Problem Chosen	2023	Team Control Number
C	MCM/ICM	2309397

Summary Sheet

揭开隐藏在世界结果中的秘密

摘要：自从《世界大战》成为一款流行的益智游戏以来，它积累了大量的数据。在本文中，我们定义了一系列的指标，并建立了几个模型来探索世界结果中的隐藏信息。

首先，在对给定数据进行预处理并分析报告结果数量的时间序列图后，我们发现变化可以分为 3 个阶段。为了预测报道结果的数量，我们开发了一个基于 ARIMA 和 BP 神经网络的加权优化模型。然后使用自举方法给出预测区间。我们将此过程封装为基于自举的 ARIMA-BP 区间预测模型。因此，我们最终预测 2023 年 3 月 1 日在 95%置信水平下获得的区间预测值约为(19504.74,20383.26)。

然后，我们定义了单词的 3 个定性属性和 4 个定量属性，并使用它们与难度模式玩家的百分比构建了一个多元线性回归模型。我们发现，当首字母从元音变为辅音时，该比例会平均下降 0.618，而单词内部距离每增加一个单位，该比例会平均增加 0.017。

之后，我们基于 LSTM 模型对报告结果进行百分比分布预测。为了确保百分比在 100%左右，我们首先使用球坐标变换对分量数据进行处理。然后我们使用它们作为输出变量，将 7 个单词属性和结果数量作为输入来训练我们的 LSTM 模型。在此基础上对 EERIE 的预测为[2%，11%，25%，24%，19%，14%，5%]。我们改变了模型的参数，并加入了噪声来做灵敏度分析。同时，我们引入 COV 来测量模型预测的不确定性，发现它在 0.4 左右。对于误差分析，我们使用 MSE、RMSE 和 R2 to 来衡量预测精度，它们的值如表 7 所示。

我们提取了 6 个指标:RDC、TE、SK、NFC、NON 和 HL 来衡量单词的难度。我们基于这些指标构建了一个 GMM 聚类模型，从而划分了 5 个难度等级。我们将 EERIE 这个词归类为难度级别 III。

此外，通过统计每个字母在五个位置的出现频率，我们发现 S 作为首字母出现频率最高，更具体的统计结果如表 9 所示。我们还使用了基于 Apriori 算法的关联规则模型来挖掘世界中的单词组合模式。理想情况下，我们发现字母 A,S,E 和 F,T,L 通常在世界语言中一起出现。

最后，我们对模型进行了评估和完善，并在给《纽约时报》the Puzzle Editor 的信中报告了这些发现。

关键词： ARIMA-BP，LSTM，GMM，先验算法，Word属性

目录

1 介绍

1.1 背景

世界大战是乔什·沃德尔在疫情期间发明的一款在线字谜游戏。以发行游戏而闻名的《纽约时报》于 2021 年 2 月收购了世界[1]。《世界》每天只允许玩一个游戏，世界上的每个玩家每天在 6 次或更少的时间内猜出同一个 5 个字母的单词[2]。玩家可以在常规模式或困难模式下进行游戏。他们可以通过 Twitter 分享自己的分数，从而吸引更多的人来玩和分享。

早在 2021 年 10 月，其网页的注册访问量不到 5000 次，而到 2022 年 1 月，流量已飙升至 4500 多万。我们中的一些人也喜欢这个游戏，图 1 显示了我们得到的一个结果。绿色的贴图表示秘解单词的字母在精确的位置。黄色瓦片表示答案有字母，但没有在正确的位置。灰色瓷砖表示这些字母根本不包含在解决方案中[3]。

图 1:世界游戏

现在我们有了一个从 2022 年 1 月 7 日到 2022 年 12 月 31 日的每日结果文件。这个文件包含了 12 个关键变量，这些变量在我们后期的研究中至关重要。在该文件中，由于四舍五入的原因，七次尝试的人数百分比之和可能不会达到 100%。

1.2 问题重述

通过分析上述背景，我们总结出需要解决的任务如下:

·任务 1:建立一个模型来解释在 Twitter 上报告分数的总人数的每日变化，并用它来给出 2023 年 3 月 1 日总人数的预测区间。

·任务 2:确定单词的属性是否会影响选择困难模式的玩家比例，并据此解释所获得的结果。

·任务 3:如果给出一个未来的日期和特定的单词，建立一个模型来预测这一天 1-X 次尝试的百分比。之后，以 2023 年 3 月 1 日的 EERIE 一词作为模型预测的具体例子，同时分析模型的不确定性和预测的准确性。

·任务 4:开发一个模型，对单词的难度进行分类，并在每个类别下识别单词的属性。用这个模型来确定 EERIE 这个词有多难?最后，讨论分类模型的准确性。

·任务 5:列出并描述数据集的其他一些有趣的特征。

1.3 我们的工作

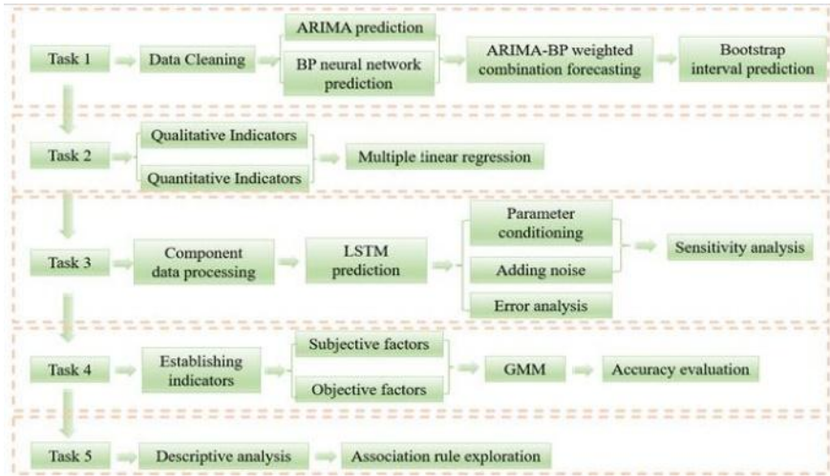


图 2:我们的工作

2 假设和注释

假设 1:数据集中每日在线游戏玩家的数量是一个时间依赖的系列集合，不受季节变化的影响。

原因:因此，我们可以使用 ARIMA 和 LSTM 模型来预测 2023 年 3 月 1 日的一系列数据。

假设 2:玩家每天在推特上报告的比分是正常且可靠的。

原因:为了确保我们基于数据集建立的模型能够可靠地预测 2023 年 3 月 1 日的数据范围。

假设 3:假设《世界大战》的开发过程符合游戏的生命周期理论。

原因:这种假设减少了外部不确定性对世界大战游戏预测的影响，从而使整个预测和分析过程更加高效。

在这项工作中，我们在模型构建中使用表 1 中的符号。其他不常用的符号一旦使用，将被引入。

表 1:符号

Symbols	Definition
w_{A}	Weights of the ARIMA model
w_{B}	Weights of the BP model
P_{Ai}	Predicted value of ARIMA model
P_{Bi}	Predicted value of BP model
L_{i}	Distance between letters
NF	Noise factor
$\mathrm{C_i}$	the ratio of players at different guess counts

