2023 MCM
Problem C: Predicting Wordle Results

曾才斌
E-mail: macbzeng@scut.edu.cn

华南理工大学 数学学院

2025 年 1 月 7 日

目录

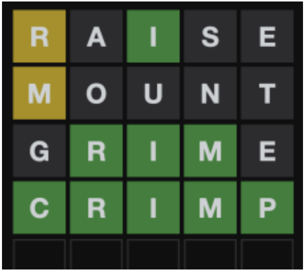
1. 问题背景与重述
2. 数据预处理
3. 模型建立与求解
4. 必要内容

问题背景

《纽约时报》提供的每日猜词游戏–Worldle。

游戏规则

- 基本规则
  1. 黄色：字母正确，位置错误
  2. 绿色：字母正确，位置正确
  3. 灰色：字母不存在
- 困难模式附加规则
  4. 对于猜对的字母（黄色或绿色），后续猜测必须使用



R A I S E
M O U N T
G R I M E
C R I M P

数据溯源 ＋ 猜词游戏

问题重述

1. **每日报告结果数量的变化**：开发一个模型来解释报告结果数的日变化，并使用该模型为 2023 年 3 月 1 日预测报告结果数的区间。此外，还需要调查单词的属性是否影响在困难模式下报告的分数百分比，如果是，如何影响；如果不是，为什么不影响。
2. **预测未来单词结果分布**：为一个给定未来日期的解决方案单词开发一个模型，以预测报告结果的分布，即预测未来日期的（1、2、3、4、5、6、X）相关百分比。需要说明模型和预测的不确定性，并为 2023 年 3 月 1 日的单词 EERIE 给出具体的预测示例，以及对模型预测的置信程度。
3. **按难度分类解决方案单词**：开发并总结一个模型，按难度分类解决方案单词。确定与每个分类相关的单词属性。使用模型判断单词 EERIE 的难度，并讨论分类模型的准确性。
4. **描述数据集的其他有趣特征**：列出并描述数据集的一些其他有趣特征。

核心

预测模型里**不确定性**做出**量化和分析**

Outline

1 问题背景与重述

2 **数据预处理**

3 模型建立与求解

4 必要内容

数据预处理的常用方法

1. 缺失数据
   - 均值填充法：对于年龄、距离等数值型数据，采用平均值方法；
     对于性别、类别等非数值型数据，采用众数方法。
   - 就近补齐法：在完整数据中找到一个与它最相似的对象的值来进行填充。
   - 聚类填充法：经典的聚类算法是 K-近邻算法。
   - 回归方程法：用不含缺失值的数据集建立回归方程来预测缺失值。
2. 异常数据
   - 检测方法：常识、统计学原理、箱线图法。
   - 处理方法：直接删除、使用缺失值处理方法来处理异常值。

