**实验1：采用Scrapy库的网络数据抓取实验报告**

姓名：江继军 学号：21373222 班级：213921

实验日期：2024年3月6日 实验地点：一号馆406 成绩：

1. **实验目的**
2. 基于Python语言和和Scrapy库，编写一个通用爬虫软件，可以通过输入网址（URL），自动抓取百度百科某一网页/多网页的某一主题内容（新闻，特色百科，知识专题），也可在新闻网抓取，写入一个纯文本（TXT）文件和JSON文件。
3. 通过此实验，初步熟悉网页信息采集和常见文件信息采集的Python编程。

**二、实验内容**

1. 环境安装

1. 安装Python编程环境，即VScode或pycharm-Professional/community, 以及Python 3.8。
2. 安装Scrapy库以及Scrapy的支持库并进行安装验证。如果爬取JavaScript动态渲染的页面，还要安装selenium或Splash（因为Scrapy不支持爬取JavaScript动态渲染页面）。

2. 通用网页爬虫程序和网页内容解析

1. 编制爬虫软件，要求能够所选定的主题内容抓取数据；
2. 对网页内容进行解析，提取内容可包含但不限于：文章标题（Title）、发表日期与时间表（Date time）、正文（Content）。提取的信息存储到可手动输入的指定文件（.TXT）中，格式清晰，存储内容易读，正文按自然段存储。

**三、实验环境**

分项列出主要实验环境（主机/操作系统、Python开发环境、Python版本、Scrapy库版本、其他库版本等）：

1. 操作系统：Windows11
2. Python开发环境：PyCharm Community Edition 2023.3.2
3. Python版本：3.9
4. Scrapy版本：Scrapy 2.6.2

**四、实验步骤**

写出生成爬虫实例、修改爬虫配置文件、存入数据的步骤。（采用1.2.3.4的序号方式列出详细步骤；写明所修改的scrapy项目中的文件和代码，所编写的代码；贴出效果示意图，例如所爬取的数据、文件采集器程序、采集的源文件、采集的中间和目标文件）。

1. 用Scrapy生成项目和爬虫实例（写明编写的代码或使用的命令，贴出带注释的实例文件）。

在终端中生成所需要的项目和爬虫实例，命令为：

scrapy startproject buaa\_cst

scrapy genspider buaa cst.buaa.edu.cn

运行结果如下图所示：

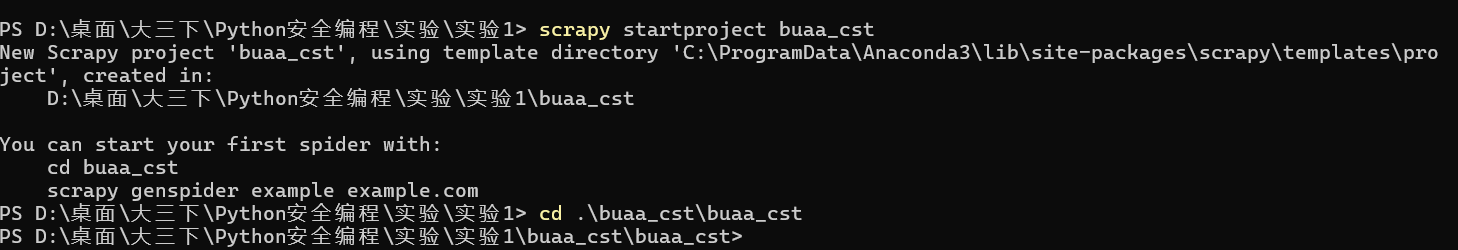


图 1用Scrapy生成项目

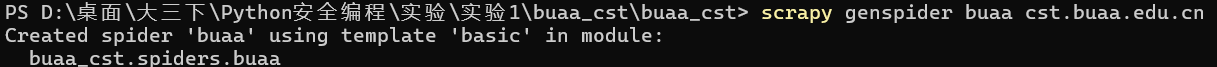


图 2 生成爬虫实例

1. 爬取网站（写明所修改的文件、代码，所添加的代码，给出爬取成功的截图）
2. 修改setting文件

首先在[北京航空航天大学网络空间安全学院 (buaa.edu.cn)](https://cst.buaa.edu.cn/)网站中调用开发者工具，查看并复制User-Agent。

