Problem Chosen	2023	Team Control Number
C	MCM/ICM	2311717
	Summary Sheet	

世界之谜:挖掘数字分数的秘密&解词

摘要：《世界之谜》是目前《纽约时报》每天提供的一个很受欢迎的谜题。简单的规则和聪明的传播特性为它的流行做出了贡献。在本文中，我们分别构建了两个预测模型来预测 Twitter 报告数间隔和结果分布，并开发了一个模型来对解词的难度进行分类。

在 TASK1 中，经过数据预处理，我们从统计学的角度建立了基于三阶高斯回归和非齐次泊松过程的世界之谜数量预测模型。其中，高斯回归用于预测报告数的趋势符号，非齐次泊松过程在此基础上预测报告数的随机波动。此外，我们使用流行度松弛函数对随机过程进行校正，从而更好地逼近流行度变化。在 75%的置信水平下，我们预测 2023 年 3 月 1 日报告数量的间隔为[7654,2015]。此外，我们根据字母数量、字母位置等提取单词的 8 个属性，发现这些属性对玩家困难模式选择的百分比没有影响。玩家对自己表现能力的信心和游戏心态可能是他们是否选择困难模式的主要原因。

在 TASK2 中:我们首先提取影响报告结果分布的数据特征，包括单词属性，以及难度模式的百分比。然后，我们构建一个 BP 神经网络，对未来某个解词的猜测结果分布进行初步预测。为了提高预测结果的泛化性能，我们构建了一个基于 Bagging 的集成 BP 神经网络。然后，我们预测 2023 年 3 月 1 日 EERIE 报告结果的分布为(0,1,6,25,31,25,13)(in %)。我们有超过 80%的置信度，对于每个可能结果的百分比，预测结果的绝对误差不超过 5%。

在 TASK3 中:首先，我们根据用户报告数据的分布，建立基于 K-Means 的单词难度归纳模型，并将难度分为 4 类。然后，我们基于 Pearson 系数探索单词属性与难度之间的关联，并将相关系数大于 0.6 的属性作为难度分类属性，构建单词难度分类模型。而且，我们发现解词的首字母和第二个字母出现的频率、发音中包含的元音数量以及单词属性的数量与难度分类有很高的相关性。最后，EERIR 的难度分类结果是最难的。

在 TASK4 中:在探索报告数量的统计属性的同时，我们发现报告数量的分布呈现出与其随时间变化的趋势相似的模式。此外，我们还注意到，在 359 天的报告结果分布数据中，3 次尝试完成游戏的百分比波动是最大的。

最后，我们对模型进行了敏感性分析，并研究了模型可变参数的变化对结果的影响。

关键词:高斯回归;泊松过程;BP 神经网络;K-Means

目录

1 介绍

1.1 问题背景

Homer 是棒球运动中的一个术语，是一个非正式的美式英语单词。令人惊讶的是，荷马(本垒打)在剑桥词典网站上被搜索了 7.9 万多次，5 月 5 日被搜索了 65401 次。由此，荷马成为了《剑桥词典》2022 年的年度词汇。你可能想知道为什么，但这要从世界大战说起，这是一款在海外非常流行的猜字游戏。2022 年，在线益智游戏《世界大战》风靡社交媒体。而世界大战那天的答案是荷马，对于不熟悉这个词的非美国用户来说，这很困难。

《世界大战》目前是《纽约时报》每日提供的热门谜题，并且越来越受欢迎，有 60 多个版本可供选择。玩家可以在“普通模式”和“困难模式”中进行选择。玩家尝试在 6 次或更少的尝试中猜出一个 5 个字母的单词来解决这个谜题，每次猜出都会收到反馈，并改变贴图的颜色(绿色、黄色、灰色)。注意:每次猜出的单词必须是真实的英文单词。未被大赛识别为单词的猜测是不允许的。

绿色方块表示该方块中的字母在单词中并且在正确的位置。黄色瓦片表示该瓦片中的字母在单词中，但位置错误。:灰色瓦片表示该瓦片中的字母不包含在单词中。

1.2 问题重述

综合本文件的背景资料和结果，我们需要解决以下问题:

