| 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 + +

目录

# 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。

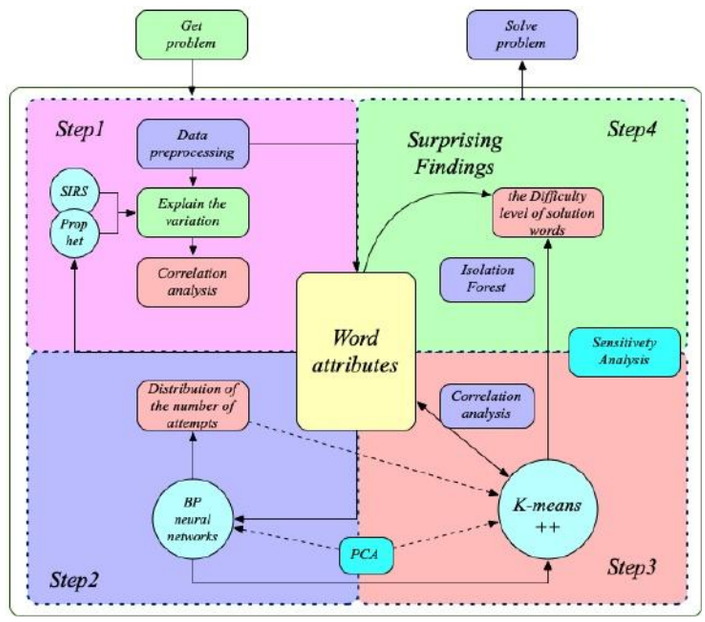


图 1:本文中的流程图

# 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。

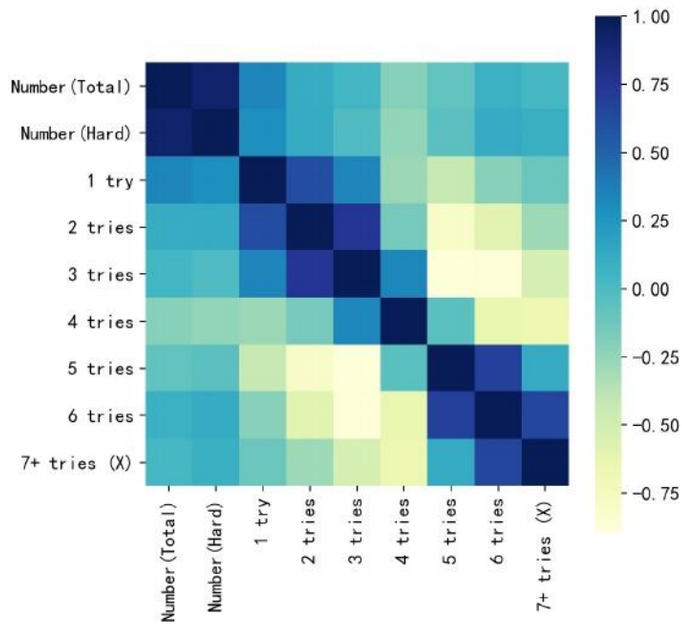


图 2:相关矩阵

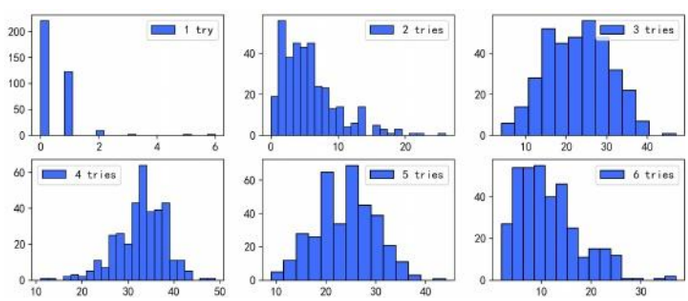


图 3:分布直方图

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

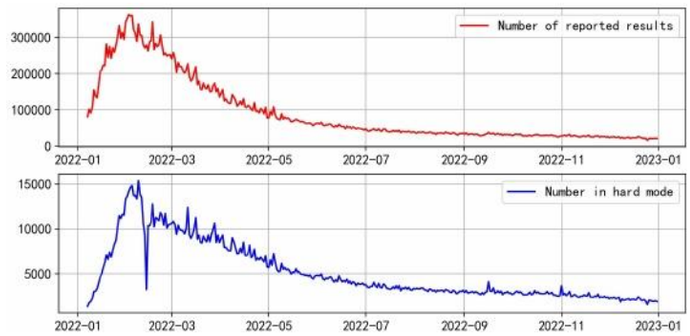


图 4:数量曲线

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 | |
| $\tau$ | 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。