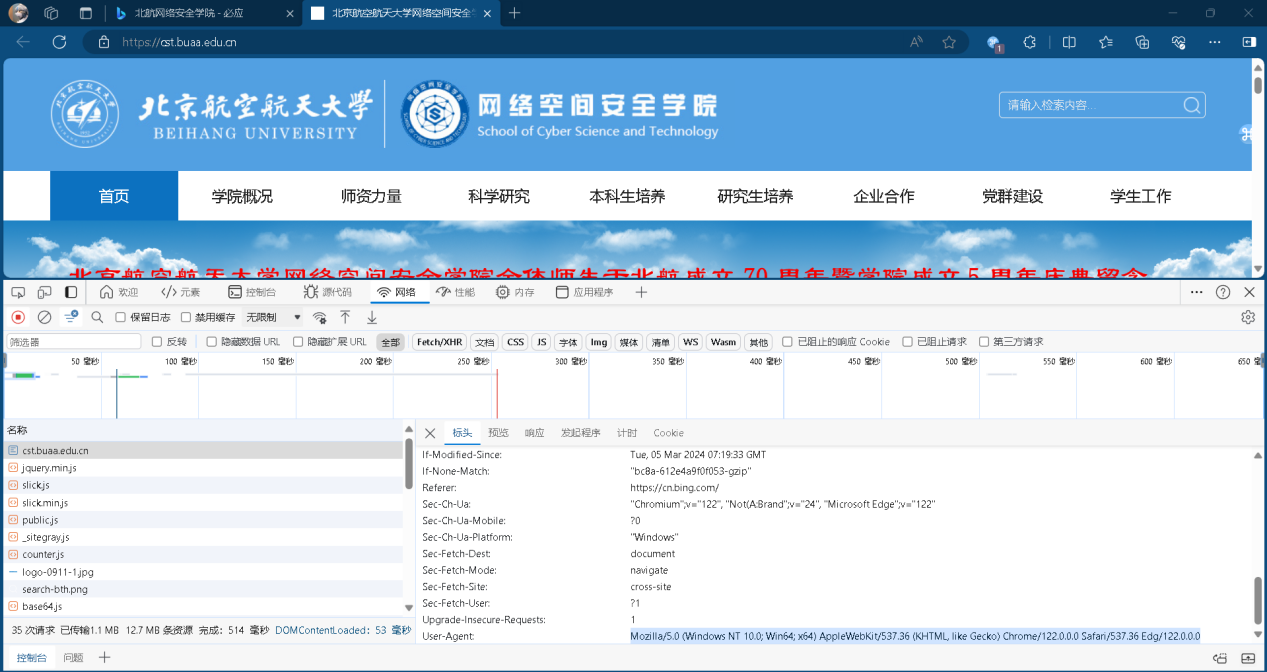


图 3 北航网络空间安全学院的User-Agent

然后，修改setting文件中的USER\_AGENT，并将ROBOTSTXT\_OBEY更改为False。



图 4 setting文件修改内容

1. 修改item文件

导入Scrapy框架，定义BuaaCstItem类用于存储结果。



图 5 item文件修改内容

1. 修改buaa文件

定义BuaaSpider类，在网页源代码中查看HTML代码，用于获取XPath，完善parse函数。



图 6 buaa文件修改内容

1. 创建run文件

创建run文件用于运行爬虫程序，将爬取的结果存在buaa.json文件。