3 基于 Bootstrap 的 ARIMA-BP 区间预测模型

3.1 数据预处理

通过审查给定的数据，我们发现没有缺失值，但有五个异常值。其中一个不存在，其中两个少于 5 个字母，由于难以获得这些单词的真实值，我们删除了这三行数据。剩下的两个异常值是由于数据输入错误造成的。为了避免数据量的过度减少，我们通过结合之前和之后的数据以及它们的语义来改变它们。另外，原始数据中各百分比之和都在[98%，102%]，与 100%相差不大，所以是合理的，不需要处理。综上所述，原始数据的预处理情况汇总在表 2 中。

表 2:数据预处理

Outliers	Variable Type	Contest number	Preprocessing method
marxh	Word	473	Delete an entire row
clen	Word	525	Delete an entire row
tash	Word	314	Delete an entire row
2569	Number of reported results	529	Change to 22569
rprobe	Word	545	Change to probe

3.2 基于 ARIMA-BP 联合模型的点预测

3.2.1 变异解释

在数据预处理中，虽然存在输入词错误，但其对应的报告结果数不受影响，因此在本问题中分析完整的报告数据。通过观察和分析 2022 年 1 月 7 日至 2022 年 12 月 31 日报告结果数量变化的特点，我们发现它们可以分为 3 个阶段，如图 3 所示。

基于游戏生命周期和玩家生命周期理论[4]，并结合《世界大战》的游戏特点，我们将这些变化的原因解释如下。

第一阶段:快速成长期(2022 年 1 月 7 日- 2 月 2 日)

自从《世界大战》的网页版推出以来，它每天只更新一个谜题。这种人为的稀缺性增强了玩家对挑战的渴望和期待。世界大战的分享功能使用表情块来指代游戏结果，辨识度很高，易于传播，同时也避免了剧透，从而吸引更多的新玩家来玩世界大战。此外，传统流行的填字游戏和易于玩的游戏机制也为世界的进一步流行做出了贡献。在此期间，世界大战报告的结果总数增长了 348.85%，在 2 月 2 日达到了 361,908 个的峰值。

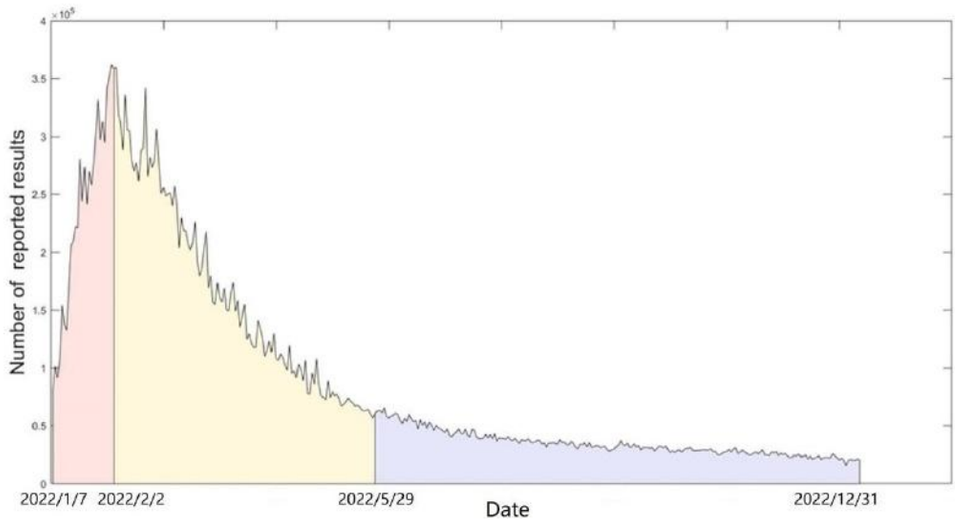


图 3:报告结果数量的变化

第二阶段:快速衰退期(2022 年 2 月 3 日- 2022 年 5 月 29 日)

在此期间，世界在报告结果的数量上总体下降了 84.29%，这是互联网时尚的普遍特征。当游戏的受欢迎程度达到顶峰时，由于玩家对游戏的兴趣下降和厌倦，它面临着玩家的大幅流失。《世界大战》游戏机制的奇异性也促成了这一点。此外，市场上竞争游戏的出现，如“Words with Friends”和世界的盗版游戏，使世界进一步失去了玩家。

Phase3:稳定减持期(2022 年 5 月 30 日- 2022 年 12 月 31 日)

在这一阶段，世界报告的结果总数减少了 64.14%。在第二阶段迅速流失的玩家通常会去玩世界，因为它在普通大众中很受欢迎，当世界不再热门或下一个趋势出现时，他们会很快离开。剩下的玩家通常是忠实的《世界大战》玩家，或者因为游戏的社交圈而产生用户粘性的玩家。此时，举报数量仍会下降，但降幅不会太大。

3.2.2 模型选择的原因

预测报告结果的数量是时间序列分析领域的一个问题。传统的 ARIMA 模型能够从历史数据中提取确定性信息，以预测变量随时间的变化趋势，但该模型在应用前需要进行大量的测试，并且在确定模型顺序的过程中存在一定的主观性[5]。近年来，随着人工智能的发展，机器学习算法被广泛应用于数据分类和预测领域。其中，BP 神经网络模型高效、便捷，能有效打破传统时间序列预测模型的局限性。考虑到这些因素，我们决定将传统时间序列模型和机器学习模型的优点结合起来。这意味着我们首先使用 ARIMA 模型和 BP 神经网络模型分别预测报告结果的数量，然后使用加权平均方法将两个模型结合起来，得到一个更可靠的 ARIMA-BP 联合预测模型。

3.2.3 ARIMA 预测

我们首先使用传统的 ARIMA 模型对报告结果的个数进行点预测，其过程如图 4 所示:

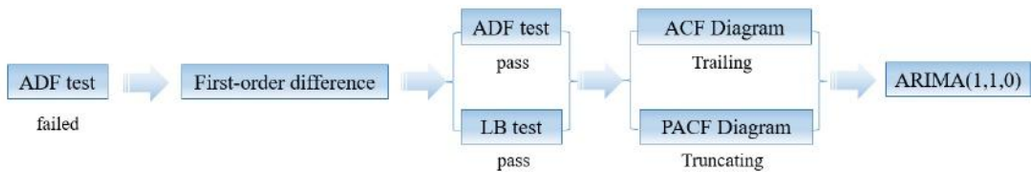


图 4:ARIMA 模型构建过程

首先，我们对预处理数据集中的变量“报告结果的数量”进行 ADF 平滑性测试，发现它并不平滑。因此，我们将该序列差分到一阶，并再次对差分序列进行测试。之后，我们观察差分序列的 ACF 和 PACF 图，发现 ACF 图表现为尾尾特征，PACF 图表现为一阶截尾特征。因此，可以对原系列开发如下 ARIMA(1,1,0)模型:

$$\begin{cases}(1-\phi_1 B)(1-B)x_t=\varepsilon_t \\ E(\varepsilon_t)=0, Var(\varepsilon_t)=\sigma_\varepsilon{}^2, E(\varepsilon_t\varepsilon_s)=0, s\neq t, \\ E(x_s\varepsilon_t)=0, \forall s<t\end{cases}$$

