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。

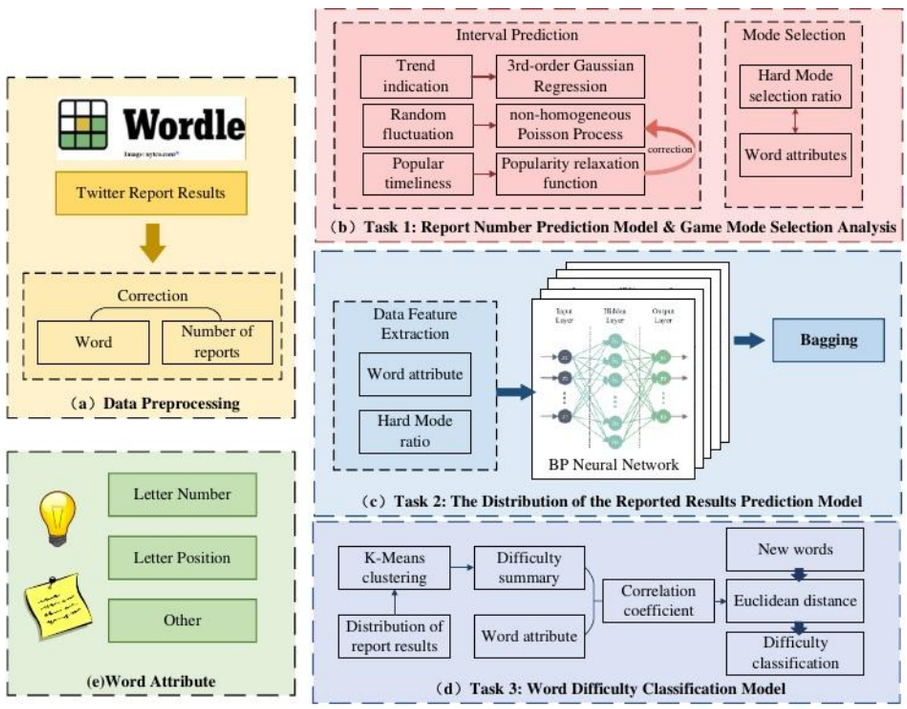


图 1:我们论文的结构

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。

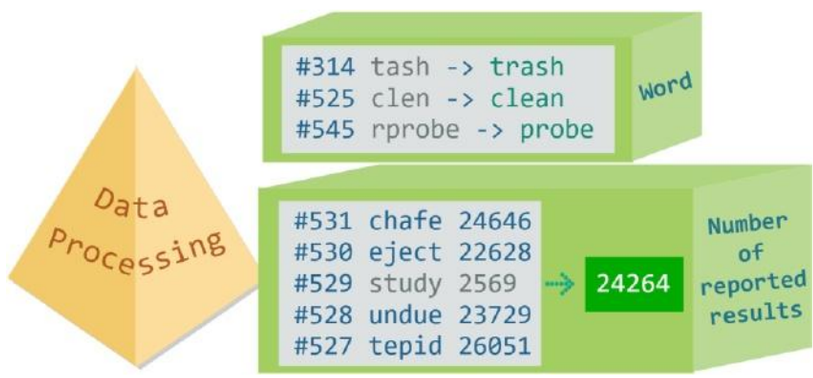


图 2:数据预处理

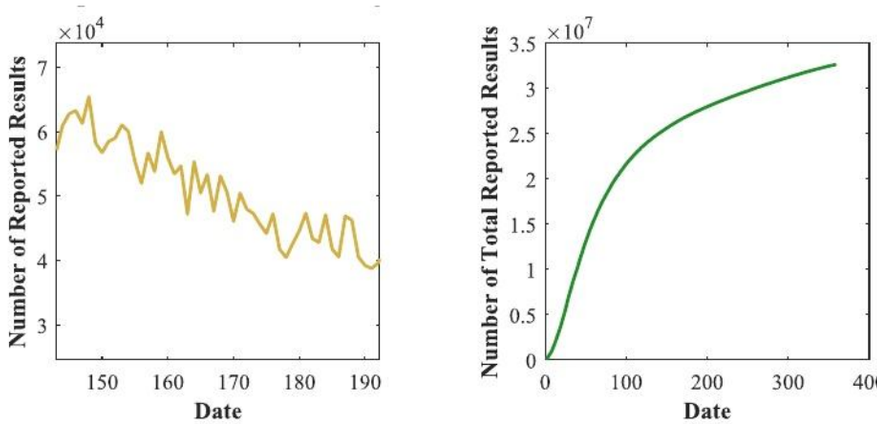
5 任务 1:报告数量预测模型&游戏模式选择

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。

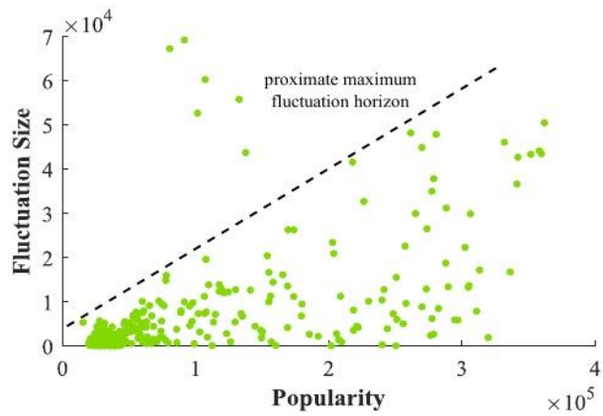


图 4:报道数波动大小与游戏受欢迎程度的关系

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：

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

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。

