实验报告

实验名称：1.概述 & 预处理，统计文章单词词频

实验目的：通过对数据处理、统计，获得出现单词的出现频率

实验过程：

# 1st. Data read in & Data Process & Frequency FILE Create #

## 1. Variables Declairation 声明变量

## 2. Data read-in & data Pre-process 文件读入与数据预处理

## 3.Build Dictionary & Calculate Frequency 建立字典与计算频率

## 4. Sortig Data 对结果排序

## 5. Texting Data 数据文本化（方便后续写入文档）

## 6. Create a WordFrequence File 创建字频文档

# 2nd. Data Data Visualization with Matplotlib

## 1. Variables Declairation 声明变量

## 2.Search Top-5 Words 统计出现次数最多的五个单词

## 3. Ploting Figure 作图

## 4.File Renaming & Figure Saving 文件重命名与图像储存

程序实现：

1st. Data read in & Data Process & Frequency FILE Create

import math

import re

1. Variables Declairation 声明变量

data = []

total = 0

lstFrequency = []

txtFrenquency = []

dicFrequency = {}

2. Data read-in & data Pre-process 文件读入与数据预处理

with open('/Users/jason/Documents/GitHub/Information\_Search/20210309\_WordFrequency/Test\_Data/sw2078-ms98-a-trans.txt', 'r+') as dataFile:

for line in dataFile**.**readlines():

if line != None:

data**.**append(line**.**strip('\n'))

dataProcessed = ' '**.**join(data)**.**split(' ')

3.Build Dictionary & Calculate Frequency 建立字典与计算频率

for word in dataProcessed :

dicFrequency[word] = dicFrequency**.**get(word, 0) + 1

total += 1

4. Sortig Data 对结果排序

lstFrequency = list(dicFrequency**.**items())

print(lstFrequency)

lstFrequency**.**sort(key = lambda p: (-p[1], p[0]))

5. Texting Data 数据文本化（方便后续写入文档）

for fr in lstFrequency:

fr = list(fr)

txtFrenquency**.**append(fr[0] + ':' + ' ' + str(fr[1]) + '\n')

Result = ''**.**join(txtFrenquency)

6. Create a WordFrequence File 创建字频文档

fileName = dataFile**.**name

dot = re**.**compile('\.') # *Renaming*

fileName = re**.**sub(dot, '\_Frequency.',fileName) # *Renaming*

ResultFile = open(fileName, 'w')

ResultFile**.**write(Result)

dataFile**.**close()

ResultFile**.**close()

2nd. Data Data Visualization with Matplotlib

# *2nd. Data Data Visualization with Matplotlib*

import matplotlib**.**pyplot as plt

import numpy as np

import pandas as pd

1. Variables Declairation 声明变量

#*# 1. Variables Declairation 声明变量*

topWords = []

totalOrigin = total

class topWord:

def \_\_init\_\_(self, word, frequency):

self**.**word = word

self**.**frequency = frequency

2.Search Top-5 Words 统计出现次数最多的五个单词

#*# 2.Search Top-5 Words 统计出现次数最多的五个单词*

for i in range(0, 5):

tmp = topWord(lstFrequency[i][0], lstFrequency[i][1])

topWords**.**append(tmp)

total -= lstFrequency[i][1]

print(topWords, total)

3. Ploting Figure 作图

#*# 3. Ploting Figure 作图*

plt**.**figure(figsize=(3, 2), dpi=300, facecolor='#f2f2f2')

plt**.**title('Words Appearance Proportion', fontdict={'fontweight': '600', 'fontsize': '10'})

Labels = [topWords[0]**.**word, topWords[1]**.**word, topWords[2]**.**word, topWords[3]**.**word, topWords[4]**.**word, 'Others']

Values = [topWords[0]**.**frequency, topWords[1]**.**frequency, topWords[2]**.**frequency, topWords[3]**.**frequency, topWords[4]**.**frequency, total]

