|  |  |  |
| --- | --- | --- |
| **Problem Chosen** C | **2023 MCM/ICM Summary Sheet** | **Team Control Number**  2309601 |

**关于wordle报道的综合研究**

**Summary**

|  |
| --- |
| Twitter提供了用户反馈渠道收集了Wordle近乎一年的反馈数据，其中包括Date，Word，Number in hard mode等信息。我们运用数学公式和机器学习模型，对纽约时报提出的要求进行合理化分析。  Firstly，我们观察数据的变化趋势，用两段分段函数分别对其前后数据进行拟合。确定第一段使用傅里叶多项式，第二段使用指数函数，**最终分段函数(Q1 Segmentation Function Y Model)**见公式(11)。**方法1推测出2023年3月1日interval for the number of reported results为[1320,2542]**。第二种思路是构建时序关系数据集获得lstm模型。**最终时序模型Q1LSTM Model**的相对误差仅为5%，**方法2推测出2023年3月1日interval for the number of reported results为[1313, 1454]**。我们对第二小问的理解是计算单词的某种属性对PSRTW的相关系数以此说明相关性。通过绘制相关性热力图(图10)发现属性与PSRTW相关度很低，也就是我们选取的属性对PSRTW几乎不影响。  Then，问题2我们用字母频率对单词进行量化。时间序列认为是等差数列。选用深度学习方法，用BP神经网络建立输入为日期序列和单词与输出为7个百分比之间的联系，并使用遗传算法对其进行优化。**Q2GA-BP Model**的相对误差见表5。我们**预测2023年3月1日”EERIE”这个词的相关百分比为[0,4,27,33,22,7,3](%)**。我们的信心程度取决于除1 try和7 or more tries (X)相对误差外其余相对误差的平均值与1差值的百分比。也就是说我们**有63%的信心确定结果准确**。  Next，我们选择3 tires，4tries和5tries作为难易程度的评价尺度。对所有单词依据这三个指标进行类别为4的Kmeans聚类，得到模型**Q3K-means Model。**通过计算聚类中心与原点之间的欧氏距离把所有单词由难到易分为ABCD类。Q3 K-means Model对**”EERIE”预测的难易程度为：B；**此外，我们训练11种常见词性的one-hot编码与难易程度的决策树和随机森林分类器(**Q3DT RF Model**)。**两个分类器的结果都为：B**。两种方法预测结果一致。  Finally，我们觉得Number of reported results很有可能与工作日或者周末有关。我们发现所有时间Number of reported results与周末并无相关。而自统计之后的150天起(此时已经趋于稳定)，星期几与得分存在较大的负相关关系。 |

**Keywords: 时序模型; BP神经网络; K-means聚类; 机器学习**

[1 Introduction 4](#_Toc127842135)

[1.1 Problem Background 4](#_Toc127842136)

[1.2 Restatement of the Problem 4](#_Toc127842137)

[1.3 Literature Review 4](#_Toc127842138)

[1.4 Our Work 4](#_Toc127842139)

[2 Assumptions and Justifications 5](#_Toc127842140)

[3 Notations 5](#_Toc127842141)

[4 Task1：建立日期与Number of reported results的关系映射 5](#_Toc127842142)

[4.1 Problem Analysis 5](#_Toc127842143)

[4.2 The Establishment of Model 6](#_Toc127842144)

[4.2.1 Q1 Segmentation Function Y Model 6](#_Toc127842145)

[4.2.2 Q1 LSTM Model 6](#_Toc127842146)

[4.2.3属性与百分比相关性分析 7](#_Toc127842147)

[4.3 The Solution of Model 7](#_Toc127842148)

[4.3.1 Q1 Segmentation Function Y Model 7](#_Toc127842149)

[4.3.2 Q1 LSTM Model 9](#_Toc127842150)

[4.3.3 属性与百分比相关性计算与分析 10](#_Toc127842151)

[5 Task2：建立日期和单词对百分比的关系映射 11](#_Toc127842152)

[5.1 Problem Analysis 11](#_Toc127842153)

[5.2 The Establishment of Model 11](#_Toc127842154)

[5.2.1 Q2 GA-BP Model 11](#_Toc127842155)

[5.3 The Solution of Model 12](#_Toc127842156)

[5.3.1 Q2GA-BP Model 12](#_Toc127842157)

[6 Task3：建立单词难易程度与单词属性的关系映射 14](#_Toc127842158)

[6.1 Problem Analysis 14](#_Toc127842159)

[6.2 The Establishment of Model 15](#_Toc127842160)

[6.2.1 Q3 K-means Model 15](#_Toc127842161)

[6.3 The Solution of Model 15](#_Toc127842162)

[6.3.1 Q3 K-means Model 15](#_Toc127842163)

[6.3.2 Q3 DT RF Model 16](#_Toc127842164)

[**6.3.3 结果分析** 16](#_Toc127842165)

[7 Task4：挖掘其他有意思的信息 16](#_Toc127842166)

[8 Sensitivity Analysis 17](#_Toc127842167)

[9 Model Evaluation and Further Discussion 17](#_Toc127842168)

[9.1 Strengths 17](#_Toc127842169)

[9.2 Weaknesses 17](#_Toc127842170)

[9.3 Further Discussion 18](#_Toc127842171)

[11 Memorandum 18](#_Toc127842172)

[References 19](#_Toc127842173)

# Introduction

## Problem Background

Wordle作为一款传统的文字游戏，由于具备独特的游戏机制和界面设计而备受人们喜爱，受到了全球范围的广泛关注[1]。文字信息和交互的特点，促使越来越多的玩家开始尝试Wordle，但受到单词特点的限制，每日的游戏通关人数会发生不同幅度的波动。分析影响玩家人数的相关因素，研究该游戏的发展趋势，需要建立合适的数学模型，以此实现对该游戏的发展趋势预测。

## Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

* Problem 1：建立数学模型解释报告结果的数量变化，并使用该模型获得2023年3月1日的报告结果的预测区间，同时分析单词属性对困难模式参与人数的百分比的影响。
* Problem 2：开发一个可以预测未来日期的相关百分比的模型，分析模型和预测结果间存在的不确定性，列举具体例子说明对“EERIE”这个词的预测结果。
* Problem 3：建立可以根据词汇难度对单词进行分类的模型，识别与每个分类相关的给定单词的属性，确定单词“EERIE”在本模型中的难度，并讨论分类模型的准确度。
* Problem 4：列出并描述该给定数据集的一些其它有趣的特征。

## Literature Review

随着互联网技术的普及和应用，网络游戏日渐成为人们日常生活的重要组成元素，显示出巨大的增长潜力[2]。目前，已有部分方法可以较好地应用于游戏玩家数量增长趋势的预测[3,4]。Africa等[5]开发了一个生存集成模型，利用该模型预测玩家的流失率，结合实验证明了该方法在提高传统分析(如Cox回归)准确度方面的有效性。Vafeiadis等[6]总结了几种应用于客户流失预测的机器学习方法，包括人工神经网络、SVM等，验证并评估了各类方法在相关问题上的适用性。

