统计历年四六级真题中的高频词

题目要求

应用所学知识，从2015年6月-2020年12月的大学英语四级考试真题中找出前300个高频英文单词，上面代码已经把试题文本保存在变量`cet4_text`中，请你在变量`cet4_text`的基础上进行后续操作。**题目资料见homework_4.zip**。

下载并解压`homework_4.zip`压缩文件包，解压后包含一个文件夹`cet-4-txt`，一个txt文本文件`stopwords.txt`，一个jupyter文件`home_work.ipynb`。在 jupyter中打开`home_work.ipynb`查看作业具体要求。在`home_work.ipynb`中完成作业后提交该文件。

注意:

- 就在该jupyter文件内作答；
- 可以使用课上未讲授过的python知识，如python的字典数据类型
- 不要把全部代码写在一个cell里；
- 为你的代码添加适当的注释；
- 最后的高频词表里不能包含停用词；
- 最后的高频词表既要给出单词，也要给出该单词的出现频率；
- 最后结果以列表形式呈现，如，`freq_300 = [(word_1, frequency_1), (word_1, frequency_1), ...]`。

题目背景知识

In [1]:
```
# 指定windows平台下Python运行时的默认编码类型为UTF-8
import _locale
_locale._getdefaultlocale = (lambda *args: ['zh_CN', 'utf8'])
```

In [2]:
```
# 导入os模块
import os
```

In [3]:
```
# 使用相对路径指定文件夹路径
data_dir = './cet-4-txt/'
```

In [4]:
```
# 使用os模块的listdir函数返回文件夹中的所有文件
files = os.listdir(data_dir)
```

In [5]:
```
# files
```

In [6]:
```
files[0]
```

Out[6]:
```
'2020年09月四级真题第1套.txt'
```

In [7]:
```
# 指定文件路径
file_path = data_dir + files[0]
```

In [8]:
```
file_path
```

Out[8]:
```
'./cet-4-txt/2020年09月四级真题第1套.txt'
```

In [9]:
```
# 使用open函数打开文本文件，r表示以阅读模式打开
f = open(file_path, 'r')
```

In [10…
```
# 调用read方法读取文件中的全部内容
content = f.read()
```

In [11…
```
# 关闭文本文件
f.close()
```

In [12…
```
type(content)
```

Out[12…
```
str
```

In [13…
```
# 查看文本内容的前500个字符
content[:500]
```

Out[13…
```
'机密*启用前\n大 学 英 语 四 级 考 试\nCOLLEGE ENGLISH TEST\n—Band Four—\n(2020年9月第1套)\n试 题 册\n\n敬 告 考 生\n\n一、在答题前，请认真完成以下内容:\n1. 请检查试题册背面条形码粘贴条、答题卡的印刷质量，如有问题及时向监考员反映，确认无误后完成以下两点要求。\n2. 请将试题册背面条形码粘贴条揭下后粘贴在答题卡1的条形码粘贴框内，并将姓名和准考证号填写在试题册背面相应位置。\n3. 请在答题卡1和答题卡2指定位置用黑色签字笔填写准考证号、姓名和学校名称，并用HB-2B铅笔将对应准考证号的信息点涂黑。\n二、在考试过程中，请注意以下内容:\n1. 所有题目必须在答题卡上规定位置作答，在试题册上或答题卡上非规定位置的作答一律无效。\n2. 请在规定时间内在答题卡指定位置依次完成作文、听力、阅读、翻译各部分考试，作答作文期间不得翻阅该试题册。听力录音播放完毕后，请立即停止作答，监考员将立即收回答题卡1，得到监考员指令后方可继续作答。\n3. 作文题内容印在试题册背面，作文题及其他主观题必须用黑色签字笔在答题卡指定区域内作答。\n4. 选择题均为单选题'
```

In [14…
```
# 查看文本内容的后500个字符
content[-500:]
```

Out[14…
```
'the boiling water immediately gave off pleasant fragrance. He drank a little water, feeling very refreshed. Therefore, tea was found in this way. Since then, tea has begun to be popular
```

```
in China. Tea gardens spread all over the country, and tea mer
chants became rich. Expensive and elegant tea sets have become
a symbol of social status. Today, tea is not only a healthy dr
ink, but also a part of Chinese culture. More and more interna
tional tourists learn about Chinese culture when they drink te
a.\n'
```

In [15…
```
# 打印输出文本的全部内容
# print(content)
```

In [16…
```
# 遍历文件夹中的全部英语四级试题文件，读取其中的文本内容，以字符串形式赋值
cet4_text = ''
for txt_file in files:
    txt_path = data_dir + txt_file
    f = open(txt_path, 'r')
    txt_str = f.read()
    f.close()
    cet4_text += txt_str + '\n'
```

In [17…
```
type(cet4_text)
```

Out[17…
```
str
```

In [18…
```
# 查看一下全部试题文本的字符长度
len(cet4_text)
```

Out[18…
```
842360
```

In [19…
```
# 从stopwords文件中读取英语中的停用词，赋值给列表变量stopwords_list
f = open('./stopwords.txt')
stopwords_str = f.read()
stopwords_list = eval(stopwords_str)
```

In [20…
```
type(stopwords_list)
```

Out[20…
```
list
```

In [21…
```
len(stopwords_list)
```

Out[21…
```
179
```

停用词：语言中包含的功能词，这些功能词极其普遍，与其他词相比，功能词没有什么实际含义，如系动词、助动词、介词、代词等。

In [22…
```
# 看一下停用词表包含哪些词语
# stopwords_list
```

题目作答区

注意：请先运行题干代码，然后在下面写你的答案。

In [23…

```
# 请作答
```