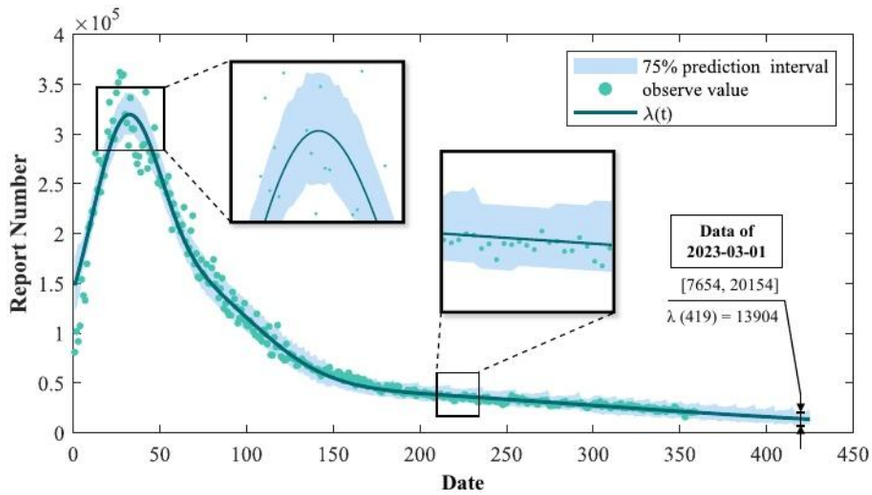


图 5:报告数字趋势和 75%置信水平区间预测

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。

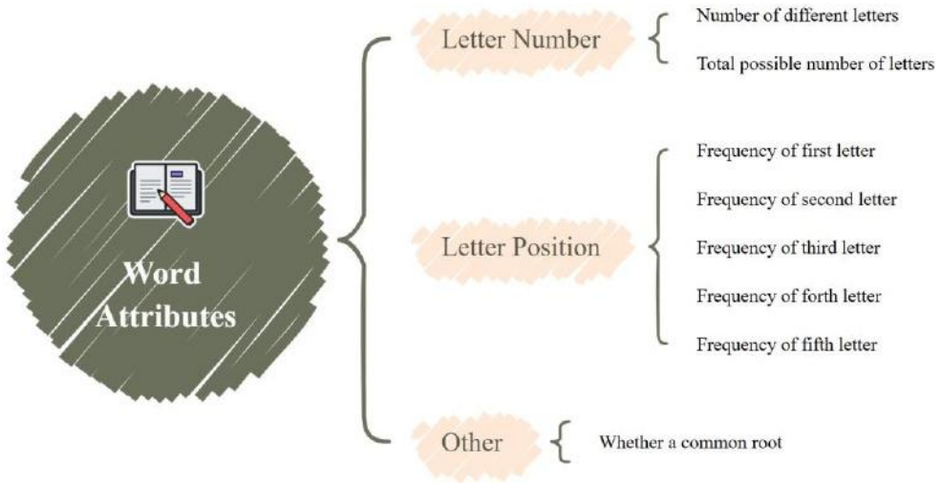


图 6:词属性分析

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+f α +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。

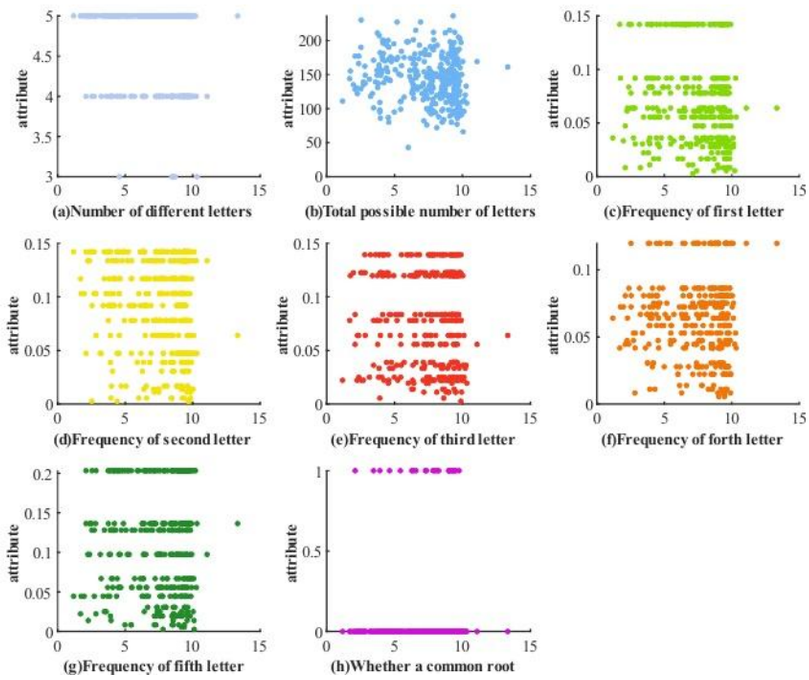


图 7:困难模式选择比例与单词属性之间的 相关性

那么与困难模式选择比率相关的主要因素是什么呢?

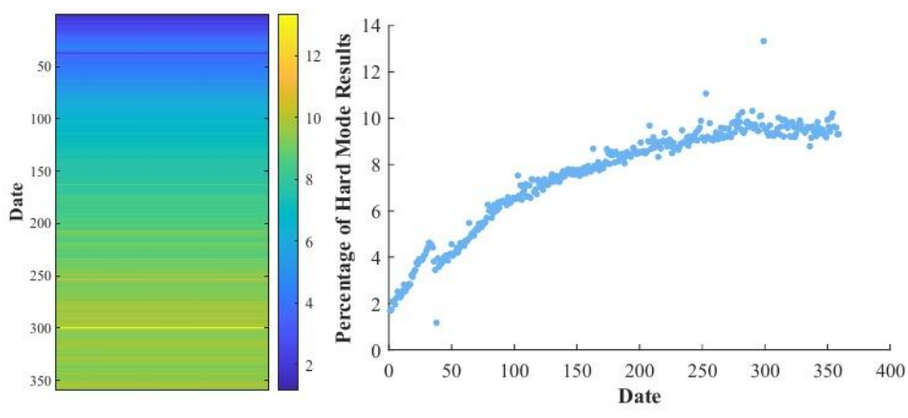
表 4 提供了一些困难模式百分比的单词内容和日期的进一步可视化。我们可以发现，在比例排名中，排名前十的报告和排名后十的报告都没有明显的词属性特征。但是，困难模式选择比例较高的数据往往出现在 2022 年的最后三个月，而比例较低的数据往往出现在 2022 年的第一个月，且与时间的相关性较高。

