**网络爬虫实训报告**

**学号： 5210162657 姓名： 肖光阳 班级： 计科1603**

**一、实训目的**

1、掌握字符串、列表、元组、字典、集合的应用。

2、掌握python程序控制结构的应用。

3、掌握函数、方法的应用。

4、掌握类、对象的应用。

5、掌握开发一个Python小工具项目所遵循的规范。

6．掌握python中模块的导入和引用。

7、掌握爬虫的概念原理及爬虫的过程。

8、了解有关爬虫的框架。

**二、实训内容**

1、分析爬取目的，确定抓取页面

2、整理爬虫常规思路

3、创建spider.py模块，用面向对象的方法写一个spider类

4、Html结构分析基本原则三条

5、正则分析HTML

6、正则分析获取主播姓名和观看人数

7、数据精炼

8、业务处理（数据排序）

9、案例总结（数据美化）

10、知识拓展

**三、设计方法**

**从起点中文网爬取一个小说的前5章**

**总体思路**

1、使用request库和beautifulsoup库爬取

2、设计一个小说爬取类，其中包含网页访问，文本处理等方法，存为一个模块文件Spider.py

3、在主程序文件main.py中使用这个类爬取5章小说内容

**四、设计代码**

Spider.py文件

*from* urllib *import* request

*from* bs4 *import* BeautifulSoup

class NovelGet(object):

def \_\_init\_\_(*self*, *url*):

self.html = ''

self.tag\_string = ''

self.url = url

def get\_html(*self*):

rep = request.urlopen(self.url)

html\_content = rep.read()

html\_content = str(html\_content, *encoding*='utf-8')

self.html = html\_content

*return* self.html

def get\_text(*self*):

soup = BeautifulSoup(self.html, 'html.parser')

read\_tag = soup.find('div', *class\_*='j\_readContent')

self.tag\_string = str(read\_tag.contents[1])

self.tag\_string = self.tag\_string.replace('<p>', '\n').replace('</p>', '')

*return* self.tag\_string

Main.py文件

*from* Spider *import* NovelGet

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

novel = ''

urls = ['https://read.qidian.com/chapter/-WkNjDCxgv6RTIpqx7GUJA2/YsQf3E-oiR62uJcMpdsVgA2',

'https://read.qidian.com/chapter/-WkNjDCxgv6RTIpqx7GUJA2/jThFps4zB\_Fp4rPq4Fd4KQ2',

'https://read.qidian.com/chapter/-WkNjDCxgv6RTIpqx7GUJA2/x2yh\_sYZBzX4p8iEw--PPw2',

'https://read.qidian.com/chapter/-WkNjDCxgv6RTIpqx7GUJA2/-LWEHxMgLab4p8iEw--PPw2',

'https://read.qidian.com/chapter/-WkNjDCxgv6RTIpqx7GUJA2/z8mhFSiOVO62uJcMpdsVgA2']

*for* index in range(5):

novel = novel + '\n================第' + str(index + 1) + '章================'

chapter = NovelGet(urls[index])