原始单词	校正单词	位置
rprobe	probe	545
clen	clean	525
tash	trash	314

Outline

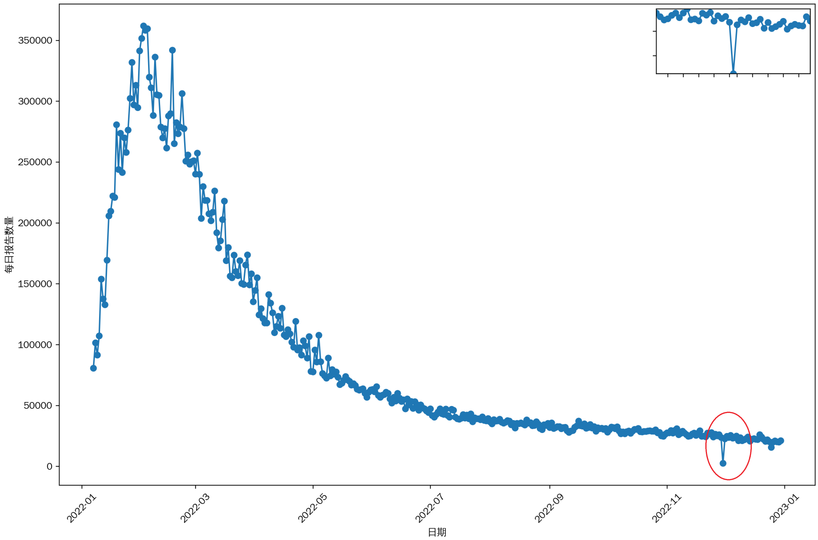
1 问题背景与重述

2 数据预处理

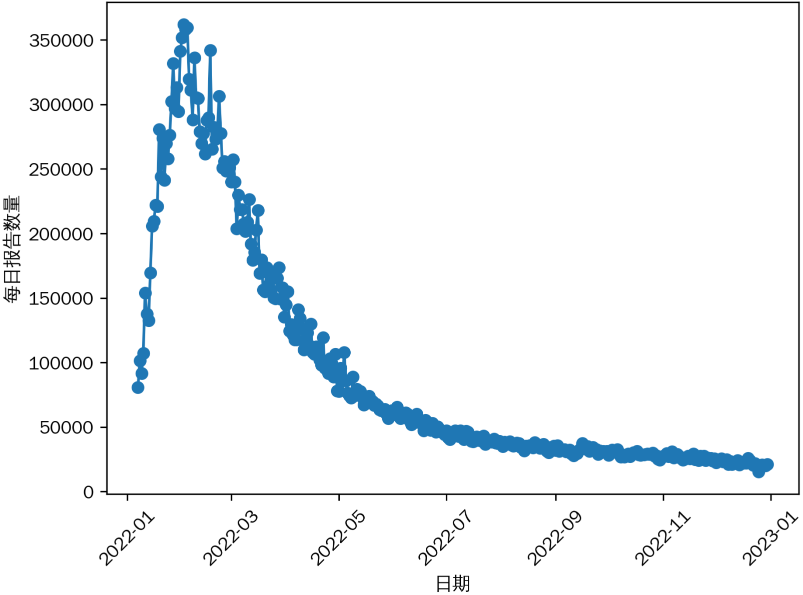
3 **模型建立与求解**

4 必要内容

问题 1：每日报告结果数量的变化



问题 1：每日报告结果数量的变化



问题 1：每日报告结果数量的变化

基本问题

时间序列预测问题：利用历史数据来预测未来值。

常见模型

- 自回归模型 (AR)
- 移动平均模型 (MA)
- 自回归综合移动平均模型 (ARIMA)
- 季节性自回归综合移动平均模型 (SARIMA)
- 指数平滑模型
- Facebook 开发的 Prophet 模型
- 长短期记忆网络模型 (LSTM)
- 门控循环单元模型 (GRU)
- 向量自回归模型 (VAR)
- 状态空间模型
- 卡尔曼滤波器

ARIMA(p, d, q) 模型的典型步骤

1. 平稳性检验
2. 差分处理，确定差分的阶数 d
3. 模型定阶，确定 AR 阶数 p，MA 阶数 q
4. 模型检验（残差检验 + 相关性检验）
5. 模型预测

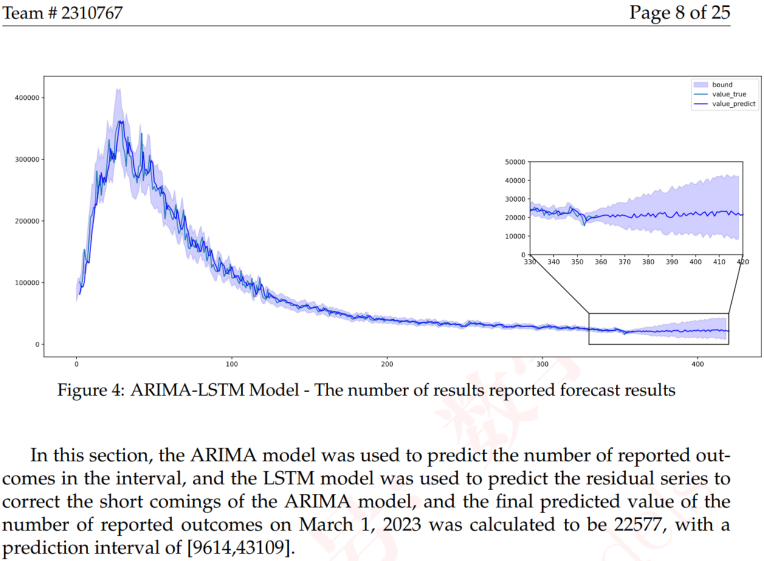
创新点

- 自相关函数 + 偏自相关函数 $\longrightarrow$ ADF 检验或 KPSS 检验
- AIC 准则 $\longrightarrow$ BIC 准则或 AIC-BIC 准则
- 残差检验 $\longrightarrow$ LSTM 模型

