Problem Chosen	2023	Team Control Number
C	MCM/ICM	2301192
	Summary Sheet	

破解世界大战:词属性分析方法

摘要：在过去的 600 天里，一款名为“世界”的五字母益智游戏在 Twitter 上掀起了一阵热潮。世界大战玩家的得分报告对管理者来说至关重要，因为它们为评估游戏难度、预测玩家数量和及时做出调整提供了有价值的信息。为了更好地分析报告并提供游戏改进建议，我们从多个角度和层面对这一主题进行了深入而密切的研究。

首先，为了解释世界大战报告数量的变化并做出预测，我们将玩世界大战与传染病的传播进行了类比。我们将玩世界大战与感染进行比较，将玩家与被感染的个体进行比较，将长期不玩世界游戏的个体与易感个体进行比较，将厌倦游戏的个体与康复的个体进行比较，将 Twitter 上的分享与传播进行比较，将退出游戏与康复进行比较。基于这些假设，我们使用 SIRS 模型来拟合曲线并解释总体趋势。我们还使用 Prophet 模型插入断点来解释数据振荡，并为未来数据提供预测区间。模型评价结果表明，我们的模型具有较高的可解释性和准确性。

接下来，我们从包含大量语料库信息的词数据库中提取各种词属性，并使用多元线性回归来研究词属性与 Hard-Mode 分数之间是否存在关系。然后，我们基于 f 统计量对模型的显著性进行检验。结果显示，这两个因素之间没有显著的相关性。

此外，我们基于之前提取的词属性构建 BP 神经网络模型来预测猜词数的分布。评价结果表明，该模型具有较高的预测精度和效率，为下一步的分析奠定了坚实的基础。

进一步，我们使用 k - means ++聚类算法将单词分为简单、中等和困难三类。我们分析单词属性和难度之间的关系，按难度对解词进行分类。我们发现，一个词的难度与该词中不同字母的数量、字母频率的总和以及该词在不同领域的使用广度密切相关，但没有明显的证据表明难度与词频之间存在相关性。结合之前的 BP 神经网络模型，可以对单词进行准确的分类。

此外，我们发现“木乃伊”、“手表”等常识词的猜测难度较高，单词的首字母与其猜测难度也存在一定的相关性。

最后，我们向《纽约时报》的编辑提供预测数据和改进建议，以帮助他们改进《世界大战，提升游戏的吸引力。

关键词:先知;SIRS;多元线性回归;BP 神经网络;k - means + +

公众号·数学建模老哥

目录

公众号：数学建模老哥

公众号：数学建模老哥

1 介绍

1.1 背景

最近，推特掀起了一股分享《世界》报告的风潮。过去的解谜游戏开发者往往不太清楚自己游戏面向大众的难度。难度太大的游戏会让人受挫，而太简单的游戏又会让人觉得无聊。随着信息技术的发展，利用大数据分析来控制谜题难度，成为让谜题变得更有趣的关键。《纽约时报》的世界游戏收集了玩家尝试次数和推特上报道次数的统计数据。这些数据可以用来评估玩家数量和特定单词的难度，保持玩家的热情，让游戏更具吸引力。

1.2 重述问题

《纽约时报》收集了 359 天的《世界大战》玩家得分报告，包括报告时间、次数、困难模式报告的百分比和尝试次数。为了控制玩法和估计玩家数量，需要分析报告次数的趋势，挖掘单词属性中包含的信息，测量单词的难度。要实现这些目标，我们需要:

•分析大时间尺度(整体趋势)和小时间尺度(数据突变)报告数量变化的原因。

•收集和挖掘潜在的词属性。

•分析高难度模式报告的百分比是否与词属性相关。

•分析尝试的分布及其与词属性的潜在关系。

•识别单词属性对难度的影响。

•挖掘其他有助于改善世界的信息。

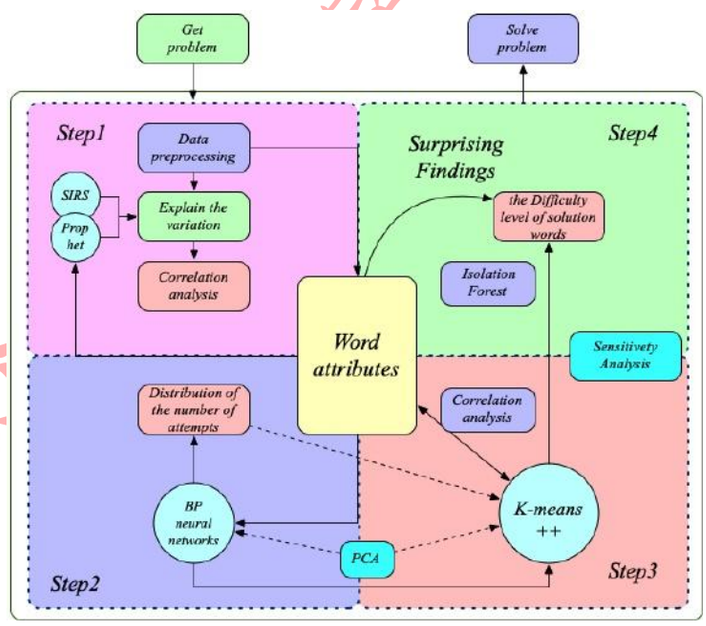

图 1:本文中的流程图

2 假设和符号

2.1 假设

为了简化模型，我们做了几个假设。然而，我们可能需要放宽其中的一些假设，以优化模型并增加其在复杂的现实环境中的适用性。

•Twitter 用户的数量基本上是恒定的，每个用户接收到与世界相关信息的概率是相等的。

•所有世界大战玩家都是 Twitter 用户，所有 Twitter 用户都是潜在的世界大战玩家。

•世界大战中每一天的单词是完全随机的，并从所有五个字母的单词中选择。

•在 Twitter 上报告游戏结果的玩家是所有玩家的随机样本。

•人们可能会厌倦玩《世界》，但他们最终可能会在很长一段时间后想再玩一次。

2.2 符号

表 1:18 词性符号

Symbols	Definition
NN	Noun, singular or mass
JJ	Adjective
RB	Adverb
VBP	Verb, non-3rd person singular present
VBD	Verb, past tense
NNS	Noun, plural
VBN	Verb, past participle
VB	Verb, base form
IN	Preposition or subordinating conjunction
VBZ	Verb, 3rd person singular present
VBG	Verb, gerund or present participle
MD	Modal
PRP	Possessive pronoun
RBR	Adverb, comparative
CC	Coordinating conjunction
JJR	Adjective, comparative
DT	Determiner
JJS	Adjective, superlative

表 2:论文中使用的单词属性注释

Symbols	Definition
Freq	Word Frequency
SLF	the Sum of Letter Frequencies
BU	the Breadth of Usage of a Word
NDLW	the Number of Different Letters in a Word
a-z	the Number of Letters from a to z in a Word

3 模型 1-基于 Prophet 和 SIRS 的解释与预测集成模型

3.1 数据预处理与探索性分析

3.1.1 数据采集与预处理

在解决任务 1 时，分析与问题相关的词的属性并收集相关数据是必不可少的。可能的因素包括频率、不同领域使用的广度、单词中不同字母的数量和词性。在一般的自然语言处理(NLP)中，有 36 种常用词类[2]，我们从中选择了 18 种与本任务相关的类型，如表 1 所示。

为了处理原始数据集中的缺失值、异常值和重复观测值，我们采用了一系列数据处理方法:数据清洗、离散变量的虚拟变量建立、报告数量的对数变换和新属性的设置。这四个步骤可以消除多余的信息，方便从数据集中识别和提取相关信息。