表 4:在困难模式中报告的分数百分比

Top 10			Bottom 10		
Date	Word	Percentage	Date	Word	Percentage
2022/11/1	piney	13.33%	2022/1/16	solar	2.36%
2022/9/16	parer	11.07%	2022/1/14	tangy	2.35%
2022/11/30	study	10.66%	2022/1/15	panic	2.26%
2022/10/23	mummy	10.32%	2022/1/12	favor	2.23%
2022/10/15	catch	10.27%	2022/1/10	query	2.09%
2022/12/26	judge	10.21%	2022/1/9	gorge	2.09%
2022/10/30	waltz	10.12%	2022/1/11	drink	1.96%
2022/10/12	ionic	10.11%	2022/1/8	crank	1.74%
2022/10/29	libel	10.08%	2022/1/7	slump	1.69%
2022/12/25	extra	10.04%	2022/2/13	robin	1.17%

为了进一步探索这一模式，我们可视化了困难模式选择百分比随时间的变化，如图 8(a)所示，随着时间的增加，热图的颜色从蓝色逐渐变为绿色，只有少数日期出现异常波动，即

从图 8(b)的散点图中也可以看出，困难模式选择呈现出增加的趋势。有一个非常明显的趋势是，困难模式选择的百分比随着时间的推移而增加，不仅是增量的，而且增长速度也在逐渐放缓。



(a) Heat map over time (b) Trends over time

图 8:困难模式选择百分比描述

我们认为，这一方面是因为随着玩游戏时间的增加，用户在游戏中的表现能力也得到了充分的提升，所以一些热爱挑战的用户会逐渐倾向于选择难度更高的模式。另一方面，很多用户可能会以更休闲的心态玩游戏，即使能力提高了，也不会考虑提高游戏的难度。换句话说，玩家对自己表现能力的信心水平和游戏心态可能是他们是否选择困难模式的主要原因，而不是单词的属性。

6 任务 2：一个关于报告结果分布的预测模型

为了预测未来报告结果的分布，我们首先提取并构造数据特征。然后，我们建立了一个 BP 神经网络模型，以 7 个数据特征作为输入，输出 7 个猜词结果的分布。最后，采用 Bagging 算法对多个 BP 神经网络进行整合，通过硬投票机制得到最终的预测结果，以减小预测结果的泛化误差。

6.1 建立基于 BP 神经网络的猜词结果分布预测模型

考虑到推特上分享的猜词次数的结果分布很可能与整个玩家群体的结果分布不同，并且与许多难以量化和计数的因素有关，包括玩家的心态以及大多数玩家对当天单词的熟悉程度。因此，对猜词结果分布进行机械建模可能是困难和不现实的，我们决定对这个问题的数据进行建模。当然，该方法隐含地假设了一个条件，即历史数据很好地代表了世界上所有可能的问题和玩家的答案，这是我们模型最大的不确定性。

给定解决方案单词和相应的日期，我们可以访问单词的属性以及基于时间进程可以预测的所有信息。历史数据的总数为 359，这对于深度学习算法来说可能会造成欠拟合问题，而数据的大小对于 BP 神经网络来说是可以接受的。因此，我们决定建立一个基于 BP 神经网络的报告分布预测模型。

6.1.1 数据特征的提取与构建

在具体构建神经网络之前，我们首先提取和构建数据特征。如前所述，我们希望根据解词和相应的日期来预测猜词结果的分布。此时的数据特征来源包括单词本身和可预测的时间过程。

可以从单词本身获得的信息已经在 5.3.1 节中进行了探讨，即单词属性。我们将词属性分为 3 类，包括字母频率、字母位置和常用词根词。其中，我们对是否包含常用词根的属性进行统计导致了数据的不平衡。作为一个布尔值，这个属性的 359 个数据中只有大约 20 个是真值。因此，这个属性对于神经网络的训练是没有价值的，我们不保留这个特征。此外，为了压缩数据维度，也去掉了具有一定重复性的字母总数(Total possible numberof letters)这个特征。

最后，我们还选择了选择 Hard Mode 的人的百分比作为特征，因为游戏的困难选择也影响了猜测次数的分布。如 5.3.2 节所分析，该特征具有明显的趋势，且随时间波动较小，因此该特征在确定的未来日期内是可预测的。2022 年下半年这一特征值大多在 9.5%左右波动，变化不再显著。因此，我们做了一些简化，并假设在未来一段时间内将保持这种状态，而不做更精确的预测。我们将后 59 天的难模选择比的平均值作为预测值。

6.1.2 BP 神经网络的构建

对于报告结果分布预测问题，如 Number of different letters 等 7 个数据特征构成输入空间，7 个报告结果分布构成输出空间。该问题的数学本质是拟合一个从 7 维输入空间到 7 维输出空间的映射。对文献[6] 的回顾表明，为了拟合这样的有限维空间映射，构建神经网络通常只需要一个隐藏层。因此，该模型中的 BP 神经网络共有三层，分别是输入层、隐藏层和输出层。