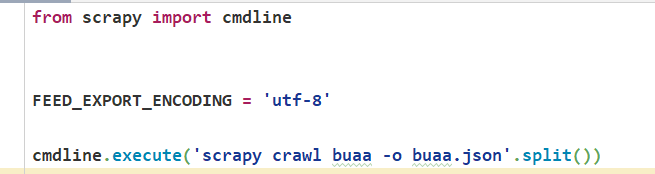


图 7 run文件内容

爬取的结果为：



图 8 爬取的新闻内容

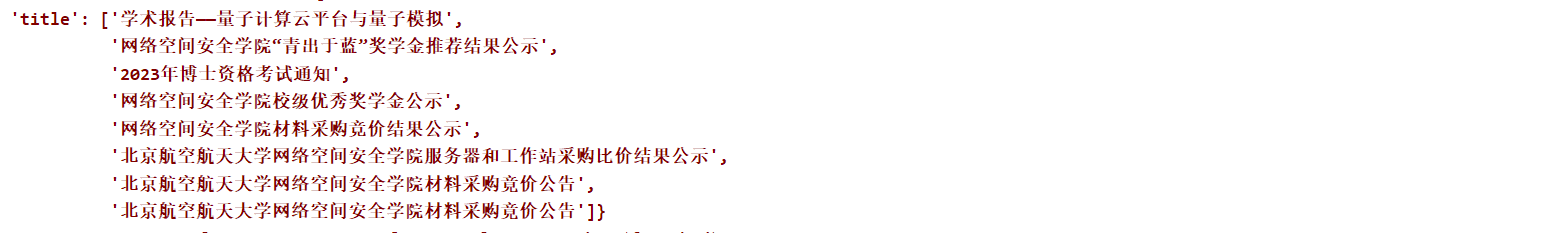


图 9 爬取的新闻标题

网站新闻内容为：



图 10 网站新闻内容

1. 内容的提取和存储（所编写的代码；贴出效果示意图，例如所爬取的数据，存入的文件）
2. 修改piplines文件，将爬取的内容转换为txt文件。

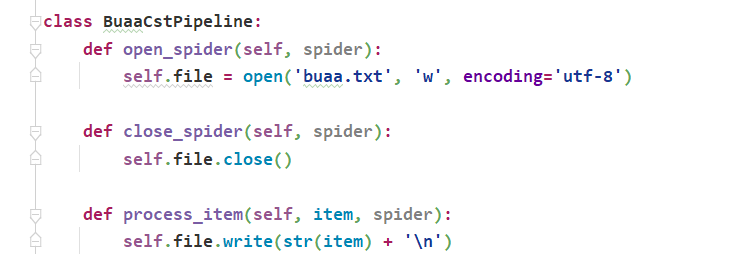


图 11 piplines文件修改内容

1. 修改setting文件，开通数据管道

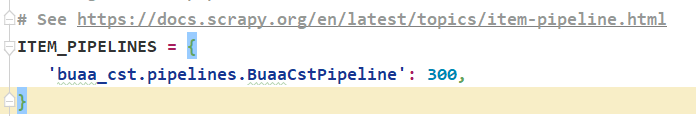


