

利用语言学特征分析Wordle的难度，确定Twitter^{玩家}的平均成功率*。

伊万-李

2022年5月7日

摘要

利用Twitter的API，收集了记录Wordle玩家在一个半月内的公开得分的数据。根据成功的程度、地理位置和单词的语言属性，获得并分析了每天得分的频率。研究发现，单词的语言属性，如共性，对玩家的成功有影响。这些结果可能在文学和语言学分析中具有意义。

内容

简介	1
数据	2
数据集	2
变量	2
缺失的数据.....	3
地块	3
模型	10
特点	11
模型关注的问题.....	12
结果	12
建模平均得分	12
建立失败率模型.....	13
讨论	14
数据和模型的发现	14
弱点和下一步措施.....	15
附录	16
数据表	16
参考文献	22

简介

Wordle是一个在线猜词游戏，与历史上流行的文字谜题相似，如《纽约时报》的填字游戏和棋盘游戏Mastermind。玩家每天会得到一个五个字母的单词，并被要求在六次尝试中猜出该单词。如果在前面的尝试中没有猜出这个词。

—*代码和数据见：<https://github.com/Ivannoar/Twilight>。

你的单词的字母会变成灰色、黄色或绿色：灰色表示单词中没有找到该字母，黄色表示有该字母但位置不对，绿色表示字母和位置都正确。游戏的目标是猜出五个绿色字母，从而猜出正确的单词。该游戏由软件工程师乔希·沃德尔（Josh Wardle）创造，他为自己和伙伴创造了这个游戏，以便在2021年COVID-19大流行期间打发时间。他的作品受欢迎程度急剧上升，这要归功于一个功能，即玩家可以轻松复制和粘贴他们的结果，与他们的朋友和家人分享。沃德尔的游戏在2021年12月底找到了国际传播，这要归功于多种因素，包括分享结果的便利性和游戏的日常性，以及其简单性。Wordle的受欢迎程度在Twitter上可以看到得淋漓尽致，每天有数百万用户通过网站的内置功能粘贴他们的分数来分享他们的结果。

本文旨在分析Wordle在Twitter上流行的几个指标。由于每天有数以百万计的用户分享他们的分数，因此可以而且已经对包括平均分数在内的变量进行了分析--最近的一项研究发现，瑞典是平均猜测次数最低的国家，在最多6次的猜测中只有3.72次（“世界上哪里的人最擅长解决Wordle？”^{n.d}）。在这里，我们希望寻求的一些关键发现包括尝试建立一个单词的某些特征是否与它的平均难度相关的模型，以及Twitter用户的平均得分与他们的地理位置，如他们的居住国有关。这些指标在之前的论文中并没有深入讨论，可能会引起未来谜题制作者、语言学家和游戏理论家的兴趣。

利用Twitter的API，收集了与Wordle有关的推文，并将其汇总为多个数据集，用于分析。这些推文是在2022年4月4日至10日和2022年4月17日至23日的间隔期间收集的）。用户的分数是通过搜索网站领域所使用的共同模板来收集的，这些模板在他们的推文中分享分数。除了所使用的分数外，还获得了发布时间、地理位置和其他由Twitter提供的数据。这些结果被转化为图形数据，并根据我们的关键任务进行分析。

在数据部分，我们讨论了从Twitter API收集的数据，以及如何清理和安排这些数据以进行适当的统计分析。我们讨论了使用的变量、方法和显示数据分布的图表。在模型部分，我们讨论了我们的模型和它对我们如何解释我们的结果的影响。在结果部分，根据我们在分析过程中发现的关键目标和结论，显示并使用图形数据来展示我们的故事。最后，我们在讨论部分讨论了所做的工作、其意义以及本文的不足之处。

数据

数据集

为了完成本文所设定的目标，所使用的数据由社交媒体网站Twitter的推文样本组成。Twitter托管并存储了大量的信息，可以通过网站的API进行访问；但是，在访问过程中，除非得到网站的许可，否则对推文数据的访问仅限于过去7天。本文收集的原始数据包含504000条推文和90个关于推文及其发件人的各种属性的变量，如消息、发布日期和时间、发件人账户的细节、流行度量以及推文的在线直接链接。推文是在两个时间段内搜索和收集的：4月4日至10日的推文，以及4月17日至23日的推文。这是用编程软件R（R核心团队2020）和推文收集包rtweet（Kearney 2019）完成的。关于如何获得和清理数据的完整细节，可以在附录中的数据表格中找到。

变量

数据集由90个变量组成，这些变量需要被精简到一个可管理的相关数据列表中。这些变量包括独特的推特标识符和所包含的确切文本，有关推特是原创还是作为回复或回应另一个推特的细节，初始或回应帖子的喜欢和转发数量，用于发送帖子的设备，以及每个观察的帖子和账户的链接。我们只寻求利用用户在游戏中取得的分数、他们的国家/地点、游戏的日期、以及他们在游戏中的表现。

表1:显示Wordle相关推文的数据集的前十行

日期	国家	词条得分	哈德莫德
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。		
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。		
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。	0	自行
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。		
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。	0	六四
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。		
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。	4	自行
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。		
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。	0	六四
2022-04-04	本报讯："我们的目标是，在未来的日子里，让我们的生活更美好。	5	暂停

信息，以及在玩的时候存在'困难模式'，这就增加了一个条件，即你必须使用你从先前的猜测中得到的提示。为了达到这个目的，需要对变量进行改变和操作。从使用Unix时间来存储时间数据的原始数据中，使用tidyverse（Wickham等人，2019）和lubridate（Grolemund和Wickham，2011）软件包来创建新的变量。从Unix时间中，日期被从原始数据中提取出来，并以YYY-MM-DD格式存储。创建了一个与用户是否使用硬模式相对应的指标变量，并根据其推文的内容给每个观察点进行分类。最终的数据框架包含4个变量，这些变量是每个推文实例的日期、国家、分数和游戏模式。

缺失的数据

本文收集的数据有一定的局限性，这将对我们的结论产生影响。由于Twitter的规定和所提供的API的限制，历史数据无法用于分析，因为该网站只允许大量收集过去一周的推文。本文对无法进行长期的历史分析表示遗憾。此外，由于rtweet软件包只能批量收集最近的推文，因此无法处理按时间顺序排列的数据。这样做的结果是，由于硬件的限制和可行性，无法在很长的时间间隔内持续收集数据。虽然数据不应该受到负面影响，但它不可能正确地进行时间顺序分析，而且由于大部分的观察都是在一个时间段进行的，所以仍然有可能出现时间间隔的偏差。

由于推文的性质，数据集中也有固有的偏见。并非所有的Wordle玩家都在网上报告他们的结果，在Twitter上的子集更少，这就使得所得出的结论可能并不适用于所有Wordle玩家的一般人群。充其量，我们可以从Twitter用户的样本中确认我们的结论，并试图推断到更大的人群。由于Wordle是一个基于英语的词，地理数据很可能也会偏向于英语国家，因为不是以英语为主要语言的国家可能有少得多的Wordle玩家，考虑到数据集的性质，这有一个放大的影响。更多的偏见存在于那些经历了较低分数（较高的平均猜字数）或未能完成游戏的用户可能不会在Twitter上发布这些结果。所有收集到的数据都必须是自我报告的，这就导致了非应答性的偏差。用户也可能通过改变他们报告的分数或模仿原始Wordle信息来说谎或提供虚假数据，这可能影响我们的分布形状。

地块

为了熟悉所获得的数据并为进一步的分析提供动力，我们在经过清理的数据集中创建了与所选变量相对应的图。首先，我们希望绘制整个数据集中每个结果/分数的频率。我们比较每个结果的频率，并将其绘制为原始频率和每个分数在我们的数据集中出现的百分比。请注意，失败，即用户在最多6次尝试中无法猜出这个词，被绘制为7分。我们观察到

