中文词表自动生成 实验报告

一、实验题目解读

对于所给的中文文本文件（共464M），进行如下操作。

1. 统计1字串、2字串、3字串、4字串的出现次数，并分别将字串和出现次数输出到4个文件中

2. 分别在这4类字串中设定阈值，出现频率>=阈值的字符串认为是词汇候选，将这些词输出到1个词典文件中

3. 记录实验程序的运行耗时、内存占用量，判断词典文件中的前100个，前200、前400，前600个词汇的正确率

4. 尝试提高前600个词汇正确率的方法。

二、需求分析

**统计1、2、3、4字串的词频，需要先生成这些1、2、3、4字串，如何生成？**

1.由于项目文件较大，选择将它按行读入，建立一个string类型数组line[MAXLINE],储存所有行的信息。

2.如何区分全角半角？

对于每一行，在读入时进行判断，假如某一位编码在0~127之间，说明它是半角字符，舍弃，读下一位，最终把这行里面所有的汉字存到char str[]中，再把这个str[]赋给line[k]， k++。

在后续测试中，发现中文文本.txt大概有260~270万行，于是设置MAXLINE为2800000

3.怎样用line[]生成1、2、3、4字串？又如何排除全角符号呢？

在测试中，我发现生成的字串很大，略作计算，这个文本大概有250万行，以一字串为例，每行至少有20个，那如果建立数组的话元素可能有50,000,000个，四个字符串就是2亿个，测试中出现了占用内存过大的现象，于是我选择把对四种字串的处理分到四个cpp文件中，每个文件处理生成一种字串。

另一方面，考虑到要对生成的字串做求出词频的操作，我最终选择使用map，因为这样key中的值，即我们要统计的单词，不会重复，而对应的value值就是它的词频。

**如何进行排序呢？**

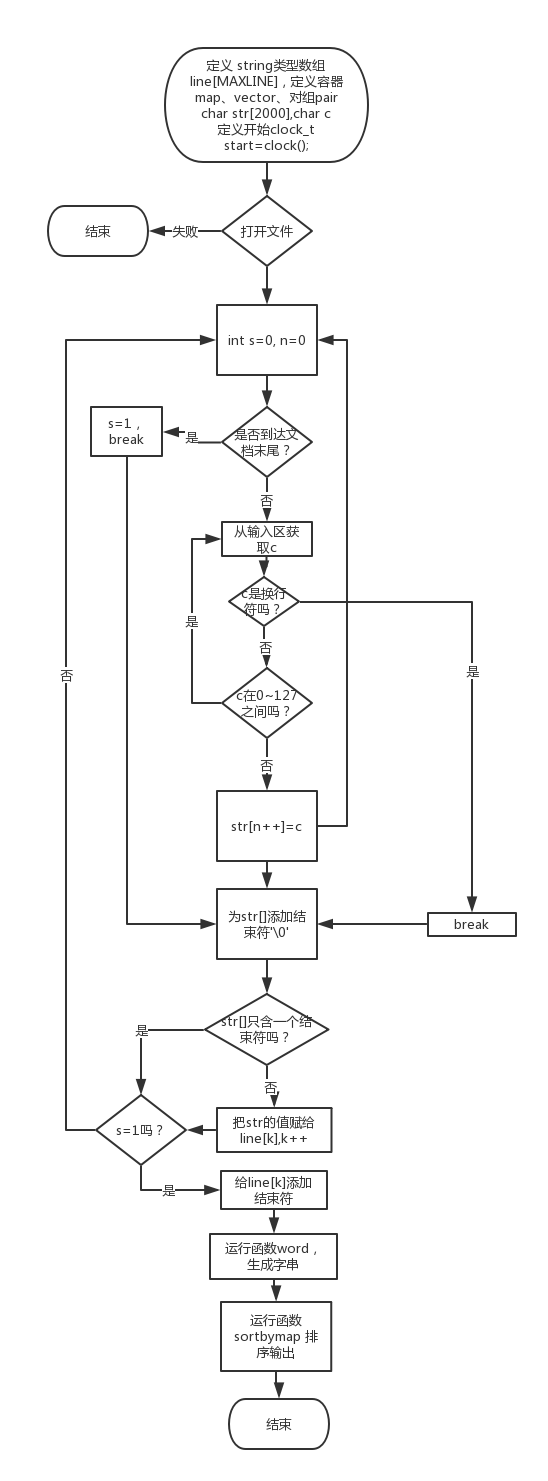
此时，问题转化为对map容器中元素按照value值进行排序，由于map是关联容器，无法使用sort，所以我们要把map中的元素放到一个顺序容器vector中。然而，vector中的元素必须满足一个条件，就是它们之间必须是可比较的，于是我们可以把它存到一个pair类型的vector中，再对vector按照一定的规则排序即可。