开发一个模型来解释报告结果数量的变化，并为 2023 年 3 月 1 日的报告结果数量创建一个预测区间。分析单词属性对玩家模式选择的影响程度。

建立一个模型来预测报告结果的分布。分析模型和预测中存在的不确定性因素。

开发一个模型，按难度对解词进行分类。识别与每个分类相关的词的属性。

描述数据集的其他有趣特征。

1.3 文献综述

近年来，随着互联网的普及，社交网络逐渐成为讨论现实世界中正在发生的事情的主要媒介，用户可以在社交平台(如 Twitter)上生成和传播丰富的数据流，从而洞察正在发生的热点事件。人气建模和预测在市场营销、舆情监测、广告等场景中有着广泛的应用，基于时间序列的趋势分析是近年来在数据挖掘和社交网络分析领域备受关注的研究课题。这类研究的思路主要借鉴了金融和流行病学模型。Shen 等[1] 人首先建立了一个增强泊松模型。

过程(RPP)模型采用异质泊松过程模型预测动态患病率，并认为“富者愈富”。Zhao 等[2] 人基于自激点过程理论，假设过去的流行程度会影响过程的未来演变，开发了一个 SEISMIC 模型，并使用双重随机过程来描绘信息的传染。Wu 等[3] 人基于时间特征、用户特征和网络结构特征提出了基于贝叶斯网络的人气预测模型(EPAB)，并提出了早期模式的概念，建立了早期特征信息与未来热度变化之间的关系。

但是，时间序列模型要求数据集包含时序信息，不满足这一条件的数据集无法建模。同时，序列模型和基于节点行为动态的深度学习方法不适用于仅基于报告数据的本任务的预测情况。一方面，现有的数据集不包含具体的信息，比如报告者是谁、在任何给定时间有多少玩家等，因此

无法基于该数据集构建节点模型。另一方面，深度学习等技术的可解释性不佳，无法从数学上解释热度变化的趋势，需要更多的训练数据。

在本文中，我们尽量从数据文件中提取所有的信息。针对 Wordle 的具体应用场景，我们不仅实现了对未来报告数量的区间预测，还对报告结果的分布和单词难度的分类进行了进一步的分析。

1.4 我们的工作

我们提出了三种模型来挖掘报告结果数据的信息。我们论文的结构如图 1 所示。

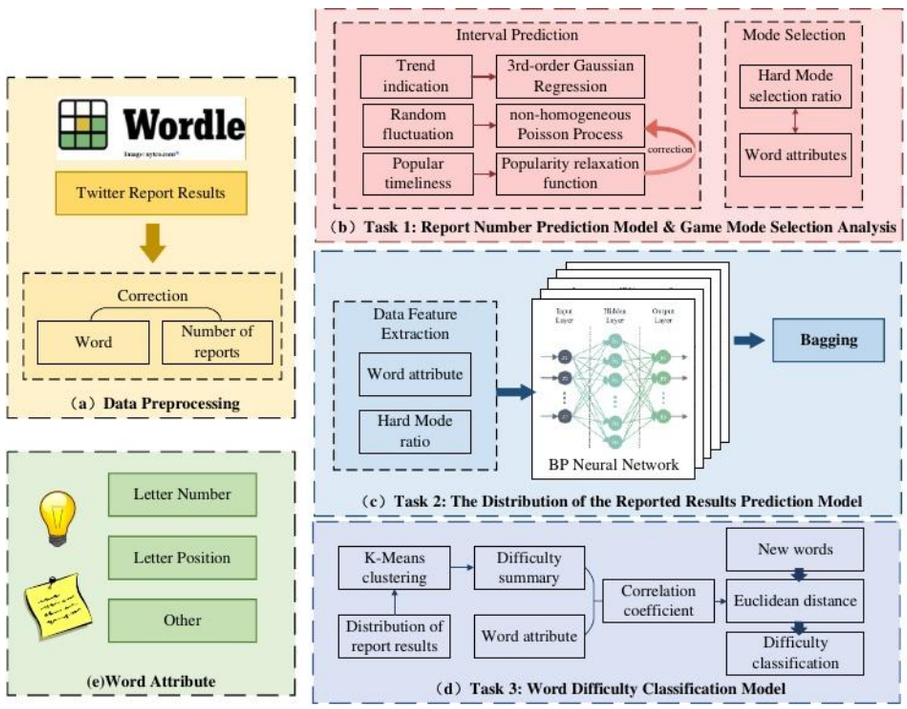


图 1:我们论文的结构

本文的其余部分组织如下。在第二节中，我们介绍了前提假设和论证，公式中的常见变量在第三节中提到。在第四节中，进行建模前的数据预处理。第五节建立了报告编号区间的预测模型，探讨了词属性与模式选择之间的关系。第六节建立了报告结果分布预测模型。在第七节中，我们提出了单词难度分类模型。第八节继续探讨数据文件的有趣特性。在第 IX 节和第 X 节中，我们分析了模型的敏感性，并进一步评估了模型的优缺点。最后，第 XI 节给出了结论。

2 假设和理由

我们做了一些一般性的假设来简化我们的模型。这些假设连同相应的理由列在下面:

1.假设报告中用户数量的变化是实际情况下玩家变化的真实反映。

可能有些玩家对游戏很感兴趣，但却不会在 twitter 上发布结果，所以报告的用户数量往往低于实际价值。不过，我们假设玩家愿意分享他们的游戏结果。

2.假设游戏每人每天只能玩一次，问题每天在美国东部时间 0:00 更新。

这个假设作为游戏的既定规则。这一规则反映了报告数据的可分析性。同时，也体现了游戏设计师 Wardle“不希望玩家每天花在游戏上的时间超过 3 分钟”的初衷。

3.假设在游戏的设定中，玩家被视为具有一定文化水平和解决问题能力的人。

每个游戏中给出的单词之间没有特别的联系，但玩家对词汇的掌握程度直接决定了答案的步骤、速度和正确性。我们假设玩家有解决问题的能力，在猜不出答案的情况下，可以选择在网上找到答案。

4.假设历史数据是所有可能的世界规则问题和玩家答案的良好代表。

由于我们只有 2022 年 359 天的报告结果数据，并作为唯一的参考数据集。数据可能不具有代表性，为了便于分析，我们假设它可以在一定程度上显示出问答模式。

3 记号

本文使用的关键数学符号列于表 1。

表 1:本文使用的记号

Symbol	Description
t_i	time, where i represents the number of days from that date to January 7, 2022
$y(t_i)$	number of results reported on the day t_i
$\lambda(t)$	the mean value of the number of reports on the day t
f_α	frequency of a given letter α in 359 words of result data
p_{mn}	the ratio of words with the n th letter m to all words
k	number of clustering algorithm centers of mass

4 数据预处理

在建立模型之前，需要对报告中的数据进行初步检查。根据世界规则，每个单词有 5 个字母长。但数据中却有不寻常的 4、6 个字母的统计。单词中的错误会干扰后面对单词属性的分析，所以根据过去的答案数据对单词数据 1 进行了修正。根据报告数之间的关系，我们发现 529 号的结果数与前后日期的数值存在较大偏差。因此，我们将其视为异常数据，并通过取前后两天各数据的平均值进行修正。整体的预处理过程如图 2 所示。

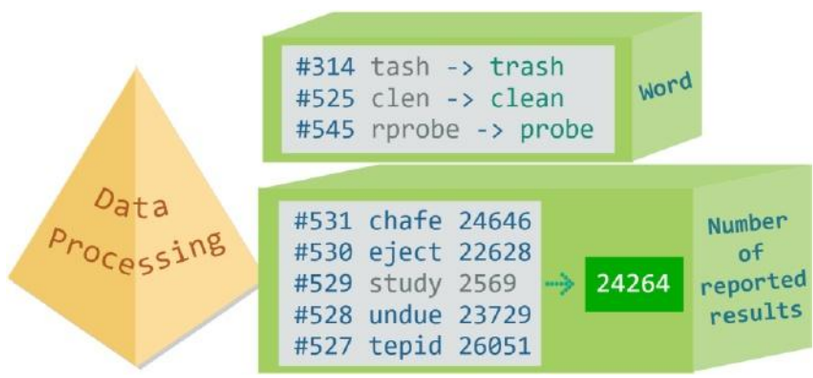


图 2:数据预处理

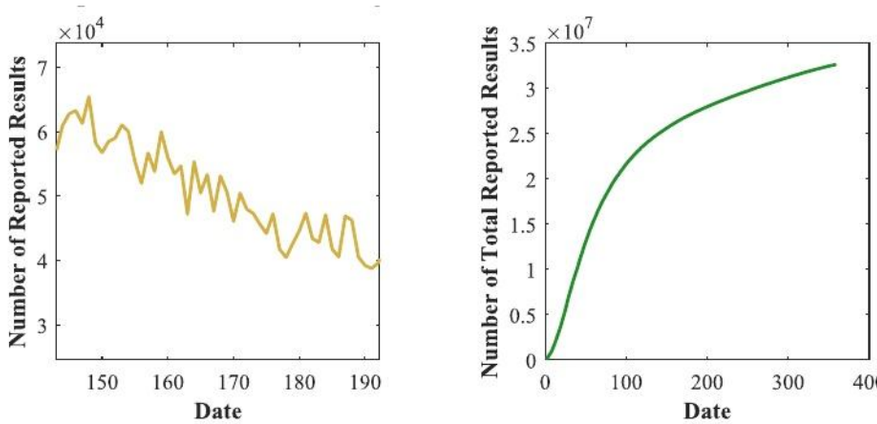
5 任务 1:报告数量预测模型&游戏模式选择

为了探索从 Twitter 获得的报告结果数量随时间的变化模式，我们首先开发了一个可解释的模型，用于描述和预测报告的数量。从统计学的角度来看，我们分别基于三阶高斯回归和非齐次泊松过程描绘了报告数量的长期时间趋势和随机波动。此外，我们观察到报告数量随机波动的大小不仅与时间有关，而且与当前的热量水平有关，因此我们引入了流行度松弛函数来修改随机过程

模型。最后，我们列举了与单词相关的 8 个属性，并分析了其影响单词属性对玩家通过散点图选择游戏模式的影响。

5.1 数据探索

报道结果数量随时间不断变化，图 3 从人数角度展示了游戏热度随时间的动态格局(日期以 1 月 7 日为起点)。总的来说，在时间尺度上的报道数量上存在一定的流行传播规律规律。当世界大战在早期爆发时，数量显著增加;然而，当流行期过去后，数量呈现下降趋势，趋于平稳，如图 3(a)所示。值得注意的是，每天的报道数与整体趋势之间存在一个小的随机波动。此外，图 3(b)描绘了累计报告数随时间(共 359 天)的增长统计。也就是说，报告数量随时间的变化过程可以分为两部分，分别是趋势信号和随机波动。



(a) Short-term random fluctuations (b) Long-term cumulative trend

图 3:报告数量的描述

同时，我们发现短时间内的报道数波动与该时期游戏的报道数之间存在相关性。考虑到推特上分享游戏报道的社交属性，我们近似认为某一时间段内的报道数代表了该游戏近期的热度。

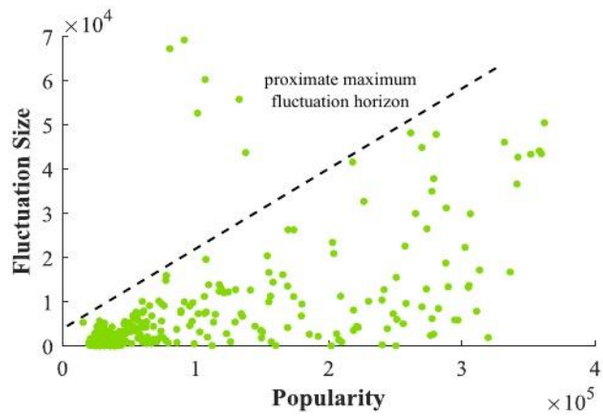


图 4:报道数波动大小与游戏受欢迎程度的关系

如图 4 所示，其横轴为该数字的两期移动平均线的游戏报道，其纵轴为每日报道数相对于当日两期移动平均线的波动。可以看出，随着报道数滑动窗口均值的增大，波动幅度变大，波动幅度的边界大致呈线性。在游戏火爆的这一时期，要准确预测游戏报道的数量就比较困难了。