大多数Twitter用户在4次尝试中完成了游戏，而且猜测的分布是合理的单模态的。该分布略微偏右，数据集显示，失败的人数是第一次尝试就猜中单词的人数的三倍多，这是预期的，因为与失败的机会相比，在没有先验信息的情况下，猜中准确单词的概率很低。这也反映在2分和6分之间的比例上。

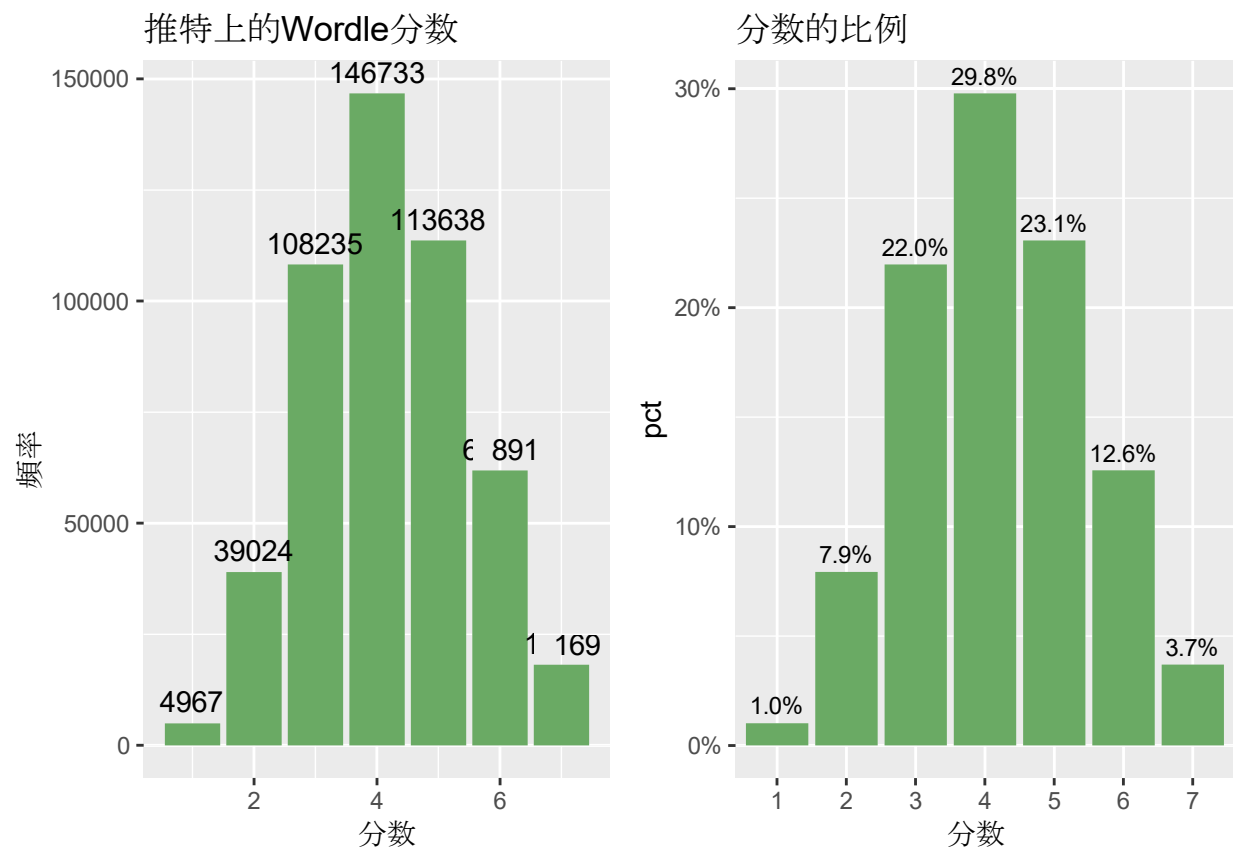


图1：Twitter Wordle玩家的得分情况

接下来，我们要比较使用硬模式的用户和不使用硬模式的用户之间的分布情况。必须指出的是，不启用硬模式的人的数量与启用硬模式的人相比相形见绌。然而，不报告使用硬模式的人中，有可能有不重要的大多数人是有意或无意地将硬模式条件强加给自己。无论如何，我们绘制了硬模式用户与正常用户的比例，看到我们的样本中94%的用户不使用硬模式，这可以解释为硬模式是选择进入的，用户可能出于多种原因不愿意费心启用它。然后，我们重复前一组图表中的分数分布，但按用户的游戏模式分开结果。从我们的图表中，我们看到我们的样本玩家在硬模式下更经常达到3和4分，但也比普通玩家有更高的失败率。普通玩家使用的策略是无视前2或3个单词的提示，在作出现实的猜测之前，尽量“使用”尽可能多的普通字母来揭示单词中尽可能多的字母，而硬模式玩家必须使用他们得到的提示，这降低了策略的效率，但意味着他们使用的每个单词都有相对较高的正确率。然而，当难以使用所提供的提示或该词与其他常见的词有相同的结构时，这就会遇到问题（例如，Wordle #284是'stove'：玩家很容易在'stone'、'store'、'stoke'等上面浪费猜测），从而导致更高的失败率。

我们还显示了国家在我们的数据集中出现的频率，使用的是《世界日报》提供的地理数据。

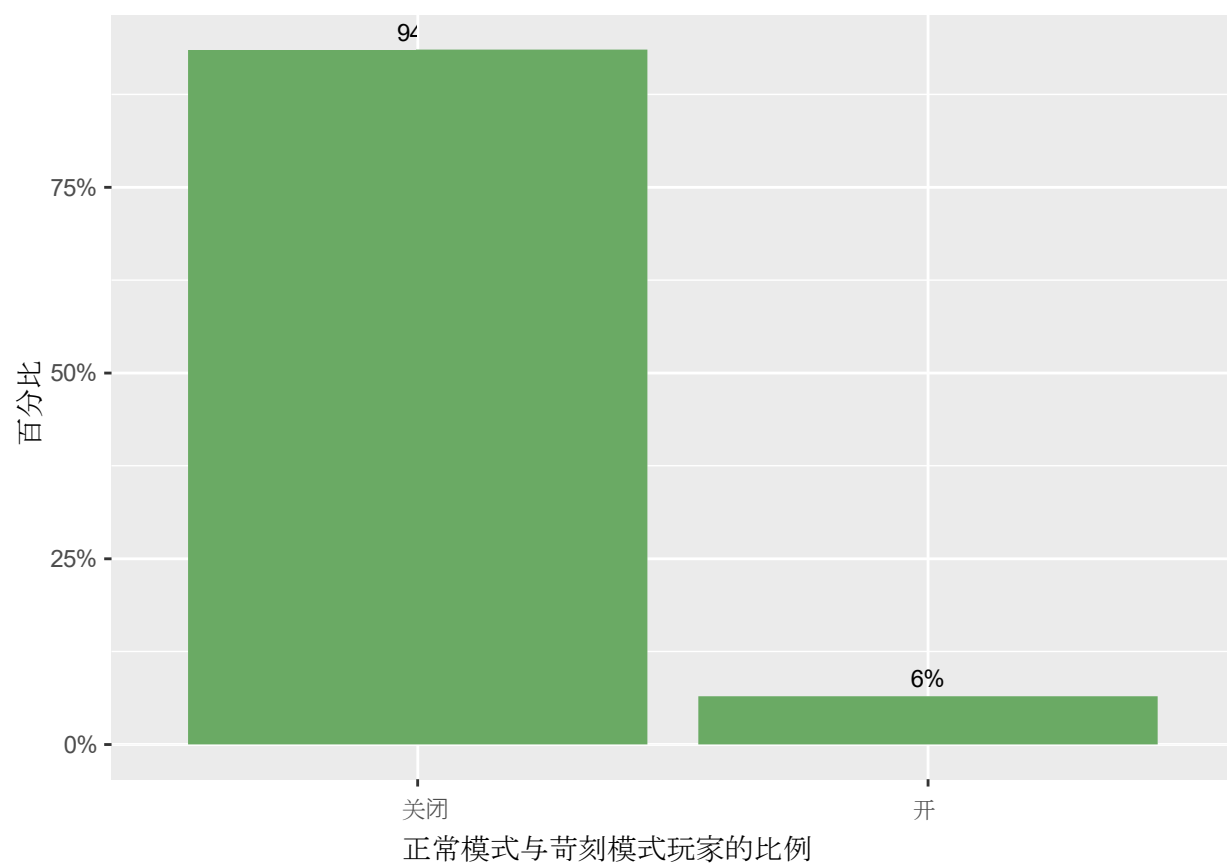


图2：Twitter Wordle玩家的游戏模式的比例

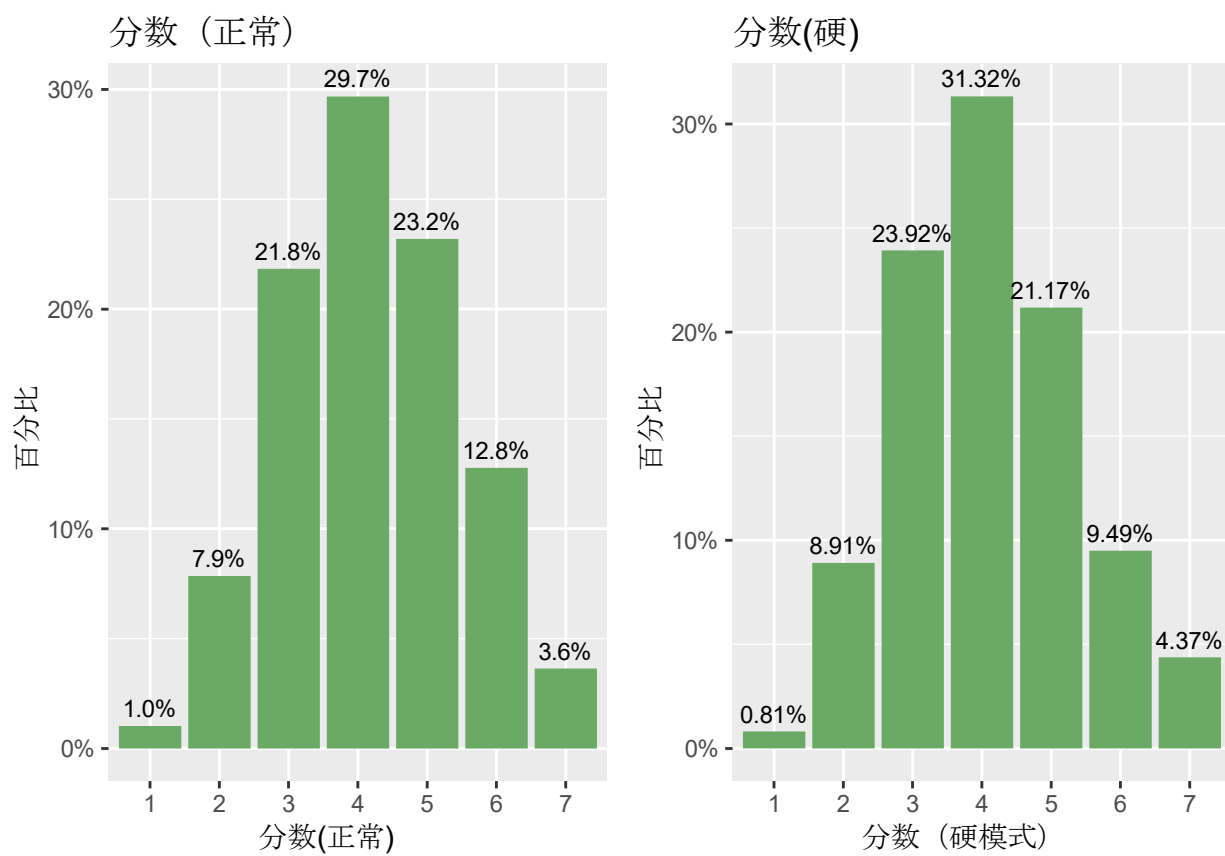


图3：不同游戏模式下Twitter Wordle玩家的得分比例

表2：在数据集中发现的最常见的10个国家

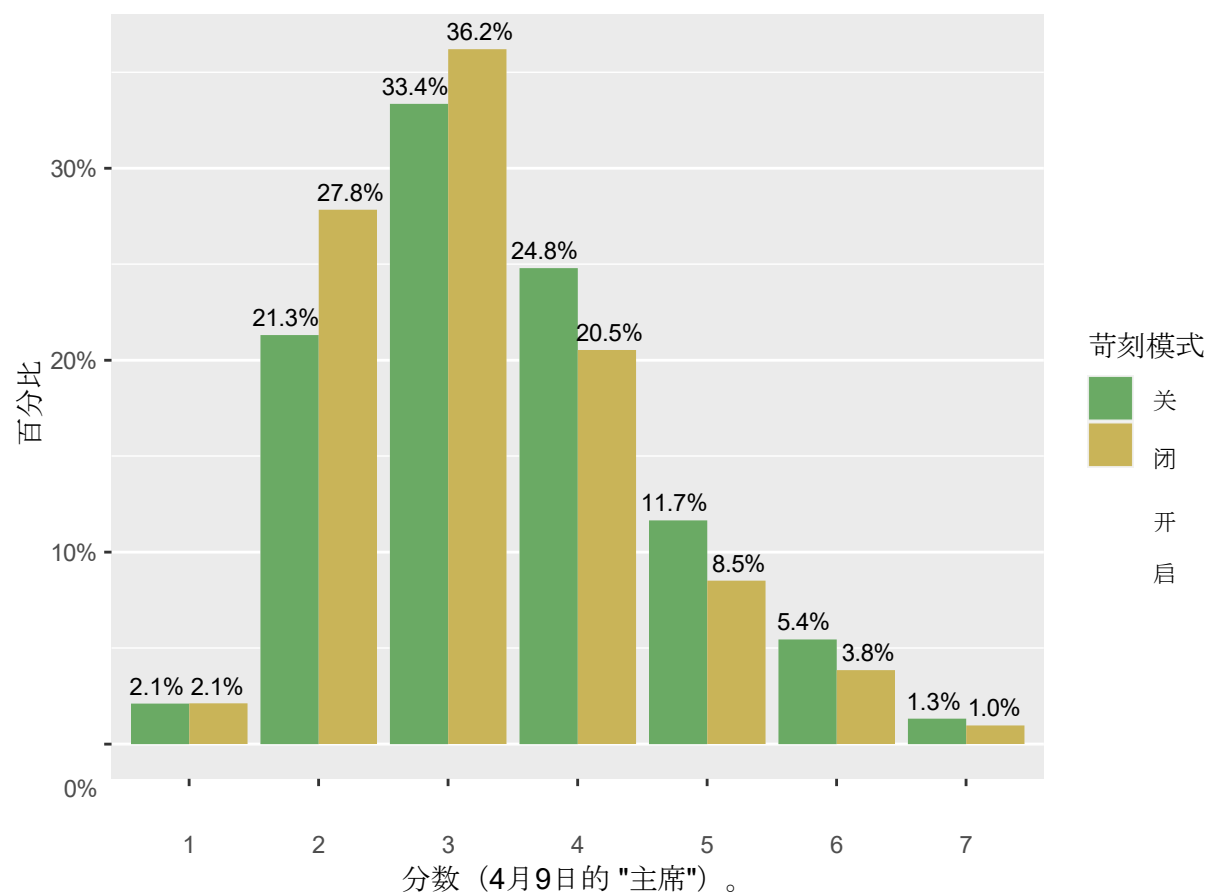
国家	频率
美国	4020
英国	679
加拿大	589
爱尔兰	150
印度	83
菲律宾共和国	60
南非	50
特立尼达和多巴哥	49
牙买加	47
巴西	43

