

按以下要求编写程序

题目说明

请各位考生从课程信息发布网站下载数据文件 **data.txt**，然后将该数据文件**手动**保存在 **D 盘根目录**下。该文件中的数据以文本形式存储，每行包含 3 列数据，格式如下：A 列和 B 列都是单词，C 列是同一行 A 列和 B 列中单词共同出现的次数，列之间用 **tab** 隔开。文件总行数在千行以上，A 列或 B 列的单词可能重复，如下面例子所示：

A	B	C
打	电话	2
喝	啤酒	3
打	人	5
吃	苹果	6
打	电话	1
...

上面第一行表示“打”和“电话”两个单词在文档 1 中共同出现了 2 次，最后一行表示这两个单词在文档 2 中一起出现了 1 次，诸如此类。

定义结构体 **unit** 和 **unita**，用来存储如下数据：

```
typedef struct {  
    string a; //存储A列的单词  
    string b; //存储B列的单词  
    int cab; //存储单词a和单词b在所有文档中共同出现次数的总和  
} unit;
```

```
typedef struct{  
    string a; //A列的单词  
    int ca; //单词a在所有文档中出现的次数总和  
}unita;
```

比如在上面的示例中，“打”和“电话”共同出现的次数总和为 $2+1=3$ 次，“打”单独出现的次数总和为 $2+5+1=8$ 次。

请按要求依次完成如下操作：

- 1、编写一个函数 `read_data`，读取 `data.txt` 文件中的所有数据，并将它们存放在一个 `unit` 的向量 `vunit` 中，要求向量里所有单元的 `a` 和 `b` 都不重复，对重复的 `a` 和 `b`，将它们共同出现次数累加到 `c` 中。
(注意：向量不同单元中的单个 `a` 或单个 `b` 可以相同，但不允许 `a` 和 `b` 同时相同)
- 2、编写一个函数 `cal_counts`，以引用方式将 `vunit` 传递给该函数，并计算 A 列单词在所有文档中出现的次数，将计算结果保存在另一个向量 `vector<unita> vunita` 中。
- 3、编写一个排序函数 `sort_units`，将向量 `vunit` 中的元素按照出现次数 `cab` 由高到低排序。
- 4、编写一个重载的排序函数 `sort_units`，将向量 `vunita` 中的元素按照出现次数 `ca` 由高到低排序。

- 5、编写一个函数 `print_top_units`，将排序后的向量 `vunita` 中的前 20 个元素按下面显示格式输出到屏幕上：

显式 a，显示宽度为 10，不足的以“S”补齐，向左对齐	显式 ca，显式宽度 5，不足的以“\$”补齐，向左对齐
打	\$\$\$\$8
吃	\$\$\$\$6

- 6、编写一个条件概率计算函数 `cal_probs`，以引用方式将 `vunit` 和 `vunita` 传递给该函数，计算一个概率表，并将该概率表输出到文件 `prob.txt` 中（同样存放在 d 盘根目录下），文件每一行格式为：A 列单词 a ||| B 列单词 b ||| $p(b|a)$ ，即给定 A 列一个单词的情况下，B 列某个单词出现的概率 $p(b|a) = c_{ab}/c_a$ ，如按照上面的例子 $p(\text{电话}|\text{打}) = 3/8 = 0.375$ 。按照 `vunit` 中元素的顺序将以上概率输出到文件 `prob.txt` 中。（注意，只需计算 `vunit` 中出现的 A 列单词和 B 列单词对的概率，`vunit` 中没有出现的 A、B 列单词组合无需考虑）
- 7、main 函数如下：

```
int main() {
    string rfn = "d:\\data.txt";
    string wfn = "d:\\prob.txt";
    vector<unit> vunit;
    vector<unita> vunita;

    read_data(rfn, vunit);
    cal_counts(vunit, vunita);
    sort_units(vunit);
    sort_units(vunita);
    cout << "A列单词出现次数前20: " << endl;
    print_top_units(vunita);
    cout << "计算条件概率并输出到: " << wfn << endl;
    cal_probs(vunit, vunita, wfn);

    return 0;
}
```