第一步:在数据清洗阶段，我们使用 Python 检查缺失值、离群值和重复值。通过测量单词的长度，我们检查是否有空值或异常长的值。我们发现没有空值，只有三个异常值:“tash”、“clen”和

“rprobe”。在网上搜索比较后，我们将这些词纠正为“垃圾”、“干净”和“探测”。此外，使用“duplicate()”方法，我们检查没有发现重复值的重复值。

步骤 2:为了使词性的离散变量更容易被模型处理，我们构造了 17 个虚拟变量，将离散变量转换为二元变量。

步骤 3:我们计划使用时间序列模型来预测 2023 年 3 月 1 日的报告数量。在这些类型的模型中，消除数据中的异方差是至关重要的。对数据取对数并不会改变其性质或相关性，但会压缩变量的尺度。通过压缩数据的绝对值，更容易消除异方差的问题。因此，我们对报告量进行对数变换。

步骤 4:为了全面探索各种词属性对报告的 hard - mode -play 分数的影响，我们进一步提取词的属性并建立几个新的变量。这将在 3.4 节中详细阐述。

3.1.2 数据描述和探索性分析

将数据可视化，挖掘其内在规律，有助于建模。图 2 描述了变量之间的相关性，而图 3 则以直方图的形式呈现了尝试次数的分布。图 4 显示了报告总数和报告在困难模式下所占比例随时间的变化曲线。

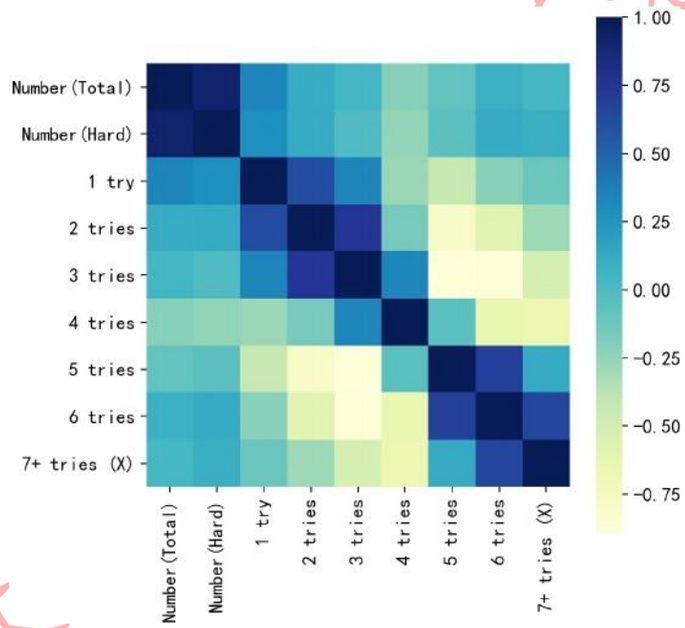

图 2:相关矩阵

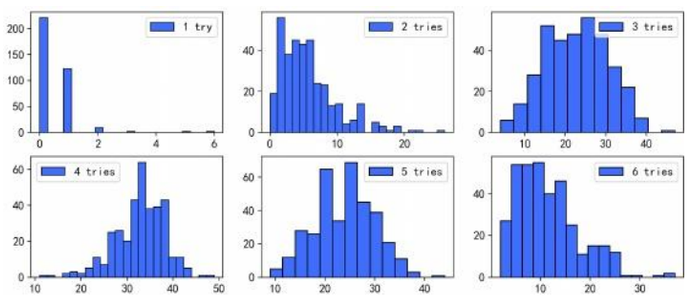

图 3:分布直方图

我们可以看到，变量之间的相关性普遍较弱，尝试次数的分布呈现两端低，中间高的状态。数量曲线的走势与感染曲线有些相似，我们将在接下来的步骤中进行详细分析。

3.2 先知模型

Facebook 提供的 Prophet 算法[3]不仅可以处理有一些离群值的时间序列数据，还可以处理部分缺失值。它几乎可以自动预测时间序列的未来趋势。它基于时间序列分解和机器学习拟合，使用开源工具 pyStan 对模型进行拟合，因此可以快速获得预测结果。

在对数据进行对数变换(在 3.1.1 节中详细阐述)之后，我们使用 Prophet 建立了一个乘法模型，其参数如表 3 所示，其中τ为 a

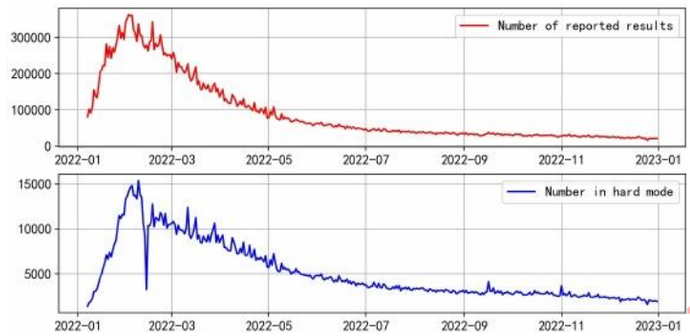

图 4:数量曲线

控制线性函数在断点处的斜率的参数。对于变点处的变化率，记为Δ，则得出Δ~Laplace(0，τ)。随着τ减小，Δ趋近于 0。因此，增大τ会拓宽预测值的上限和下限。趋势项使用默认的分段线性函数。设置更多的变点，增加断点的范围，使得模型对时间序列数据的变化更加敏感，从而提高了拟合效果。

表 3:先知模型参数设置

the Number of Changepoints	60	
τ	0.8	
the Range of Changepoints	0.9	
Holidays	Valentine	2022/02/14
	Easter	2022/04/24
	Halloween	2022/10/31
	Thanksgiving	2022/11/24
	Christmas	2022/12/25

先知模型通常由趋势项 g(t)、季节项 s(t)、假日效应项 h(t)和残差项ε(t)组成。G(t)是一个分段线性函数，它满足:

$$g(t) = (k + a(t)\Delta)t + (m + a(t)^{\intercal}\gamma) \tag{1}$$

其中 k 表示增长率，Δ表示增长率的变化，m 表示偏移量参数。S(t)包含每周的周期性变化:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right) \tag{2}$$

其中 P 为周期时间，(a_n, b_n)， (n = 1…N)服从正态分布。H(t)说明了假期对结果的潜在影响:

$$h(t) = \sum_{i=1}^{L} k_i * l_{\{t \in D_i\}} \tag{3}$$

其中 ki, I = 1…L 服从正态分布。基于上述参数和函数，建立乘法模型:

$$y(t) = g(t) * s(t) * h(t) * \varepsilon(t) \tag{4}$$

我们将 2022-01-07 至 2022-11-21 的数据作为训练集，将 2022-11-21 至 2022-12-31 的数据作为测试集。拟合结果如图 5 所示，其中红色竖线代表我们设置的断点。