**如何输出为字典文件呢？**

在生成了字符串以后，我们选定阈值，写一个新的程序，先读入四个字串文件，再按照每个字串25个的规则输入到词典文件中。

三、主函数流程图

在生成字串程序中，主函数中主要是进行line[]数组赋值

字符串处理程序主函数

词典生成程序主函数



四、功能函数说明

1.生成字串的函数void word(string line[], int n)

以一字串为例，设一个string类型变量temp，循环把line[][]中的相邻四个值输入到temp中，这样temp中就含有一个汉字，考虑到汉字的高位字节和全角标点的高位字节有所不同，所以可以进行判断，如果是汉字的话，就把temp赋值给word，再生成一个pair值（word，1），把它插入到map容器中。如此进行循环

代码如下：

void word(string line[], int n)

{

int i, j, k; string word, temp;

for (i = 0; i<n; i++) {

for (j = 0; line[i][j]; j++)

;

k = j;

for (j = 0; j<k - 1; ) {

temp = line[i][j];

temp += line[i][j + 1];

if (('0' - temp[0]) <= 128) {

word = temp;

ret = tMap.insert(make\_pair(word, 1));

if (!ret.second)

++ret.first->second;

}

j = j + 2;

}

}

}

2.对map容器进行排序的函数void sortMapByValue(map<string, int>& tMap, vector<pair<string, int> >& tVector)

把map容器中pair类型的值push进入tVector容器中，进行排序。

排序时要用到一个判定排序标准的函数cmp，代码如下：

int cmp(const pair<string, int>& x, const pair<string, int>& y)

{

return x.second > y.second;

}

当x元素的value值大于y元素的value值时返回1.

sort函数以cmp的值为标准进行排序，这样就得到了1字串按照词频进行排序的结果。

整个排序函数代码如下：

void sortMapByValue(map<string, int>& tMap, vector<pair<string, int> >& tVector)

{

for (map<string, int>::iterator curr = tMap.begin(); curr != tMap.end(); curr++)

tVector.push\_back(make\_pair(curr->first, curr->second));

sort(tVector.begin(), tVector.end(), cmp);

}

3.词典生成程序中输入函数input，按照阈值输入词

void input(const char s[], Word word[], const int k)

{

Word t;

int n = 0;

ifstream infile(s);

if (!infile) {

cerr << "打开文件" << s << "失败" << endl; exit(0);

}

while (infile >> t.s) {

infile >> t.num;

if (t.num < k) break;

else word[n++] = t;

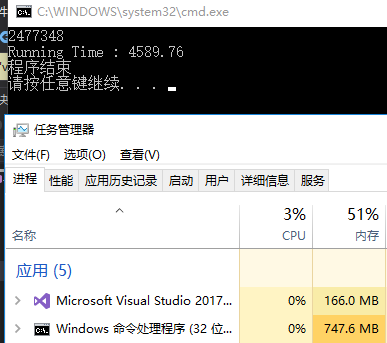
}

word[n].s = "0";

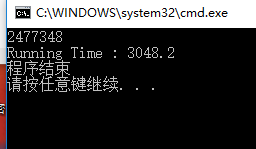
}

四、程序运行时耗时和内存分析

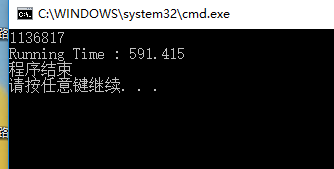
说明：由于本人电脑的内存限制（某个进程占用内存超过2GB时会自动停止运行），在处理1、2字串时采用了400M的中文文本，处理3、4字串时采用了100M的中文文本。



