正确答案的主观概率，每次填写概率高的词。

信息第一原则。玩世界游戏本质上是一个获取信息，缩小可能单词范围，最终得到唯一正确答案的过程，所以每次都把有很大概率提供更多信息的单词填进去，也是一个很好的策略。

基于这两种策略，我们提出了两种相应的选择第一个填空词的原则，从而得到了两种策略下当天答案影响的初始分布。此外，我们给出了一种量化参与人在当前状态下获得的信息的方法，假设两种状态之间的传递概率是对应信息增量的简单函数，并受参与人所采用的策略的影响，从而得到了该马尔可夫链的传递概率。此时，我们就可以计算出两种策略下报告结果的分布。

为了更好地模拟世界博弈，我们假设使用这两种策略的人的比例每天都在变化，并给出一种从历史数据中估计当天策略比例的方法。最后，我们利用对历史数据的比例估计值，结合高斯过程回归模型，给出对未来策略比例的估计值，并利用预测的未来策略比例结合不同策略下报告结果的分布，实现对报告结果分布的预测。

5.1 状态空间

玩家在世界游戏中的状态可以用已经获得的绿、黄、灰三色块的集合来表示，它代表了玩家拥有的关于当前状态的所有信息，玩家的决策完全基于这个集合(假设(1))。该集合中的绿色和黄色块同时携带位置和字母信息，因此各有 5 个(可能的位置)×26(可能的字母)= 130 个，而灰色块仅代表字母信息，总共有 26 个。为了简化模型，将集合简化为三组(ngreen,nyellow,ngrey)，其中 ncolor 表示玩家当前获得的“颜色”块的数量。

根据假设(3)，玩家将在下一次猜测中保留已经获得的绿色块，即已知绿色块的数量不会减少，获得的新颜色块的总数不会超过(5 -已经获得的绿色块的数量)。显然，已知的黄色方块只能转化为绿色方块，所以绿色方块和黄色方块的总和不会减少。此外，灰色块的数量也不会减少。设 $A_i = \{j : j \text{ is accessible from } i \text{ in one step}\}$。综上所述，对于状态 $i = (n_{green}, n_{yellow}, n_{grey})$ 和状态 $j = (n'_{green}, n'_{yellow}, n'_{grey})$，

$$j \in A_i \Leftrightarrow$$

$$\begin{cases} n_{green} \neq 5 \\ n'_{green} \geq n_{green} \\ n'_{grey} \geq n_{grey} \\ n'_{green} + n'_{yellow} \geq n_{green} + n_{yellow} \\ n'_{green} + n'_{yellow} + n'_{grey} \leq n_{yellow} + n_{grey} + 5 \end{cases} \text{ or } \begin{cases} n_{green} = 5 \\ n'_{green} = n_{green} \\ n'_{yellow} = n_{yellow} \\ n'_{grey} = n_{grey} \end{cases} \quad (5)$$

设 bi 为玩家在第 i 次猜测后可能达到的状态集合，那么

$$B_0 = \{(0,0,0)\}, B_{i+1} = \{j : \exists i \in B_i \; s.t. \; j \in A_i\}, i = 0, 1, 2, \cdots, \; S = \left(\bigcup_{i=0}^{6} B_i\right) \cup \{X\} \quad (6)$$

其中状态 X 表示游戏失败。S 是马尔可夫链的状态空间。总共有 384 种状态，下面以一条链为例。

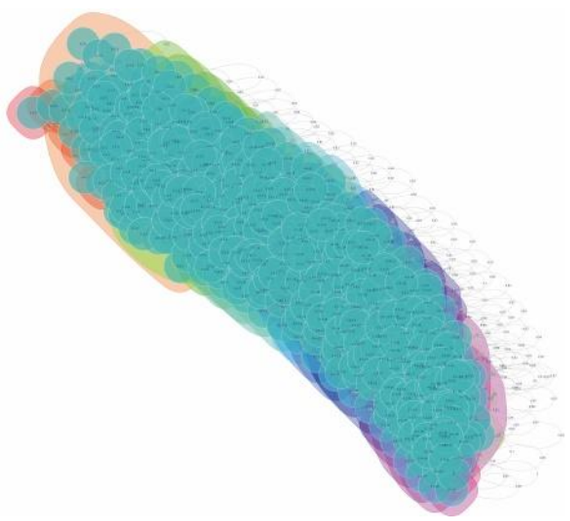

图 13:马尔可夫链

5.2 初始分布

根据假设(4)，世界玩家拥有大致相同的词库，并且对一个词是否是正确答案持有相同的主观概率。设 W 为玩家可能填写的所有单词，则某个单词 x∈W 是正确答案的主观概率为 px。