关键点

- 明确的预测至及预测区间
- 不确定性或置信水平的有效度量

结果展示 1 (ARIMA-LSTM 模型)

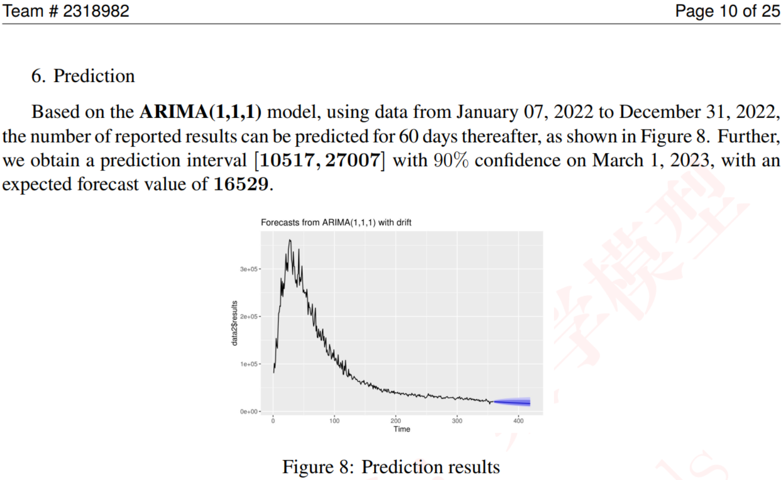


Team # 2310767 Page 8 of 25

Figure 4: ARIMA-LSTM Model - The number of results reported forecast results

In this section, the ARIMA model was used to predict the number of reported outcomes in the interval, and the LSTM model was used to predict the residual series to correct the short comings of the ARIMA model, and the final predicted value of the number of reported outcomes on March 1, 2023 was calculated to be 22577, with a prediction interval of [9614,43109].

结果展示 2 (ARIMA 模型)



Team # 2318982 Page 10 of 25

6. Prediction

Based on the **ARIMA(1,1,1)** model, using data from January 07, 2022 to December 31, 2022, the number of reported results can be predicted for 60 days thereafter, as shown in Figure 8. Further, we obtain a prediction interval $[\mathbf{10517}, \mathbf{27007}]$ with 90% confidence on March 1, 2023, with an expected forecast value of $\mathbf{16529}$.

Figure 8: Prediction results

问题 1：困难模式报告百分比的单词属性

基本问题

属性、特征、度量定义问题：开放且主观。

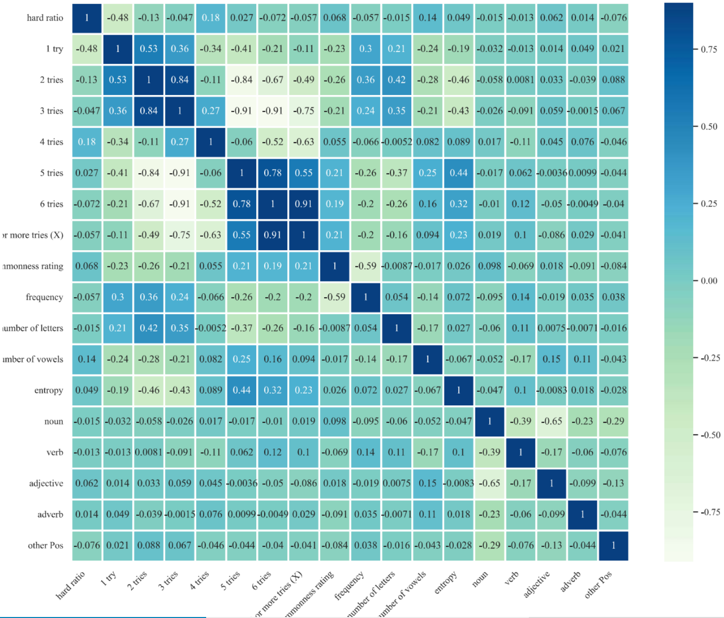
建议的属性

- 音节数量 (number of syllables)
- 单词类型 (word class)
- COCA 指数 (Corpus of Contemporary American English index)
- 字母数量 (number of letters)
- 词频 (word frequency)
- 信息熵 (information entropy)
- 情感认知性 (emotion subjectivity)