式中，xt is 为 tth time 序列值，B 为延迟算子，为移动自平均多项式的系数。

3.2.4 BP 神经网络预测

BP 神经网络是一种根据误差反向传播算法训练的多层前馈神经网络，它可以根据输入输出变量之间的关系，不断地迭代和修复自身的权值，从而最终估计出精确的函数关系[6]。它的算法用下面的伪代码来解释:

Algorithm 1: Prediction of the Number of Reported Results

Input: Training set D

Output: test set E

Learning rate η; Data normalization (normalization formula)

Create network; Train the network

repeat for D

Forward propagation; Backward propagation

until for reaches end condition

Using the network; Inverse normalization of data

BP neural network prediction with test set E

end: The completed BP neural network trained

3.2.5 结果分析

使用 ARIMA(1,1,0)模型预测得到的结果如图 5 所示，使用 BP 神经网络模型预测得到的结果如图 6 所示:

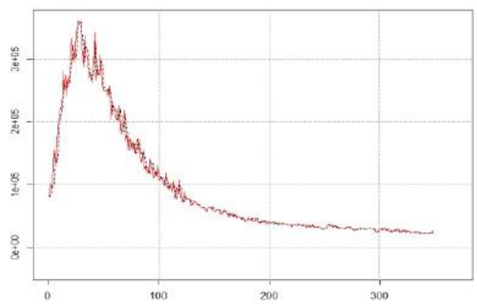


图 5:ARIMA 预测效果

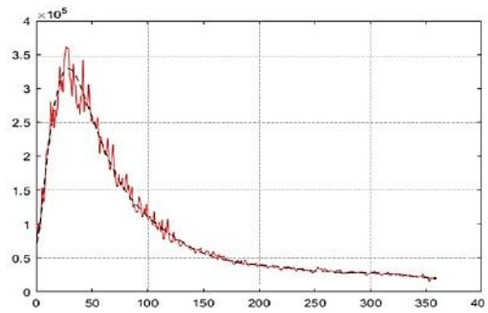


图 6:BP 神经网络预测效果

可以看出，ARIMA 模型具有较好的预测效果，使用该模型对 2023 年 3 月 1 日报告的结果个数得到的点预测 PA 约为 20598.84。BP 神经网络模型在前期存在一定的偏差，但在后期具有较好的预测效果。该模型在 2023 年 3 月 1 日得到的点预测值 B 约为 19496.05。

3.2.6 加权投资组合模型预测

为了提高模型预测的精度和鲁棒性，在 ARIMA 和 BP 神经网络模型的基础上，利用预测误差对两种模型进行加权，推导出一种新的 ARIMA-BP 联合预测模型。这种加权组合预测模型的设计思路如图 7 所示:

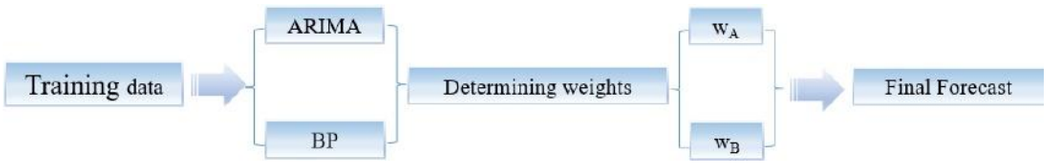


图 7:组合模型设计

我们首先计算两种预测模型各自的平均预测误差:

$$\overline{PEA^2} = \frac{1}{N}\sum_{i=1}^{N}(R_i - P_{Ai})^2,\ \overline{PEB^2} = \frac{1}{N}\sum_{i=1}^{N}(R_i - P_{Bi})^2.$$

之后，我们使用任意一个模型的平均预测误差平方在两个模型的总平均预测误差平方中的份额来反映另一个模型所占的权重，即。

$$w_A = \frac{\overline{PEB^2}}{\overline{PEA^2} + \overline{PEB^2}} \approx 0.406,\ w_B = \frac{\overline{PEA^2}}{\overline{PEA^2} + \overline{PEB^2}} \approx 0.594.$$

因此，最终得到的 ARIMAth -BP 组合模型中 i [th]序列的点预测值计算为:

$$P_i = w_A P_{Ai} + w_B P_{Bi}. \quad (1)$$

通过对上述预测误差的分析，可以明显看出 ARIMA 模型的整体预测误差较高，而 BP 神经网络模型可能存在过拟合的情况。因此，联合 ARIMA-BP 预测模型可以修正这两种模型的误差，从而使预测结果更加真实。

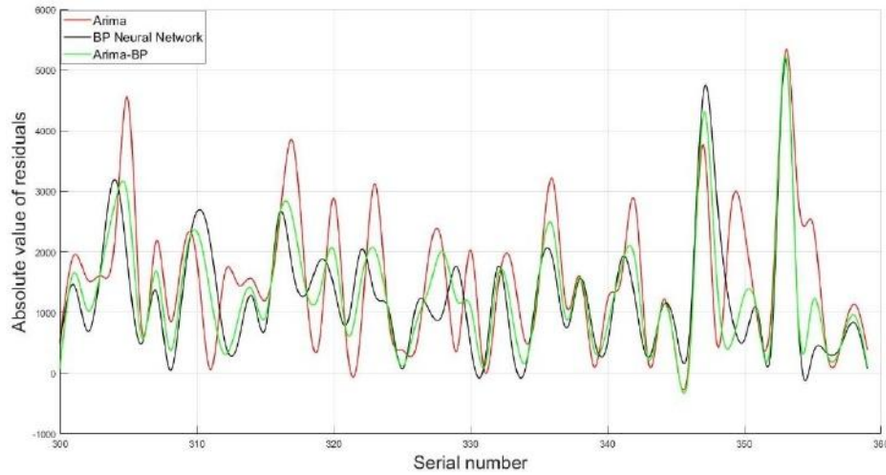


图 8:三种模型的预测误差

使用该组合 ARIMA-BP 模型对 2023 年 3 月 1 日报告的结果数量的最终点预测约为 19944。

3.3 基于 Bootstrap 方法的区间预测

Bootstrap 方法是统计学中用于区间预测的一种重要方法，它首先假设数据集服从未知分布，然后通过对给定数据集重复采样来估计样本的分布区间[7]。我们将在加权组合模型下预测的未来 60 个周期的数据作为样本集，之后对其进行放回采样 1000 次，得到 1000 Bootstrap 样本集 YB=(Y1,Y2，....1000..，Y)。

对于每个子样本，其分布与样本的分布一致，因此我们计算这 1000 个子样本集的标准差(SD)作为样本的 SD:

$$\sqrt{Var(y)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(yi - \bar{y}\right)^2}.$$

根据中心极限定理可知，进行 1000 次抽取得到的样本集近似服从正态分布。因此，可以借助 z 统计量计算出样本在置信水平 1-α下的置信区间的上下限:

$$y = \hat{y} \pm Z_{\alpha/2} \times \sqrt{Var(y)}. \quad (2)$$

在 95%、90%和 80%的置信水平下，2023 年 3 月 1 日报告结果数量的最终预测区间如表 3 所示。

表 3:预测区间

Confidence level	Lower limit	Upper limit
95%	19504.74	20383.26
90%	19575.33	20312.67
80%	19656.69	20231.31