一个来自互联网的数据集可以作为 W 的近似值，其中包含了大约 36000 个最常见的五字母单词及其词频。sigmoid 函数可以将任何实数映射到[0,1]，并且具有在零附近变化更快的特点，所以对于一个单词 x，我们可以把

$$p_x = \frac{1}{1 + e^{-(freq_x - k)}} \tag{7}$$

作为 x 为正确答案的主观概率，其中 freqx 为单词 x 的词频，k 为正整数。取 k 为其词频的中位数，我们得到{px : x∈W}。

对于使用第二种策略(信息优先)的玩家，我们可以假设他们使用

$$I_x = -\log_2 \frac{\mathrm{card}(\{y \in W : \mathrm{match}(x, y) = \mathrm{match}(x, x_{true})\})}{\mathrm{card}(W)} \tag{8}$$

为了衡量在开局时填入单词 x 所获得的信息量，其中 card(·)表示集合中元素的个数，match(x, y)表示在正确答案为单词 y 时填入单词 x 所获得的匹配，x(r)true 表示第 r 天的正确单词。很明显,

$$\forall x \in W, \mathrm{E}[I_x] = \mathrm{E}[\mathrm{E}[I_x | x_{true}^{(r)}]] = \sum_{x_0 \in W} \mathrm{E}[I_x | x_{true}^{(r)} = x_0] P(x_{true}^{(r)} = x_0). \tag{9}$$

将 P(xtrue (r)=x0)替换为主观概率 px0，我们得到

$$\mathrm{E}[I_x] \approx \sum_{x_0 \in W} -\log_2 \frac{\mathrm{card}(\{y \in W : \mathrm{match}(x, y) = \mathrm{match}(x, x_0)\})}{\mathrm{card}(W)} p_{x_0} \triangleq \tilde{\mathrm{E}}[I_x]. \tag{10}$$

我们可以假设使用第二种策略的玩家选择单词 x 的概率为 $\frac{\tilde{\mathrm{E}}[I_x]}{\sum_{y \in W} \tilde{\mathrm{E}}[I_y]}$

作为第一个填入的单词。

可以创建一个映射

$$f : W \to S,\ f(x) = \begin{pmatrix} \text{The number of green blocks in match}(x, x_{true}^{(r)}) \\ \text{The number of yellow blocks in match}(x, x_{true}^{(r)}) \\ \text{The number of grey blocks in match}(x, x_{true}^{(r)}) \end{pmatrix}, \forall x \in W. \tag{11}$$

设第 r 天世界博弈的马尔可夫链为{Xi(r)∈S: i = 0,1,2，…}，则对于使用第一种策略的玩家成立

$$P(X_0^{(r)} = i) = \sum_{x \in W : f(x) = i} p_x, \forall i \in S, \tag{12}$$

对于使用第二种策略的玩家来说

$$P(X_0^{(r)} = i) = \frac{\sum_{x \in W : f(x) = i} \tilde{\mathrm{E}}[I_x]}{\sum_{y \in W} \tilde{\mathrm{E}}[I_y]}, \forall i \in S. \tag{13}$$

这给出了当天在马尔可夫链上使用不同策略的玩家的初始分布。

5.3 转移概率

$$I_i = -\log_2 \frac{\mathrm{card}(\{x \in W : \bigcup_{y \in G_i} \mathrm{match}(x, y) = i\})}{\mathrm{card}(W)}, \tag{14}$$

其中 Gi 是玩家在状态 i 时猜出的单词集合。然后测量 I i 状态 i 的信息量(在知道匹配结果为 i后，备选单词数量减少的程度)。

表示马尔可夫链{Xi(r)∈S 中状态 i 到 j 的转移概率:i = 0,1,2，…}表示当使用第 k 个策略(k = 1,2)时，p(r)k (i, j)。通过假设(5)，并注意到与第一个策略(K=1,2)相比，第二个策略(信息优先)更侧重于获取更多的信息，我们可以简单地让

$$p_k^{(r)}(i,j)=\begin{cases}\dfrac{e^{k(I_j-I_i)+C}}{\sum_{j:j\in A_i}e^{k(I_j-I_i)+C}}, & j\in A_i\\ 0, & j\notin A_i\end{cases}, C\in\mathbb{R}^+. \tag{15}$$