5.2 世界报表数量预测模型

5.2.1 报表数量预测模型的建立

我们想要基于现有的数据建立一个数学模型来描述 Twitter 上的报告结果数量随时间变化的过程，并预测未来某一时期的人气，该模型对变化过程进行解释。这个问题是近年来经常被讨论的人气预测问题。

通过回顾文献[4]，我们了解到业界常用的两类热预测算法，包括基于节点行为动态的时间模型和基于深度学习的方法。然而，它们并不适用于本文研究的场景。这主要是由于以下两个原因:

1)现有数据集不包含具体的信息，比如记者是谁，总共有多少人。基于这个数据集构建节点模型是不够的。

2)深度学习没有很好的可解释性，需要更多的训练数据才能获得更好的预测结果。

因此，我们从统计学的角度建立了基于三阶高斯回归和非齐次泊松过程的世界报告数量预测模型。

基于高斯回归的趋势预测模型

在数据文件中，报告数量的时间序列有明显的趋势。我们尝试了几种回归算法来拟合报告数量随时间变化的趋势，最好的结果是三阶高斯回归，回归方程为：

$$G(t;\boldsymbol{\theta}) = A_1\exp\left[-\left(\frac{t-B_1}{C_1}\right)^2\right] + A_2\exp\left[-\left(\frac{t-B_2}{C_2}\right)^2\right] + A_3\exp\left[-\left(\frac{t-B_3}{C_3}\right)^2\right] \quad (1)$$

其中 $\boldsymbol{\theta} = [A_1, A_2, A_3, B_1 \ldots C_1, C_2, C_3]$ 为回归系数，t 为以天为单位的时间。

然后，我们用最小二乘法对其进行回归，设对每日报告数的观察结果为 $y(t_i)$，回归结果为：

$$\hat{G}(t;\hat{\boldsymbol{\theta}}) = \sum_{n=1}^{3} \hat{A}_n \exp\left[-\left(\frac{t-\hat{B}_n}{\hat{C}_n}\right)^2\right]$$

则其损失函数为:

$$L(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{359}\left[y(t_i) - \hat{G}(t_i;\hat{\boldsymbol{\theta}})\right]^2 \quad (2)$$

我们以 $\min_{\hat{\theta}} L(\hat{\boldsymbol{\theta}})$ 为回归结果，对应的 $\hat{G}(t;\hat{\boldsymbol{\theta}})$ 为预测趋势。

基于非齐次泊松过程的报告数预测模型

泊松分布描述了一定数量的事件发生的概率在一段时间内事件发生率恒定的条件下，从而可以描述一天内上传一定数量报告的概率。假设每天的报告数量服从泊松分布，则这些泊松分布在时间上形成非均匀的泊松过程，即到达强度随时间变化的泊松过程。

当天的报告数量是一个随机过程 $X(t)$，服从具有到达强度的非齐次泊松过程 $\lambda(t)$。当天的报告数量的概率为 t：

$$P_k(t) = P\{X(t)=k\} = \frac{\lambda(t)^k}{k!}e^{-\lambda(t)} \quad (3)$$

其中 $\lambda(t)$ 的含义为当天报告数的平均值 $m_X(t)=E[X(t)]$。然而，均值函数无法从可用的数据中推导出来统计数据。因此，我们退一步使用前面的趋势预测结果 $\hat{G}(t)$ 。用高斯回归来近似报告数量的均值函数M(XT)，因此，通过引入非齐次泊松过程可以很好地描述所报告数的随机波动 $\bar{\lambda}(t)=\hat{G}(t)$。

