10. 字典+集合 课后作业

1）电话簿管理 （20 分）

编写程序，完成对快递公司(见下表)的电话进行管理，要求完成以下功能：

- 快递公司查询
- 快递公司增加
- 快递公司修改
- 快递公司删除
- 快递公司遍历

快递公司名称	电话号码
顺丰速运	95338
申通快递	95543
韵达快递	95546
圆通速递	95554
中通速递	95311
天天快递	4001888888
京东物流	950616
百世快递	95320

运行效果：

```
1--快递公司查询
2--快递公司增加
3--快递公司修改
4--快递公司删除
5--快递公司遍历
0--退出
输入您的选择(0~5)：1
输入快递公司名称：京东物流
950616
------------------------------
1--快递公司查询
2--快递公司增加
3--快递公司修改
4--快递公司删除
5--快递公司遍历
0--退出
输入您的选择(0~5)：2
新增加的快递公司名称：兄弟快递
该公司的电话号码：121212
已经增加了兄弟快递。
------------------------------
1--快递公司查询
2--快递公司增加
3--快递公司修改
4--快递公司删除
5--快递公司遍历
0--退出
输入您的选择(0~5)：3
欲修改的快递公司名称：兄弟快递
该公司的电话号码：212121
已经修改了兄弟快递
------------------------------
1--快递公司查询
2--快递公司增加
3--快递公司修改
4--快递公司删除
5--快递公司遍历
0--退出
输入您的选择(0~5)：4
欲删除的快递公司名称：兄弟快递
已经删除了兄弟快递。
------------------------------
1--快递公司查询
2--快递公司增加
3--快递公司修改
4--快递公司删除
5--快递公司遍历
0--退出
输入您的选择(0~5)：5
顺丰速运          95338
申通快递          95543
韵达快递          95546
圆通速递          95554
中通速递          95311
天天快递       4001888888
京东物流         950616
百世快递          95320
------------------------------
1--快递公司查询
2--快递公司增加
3--快递公司修改
4--快递公司删除
5--快递公司遍历
0--退出
输入您的选择(0~5)：0
```

2）词频分析 （20 分）

对较长文章的词频进行统计分析是自然语言处理中的一个重要任务，为后续的文本分析、情感分析、机器翻译等任务做准备。下面请对莎士比亚的喜剧《罗密欧与朱丽叶》进行词频统计分析。要求：

- `txt = open('Romeo+Juliet.txt', 'r', encoding='utf-8').read()`
- 同一单词有大小写不同的形式，计数时不区分大小写。
- 英文单词的分隔可能是：空格、标点符号、特殊字符等，为了统一分隔方式，先将各种符号转换成空格。假设有：`'!"#$%&()*+-,./:;<=>?@[\\]^_{|}~`'` 用字符串的 **`txt.replace()`**方法。
- 用字符串的 **`txt.split()`**方法将每个单词分隔出来。
- 统计每个单词的出现次数。
- 按照由大到小的次序显示出前 30 个单词。

显示结果：

```
1   the         812
2   and         776
3   to          613
4   i           593
5   a           537
6   of          487
7   in          382
8   is          376
9   my          376
10  that        354
11  romeo       307
12  you         306
13  with        288
14  thou        279
15  not         272
16  me          267
17  for         255
18  it          235
19  this        233
20  be          221
21  but         190
22  juliet      178
23  thy         171
24  as          168
25  what        167
26  her         166
27  his         157
28  nurse       154
29  will        153
30  so          152
```

观察输出结果可以看到，高频单词大多是冠词、代词、连词等，这说明作者喜欢使用定冠词 the、连词 and。还可以看到，Romeo 出现了 307 次，几乎是 Juliet 出现 178 次的一倍。

3）学生信息管理 （20 分）

根据下面的表格，应用嵌套字典（学号为键，其余为值，值为内层字典）的方法，编写程序，完成以下功能：

- 显示 ID 是 202201 的姓名，及其全部信息；
- 按照输入的 ID 进行查询；
- 显示全部信息；
- 统计男女各有多少人，输出年龄大于 18 岁的学生姓名；
- 显示成绩最高的学生信息；
- 退出程序。

学号	姓名	性别	年龄	成绩
200001	张三	男	19	598
200016	芳芳	女	19	605
202201	圆圆	女	18	586
202336	李四	男	18	635
202318	王五	男	18	618
202112	佳佳	女	18	620