根据假设(6)，对于状态i=（ngreen，nyellow，ngrey）∈S，Ii可以用

$$\frac{\mathrm{card}(\{(x_i,x_j)\in W\times W:\mathrm{match}(x_i,x_j)=i\})}{\mathrm{card}^2(W)}, \tag{16}$$

给出了马氏链{Xi(r)∈S: i = 0,1,2，…}在第 k 个策略下。

5.4 报告结果的分发

对于马尔可夫链{Xi(r)∈S: i = 0,1,2，…}，表示其在策略 k 下的传递概率矩阵为 Pk(r)

$$T_j^{(r)}=\inf\{n\geq 1:X_n^{(r)}=j\},\ \rho_{ijn}^{(r)}=P(T_j^{(r)}=n|X_0^{(r)}=i),\ \mathbf{G}_n^{(r)}=[\rho_{ijn}^{(r)}]_{i,j\in S} \tag{17}$$

$$c=\mathrm{card}(S),\ \mathbf{E}=\mathbf{1}_c\mathbf{1}_c^T-\mathbf{I}_c, \tag{18}$$

现在，下面的等式成立

$$\mathbf{G}_n^{(r)}=\underbrace{\mathbf{P}_k^{(r)}(\ldots(\mathbf{P}_k^{(r)}(\mathbf{P}_k^{(r)}(\mathbf{P}_k^{(r)}}_{n}\cdot\mathbf{E})\cdot\mathbf{E})\cdot\ldots)\cdot\mathbf{E}), \tag{19}$$

其中·表示矩阵的逐元素乘法。显然这个马尔可夫链很吸引人。

$$C_k^{(r)}=\{i\in S:p_k^{(r)}(i,j)=0,\forall j\neq i\},\ T_{absorbed}^{(r)}=\inf\{n\geq 1:X_n^{(r)}\in C_k^{(r)}\}, \tag{20}$$

$$P(T_{absorbed}^{(r)}=n|X_0^{(r)}=i)=\sum_{j\in C}\rho_{ijn}^{(r)},\ P(T_{absorbed}^{(r)}=n)=\sum_{i\in S}\sum_{j\in C}\rho_{ijn}^{(r)}P(X_0^{(r)}=i) \tag{21}$$

现在我们有了 T(r)absorbed的分布，它给出了策略 k 下报告结果的分布

$$P_k^{(r)}(\text{reported results}=\mathrm{X})=P(T_{absorbed}^{(r)}\geq 7)=1-\sum_{i=1}^{6}P(T_{absorbed}^{(r)}=i), \tag{22}$$

$$P_k^{(r)}(\text{reported results}=i\text{ tries})=P(T_{absorbed}^{(r)}=i),i=1,2,\ldots,6. \tag{23}$$

5.5 使用两种策略的比例

根据假设(7)，所有玩家在第 r 天使用策略 k 的比例为 qk(r)， k = 1,2，其中 q1(r)+ q2(r) = 1。当天所有玩家报告结果的理论分布为

$$\mathfrak{I}^{(r)}(\text{reported results}=\mathrm{X})=\sum_{k=1}^{2}q_k(r)P_k^{(r)}(\text{reported results}=\mathrm{X}), \tag{24}$$

$$P^{(r)}(\text{ reported results}=i\text{ tries})=\sum_{k=1}^{2}q_k(r)P_k^{(r)}(\text{reported results}=i\text{ tries}),i=1,2,\ldots,6. \tag{25}$$

假设当天报告结果的真实分布为

$$P_{true}^{(r)}(\text{reported results} = \text{X}) = l_X^{(r)}, \tag{26}$$

$$P_{true}^{(r)}(\text{reported results} = i \text{ tries}) = l_i^{(r)}, i = 1, 2, \ldots, 6. \tag{27}$$