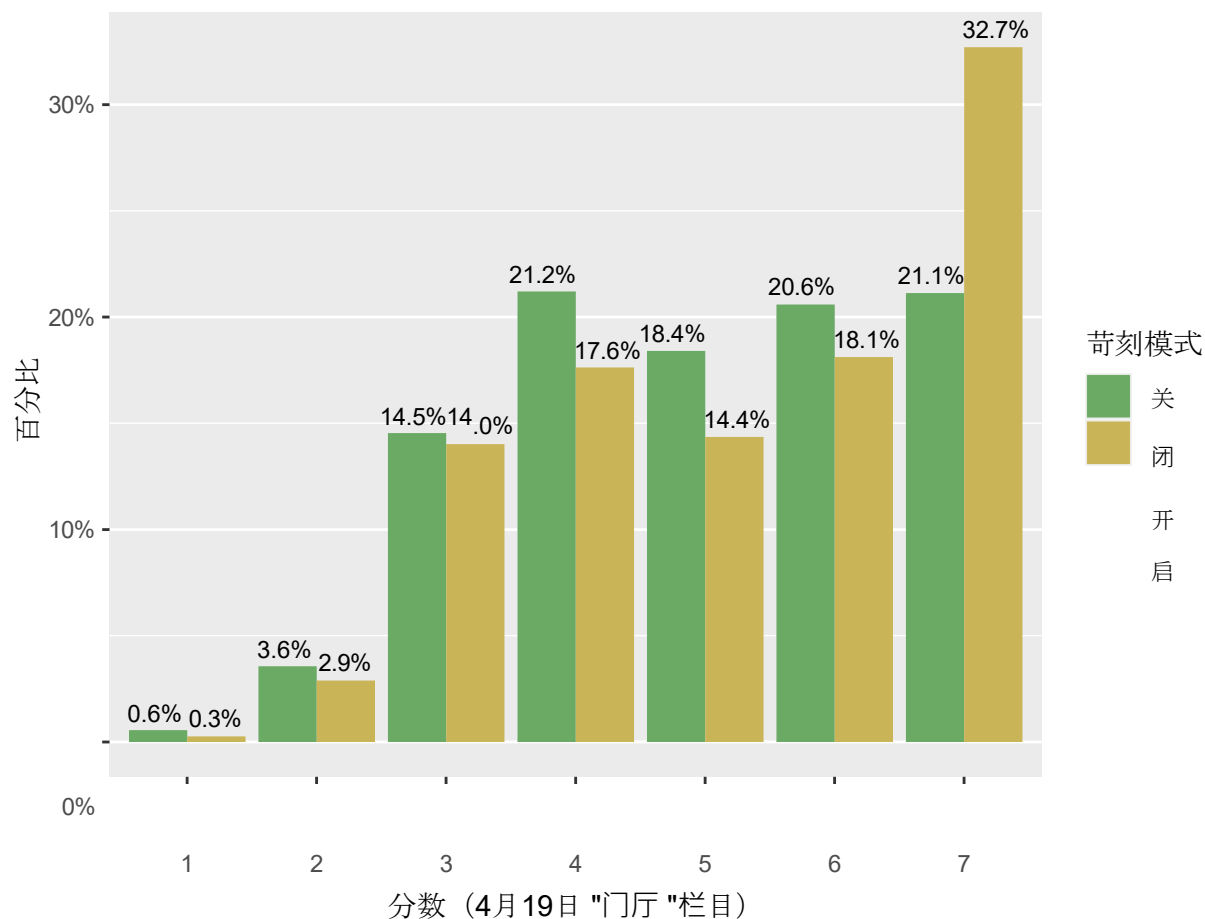
tweets。我们展示了在我们的数据集中发现的前10个国家，并观察到美洲和英国占了数据中发现的推文的大部分。

为了进行进一步的分析，我们试图根据不同日期和不同国家的用户的平均得分来绘制数据。显示的是一个表格，其中显示了每天给出单词的日期，我们整个数据集的Twitter用户的平均得分，以及当天的单词。此外，还绘制了由平均分决定的 "最容易 "和 "最难 "的单词的分布，以百分比显示，并按所用模式分组。

表3：Twitter Wordle玩家的平均得分

日期	平均得分	每日字数
2022-04-04	4.4	披肩
2022-04-05	4.5	产地
2022-04-06	4.6	逗号
2022-04-07	4.6	觅食
2022-04-08	4.1	吓人
2022-04-09	3.4	梯级
2022-04-10	3.7	黑色
2022-04-17	4.2	充足
2022-04-18	3.9	炫耀
2022-04-19	5.0	门厅
2022-04-20	4.4	货物
2022-04-21	4.6	氧化物
2022-04-22	3.5	厂房
2022-04-23	3.9	橄榄色





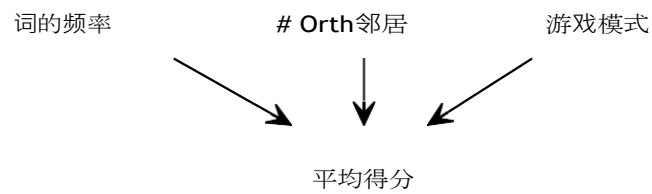
最后，我们对游戏中使用的单词本身进行分析。下表显示了该词被使用的日期，用户对该词的平均得分，当天的单词，根据当代美国英语语料库 (COCA) ("Corpus of Contemporary American English," n.d.) 该词在文本中的频率，以及根据MCWord ("MCWord: An Orthographic Wordform Database," n.d.) 该词的正字法邻居的数量。正字法邻接词被定义为长度相同但相差一个字母的词 (例如，scare和stare)。

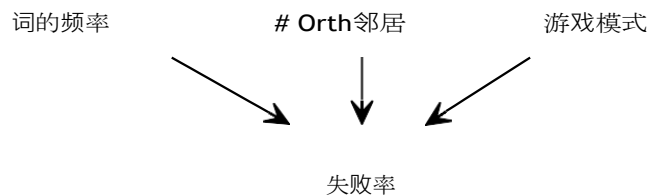
表4：每日词条的语言学属性

日期	平均得分	每日单词	词频排名	正字法邻居	2022-04-04
	4.4	披肩	11938	2	
2022-04-05	4.5	宿命	26852		3
2022-04-06	4.6	逗号	11931		1
2022-04-07	4.6	觅食	10467		1
2022-04-08	4.1	恐慌	3837		10
2022-04-09	3.4	阶梯	2880		2
2022-04-10	3.7	黑	253		5
2022-04-17	4.2	充足	6374		3
2022-04-18	3.9	叭儿	11502		1
2022-04-19	5.0	门厅	9811		2
2022-04-20	4.4	货物	4953		1
2022-04-21	4.6	氧化物	10008		0
2022-04-22	3.5	厂房	623		5
2022-04-23	3.9	橄榄球	6242		1

模型

在本文中，我们试图使用回归模型来确定推特用户每天的平均得分与当天给定单词的属性之间是否存在关系。我们首先构建了有向无环图，将我们希望讨论和建模的变量可视化，并清楚地显示我们认为它们之间存在的关系。本文以DAG为视觉，旨在表明该词在文献中的频率、正字邻接的数量和玩家的游戏模式都对预测玩家的平均得分有重要意义。