## Our Work

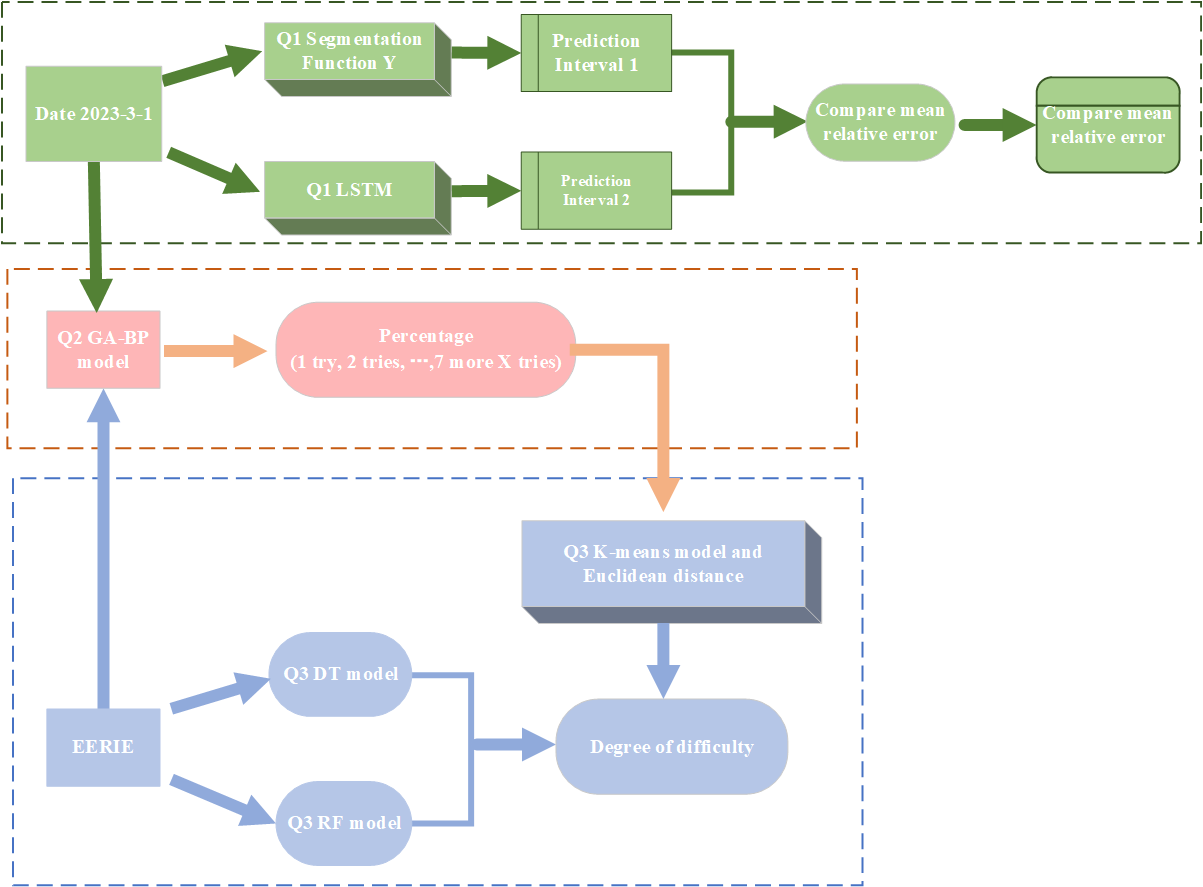


Figure 1: Overall structure of our work

# Assumptions and Justifications

In order to facilitate the establishment of the optimal investment strategy model, we make the following assumptions and simplifications according to the actual situation and classical theory

假设平均相对误差可以说明任何一种模型的好坏；

假设问题2中按照字母出现频率对单词进行量化与游戏规则中填充字母之间存在关系并不冲突；

假设问题3中3 4 5tries百分比对难易程度的影响最大；

假设问题4中自2022年1月7日150天之后”Number of reported results”趋于平稳。不再有剧烈的起伏趋势。

# Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

|  |  |  |
| --- | --- | --- |
| **Symbol** | **Description** | **Unit** |
| *x* | Date |  |
| *y* | Number of reported results |  |
|  | 预测的报告数量结果 |  |
|  | 误差(实际值和预测值相减的绝对值) |  |
|  | 相对误差 |  |
| θ | 平均相对误差 |  |
| epoch | 训练次数 |  |
| Neu\_num | LSTM层神经元个数 |  |
| *n1* | 输入层神经元个数 |  |
| *n2* | 隐藏层神经元个数 |  |
| *Xi(i=1,…,n)* | 第*i*个对象 |  |
| *Ck(k=1,…,n)* | 第*k*个聚类中心 |  |
| *Xit*(1≤*t*≤*m)* | 第*i*个对象的第*t*个属性 |  |
| *Cjt* | 第*j*个聚类中心的第*t*个属性 |  |
| *Sk* | 第k个类簇 |  |
| *Cl*(1≤*l*≤*k*) | 一个聚类的中心 |  |
| |*Sl*| | 第*l*个类簇中对象的个数 |  |
| *Xi*(1≤*i*≤|*Sl*)| | 第*l*个类簇中第*i*个对象 |  |

**阈值日：**Number of reported results随日期变化的趋势(先上升再下降)出现转折的日期，设定为2023年2月4日，即统计日期的29天后。

**SQx(Sub question x)：**在某问题分析的过程中存在的第x个子问题。

**PSRTW(the percentage of scores reported that were played in Hard Mode)：**单词的此属性定义为Number in hard mode除以Number of reported results。

# Task1：建立日期与Number of reported results的关系映射

## Problem Analysis

问题1可分为以下两个子问题：

**SQ1**：建立一个模型来描述” Number of reported results”与” Data”的关系并预测2023年3月1日” Number of reported results”的一个可能存在的区间。

**SQ2**：研究词语的某种属性是否会影响the percentage of scores reported that were played in Hard Mode ,说明原因。

**SQ1分析**：本题实际上是寻找变量与时间的对应关系。我们计划用两种方法解决该问题，并对比两种方法产生的结果。方法1首先需要可视化数据(数据的变化趋势见图1)，观察数据随日期的变化趋势，再尝试用合适的数学公式进行拟合，继而推出指定日期的” Number of reported results”的预测空间；方法2利用已知数据” Number of reported results”，通过相邻项差分构造时序关系数据集，训练符合SQ1的lstm时序模型并进行推理。

**SQ2分析：**重点在于计算单词的某种属性对PSRTW(the percentage of scores reported that were played in Hard Mode)的相关系数以此说明相关性。单词的属性需要自己指定或定义。我们选择单词的11种常用词性以及单词频数(与文件中字母出现频率有关)，通过绘制相关性热力图观察所选属性与PSRTW的相关度。