关键点

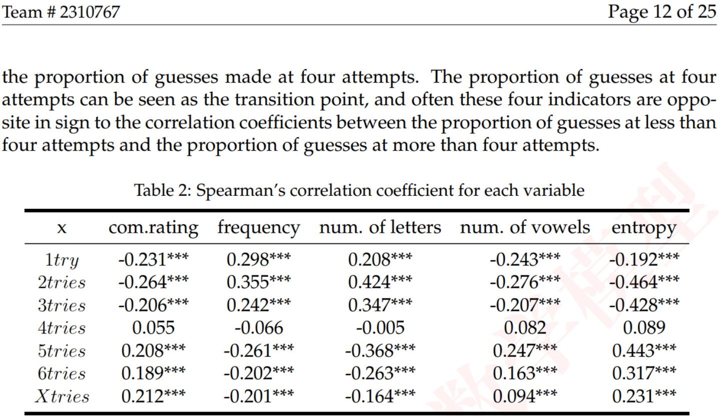
- 合理的属性定义及量化
- 结果的可视化：相关矩阵、热图等

结果展示 3 (热图)



	hard ratio	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	r more tries (X)	nmonness rating	frequency	umber of letters	umber of vowels	entropy	noun	verb	adjective	adverb	other Pos
hard ratio	1	-0.48	-0.13	-0.047	0.18	0.027	-0.072	-0.057	0.068	-0.057	-0.015	0.14	0.049	-0.015	-0.013	0.062	0.014	-0.076
1 try	-0.48	1	0.53	0.36	-0.34	-0.41	-0.21	-0.11	-0.23	0.3	0.21	-0.24	-0.19	-0.032	-0.013	0.014	0.049	0.021
2 tries	-0.13	0.53	1	0.84	-0.11	-0.84	-0.67	-0.49	-0.26	0.36	0.42	-0.28	-0.46	-0.058	0.0081	0.033	-0.039	0.088
3 tries	-0.047	0.36	0.84	1	0.27	-0.91	-0.91	-0.75	-0.21	0.24	0.35	-0.21	-0.43	-0.026	-0.091	0.059	-0.0015	0.067
4 tries	0.18	-0.34	-0.11	0.27	1	-0.06	-0.52	-0.63	0.055	-0.066	-0.0052	0.082	0.089	0.017	-0.11	0.045	0.076	-0.046
5 tries	0.027	-0.41	-0.84	-0.91	-0.06	1	0.78	0.55	0.21	-0.26	-0.37	0.25	0.44	-0.017	0.062	-0.0036	0.0099	-0.044
6 tries	-0.072	-0.21	-0.67	-0.91	-0.52	0.78	1	0.91	0.19	-0.2	-0.26	0.16	0.32	-0.01	0.12	-0.05	-0.0049	-0.04
r more tries (X)	-0.057	-0.11	-0.49	-0.75	-0.63	0.55	0.91	1	0.21	-0.2	-0.16	0.094	0.23	0.019	0.1	-0.086	0.029	-0.041
mmonness rating	0.068	-0.23	-0.26	-0.21	0.055	0.21	0.19	0.21	1	-0.59	-0.0087	-0.017	0.026	0.098	-0.069	0.018	-0.091	-0.084
frequency	-0.057	0.3	0.36	0.24	-0.066	-0.26	-0.2	-0.2	-0.59	1	0.054	-0.14	0.072	-0.095	0.14	-0.019	0.035	0.038
umber of letters	-0.015	0.21	0.42	0.35	-0.0052	-0.37	-0.26	-0.16	-0.0087	0.054	1	-0.17	0.027	-0.06	0.11	0.0075	-0.0071	-0.016
mber of vowels	0.14	-0.24	-0.28	-0.21	0.082	0.25	0.16	0.094	-0.017	-0.14	-0.17	1	-0.067	-0.052	-0.17	0.15	0.11	-0.043
entropy	0.049	-0.19	-0.46	-0.43	0.089	0.44	0.32	0.23	0.026	0.072	0.027	-0.067	1	-0.047	0.1	-0.0083	0.018	-0.028
noun	-0.015	-0.032	-0.058	-0.026	0.017	-0.017	-0.01	0.019	0.098	-0.095	-0.06	-0.052	-0.047	1	-0.39	-0.65	-0.23	-0.29
verb	-0.013	-0.013	0.0081	-0.091	-0.11	0.062	0.12	0.1	-0.069	0.14	0.11	-0.17	0.1	-0.39	1	-0.17	-0.06	-0.076
adjective	0.062	0.014	0.033	0.059	0.045	-0.0036	-0.05	-0.086	0.018	-0.019	0.0075	0.15	-0.0083	-0.65	-0.17	1	-0.099	-0.13
adverb	0.014	0.049	-0.039	-0.0015	0.076	0.0099	-0.0049	0.029	-0.091	0.035	-0.0071	0.11	0.018	-0.23	-0.06	-0.099	1	-0.044
other Pos	-0.076	0.021	0.088	0.067	-0.046	-0.044	-0.04	-0.041	-0.084	0.038	-0.016	-0.043	-0.028	-0.29	-0.076	-0.13	-0.044	1