从统计的角度来看，2023 年 3 月 1 日在 Twitter 上报告 worlddle 成绩的人数将大致在(19,504.74,20,383.26)的范围内，与目前的情况相比会有所减少。

4 探索 Word 属性对硬模式的影响

基于多元线性回归模型

探索单词的某些属性对选择困难模式的玩家比例的影响是一个多输入变量、单输出变量的建模问题。我们决定开发一个多元线性回归模型，以得出单词属性是否影响在困难模式下打出的分数百分比及其影响程度。输入变量 X1-X7 的定义如下。

4.1 定义词的定性属性

首先我们对数据集中的所有单词进行了概述，并对图 9 中预处理后的 356 个单词绘制了单词云图。我们发现每个词都是不同的，它的频率是 1，所以它在词云图中的大小是差不多的。

图 9:357 个单词的单词云图

通过搜索相关文献，我们将词属性分为 3 个定性属性和 4 个定量属性，共 7 个词属性。

词性(X1)

词性是一个词最基本的属性，大类有名词、动词和形容词[8]。通过分析这 356 个单词的词性，我们将单词的词性分为名词、动词、形容词和其他四类。

不常见程度(X2)

不常见程度也是词语的定性属性之一。我们使用 COCA 语料库来查找每个单词在口语、小说书籍、杂志和报纸、学术期刊等实际应用中的出现频率。我们以 3000 作为分界点，将单词分为两类:常见和不常见。

首字母是元音还是辅音字母(X3)

以元音或辅音字母开头，也是一个词的定性属性。元音字母是 a, e, i, o, u，这可能会影响玩家使用困难模式的比例。

4.2 定义单词的定量属性

元音字母数量(X4)

元音字母的数量是单词的数量属性之一。通过计算每个单词的这个数字，我们发现它的取值范围是[0,3](该值是一个整数)。

字母重复的最大次数(X5)

一个单词可以由不同的字母组成，也可以包含重复的字母，字母重复的最大次数也可以作为一个单词的定量属性。通过计算每个单词的这个数字，我们发现它的取值范围是[1,3](该值为整数)。

情绪极性值(X6)

每个单词都倾向于暗示某种情绪态度，我们使用 SenticNet 情绪词典获得了 356 个单词的情绪极性值。其中，负极性值代表消极情绪，正极性值暗示积极情绪。

Word 内部距离(X7)

我们将单词的内部距离(WID)定义为其中每个相邻字母的距离之和(参见公式 3)，字母距离是它们在字母表中的位置(L)之间的距离，例如 a 和 C 的字母距离为 2,ABBEY 的内部距离为 24。

$$WID_n = \sum_{i=1}^{4} |L_{i+1} - L_i|, \quad n = 1, 2, 3 \cdots 357 \tag{3}$$

4.3 建立多元线性回归模型(Multiple Linear Regression Model)

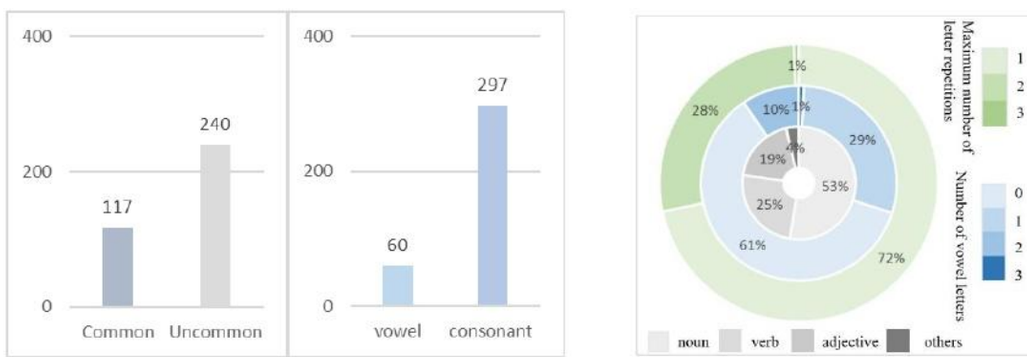
目前我们已经获得了 7 个自变量的数据，但由于 3 个定性数据:X1-X3 是分类变量，所以我们需要在将它们代入模型之前给它们赋值。我们给每个类别赋值如下。

$$X1 = \begin{cases} 1, noun \\ 2, verb \\ 3, adjective \\ 4, others \end{cases}, \quad X2 = \begin{cases} 1, commom \\ 2, uncommom \end{cases}, \quad X3 = \begin{cases} 1, vowel \\ 2, consonant \end{cases}$$

因变量(Y)是报告的在 Hard Mode 下玩的分数的百分比，可以根据公式(4)计算。

$$Y = \frac{Number\ in\ hard\ mode}{Number\ of\ reported\ results} \times 100\% \tag{4}$$

所识别的所有变量的描述性统计结果如图 10 和表 4 所示。如(a)所示，数据集中的不常见单词多于常见单词，带有辅音字母的单词多于带有元音字母的单词;如(b)所示，数据集中的名词最多，占 53%，只有一个元音字母的词占 61%，和 29%的单词有超过两个最大重复字母。



(a) X2 – X3　　(b) X1, X4 – X5

图 10:变量的描述性统计

如表 4 所示，356 个词中情绪极性值的平均值为 0.0355，为正;词内距离的平均值为 33.3726;报告玩硬模式的人的平均百分比为 7.52%

表 4:变量的描述性统计

Variable	Minimum	Maximum	Average	Standard deviation
X6	-1	0.999	0.0355	0.6448
X7	7	71	33.3726	11.9655
Y	1.17%	13.33%	7.52%	0.0223

然后，可以建立如下的多元线性回归模型。

$$Y = \beta_0 + \sum_{i=1}^{7} \beta_i X_i + \varepsilon$$

代入变量数据，最终拟合的回归模型为:

$$\hat{Y} = 6.855 - 0.007X1 + 0.305X2 - 0.618X3 + 0.136X4 + 0.394X5 - 0.025X6 + 0.017X7$$

4.4 结果分析

模型整体显著性检验结果和各自变量回归参数显著性检验结果整理于表 5。

表 5:模型与参数显著性结果

Variable	t-value (F-value)	P-value	Variable	t-value	P-value
Model	1.747	0.097	X4	0.662	0.509
X1	-0.053	0.958	X5	1.551	0.122
X2	1.210	0.227	X6	-0.138	0.890
X3	-1.798	0.073	X7	1.655	0.099

首先，整个模型的 p 值为 0.097，在 10%水平下显著的显著性，表明模型是显著的。其次，在 7 个自变量中，只有 X3 和 X7 的 p 值在 10%水平上显著，分别为 0.073 和 0.099，说明这两个变量的回归系数显著。其他变量的 p 值均大于 0.1，说明它们的回归系数不显著。

由此可见，单词的首字母是元音还是辅音字母、单词内部距离这两个单词属性对报告玩难模式的玩家比例有显著影响，而词性、不常见程度、元音字母数量、字母最大重复次数、情绪极性值这五个变量对这一比例没有显著影响。在保持其他变量不变的情况下，当首字母由元音变为辅音时，该比例将平均下降 0.618;单词内部距离每增加一个单位，比例平均增加 0.017。