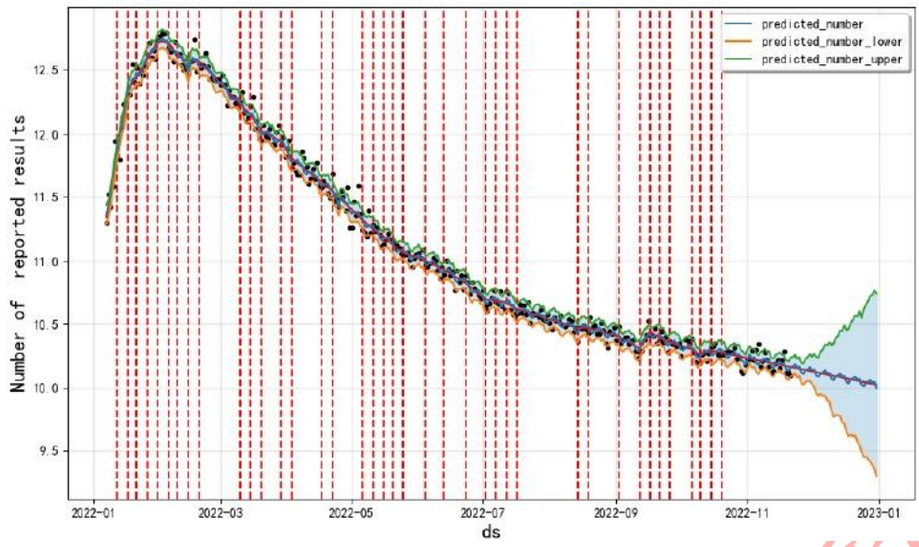

图 5:先知预测(Prophet Forecasting)

我们使用四个指标来评估模型的有效性:r 平方、MSE、RMSE 和 MAPE，具体如下:

$$
\begin{aligned}
MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\
RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \\
R^2 &= 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \\
MAPE &= \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right|
\end{aligned}
\tag{5}
$$

其中，yi （∧） =拟合值，y_i =实际值。结果如表 4 所示。r 平方值接近于 1，表明模型的拟合非常好。由于我们的最终结果是通过取对数变换数据的指数得到的，因此可以考虑较小的 RMSE 和 MSE。MAPE 为 4.8%表明平均绝对百分比误差较小。总体而言，所建立的模型适合于预测。

基于上述数据，我们减小τ以提高预测区间的精度。然后我们重新建立模型并预测 2023 年 3 月 1 日报告的结果数量

表 4:先知的评价

R-squared	MSE	RMSE	MAPE
0.9924	60340502	7767.9149	4.8002

结果为 14534，预测区间为(13175,16128)(95%置信水平)。这些预测结果表明，随着时间的推移，世界大战的受欢迎程度正在下降。

3.3 报告数量变化的解释

报表数量的变化可以分解为趋势、季节、节假日三个部分，如图 6 所示。我们将从这三个方面来解释报表数量的变化。

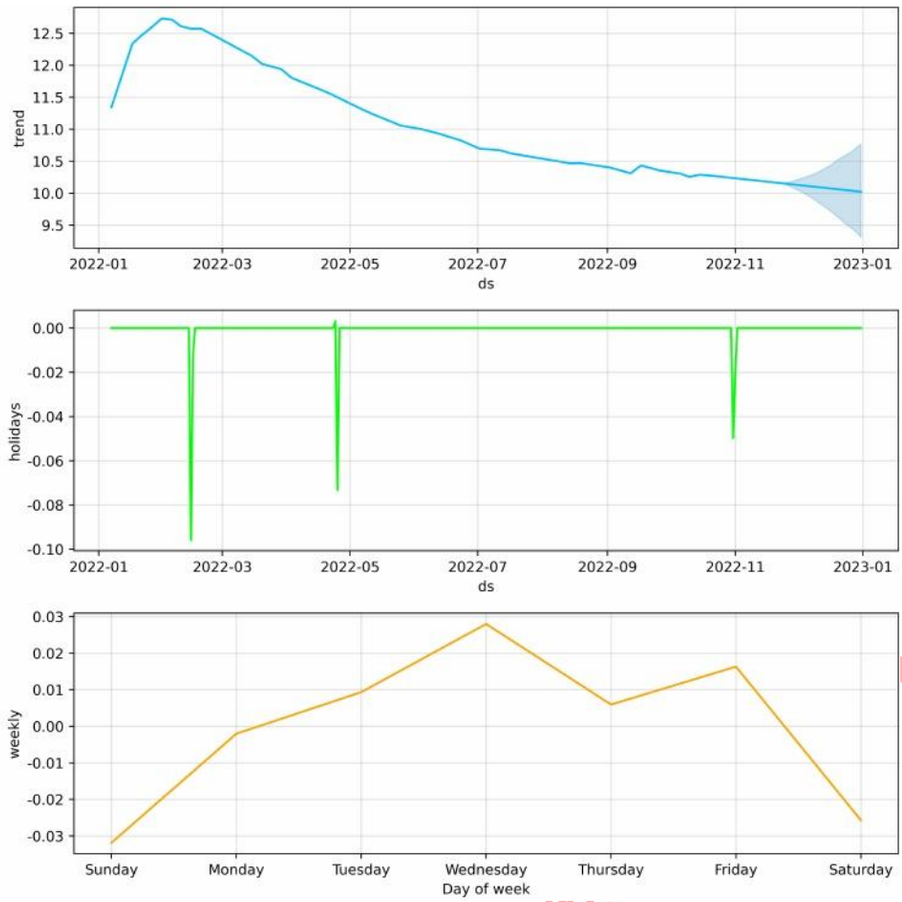

图 6:时间序列分解图

季节性和假日效应:

节假日导致报告数量减少，比如线性趋势图中情人节前后报告数量略有下降。在周效应中，报告数量从周日到周三增加，从周三到周六减少(周五有反弹)。这表明人们倾向于在工作日将《世界大战》作为一种消遣，而在假期则不太感兴趣。

整体变异解释:

SIRS 传染病模型可以很好地解释趋势成分的变化。我们的假设如下:

假设 1:所有 Twitter 用户 A(t)可以分为三组:

(1)普通 Twitter 用户 S(t)。他们可能会因为在 Twitter 上看到一些《世界大战》玩家的得分报告而受到影响，并有可能成为《世界大战》玩家。他们对应的是“易感个体”;

(2)世界选手 I(t)。一些玩家会在 Twitter 上发布报告，这将吸引其他人成为《世界大战》玩家。他们对应的是“被感染的个体”;

(3)疲倦玩家 R(t)。他们在一段时间内不会玩《世界大战》，但在这段时间后可能会重新开始玩。他们对应的是“康复个体”。

假设 2:普通玩家 S 可能有λ被感染的概率;在玩家 I 中，他们有可能会厌倦《世界大战》，并在一段时间内不玩游戏;在疲惫的玩家 R 中，有可能会受到外部因素的影响而重新开始玩《世界大战》。普通玩家 S、玩家 I、疲惫玩家 R 可能都有自然移除一定θ的概率。

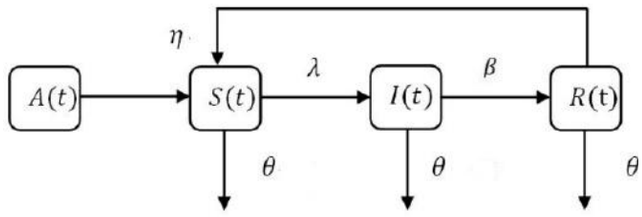

图 7:玩家状态转换

在上述假设的基础上，设置好参数后，通过求解微分方程来拟合玩家人数，然后乘以一定的比例来计算 Twitter 上的得分报告数量。报告数量对应的拟合曲线如图 8 所示，符合先知的趋势曲线。因此，SIRS 模型可以用来解释变化的总体趋势。世界大战从 2022 年 1 月开始流行，玩家数量在 2 月左右达到峰值(报道数量也达到峰值)。之后，游戏逐渐降温，玩家数量减少，报道数量也随之减少。

3.4 提取单词的属性

为了研究单词属性对具有挑战性模式的报告比例的影响，我们首先需要提取单词的各种有用属性。

1.一个单词中不同字母的数量(NDLW)

一般来说，一个单词的不同字母越少，在测试中猜出一个字母的概率就越低，谜题难度也就越大。我们统计了不同字母数单词的分布，以及尝试 5 次以上的人的平均比例，结果如表 5 所示。从表中可以看出，尝试次数越少