本文预测，单词的语言特性，如正字邻接的数量，会根据玩家的游戏模式对其平均得分产生不同的影响；当日常单词与其他可能的猜测有类似的结构时，假设硬模式的玩家会有更高（更差）的分数。为了尝试对这种影响进行建模，我们建立了多元回归模型，每个模型有2个预测因子，如下所示。

$$Y_1 = \beta_0 + \beta X_{11} + \beta X_{22} + c$$

在我们的第一个多元回归模型中， Y_1 代表一个正常玩家完成一个给定的Wordle谜题所需的平均猜测次数。 X_1 代表游戏中使用的某个单词的频率等级， X_2 代表该单词的正字法邻居数量， β_0 代表当频率等级和邻居数量为0时的截距或平均得分， β_1 和 β_2 代表回归系数。频率等级每增加一个，邻域数量每增加一个， Y_1 分别增加 β_1 和 β_2 。

$$Y_2 = \beta_0 + \beta X_{11} + \beta X_{22} + c$$

第二个多元回归模型也是类似的。 Y_2 代表玩家在困难模式下完成一个特定谜题所需的平均猜测次数，其他预测因子和系数具有相同的含义。然后，我们希望根据相同的游戏模式和语言特性的分组来建立完成游戏失败的概率模型。这些模型如下。

$$Y_3 = Pr(y_i = 1) = \text{logit}^{-1} (\beta_{0H} + \beta X_{1H1H} + \beta X_{2H2H})$$

$$Y_4 = Pr(y_i = 1) = \text{logit}^{-1} (\beta_{0H} + \beta X_{1H1H} + \beta X_{2H2H})$$

Y_3 和 Y_4 分别代表一个正常的Wordle玩家和一个在硬模式下的Wordle玩家在6次猜词中失败的平均概率。其他预测因子和系数的含义与先前的模型相同。

特点

我们对单词的语言属性以及它们如何转化为在Wordle背景下猜测它的难度感兴趣，因此我们决定为我们的模型包括与我们的目标有关的严格的预测变量。根据COCA的报告，一个词的频率等级与它在文学、演讲和学术界的使用直接相关，这里用来表示一个词的常见程度以及人们在生活中遇到它的可能性。我们认为，更熟悉的词会更容易回忆和猜测，因此得分较低，而不太常见的词则需要更长时间来猜测。我们还将正字旁的数量作为一个预测因素：结构相似的词可以起到“提醒”选手的作用

词的存在，并引导他们进行正确的猜测，但它们也可以迷惑和打击那些不断猜错邻居的玩家。没有邻居也可以帮助玩家，因为与其他词缺乏结构上的相似性是 "唯一性 "的代表，因此不能与其他词相混淆：例如，如果要求人们填写"_nique"，就会发现只有 "unique "才能完成这个词，而且一般来说unique有0个邻居，结构上完全不同。这可能有助于避免玩家的混淆，并减少赢得游戏所需的猜测量。

模型关注的问题

建立这些关系的最令人担忧的因素是我们数据库中的词汇量的样本大小。由于Twitter的限制和约束，收集到的单词数量非常少，这几乎不足以形成一个我们希望预测的准确模型。每天的数据点数量是合适的，可以减少每个词的变异，但是小数量的词意味着每一天在数据中都显示出极大的意义，并不能得出准确的结论。该模型将受益于一个跨度更大的时间范围的数据集，以使更多的词被分析并适合于该模型。

结果

我们创建了两个模型的变体，分析了总共四个完整的模型，并通过分析预测系数的值及其对整个模型的影响和它们所起的预测作用的值进行解释。我们继续分享每种类型的模型的结果，以及它们的发现如何相互比较。

建模平均得分

第一种形式的模型是根据单词的频率和该单词的正字邻接数来预测一个Wordle玩家的平均得分。我们的结果显示，如果玩家正常游戏，即不使用硬模式，他们的平均猜测次数从1开始每排一次就会增加0.00004126。这对应于我们模型方程中的 $Y_1 = 0.00004126$ 。例如，一个在10000名左右的词会使平均猜测次数增加0.4126。此外， Y_2 的值被确定为-0.04117。这可以被解释为。当天的单词每有一个正字邻居，一个正常的Wordle玩家的平均猜测次数就会减少0.04117次。最后，截距被发现为 $\beta_0 = 3.954$ 。对这一结果的解释是，一个排名为"0 "且有0个邻居的词将平均被猜中3.954次；它代表了其他两个预测因素保持不变时的平均猜测次数。

```
##
##调用。
## lm(formula = score ~ rank + neighbournum, data = masterdata_clean_modelmain_norm) ##
lm(formula = score ~ rank + neighbournum)
## 残留物。
## Min           Median           Max    ## -
3.9382 -0.9382 -0.1170 0.8298 3.2996 ##
## # # 系数。
##           估计值 标准误差 t值 Pr(>|t|)
## (截距)      3.954e+00  4.192e-03  943.21  <2e-16 ***
##等级        4.126e-05  3.021e-07  136.56  <2e-16 ***
## neighbournum -4.117e-02  7.810e-04  -52.72  <2e-16 ***
## ---
## Signif.代码.    0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' ' 1
##
## 残余标准误差. 1.248 on 458509 degrees of freedom ## Multiple R-
squared:          0.06022,      调整后的R平方.          0.06021
```

```
## F-statistic: 1.469e+04 on 2 and 458509 DF, p-value: < 2.2e-16
```

第二个模型使用相同的预测变量，但使用的数据集只包括使用Wordle硬模式的玩家。这意味着变量的解释保持不变，但它们具有改变硬模式玩家而非普通玩家的平均猜测次数的作用。这对系数的数值有影响，这在模型中得到反映。在二级模型中，截距值为3.843，排名系数的值为0.00004232，而邻居系数的值为-0.04185。总的来说，每个模型之间的系数值没有急剧变化。第二个模型的特点是截距较低，意味着硬模式玩家完成游戏的速度略快，但他们的结果受每天的单词频率影响非常小。

```
##
##调用。
## lm(formula = score ~ rank + neighbournum, data = masterdata_clean_modelmain_hard) ##
lm(formula = score ~ rank + neighbournum)
## 残留物。
##      闵行区      1Q 中位数      3Q      最大
## -3.8534 -0.8808 -0.0649   0.7563  3.4135
##
## 系数。
##              估计值 标准误差 t值 Pr(>|t|)
## (截点)         3.843e+00 1.602e-02 239.82 <2e-16 ***
## 级别           4.232e-05 1.184e-06  35.74 <2e-16 ***
## 邻居-4.185e-02          3.047e-03 -13.74 <2e-16 ***
## ---
## 符号代码.      0 '***', 0.001 '**', 0.01 '*', 0.05 '.' 0.1 ' ' 1
##
## 残差标准误差。1.243, 31637个自由度
## 多重R平方。      0.05952,      调整后的R平方。      0.05946
## F-statistic。 1001对2和31637DF。      p值 : < 2.2e-16
```

建立失败率模型

第二种形式的模型是使用与第一种模型相同的指标来预测两种游戏模式中玩家的失败率。由于模型的简单性和预测变量的数量较少，这些模型的结果具有相似的结构，但数值及其解释是不同的，并揭示了与之前模型不同的结果。第三个模型，预测正常Wordle玩家的失败率，显示截距的值为 $\beta_0 = 0.02647$ 。因为这个模型预测的是一个百分比，对截距的解释表明，如果一个词有0个等级和0个正字旁，根据这个模型，一个正常的Wordle玩家平均会有2.647%的失败机会。等级系数的值为 $Y_1 = 0.000001314$ 。一个词每有一个等级，正常人在6次尝试中猜不出这个词的百分比机会就会增加0.0001314。如果一个词的等级是10000，那么玩家失败的机会就会比截距值增加1.314。邻居系数被记录为 $Y_2 = -0.0004534$ ，表明一个词每有一个邻居，失败的几率就会下降0.04534%。

```
##
##调用。
## lm(formula = failed ~ rank + neighbournum, data = masterdata_clean_modelmain_norm) ##
## 残留物。
##      Min1Q Median3Q      Max ##      -
## 0.06040 -0.04113 -0.03422 -0.02935 0 .97547 ##
## ## 系数。
##              估计值 标准误差 t值 Pr(>|t|)
```

```
## (截获)      2.647e-02  6.275e-04  42.177  < 2e-16 ***
## 级别      1.314e-06  4.523e-08  29.058  < 2e-16 ***
## 邻居      -4.534e-04  1.169e-04  -3.878  0.000105 ***
## ---
## Signif.代码。    0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' ' 1
##
## 残差标准误差。 0.1868 on 458509 degrees of freedom ## Multiple R-
squared:      0.00228,      调整后的R平方。      0.002276
## F-statistic:    524 on 2 and 458509 DF,  p值: < 2.2e-16
```