5 基于 LSTM 模型的百分比分布预测

5.1 成分数据的处理

目前，我们知道在 Twitter 上报告分数的玩家中，每天有 1 到 6 次尝试猜中单词或无法解决谜题的百分比，这些百分比加起来大约是 100%。

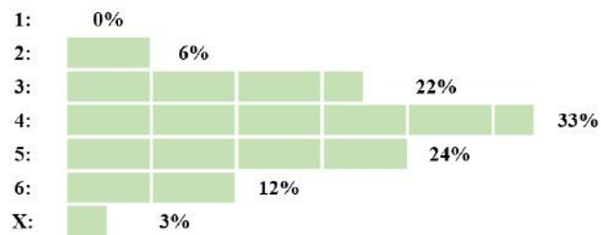


图 11:成功尝试的百分比

可以看出，这些百分比数据共同构成了一组分量数据 X=(X12,X7，.....， X)， X 中七个分量的值近似满足约束:

$$\sum_{i=1}^{7} x_i \approx 1,\ 0 \leqslant x_i < 1. \qquad (5)$$

因此，对未来日期相对于(1,2，…，6,X)的百分比的预测可以认为是对成分数据的预测问题。为了满足式(5)的限制，我们没有直接使用原始数据集作为预测的输出变量，而是首先对成分数据进行了球坐标变换。

$$\begin{cases}\theta_7 = \arccos y_7 \\ \theta_6 = \arccos\left(\dfrac{y_6}{\sin\theta_7}\right) \\ \theta_5 = \arccos\left(\dfrac{y_5}{\sin\theta_7\sin\theta_6}\right) \\ \vdots \\ \theta_2 = \arccos\left(\dfrac{y_2}{\sin\theta_7\sin_6\cdots\sin\theta_3}\right)\end{cases}$$

模型对其进行预测。预测结果得到后，可以通过逆变换得到最终的分量数据 x7 1- xx。

5.2 如何预测

对于玩家在指定的未来日期内尝试次数的百分比预测，我们将其视为多个输入和多个输出之间的信息映射问题。在这个过程中，我们将输出变量视为我们使用球坐标进行转换的百分比分量数据。至于输入变量，我们主要考虑两个方面:球员的数量和单词的属性。从之前的分析中，我们发现玩家的数量会随着时间的推移而变化，这可能会导致玩家尝试的百分比发生相应的变化。另一方面，单词的属性会直接影响玩家猜测的成功率，从而影响百分比的分布。因此，除了每天的玩家数量外，我们还将前面已经定义的 7 个单词的属性作为输入变量。最后，对于预测模型的选择，我们选择了 LSTM 深度学习模型，下面我们将详细描述该模型的应用。

LSTM 模型是一种长短期记忆网络模型，它允许使用时间序列对输入数据进行分析，并利用单元格状态运算的核心思想预测这些输入数据的未来趋势[9]。该模型的工作原理可以如图 12 所示，其中黄色表示算术运算，蓝色表示学习后的神经网络。

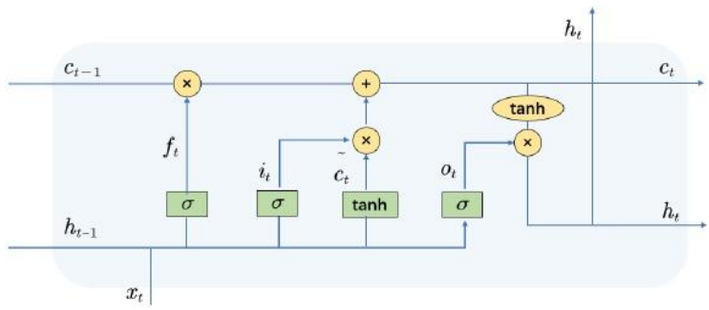


图 12:LSTM 的工作原理

LSTM 模型具有向每个模块删除或添加信息的能力

设置“大门”，总的来说包括四个步骤:

第一步:确定单元格中丢弃的信息:这个操作首先读取当前输入的 xt 和预神经元信息 h_{t-1}，之后遗忘门 f_t 确定丢弃的信息:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

步骤 2:确定存储在细胞状态中的新信息:该操作使用 sigmoid 层作为输入门层来确定更新的值，并通过 tanh 层创建一个新的候选值向量，然后将其添加到状态中:

$$= \sigma(W_i[h_{t-1}, x_t] + b_i),\ \tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

步骤 3:更新单元状态。此操作更新 c_{t-1} to c_t 并将旧状态与 f_t 相乘，同时丢弃需要丢弃的信息，之后确定新的候选值。

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t$$

第四步:确定输出:此操作使用 sigmoid 层来确定输出的单元格状态部分，然后通过 tanh 处理单元格状态，并将其与 sigmoid 层的输出乘以。

$$o_t = \sigma(W_0[h_{t-1}, x_t] + b_0),\ h_t = o_t \cdot \tanh(c_t)$$

之后，我们以单词 EERIE 为例，使用 ARIMA-BP 加权模型预测当天报告比分的玩家数量约为 19944 人，而该单词的属性向量集为(3,2,4,3,1,1,26)。将这些输入属性带入我们的 LSTM 模型，同时对得到的结果进行逆变换，我们可以得到 2023 年 3 月 1 日玩家尝试次数的最终百分比分布:

表 6:结果

1 try	2 tries	3 tries	4 tries	5 tries	6 tries	X
2%	11%	25%	24%	19%	14%	5%

可以看到，在 2023 年 3 月 1 日，对于单词 EERIE，大约 13%的玩家可以在 2 次尝试中猜出单词，大约 49%的玩家可以在 3 到 4 次尝试中猜出单词，而 5%的玩家在要求的尝试次数中没有猜出单词。

5.3 灵敏度分析

我们通过敏感性分析探索了我们的模型和预测的不确定性，敏感性分析是通过两种方法分别对模型进行的:调整模型参数和以正态分布的形式向变量添加噪声。

调整模型参数

在参数调整方面，我们对 LSTM 模型的 batchsize 和 epoch 参数进行了 5 次调整，具体调整值分别为[25,250]、[50,500]、[75,750]、[100,1000]和[125,1500]。我们考虑了 7 分球员的百分比

将每个单词的尝试类型作为预测整体，并使用变异系数(冠状病毒)来测量 7 个百分比值之间的分散程度。冠状病毒的公式为:

$$COV = \frac{s}{\bar{x}}$$

基于上述 LSTM 模型的预测思想，我们在调整参数后使用该模型对数据进行重新预测，并计算预测数据的冠状病毒。如果冠状病毒的总体分布与调整后的模型相似，则模型的不确定性较低，反之则较高。随着 batchsize (bs)和 epoch (ep)的变化，356 个单词的冠状病毒散点图如图 13 所示。

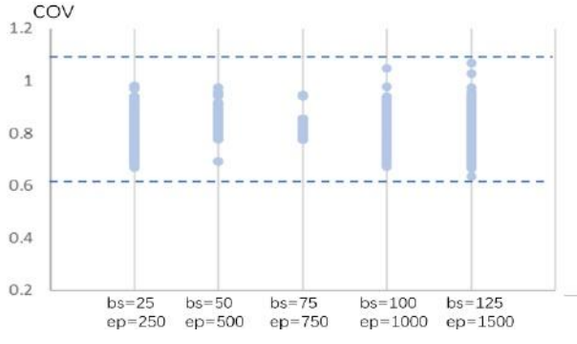


图 13:敏感性和不确定性分析(1)