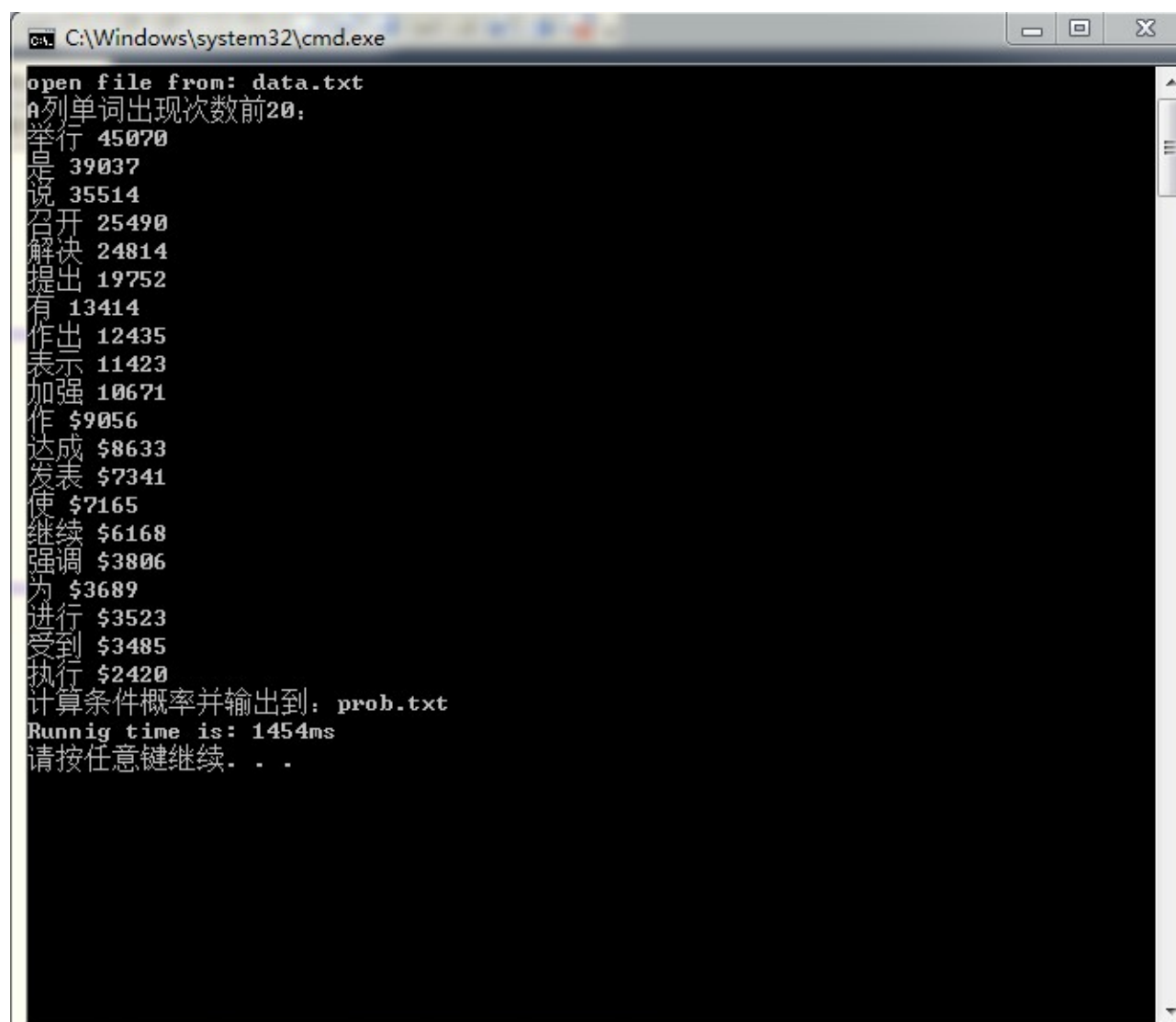
注意：不允许修改 main 函数, 每修改一处, 扣 3 分;

评分标准

(编程题满分为 80 分)

大项	子项	评分项	应得分	实得分
正 确 性 70 分	结果（70 分） 含编译子项 5 分	文件读取函数 read_data	15	
		次数统计函数 cal_counts	15	
		两个向量排序函数 sort_units	共 10 分，每个 5 分	
		向量打印函数 print_top_units	10	
		概率计算和输出函数 cal_probs	15	
		上述各项都不得分	见编译子项	本项不得分
		程序运行出现异常	-10	
		程序死循环	-10	
		修改 main 函数	-3（每处修改）	
	编译（5 分）	编译连接均通过（无 warning）	5	
		编译连接均通过（有 warning）	4	
		编译通过、连接不通过	3	
		编译、连接均不通过	0	
可 读 性 10 分	缩进对齐（4 分）	正确运用缩进对齐规则	4	
		有缩进对齐但不完全符合要求	2	
		没有使用缩进对齐规则	0	
	注释（3 分）	有详细且正确的注释	3	
		有注释，但不够详细	2	
		完全没有注释	0	
	变量命名（3 分）	变量命名有规则	3	
		变量命名有规则、但规则使用不一致	2	
		变量命名无规则	0	
总分（满分 80 分）				

程序运行结果截图：



A screenshot of a Windows command prompt window titled "C:\Windows\system32\cmd.exe". The window has standard Windows window controls (minimize, maximize, close) in the top right corner. The command prompt displays the following text:

```
open file from: data.txt
A列单词出现次数前20:
举行 45070
是 39037
说 35514
召开 25490
解决 24814
提出 19752
有 13414
作出 12435
表示 11423
加强 10671
作 9056
达成 8633
发表 7341
使 7165
继续 6168
强调 3806
为 3689
进行 3523
受到 3485
执行 2420
计算条件概率并输出到: prob.txt
Runnig time is: 1454ms
请按任意键继续. . .
```

The text is displayed in a monospaced font on a black background. The window has a vertical scrollbar on the right side.