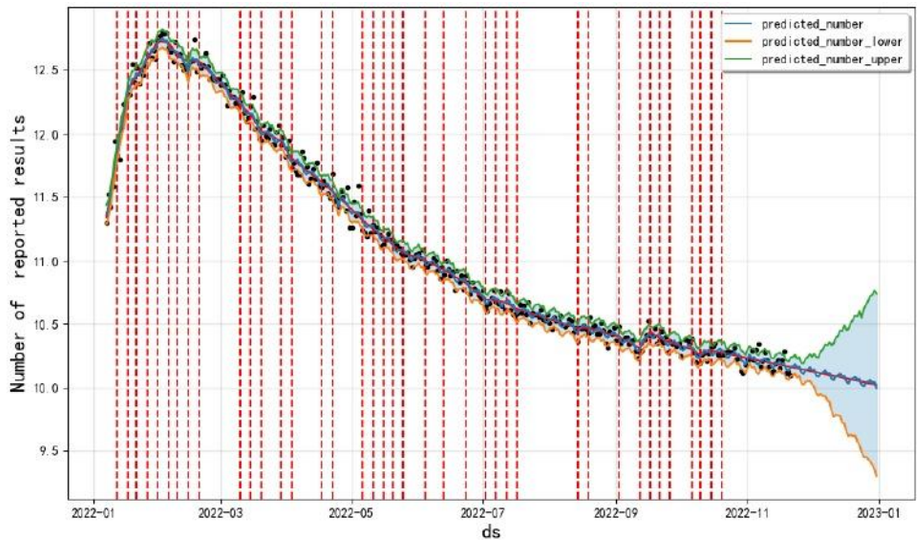


图 5:先知预测(Prophet Forecasting)

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:

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

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。

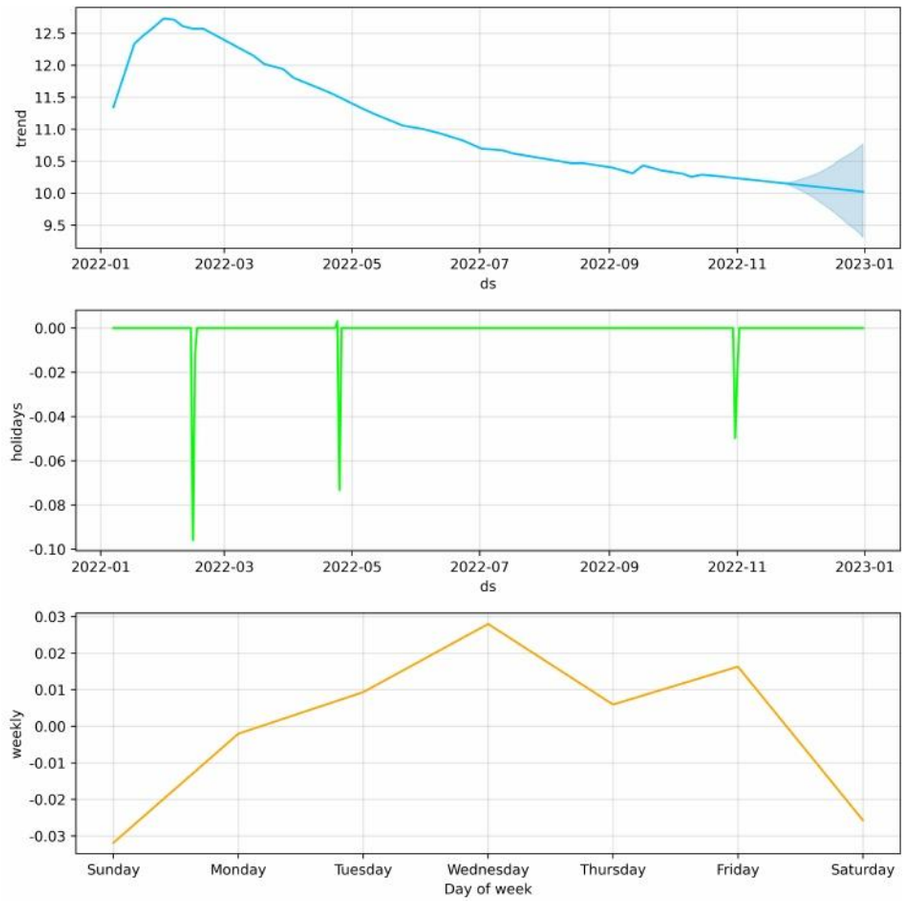


图 6:时间序列分解图

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。

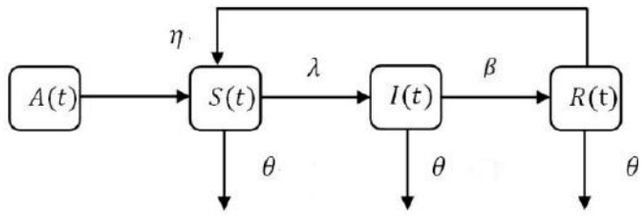


图 7:玩家状态转换

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

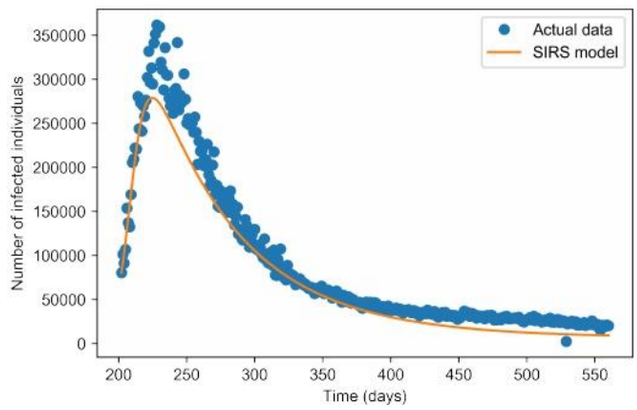


图 8:SIRS 拟合曲线

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。