然后，与 MLE 方法类似，可以将理论分布与真实分布之间的交叉熵的 q1,q2 最小化作为 qk(r)的估计，即。

$$\hat{q}_1(r) = \arg\min_{q\in[0,1]}\{-\ln\left[qP_1^{(r)}(\text{reported results} = \text{X}) + (1-q)P_2^{(r)}(\text{reported results} = \text{X})\right] l_X^{(r)}$$

$$-\sum_{i=1}^{6}\ln\left[qP_1^{(r)}(\text{reported results} = i) + (1-q)P_2^{(r)}(\text{reported results} = i)\right]\}, \tag{28}$$

$$\hat{q}_2(r) = 1 - \hat{q}_1(r). \tag{29}$$

5.6 预测未来报告结果的分布

高斯过程(GP)是一种通用的监督学习方法，旨在解决回归和概率分类问题，而高斯过程先验的连续值推理被称为高斯过程回归(GPR)。

我们根据上述模型估计 2022 年 1 月 7 日至 2022 年 12 月 31 日每天的 q1°(r)，得到时间序列，并与 GPR 模型拟合，得到 2023 年 3 月 1 日 q1(r)的预测值和置信区间。此外，我们还可以将上述模型与 2023 年 3 月 1 日的正确单词“eerie”一起使用，得到两种策略下报告结果的分布。

使用上述方法，我们得到的 q1(r)预测值为 0.63,95%置信区间为[0.59,0.67]，两种策略下报告结果的分布分别为(0,0,12.6,22.0,39.6,21.6,4.2)和(0,0.4,8.4,39.4,28.4,20.4,3.0)。

我们对报告结果分布的最终预测值为(0.000,0.148,11.046,28.438,35.456,21.156,3.756)。

6 任务 3:解词分类

单词的难易程度是由单词本身决定的，并反映在单词的多个属性上。我们推测，根据属性对单词进行聚类，会给我们提供每个类别中嵌入的单词的难度信息。

6.1 难度评分

从之前的研究中可以清楚地看出，(1,2,3,4,5,6,X)的百分比只与单词本身有关，与其他因素无关。由于 X 代表 7 次或更多猜测的结果，除非玩家故意多次猜测一个 5 个字母的单词而没有得到结果，否则猜测次数的平均值估计为 8。我们将每个可能的猜测次数的百分比结合起来，得到计算难度分数的权重公式:

$$\text{score} = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 + 6x_6 + 8x_7, \tag{30}$$

其中，x 1 表示 1 次尝试的百分比，x 2 表示 2 次尝试的百分比，x 3 表示 3 次尝试的百分比，x4 表示 4 次尝试的百分比，x 5 表示 5 次尝试的百分比，x 6 表示 6 次尝试的百分比，x 7 表示 7 次或更多尝试的百分比。

6.2 k - prototype 聚类

6.2.1 求解步骤

从数据集 X 中为每个聚类选择一个原型(中心样本)。

将 X 中的样本分配给最接近原型的类别，并在每次分配后更新类别的原型。

将所有样本分配到各自的类后，重新计算样本到当前原型的距离。如果某个样本比原始原型更接近新原型，则将其重新分配到新原型的类别中。

重复上一步的过程，直到没有样本改变类别。

6.2.2 结果

图 14:聚类结果

根据 k - prototype 聚类得到三类词，它们的类特征由它们的聚类中心总结，如图 14 所示。

6.3 解词难度分类

从难度评分密度图 x 可以看出，三种分布差异明显;计算了三类之间的 Kullback-Leibler 散度，分别为 109.59、271.9 和 864.3，证明了我们的模型对难度的分类是准确的。从上面的分析来看，根据词属性得到的分类能更好的区分词的难度。