## The Establishment of Model

### 4.2.1 Q1 Segmentation Function Y Model

根据问题描述，我们首先对” Number of reported results”随时间的变化趋势进行可视化，结果如图2所示。

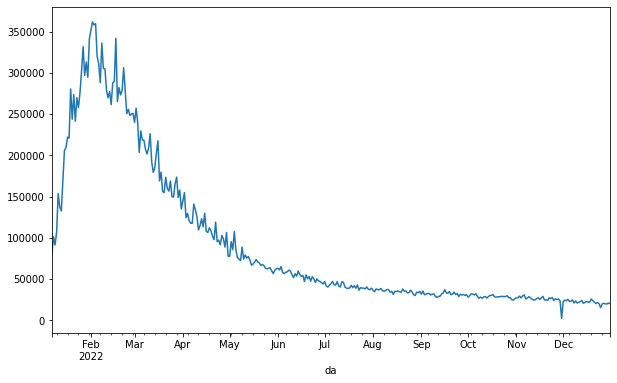


Figure 2： Number of reported results as a function of date

可以发现，随着时间的发展，参与该项游戏的人数急剧增加，在阈值日参与该项游戏的人数达到最大，之后逐渐降低。为了客观描述上述情况，我们考虑采用分段函数对其进行量化。可以将整个过程分为” Number of reported results”的增长期和下降期两阶段。

#### 4.2.1.1 增长期模型建立

设“Number of reported results”的数值为*y*,“Date”数值为*x*。为了便于计算，设当时间为2022年1月7日时，*x*=1，当时间为2022年1月8日时，*x*=2，以此类推。图3为该区间内的散点图，对该散点图进行傅里叶拟合。设“增长期”函数为：



#### 4.2.1.2 下降期模型建立

图4是2022年2月5日至2022年12月31日的数据散点图。分析数据发现，该散点图符合指数函数的变化趋势。使用指数函数(exponential)对该阶段的散点图进行拟合。设函数为



|  |  |
| --- | --- |
|  |  |
| Figure 3： 增长期内散点图 | Figure 4: 下降期内散点图 |

### 4.2.2 Q1 LSTM Model

长短时记忆网络（Long Short Term Memory）模型，本质上是一种特定形式的循环神经网络（Recurrent Neural Network)。它能够将原始数据作为模型输入，自动学习数据特征，输出预测值，是一种“端对端”的模型[7]。LSTM模型在RNN模型的基础上通过增加门限（Gates）来解决RNN短期记忆的问题，使得循环神经网络能够真正有效地利用长距离的时序信息。其具体结构如下图所示。

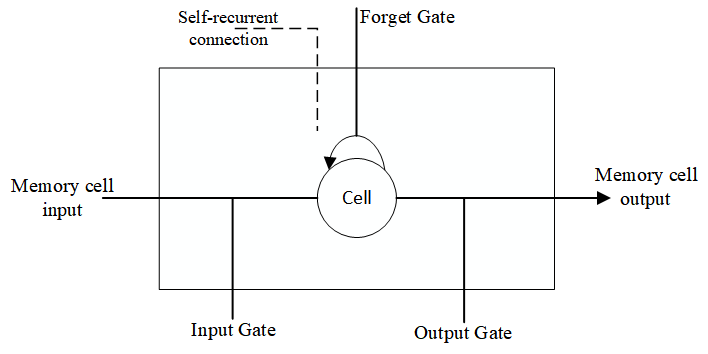


Figure 5: LSTM概念图

### 4.2.3属性与百分比相关性分析

首先计算所有单词的PSRTW，然后找出所有单词的词性，对11种常见词性进行one-hot编码(某单词属于该词性则为1，反之为0)。这样就可以获得单词的一个11维的数字特征。此外，还可以定义单词频数也是一种单词属性。单词频数是单词内的5个字母在文件中出现次数之和。最后，对这12个特征计算与PSRTW和相关百分比(1 try,2 tries…)的相关系数矩阵。由相关系数判断相关性。

## The Solution of Model

### 4.3.1 Q1 Segmentation Function Y Model

#### 4.3.1.1增长期模型求解

将数据带入(1)式进行傅里叶(Fourier)拟合。得到函数参数在95%内的置信区间的结果为：









将参数带入公式(1)，即可得到增长期的函数表达式为:



拟合结果如图6所示。

为了衡量该模型的拟合程度并说明此模型最好，我们计算了多个模型的“决定系数（R-Squared）”和“校正决定系数（Adjusted R-Squared）”。用它们对模型进行评估。

不同模型的评价指标如表2所示。可见在前半部分傅里叶(Fourier)拟合模型的R-square 和Adjusted R-Squared的值相较于其他模型更接近于1，拟合效果最佳。

**Table 2: 增长期各模型的R-square和Adjusted R-Squared**

|  |  |  |
| --- | --- | --- |
|  | R-square | Adjusted R-square |
| 傅里叶 | 0.9584 | 0.9535 |
| 高斯 | 0.9471 | 0.9432 |
| 多项式 | 0.9398 | 0.9352 |

#### 4.3.1.2 下降期模型求解

将数据带入(2)式进行指数函数(exponential)拟合。得到函数参数在95%内的置信区间的结果为：





即可得到增长期的函数表达式为



拟合结果如图7所示，不同模型的评价指标R-square 和Adjusted R-Squared的值如表3所示，可见在后半部分指数函数(exponential)拟合模型的R-square 和Adjusted R-Squared的值相较于其他模型更接近于1，表示拟合效果最佳。

**Table 3: 下降期各模型的R-square和Adjusted R-Squared**

|  |  |  |
| --- | --- | --- |
|  | R-square | Adjusted R-square |
| 指数 | 0.9610 | 0.9609 |
|  |  |  |
| Power | 0.9527 | 0.9520 |
| 傅里叶 | 0.9251 | 0.9248 |

|  |  |
| --- | --- |
|  |  |
| Figure 6: 增长期傅里叶拟合 | Figure 7: 下降期的指数拟合 |

#### 4.3.1.3 综合求解结果及预测区间

综上可得所建立的分段函数为：



设实际的“Number of reported result”值为y，预测的“Number of reported result”值为，误差为w，相对误差为，平均相对误差为，即：







计算可得平均相对误差为。这个误差非常小，在可接受的范围之内。

由以上分析可知，当时间为2023年3月1日时，*x*=419，带入公式可得y=1931.1，即2023年3月1日的报告结果数量为1931。另外，模型的相对误差θ=0.316449587，因此可以计算得到报告结果数量的预测区间为





带入y和θ的数值，可以得到计算结果为：=2542.19，=1320。预测得出2023年3月1日的报告结果数量的预测区间为[1320,2542]。

### 4.3.2 Q1 LSTM Model

#### 4.3.2.1 模型实例化

” Number of reported results”作为一组时间序列，它内部隐含了许多关于时间的信息。我们将遵从以下3个步骤将LSTM模型适配SQ2：

1.预处理：将”Number of reported results”序列进行差分后，构造有监督学习数据集。当前项差分为前项差分的label，并正则化到[-1,1]，易于进行LSTM训练。