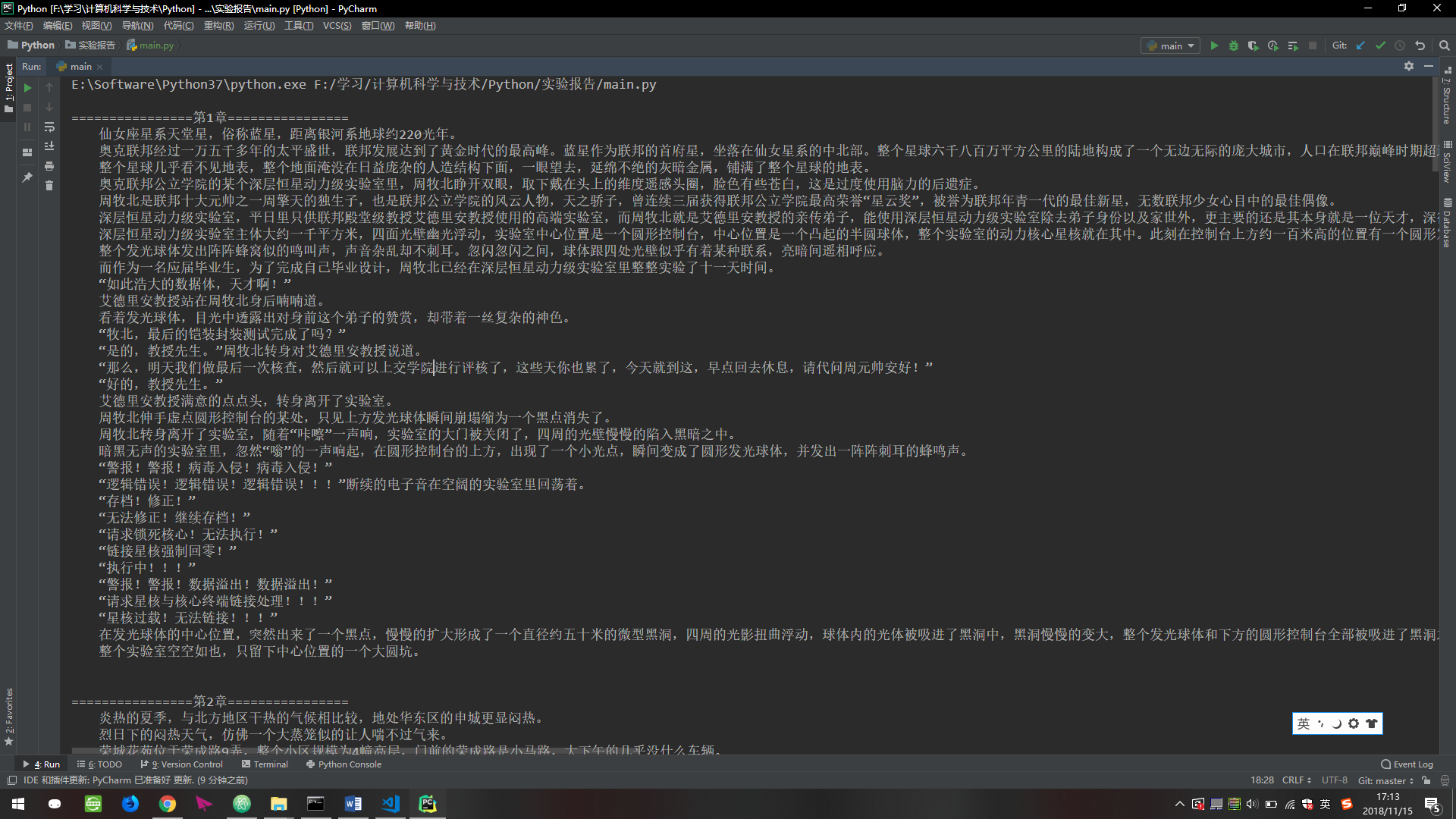
chapter.get\_html()