预测硬模式玩家失败率的最终模型显示出与之前预测正常玩家失败率的模型明显不同的结果。我们观察到截距为0.0301，等级系数为0.000001598，邻居系数为-0.00134。一个表格显示了两个比较的并列情况，以及按数值和百分比的差异。

```
##
##调用。
$lm(formula = failed ~ rank + neighbournum, data = masterdata_clean_modelmain_hard) $
## 残留物。
##      闵行      1Q  中位数      3Q      最大
##      区
## -0.07252 -0.04982 -0.04185 -0.03535      0.97225
##
## 系数。
##      估计值 标准误差 t值 Pr(>|t|)
## (截点)    3.301e-02  2.624e-03  12.580  <2e-16 ***
## 级别      1.598e-06  1.939e-07   8.241  <2e-16 ***
## 邻居 -1.134e-03      4.990e-04  -2.272  0.0231 *
## ---
## 符号代码。    0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' ' 1
##
## 残差标准误差。 0.2037, 31637个自由度
## 多重R平方。      0.003012,      调整后的R平方。      0.002949
## F统计量： 47.79对2和31637 DF。      p值： < 2.2e-16
```

讨论

数据和模型的研究结果

在撰写本文时，由于自2021年底COVID-19大流行以来，Wordle的流行程度如日中天，因此已经有许多关于Wordle各方面的报道和文章。本文试图在分析上更进一步，涵盖这个流行的文字游戏的语言学方面，以及是否可以从作者的研究结果中得出任何重要的结论。本文构建的模型所得出的结论既能反映在数据中，又能让人乍看之下感到惊讶，但也能理解。本文的假设是，一个词的正字旁数量会对选手的表现产生负面影响；失败率和猜测次数都会增加，一个词与其他词越相似。这是从作者自己的经验中得出的结论，并认为共享高相似度的单词会导致在邻居上使用许多错误的猜测。然而，构建的四个模型显示，虽然一个词越不常见，它就越难猜，正如预测的那样，邻居的数量与平均猜测的数量以及两种类型的选手的失败率有负相关。一些解释可能包括：有更多的邻居意味着这个词的结构更容易回忆起来，因此如果被猜中，就能为玩家提供早期的领先优势；或者说，这个词的常见结构与玩家喜欢用的第一个或第二个词有关，因此在游戏开始时更容易获得提示。

数据曾显示，平均而言，硬模式的玩家的平均猜测值比正常玩家低，但总体上表现出更高的失败率。这反映在论文的模型中--虽然硬模式数据集的模型在第一类模型中显示出较低的截距值，但在第三和第四类模型的截距中反映出这两类玩家的失败率有明显的差异。由于多种原因，预计在硬模式下的玩家的平均猜测率会较低。论文中假设的一个原因是，总是利用游戏给出的提示会导致不断的进步，最终推导出单词。另一个原因是，在困难模式下的玩家可能会更认真地对待游戏或进行竞争，因此可能会因为使用其他玩家无动于衷的技术或策略而降低平均数。

弱点和下一步措施

本文的弱点主要涉及到所使用的数据集以及用于预测游戏的模型的意义和验证。本文所使用的数据集是从作者那里收集来的，在很大程度上是不完整的观察。虽然Wordle已经发布，而且它的流行程度在几个月前就已经出现了，但数据集只涵盖了两个星期，而且在这两个星期之间的观察也有差距。这导致了一个极小的样本量，不适合用于建模和分析样本统计。论文中引用的先前文章对Wordle游戏进行了几周或几个月的分析，这将是更合适的时间框架，以便收集足够的数据对平均游戏统计进行适当的分析。虽然数据集在天数方面缺乏样本量，但每天记录的推文量足以准确衡量记录天数的公众意见，如地理和游戏模式比例等统计数据应该是成立的。然而，为了确保本文的结果成立，还需要更多关于文字量和分数的数据。

由于数据集和预测变量的缺乏，论文中创建的模型也很难成立。当与Wordle游戏和单词的全部人口相比较时，模型很可能不成立，因为我们只取了一小部分单词样本，而这些单词容易产生大量的变异。这些模型没有经过验证或回归系数的测试，这可能会导致一个没有根据的、不准确的模型。未来的工作可能涉及模型的完善和增加更多的预测变量，以便准确地建立所选词语和玩家表现之间的语言关系。

作者希望本文的研究结果能够作为一个开端，并激励人们继续研究基于文字的游戏和一般的益智游戏。作为一个病毒式的游戏，Wordle在国际上和网络上都是一个流行的讨论话题，而且已经有许多游戏的衍生品看到了自己的成功。未来的游戏设计者和益智游戏制作者可能会寻求利用Wordle的结果和影响来定制自己的作品，为玩家提供有趣的挑战，并且应该继续研究如何在难度和挫折之间找到一个平衡点，正如Wordle的玩家经常经历的那样。语言研究者也可能对Wordle感兴趣，因为它是收集公众对英语的知识和意见的一种方法。随着语言的不断发展，单词有了新的含义，或者变得过时和被遗忘，Wordle可能在教育和提醒普通人识字方面很有用，就像过去的填字游戏。流行趋势不仅可以在普通民众中找到成功，并提供一个粘合和讨论的来源，而且还可以成为研究的来源，以便从这些趋势中学习并利用其影响达到进一步的目的。

附录 数据表动机

1. 创建该数据集的目的是什么？是否有一个特定的任务在脑海中？是否有一个具体的空白需要填补？请提供一个描述。
 - 创建该数据集是为了对有关流行的网络游戏Wordle的Twitter趋势进行分析。推特是一个方便的网站，可以收集大量的玩家统计数据，因为用户定期和每天都会以一种容易分析的形式分享他们的分数。
2. 谁创建了数据集（例如，哪个团队、研究小组），代表哪个实体（例如，公司、机构、组织）？
 - 该数据集是由本文作者创建的，以服务于本文的目的。
3. 谁资助了该数据集的创建？如果有相关的资助，请提供资助者的名字以及资助名称和编号。
 - 在创建数据集的过程中不需要货币成本；它是使用rtweet（Kearney 2019）软件包和编程软件R（R核心团队2020）免费创建的。
4. 还有什么意见吗？
 - 由于数据收集错误，数据集的时间数据存在差距。大部分的推文是在与获得推文的时间相关的一般时间点上收集的。因此，推文在时间上的分布是不均匀的。

