# 实验项目四
# 中文词典的自动生成

尹存燕

2015年12月

# 实验说明：

- 实验数据文件：100M的文本文件
  - 一行是一个中文句子，句子中的汉字或中文标点符号均用空格分隔。
  - 数字及数字相关字符（比如：%)和英文字母字符都是半角字符，英文单词有的用空格分隔，有的没有。网址有的是连续的，有的中间有空格。
  - 每一行的**有效**字符长度不超过100；

# 实验内容（一）

- 在数据文件的基础上，统计文件中汉字字符的出现次数，并按照次数，从多到少排序，输出到D:\\token.txt文本文件中。一个字和其出现次数占文本文件的一行。
- 统计相邻两个汉字的出现次数，以此生成中文二字词词典，并按照次数，将次数大于1的词，从多到少排序，输出到D:\\word_2.txt文本文件中。
- 统计相邻三个汉字的出现次数，以此生成中文三字词词典，并按照次数，将次数大于1的词，从多到少排序， 输出到D:\\word_3.txt文本文件中。
- 统计相邻四个汉字的出现次数，以此生成中文四字词词典，并按照次数，将次数大于1的词，从多到少排序， 输出到D:\\word_4.txt文本文件中。

# 实验内容（二）

- 在统计结果文件（token、word_2、 word_3、 word_4）基础上，设定合适的规则，选择相关结果生成中文词典文件，按照出现次数，从多到少，输出到文本文件D:\\dictionary.txt 。文件中一个词及其出现次数占一行。
- 在中文词典文件中统计：按照你设定的规则选择出的中文词汇正确率，所谓“正确”的词汇就是中文实际使用的词汇。根据正确率的统计，将出现次数排名前1000的词汇用excel表列出，字段分别是：次数、词汇、是否正确。Excel文件名为：statistics.xlsx
- 实验报告要求：
  - 写明各个统计文件的生成算法，以及各个统计算法实际耗时、运行时占用内存大小。
  - 写明在统计文件基础上，选择中文词汇的规则。
  - 在statistics.xlsx文件的基础上，完成正确率的统计表格，表格模板如下页所示，并根据表格画出正确率曲线图。

# 词汇正确率统计表

| 出现次数 | 词典的词汇个数 | 正确的词汇数 | 正确率（%） |
| --- | --- | --- | --- |
| 排名前100 | | | |
| 排名前200 | | | |
| 排名前400 | | | |
| 排名前600 | | | |
| 排名前800 | | | |
| 排名前1000 | | | |

# 实验评分说明

- 程序、报告、答辩各占5分。
- 如果程序中使用标准模板库（STL），各项评分最高为3分。
- 在截止时间之后提交程序和报告的，如果没有使用STL，最高4分。
- 程序和报告的提交截止时间是：12月28日晚上9点。
  - 提交5个程序文件，分别对应5个文本文件的生成：token.cpp、word_2.cpp、 word_3.cpp、 word_4.cpp、dictionary.cpp
- 答辩时间是：12月29日下午2-4点。
- 如果要求提前答辩，请先将程序和报告提交在网站上，并和相关助教email联系，约定答辩时间。
- 最晚答辩时间为1月5日下午2-4点。

# 助教负责学号范围

- 学号121130086~131220044：林木丰（forest.sky.sea@gmail.com）
  学号131220045~131220090：吴小同（csu_wxt@126.com）
  学号131220092~131220146：李其玮（liqiwei_nju@163.com）
  学号131220147~158354004：黄璐宸（luvinahlc@gmail.com）