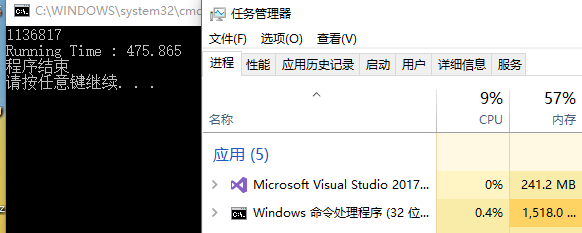
处理1字串时，内存占用747.6，总共耗时4589s



处理2字串时，内存占用约1200MB，总共耗时3048.2s



处理3字串时，内存占用1354MB，总共耗时591.415s



处理4字串时，内存占用1518MB，总共耗时475.865s

五、实验结果分析

选定的阈值：

1字串 出现18000次以上的大概有1500个左右

2字串 出现5000次以上的大概有1800个左右

3字串 出现 400次以上的大概有600个左右

4字串 出现120次以上的大概有800个左右

这些词可以作为词汇候选。

统计正确率

前600个词分段统计正确的词的数量和正确率

1字词 2字词 3字词 4字词 正确率

20/25 21/25 13/25 10/25 0.64

16/25 24/25 5/25 12/25 0.57

19/25 20/25 13/25 13/25 0.65

18/25 21/25 10/25 12/25 0.61

22/25 19/25 8/25 8/25 0.57

12/25 21/25 4/25 14/25 0.41

18/25 18/25

11/25 17/25

10/25 22/25

22/25

20/25

20/25

14/25

前100 0.64

前200 0.605

前400 0.6175

前600 0.575

前600

六、思考与改进

如何提高前600个词汇正确率？

1.分析统计数据我们发现2字词的正确率最高，1字词次之，而4字词和三字词的正确率都比较低，所以可以增加1、2字词的输出比例

在前600个词中，1字词正确率为0.71,

2字词为0.84,

3字词为0.35

4字词为0.46

那么我们划定比例为 0.21：0.54:0.12:0.13

2.另外，我们注意到对于3、4字串，如果3字串的后两个字以及4字串的中间两个字是一个二字词，那么这个字串是词的概率较小，所以我们在输出时可以进行筛选

经过优化后的正确率：

前600个词正确率提高到0.73

前六百个词汇：

的:

一:

在:

是:

人:

有:

中:

国:

不:

了:

年:

大:

为:

上:

和:

发:

日:

会:

这:

个:

出:

记者:

中国:

公司:

月日:

一个:

市场:

工作:

发展:

表示:

没有:

企业:

我们:

国家:

目前:

问题:

政府:

新华:

北京:

经济:

进行:

年月:

美国:

今年:

自己:

可以:

社会:

价格:

已经:

他们:

建设:

人员:

可能:

时间:

管理:

认为:

同时:

部门:

的一:

通过:

投资:

方面:

项目:

情况:

国际:

服务:

人民:

安全:

这个:

产品:

一些:

大学:

如果:

研究:

成为:

年月日:

新华社:

资料图:

社记者:

华社记:

新华网:

自己的:

有限公:

限公司:

房地产:

北京市:

进一步:

新华社记:

华社记者:

有限公司:

月日下午:

月日上午:

工作人员:

新华网北:

资料图片:

新华社发:

告诉记者:

可吸入颗:

本报记者:

月日消息:

行:

地:

市:

时:

到:

业:

公:

对:

生:

以:

家:

来:

要:

月:

成:

者:

新:

多:

能:

于:

部:

城市:

对于:

就是:

万元:

学生:

全国:

开始:

影响:

出现:

现在:

其中:

部分:

教育:

政策:

是一:

这些:

第一:

地区:

文化:

因为:

发现:

了一:

增长:

主要:

要求:

相关:

银行:

亿元:

消费:

显示:

信息:

的是:

世界:

发生:

需要:

组织:

活动:

由于:

上海:

地产:

根据:

保障:

这一:

重要:

销售:

孩子:

不是:

以及:

这样:

去年:

技术:

的人:

生活:

中心:

利比亚:

消费者:

委员会:

负责人:

工作人:

报记者:

月日下:

国务院:

月日上:

人民币:

告诉记:

月日晚:

业内人士:

数据显示:

犯罪嫌疑:

罪嫌疑人:

食品安全:

保险公司:

媒体报道:

全国人大:

月日凌晨:

图片来源:

巴基斯坦:

记者采访:

股份有限:

方:

现:

经:

作:

前:

后:

学:

工:

用:

高:

开:

也:

过:

下:

他:

产:

全:

我:

将:

分:

进:

调查:

大的:

提供:

到了:

生产:

开发:

日本:

什么:

年的:

媒体:

报道:

使用:

很多:

还是:

资金:

机构:

发布:

标准:

网络:

负责:

有关:

产业:

这是:

不过:

行业:

人的:

但是:

国内:

作为:

规定:

人士:

增加:

有一:

美元:

个人:

数据:

集团:

合作:

包括:

食品:

其他:

不能:

收入:

一步:

了解:

中的:

方式:

为了:

达到:

介绍:

上的:

提高:

来的:

的时:

作人员:

这样的:

是一个:

俄罗斯:

华社发:

的时候:

平方米:

委书记:

诉记者:

在北京:

日下午:

的问题:

中国共产:

新闻发布:

澳大利亚:

人民法院:

国共产党:

同比增长:

月日至日:

关负责人:

人大常委:

中共中央:

有关部门:

闻发布会:

中国人民:

可:

就:

动:

元:

长:

说:

民:

记:

们:

价:

场:

机:

政:

自:

电:

实:

等:

法:

同:

资:

关:

单位:

超过:

这种:

也是:

新闻:

计划:

我国:

工程:

一次:

学校:

系统:

住房:

代表:

实现:

虽然:

微博:

地方:

个月:

领导:

以来:

资源:

都是:

交易:

以上:

希望:

正在:

关注:

房地:

们的:

环境:

基金:

样的:

己的:

委员:

中央:

关系:

分析:

此外:

继续:

专家:

上涨:

还有:

不同:

行为:

一直:

因此:

基本:

接受:

网站:

有的:

进入:

选择:

业的:

网友:

开发商:

日上午:

主持人:

据了解:

的情况:

为什么:

进行了:

越来越:

今年月:

亿美元:

本报记:

入颗粒:

大常委会:

委副书记:

市公安局:

店内提供:

年月至年:

日时至日:

羊城晚报:

上市公司:

经济发展:

地方政府:

与此同时:

者采访时:

京华时报:

本:

面:

车:

理:

加:

房:

网:

主:

而:

其:

力:

事:

与:

定:

品:

区:

子:

重:

表:

小:

体:

非常:

第二:

有限:

存在:

实施:

设计:

进一:

会议:

之后:

报告:

时候:

并不:

汽车:

不会:

近日:

最大:

全球:

未来:

医院:

上市:

一定:

事件:

应该:

加强:

改革:

所以:

电视:

能力:

看到:

公开:

国的:

不少:

公布:

旅游:

责任:

提出:

造成:

甚至:

的大:

水平:

支持:

品牌:

解决:

出了:

制度:

按照:

左右:

当地:

消息:

据了:

联合:

告诉:

不断:

一种:

公司的:

新京报:

我们的:

的工作:

嫌疑人:

在美国:

最大的:

大学生:

办公室:

中国人:

发展的:

互联网:

月日时分:

公安机关:

研究中心:

新华社路:

的情况下:

个百分点:

保障性住:

人民政府:

南方日报:

今年月日:

全国政协:

障性住房:

部分地区:

还:

天:

员:

目:

建:

都:

得:

最:

内:

合:

保:

通:

美:

金:

当:

度:

利:

报:

万:

比:

之:

保护:

费者:

参加:

现场:

随着:

期间:

月份:

此次:

原因:

交通:

在这:

行的:

所有:

也不:

财政:

获得:

金融:

指出:

比较:

如何:

不仅:

完成:

持续:

特别:

一位:

需求:

历史:

只有:

科学:

过程:

举行:

平方:

方米:

规模:

家庭:

群众:

是否:

在一:

的发:

土地:

的情:

两个:

商品:

结果:

手机:

的重:

多的:

新的:

广州:

业务:

好的:

而且:

华社:

生的:

在中国:

公安局:

新闻发:

保障房:

的一个:

联合国:

发布会:

的发展:

中国的:

卡扎菲:

共产党:

关部门:

中国经济:

领导干部:

年上半年:

份有限公:

参考价格:

乌鲁木齐:

记者了解:

记者表示:

相关部门:

中小企业:

人大代表:

社会主义:

华社路透: