2022年12月－实验七

【实验名称】

数据及文件操作练习

【实验目的】

1. 统计hamlet.txt 中每个单词的出现频次，输出频次最高的10个单词。  
 1）保留介词、冠词、连词的情况下统计词频，并输出结果  
 2）去除介词、冠词、连词的情况下统计词频，并输出结果

2. 将整数12345分别写入文本文件test.txt和二进制文件test.dat，并比较两个文件的不同输出。

【实验内容】

1. 程序清单

**（1）统计每个单词的出现频次  
 1）保留介词、冠词、连词的情况下统计词频**

import re

res = {}

# 只读打开文件

with open('./hamlet.txt', 'r') as f:

    txt = f.read()

for line in txt.splitlines():

    line = re.sub(r'[+=^$#!]', ' ', line)  # 去除所有标点符号

    for word in line.split():

        flag = False

        if word[-1] == '-':

            up = word[:-1]

            flag = True

            break

        if flag:

            word = up + word  # 拼接末位单词

            flag = False

        res.setdefault(word.lower(), 0)

        res[word.lower()] += 1

values = sorted(res.values())

sortedres = sorted(res.items(),  key=lambda d: d[1], reverse=True)

for i in range(0, 10):

    print(sortedres[i][0])

**2）去除介词、冠词、连词的情况下统计词频**

import re

res = {}

# 只读打开文件

with open('./hamlet.txt', 'r') as f:

    txt = f.read()

for line in txt.splitlines():

    line = re.sub(r'[+=^$#!]', ' ', line)  # 去除所有标点符号

    for word in line.split():

        flag = False

        if word[-1] == '-':

            up = word[:-1]

            flag = True

            break

        if flag:

            word = up + word  # 拼接末位单词

            flag = False

        res.setdefault(word.lower(), 0)

        res[word.lower()] += 1

values = sorted(res.values())

sortedres = sorted(res.items(),  key=lambda d: d[1], reverse=True)

lista = ['the', 'a', 'an', 'at', 'on', 'behind', 'during', 'from',

         'into', 'and', 'but', 'or', 'so', 'however', 'although']

count = 0

for i in range(0, 20):

    if sortedres[i][0] not in lista:

        print(sortedres[i][0])

        count += 1

    if count == 10:

        break

**（2）将整数12345分别写入文本文件test.txt和二进制文件test.dat**

a = '12345'

with open('./test.txt', 'w') as f:

f.write(a)

with open('./test.dat', 'w') as f:

f.write(a)

with open('./test.txt', 'r') as f:

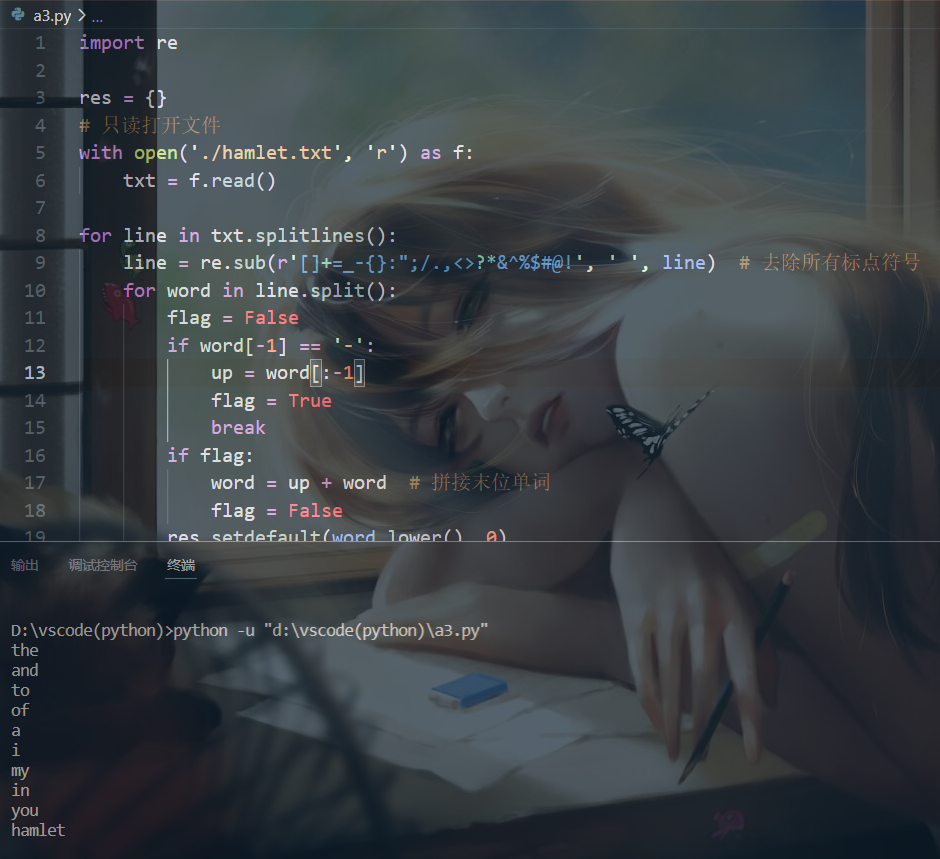
print(f.read())

with open('./test.dat', 'r') as f:

print(f.read())

1. 结果截图

**（1）统计每个单词的出现频次  
 1）保留介词、冠词、连词的情况下统计词频**

****

**2）去除介词、冠词、连词的情况下统计词频**

****

**（2）将整数12345分别写入文本文件test.txt和二进制文件test.dat**



【实验体会】

通过此次实验，使我更加熟悉了python中对于文件的操作和对于数据的处理，在数据的处理中，回顾了re库——正则表达式的使用，收获较大。