基于流行度松弛函数的随机过程修正

由于上述随机过程X(t)在实践中并不是一个完全独立的增量过程，因此其随机波动的大小受到其受欢迎程度的影响。本文借鉴网络舆论[5]的生命周期，划分了 Wordle 流行趋势的生命周期。考虑到这些数据是从 1 月 7 日开始计算的，因此省略了初始的“形成”阶段。

1.爆发期:由于人气的增长和社交平台上的成果分享，玩家数量激增。推特用户对这类话题的关注和行动指数飙升至峰值，波动较大，其范围的不确定性较大。

2.退潮期:随着游戏的新鲜感对玩家来说已经过去，游戏的热度是时间敏感的。而且，玩家分享成就的欲望会减弱，但这并不意味着此时玩家数量会减少。尽管如此，与爆发期相比，人气的整体波动还是要低一些。

3.休眠期:人气趋于平稳，游戏依然有很多忠实玩家，话题依然存在。总体来说，起起伏伏的变化不大。

在本文中，我们观察到，当游戏流行时，报告数量的随机波动并不完全服从泊松分布，波动明显更大。随着游戏受欢迎程度的减弱，波动也随之减弱。因此，我们引入人气松弛函数来修改随机过程模型。

如 5.1 节所述，人气松弛现象的边界可以近似地简化为线性边界，因此我们定义人气松弛函数其中 $f(k)=l\cdot k+m$ 为报道数，l,m 为常数。

$$f(k)=l\cdot k+m$$

修改后的随机过程得到强度函数为:

$$\mathring{\lambda}_k(t)=\lambda(t)\cdot f(k) \tag{4}$$

因此，报告数量在方程(3)修正后的当天的概率为:

$$\mathring{P}_k(t)=\frac{\mathring{\lambda}_k(t)^k}{k!}e^{-\mathring{\lambda}_k(t)} \tag{5}$$

我们可以基于一定置信度 $[\lambda(t)-lb,\ \lambda(t)+rb]$ ，计算出当天t报告数的预测区间β，如式(6)所示 $\mathring{P}_k(t)$

$$\begin{cases} lb=\arg\min\limits_{N}\left|\dfrac{\beta}{2}-\sum\limits_{n=1}^{N}\mathring{P}_{\lambda(t)-n}(t)\right| \\ rb=\arg\min\limits_{M}\left|\dfrac{\beta}{2}-\sum\limits_{m=1}^{M}\mathring{P}_{\lambda(t)+m}(t)\right| \end{cases} \tag{6}$$

5.2.2 建立未来报告数量结果的预测区间

基于高斯回归的趋势预测模型，我们预测了长期。回归系数如表 2 所示，我们将与预测区间一起展示具体的趋势预测效果。

表 2:趋势预测模型的回归系数

$\hat{\theta}$					
A_1	1.57e+05	B_1	33.01	C_1	30.79
A_2	9.69e+04	B_2	48.2	C_2	75.31
A_3	4.846e+04	B_3	5.864	C_3	386.7

然后，通过修改非齐次泊松过程的报告数预测模型，得到 75%置信度的报告数预测区间。图 5 显示了该模型对具有未来预测的报告数结果的当前描述，横坐标为截至 2022 年 1 月 7 日的天数。

