爬虫的解析与存储

作业要求：

本次作业的要求如下：

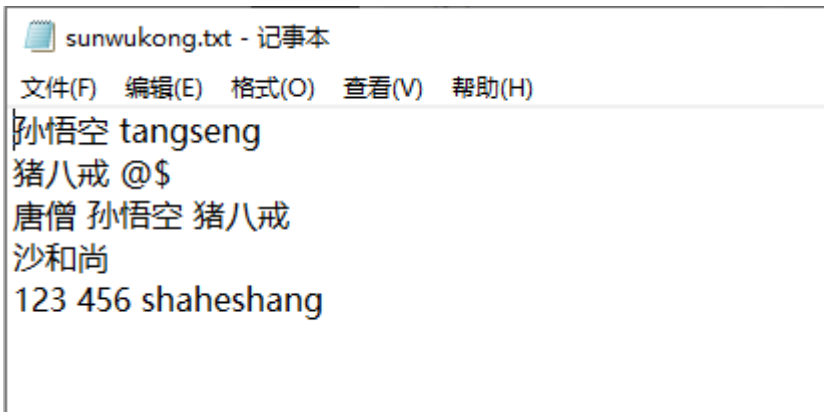
1. 爬取一万个网页（以豆瓣电影排行榜为例，一页 25 个电影条目视作是一个网页。网页内容自选，可以从一个或多个网站爬取。一些大网站如豆瓣有相应的反爬机制，爬取人数较多时可能会出现问题，同学们可以合理得自行选择要爬取的网站）。

   QQ 群文件已经上传了 Scrapy 爬虫教程、周末的线上教学录屏以及参考代码，供大家学习和参考。

2. 对网页内容进行解析。提取想要保留的网页信息。

3. 将所有爬取的网页内容合并成为一个大文件存储下来，文件格式为.txt。存储时要进行分词（中英文皆可，每个词之间用空格分开，可以有换行。）存储的文件要作为之后实验课的实验材料，所以请大家按要求做好分词。

   文件示例如下：



提交内容：

提交的内容包括爬虫程序代码包和作业报告，网页文件无需提交，留待实验使用。作业报告的格式参考群文件中的模板。完整的提交文件打包命名为‘学号

-姓名-作业 1.zip’ 。

提交方式：

将作业压缩包发送至课程邮箱：bigdata2021fall@163.com

提交日期：

作业一的最终截止日期定为第十二周周一（11.15）的下午 15：00，请同学们注意时间按时提交。