运行效果：

```
1--显示ID是202201的姓名，及其全部信息
2--按照输入的ID进行查询
3--显示全部信息
4--统计男女各有多少人，输出年龄大于18岁的学生姓名
5--显示成绩最高的学生信息
0--退出程序
请输入您的选择：1
ID是202201的姓名：圆圆
{'姓名': '圆圆', '性别': '女', '年龄': 18, '成绩': 586}
----------------------------------------
1--显示ID是202201的姓名，及其全部信息
2--按照输入的ID进行查询
3--显示全部信息
4--统计男女各有多少人，输出年龄大于18岁的学生姓名
5--显示成绩最高的学生信息
0--退出程序
请输入您的选择：2
请输入学生的ID：200001
{'姓名': '张三', '性别': '男', '年龄': 19, '成绩': 598}
----------------------------------------
1--显示ID是202201的姓名，及其全部信息
2--按照输入的ID进行查询
3--显示全部信息
4--统计男女各有多少人，输出年龄大于18岁的学生姓名
5--显示成绩最高的学生信息
0--退出程序
请输入您的选择：3
200001 张三 男 19 598
200016 芳芳 女 19 605
202201 圆圆 女 18 586
202336 李四 男 18 635
202318 王五 男 18 618
202112 佳佳 女 18 620
----------------------------------------
```

```
1--显示ID是202201的姓名，及其全部信息
2--按照输入的ID进行查询
3--显示全部信息
4--统计男女各有多少人，输出年龄大于18岁的学生姓名
5--显示成绩最高的学生信息
0--退出程序
请输入您的选择：4
男=3，女=3
>18岁： ['张三', '芳芳']
----------------------------------------
1--显示ID是202201的姓名，及其全部信息
2--按照输入的ID进行查询
3--显示全部信息
4--统计男女各有多少人，输出年龄大于18岁的学生姓名
5--显示成绩最高的学生信息
0--退出程序
请输入您的选择：5
最大值： 202336
{'姓名': '李四', '性别': '男', '年龄': 18, '成绩': 635}
----------------------------------------
1--显示ID是202201的姓名，及其全部信息
2--按照输入的ID进行查询
3--显示全部信息
4--统计男女各有多少人，输出年龄大于18岁的学生姓名
5--显示成绩最高的学生信息
0--退出程序
请输入您的选择：0
```

4） 调查问卷（20 分）

复旦大学为了合理安排后续教学，需对学生曾经学过的程序设计语言进行统计。下表是对 Python、VB、C 语言的调查问卷。编写程序，完成以下功能：

- 统计参加调查问卷的所有学生名单，并输出；
- 统计学过两门计算机语言的学生名单，并输出；
- 统计**仅**学过 Python 语言的学生名单，并输出；
- 统计**仅**学过 VB 语言的学生名单，并输出；
- 统计**仅**学过 C 语言的学生名单，并输出；
- 统计**仅**学过 1 门计算机语言的学生名单，并输出。

学生姓名	Python	VB	C
赵	√		√
钱	√	√	√
孙		√	
李	√		
周	√		√
吴		√	
郑		√	
王			√
冯	√		
陈			√
褚	√	√	
卫		√	√

运行效果：

```
参加调查问卷的所有学生名单：
吴，王，孙，钱，陈，李，褚，周，郑，卫，赵，冯，

学过两门计算机语言的学生名单：
周，卫，赵，钱，褚，

仅学过Python语言的学生名单：
李，冯，

仅学过VB语言的学生名单：
吴，郑，孙，

仅学过C语言的学生名单：
王，陈，

仅学过一门计算机语言的学生名单：
吴，王，孙，陈，李，郑，冯，
```

5）红楼梦 （20 分）

中文词汇量是衡量一个人中文水平的重要指标之一。中文词汇量的计算方法通常是将一个人所掌握的所有中文单词相加，这些单词可以包括常用词汇、专业词汇、成语、俗语等。不同的人在不同的领域和环境下所掌握的中文词汇量会有所不同。一般来说，一个普通人的中文词汇量在 5000 个左右，如果能达到 1 万个就比较高了。

通过对《**红楼梦**》一书的词汇数量统计，从一个侧面看看**曹雪芹**的词汇量有多少？

请编写代码，首先打开“红楼梦.txt”文件，然后用结巴分词库分词为列表，再去掉重复的词汇，还需要去掉常用的中文标点符号，最后统计词汇量，并显示出来。

中文的标点符号主要有：

'。', ',', '，', '、', '？', '！', '：', '；', '“', '”', '《', '》', '——', '……'

运行结果如下：

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\Sam2023\AppData\Local\Temp\jieba.cache
Loading model cost 0.689 seconds.
Prefix dict has been built successfully.
《红楼梦》中包含的词汇量：45,506个。
```

观察输出结果可以看到，**曹雪芹**的写作词汇量几乎达到了普通人的十倍。一般来说，一个人在听、说、读、写过程中，写作词汇一般是最少，听力词汇是最多的，也就是说，你能听懂的，不见得你能自己写出来。