结果展示 4 (相关系数)



the proportion of guesses made at four attempts. The proportion of guesses at four attempts can be seen as the transition point, and often these four indicators are opposite in sign to the correlation coefficients between the proportion of guesses at less than four attempts and the proportion of guesses at more than four attempts.

Table 2: Spearman's correlation coefficient for each variable

x	com.rating	frequency	num. of letters	num. of vowels	entropy
$1try$	-0.231***	0.298***	0.208***	-0.243***	-0.192***
$2tries$	-0.264***	0.355***	0.424***	-0.276***	-0.464***
$3tries$	-0.206***	0.242***	0.347***	-0.207***	-0.428***
$4tries$	0.055	-0.066	-0.005	0.082	0.089
$5tries$	0.208***	-0.261***	-0.368***	0.247***	0.443***
$6tries$	0.189***	-0.202***	-0.263***	0.163***	0.317***
$Xtries$	0.212***	-0.201***	-0.164***	0.094***	0.231***

问题 2：预测未来单词结果分布

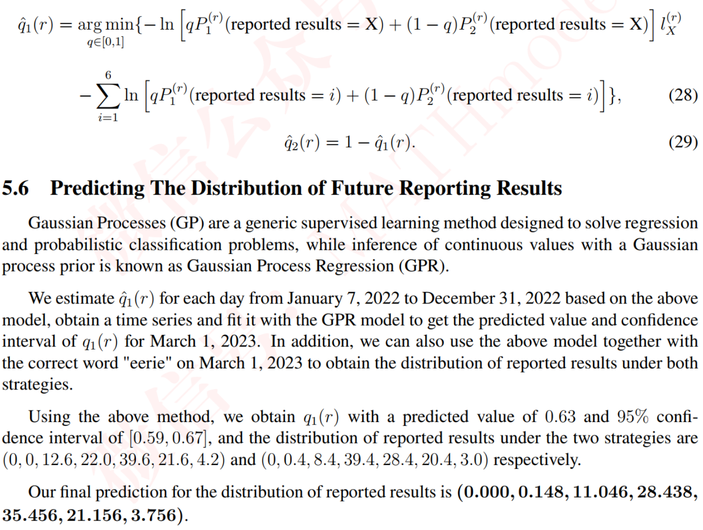
基本问题

回归分类问题

常用模型

- Ridge 回归
- Lasso 回归
- Gauss 过程回归
- XGBoost 决策树模型
- LightGBM 决策树模型
- Markov 链模型
- 随机森林搜索算法

结果展示 5 (Markov 链模型)



$$\hat{q}_1(r) = \arg\min_{q\in[0,1]}\{-\ln\left[qP_1^{(r)}(\text{reported results} = \text{X}) + (1-q)P_2^{(r)}(\text{reported results} = \text{X})\right] l_X^{(r)}$$

$$-\sum_{i=1}^{6}\ln\left[qP_1^{(r)}(\text{reported results} = i) + (1-q)P_2^{(r)}(\text{reported results} = i)\right]\}, \tag{28}$$

$$\hat{q}_2(r) = 1 - \hat{q}_1(r). \tag{29}$$

5.6 Predicting The Distribution of Future Reporting Results

Gaussian Processes (GP) are a generic supervised learning method designed to solve regression and probabilistic classification problems, while inference of continuous values with a Gaussian process prior is known as Gaussian Process Regression (GPR).