由于输入空间和输出空间都是 7维的，所以输入层的神经元数Ni和输出层的神经元数No都是7维的考文献[7]中提出的经验公式，我们设置了隐藏层神经元的个数 N_h 满足式(7)。

$$N_h = \frac{N_s}{\alpha(N_i + N_o)},\ \alpha = [2, 10] \tag{7}$$

其中为训练集的大小，为常数。

至此，我们已经完成了 BP 神经网 络的构造、结构，如图 9 所示。

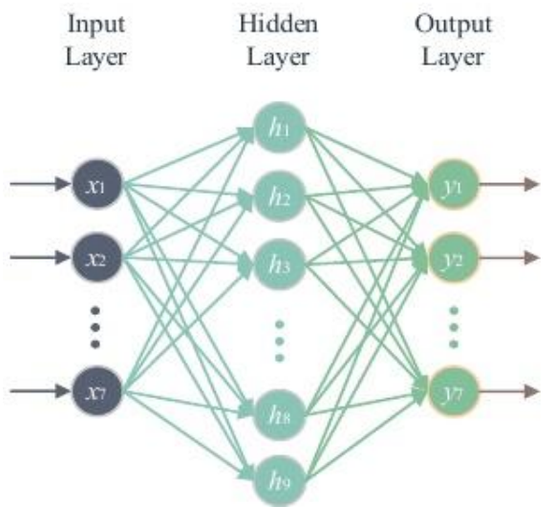


图 9:BP 神经网络的结构

6.1.3 基于 bagging 的综合 BP 神经网络预测模型

在 BP 神经网络的训练中，我们根据 85%随机选择训练集。经过几次测试，我们发现每次训练得到的神经网络在测试集中的极少数样本上总是存在不可接受的误差。而且，对于不同的神经网络，具有较大预测误差的样本的重复率并不高。因此，我们决定采用 Bagging 算法对多个 BP 神经网络进行整合。然后，通过硬投票机制推导出最终结果，具有降低预测结果泛化误差的效果。

在积分算法中，我们通过随机抽取总体数据次数的 85%来获得不同的子训练集，每次抽样的百分比也是 85%。然后，基于这些用相同参数训练的子训练集得到神经网络。接下来，这些神经网络被用来预测猜词的分布。所有神经网络直接输出各自的预测结果，最终的预测结果由少数多数投票机制[8]获得。这种集成算法的流程如图 10 所示。

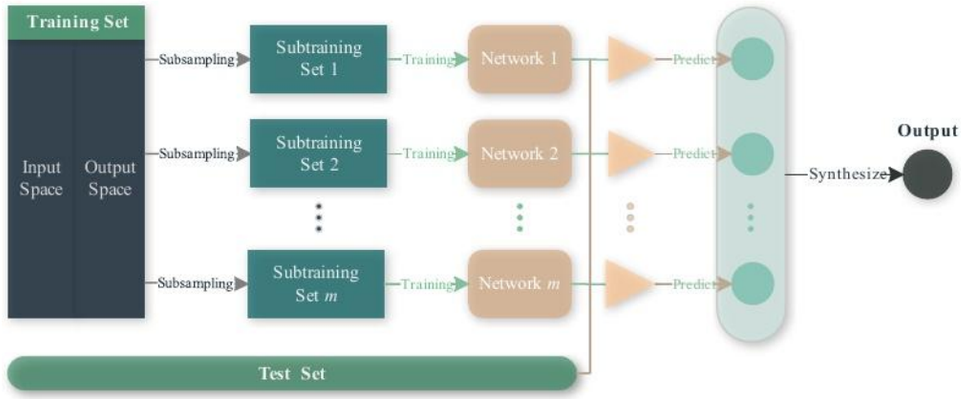


图 10:集成 BP 神经网络示意图

为了避免偶然误差对投票结果的影响，被用于训练。

6.2 影响模型的不确定性分析

我们拥有的数据集和选择的数据特征可能不能很好地代表 Wordle 的提问和玩家的回答，这将导致我们模型的泛化性很差。

此外，用于训练的数据集的大小虽然可以接受，但仍然相对有限。在用积分算法随机选择子训练集的过程中，子训练集中相同样本的比例可能会相当高。这可能会使单个神经网络之间的同质性过高，无法起到整合和投票机制的作用。

6.3 预测模型的结果分析

首先，我们按时间降序排列数据集，选择前 85%的数据作为总训练集。剩下的 15%的数据将作为测试集来验证模型效果。然后，每次随机选取总训练集的 85%作为子训练集，对神经网络进行训练。最后，对这些神经网络进行整合，用测试集测试预测效果。

经过几次尝试，我们发现我们能够获得更好的预测结果 7 个猜词结果的分布。其中，3 次尝试的预测结果误差最大，但均方误差仍不超过 5%。

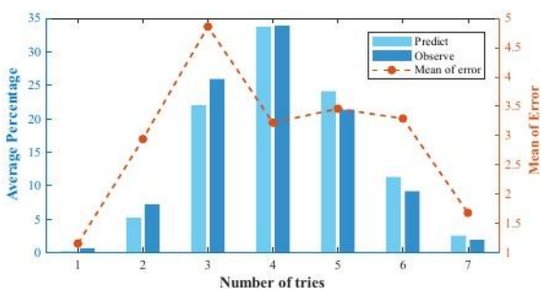


图 11:均值比较和均方分布

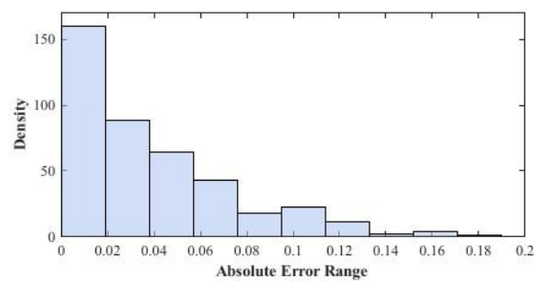


图 12:绝对误差分布

还统计了所有误差的分布情况。如图 12 所示。虽然误差是不可避免的，但大部分都维持在 6%以内。其中，2%以内的绝对误差约占 38%，5%以内的绝对误差达到 80%以上。有- - - - - -