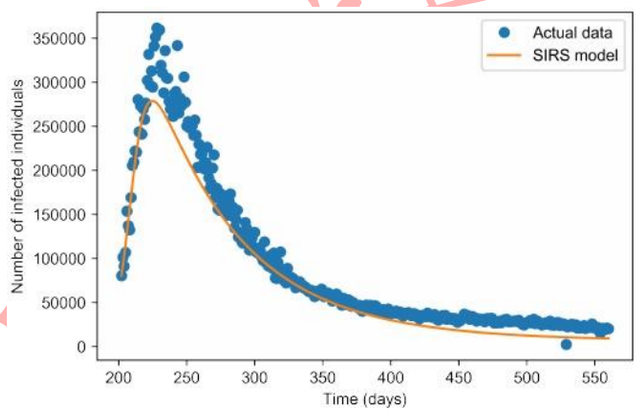

图 8:SIRS 拟合曲线

一个单词的字母不同，尝试 5 次以上的人比例越高，说明这个谜题难度越大。因此，一个单词中不同字母的数量是这个单词的一个重要属性。

表 5:单词中字母的种类和 5+尝试的比例

Different Letters	Number of Words	Proportion of 5+ Tries
3	6	62.50%
4	94	45.10%
5	259	34.90%

2.词汇在日常生活中的使用频率(Freq)

一般来说，一个词在日常生活中使用的频率越高，人们对它的熟悉程度越高，反之亦然。而字谜中越不熟悉的单词，字谜难度就越大。因此，单词在日常生活中的使用频率也是一个必不可少的属性。我们使用来自 Wolfram[4]的词频数据，它是从 Google Books 数据集中计算出来的。

3.不同领域词汇使用的广度(BU)

一个词的使用越广泛，人们对它的熟悉程度就越高，反之亦然。人们对字谜中的单词越不熟悉，字谜就会变得越难。一个词的流行度定义为该词在 100 个语料库中出现的语料库数量(数据来自《书面英语和口语中的词频》)。

4.字母使用频率总和(SLF)

在玩“世界”游戏时，玩家通常会尝试包含更多常见字母的单词来获取更多信息。因此，单词中的字母是否常见，也是衡量单词难度的一个重要属性。我们定义 SLF 来描述一个词的这个属性:

$$SLF = \sum_{i=1}^{5} frequency_i \tag{6}$$

其中 frequencyi 表示单词中字母 i th 出现的频率。字母频率数据来源于网站 Algoritmy[1]。

5.一个单词中一个字母的总和

单词中字母的总和也是单词的一个属性，因为字谜由五个相同或不同的字母组成。

6.一个词的词性

词的词性是一个词最常见的属性之一。

3.5 词属性对硬模式报表比例的影响

硬模式下的报表占比定义如下:

$$percentage_{hard} = \frac{number_{hard}}{number_{reported}} \tag{7}$$

我们建立了基于最小二乘法的多元线性回归模型，并利用回归方程的显著性检验(即 f 检验)来研究词属性是否对硬模式报告的比例有影响。

3.5.1 模型建立

多元线性回归描述因变量 y 与自变量 x1, x2，…的关系。， xm 用下面的方程表示:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \tag{8}$$

式中 0，β为常数项 k，β为 KTH 自变量的回归系数，ε为随机误差项。我们以 Freq、SLF、NDLW、BU、the Sum of a Letter in a Word、Part-of-Speech of a Word 作为自变量，以 percentagehard as 作为因变量，进行多元线性回归。由于得到的回归方程较长，故列入附录 A。

3.5.2 回归方程的显著性检验

1.假设配方:

Null hypothesis: $H_0 : \beta_0 = \beta_1 = \cdots \beta_m = 0;$

备选假设:H1 :0 β、1β、…、m β不都等于 0。

2.计算 f 统计量:

$$F = \frac{SSR/m}{SSE/(n-m-1)} \sim F(m, n-m-1) \tag{9}$$

其中SSR表示回归后的平方和，SSE表示其中的残差平方和。

3. 基于给定的显著性水平α= 0.05，测试的拒绝区域为 Fα > F(m, n－m－1)。我们建立了一个多元线性回归模型，以单词属性为自变量，以硬模式下的报告比例为因变量。

回归方程的 f 统计量为 1.058，对应的 p 值为 0.379 (>α= 0.05)，说明回归方程不具有统计学意义。因此，我们得出结论，在硬模式下，单词属性对报告的占比没有显著影响。

4 模型 2-基于 BP 神经网络的分布预测模型

4.1 BP 模型的建立

我们首先将预训练与 GloVe 模型(Global Vectors model)相结合，将降维与主成分分析法(PCA)相结合，对数据进行预处理。词嵌入是一种将词映射到实值向量的技术，是自然语言处理中的基本应用。GloVe 模型是一种词嵌入模型，它采用平方损失，将词向量拟合到基于整个数据集计算的全局统计信息中。我们使用来自 GloVe 模型的预训练词向量作为特征。

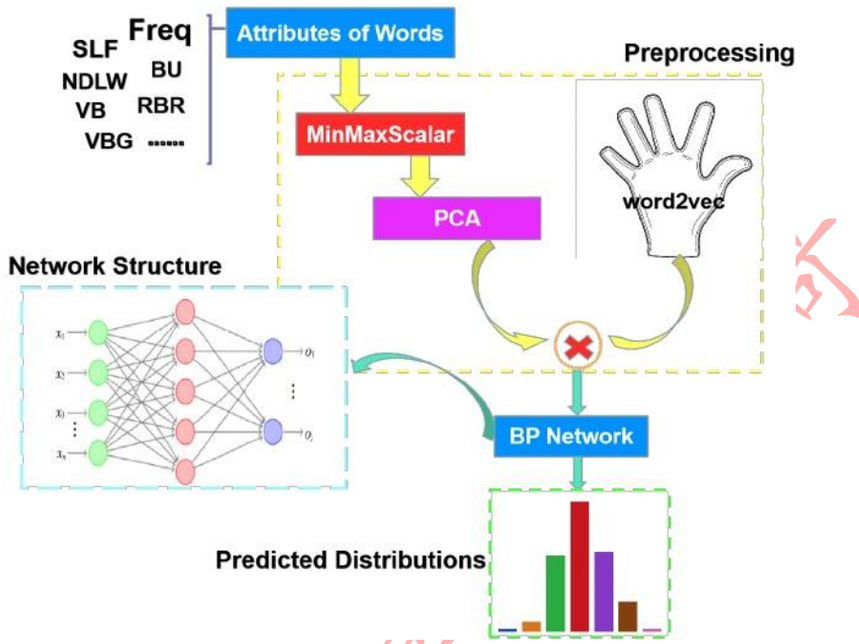

图 9:实现过程

在解决任务 2 时，我们使用从第一个子问题中获得的单词属性作为特征的一部分，对它们进行规范化，并使用 PCA 提取主成分。我们将提取的主成分与使用 GloVe 模型预训练的词向量结合起来，并将它们作为 BP 神经网络的输入特征。在 BP 神经网络中，我们输入单词特征，并使用每个单词的百分比分布作为标签。我们选择 80%的数据作为训练集，20%作为测试集来训练神经网络并测试其效果。由于给定的数据量很小，我们选择建立一个低复杂度的网络，它包括一个输入层、一个单个隐藏层和一个输出层。隐藏层包含 1024 个隐藏单元，使用 ReLU 函数作为激活函数。在训练过程中应用 Dropout 来丢弃 50%的网络单元，以对抗过拟合。选择 l2 范数作为损失函数，反向传播过程中使用 Adam 优化器进行梯度优化，学习率设置为 0.05。使用泽维尔随机初始化。