We estimate $\hat{q}_1(r)$ for each day from January 7, 2022 to December 31, 2022 based on the above model, obtain a time series and fit it with the GPR model to get the predicted value and confidence interval of $q_1(r)$ for March 1, 2023. In addition, we can also use the above model together with the correct word "eerie" on March 1, 2023 to obtain the distribution of reported results under both strategies.

Using the above method, we obtain $q_1(r)$ with a predicted value of 0.63 and 95% confidence interval of $[0.59, 0.67]$, and the distribution of reported results under the two strategies are $(0, 0, 12.6, 22.0, 39.6, 21.6, 4.2)$ and $(0, 0.4, 8.4, 39.4, 28.4, 20.4, 3.0)$ respectively.

Our final prediction for the distribution of reported results is $\mathbf{(0.000, 0.148, 11.046, 28.438, 35.456, 21.156, 3.756)}$.

问题 3：按难度分类解决方案单词

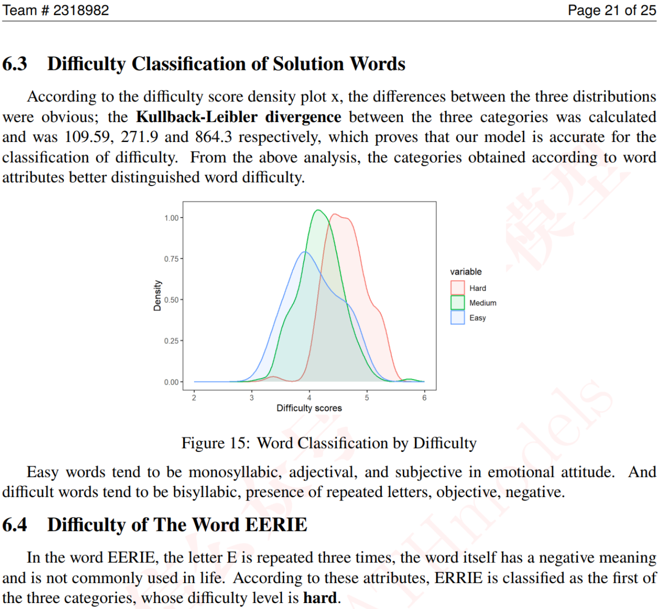
基本问题

聚类问题

常用算法

- 基于划分的聚类方法：K-Means++、K-Medoids
- 基于层次的聚类方法：凝聚层次聚类、分裂层次聚类
- 基于密度的聚类方法：DBSCAN 算法、OPTICS 算法
- 基于网格的聚类方法：STING 算法、CLIQUE 算法
- 基于模型的聚类方法：Gauss 混合模型、模糊 C-Means 算法

结果展示 6 (K-Means++ 算法)



6.3 Difficulty Classification of Solution Words

According to the difficulty score density plot x, the differences between the three distributions were obvious; the **Kullback-Leibler divergence** between the three categories was calculated and was 109.59, 271.9 and 864.3 respectively, which proves that our model is accurate for the classification of difficulty. From the above analysis, the categories obtained according to word attributes better distinguished word difficulty.

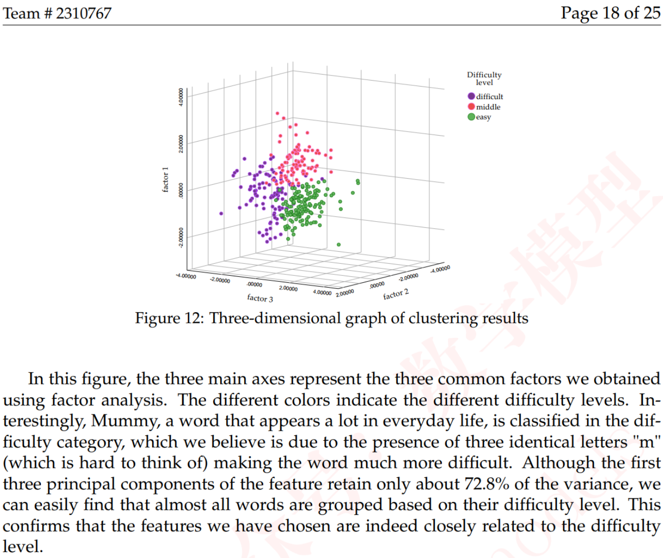
Figure 15: Word Classification by Difficulty

Easy words tend to be monosyllabic, adjectival, and subjective in emotional attitude. And difficult words tend to be bisyllabic, presence of repeated letters, objective, negative.

