**广东外语外贸大学信息科学与技术学院**

**《文本信息处理工程实践》课程作业**

**第1次作业：字典实现词频统计**

姓名: \_\_\_高子雄、游畅\_\_\_\_\_\_\_\_\_\_\_

学号： \_20181002989、20181003005\_

所在班级：\_\_\_软工1804\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **作业算法描述**

本次作业的的实现思路如下：

1. 将文本读入一个字符串；
2. 将字符串按空格进行切分，存进列表中；

3、建立空字典，并以词作为key，出现次数作为value；

4、使用sorted函数，分别按照字母表、词频输出字典。

5、输出字符串列表长度作为总词数。

**2.实现要点**

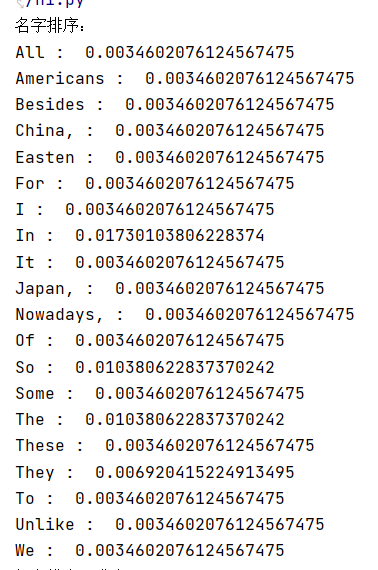
1. 将文本读入字符串后，使用strip与slipt函数进行处理以便去除“看不间的”换行符、制表符等；
2. 使用单层的for语句遍历处理后的字符串，若str[i]首次出现在dict.key()中，将其加入字典，并为value赋值1；若str[i]已经是dict.key()，其vale自增1；
3. 排序部分使用sorted函数结合lambda表达式完成。sorted()函数需要三个参数(a, b, c)，其中：

a为排序对象；

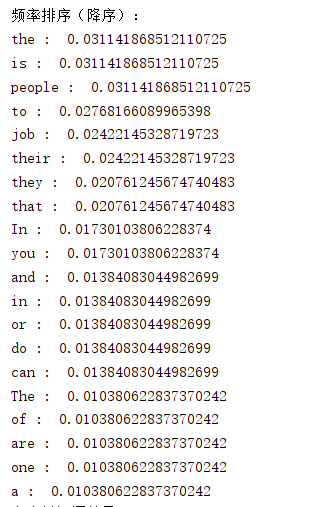
b为排序依据（可以是根据排序对象的特性设置，本次作业便分别使用了字典的key与value。）；

c为选择升序或降序。

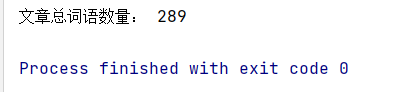
**3.程序演示**



图（1）：字母表排序，前二十位的词；



图（2）：依照字频排序（降序），前二十位的词；



图（3）：文章的总词数；

**4.总结和体会**

**高子雄：**

本次作业相对简单，各种知识在课堂上都有所讲解。不过要注意的是字符串文本中存在着单引号和双引号，所以单纯使用单引号和双引号将文本变成字符串是行不通的，还需要利用到转义字符。

**游畅：**

本次实践当中，遇到的一个小麻烦就是如何字符串读入字典。起先思路比较混乱，认为可以像赋值一样直接完成，随后就遇到了“AttributeError: 'list' object has no attribute 'keys'”的报错，原因是列表对象并没有下标，不能满足字典的结构需要。