信息熵计算实验 实验报告

2013011427　刘智峰　计 31

一、 实验目的

由于本次是自选实验，主题是信息熵的计算，所以我给自己设定了一个实验目标：网上都能搜到汉字和 26 个英文字母的使用频率，根据这些信息，可以计算出汉字和 26 个字母的信息熵。在此基础上，分别实现对一部长篇中文小说和一部长篇英文小说的单词统计，并计算信息熵，将结果与网上得到的数据计算出的结果进行对比分析，看看是否符合网上所陈述的数据的分布。

二、 实验过程

本次实验使用的熵公式为：

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

选取的中文样本为：三体 1-3 部合集，保存在 santi.txt 中；选取的英文样本为：哈利波特 1-4 部英文合集，保存在 HarryPotter.txt 中。

使用 python 编写代码，实验环境为 pycharm 4.5.4. 代码思路为：维护并统计所有非重复的汉字或单词的个数信息，同时统计所有单词或汉字的总和，然后计算各个汉字、字母的频率，计算信息熵。由于汉字博大精深，在进行中文信息熵计算时，只取了概率高的 50 个汉字。

中文前 50 个字的概率统计结果，位于 countChinese.txt 中，最终的信息熵由程序 calChinese 输出。

英文 26 个字母的概率统计结果，位于 countEnglish.txt 中，最终信息熵结果由程序 calEnglish 输出。

同时，汉字使用频率前 50_百度文库.xlsx 和英文字母统计频率.xlsx 这两个 xlsx 文件，为对网上的汉字、字母信息进行陈列，并计算出信息熵。

三、 实验结果

汉字信息熵：

针对三体小说的统计结果为 1.1934

```
E:\Python27\python.exe E:/Pycharm_workspace/信息熵/caliChinese
886881
29243
the answer is:
1.19344105782
```

针对网上得到的数据，统计结果为 2.1884

英文字母信息熵：

针对哈利波特 1-4 部的统计结果为 4.1923

```
alEnglish
E:\Python27\python.exe E:/Pycharm_workspace/信息熵/calEnglish
3176780
26
the answer is:
4.19233418045
```

针对密码学课件给出的统计结果，结果为 4.1802

四、 分析与体会

从上述结果可以看出，汉字信息熵的结果与用网上给出的各个汉字出现频率计算出的信息熵差距比较大，原因可能是，汉字实在太多了，本

次实验我只取了前 50 个字，还远远不够全面。而且，三体小说的总体篇幅也不是很长，从上述截图可以看出，总共才 88 万个字，其中只有 2.6 万个不同。再加上汉字的博大精深，很多词都是由看似不着边际的几个字组成的，规律性较小。而英文就不一样了，由于英文只有这 26 个字母，取 26 个字母的统计之和，肯定比取 50 个汉字要来的全面的多。而且，从英文结果截图可以看出，哈里波特样本总共有 317 万个字母，样本量足够 ,所以最后得出的结果和密码学课件中给出的频率计算的结果很接近。

本次实验的数据可能还不够准确，但通过实验过程，我更加深入地了解掌握了信息熵的计算方法 ,对老师上课讲授的内容也有了更多的理解，收获还是很大的。