在此之前，我们有超过 80%的置信度，预测结果的绝对误差不超过 5%。

最后，我们预测 2023 年 3 月 1 日单词 ERRIE 的猜测次数分布，如表 5 所示。

表 5:ERRIE 一词的结果分布预测

1 try	2 tries	3 tries	4 tries	5 tries	6 tries	(X)
0%	1%	6%	25%	31%	25%	13%

7 任务 3:单词难度分类模型

为了对解词进行合理分类，我们首先基于 K- Means 聚类算法对难度进行分类。然后，我们基于 Pearson 相关系数探索词属性与难易度分类之间的关联，构建了词难易度分类模型。最后，可以根据这种相关性对新单词进行分类。

7.1 建立单词难度分类

7.1.1 基于 K-Means 聚类的词难度归纳模型

在按难度对解词进行分类之前，我们需要先定义难度。为了使模型结果更接近用户的游戏体验，我们决定根据用户猜测次数分布来定义一个难度划分。需要注意的是，这只是难度的指示，并不是确定解词难度的原因。

根据我们对难度的定义，猜测的分布反映了这个词的难度。因此，对词汇难度进行分类的第一步是将猜测次数的分布归纳为一定的类别。这个问题的本质是探索历史猜测词的频率分布的同质性和差异性。从数学的角度来看，数据的同质性和差异性可以用距离来描述，数据可以用距离来分组。因此，我们决定使用 K-Means 算法来总结单词的难度。

首先，完成猜测的百分比越高，说明这个谜题的难度越高。因此，猜测次数分布的难度差异是一种绝对差异，可以用欧几里得距离来描述。则猜测结果A与猜测结果B的差值体现为两者分布向量的欧几里得距离，如式(8)所示。

$$D_E(\mathbf{A},\ \mathbf{B}) = \sqrt{\sum_{i=1}^{7}(A_i - B_i)^2} \tag{8}$$

然后，我们需要确定分类的难度等级，即确定 K-Means 聚类算法的质心数量。我们会找到最好的结果 k,通过多次尝试。目标是使难度分类无重复、无遗漏。

接下来，在整个样本集中分别随机抽取样本 $x = \{x_1, x_2 \dots x_{359}\}$,作为初始聚类中心，并标注为：

$$\mu_1^{(0)}, \mu_2^{(0)}, \cdots, \mu_k^{(0)}$$

由于我们决定使用欧几里得距离来度量样本之间的差异，因此我们将式(9)作为算法的优化目标，即使所有样本到其聚类中心的欧几里得距离最小。

$$J(c,\ \mu) = \min \sum_{i=1}^{359} \|x_i - \mu_{c_i}\|^2 = \min \sum_{i=1}^{359} D_E(x_i,\ \mu_{c_i}) \tag{9}$$

其中 c_i 为样本所属的聚类。

迭代计算每个样本点与每个质心之间的距离。同时，将样本分配到离相应质心距离最小的聚类中:

$$c_i^{(t)} = \arg\min_m \|x_i - \mu_m^{(t)}\|^2$$

在每次迭代结束时，将每个聚类中样本点的平均距离计算为下一次迭代的质心:也得到了它的聚类中心。

$$\mu_n^{(t+1)} = \frac{1}{b} \sum_{i:c_i^{(t)}=n}^{b} x_i$$

同时，将本次迭代的目标函数值$J^{(t)}$ 与前一次迭代的目标函数值$J^{(t-1)}$ 进行比较。目标函数收敛，聚类结束。最后是每个样本的难度等级，与典型相对应

此外，还得到了每个样本与其聚类中心之间的距离矩阵。

7.1.2 基于 Pearson 系数的词属性与难度等级的相关分析

在获得难度评分结果后，我们使用 Pearson 相关系数分析单词各属性与难度评分之间的关联。这是因为要确认属性和难度之间的因果关系比较困难，但在分类时可以使用相关性来代替。

首先，我们选择字母在 5 个位置出现的频率，如 5.3.1 节提到的第一个字母的频率(F1)、第二个字母的频率(F2)，作为要分析的单词属性。在此基础上，我们进一步统计了数据集中每个解词所包含的词类(WCN)数量和元音(VN)数量。到目前为止, 我们已经获得了 7 个待分析的词属性。

然后，我们选择离各自聚类中心最近的样本作为典型样本从难度等级。得到具有代表性的典型 $S_j=[a_{j1}\ a_{j2}\ \dots\ a_{j7}]$,式中， $i=1,2,\dots 359$ i为第个样本, $j=1,2,\dots,k,$ j为第2个 难度分级的典型样本序列号。

对于第 t 个样本的第 i 个属性，我们分别计算样本属性与典型样本k对应属性之间的欧氏距离。

$$\widehat{dist}_{ij}^{(t)}=D_E(a_{it},\ a_{jt})$$

我们让 $W_i^{(t)}=[dist_{i1}\ dist_{i2}\ \dots\ dist_{ik}]$ ，表示为属性距离向量。让 $\widehat{W}_i^{(t)}=\left[\widehat{dist}_{i1}^{(t)}\ \widehat{dist}_{i2}^{(t)}\ \dots\ \widehat{dist}_{ik}^{(t)}\right]$ ，表示聚类距离向量。

接下来，我们计算属性距离向量的 Pearson 相关系数 $\widehat{W}_i^{(t)}$ 和聚类距离向量 $W_i^{(t)}$ ，得到相关系数

$$\rho_i^{(t)}=\frac{\operatorname{cov}\left(\widehat{W}_i^{(t)},\ W_i^{(t)}\right)}{\sigma_{W_i^{(t)}}\sigma_{W_i^{(t)}}}$$