Colors = ['#F25C5C', '#FFCA40', '#6B98F2', '#2E8C83', '#402CBF', '#c6cdd7']

patches, texts, autotexts = plt**.**pie(Values, labels = Labels, colors=Colors, autopct = '%.2f %%', pctdistance = 0.8, labeldistance = 1.2, startangle = 30, textprops={'size': '6'}, counterclock=False)

plt**.**setp(autotexts, size='3')

autotexts[3]**.**set\_color('white'),autotexts[4]**.**set\_color('white')

plt**.**legend(loc="upper right",fontsize=4,bbox\_to\_anchor=(-0.1,1),borderaxespad=0.3)

4.File Renaming & Figure Saving 文件重命名与图像储存

#*# 4.File Renaming & Figure Saving 文件重命名与图像储存*

fileOriginName = dataFile**.**name

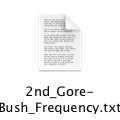
form = re**.**compile('(\.).\*')

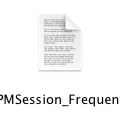
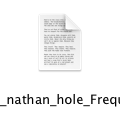
PltName = re**.**sub(form, '\_Proprotion.jpeg',fileOriginName)

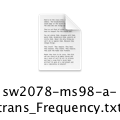
plt**.**savefig(PltName , dpi=300)

plt**.**show()