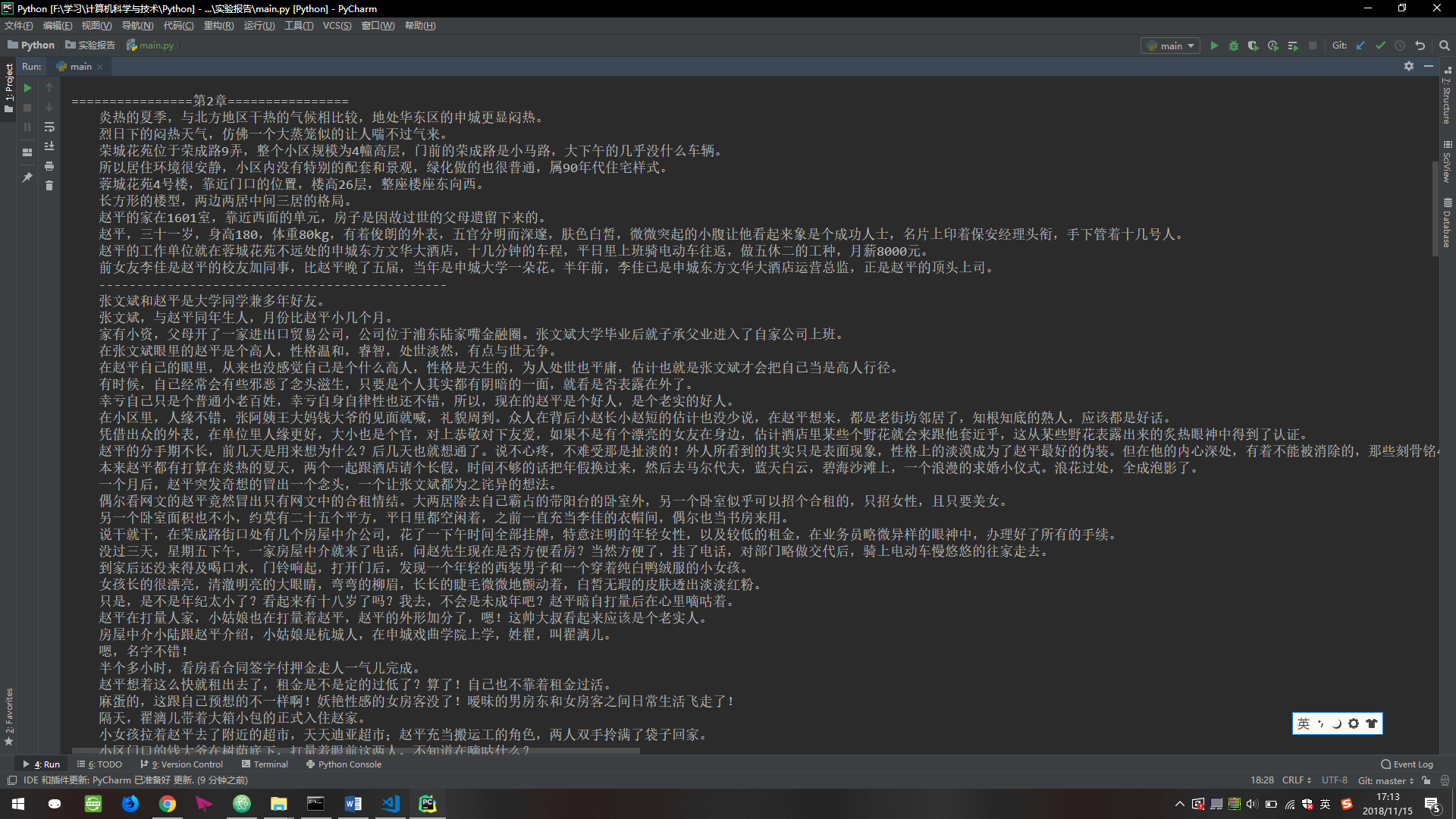
chapter.get\_text()