如图 13 所示，模型参数(bs, ep)的变化对预测数据的冠状病毒分布有影响。各模型参数下冠状病毒的分布基本相似，但 bs=75、ep=750 分布较为集中，说明模型和预测的不确定性较低。

调整噪声

为了添加噪声，我们首先随机生成一个正态分布矩阵 M356×7 ，然后乘以一个噪声系数 NF，并添加到一个矩阵 T7×356 中，该矩阵由原始数据组成，从 1 次尝试到 X 次尝试计算模型的新输出数据集 N。公式表示如下。我们将 NF 值分别调整为 0.1、0.15、0.2、0.25 和 0.3。

$$N = T_{356\times7} + NF \times M_{356\times7}$$

同样，我们在加入噪声后使用模型重新预测数据，并计算预测数据的 cov。随着 NF 的变化，356 个单词的 cov 散点图如图 14 所示。

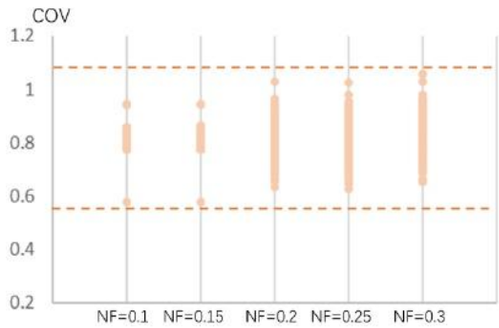


图 14:敏感性和不确定性分析(2)

如图 14 所示，噪声因子(NF)的变化也会影响预测数据的 cov 分布。当 NF 值为 0.2 及以上时，cov 在单词间的分布更为一致，这也表明此时模型和预测的不确定性较低。我们定义了 COV 的整体极端差异(式 6)来表示不确定程度。

$$Uncertainty = \max(COV) - \min(COV) \tag{6}$$

从模型参数的变化来看，模型与预测的不确定性为 0.4321;从噪声系数的变化来看，这种不确定性为 0.4794。

5.4 误差分析

我们用误差分析来衡量模型预测的置信度。预测误差越小，我们对所建立的模型的置信度就越高。通过查阅文献[10]，我们决定使用平均绝对误差(mean absolute error, MAE)、均方根误差(root mean squared error, RMSE)和决定系数(coefficient of determination, R2)来评价模型预测的有效性。

平均绝对误差(MAE)

MAE 是模型的预测值()与真实值()之差绝对值的加权平均值。均值绝对误差不考虑

误差的方向，而只是平均值的误差量级的预测值。因此，这个指标可以避免相互抵消的问题，进而可以更准确地反映实际预测误差的大小。MAE 的取值范围为[0，+∞)，MAE 越小，模型拟合越好。公式如下。

$$MAE = \frac{1}{N}\sum_{n=1}^{357}|y_n - \hat{y}_n|$$

均方根误差(RMSE)

RMSE 是真实值与预测值之差的平方的期望值的平均值的根。它衡量的是预测值与真值之间的偏差，在数量级上更为直观。RMSE 的取值范围为[0，+∞)，RMSE 值越小，模型的预测能力越好。公式如下。

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{357}(y_n - \hat{y}_n)^2}$$

决定系数(R2)

Ris2 用于衡量模型预测与实际值的拟合优度。它的取值范围为[0,1]，接近 1 的值表明模型拟合良好;如果该值接近 0，则表示模型拟合较差。公式如下。

$$R^2 = 1 - \frac{\sum_{n=1}^{357} (y_n - \hat{y}_n)^2}{\sum_{n=1}^{357} (y_n - \bar{y})^2}$$

根据以上 3 个指标的公式，我们分别计算出玩家进行 1 次尝试- X 次尝试的百分比的预测误差，然后求出它们的平均值。计算结果整理在表 7 中。

表 7:误差分析

Indicators	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	X tries	Average
MAE	0.543	3.361	7.168	4.162	5.350	5.443	2.403	4.062
RMSE	0.911	4.414	8.868	5.423	6.590	6.865	4.302	5.339
R^2	0.899	0.986	0.993	0.989	0.986	0.993	0.998	0.978

从表 7 中可以看出，所有预测结果的 MAE 都在 8 以内，MAE 平均值为 4.062,RMSE 平均值为 5.339，均值较小，说明模型的预测误差较小。此外，模型的整体 R2 of 为 0.978，说明模型拟合非常好，模型具有良好的预测效果。因此，我们在很大程度上对我们的模型有信心。

6 基于 GMM 模型的单词难度评估

6.1 单词难易度评价指标描述

单词的难度受其自然属性的影响，也可以通过玩家的尝试次数来反映。因此，在对单词的难度进行分类之前，我们首先定义了一些可以从主观性(S)和客观性(O)两方面衡量单词难度的指标。

计算相对难度系数(S)

每次尝试类型的玩家数量可以从报告结果的数量和后期百分比中得到，为了减少数量减少对时间序列变化的影响，我们使用极端差分归一化对玩家数量(NP)进行预处理:

$$NP_{ij}' = \frac{NP_{ij} - NP_{\min}}{NP_{\max} - NP_{\min}}.$$

我们将一次猜测- x 次猜测的得分定义为[1,7]，得分越高表示单词难度越大。将分数乘以人数，得到每个单词在不同猜测下的总分(TS)。然后基于 AHP 对不同猜测次数对单词难度的重要性进行赋值，通过特征值法得到的权重 Wi 分别为[0.0285,0.0428,0.0662,0.1016,0.1558,0.2404,0.3648]。一致性指标 C.R 为 0.0176<0.1，说明基于判断矩阵权重计算的权重是有效的。由此我们定义单词的相对难度系数(RDC)如下。

$$RDC = W_i \times TS \tag{7}$$

尝试期望(S)

我们将玩家在不同猜测次数下的比例(Ci)视为该计数出现的概率(Pi)，并定义尝试期望(TE)，即每个单词猜测的平均次数，如下所示。更难的单词 TE 更大。

$$TE = \sum_{i=1}^{7} P_i \times C_i \tag{8}$$

偏度系数(S)

偏度系数(SK)可以反映数据的分布。对于尝试次数的数据，当 SK 小于 0 时单词更难，而数据是左偏分布的，反之亦然。其计算公式如下。

$$SK = \frac{n\sum (x_i - \overline{x})^3}{(n-1)(n-2)s^3}$$

词频系数(O)

基于 COCA 语料库，我们统计了 Wordle 中超过 350 个单词的词频，然后将频率最高的单词的词频系数定义为 1，频率最低的单词的词频系数定义为 0，频率中间的单词的词频系数定义为 0.5。这样，所有的词频都可以映射到区间[0,1]上。下区间词频系数的转换方程为:

$$\underline{FC} = 0 + \frac{FRE - FRE_{\min}}{FRE_{mid} - FRE_{\min}} \times 0.5 .$$

上区间词频系数的转换方程为:

$$\overline{FC} = 0.5 + \frac{FRE - FRE_{mid}}{FRE_{\max} - FRE_{\mathrm{m}id}} \times 0.5 .$$

正字法邻接数(O)

正字法邻词定义为长度相同但相差一个字母的单词(如 scare 和 stare)[11]。我们从“MCWord:一个正字法字型数据库”中收集每个单词的数据。一个单词的正字法邻居越多，玩家定位这个单词的难度就越大，难度也就越大。

和谐度(O)

我们从世界玩家认知能力的角度，用和谐度来表示一个词的读与写的和谐程度。首先，我们用 H 来表示单词长度与其音标长度的比值，即。

$$H = \frac{LW}{LT}$$

H 的值越接近 1，说明这个词的谐音水平越好。因此，我们可以计算出 H 与 1 的差值的绝对值。我们定义绝对值最小对应的词的和谐度为 1，绝对值最大对应的词的和谐度为 0。这样，所有单词的和声级别也可以映射到区间[0,1]上。

6.2 建立 GMM 聚类模型

在确定了表示单词难易程度的指标后，我们选择 GMM 模型对这些指标值进行聚类，从而对单词的难易程度进行分类。与传统的 K-means 聚类相比，GMM 聚类属于混合聚类，它利用 M '距离来度量不同变量之间的约束关系，可以认为是一种更好的聚类方法。

GMM 算法假设样本服从 k 个混合高斯分布函数，对于每个样本点，其属于第 k 个多元高斯分布的概率可以计算如下:

$$p(x^{(i)}|z^{(i)} = j) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}} e^{-\frac{1}{2}(x^{(i)} - \mu_j)\sum^{-1}(x^{(i)} - \mu_j)^T}$$

之后，可以通过 EM 算法更新各多元高斯分布函数的参数:

$$\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}},\ \sum j = \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_i)(x^{(i)} - \mu_i)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

根据上述过程，继续迭代循环，直到高斯收敛，循环停止。然后，我们可以通过计算得到的高斯参数对样本进行分类。

6.3 聚类结果分析

基于 GMM 模型的聚类结果如表 8 所示。由于相对难度系数(relative difficulty coefficient, RDC)的评价标准在这 6 个指标中较为全面，它是通过提炼玩家在不同猜测情况下的比例特征而得出的权重分数。相对于其他指标，它对单词难度的判断更具有决定性。因此，我们主要按照每个类别的 RDC 大小来划分难度等级，划分结果如下:

表 8:聚类结果

Difficulty Level	TE	SK	RDC	WFC	NON	HL
I	4.1548	0.6371	0.4158	0.351	0	0.3324
II	4.1920	0.4875	0.4345	0.3684	3.7589	0.3705
III	4.1866	0.5483	0.4656	0.3571	1.3984	0.3496
IIII	4.0888	0.1631	0.5786	0.4379	11.3750	0.4167
IIIII	4.2992	0.2261	0.5941	0.3782	6.8958	0.3524

难度等级 1 的单词相对于其他类别的单词是最简单的，其 RDC 最低;偏度系数(SK)最高，呈右偏态分布，这表明更高比例的玩家用更少的尝试猜对了这一类别的单词;正字法邻居(NON)的数量为 0，表明玩家能够很容易地定位这些单词。

难度等级 IIIII 的 ROC 最高，为 0.5941;尝试期望也是最高的，玩家猜出这个单词的平均次数类别为 4.2992。其余难度类别的具体属性值见表 8。

根据 6 个指标的定义和公式，得到 EERIE 的 6 个属性值，如图 15 所示。通过计算该词的 6 个指标值与分类的 5 个难度等级对应的指标之间的马克思主义距离，我们发现该词与难度等级 III 的距离最接近，因此我们认为 EERIE 属于难度等级 III。

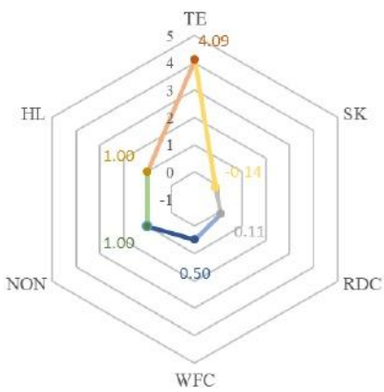


图 15:EERIE 的六个属性值

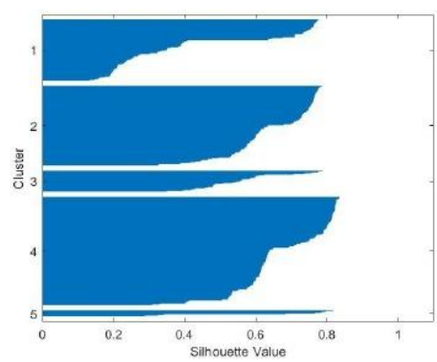


图 16:模型的精度

如图 16 所示，廓形值可以反映 GMM 模型聚类的准确性。廓形值越接近 1，模型的精度越高。结果表明，我们所汇总的 5 个类别的剪影值都在 0.8 的附着范围内，这表明我们使用 GMM 模型获得的聚类效果具有较高的准确性。

7 数据集其他值得探索的有趣特征

7.1 5 个字母位置词频统计

我们统计了 356 个单词中每个单词在 5 个字母位置出现的字母频率，表 9 展示了 5 个字母位置出现频率最高的 8 个字母及其频率。我们发现“s”作为首字母出现的频率最高，“o”作为第 2 和第 3 个字母出现的频率最高，“e”作为第 4 和第 5 个字母出现的频率最高。

表 9:5 个字母位置中出现频率最高的 8 个字母

First Letter	Frequency	Second Letter	Frequency	Third Letter	Frequency	Fourth Letter	Frequency	Fifth Letter	Frequency
s	51	o	51	o	50	e	43	e	73
c	32	a	47	i	44	a	30	y	49
t	29	l	41	a	42	r	29	t	46
a	28	r	36	e	29	t	27	r	35
p	23	h	33	u	28	n	26	l	24
f	22	e	28	n	23	l	24	d	20
b	20	i	23	r	19	i	23	h	18
m	19	n	17	p	14	c	19	k	16

7.2 单词之间的组合模式探索

一般来说，组成一个单词的字母并不是相互独立的，但是有一些来回的联系。这些具有来回关联的字母可以形成一个资源池，玩家可以从中选择字母进行猜测，帮助提高正确率。因此，从这个角度出发，我们的目标是探索单个单词之间的组合模式。

为了减少讨论的复杂性，我们假设不考虑单词出现的顺序，而只探索不出现重复字母的 255 个单词的组合模式。我们在关联规则模型的帮助下解决了这个问题，同时使用以下三个定义来确定字母之间的关联强度。