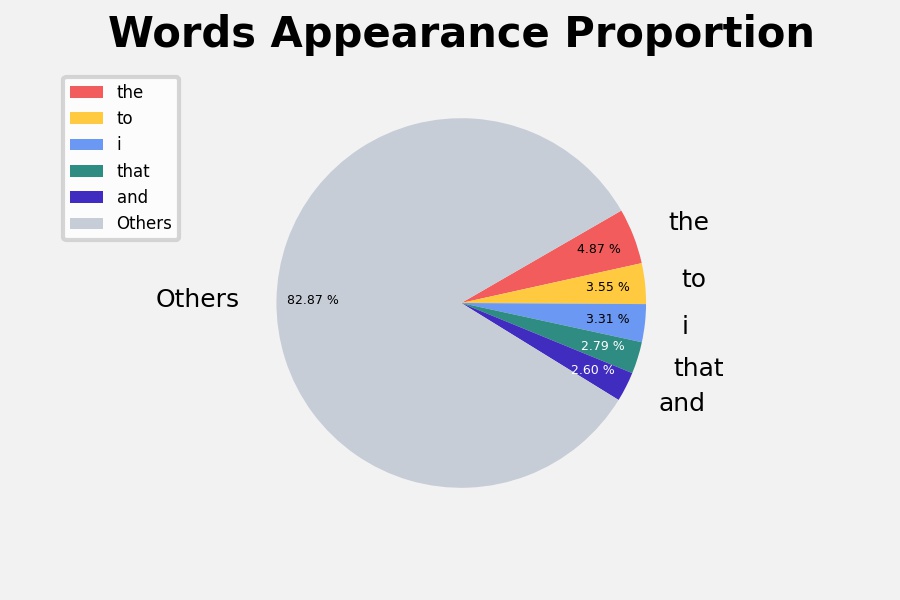
实验结果：通过读入文件统计所有文件的词频并生成\_Frequency 文件

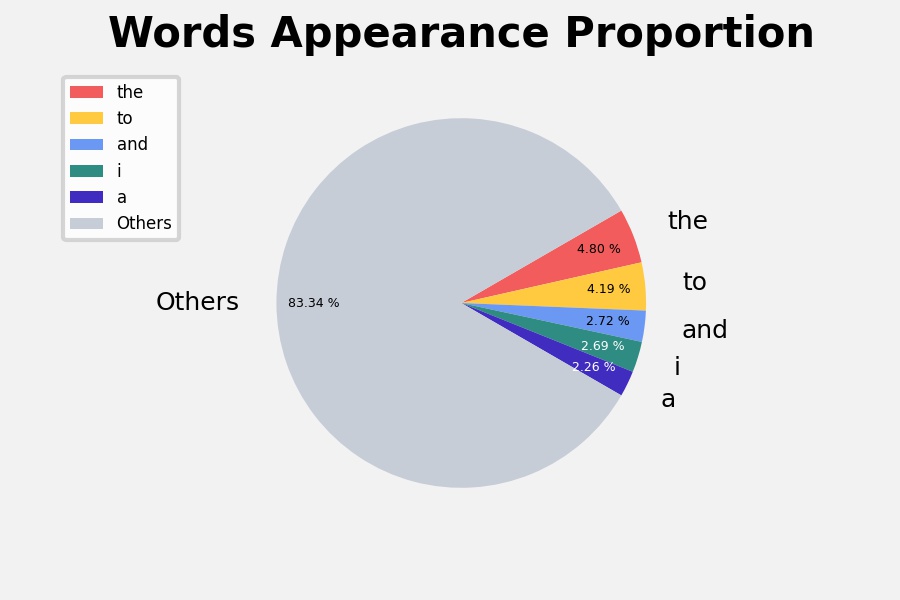
  

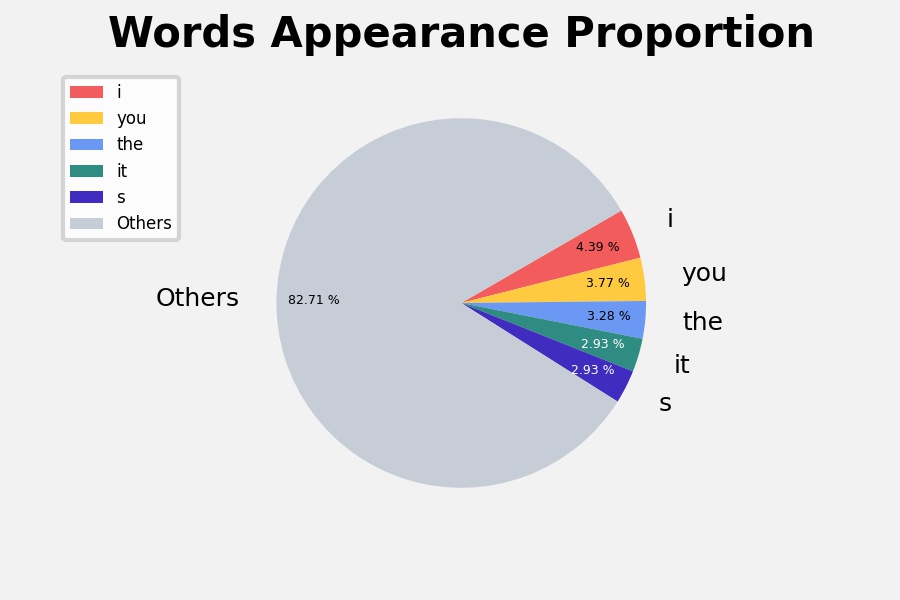
 

数据统计：使用matplotlab制作饼状图绘图

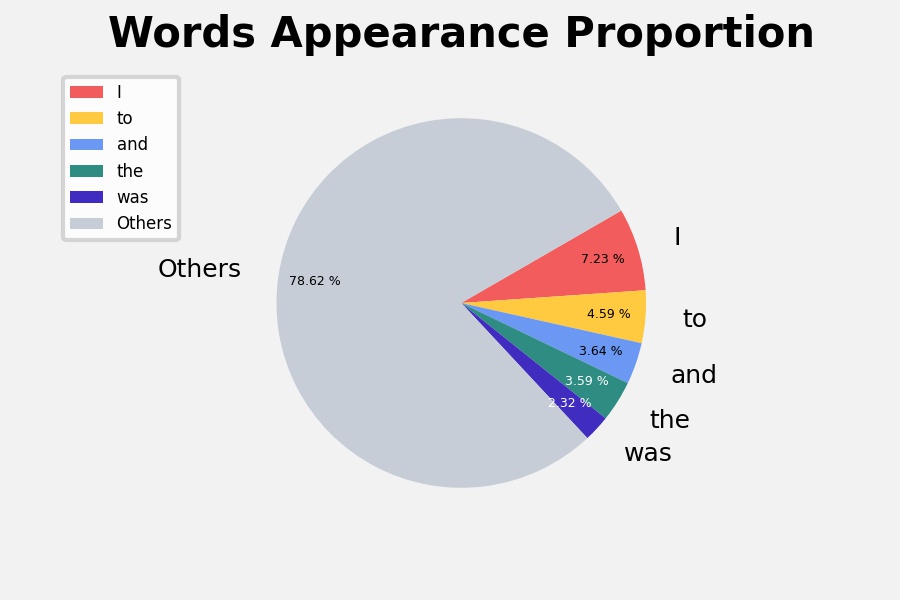
2nd\_Gore-Bush\_Frequency:

3rd\_Bush-Kerry\_Frequency:

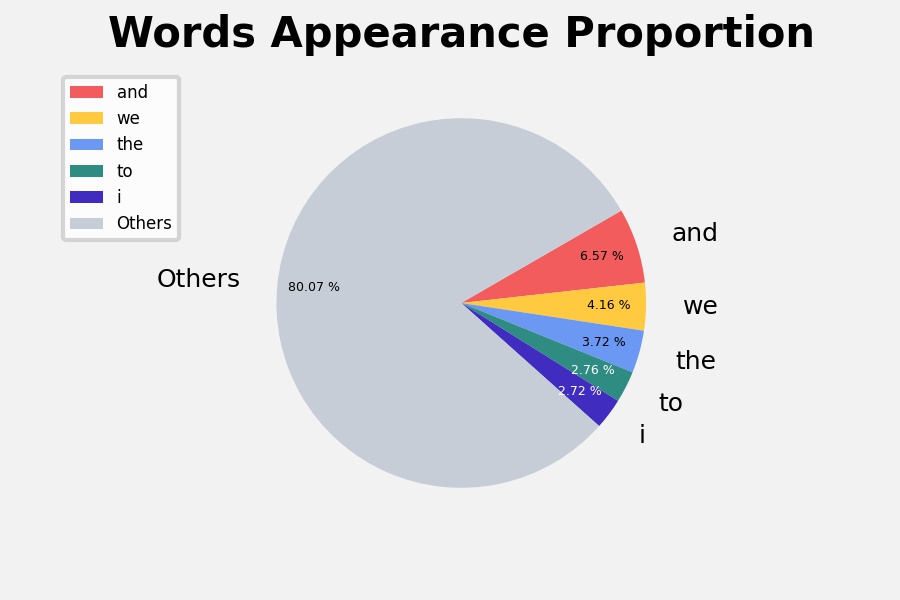


Bed012\_Frequency:

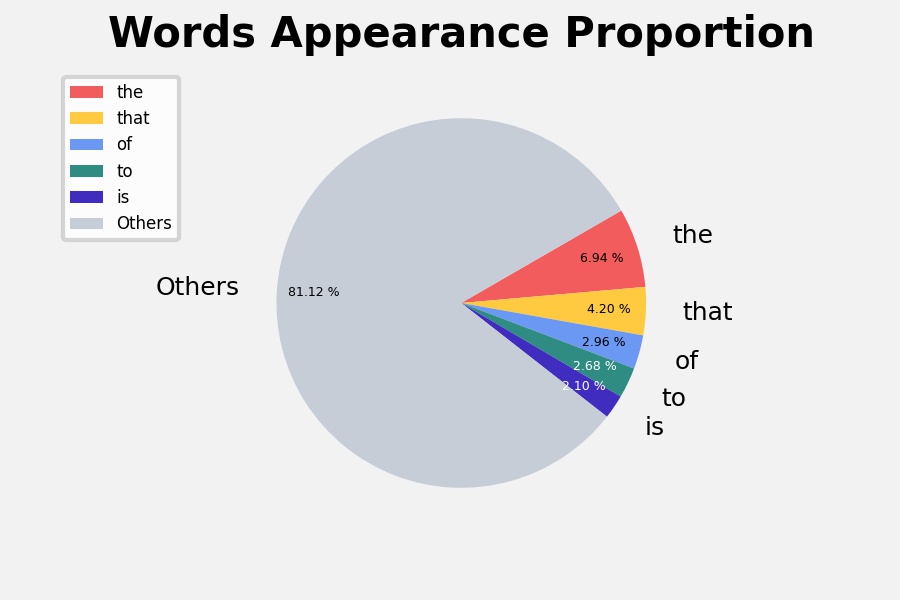
Bmr021\_Frequency:



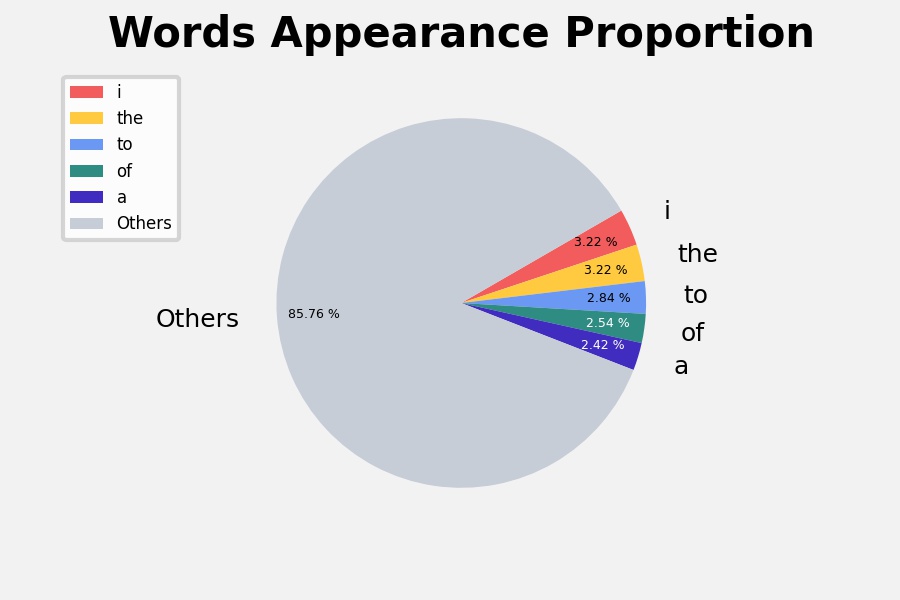
Day3PMSession\_Frequency:



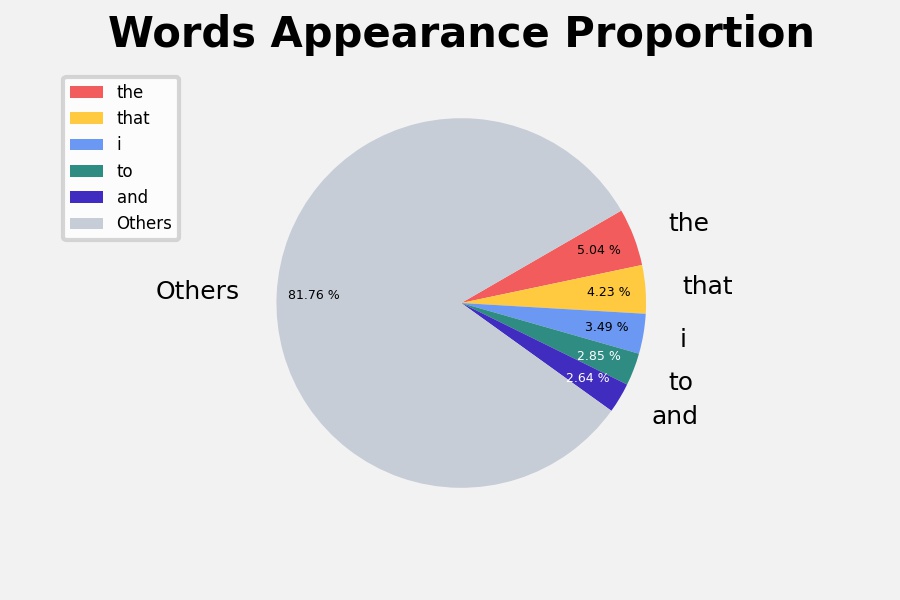
interview\_nathan\_hole:



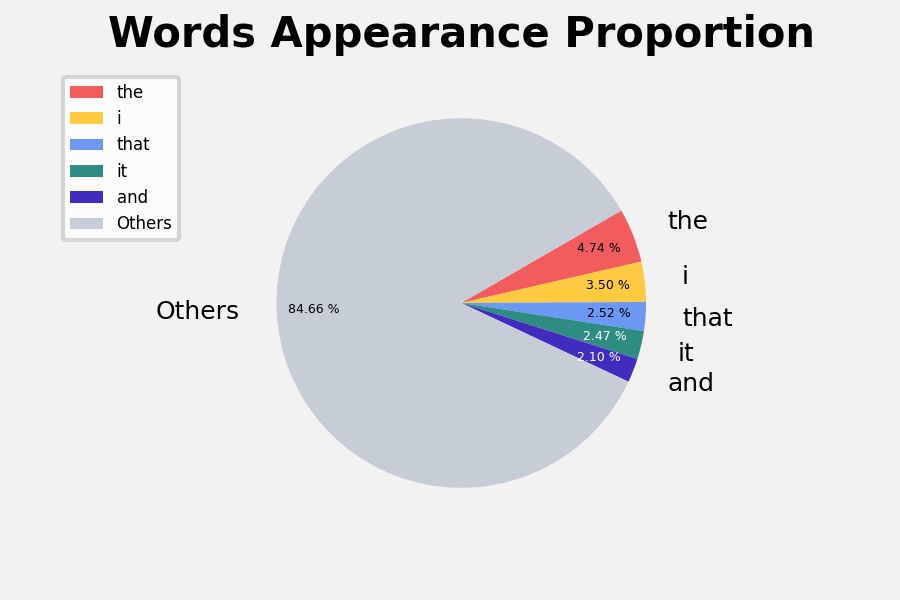
Lessig-court-transcript\_Frequency



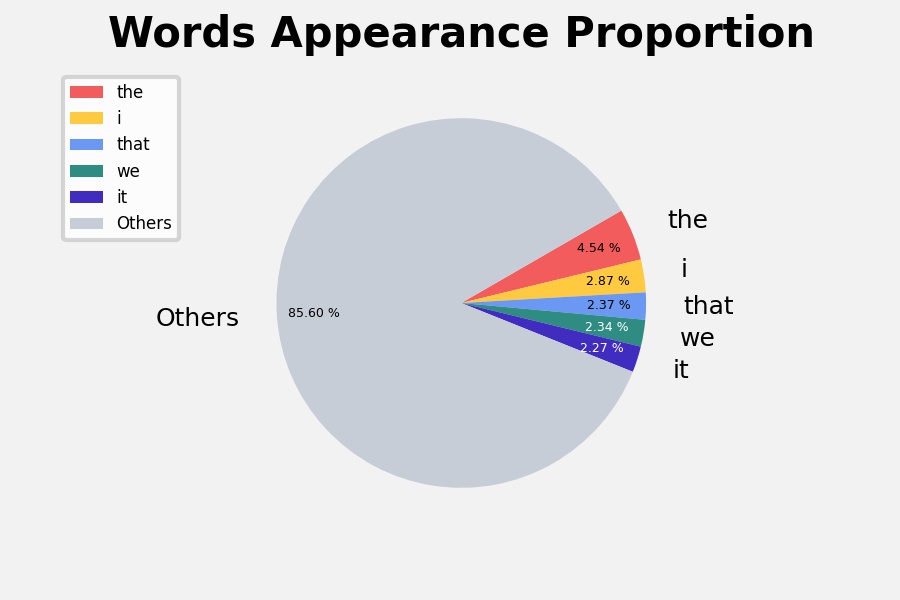
NapierDianne\_Frequency：



ReidSandra\_Frequency



sw2078-ms98-a-trans\_Frequency：



实验小结：通过本次实验，我了解了词汇的基本统计方法，熟悉了文件处理，以及可视化工具的使用。