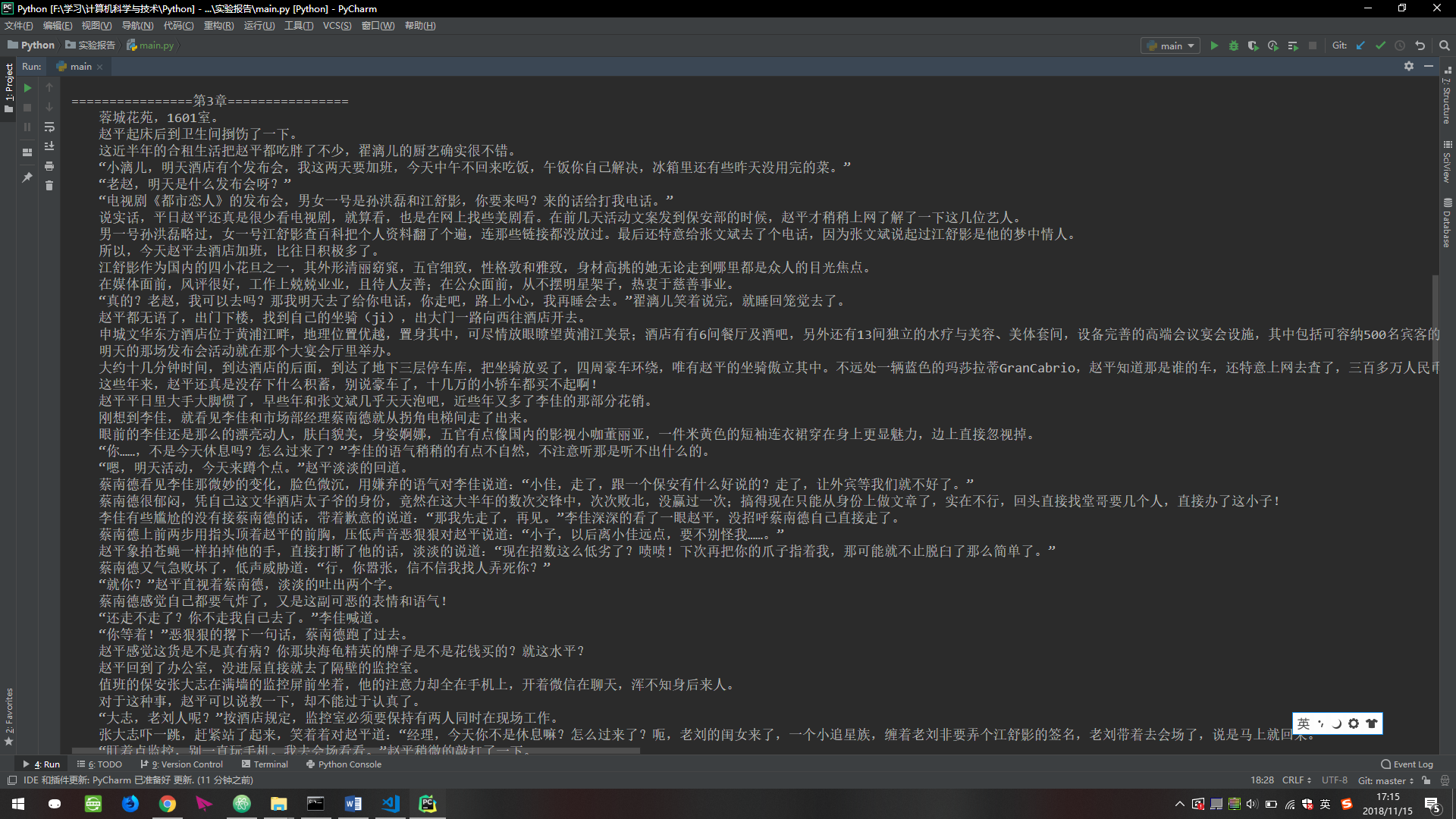
novel += chapter.tag\_string

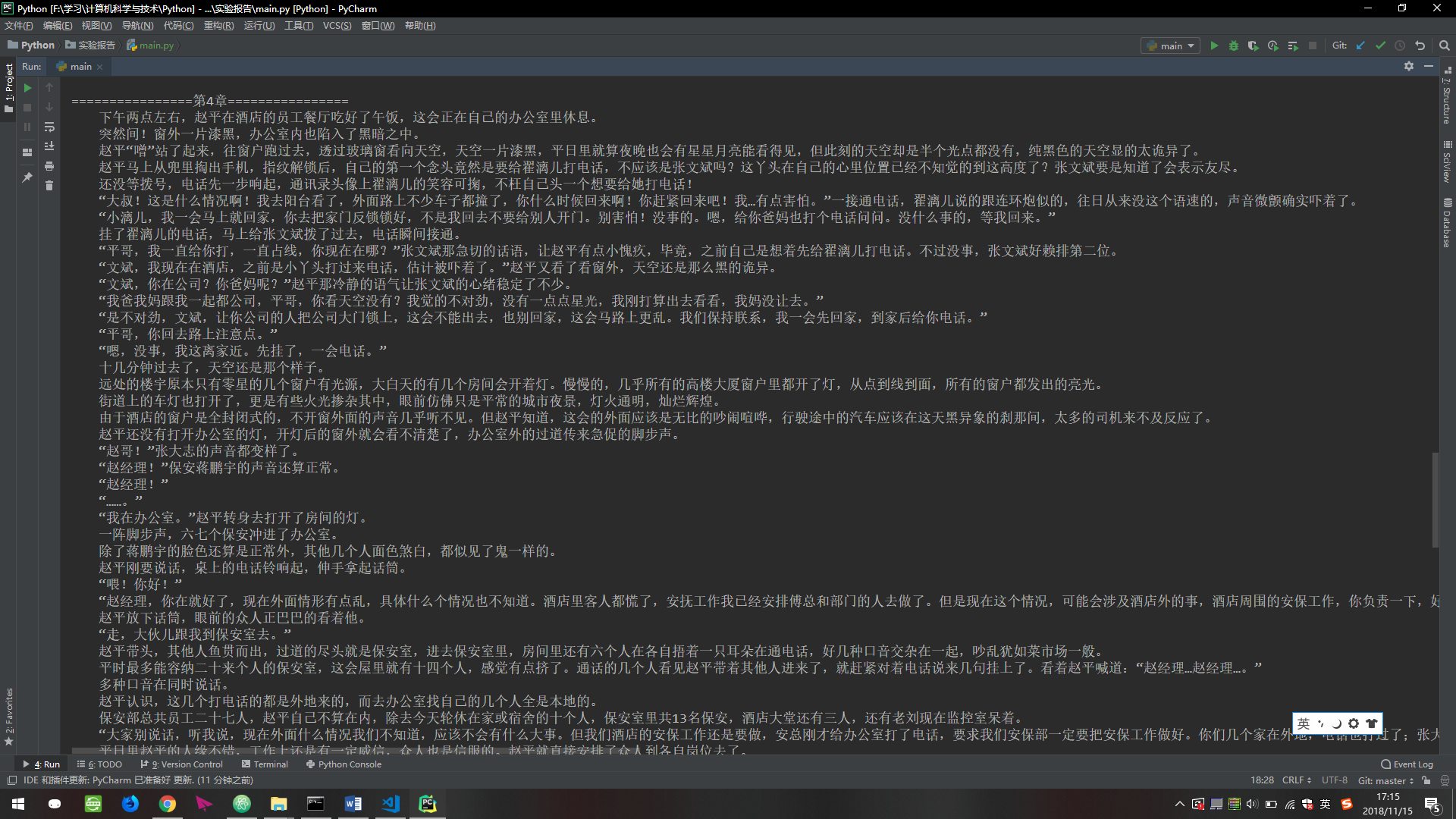
print(novel)

