爬取百度百科5A景点摘要并实现分词

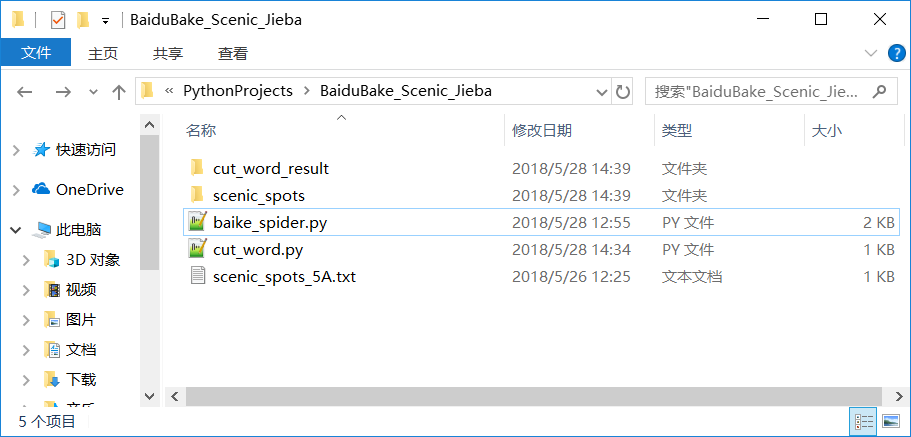
#一、编程环境

操作系统：Win 10

语言：Python 3.6

分词工具：结巴分词

#二、程序目录



这里baike\_spider.py用来爬取景点摘要，内容放在senic\_spots目录中

cut\_word.py用来分词，分词结果放在cut\_word\_result中

scenic\_spots\_5A.txt中列出了所要爬取的景点的名称，具体内容如下：

```

北京故宫

天坛公园

颐和园

八达岭

慕田峪长城

明十三陵

恭王府

北京奥林匹克公园

```

注意，scenic\_spots和cut\_word\_result这两个文件夹不需要提前创建，程序运行时会自动创建。

#三、爬取景点摘要

baike\_spider.py中的代码：

```

import os

import time

import codecs

import shutil

from selenium import webdriver

from selenium.webdriver.common.keys import Keys

driver = webdriver.Chrome(executable\_path="C:\Program Files (x86)\Google\Chrome\Application\chromedriver.exe")

def getInfoBox(spotname, filename):

try:

print(filename)

info = codecs.open(filename,'w','utf-8')

driver.get("http://baike.baidu.com/")

elem\_input = driver.find\_element\_by\_xpath("//form[@id='searchForm']/input")

time.sleep(2)

spotname = spotname.rstrip('\n') # 景点名称是从文件中读取的，含有换行符（最后一行的景点名称可能不含护身符）

elem\_input.send\_keys(spotname)

elem\_input.send\_keys(Keys.RETURN)

info.write(spotname + '\r\n') # codecs不支持'\n'换行

print (driver.current\_url)

print (driver.title)

elem\_value = driver.find\_elements\_by\_xpath("//div[@class='lemma-summary']/div")

for value in elem\_value:

print (value.text)

info.writelines(value.text + '\r\n')

time.sleep(2)

info.close()

except Exception as e:

print ("Error: ", e)

finally:

pass

def main():

# 创建路径

path = "scenic\_spots\\"

if os.path.isdir(path):

shutil.rmtree(path, True)

os.makedirs(path)

source = open("scenic\_spots\_5A.txt", 'r')

num = 1

for scenicspot in source:

name = "%03d" % num

fileName = path + str(name) + ".txt"

getInfoBox(scenicspot, fileName)

num += 1

print ('End Read Files!')

time.sleep(10)

source.close()

driver.close()

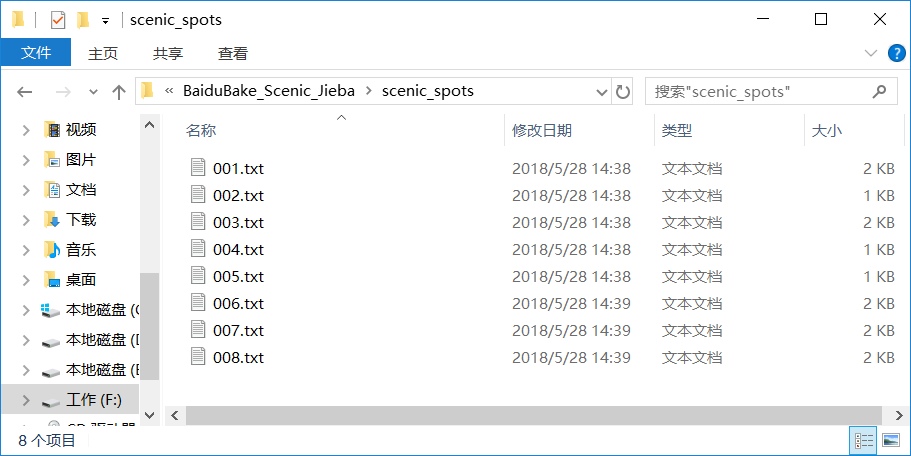
if \_\_name\_\_ == '\_\_main\_\_':

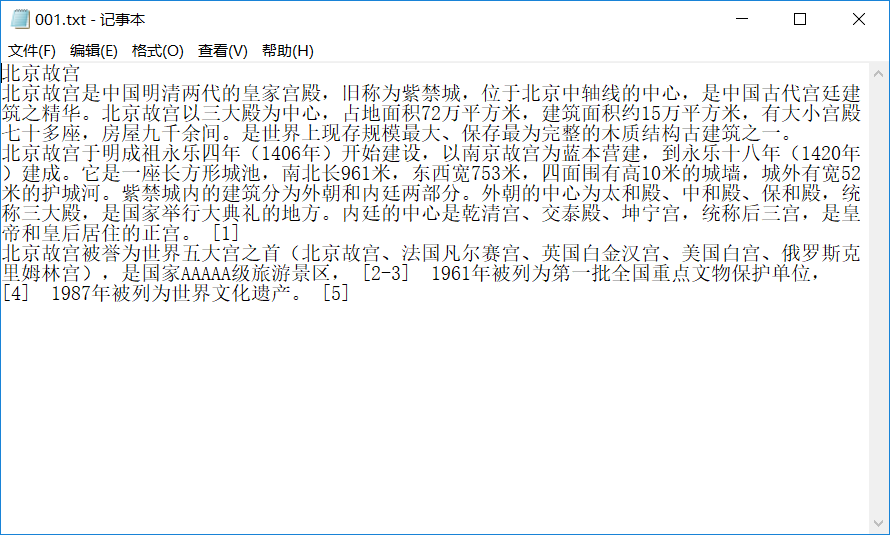
main()

