**实验报告**

姓名：周蓓 学号：117060400109

**实验名称：**文本词频统计

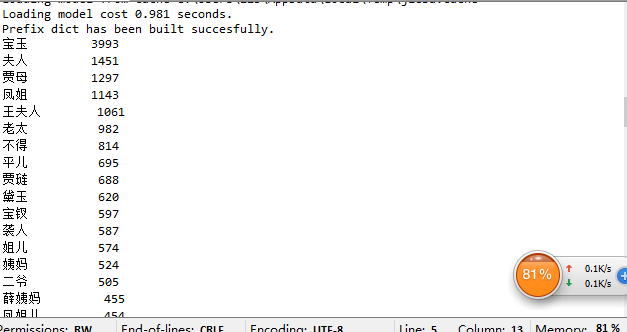
**实验目的：**使学生充分了解jieba库，熟悉jieba库的安装的相关操作。通过实际的操作，让学生掌握jieba库的使用和文本词频统计这两节的内容。巩固前面学习的内容，加强学生的自主操作能力。

**实验内容原理：**生日悖论分析、安装jieba库以及对红楼梦人物进行统计。由于jieba库是第三方库，不是python安装包自带，因此，需要通过pip指令安装。jieba库主要提供分词功能，其原理是利用一个中文库，将待分词的内容与分词词库进行比对，通过图结构和动态规划方法找到最大概率的词组。jieba库安装完就可以进行文本词频的统计。统计红楼梦中出场次数最多的人。

实验代码：

import jieba  
excludes = {'什么','一个','我们','你们','如今','说道','老太太','知道','姑娘','起来',  
            '这里','出来','众人','那里','奶奶','自己','太太','一面','只见','两个',  
            '没有','怎么','不是','这个','听见','这样','进来','咱们','就是','不知',  
            '东西','告诉','回来','只是','大家','老爷','只得','丫头','这些','他们',  
            '不敢','出去','所以','不过','不好','姐姐','的话','一时','过来'}  
txt = open('红楼梦.txt','r', encoding='GB18030').read()  
words = jieba.lcut\_for\_search(txt)  
counts = {}  
for word in words:  
    if len(word) == 1:  
        continue  
    elif word == '凤姐' or word == '熙凤':  
        rword = '凤姐'  
    elif word == '元春' or word == '贵妃':  
        rword = '元春'  
    else:  
        rword = word  
    counts[rword] = counts.get(rword,0) + 1  
  
for word in excludes:  
    del(counts[word])     
       
items = list(counts.items())  
items.sort(key=lambda x: x[1], reverse = True)  
for i in range(50):  
    word, count = items[i]  
    print("{0:<10}{1:>5}".format(word, count))

显示结果：



生日悖论代码：

from datetime import datetime

from random import \*

def generateSamples1(n:int):

birthdays = []

days = [31,29,31,30,31,30,31,31,30,31,30,31]

for i in range(n): # 随机生成n个人的生日（月，日）

month = randint(1,12)

day = randint(1,days[month-1])

someday = (month,day)

birthdays.append(someday)

return birthdays

def generateSamples2(n:int):

birthdays = []

days = [31,28,31,30,31,30,31,31,30,31,30,31]

year = randint(1950,2000)

for i in range(n): # 随机生成n个人的生日（月，日）

month = randint(1,12)

if (year%400==0) and (year%4==0 and year%100 != 0):

days[1] = 29

else:

days[1] = 28

day = randint(1,days[month-1])

someday = (month,day)

birthdays.append(someday)

return birthdays

def calSameBirthdayProb(birthdays:list, n:int):

num = 0

for i in range(n):

people = sample(birthdays,23)

pset = set(people)

if len(pset) != len(people):

num += 1

return num/n

def main():

while True:

n = int(input("输入一个整数："))

if n < 23:

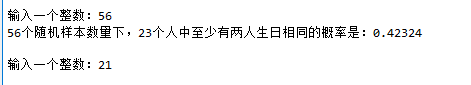
break

birthdays = generateSamples2(n)

print("{}个随机样本数量下，23个人中至少有两人生日相同的概率是：{}".format(n, calSameBirthdayProb(birthdays, 100000)))

main()

结果：



**实验总结：**经过实验课自己动手操作学会了安装第三方库（jieba,不是python安装包自带的），了解jieba库常用的函数，掌握了人物出场统计。通过自己动手，不断把程序完善。编写人物出场次数的程序，采用集合类型构建一个排除词汇库excludes，排除一些人名无关的词，进一步完善代码。通过运用jieba库解决人物出场次数问题，更深一步的了解程序基本结构这门课。