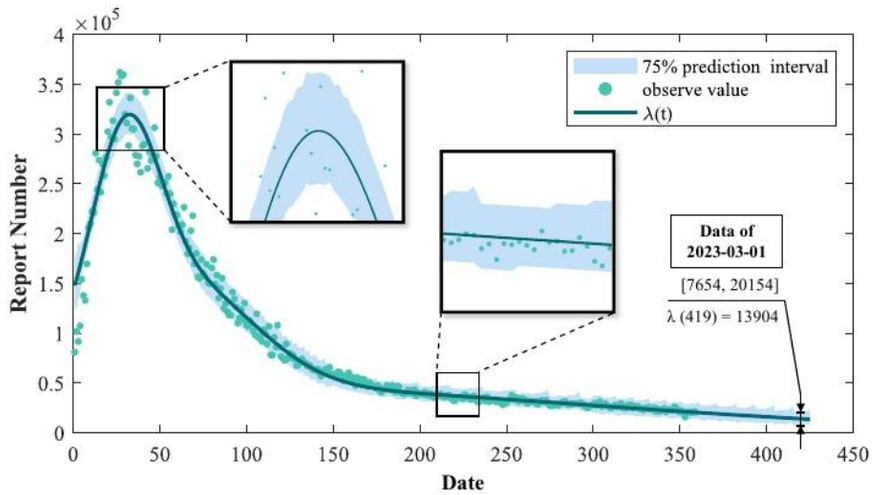


图 5:报告数字趋势和 75%置信水平区间预测

从上图可以看出，我们的模型可以更准确地预测报告数量的长期趋势，也可以大致估计每天报告数量的随机波动区间。值得注意的是，预测区间很好地反映了热度的时效性，当“日期”为 40 左右时，由于用户数量处于“爆发期”，用户数量呈现激增和显著波动。“日期”在 220 左右时，用户数量处于“休眠期”，数量和波动相对较小，趋于稳定。

我们预测 2023 年 3 月 1 日报告结果的数量收敛到 13904(图 5 中横坐标的值为 419 时)，不同置信水平下的预测区间结果如表 3 所示。

表 3:报告数量的预测区间(3 月 1 日)

Confidence level	Left border of the prediction interval	Right border of the prediction interval
75%	7654	20154
85%	5434	23657

一般来说，报道数的整体变化格局是由游戏的社会属性和社会规律决定的。而这种变化格局呈现出明显的趋势，因此通过回归模型可以获得较好的预测效果。

从整体变化趋势来看，报告数量也存在一定程度的随机波动。这种随机波动具有随时间和热度变化的统计特征。因此，可以用随机过程来描述它。

5.3 博弈模式选择分析

5.3.1 词属性分析

首先，我们对可能涉及到的词属性进行分析，这些词属性可以从现有数据集中主要从三个方面挖掘:字母频率、字母位置和常用词根等。

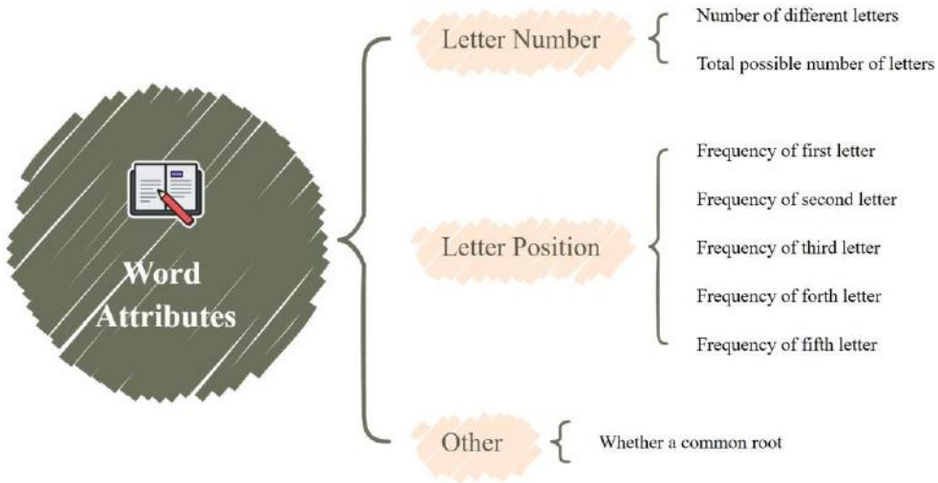


图 6:词属性分析

不同字母的数量:表示为一个单词中不同字母的数量。统计上，取值范围为 3～5。例如，单词“happy”的这个属性值是 4。这个属性反映了这个词的内部可变性。

可能出现的字母总数:表示所有字母出现频率的总和

一个字。假设在 359 个单词的结果数据中，字母“α”出现的频率为 fh+fα+2fp+fy，表明了这个词的整体使用趋势。

首字母出现频率:表示单词首字母出现的频率。例如在 359 个数据项中，每个单词首字母出现的总次数为 359 次，而首字母为“h”的单词的百分比为Ph1，

单词“happy”的值为 ph1/359。这个属性反映了这个词的局部位置倾向。

第二个字母出现的频率:表示单词第二个字母出现的频率。例如，在 359 个数据项中，每个单词的第二个字母出现的总次数为 359 次，且第二个字母“a”的单词所占的百分比为 Pa2。这个属性的值为 Pa2/359。“快乐”这个词是。这个属性反映了这个词的局部位置倾向。