4.2 BP 的模型不确定性

神经网络具有相当大的随机性，神经网络的泽维尔方法的初始化参数是从均匀分布中采样的。此外，dropout 在隐藏层中随机丢弃神经元。这意味着神经网络的训练结果可能每次都有所不同。为了解决这个问题，我们尝试对模型进行多次训练，并选择最好的模型。

模型输出可能是负的，为了解决这个问题，我们选择将负值调整为 0。

由于输出值不能直接用作百分比，因为它们的总和可能大于或小于 100，因此我们将每个输出除以总总和，以获得最终预测的百分比。

4.3 BP 的模型评价

在测试集上的评价结果如下:

表 6:对测试集的评估

MAE	MSE	MAPE
3.2302	21.857	32.9194

在测试集上，神经网络的平均绝对误差(MAE)在 4 左右，表明预测值与真实值的平均绝对差为 4，表明模型具有较高的精度。其他指标也支持这一结论。

4.4 BP 的模型预测

我们预测了尝试不同次数的人的分布，结果误差在 3%以内。

5 基于 K-Means++的 Model 3-难度分类

5.1 基于 K-Means++的聚类分析

K-Means 算法是一种无监督学习方法，是一种基于分区的聚类算法。它通常使用欧几里得距离作为度量数据对象之间相似度的度量，相似度与数据对象之间的距离成反比。相似度越大，距离越小。该算法需要预定的初始聚类个数 k 和 k 个初始聚类中心。该算法基于数据对象与聚类中心的相似性，不断更新聚类中心的位置，降低聚类的平方和误差(sum of squared errors, SSE)。

Try Times	Percentage(%)
1	0
2	6
3	18
4	29
5	27
6	15
7+	5

表 7:“EERIE”预测结果

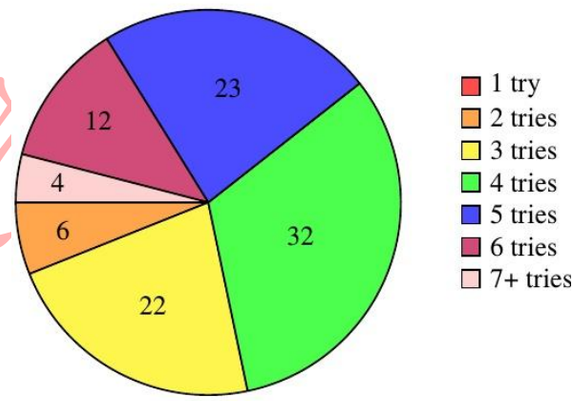

图 10:“EERIE”的预测结果

当 SSE 较长的变化或目标函数收敛，聚类结束，得到最终结果。

“1 次尝试”、“2 次尝试”、直到“7+尝试”的类别很好地反映了谜题的难度。我们使用这些变量作为输入，并采用 k - memeans ++算法对单词的难度进行分类。具体过程如下:

第一步:确定集群个数 k，初始化 k 个集群中心。

步骤 2:计算数据点与 k 个初始聚类中心之间的欧氏距离，并根据最小距离进行聚类划分，得到 k 个区域。

步骤 3:计算上一步得到的每个聚类的中心位置，并将其作为下一次迭代的聚类中心。

步骤 4:重复上述步骤，直到最后两个聚类结果之间的变化满足精度要求。

图 11 中的肘规则图结合我们区分游戏难度的经验来确定聚类的数量。我们选择集群的数量为 k=3，代表三个难度级别:难、中、易。

最后，我们得到了聚类结果，聚类结果显示，聚类 1 包含 135 个单词，聚类 2 包含 156 个单词，聚类 3 包含 68 个单词。聚类中每个属性的均值和标准差的统计结果如表 8 所示。通过计算每个聚类中 5+次尝试的平均比例，我们得到表 9。通过观察表 8 和表 9，我们将聚类 1、聚类 2 和聚类 3 中的单词分别分类为简单、中等和困难。

5.2 单词属性和难度等级之间的关系

5.2.1 难度等级与 NDLW 的关系

不同 NDLW 词的词难度分布如图 12 所示，不同难度 NDLW 所占比例如表 10 所示。我们发现，不同字母较少的单词比例随着难度等级的增加而增加。在本研究提供的数据集中，有 6 个单词只有 3 个不同的字母。

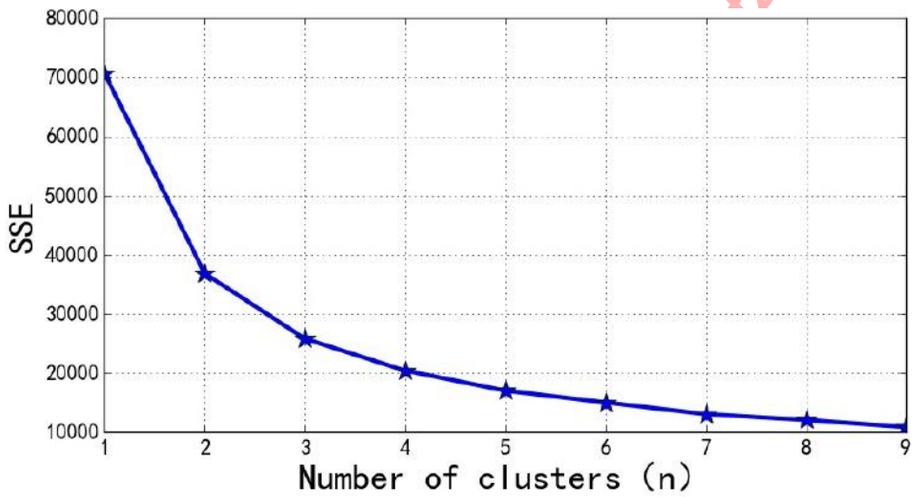

图 11:k 值优化

表 8:不同类别 K-means 聚类的结果和显著性检验

	Cluster Categories (means ± sd)				
	1 (n=135)	2 (n=156)	3 (n=68)	F-value	P-value
1 try	0.8 ± 1.057	0.269 ± 0.459	0.279 ± 0.452	21.307	**0.000*****
2 tries	9.459 ± 4.168	4.013 ± 1.749	2.868 ± 1.962	168.146	**0.000*****
3 tries	30.748 ± 3.81	20.212 ± 3.516	12.574 ± 4.108	594.862	**0.000*****
4 tries	33.637 ± 3.824	35.487 ± 3.819	25.647 ± 4.485	150.08	**0.000*****
5 tries	17.807 ± 3.159	26.436 ± 3.115	28.794 ± 5.732	268.933	**0.000*****
6 tries	6.43 ± 2.261	11.596 ± 3.05	21.662 ± 4.188	565.646	**0.000*****
7+ tries (X)	1.081 ± 0.931	1.974 ± 1.169	8.132 ± 7.035	119.123	**0.000*****

表 9:不同类别 5+尝试的平均比例

Cluster categories	Average Proportion of 5+ tries
Cluster 1	25.35%
Cluster 2	39.92%
Cluster 3	58.59%

表 10:不同难度等级的 NDLW 比例

Cluster Categories	Proportion of NDLW		
	5	4	3
Easy	90.78%	9.22%	0%
Medium	63.13%	35.63%	1.24%
Hard	52.94%	41.18%	5.88%