2.训练：训练得到Q1 LSTM Model。

3.预测：预测需要初始差分输入，我们选定为-824=20380-21204。这是2022年12月31日的单词“manly”和2022年12月30日“molar”的”Number of reported results”的差分结果。输入进LSTM后会输出2023年1月1日的”Number of reported results”。记录下该数值，并与2022年12月31日的单词“manly”的”Number of reported results”(20380)做差分继续输入进LSTM。如此循环60次，可以得到2023年3月1日的”Number of reported results”数值。

针对SQ2的处理流程见图8：

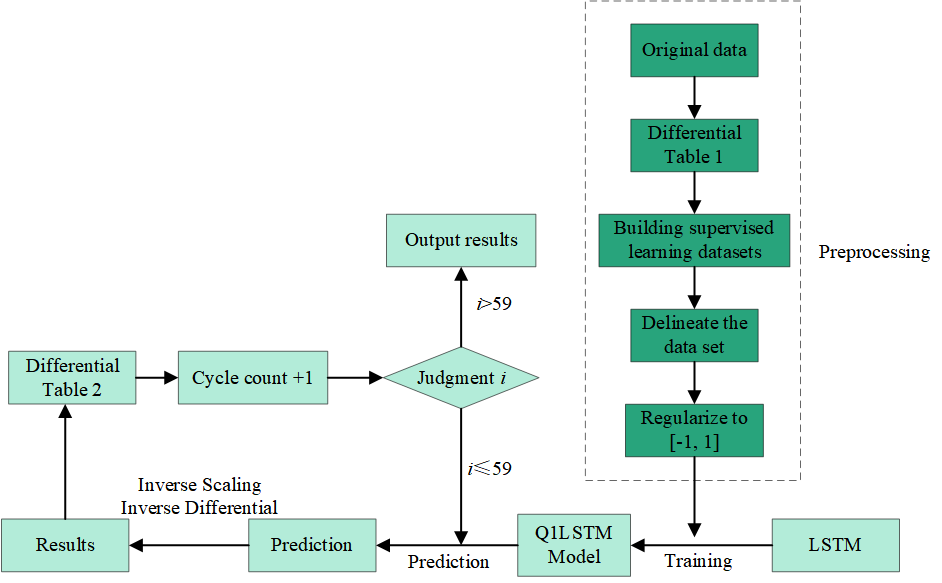


Figure 8: 如何生成Q1 LSTM Model

图8的说明如下：

原始数据：附件中的” Number of reported results”序列。

差分表1：存储原始数据中当前日期减前一天的数据。自2022年1月8日开始，首项差分为101503-80630=20873，以此类推，共358项。

差分表2：根据Results表计算当前项与前项的差分，存储预测结果的差分序列，共60项。差分表2具有首项差分值[1203, -824]，用于启动程序循环。

监督学习数据集：以前项差分为输入，后项差分为输出构造的二维数组。如[ [20873,-10026]，[-10026, 15657]]。

#### 4.3.2.2 结果分析

我们用LSTM模型在测试集上的平均相对误差θ(见公式12 13 14)来描述模型的准确度。通过大量实验对比选择使得平均相对误差θ最小的超参数，并由此计算合适的预测区间(见公式15 16)。实验数据如表4：

**Table 4: lstm超参数与平均相对误差θ对比**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| batch\_size | epoch | Neu\_num | θ | 预测区间 |
| 1 | 2 | 2 | 0.0581 | [4876,5478] |
| 1 | 2 | 3 | 0.0516 | [10052,11146] |
| 1 | 3 | 3 | 0.0608 | [9384, 10598] |
| 1 | 3 | 4 | 0.0512 | [1313, 1454] |
| 1 | 4 | 4 | 0.0563 | [1812,2028] |

可见训练次数(epoch)=3，LSTM层神经元个数(Neu\_num)=4时训练的LSTM θ最小，为5.12%。远小于Q1 Segmentation Function Y Model的31.64%。所以我们选择预测区间[1313, 1454]为最终答案。

