输入：待评价的单词

输出：单词的被联想概率

步骤一：切分单词

1；A Basic(S),其中，Basic(S)表示获得单词的切分序列A=（空集，w\_1,w\_2…w\_i，空集），

2；for each letter groups w\_i in A：

统计w\_i在训练语料库中出现的频数，记为N\_i,

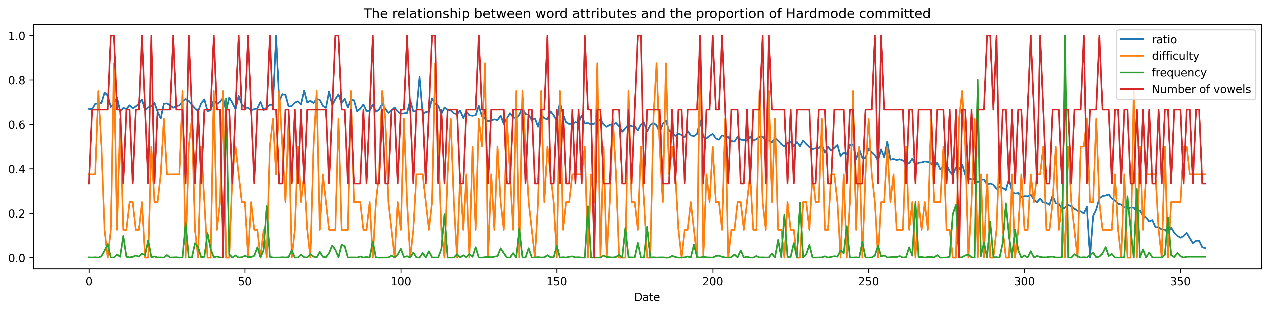
评价EERIE的难度：

应用我们建立的trigram模型，把EERIE代入模型中进行计算，得到log\_p(EERIE)=-22.0632,这是一个很低的数值，通过聚类后，该单词被分为单词迷雾等级为7的一个生僻的单词

单词属性

假设：在统计单词包含的元音字母时，只考虑“a”,’i’’u’’e’’’o的出现次数。

对于单词的难度，在语料库中的频数，以及包含的元音字母个数，我们做了相关分析和图像展示来表现他们直接的关系，但遗憾的是，单词的属性和困难模式提交次数的占比没有直接关系。



通过计算相关系数r：

R=

得到：

|  |  |  |  |
| --- | --- | --- | --- |
|  | Difficulty | Frequency | Number of vowels |
| Ratio | 0.085 | -0.094 | 0.09 |