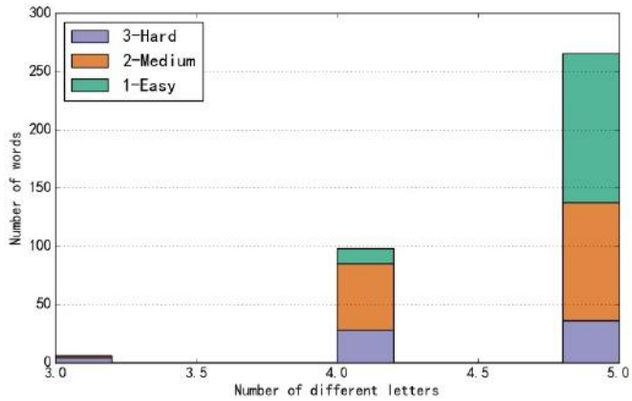

图 12:难度等级与 NDLW 的关系

其中 2 个为中难度，即“座右铭”和“夫人”，另外 4 个为难度，即“绒毛”、“木乃伊”、“可可”、“生动”。基于以上分析，我们有足够的证据表明，一个单词包含的不同字母越少，猜测的难度就越大。

5.2.2 难度等级与 SLF 的关系

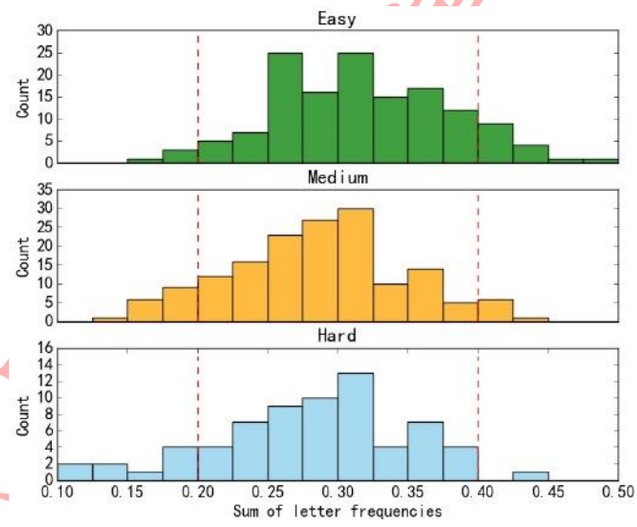

图 13:难度等级与 NDLW 的关系

不同难度单词的 SLF 分布如图 13 和表 11 所示。我们主要关注 SLF 小于 0.2 和大于 0.4 的部分，因为它代表了难度水平和大多数字母频率和之间的关系。可以看出，随着难度的增加，SLF 小于 0.2 的比例增加，而 SLF 大于 0.4 的比例减少。这意味着，一个单词中使用频率越高的字母越容易被猜出，反之亦然。

5.2.3 难度等级与 BU 和 Freq 的关系

表 12 为不同难度等级的词宽分布情况。一个词的宽度定义为该词在总共 100 个语料库中出现的数量

表 11:SLF 在不同难度级别的分布

Cluster Categories	Proportion of SLF		
	< 0.2	0.2-0.4	> 0.4
Easy	2.84%	86.52%	10.64%
Medium	10%	85.63%	4.37%
Hard	13.24%	85.29%	1.47%

语料库，它取 0 到 100 之间的整数值。该表表明，一个词在不同领域的使用越广泛，对应的谜题就越容易解，反之亦然。

同时，我们试图找到一个词的难易程度和它在日常生活中的使用频率之间的关系。虽然不同难度等级的平均使用频率存在差异，但平均值对异常值很敏感。因此，我们首先按频率对单词进行排序，然后使用直方图来显示它们的分布，如图 14 所示。最终，我们发现，在日常生活中，难度等级和使用频率之间并没有明显的关系，因为不同难度等级的词频分布是相当均匀的。

表 12:BU 在不同难度水平上的分布

Cluster Categories	BU		
	0-33.33	33.33-66.66	66.66-100
Easy	9.93%	12.77%	77.30%
Medium	19.38%	21.87%	58.75%
Hard	22.06%	26.47%	51.47%

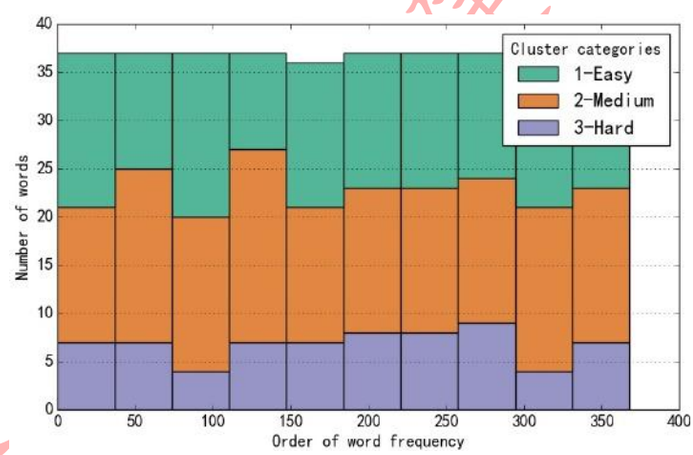

图 14:难度等级与频率的关系

5.3PCA 对模型分类精度的探讨

关于模型分类精度的讨论可以分为两部分。第一部分是每个聚类中的单词之间是否存在显著的难度差异，第二部分是 K-Means 聚类的有效性。如前所述，在第一部分中，

每个聚类单词的平均尝试次数为 5+的比例分别为 25.35%、39.92%和 58.59%，这表明聚类之间存在显著的难度差异。

在第二部分，我们对“1 次尝试”、“2 次尝试”到“7+尝试”这 7 个特征进行了主成分分析(PCA)。我们发现，前两个主成分解释的方差达到 82.88%(>80%)。因此，我们取前两个主成分创建散点图，如图 15 所示。从这个散点图中可以看出，单词的分化很好，K-Means 聚类效果很好。

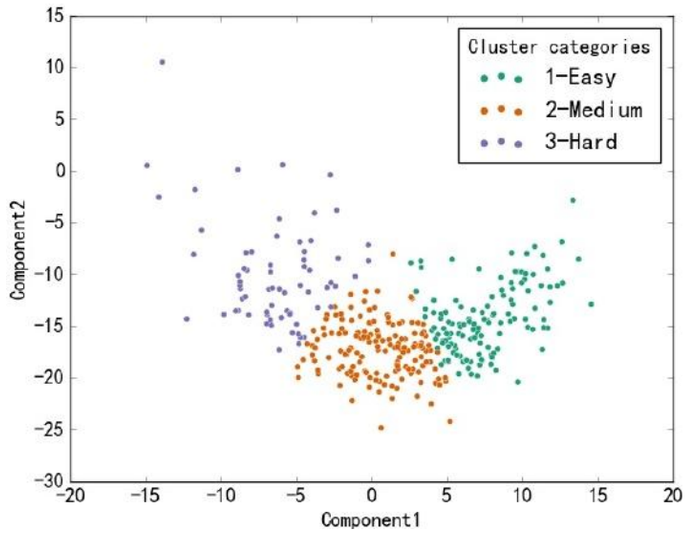

图 15:PCA 对模型分类准确率的影响

从评价指标的角度，我们将 K-Means++与其他聚类方法如 Partitioning Around medioids (PAM)和 Gaussian Mixture Model(GMM)进行了比较，结果如表 13 所示。从比较结果可以看出，K- means++具有更好的聚类效果。

表 13:模型评价(Model Evaluation)

Model	Silhouette Coefficient	CH Score
K-Means++	0.372	309.326
PAM	0.347	303.049
GMM	0.347	291.637

5.4 确定“EERIE”的难度等级

根据第二题的结果，我们得到了 EERIE 的相关百分比(1,2,3,4,5,6,X)的分布，即[0,6,18,29,27,15,5]。通过将这个分布输入到我们的模型中，我们确定“EERIE”的难度等级为“中等”。

6 有趣的惊喜

6.1 这些单词真的那么难吗?