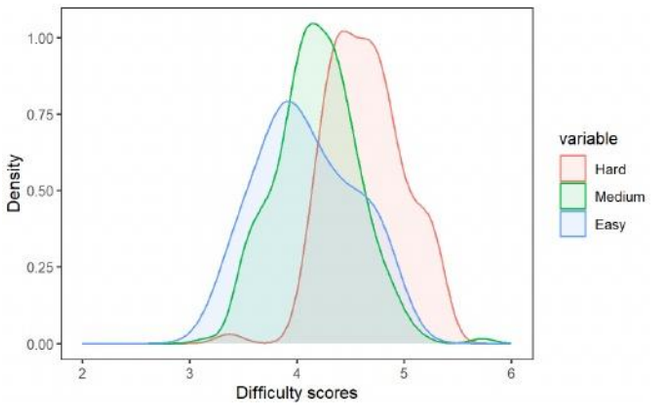

图 15:根据难度进行单词分类

容易的词往往是单音节的、形容词的、情感态度主观的。而难词往往是双音节的，存在重复字母，客观、否定。

6.4 EERIE 这个词的难度

在单词 EERIE 中，字母 E 重复出现了三次，这个单词本身有否定的意思，在生活中不常用。根据这些属性，ERRIE 被归类为三类中的第一类，难度等级比较难。

7 任务 4:其他有趣的特征

在对数据集进行仔细分析后，我们观察到几个值得注意的特征。•报告结果数量的波动性

如前文图 4 所示，每日报告结果数与前一天的差值具有非常不确定的波动率，因此我们选取波动率绝对值大于 40000 的日期，呈现如下表 5。

表 5:波动率

	Date
Sudden increase	2022-2-18, 2022-3-17
Sudden decrease	2022-1-19, 2022-2-17, 2022-2-8, 2022-1-11, 2022-1-31

然后我们选择了 2022-2-18 和 2022-1-11 来分析波动的原因。

对于 2022-2-18 上报的数量大幅下降，可能是由于躲闪这个词难度太大，大家猜不出来，因此没有选择分享自己的分数。

对于 2022-1-11 来说，报告分数的数量突然增加，可能是因为喝这个词很容易，人们很快就猜到了，所以他们选择分享自己的分数。

最简单的单词

如果只考虑第一次正确猜测，那么有 15 个单词的 1 次正确猜测的百分比大于 0.01。结果发现，这些单词都没有重复的字母，出现频率最高的字母是“a”，出现了 11 次，其次是“e”和“t”，都出现了 10 次。由于这是第一次猜测，而且问题是随机的，所以人们会最优地选择猜测一个有 5 个不同字母的单词，从而达到最大的消除效率。此外，由于“a”、“e”和“t”是构成这 5 个单词的单词中出现频率更高的字母，所以人们的猜测也大多包含了它们。

图 16:初步猜测

图 17:前两次猜测

考虑到第一次猜测的随机性，如果我们考虑前两次正确猜测大于 0.15 的 20 个单词，其中 2 个包含相同的字母，“treat”和“逗”。这些单词大部分也有 1 个猜对大于 0.01，出现次数最多的字母仍然是“a”，达到 14 次，出现次数最多的字母是“a”，出现次数为 14 次，其次是“t”，出现次数为 11 次。出现这种情况的可能原因与第一次猜对的原因大致相同。

最难的单词

7 次及以上尝试比例最大的单词是“parer”(0.48)，远远领先于第二大单词“foyer”(0.26)。这两个词都可以用在建筑学中，指的是较大建筑物内的特定区域或空间。它们也都有相似的词源。Parer 来自拉丁语“parare”，意思是“准备”，而 foyer 来自法语“fouyer”，意思是“壁炉”。

失实陈述或答案泄露

可能存在不实陈述或答案泄露的情况。例如，对于像 poise 这样的单词，在 task2 中的两种策略中，第一次答对的概率远低于 1%。很难想象 2%或更多的人只使用一次猜测就猜对了。