$\sigma_{\widehat{W}_i^{(t)}},\ \sigma_{W_i^{(t)}}$ 式中分别为 $\widehat{W}_i^{(t)}$ 和的方差 $W_i^{(t)}$ 。

最后，我们计算了每个属性相关系数的平均值 $m_\rho^{(t)}$ 对于所有的样本。当考虑到 第 t单词属性和难度相关时，我们也设置了边界。最后，我们对属性进行过滤，得到 属性向量。$attributes_i=[a_{il_1}\ a_{il_2}\ \dots\ a_{il_u}]$

7.1.3 基于欧几里得距离的词难度判别

对于未来的解词，我们可以通过计算其与每个典型样本的相似度来判断其难易程度。

因为我们建立了一个预测模型来预测数量的分布对于 6.2 节中给定的解决方案单词在未来日期的猜测，我们有两个判断基础单词的难度。一种是基于猜测单词数量的预测分布X，另一种是基于解词的属性向量C。

$$\hat{c} = \arg\min_{i} \left\| \hat{x} - \mu_i \right\|^2$$

式中 μ_i 是第 i 个难度的重心，$i=1,2,...k.$

基于属性向量的难度确定方法:首先计算解词的属性向量，然后计算解词的难度c是根据欧几里得距离和 K-Means 的赋值原则来判断的算法。

$$\hat{c} = \arg\min_{j} \left\| attributes_x - attributes_j \right\|^2$$

其中 $attributes_j$ 为第j个难度的典型样本属性向量 $j=1,2,...,k$ 。

7.2 单词难度分类结果分析

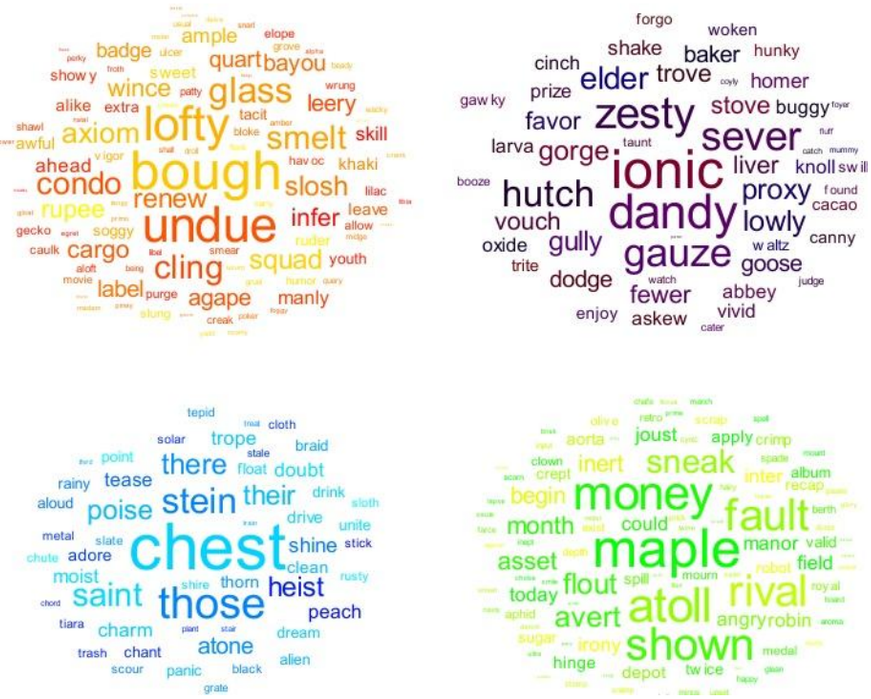


图 13:难度分类结果词云

通过 K-Means 聚类得到四类难度的结果，按照难度从小到大的顺序分为 Easy、Normal、Hard、Master，并选择一个典型单词进行可视化分析，如图 14 所示。

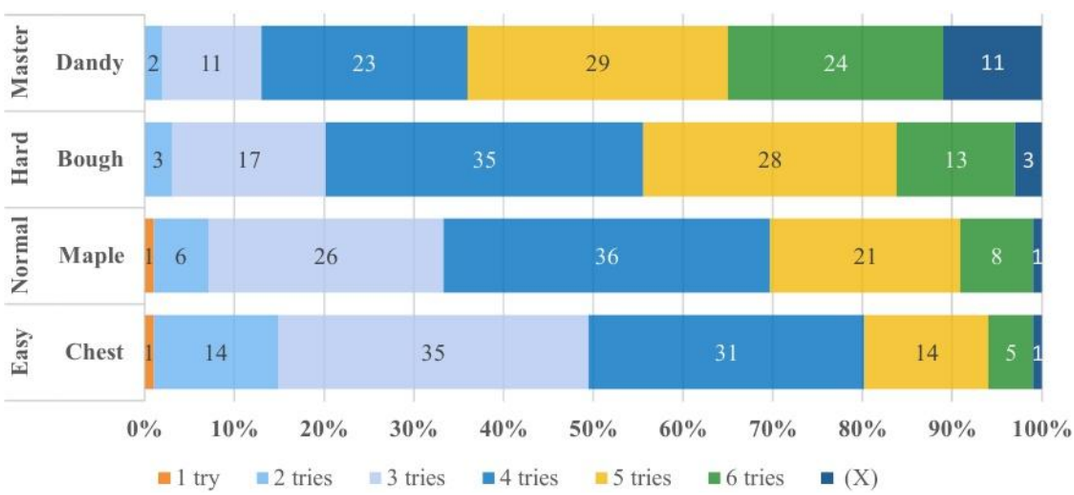


图 14:单词难度分类结果条形图

从上图中我们可以看到，当谜题更容易时，报告的分布倾向于猜测解决的单词的频率更低，反之亦然。比如“Chest”这个词的报告分布主要是在 2 次、3 次和 4 次，6 次和 7 次的分布较小。