第三个字母出现的频率:表示单词中第三个字母出现的频率。例如，在 359 个数据项中，每个单词的第三个字母的总出现次数为 359 次，而含有第三个字母“p”的单词的百分比为，则此属性的值为 Pp3,这个属性反映了这个词的局部位置倾向。

第四个字母的频率：它表示一个单词的第四个字母的频率。例如，在359个数据项中，每个单词的第四个字母的总数为359个，具有第四个字母“p”的单词的百分比为。这个属性对单词“快乐”一词的价值是Pp4。这个属性反映了这个词的局部位置趋势。

第五个字母的频率：它表示一个单词的第五个字母的频率。例如，在359个数据项中，每个单词的第五个字母的总数也是359个，带有第五个字母“y”的单词的百分比为，那么单词“fappy”的这个属性的值为Py5。这个属性反映了这个词的局部位置趋势。

是否有共同词根:表示一个单词内部是否有共同词根。例如，如果单词“manly”包含词根“-ly”，则该单词的值为 1;否则，它的值为 0。这个属性反映了单词的局部规律性。

5.3.2 词属性对模式选择的影响分析

我们想弄清楚前一节列出的单词的 8 个属性是否会影响用户对游戏模式的选择。因此，对于每个属性，图 7 绘制了一个散点图，比较日常单词属性和困难模式选择之间的关系。

在下图中，每个散点图的水平坐标是困难模式选择的百分比(单位为%)。可以注意到，单个属性与困难模式的百分比没有很强的相关性。呈现的单词属性不会影响困难模式报告数据的比例。

我们认为造成这种现象的原因是玩家没有提前获知解词。也就是说，在大多数玩家选择游戏模式之前，解词属性是未知的，因此玩家的选择与其没有高度的相关性。

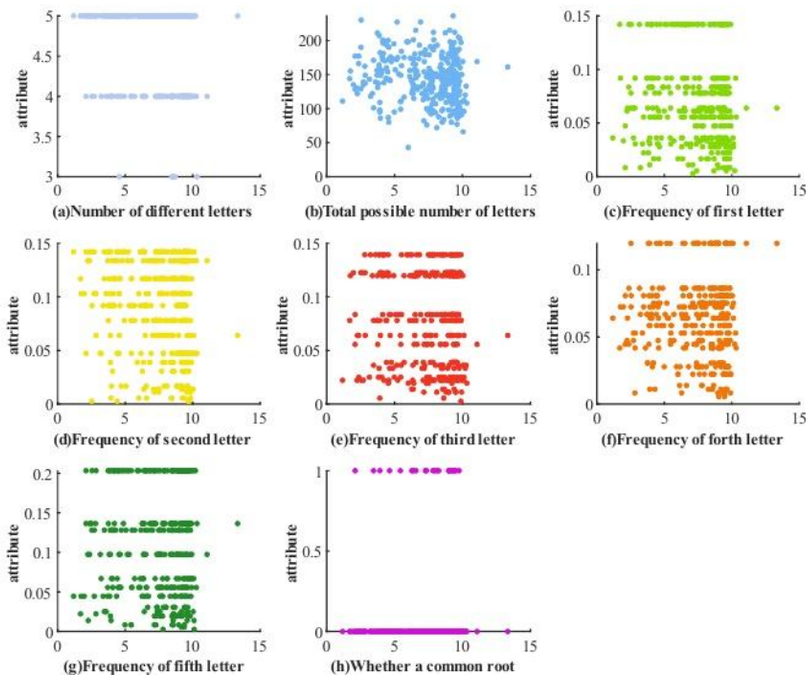


图 7:困难模式选择比例与单词属性之间的 相关性

那么与困难模式选择比率相关的主要因素是什么呢?