组成

1. 构成数据集的实例代表什么（例如，文件、照片、人、国家）？是否有多种类型的实例（例如，电影、用户和评级；人和他们之间的互动；节点和边）？请提供一个描述。
 - 观察的每一个实例都代表了一条被称为“推特”的信息。这条信息类似于文本信息，可能包含表情符号、图片和GIF，与信息相连。推文也可以作为对其他推文的回复来发送，或者作为引用推文来发送，它是独立的，但与最初的信息有关。
2. 总共有多少个实例（每种类型，如果合适）？
 - 总共有508000个实例和508000条收集的推文。
3. 数据集是否包含所有可能的实例，还是一个更大的实例集的样本（不一定是随机的）？如果数据集是一个样本，那么这个更大的集合是什么？样本是否代表更大的集合（例如，地理覆盖面）？如果是，请描述如何验证/核实这种代表性。如果它不能代表更大的集合，请描述为什么不能（例如，为了涵盖更多不同的实例，因为实例被扣留或无法获得）。
 - 该数据集并不包含所有可能的实例。Twitter每天发送和处理数十亿条信息，不可能对如此巨大的数据流进行处理和过滤，只获得相关数据。该数据集是与Wordle相关的推文的样本，更大的数据集将包括数据收集期间每天的所有相关推文。由于数据收集的局限性，这个样本在时间上不能代表整个数据集。rtweet的局限性在于它只能收集与请求发送时间相关的“最近”的推文。由于这个原因，无法收集到均匀分布的时间范围。
4. 每个实例由什么数据组成？是“原始”数据（例如，未经处理的文本或图像）还是特征？在这两种情况下，请提供一个描述。
 - 每个实例都由以文本和unicode字符发送的信息组成，其中可能附有图片或GIF。在数据集中还有推文和发件人账户的属性，如时间顺序数据。
5. 是否有一个与每个实例相关的标签或目标？如果有，请提供说明。
 - 每条推文都有一个状态ID，显示在原始数据中。这个状态ID是唯一的，直接对应于Twitter数据库中的相关信息。
6. 个别案例中是否缺少任何信息？如果是的话，请提供一个说明，解释为什么缺少这些信息（例如，因为无法获得这些信息）。这并不包括故意的

删除的信息，但可能包括，例如，经编辑的文本。

- 没有任何信息是被Twitter或我们的数据收集程序所屏蔽的。诸如推文发件人的位置或账户信息等可能因用户的隐私问题而被有意隐瞒。

7. 单个实例之间的关系是否明确（例如，用户的电影评分、社交网络链接）？如果是，请描述这些关系是如何明确的。

- 个体实例之间的关系是明确的。部分原始数据包含了关于该实例是否是对初始推文的回复或“引用推文”的信息。这些初始推文可能在数据集中，也可能不在，这可以通过使用初始信息的状态ID并检查它是否在数据集中来验证。

8. 是否有推荐的数据分割（例如，训练、开发/验证、测试）？如果有，请提供这些分割的说明，解释其背后的理由。

- 训练和测试数据集的80/20的数据分割将被用来创建和测试数据模型。这些分割将由随机抽样的数据分区组成，这样做是为了减轻所创建的模型的偏差。

9. 数据集中是否存在错误、噪音源或冗余？如果有，请提供说明。

- 一些噪音的来源可能包括符合我们的相关标准的推文，并获得了这些推文，但并没有提供任何可用的信息。也可能有模仿我们希望获得的数据的“玩笑”结果，但其本身并不是准确的信息。

10. 数据集是自成一体的，还是链接或依赖外部资源（如网站、推特、其他数据集）？如果它链接或依赖外部资源，a) 是否有保证它们会存在，并随着时间的推移保持不变；b) 是否有完整数据集的官方存档版本（即包括数据集创建时存在的外部资源）。

c) 是否有任何与可能适用于数据集消费者的外部资源有关的限制（例如，许可证、费用）？请提供所有外部资源的描述和与之相关的任何限制，并酌情提供链接或其他访问点。

- 该数据集依赖于外部资源，因为我们正在研究推文。由于Twitter及其用户有权在任何时候删除和隐藏推文，因此不能保证它们在未来会存在。然而，档案服务是存在的，这将允许信息被保存下来。同样，被删除的推文也储存在Twitter的数据库中，尽管公众无法获得。

11. 数据集是否包含可能被视为机密的数据（例如，受法律特权或医患保密保护的数据，包括个人非公开通信内容的数据）？如果是，请提供说明。

- 如果用户在他们的推特上明确地、自愿地提到这一点，数据集可能包含机密数据。

12. 数据集是否包含如果直接查看可能会冒犯、侮辱、威胁，或可能引起焦虑的数据？如果是，请说明原因。

- 如果用户选择在他们的原始推文中包含攻击性、侮辱性和恶意的文字，那么该数据集可能包含这些文字。

13. 数据集是否确定了任何次级人群（例如，按年龄、性别）？如果是，请描述如何识别这些子人群，并说明他们在数据集中各自的分布情况。

- 该数据集没有确定任何子人群，因为这一数据无法从推文中获得。

14. 是否有可能从数据集中直接或间接（即与其他数据结合）识别个人（即一个或多个自然人）？如果是，请说明如何。

- 如果有足够的信息在他们的账户细节和他们的信息中，就有可能从数据集中识别出个人，因为这些数据是与推文实例一起包括和收集的。

15. 数据集是否包含可能被认为是任何形式的敏感数据（例如，揭示种族或民族血统、性取向、宗教信仰、政治观点或工会会员资格或地点的数据；财务或健康数据；生物识别或遗传数据；政府身份识别形式，如社会保险号码；犯罪历史）？如果是，请提供说明。

- 这些数据包含了发送推文的地理位置和地理信息。

- 用户的位置。没有包括其他敏感信息。
16. 还有什么意见吗？
- 无

收集过程

1. 与每个实例相关的数据是如何获得的？这些数据是可以直接观察到的（例如，原始文本，电影评分），是由受试者报告的（例如，调查回答），还是间接推断来自其他数据的（例如，部分语音标签，基于模型的年龄或语言猜测）？如果数据是由受试者报告或间接推断衍生自其他数据，那么数据是否经过验证/核实？如果是，请描述如何验证。
 - 这些数据可以从每条推文和用户那里直接观察到，因为推文是以原始文本形式发送的，其他细节可以从时间戳和自我提供的用户信息中直接获得。
2. 使用什么机制或程序来收集数据（例如，硬件设备或传感器，人工整理，软件程序，软件API）？这些机制或程序是如何验证的？
 - 这些程序是通过手动检查程序是否使用状态ID和推文和账户的引用来收集有效的推文来验证的。
3. 如果数据集是一个较大集合的样本，那么抽样策略是什么（例如，确定性的，有特定抽样概率的概率性的）？
 - 抽样策略是抽取截至请求时符合搜索标准的最近推文。最近的推文是按照发布时间的顺序显示的。
4. 谁参与了数据收集过程（例如，学生、群众工作者、承包商），他们是如何得到补偿的（例如，群众工作者的报酬是多少）？
 - 只有论文的作者参与了数据收集。
5. 数据是在什么时间范围内收集的？这个时间框架是否与与实例相关的数据的创建时间框架一致（例如，最近抓取的旧新闻文章）？如果不是，请描述创建与实例相关的数据的时间框架。
 - 这些数据是在两个时间段内收集的--2022年4月10日和4月23日。该时间段与数据的创建时间段大致相符。创建时间段大致是2022年4月4-10日和4月17-23日。这些时间是收集每日推文的时间。
6. 是否进行了任何伦理审查程序（例如，由机构审查委员会进行）？如果是，请提供这些审查过程的描述，包括结果，以及任何支持文件的链接或其他访问点。
 - 无
7. 你是直接从有关个人收集数据，还是通过第三方或其他来源（例如网站）获得数据？
 - 数据是从其他来源收集的，使用rtweet软件包。
8. 有关个人是否被告知数据收集的情况？如果有，请描述（或用屏幕截图或其他信息显示）如何提供通知，并提供一个链接或其他访问点，或以其他方式复制通知本身的确切语言。
 - 有关个人没有得到有关数据收集的通知。
9. 有关个人是否同意收集和使用他们的数据？如果是，请描述（或用截图或其他信息显示）如何要求和提供同意，并提供链接或其他访问点，或以其他方式复制个人同意的确切语言。
 - 有关个人同意在Twitter的服务条款中收集和使用他们的数据，所有用户必须同意该条款才能访问网站。该链接可以在这里找到：<https://twitter.com/en/tos>
10. 如果获得了同意，是否向同意的个人提供了一种机制，以便在今后或针对某些用途撤销其同意？如果有，请提供说明，以及该机制的链接或其他接入点（如果合适）。
 - 同意的个人可以选择停用他们的账户，这将删除他们从平台发送的所有信息和推文。
11. 对数据集及其使用对数据主体的潜在影响进行了分析（例如，一个