通过对单词的难易程度进行分类，我们发现“木乃伊”这个单词的难易程度是“难”。当我们计算已经尝试过“5 次及以上”的单词的比例，并将其按降序排序时，我们惊讶地发现，“木乃伊”在难度上排名第二，有 82%的人尝试了 5 次及以上。由于我们从小就对“木乃伊”这个词很熟悉，所以我们主观地认为这是一个简单的词，这使得相反的结果更加令人惊讶。除此之外，其他类似的词还有“watch”、“catch”、“prize”等。事实上，这也印证了我们的观点，即“词频和难度等级之间没有显著的关系”。

6.2 哪个首字母对解词的难度最大?

我们使用假设检验来确定哪个首字母是最困难的字谜。首先，我们统计所有单词中每个首字母出现的频率。接下来，我们将一个单词的难度系数定义为“5 次尝试”、“6 次尝试”和“7 次或更多尝试(X)”类别的和，并计算所有单词的难度系数。在对难度系数进行排序并选出难度最高的前 20%的单词后，我们对这些单词的首字母进行统计。然后我们计算每个字母是进入前 20%最难列表的单词的首字母的概率。部分结果如表 14 所示。

假设以某个字母开头的单词进入前 20%最难列表的概率为 0.2，即

$$H_0: \quad p = 0.2 \tag{10}$$

设 n 为该字母作为第一个字母出现的总次数，k 为该字母作为第一个字母出现在进入前 20%最难列表的单词中的次数。然后我们有

$$k \sim B(n, 0.2) \tag{11}$$

由于 n 和 k 都是已知的，我们可以计算出这种情况发生的概率 P。取显著性水平α为 0.05，当 P < α= 0.05 时，我们拒绝原假设，认为 P 不等于 0.2。我们已经计算了所有字母的 P 值，并将它们按升序排序。部分结果如表 15 所示。

从表 15 可以看出，对于字母 e、s、f、w、a，它们对应的 P 值都较小

表 14:首字母与难度的关系

the First Letter	Total	Hard(20%)	Easy(20%)	Hard Rate	Easy Rate
a	28	2	5	0.071	0.179
b	20	4	3	0.200	0.150
c	33	7	10	0.212	0.303
d	12	4	5	0.333	0.418
e	10	6	0	0.600	0.000
f	22	8	2	0.364	0.091
g	17	5	1	0.294	0.059
h	11	4	2	0.364	0.182
...	...	...	...	...	...

表 15:难词首字母统计

the First Letter	Total	Hard(20%)	Hard rate	P value
e	10	6	0.6	0.006
s	51	4	0.078	0.011
f	22	8	0.364	0.036
w	11	5	0.455	0.039
a	28	2	0.071	0.046
r	13	0	0	0.055
t	30	3	0.1	0.079
...	...	...	...	...

比 0.05，所以 p ≠0.2 有统计学意义。首字母为 e、f、w 的单词的 Hard rate 都大于 0.2。因此，我们有充分的理由相信，以 e、f、w 开头的单词是困难的，以 e 开头的单词是最难的。同样地，我们可以发现字母 t 作为首字母时对应的是最简单的单词。具体的数据可以在表 16 中找到。

表 16:易词首字母统计

The first letter	Total	Easy(20%)	Easy rate	P value
t	30	16	0.533	0
d	12	5	0.417	0.053
c	33	10	0.303	0.056
s	51	13	0.255	0.081
...	...	...	...	...

6.3 哪些词能让世界继续流行?

一个向上的突变点表示一种本应下降的趋势突然上升。这可能意味着向上突变点对应的单词更容易刺激人与人之间的交流和传播，从而导致当天的报道数量增加。为了检测向上突变点，我们首先使用隔离森林算法检测所有突变点，然后使用一阶差分的正负性来搜索向上突变点。同时，我们计算突变率，选择突变率大于 5%的突变点作为最终点。结果如下表所示。

Date	Word	Date	Word	Date	Word
2022/2/8	frame	2022/3/18	saute	2022/4/18	flair
2022/2/15	aroma	2022/3/21	their	2022/4/22	plant
2022/2/17	shake	2022/3/24	chest	2022/4/26	heist
2022/2/19	swill	2022/3/27	nymph	2022/4/29	tarsh
2022/2/22	thorn	2022/3/30	stove	2022/5/2	story
2022/3/2	nasty	2022/4/1	snout	2022/5/4	train
2022/3/5	brine	2022/4/2	trope	2022/5/9	shine
2022/3/11	watch	2022/4/8	scare	2022/5/11	farce
2022/3/15	tease	2022/4/13	chunk	2022/12/26	judge

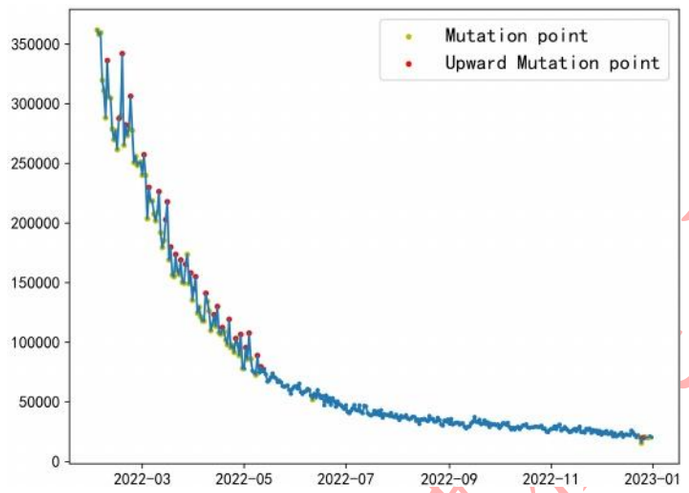

图 16:灵敏度分析

《纽约时报》或许可以借鉴这些热潮词的特点，提升 world 的趣味性。

7 敏感性分析

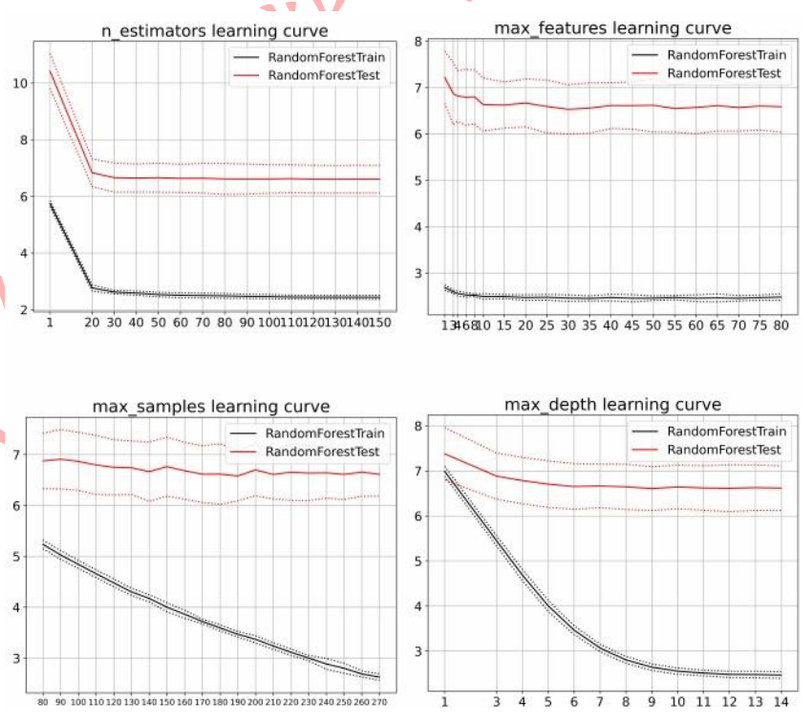

图 17:敏感性分析

在改进 BP 神经网络模型时，我们使用随机森林模型对三次尝试的结果进行回归。我们选择了参数 nestimators(弱评估器的数量)、maxfeatures(采样特征的最大数量)、maxsamples(随机采样的最大样本量)和 maxdepth(树的最大深度)进行敏感性分析。在每个参数分析中，我们只改变该参数，同时保持另一个 pa-参数设置为默认值。我们对数据集进行了 5 倍交叉验证，并将训练集和测试集上的 RMSE