```

运行结果：

在scenic\_spots目录下，生成了8个txt文件，每个文件存放一个景点的摘要内容





#四、实用结巴工具实现分词

cut\_word.py中的代码：

```

import sys

import codecs

import os

import shutil

import jieba

def read\_file\_cut():

#create path

path = "scenic\_spots\\"

respath = "cut\_word\_result\\"

if os.path.isdir(respath):

shutil.rmtree(respath, True)

os.makedirs(respath)

num = 1

while num <= 8:

name = "%03d" % num

fileName = path + str(name) + ".txt"

source = open(fileName, 'r', encoding = 'utf-8')

line = source.readline()

line = line.rstrip('\n')

resName = respath + str(name) + ".txt"

if os.path.exists(resName):

os.remove(resName)

result = codecs.open(resName, 'w', encoding = 'utf-8')

while line != "":

seglist = jieba.cut(line,cut\_all=False) #精确模式

output = ' '.join(list(seglist)) #空格拼接

print (output)

result.write(output + '\r\n')

line = source.readline()

else:

print ('End file: ' + str(num))

source.close()

result.close()

num += 1

else:

print ('End All')

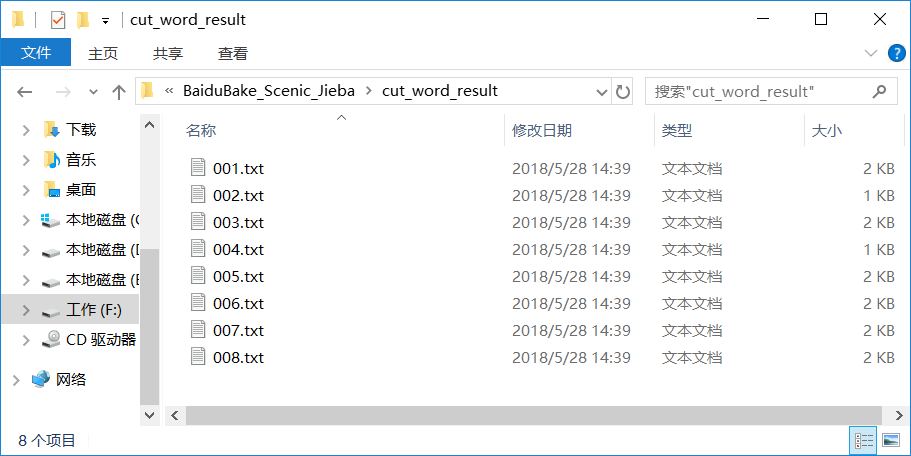
if \_\_name\_\_ == '\_\_main\_\_':

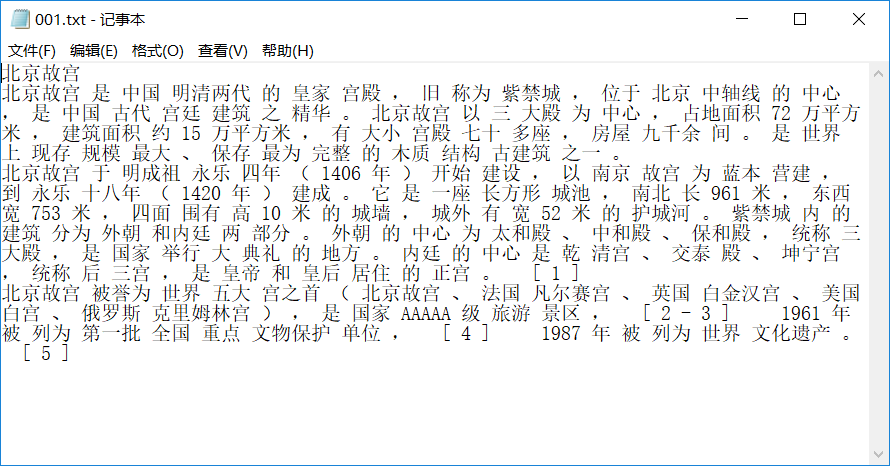
read\_file\_cut()

```

运行结果：

在cut\_word\_result目录下，生成了8个文件，每个文件存放的是分词后的内容：





#五、参考

https://blog.csdn.net/eastmount/article/details/50256163