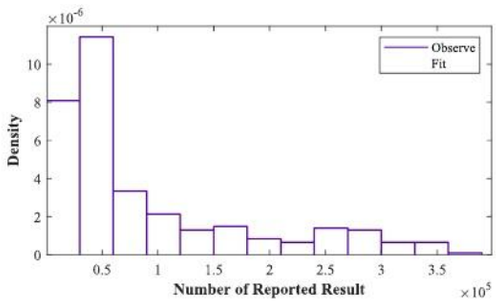
表 6:单词属性与难度的相关系数

m_ρ						
F1	F2	F3	F4	F5	WCN	VN
0.637634	0.813854	0.336972	0.417339	0.311806	0.81528	0.8989

在计算单词属性与难度等级之间的 Pearson 相关系数后，我们得到了表 6 中的结果。上表显示了一些属性具有与难度有很强的相关性。我们将大于 0.6 的单词属性作为难度分类属性，形成属性向量。最后，我们计算了单词 EERIR 的属性向量到 attributesi=[ai1,ai2,ai6,ai7]的距离第 th 难度典型样本的致敬向量，即 R1 = 4.2385 R2 = 3.0021,R3 = 2.1823,R4 = 0.6250，与“丹迪”的距离最近。因此，EERIR 单词的难度为“Master”。

8 任务 4:其他有趣的特征

在探索报告数量的统计属性时，我们发现 359 天内报告数量分布的密度与其随时间的趋势表现出类似的模式。然后，我们对报告数量的分布进行拟合。如图 15 所示，使用对数正态分布对报告数量进行拟合，可以得到很好的结果。参考图 16 可以发现，通过拟合得到的分布与报告数量随时间的变化具有相当高的相似性。



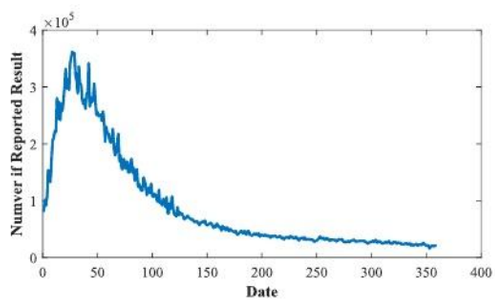


图 15:拟合报告编号的分布图 16:报告编号随时间的变化

但是，我们注意到，报告数量的分布在时间上并不均匀，而且这种分布没有考虑到时间因素。因此，这种拟合并不能为预测报告数量提供太多信息。这导致我们没有更深入地研究这一现象。然而，我们推测，在这种现象之下，可能存在一些有趣的统计性质。

此外，当我们探索用户报告结果的分布时，我们发现 3 次尝试完成游戏的百分比波动最大。如图 17 所示，我们认为这一现象与我们的 3 次尝试游戏完成度报告分布预测模型的预测精度较低有很强的关系。

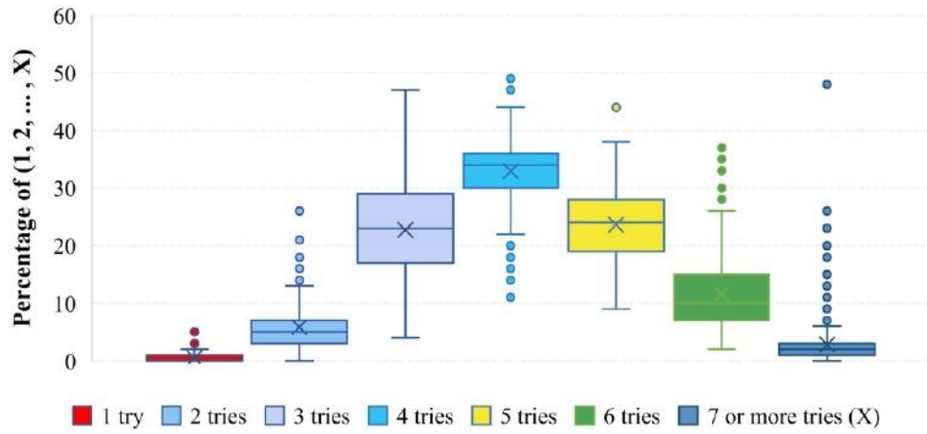


图 17:报告分布方框图

9 敏感度分析

对于 Task 1 中报告数量的预测模型，我们通过将流行松弛函数的截距乘以某个比例因子来微调，以观察模型的灵敏度δ。然后将置信水平为 75%时对应预测区间的平均变化计算为:

$$m_\delta = \frac{\sum_{n=1}^{N}\left(lb_n + rb_n - \widehat{lb_n} - \widehat{rb_n}\right)}{2N}$$

其中 N 为序列的长度。然后，我们分别进行观察 $\delta = 1.1,\ 1.5,$ 时，预测区间的平均变异量达到约 8.3%。当δ=1.5时，预测区间的平均变异约为 32.71%。当δ1.8时，预测区间的平均变异约为 78.2%，且此时预测区间的左边界接近于 0。

我们开发的模型具有一定的鲁棒性，但对于即使在稳定时期也存在较大随机波动的情况，并不十分适用。

10 模型评估和进一步讨论

10.1 优势

我们将时间序列分解为趋势符号和随机波动，然后分别进行预测。据此建立的模型对时间序列的模式有很好的解释。

我们不自己定义单词的难度，而是根据用户玩的结果来概括难度。因此，我们获得的难度评级与用户的游戏体验更相关。

我们应用了积分算法来削弱由于神经网络训练导致的偶然性造成的泛化误差。最终，实现了比单个神经网络更好的预测精度。

10.2 缺点

对于游戏的流行阶段，我们报告的体积预测模型给出的预测区间比其他阶段宽得多。这导致我们的模型在这些阶段的预测结果可能不是很有信息量。

我们预测报告结果分布的模型在很大程度上依赖于数据集的代表性。然而，由于所提供数据集的规模，要保证整个样本空间的代表性就比较困难了。

10.3 进一步讨论

我们建立的模型仅基于报告中给出的现有数据，但其他可用信息可以进一步整合。在玩家数量的预测模型中，除了使用时间序列信息外，我们还可以根据用户发布的文字信息分析游戏的吸引力，获得玩家对游戏的情感体验;基于分享结果的用户的个人信息，我们可以将用户模拟为节点，构建基于社交网络分析的信息传播模型，进而可以刻画影响游戏传播热度的用户行为，做出更全面的人气预测。

