2.18

预解决问题：

预测单词难度以及对单词的各分数段（尝试次数）百分比做出预测

思路：

创建一个算法模型，模拟人类玩填字迷游戏。以该算法来对人类的游戏行为做出结果预测

方法：

1. 选取指标：

   影响人类在填字谜游戏中选择单词的因素很多，如单词的词性、所具备的情感 以及单词的常见程度等 我们首先选取以上三个指标作为标准 **查看其与真实数据（分数期望）之间的关系：**

   两张条形图：-->单词的情感和词性属性在填字谜中对结果对数据的影响较小

   - 情感分析：选取出全部数据集中的词汇，利用python NTLK工具包 对每个单词情感进行打分，（区间【-1，1】）越大说明情感的积极性越强

     ```
     array([4.14691606, 4.09017473, 4.1697104 , 4.23213311, 4.10668029])
     array(['NN', 'JJ', 'RB', 'VBP', 'VBD'], dtype=object)
     ```

   - 词性分析 选取出全部数据集中的词汇，利用python NTLK工具包 对每个单词词性进行分类，在分类得到的多个数据中 主要以 名词（NN）形容词（JJ）动词（VB）为主，大部分类型的单词的数据在10以下，为提升稳健性和数据的可靠性，我们选取数据集在10以上的数据进行讨论

     ```
     [4.0560515949076485,4.148104614766078,4.0560515949076485],[pos nut,neg]
     ```

   -->二者均不存在显著的关联性，排除这两组数据

   选取单词的使用频率对得分之间的关系作为指标

2. 频率数据获取

   利用wolftram (https://reference.wolfram.com/legacy/language/v13/ref/WordFrequencyData.html.zh)的单词频率数据集为基础 截取数据

   再获取wordlist的词库数据

   Wordle Words - All 2309 Words (Not in Order) No Spoilers! (wordunscrambler.net)

   在指定的2309个单词内，将频率转换为概率，依照概率选取满足条件的下一个指标

   选取指标后截取数据 依据条件反馈缩减答案的可能的词库

   -->.....

   拟合出结果

   -->图1

   由图像可得，预测数据的均值与方差略高于真实数据，由此对预测数据做一定的调整与修改 使之更好地拟合于真实数据

1. 调整期望：取前250个单词，求出真实数据与预测数据的期望差，将预测数据减去差值后 得到新的预测数据（new=old-difference）用后109个数据核对：误差在0.01以内（见图）
2. 调整方差：

   在不改变相关性的情况下 用线性的方式将方差大的数据尽可能向中心靠拢，采用公式：

   f(x)=[f(x)-E(x)]*k+E(x)，k为常数，定义为

   1-[|f(x)-g(x)|]*4/[f(x)+g(x)]
3. 得到新的拟合数据结果（见图）

3. 数据各区间百分比预测：

   一个目标单词 按照上述规则随机输入输出100次 看其概率分布 以其概率分布来模拟在各分数段的数据分布
4. 结果：

   见bigMountain.csv 和smallMountain.csv

   还没有做拟合结果好坏分析

单词按照字母出现频率打分分析：

数据结果：finalLetterData.csv