图9为此模型在训练集上的表现，可以非常直观地发现拟合效果远强于Q1 Segmentation Function Y Model。

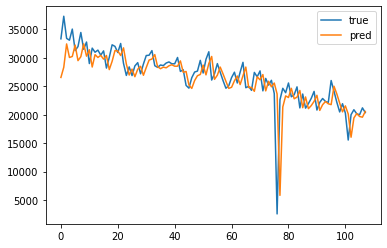


Figure 9: Q1 LSTM Model在测试集上的拟合效果

### 4.3.3 属性与百分比相关性计算与分析

对12个特征计算与PSRTW和相关百分比(1 try,2 tries…)的相关系数矩阵如图10。

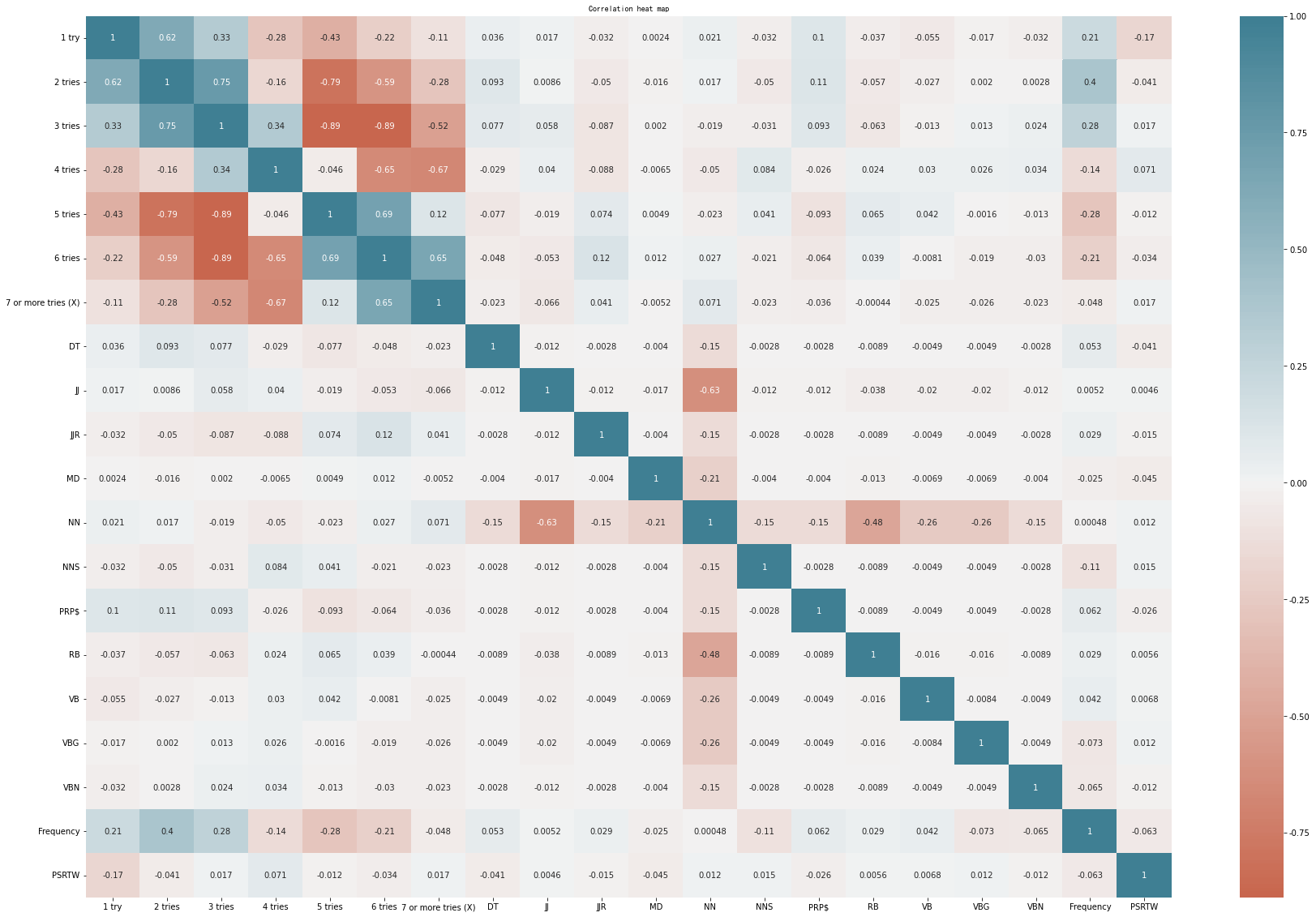


图10 相关性矩阵

通过观察该矩阵，我们可以发现单词属性也就是12个数字特征与PSRTW和相关百分比(1 try,2 tries…)的相关系数非常小，极个别超过0.1。说明我们选择的属性几乎不影响PSRTW和相关百分比(1 try,2 tries…)。

# Task2：建立日期和单词对百分比的关系映射

## Problem Analysis

因为游戏是连续进行字母输入的，我们想到可以将单词拆分为5个字母，并用字母在常用单词中出现的频率高低(图11)作为权值来对单词进行量化[8]。出现频率越高则权值越大，e最大为26，z最小为1。时间序列则处理为单调递增且差值为1的等差数列。我们选用深度学习方法，训练7个模型，用BP神经网络建立输入为日期序列和单词字母与输出为7个百分比之间的联系。由于BP神经网络的初始权重是随机分配的，为了减少训练时间，提升模型的准确度，我们使用遗传算法对其进行优化。

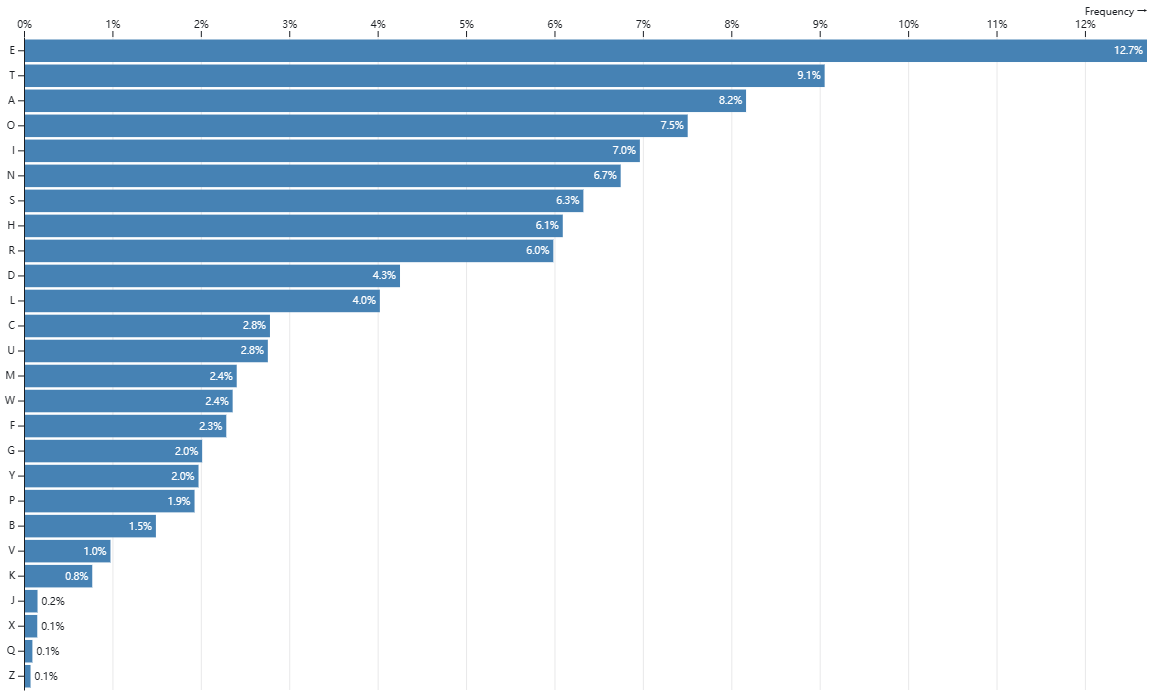


图11 英语中字母使用的相对频率

## The Establishment of Model

### Q2 GA-BP Model

神经网络可以理解为异常复杂以至于无法套用一般化公式的函数[9]。它由神经网络结构和权值阈值确定。引入遗传算法就是为了得到神经网络的最优初始权值阈值。假设权值阈值所对应的所有解是种群，而种群中的个体就是一组权值阈值。遗传算法的工作就是让这个种群中的个体去不断繁衍，在繁衍的过程中一方面会发生染色体交叉而产生新的个体[10]。另一方面，基因变异也有概率会发生并产生新的个体。接下来，通过计算个体的适应度，根据选择算子淘汰质量差的个体，保留质量好的个体，并且让这个繁衍的过程持续下去，那么最后就有可能进化出最优或者较优的个体。

算法总体流程如图12所示。

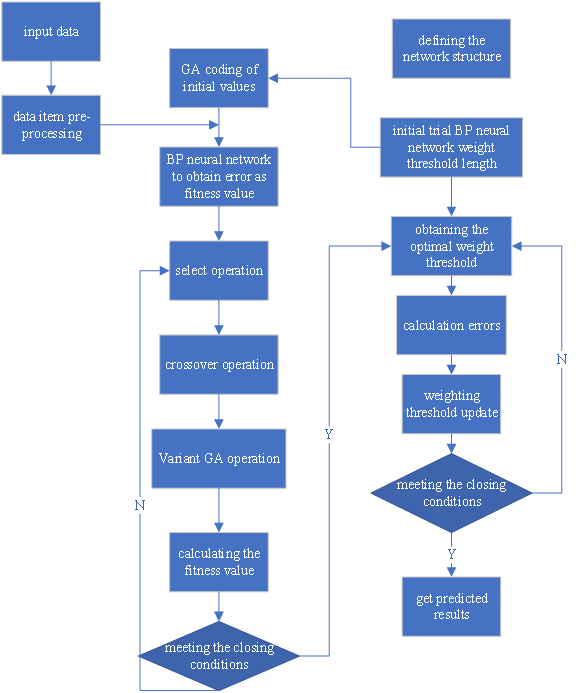


图12: GABP算法流程图

。

## The Solution of Model

### 5.3.1 Q2GA-BP Model

#### 5.3.1.1 模型实例化

**1. 确定染色体长度**

我们选用三层神经网络。现根据图13计算染色体长度：

*n1*，*n2*分别代表输入层与隐藏层神经元个数。其中样本有6个输入参数，1个输出参数，所以这里*n1*=6，由于隐藏层神经元个数与输入层神经元个数有以下的近似公式：