的平均值作为每个参数值的结果。如图 17 所示，随着参数值的增加，训练集和测试集上的 RMSE 逐渐减小并趋于稳定。各种参数变化下的 RMSE 稳定在 6.60 左右。这说明模型对参数变化不敏感，相对稳定。

8 模型评估

8.1 优势

1. 先知模型考虑了节假日对时间序列的影响，模型参数的可解释性更强，有助于我们理解报告数量的变化。

2. SIRS 模型对于 Prophet 模型的趋势项具有极好的解释力。

3. BP 神经网络具有输出多个目标的能力，并且不需要为每个目标单独建立模型。

4. k - means++模型具有简单实用的特点，适合本题的数据集。

8.2 缺点

1. Prophet 模型对于长期预测有些不足，在某种程度上类似于传统的时间序列模型。

2. BP 神经网络的参数在初始化过程中具有随机性，训练后的模型也包含了这种随机性，这意味着在某些情况下，模型的输出结果并不总是唯一的值。

3. 虽然我们正在努力寻找与谜题难度相关的词属性，但可能仍然有一些词属性被我们忽略了。

References

[1] Algoritmy.net. Letter frequency english, Accessed on 2023-02-18.

[2] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O’ Reilly Media, 2009.

[3] Sean J Taylor and Benjamin Letham. Forecasting at scale. The American Statistician, pages 0 – 0, 2017.

[4] Wolfram.com. Wordfrequencydata, Accessed on 2023-02-18.

公众号·数学建模老哥

尊敬的先生/女士:

我们很荣幸在数据分析和建模后，呈现我们对 worlddle 的报告数量和单词难度的分析。我们相信，我们有趣的发现会对您有所帮助。以下是我们的一些理论分析和数值预测。

1. 报告的数量将略有减少，但将再次上升。玩游戏的过程就像一种感染，人们总是喜欢后就会厌倦，厌倦后又会再次喜欢。玩家的数量就像一条感染曲线，随着时间的推移而增加，达到峰值，然后逐渐减少。此外，在节假日和周末，报告人数往往会减少，从而导致曲线出现小幅振荡。根据我们的模型，报告数量将在 2023 年 3 月 1 日下降到(10,452,21,454)的范围内(95%置信水平)。尽管如此，之后还会增加。

2. 没有单词属性会影响在困难模式下打出的分数所占的比例。通过拟合得到的多元线性回归模型显示，r 平方值仅为 0.129。相应的，p 值为 0.379，远远大于 0.05。这表明，单词属性与 Hard- Mode 报告的比例之间几乎没有相关性。这是合理的，因为在玩之前没有人知道这个词是什么，因此，一个词的属性并不影响一个人是否在高难度模式下玩。

3. 我们可以根据单词的属性更准确地预测答案尝试的分布。通过训练一个 BP 神经网络模型，我们可以根据单词的属性来掌握单词的回答率。例如，对于“EERIE”这个单词，人们尝试次数的分布应该是这样的:0%一次尝试通过，6%两次尝试通过，22%三次尝试通过，32%四次尝试通过，23%五次尝试通过，12%六次尝试通过，4%不通过。我们有信心，误差范围在 3%以内。

4.导致猜词困难的属性可能与你想象的不同。通过 K-means++聚类，我们根据尝试成功的比例将单词难度分为三个级别，并分析了难度与单词属性之间的关系。结合前面提到的神经网络模型，我们可以根据单词的属性直接判断单词的难易程度。人们可能会认为猜测一个词的难度与它的使用频率有关，但实际上，这是不正确的。在 worlddle 中，猜词的难度与单词中字母的种类、每个字母的使用频率之和以及在不同领域的使用广度有关。根据我们的分析，“EERIE”这个词应该被归类为中等难度。K-means++模型比其他类似模型表现得更好，从而增加了我们结果的可信度。

接下来，我们将介绍一些有趣的发现:我们打赌你从来没有想过“木乃伊”是第二难猜的单词!高达 82%的玩家需要尝试 5 次以上才能猜出来。其他类似的词还包括“watch”、“catch”、“prize”等。这是因为猜测单词的难度与单词以哪个字母开头有关，例如，以“e”、“s”、“f”、“w”、“a”开头的单词更难猜，而以“t”开头的单词更容易猜。你可以尝试利用这些有趣的特征来设计游戏。

这些都是我们团队给贵公司提供的建议和策略。感谢您宝贵的时间。希望我们的模型和这些结论能对大家有所帮助!

真诚地,

MCM 团队成员

附录

附录 A 回归方程

这里是 3.5.1 中提到的回归方程

$$
\begin{aligned}
\hat{y} = & \underset{(0.000)}{0.045} - \underset{(0.365)}{8.696}\, Freq + \underset{(0.000)}{0.007}\, SLF - \underset{(0.250)}{0.003}\, NDLW + \underset{(0.003)}{0.008}\, a \\
& + \underset{(0.133)}{0.007}\, b + \underset{(0.045)}{0.006}\, c + \underset{(0.013)}{0.009}\, d + \underset{(0.000)}{0.010}\, e + \underset{(0.011)}{0.010}\, f + \underset{(0.014)}{0.008}\, g \\
& + \underset{(0.040)}{0.007}\, h + \underset{(0.003)}{0.010}\, i + \underset{(0.018)}{0.027}\, j + \underset{(0.430)}{0.003}\, k + \underset{(0.011)}{0.006}\, l + \underset{(0.037)}{0.007}\, m \\
& + \underset{(0.069)}{0.006}\, n + \underset{(0.048)}{0.005}\, o + \underset{(0.052)}{0.007}\, p + \underset{(0.515)}{0.007}\, q + \underset{(0.123)}{0.004}\, r + \underset{(0.144)}{0.004}\, s \\
& + \underset{(0.004)}{0.008}\, t + \underset{(0.011)}{0.009}\, u + \underset{(0.047)}{0.010}\, v + \underset{(0.309)}{0.005}\, w + \underset{(0.227)}{0.011}\, x + \underset{(0.003)}{0.010}\, y \\
& + \underset{(0.032)}{0.022}\, z + \underset{(0.320)}{0.016}\, VBG - \underset{(0.062)}{0.020}\, VB - \underset{(0.477)}{0.016}\, CC + \underset{(0.079)}{0.008}\, JJ \\
& + \underset{(0.156)}{0.019}\, VBZ + \underset{(0.118)}{0.019}\, VBN + \underset{(0.052)}{0.016}\, VBD - \underset{(0.525)}{0.010}\, MD \\
& + \underset{(0.066)}{0.008}\, NN + \underset{(0.348)}{0.010}\, NNS + \underset{(0.042)}{0.014}\, RBR + \underset{(0.102)}{0.012}\, VBP \\
& - \underset{(0.700)}{0.008}\, JJS - \underset{(0.742)}{0.007}\, JJR + \underset{(0.918)}{0.003}\, PRP - \underset{(0.377)}{0.019}\, DT \\
& R^2 = 0.129, N = 359
\end{aligned}
\tag{12}
$$

公众号：数学建模