多媒体技术基础及应用信息熵计算实验报告

计 15 班 申喆 2011011313

一、实验目的

计算汉字和英文字母的信息熵。

二、实验原理

对于英文字母或者汉字，利用下面公式计算得到其信息熵。其中 p_i 代表每个英文字母或者汉字出现的频率。

$$H = -\sum_{i=1}^{n} p_i log_2 p_i$$

三、实验代码

实验代码见 entropy.py 文件。本实验所采用的语言为 Python3。其中 entropy 函数对每个英文字母或者汉字频率计算 $-p_i log_2 p_i$ 的值，cal_res 函数根据文件(汉字频率文件是 CN_entropy.txt，英文字母频率文件是 EN_entropy.txt)中数据特征获得有效数据并且计算得到相应的信息熵。其中 cal_CN.py 用于对 ChineseCharFrequencyG.txt 进行处理获得需要的每个汉字的频率值。

实验最终结果保存在 Final_Result.txt 文件中。

四、实验结果及分析

经实验得到汉字的信息熵为 9.665541472638903，英文字母的信息熵为 4.17086491107289。

可以看出汉字的信息熵远高于英文字母的信息熵。这也就说明英文比汉文更加容易读懂，或者换种说法也就是更加容易被预测。联系之前做过的密码学作业，可以联想到各种维吉尼亚密码之类的基本可以根据英文单字母以及双字母的频率以及英文常识来破解，然后我就自行脑补了一下破解汉语密文，接着不知道为什么就想起了”烫烫烫烫烫屯屯屯屯屯锟斤拷”之类的……

总之，实验结果就是英文比汉语信息熵要小，更加容易读懂，更加容易被预测，这也很符合我们的常识。所以实验结果比较科学合理。

五、实验参考资料

汉字频率统计文件来自于

http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=MO

英语字母频率统计文件来自于

http://zh.wikipedia.org/wiki/%E5%AD%97%E6%AF%8D%E9%A2%91%E7%8E%87

注：其中共统计了 9933 个汉字，覆盖率大于 99.9999994832。汉字频率数据为 2005 年左右数据。