故 *n2*=13。BP神经网络结构：6-13-1 即输入层、隐藏层、输出层的节点数分别为6个、13个、1个。

权值总数：6\*13+13\*1=91个 阈值总数：13+1=14个

遗传算法待优化参数个数(染色体长度)：91+14=105个



图13: BP神经网络参数（染色体形式）

现对遗传算法流程进行如下简要说明：

(1)导入数据，将训练数据与测试数据归一化。确定BP神经网络结构，为6-13-1，所以需要优化的参数总数为6\*13+13\*1+13+1=105，即为染色体长度。

(2)对一个种群所有个体进行初始化，即对每个个体的染色体编码。产生个数为种群大小的一组染色体，染色体存储结构为105维的实数向量。分别计算它们的适应度。并记录下初始种群的平均适应度和最优染色体。

(3)开始进化，对父代进行选择，父代到子代进行交叉，变异操作。解码计算子代每个个体的适应度，将子代适应度最小的个体染色体与父辈的最优染色体适应度进行比较，若小于父辈则替代成为最优染色体，大于则不变。同样记录该代的平均适应度和最优染色体。

(4)继续进化，直到达到预设进化代数，输出最优染色体。即为BP网络最优初始权值阈值组合。

(5)将最优权值阈值分配给BP网络，对输入数据和输出数据进行训练。即进行误差计算，权值更新的有限次循环。设定程序的最大循环次数为200，学习率为0.1，期望误差为0.00001。达到最大循环次数或误差小于期望误差时结束训练。此时的模型即为最优BP模型。

#### 5.3.1.2 结果分析

我们训练了2次，每次训练7个模型(Q2GA-BP Model)分别对2023年3月1日“EERIE”的7个百分比进行预测。Q2GA-BP Model的相对误差见表5，可见除了1 try和7 or more tries (X)平均相对误差θ较差外(Inf,可能是过于离散，无法拟合)，其余模型的误差都在可接受的范围内。最终，我们选择第1次训练模型的预测结果作为最终答案，即2023年3月1日”EERIE”这个词的相关百分比为[0,4,27,33,22,7,3](%)。图14为第2次训练Q2GA-BP Model对5个百分比的拟合效果。

**Table 5 : Q2GA-BP Model的平均相对误差θ和预测结果**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 try | 2 ties | 3 ties | 4 ties | 5 ties | 6 ties | 7 or more tries (X) |
| 第1次θ | Inf | 0.5952 | 0.3134 | 0.1346 | 0.2673 | 0.6369 | Inf |
| 第1次”EERIE”百分比 | 0 | 4 | 27 | 33 | 22 | 7 | 3 |
| 第2次θ | Inf | 0.6716 | 0.3217 | 0.1372 | 0.2852 | 0.7033 | Inf |
| 第2次”EERIE”百分比 | 0 | 6 | 21 | 33 | 23 | 8 | 3 |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |

图14: 第2次训练Q2GA-BP Model对7个百分比的拟合效果

# Task3：建立单词难易程度与单词属性的关系映射

## Problem Analysis

对单词难易程度的一种直观理解是尝试猜谜的次数越多，单词的难度越高。7个尝试次数的相关百分比似乎是一个不错的评价尺度。同时，考虑到本题还需要对”EERIE”进行难度预测，而该单词的尝试次数相关百分比由问题2给出。在问题2中预测的百分比存在误差差异，所以我们选择误差相对较小的3 tires，4tries和5tries作为最终的评价尺度。我们依据这三个指标对所有单词进行类别为4的K-means聚类，得到模型Q3 K-means Model。这个模型会将所有单词分为4类。然而聚类结果并不能说明单词难易程度，只能说明存在不同类别。对聚类结果要加上特定的语义信息才能使之转变为分类结果。所以，我们先找出这4类的聚类中心，通过计算与原点(0,0,0)之间的欧氏距离(欧式距离越大说明更多次尝试的占比越大，单词越难)来规定每种类别的难易程度。由此语义信息把所有单词分为4类，分别为：A-困难，B-较难，C-普通和D-简单。这样每个单词都具备了难易程度标准。Q3 K-means Model与Q3 DT RF Model的训练过程如图15所示，我们将”EERIE”预测的的三个百分比作为进行Q3 K-means Model的输入，可得到”EERIE”的难易程度(图16)。为了讨论方法的准确性，我们提出另一种方法对上述结果进行相互验证。对于词的属性，选择11种常见词性(DT，JJ等，词性只与相关百分比相关性低，与难易程度应该存在较高的相关性)。训练单词词性与难易程度的决策树和随机森林分类器(Q3 DT RF Model) (图15)。对已知词性的”EERIE”进行预测，其预测过程如(图16)所示，最后比较三种模型的预测结果。



图15: Q3 K-means Model与Q3 DT RF Model的训练过程



图16: “EERIE”经Q3 K-means Model与Q3 DT RF Model的预测过程

## The Establishment of Model

### 6.2.1 Q3 K-means Model

K均值聚类（K-means）算法是无监督聚类算法中的代表，其主要作用是将相似的样本自动归到一个类别中[11]。假定给定数据样本*X*，包含了*n*个对象*X*={*X1，X2，X3，…，Xn*}，其中每个对象都具有*m*个维度的属性。K-means算法的目标是将*n*个对象依据对象间的相似性聚集到指定的*k*个类簇中，每个对象属于且仅属于一个到类簇中心距离最小的类簇中。对于K-means，首先需要初始化*k*个聚类中心{*C1，C2，C3，…，Ck*}，1< *k* ≤ *n*，然后通过计算每一个对象到每一个聚类中心的欧式距离，如下式所示



上式中，*Xi*表示第*i*个对象1≤ *t* ≤ *n*，*Cj*表示第*j*个聚类中心的1≤ *j* ≤ *k*，*Xit*表示第*i*个对象的第*t*个属性，1≤ *t* ≤ *m*，*Cjt*表示第*j*个聚类中心的第*t*个属性。