8 敏感性分析

为了分析马尔可夫链模型对参数的敏感性，我们选择不同的 C 值来计算 hatq(r)k，并使用 GPR 模型对其进行预测。选取的 C 值分别为(1、2、3)。

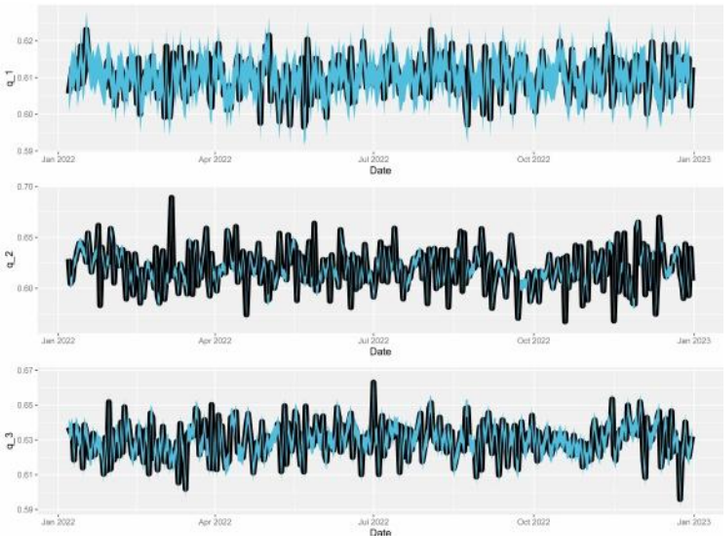

图 18:敏感性分析

从图 18 中，我们可以得出结论，C 对我们模型的预测精度影响很小，这表明我们的模型是鲁棒的，非常可靠。

9 模型评估和进一步讨论

9.1 优势

新奇。我们利用信息熵等概念对马尔可夫链的初始分布和转移概率进行建模，提出了两种博弈策略，并成功估计了两种策略在历史数据中的使用比例，并对其使用时间序列预测方法，最终将两者结合得到结果。

可解释性。我们使用 ARIMA 和 GPR 模型进行时间序列预测，k - prototype 方法进行聚类分析，马尔可夫链建模世界游戏，所有这些都具有很强的可解释性或深入的统计基础。该模型的创新部分也具有令人信服的结构。

9.2 缺点

缺少其他潜在的有用功能。我们构建的模型只使用了单词的部分特征，可能还有一些未开发的特征可以改进模型。

模型设计中存在一定程度的随意性。我们会采取一种随意性的心态，当

设计我们的部分模型。

10 给拼图编辑的一封信

亲爱的先生或女士:

根据您提供的文件，我们对数据进行了一些有趣的分析，以回答您提出的问题。

首先，我们建立了一个时间序列模型 ARIMA，它是严格选择和迭代比较参数的结果。我们有信心，这个模型在未来的预测中会有很好的表现。根据该模型，2023 年 3 月 1 日报告的结果数量将在 10517 到 27007 之间，其中 16529 的可能性特别大。由此可见，world 的生命力之强。在这个各种游戏、短视频狂轰滥炸的时代，一款游戏上线一年后，仍然拥有如此高的关注广度，实在令人惊讶。而且，通过一些可视化的方式，我们发现，在经历了 2022 年 Q1 的疯狂增长之后，世界游戏的数量在 2022 年 Q3 逐渐趋于稳定。这意味着有相当一部分玩家已经习惯了世界游戏，并将其作为日常生活的一部分，而不是出于新鲜感而尝试一次。

接下来，我们发现难度模式的百分比与解词的属性有关。我们推测玩家会根据社区分享的信息来决定是否开启困难模式。当玩家发现当今的游戏世界缺乏挑战时，他们会选择困难模式。毕竟，困难模式的规则使得接近正确答案的方式更加单一。

我们还开发了一个马尔可夫链来模拟玩家玩世界游戏的过程，并提出了两种游戏策略。最后，通过结合这些结果来实现对未来某一日期报告结果分布的预测。基于该模型，2023 年 3 月 1 日，EERIE 这个词的尝试次数的百分比分布呈现出集中在三次及以上的趋势，表明这个词的挑战性。

最终，我们根据单词的属性对其进行分类。正如我们推测的那样，得到的单词类别暗示了单词的难度信息。多音节的、重复字母的、客观的、否定的和不常见的词比单音节的、情感丰富的和常见的词更难猜。根据我们开发的马尔可夫链模型的结果，我们预测 EERIE 这个词对玩家来说将是一个巨大的挑战。

在解决问题的过程中，我们发现了这个数据集的一些有趣的特征。比如，报告词数的增加或减少，可能是由于当天这个词的难易程度;容易猜对的单词往往包含 a、t 等常见字母;可能出现结果误读或答案泄露。

很多玩家表示，world 已经成为他们醒来后大脑的第一场热身。这都得益于你们不断努力营造良好的游戏氛围和和谐的游戏社区。在这里，每个人都可以自由快乐地分享自己的成就，而不会被戏剧化。衷心感谢大家的付出，希望我们的分析对大家有所帮助。

真诚地,

团队# 2318982

References

[1] J. D. Cryer and K. S. Chan. Time Series Analysis. Time Series Analysis, 2005.

[2] Renato Feres. Notes for math 450 matlab listings for markov chains. 2007.

[3] Robin John Hyndman and George Athanasopoulos. Forecasting: Principles and Practice. OTexts, Australia, 2nd edition, 2018.

[4] Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian processes for machine learning (adaptive computation and machine learning). 2005.

[5] Xiangxiang Yan. Using arima model to predict green area of park. Computer Science, 47(S02):5, 2020.