6.4 Difficulty of The Word EERIE

In the word EERIE, the letter E is repeated three times, the word itself has a negative meaning and is not commonly used in life. According to these attributes, ERRIE is classified as the first of the three categories, whose difficulty level is **hard**.

结果展示 7 (Gauss 混合模型)



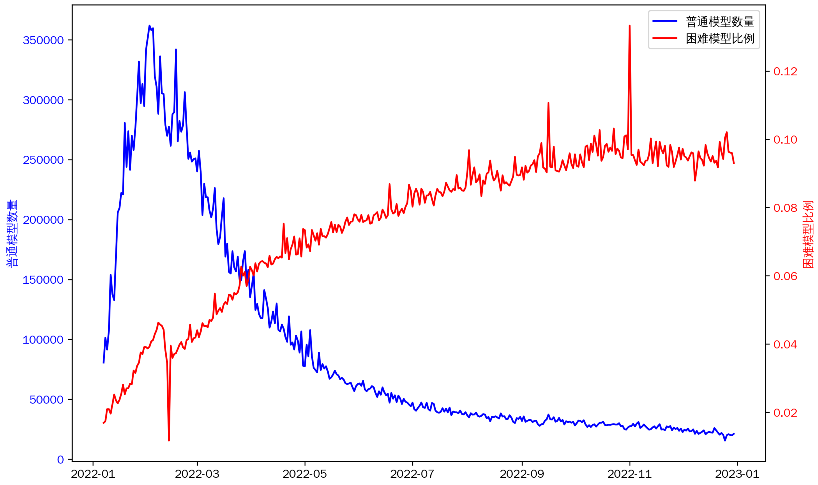
Team # 2310767 Page 18 of 25

Difficulty level
● difficult
● middle
● easy
factor 1
4.00000
2.00000
.00000
-2.00000
factor 3
-4.00000
-2.00000
.00000
2.00000
4.00000
factor 2
2.00000
.00000
-2.00000
-4.00000

Figure 12: Three-dimensional graph of clustering results

In this figure, the three main axes represent the three common factors we obtained using factor analysis. The different colors indicate the different difficulty levels. Interestingly, Mummy, a word that appears a lot in everyday life, is classified in the difficulty category, which we believe is due to the presence of three identical letters "m" (which is hard to think of) making the word much more difficult. Although the first three principal components of the feature retain only about 72.8% of the variance, we can easily find that almost all words are grouped based on their difficulty level. This confirms that the features we have chosen are indeed closely related to the difficulty level.

问题 4：描述数据集的其他有趣特征（数量变化趋势）



问题 4：描述数据集的其他有趣特征（字母在不同位置上的频率）

Team # 2310767 Page 22 of 25

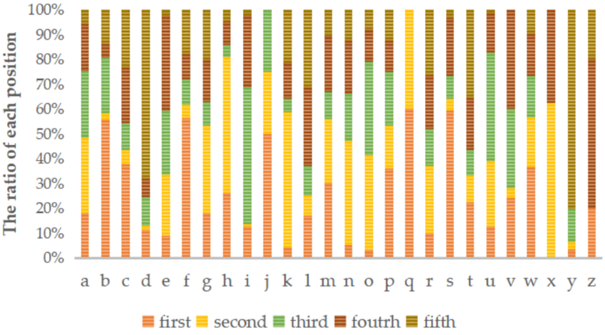


Figure 19: Proportion of the 26 letters appearing in words

问题 4：描述数据集的其他有趣特征（最困难词）

1. parer：X 的比例为 0.48，可能原因是它不是一个常见的英语单词，而是拉丁语 parare 衍生出的单词。
2. foyer：X 的比例为 0.26，可能原因是它不是一个常见的英语单词，而是法语 fouyer 衍生出的单词。

Outline

1 问题背景与重述

2 数据预处理

3 模型建立与求解

4 必要内容

其他必要的组成部分

1. 模型假设
2. 符号定义
3. 敏感性分析
4. 模型优缺点分析
5. 参考文献
6. 建议信

博學慎思明辨篤行

華南理工大學校訓

葉選平敬書

二〇〇五年七月