依次比较每一个对象到每一个聚类中心的距离，将对象分配到距离最近的聚类中心的类簇中，得到*k*个类簇{*S1，S2，S3，…，Sk*}。

K-means算法用中心定义了类簇的原型，类簇中心就是类簇内所有对象在各个维度的均值，其计算公式如下



式中，*Cl*表示一个聚类的中心，1≤*l*≤*k*,|*Sl*|表示第1个类簇中对象的个数，*Xi*表示第1个类簇中第*i*个对象，1≤*i*≤|*Sl*|。

## The Solution of Model

### 6.3.1 Q3 K-means Model

#### 6.3.1.1 模型实例化

正如Task 3问题分析所述，我们需要提取每个单词的3 tires，4tries和5tries作为单词难易程度的评价准则并依据这些准则进行类别为4的Kmeans聚类。图17是聚类结果，其中四个聚类中心由大且特殊的形状表示，小圆点代表每个单词，其*x, y, z*坐标分别代表3 tires，4tries和5tries的百分比，颜色代表不同的类别。由于3维图具有固有的深度模糊性，聚类结果并不直观。因此，在图18中对三个百分比字段两两之间显示聚类成果，可以理解为3维图在*x, y, z*平面上的投影。

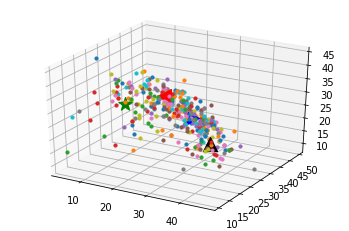


图17: Q3 K-means Model聚类结果3维图

|  |  |  |
| --- | --- | --- |
|  |  |  |

图18: Q3 K-means Model聚类结果2维图

聚类中心的坐标，距原点距离及难易程度见表6。难易程度：A-困难，B-较难，C-普通，D-简单。

**Table 6: 聚类中心信息**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | x | y | z | 距原点的欧氏距离 | 所属类难易程度 |
| 中心1 | 17.309 | 34.907 | 28.762 | 48.429 | C |
| 中心2 | 33.342 | 32.328 | 15.589 | 48.988 | A |
| 中心3 | 12.406 | 24.745 | 28.576 | 39.785 | D |
| 中心4 | 25.492 | 35.5 | 22.092 | 48.971 | B |

由第二问结果可知，“EERIE”的3 tires，4tries和5tries百分比为[27,33,32]。此数据代入Q3 K-means Model输出结果为“EERIE”所属类别为“B”，即“较难”。

### 6.3.2 Q3 DT RF Model

#### 6.3.2.1 模型实例化

研究单词的属性，我们选择11种常见词性(DT，JJ等)作为特征，对单词和词性进行one-hot编码。某个单词具有该词性则为1，否则为0。此外，由Q3 K-means Model我们可以得出每个单词的难易程度。接下来将问题理解为离散数据分类问题。选择决策树的节点分类依据为基尼指数，随机森林里包含10颗决策树[12,13]。训练单词词性与难易程度的决策树和随机森林分类器(Q3 DT RF Model)即可找到词性与难度的对应关系。

“EERIE”的词性one-hot编码为：[0,0,0,0,1,0,0,0,0,0,0]。经Q3 DT RF Model分类结果为：“B”，即“较难”。

**6.3.3 结果分析**

对于两种方法，三类分类器而言，最终的结果都是“B”，即“较难”。这既说明了我们在问题2预测的准确性，又说明在问题3预测的成功。可见我们的模型在根据词性预测难度方面预测比较准确。

# 7 Task4：挖掘其他有意思的信息

我们对该游戏有一个大胆的猜测：Number of reported results很有可能与工作日或者周末有关。因为人们往往在周末有更多时间去专注于现实生活而不是电子产品。

我们对均值进行了分析。发现在整体上也就是所有时间，Number of reported results与周末并无相关(图19)。而自统计之后的150天起(此时已经趋于稳定)，星期几与得分存在较大的负相关关系，且周末（周六周日）的得分明显小于工作日的得分情况(图20)。一种可能的原因是在一段时间后，人们的热情逐渐退却，更倾向于用Wordle在工作日打发时间。

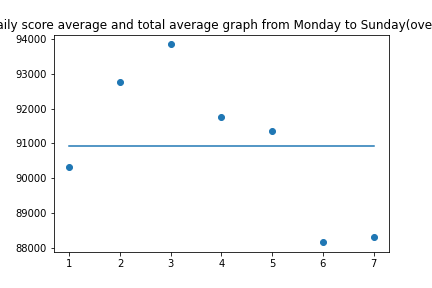
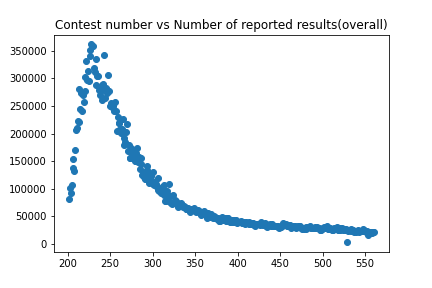


图19: 得分数-编号分布图(整体)与周一到周日每天得分均值与总均值图(整体)

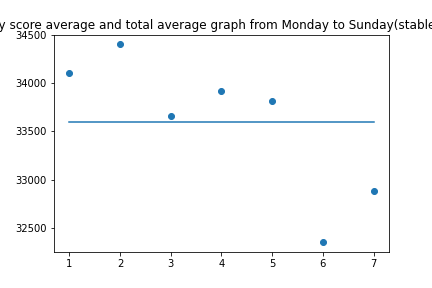
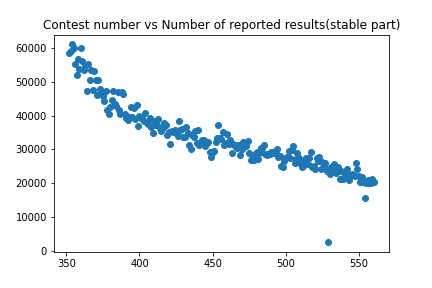


图20: 得分数-编号分布图(稳定部分)与周一到周日每天得分均值与总均值图(稳定部分)

# Error Analysis

我们的解决方案没有考虑游戏规则，单纯认为是填字游戏。问题二仅考虑了人为选取字母的可能性，把这种概率作为权重是不合理的。实际上在Wordle游戏中字母之间存在关联，这可以从颜色方块规则上体会到。单词的量化过程没有体现游戏规则势必会使最终预测结果存在很大的误差。此外，问题三中关于单词属性的选取不应该仅仅拘泥于词性。应该多关注一些数据集的特有特征，如单词频数，单词频率等。词性作为一种属性用11维数据进行描述将会夸大词性对难易程度的影响。为了避免存在过大误差，不应该对词性进行one-hot编码而是计算加权值，进而加入其他属性作为输入，更全面地体会属性对难易程度的影响。

# Model Evaluation and Further Discussion

我们的模型虽然结构简单，易于实现，但是相应地，在使用过程中可能会存在诸多问题。接下来我们将对可能遇到的风险进行讨论与分析。

## Strengths

**Task1**： Q1 Segmentation Function Y Model具有确定的数学表达式使它具有很好的可扩展性，针对未来可能遇到的如求导，积分问题也能完美解决。Q1 LSTM Model时序模型极其擅长处理时间序列。门机制的引入使它们对近期及早期的序列数据也能做到很好的关联。

**Task2**：遗传算法常用于BP神经网络的优化，用于初始权值阈值的选择。这样可以加快BP网络的拟合速度，从而有利于生成更加泛化的模型。对单词的量化依据统计学原理，以字母出现频率作为权值。符合人的对猜字的行为习惯，对具有固定长度的单词而言是一种较好且创新的量化方案。