11 结论

为了挖掘世界报告信息，我们提出了一系列新颖的模型来解决预测报告数量和结果分布的问题。同时，我们充分分析了解词的属性，可以对给定词的难易度进行分类。

该模型仅依赖于数据文件，具有良好的解释性和合理性。

1.报告数量预测模型可以将数量随时间变化的长期趋势与随机波动相结合。对报告数量进行高斯回归的结果描述了报告数量随时间变化的趋势，而非齐次泊松过程描述了基于趋势的报告数量的随机波动。根据热度的生命周期，我们还引入了人气松弛函数来修正随机过程模型。

2.玩家选择游戏模式的比例受单词属性的影响较小，但与时间高度相关。玩家对自己的表现能力和玩法的自信程度可能是他们选择困难模式的主要原因。

3.报告结果预测模型的分布可以结合影响困难模式比例的词属性特征和时间特征。我们基于 BP 神经网络对猜测结果的分布进行预测，并通过 Bagging 算法对神经网络的预测结果进行整合，提高模型的泛化性能。

4.单词难度分类模型可以将报告结果的分布与单词的属性结合起来。其中，报告结果的分布只反映了难易程度，而单词属性才是造成不同难易程度的根本原因。我们基于 K-Means 算法对难度进行了划分，并基于 Pearson 相关系数探索了单词属性与难度分类之间的相关性。最后，基于这种相关性，可以对生词进行难度分类。

References

[1] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "SEISMIC: A Self Exciting Point Process Model for Predicting Tweet Popularity," ACM, 2015.

[2] H. W. Shen, D. Wang, C. Song, and A. Barabási, "Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes," AAAI Press, 2014.

[3] Q. Wu, C. Yang, X. Gao, H. Peng, and G. Chen, "EPAB: Early Pattern Aware Bayesian Model for Social Content Popularity Prediction," in 2018 IEEE International Conference on Data Mining (ICDM), 2018.

[4] Z. Zhen, S. Shao, X. Gao, G.Chen, "Social Circle and Attention Based Information Popularity Prediction," Journal of Co puter Science, 2021.

[5] J. Feng, "Research on social network event heat prediction based on text analysis," Harbin Institute of Tec nolog , 2018.

[6] G. Ian, B. Yoshua, C. Aaron, "Deep Learning: Adaptive Computation and Machine Learning series," The MIT Press, 2016.

[7] C. D. Ari, "How to choose the number of hidden layers and nodes in a feedforward neural net-work? " StackE c ange, https://stats.stackexchange.com/questions/181/how-tochoose-the-number-of-hidden-layers-and-nodes-in-a-feedforward-neural-netw.

[8] L. Kai, L. J. Cui, "Diversity and Performance Comparison for Ensemble Learning Algorithms," Co puter Engineering, 2008.

信

亲爱的《纽约时报》拼图编辑:

我们很高兴通过这个模型来挖掘世界数据背后的奥秘。无论是为了解开谜题的乐趣，还是为了解决太难的问题，完成《世界》后立即出现的分享按钮都能让玩家将自己当下的情绪传达给更多的人。随着人们继续分享世界大战，它已经成为一种社交货币。

的过去一年在 Twitter 上收集的报告结果数据，我们构建了三个模型:报告数量预测模型，用于对未来报告数量进行区间预测;用于预测未来报告结果猜测次数分布的报告分布预测模型;以及一个单词难度分类模型，用于对给定词汇的解决难度进行分类。我们通过对报告中更深层次信息的全面挖掘，获得了一些有价值的发现，帮助大家发现去年游戏数据中的秘密，开发新的谜题。

首先，我们建立了一个报告数量预测模型来描述过去或预测未来的报告数量间隔，以帮助您更轻松地跟踪世界的流行趋势。我们发现，报告数量变化的整体模式是由游戏的社交属性和社交模式决定的，而报告数量的随机波动则具有随时间和受欢迎程度而变化的统计属性。

我们分析了日常解词的属性，其中包括字母的数量、位置和特殊词根等特征。我们初步分析了单词是否会影响玩家的游戏模式选择。我们的结论是，困难模式选择的比例随着时间的推移而增加，然后趋于平稳，而单词属性并不影响困难模式的比例。玩家对自己表现能力的信心程度和游戏心态可能是他们是否选择困难模式的主要原因。

接下来，我们基于 Bag- ging 的集成 BP 神经网络，构建了报告结果分布的预测模型。我们可以根据未来日期和该日期的解词来预测报告结果的分布。简单地说，这个模型同时考虑了时间序列和单词属性，并帮助你在你决定了谜题后的某一天预测玩家对某个单词的猜测分布。

然后，根据报道结果的分布，我们基于 K-Means 算法对单词难度进行分类，并通过 Pearson 相关系数探索单词属性与难度之间的关联。这样，我们就可以基于难度分类模型，从新给定单词的属性中计算出解词的难度。

最后，我们列出了以下结果，可能会给大家提供一些参考:

2023 年 3 月 1 日上报结果个数的间隔为[7654,20154]。2023 年 3 月 1 日 EERIR 结果分布为(0,1,6,25,31,25,13)。EERIE 单词的难度为“Master”(最难)。

感谢您在百忙之中抽出时间来阅读我的来信。希望我们的建议能有所帮助。

你诚挚的，2311717 团队