Sup:表示项目集 X 在整个数据 N 中出现的概率，即:

$$Sup(X) = \frac{sum(X)}{N}$$

Conf:表示一个变量 X 以另一个变量 Y 也发生变化为条件而发生变化的概率，即:

$$Conf(X \to Y) = \frac{Sup(X \cup Y)}{Sup(X)}$$

Lift:它可以反映一个变量 X 的变化概率对另一个变量 y 的变化的提升程度。如果它的绝对值大于 1，则意味着这两个对象之间确实存在相关性。

$$Lift(X \to Y) = \frac{Sup(XY)}{Sup(X)Sup(Y)} = \frac{Conf(X \to Y)}{Sup(Y)}$$

在使用关联规则模型探索单词之间的组合规律时，我们首先对单词进行字母拆分，并将拆分后的单词转换为 Bool 值为 0 或 1 的形式。之后，我们使用高效精简的 Apriori 算法对关联规则进行分析，指定 Sup 为 0.5，并在 Lift 大于 1 时过滤掉字母之间的关联规律和置信度值。

表 10:单词组合模式

Letters that appear	Letters that may appear together	Conf	Letters that appear	Letters that may appear together	Conf
a / k / s	e	1	e / n / r	i	1
l / p / t	a	1	p / r / t	e	1
b / e / n	i	1	g / h / t	i	1
c / r / t	e	1	i / m / s	t	1
d / p / t	e	1	d / f	e / t	0.6
h / i / s	e	1	f / t	l / o	0.5

使用关联规则模型，我们可以得到许多字母组合之间的规律，但由于页面限制，我们只展示了12种可能的情况，根据这些情况，玩家可以在某种程度上更有效地拼写特定的单词。

8 模型评估和改进

8.1 优势与劣势

在预测2023年3月1日的人数时，我们使用了ARIMA模型和BP神经网络两种模型，并基于两者的残差平方和进行了加权计算，这在一定程度上弥补了两种算法的不足。

我们在之前使用LSTM模型时，首先用球坐标对原始数据进行了变换，使得未来日期约(1、2、3、4、5、6、X)的百分比预测值之和更接近100%。

在考虑LSTM模型的不确定性时，我们加入适当噪声的干扰对模型进行扰动，然后计算它们的变异系数，从而使我们的预测结果更加可信。

在考虑数据集的特征时，我们使用了Apriori算法，通过挖掘数据之间的特征，发现每天给定的单词中某些字母的出现存在一定的相关性。

在我们的探索中，我们发现每天报告的数据集中的玩家总数在减少，但当我们受到空间限制，无法对其进行文字描述而没有深入探索时，选择困难模式的玩家比例在增加。

8.2 模型改进

在后续的研究中，在使用GMM混合模型对数据进行聚类时，我们可以深入考虑到每个指标的重要性，这样我们就可以给每个指标添加一定的权重。

我们只使用了关联规则算法来寻找一些字母组合的相关性，但在现实中，单词的组成比我们想象的要复杂得多(比如每个字母的不同位置)，所以我们也应该开始考虑建立一个更全面的模型来描述它们。

9 结论

在这篇文章中，我们使用了不同的模型来分析Wordle结果的信息。总的来说，我们得到了以下结论:

与2022年相比，在95%的置信水平下，报告结果的玩家数量将大致在(19,504.74,20,383.26)的范围内。

单词的首字母是元音还是辅音字母以及单词内部距离都会影响报告玩硬模式的玩家比例，当前者从元音变为辅音时，这一比例将平均下降0.618，而单词内部距离每增加一个单位，这一比例将平均增加0.017。

2023/3/1，对于EERIE这个词，大约49%的玩家在3到4次猜测中猜对了，这个结果的不确定性约为0.4，同时平均预测误差约为5.339。此外，我们发现EERIE属于难度等级III，为中等难度。

10 给《纽约时报》拼图编辑的信

2023年2月20日

尊敬的先生/女士:

我们给您写这封信是因为我们想向您报告我们对2022年1月31日至2022年12月31日世界游戏数据的基本发现。希望这些结果对大家有所帮助。

首先，我们发现推特上报道结果数量的变化大致可以分为 3 个阶段:快速增长期(2012.1.07 - 2012.2.07)。2022 年 2 月 2 日)、快速下降期(2022 年 2 月 3 日- 2022 年 5 月 29 日)和稳定减少期(2022 年 5 月 30 日- 12 月 29 日)31, 2022)。通过建立加权 ARIMA-BP 模型，我们预测 2023 年 3 月 1 日的报告结果数为 19,944，并借助于 Bootstrap 方法，我们得到 95%的预测区间为(19,504.74,20,383.26)。

其次，我们定义了世界大战中解词的单词属性，并将其分为 3 个定性属性和 4 个定量属性，通过建立这 7 个属性变量与难度模式玩家百分比的多元线性回归模型，我们发现单词的首字母是元音还是辅音字母以及单词内部距离对比例有显著影响。

然后，基于这 7 个定义的单词属性和报告结果的数量，我们通过 LSTM 模型建立了它们与结果分布的关系，并预测 2023 年 3 月 1 EERIE 这个单词的分布为[2%，11%，25%，24%，19%，14%，5%]，对应于 1 到 7 个以上的猜测次数。这种预测方法也可以应用于其他解词。

同样，基于玩家对每个单词的猜测次数和单词属性的数据，我们定义了三个主观指标:相对难度系数、尝试期望、偏度系数，以及三个客观指标:词频系数、正字法邻接数、和声水平，共同衡量单词的难度。利用这 6 个指标，我们用 GMM 聚类模型将世界语言中的单词划分为 5 个难度等级。以单词 EERIE 为例，我们根据其 6 个指标的数据将其划分为难度等级 III。

最后，我们统计了该词在数据集中的 5 个位置中每个字母出现的频率，发现解词往往以“s”作为首字母，第 2 和第 3 个位置出现频率最高的是字母“o”，第 4 和最后一个位置出现频率最高的是“e”。此外，我们利用关联规则算法对世界谜题中的字母间连接进行了探索，得到了“a”、“k”、“s”同时出现时，“e”也倾向出现的关联规则。

感谢大家在百忙之中抽出时间来阅读这个字母。我们希望这能帮助你更好的管理和运营世界。

真诚地,

团队# 2309397

References

[1] Match, S. (2022). The New York Times buys Wordle. The New York Times.

[2] Anderson, B. J., & Meyer, J. G. (2022). Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning. arXiv preprint arXiv:2202.00557.

[3] Short, M. B. (2022). Winning Wordle Wisely—or How to Ruin a Fun Little Internet Game with Math. The Mathematical Intelligencer, 44(3), 227-237.

[4] Kummer, L. B. M., Nievola, J. C., & Paraiso, E. C. (2017). Digital Game Usage Lifecycle: a systematic literature review. In Brazilian Symposium on Computer Games and Digital Entertainment (SBGames) (pp. 1163-1172).

[5] Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS 16th international conference on computer modelling and simulation (pp. 106-112). IEEE.

[6] Buscema, M. (1998). Back propagation neural networks. Substance use & misuse, 33(2), 233-270.

[7] Masarotto G. Bootstrap prediction in tervals for auto regressions [J]. International Journal of Forecasting. 1990,6(2),229-239.

[8] Kornfilt, J. (2020). Parts of speech, lexical categories, and word classes in morphology. In Oxford Research Encyclopedia of Linguistics.

[9] S. Siami-Namini, N. Tavakoli and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 3285-3292.

[10] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623.

[11] Li, I. (2022). Analyzing difficulty of Wordle using linguistic characteristics to determine average success of Twitter players.