**Task3**：用两种方案进行对比实验，一方面可验证Q3 K-means Model的聚类效果，另一方面可验证Q2 GA-BP Model的预测准确度。

## Weaknesses

•**对给定数据未进行清洗**。注意到相关百分比之和有的已经超出100%，但是实际上考虑到数据量稀少并未剔除此类数据。在使用中也未进行归一化，直接影响训练所得模型性能。

•**对模型的评价指标单一**。机器学习模型具有多种评价指标，不仅限于平均相对误差。平均相对误差不具有普适性。针对某些特定的模型用其评价有失偏颇。难以说明模型的好坏。

•**选择的模型过于经典固化**。本文中引用的模型大多是经典模型。有的模型过于简单，有的则过于繁杂。简单的模型往往得不到比较标准的结果，复杂的模型则会付出更多的时间。

## Further Discussion

在数据处理之前，一定要先做数据的预处理。如剔除无效值，扩展有效数据等。对训练好的模型，应该采取多种评价指标，对模型进行全方位式的多元评价。对于模型类别的选择，应该关注近些年比较流行的模型。它们大都基于传统模型并在此基础上进行了优化。在处理相同问题的情况下往往具有更优越的性能。总之，应该以更全面的眼光去看待问题，从而达到事半功倍的效果。

**11 Memorandum**

**To：**New York Times Editorial Board

**From：MCM** Team # 2309601

**Subject：**Trend analysis on the digital characteristics of wordle reports

**Date：**February 20, 2023

Dear Sir:

我们是2023年美国大学生数学建模竞赛的一支参赛队伍，针对您提出的问题，我们结合现有的wordle反馈数据，利用合理的数学公式和机器学习模型对其进行了研究，在此基础上借助合理的分析，成功得到了相应的解决方案。以下内容是我们关于本次任务的成果总结。

针对2023年3月1日的报告结果预测区间问题，我们采取了两种解决方案。首先，为了更好地发现数据的变化规律，我们借助工具将表格中的数据可视化。可以发现，数据大致呈现出先上升后下降的趋势，为此我们找到了两种变化规律中的分界点作为阈值日，在此基础上建立了分段函数并利用该函数进行数据拟合，得到预测结果。另外，我们借助时序模型同样获得了更好的结果，最后得到的时序模型(Q1LSTM Model)的相对误差仅为5%，因此我们将[1313,1454]作为最终的预测区间。

针对预测未来某天、某个单词的Percent of try times的问题，我们首先考虑了单词的量化问题。考虑到单词长度相同，我们将单个单词拆解为独立的字母，以字母在常用词汇中出现的频率高低作为权值，以此作为量化手段。在此基础上，利用BP神经网络来建立输入为日期和单词、输出为7个百分比之间的联系，同时借助遗传算法对该模型进行了优化，并得到了该模型(Q2GA-BP Model)的相对误差，结果显示：2023年3月1日单词“EERIE”的相关百分比分别为0%、4%、27%、33%、22%、7%、3%。根据该结果，可以计算得出：大概有63%的概率可以确定我们的预测结果是准确的。

在第三个问题中，我们优先考虑了衡量单词难度的评价指标。我们认为，成功猜出单词的难易程度与猜测次数相关，即猜测次数越多，难度越大。为此，我们在考虑误差的基础上，以问题二中的3 tires，4 tries和5 tries作为评价尺度。接着，利用Kmeans聚类方法对单词进行分类，建立了Q3K-means Model模型，在分类的基础上借助欧式距离来评定单词的难度(难度分别为：A-困难，B-较难，C-普通和D-简单)。最后，我们利用Q3K-means Model得出单词“EERIE”的难度为：B-较难。另外，同时采用决策树和随机森林分类器（Q3DT Q3RF Model）对词性确定的“EERIE”进行预测，二者结果均为：B-较难。与Q3K-means Model的预测结果吻合，进一步验证了结论的准确性。

关于给定数据集的特征描述问题，我们基于先验知识，猜测当日的Number of reported results应该与工作日和非工作日相关。然而，均值分析结果表明Number of reported results与是否为工作日没有直接关联，但进一步观察之后150天的得分趋势可以发现，一周中的某天与得分呈负相关，其中工作日的得分明显高于非工作日。以上分析说明，人们对于工作日和非工作日间的游戏时间分配存在较大差异，且在工作日利用wordle进行休闲和放松的人相对更多。造成这种差异的原因可能是随着时间推移，人们对wordle的看法由火爆的小游戏变成了工作间隙消磨时间的小工具。

以上就是我们针对贵社所提问题的解决方案和成果总结，感谢您抽出宝贵的时间查阅，如果您需要更多关于我们研究结果的信息，请随时与我们联系。

Sincerely,

Team #2309601

# References

[1] “wordle-The New York Times.” The New York Times, 2022. Accessed December 13, 2022 at <http://www.nytimes.com/games/wordle/index.html>.

[2] Chuang-Chun Liu. Understanding player behavior in online games: The role of gender[J]. Technological Forecasting & Social Change, 2016:265-274.

[3] Jeon JiHoon, et al. Extracting gamers' cognitive psychological features and improving performance of churn prediction from mobile games[C]//Computational Intelligence & Games. IEEE, 2017.

[4] Kim Seungwook, et al. Churn prediction of mobile and online casual games using play log data[J]. PloS one, 2017, 12(7):e0180735.

[5] Periáñez África, et al. Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles[C]//2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2016.

[6] Vafeiadis T, Diamantaras K I , Sarigiannidis G , et al. A comparison of machine learning techniques for customer churn prediction[J]. Simulation Modelling Practice & Theory, 2015, 55:1-9.

[7] Zha W, Liu Y, Wan Y, et al. Forecasting monthly gas field production based on the CNN-LSTM model[J]. Energy, 2022: 124889.

[8] Pande H. Mathematical modeling of the frequencies of letters for their occurrence in corpora, words (types) and in the initial positions of words of corpora[J]. Glottotheory, 2021, 12(1): 57-69.

[9] Han J X, Ma M Y, Wang K. Product modeling design based on genetic algorithm and BP neural network[J]. Neural Computing and Applications, 2021, 33: 4111-4117.

[10] Katoch S, Chauhan S S, Kumar V. A review on genetic algorithm: past, present, and future[J]. Multimedia Tools and Applications, 2021, 80: 8091-8126.

[11] Ghazal T M, Hussain M Z, Said R A, et al. Performances of K-Means Clustering Algorithm with Different Distance Metrics[J]. Intelligent Automation & Soft Computing, 2021, 30(2).

[12] Yariyan P, Janizadeh S, Van Phong T, et al. Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping[J]. Water Resources Management, 2020, 34: 3037-3053.

[13] Speiser J L, Miller M E, Tooze J, et al. A comparison of random forest variable selection methods for classification prediction modeling[J]. Expert systems with applications, 2019, 134: 93-101.