是否进行了数据保护影响分析？如果是这样，请提供对该分析的描述，包括结果，以及任何支持文件的链接或其他访问点。

- 无
12. 还有什么意见吗？
- 无

预处理/清洗/标记

1. 是否对数据进行了任何预处理/清理/标记（例如，离散化或桶化、标记化、部分语音标记、SIFT特征提取、去除实例、处理缺失值）？如果是，请提供说明。如果没有，您可以跳过本节中的其余问题。

- 对数据进行清理是为了在我们的数据集中获得更多的相关性。在收集了原始数据后，观察结果通过特定的信息标准进行过滤。相关的数据（Wordle分数）通过特定的短语包含在信息中，这些短语是由用户从游戏的网站上复制和粘贴的。分数采取 "Wordle /6 "的形式，其中显示了每天的单词数量和玩家所猜的数量。推文根据其推文中是否包含这个短语进行过滤。一旦这样做了，用户猜测次数的数字数据就会从信息文本中提取出来，并作为一个新的变量存储。如果用户在6次尝试中没有成功地猜到这个词，因而失败了，那么这个游戏就被算作是用了7次猜测来完成。完成这些工作后，使用lubridate（Grolemund和Wickham 2011）软件包从原始数据中提供的Unix时间中提取推文的发送日期和时间。我们还创建了一个二进制变量，表明用户是否在他们的游戏中启用了 "困难模式"，这为游戏增加了额外的条件，使其更具挑战性。这种指示在原始文本中是可用的；来自硬模式玩家的信息会在他们的推文中包含一个"/6*"，这个星星表示使用了硬模式。一旦完成这些工作，数据就被汇总到一个单一的数据框架中。还根据推文的创建时间框架进行了划分，以便将来可能使用。

2. 除了预处理/清洗/标记的数据之外，是否还保存了 "原始" 数据（例如，为了支持未预计到的未来用途）？如果是这样，请提供一个链接或其他访问点，以获得 "原始" 数据。

- 这些 "原始" 数据包含在本文所在的资料库中。

3. 用于预处理/清理/标记数据的软件是否可用？如果是，请提供一个链接或其他访问点。

- 用于清理数据的软件由R软件包组成，这些软件包已被引用并注明出处。

4. 还有什么意见吗？

- 无

用途

1. 该数据集是否已经被用于任何任务？如果是，请提供说明。

- 无

2. 是否有一个储存库可以链接到任何或所有使用该数据集的论文或系统？如果有，请提供一个链接或其他访问点。

- 是的。代码和数据见：<https://github.com/IvanNoar/Twilight>

3. 该数据集可用于哪些（其他）任务？

- 其他任务可能包括与Wordle有关的更多分析，以及使用数据集来得出基于Twitter的一般结论。

4. 关于数据集的构成或其收集和预处理/清洗/标记的方式，是否有任何可能影响未来使用的东西？例如，是否有可能是数据集消费者可能需要知道的，以避免使用可能导致对个人或群体的不公平待遇（例如，陈规定型观念、服务质量问题）或其他风险或伤害（例如，法律风险、财务伤害）？如果是，请提供说明。数据集的消费者可以做些什么来减少这些风险或危害？

- 鸣叫的地理和时间数据，以及所使用的账户的状态和描述，都可能被用来不公平地对待相关的人。一个数据集的消费者可能会避免做出这些结论，而更多地从表面上看这些数据。

5. 是否有不应该使用该数据集的任务？如果有，请提供说明。

- 该数据集不应被用于任何恶意或非法活动。
6. 还有什么意见吗？
- 无

分布情况

1. 该数据集是否会分发给创建该数据集的实体（例如公司、机构、组织）以外的第三方？如果是，请提供说明。
 - 该数据集不会明确地分发给第三方，但会在网上公开提供。
2. 数据集将如何分发（例如，网站上的压缩包、API、GitHub）？数据集是否有一个数字对象标识符（DOI）？
 - 该数据集将通过本文的存储库在GitHub上发布。
3. 数据集何时分发？
 - 该数据集将于2022年4月27日分发。
4. 数据集是否在版权或其他知识产权（IP）许可下和/或根据适用的使用条款（ToU）分发？如果是，请描述该许可和/或使用条款，并提供一个链接或其他访问点，或以其他方式复制任何相关的许可条款或使用条款，以及与这些限制有关的任何费用。
 - 无
5. 是否有任何第三方对与实例相关的数据施加基于知识产权或其他限制？如果是，请描述这些限制，并提供一个链接或其他访问点，或以其他方式复制任何相关的许可条款，以及与这些限制相关的任何费用。
 - 无
6. 是否有任何出口管制或其他监管限制适用于该数据集或个别实例？如果是，请描述这些限制，并提供一个链接或其他访问点，或以其他方式复制任何支持性文件。
 - 无
7. 还有什么意见吗？
 - 无

维护

1. 谁将支持/托管/维护数据集？
 - 该数据集将由论文的作者托管。
2. 如何联系数据集的所有者/收藏者/管理者（例如，电子邮件地址）？
 - 数据集的所有者可以通过存储库所在的GitHub账户联系。
3. 是否有勘误表？如果有，请提供一个链接或其他接入点。
 - 无
4. 数据集是否会被更新（例如，纠正标签错误、添加新实例、删除实例）？如果是这样，请描述多久更新一次，由谁来更新，以及如何将更新信息传达给数据集消费者（例如，邮件列表，GitHub）？
 - 该数据集将不会被更新。
5. 如果数据集与人有关，对与实例有关的数据的保留是否有适用的限制（例如，有关个人是否被告知他们的数据将被保留一段固定的时间，然后删除）？如果有，请描述这些限制，并解释如何执行这些限制。
 - 对与实例相关的数据的保留没有限制。
6. 数据集的旧版本是否将继续得到支持/托管/维护？如果是，请描述如何。如果不是，请说明如何将其过时的情况告知数据集消费者。
 - 数据集的旧版本可以由论文作者在其本地文件上托管。过时的情况将从论文托管的存储库中通报。
7. 如果其他人想对该数据集进行扩展/增补/构建/贡献，是否有一个机制让他们这样做？如果有，请提供说明。这些贡献是否会被验证/核实？如果是的话，请说明如何。如果不是，为什么不？是否有一个向数据集消费者传达/分发这些贡献的过程？如果有，请提供说明。

- 其他人可以通过GitHub内置的协作功能对数据集做出贡献，这些贡献将由作者核实并最终确定

8. 还有什么意见吗？

- 无

参考文献

- "Corpus of Contemporary American English." n.d. 英语。 <https://www.english-corpora.org/coca/>。
- Grolemund, Garrett, and Hadley Wickham. 2011. "用lubridate轻松实现日期和时间"。 *统计软件杂志》* 40 (3) 。 1-25。 <https://www.jstatsoft.org/v40/i03/>。
- Kearney, Michael W. 2019. "Rtweet. 收集和分析Twitter数据。" *开源软件杂志》* 4 (42)。 1829. <https://doi.org/10.21105/joss.01829>.
- "MCWord. 一个正字法词形数据库。" n.d. MCWord. *An Orthographic Wordform Database*. <http://www.neuro.mcw.edu/mcword/>.
- R核心团队。 2020. *R: A Language and Environment for Statistical Computing*. 维也纳, 奥地利: R统计计算基金会。 <https://www.R-project.org/>。
- "世界上哪里的人最会解字谜?" N.d. *Wordtips*. <https://word.tips/wordle-wizards/>。
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "欢迎来到tidyverse。" *开源软件杂志》* 4 (43):1686. <https://doi.org/10.21105/joss.01686>.