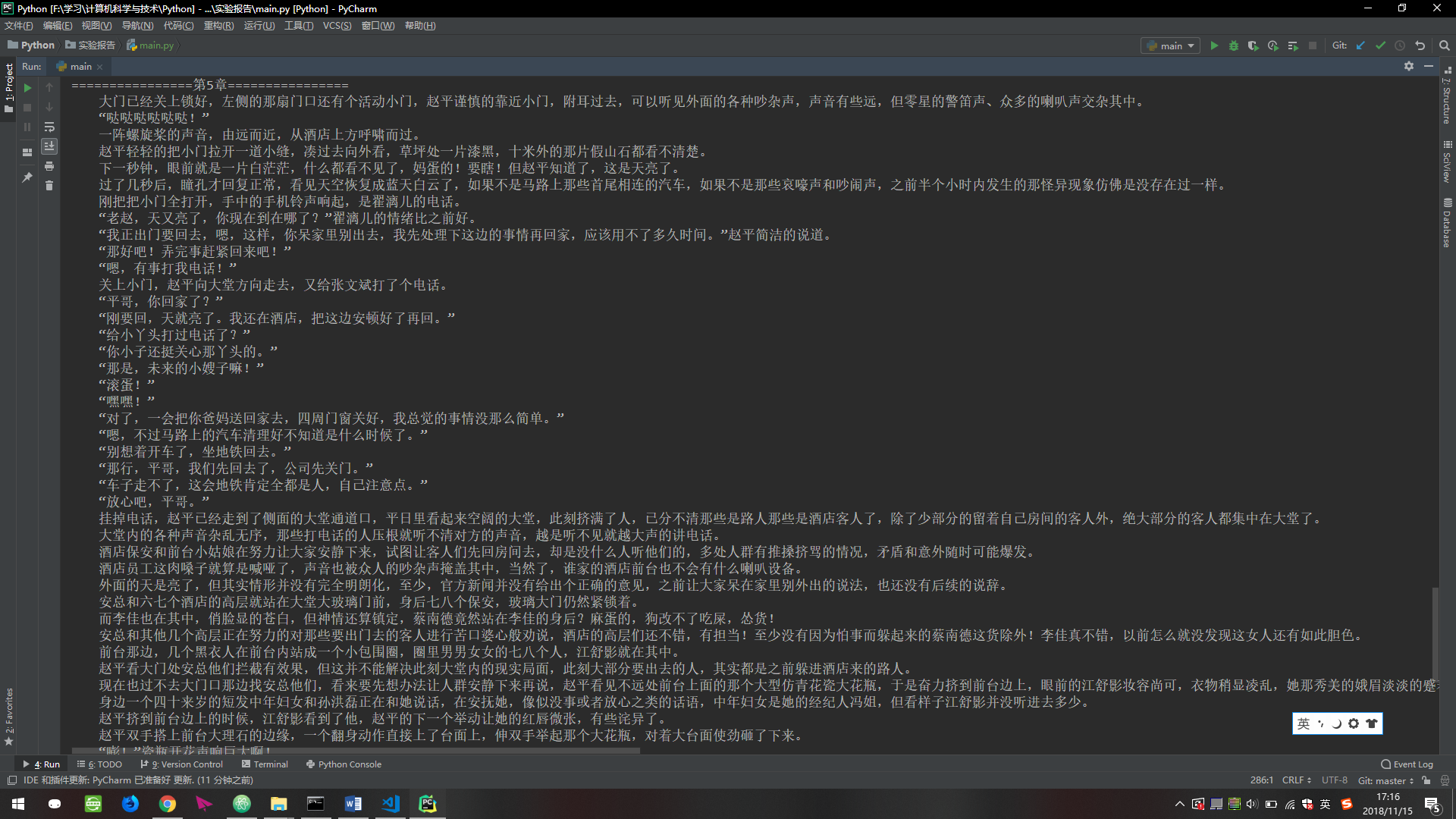
**运行截屏**











**五、实训总结**

通过实训，掌握了urllib中request模块的使用和bs4库中beautifulsoup的使用，了解了网络爬虫的基本运行模式。代码中仍有部分不足，如果网页有较大变动，则不能很好处理。