图 12 setting文件打开piplines

1. 运行run文件，得到输出。

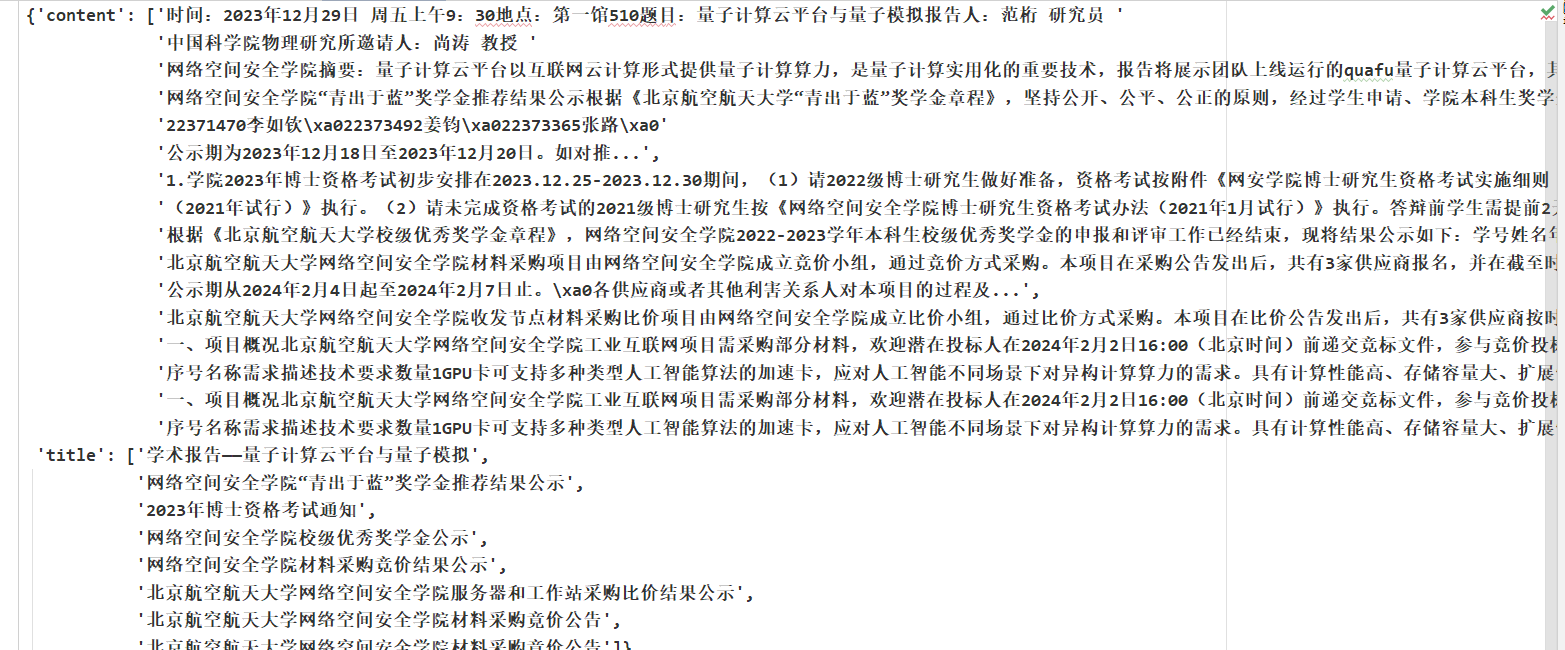


图 13 buaa.txt内容

1. 加分项（例如周期爬取+2、增量爬取+2、多页面爬取+2、多主题爬取+2、并行+3方式等，如有难度较高网站自行编写的爬虫程序+5-10，不限制项数）

周期爬取豆瓣电影Top250。

1. 用Scrapy生成新的项目和爬虫实例

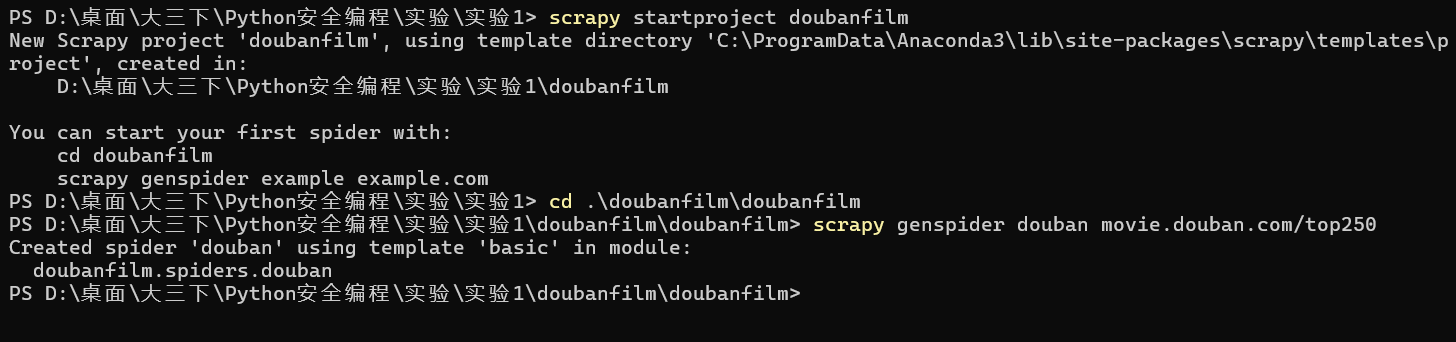


图 14 生成项目和爬虫实例

1. 修改setting文件

修改USER\_AGENT，将ROBOTSTXT\_OBEY更改为False，并将CONCURRENT\_REQUESTS设置为32，增加并发请求。

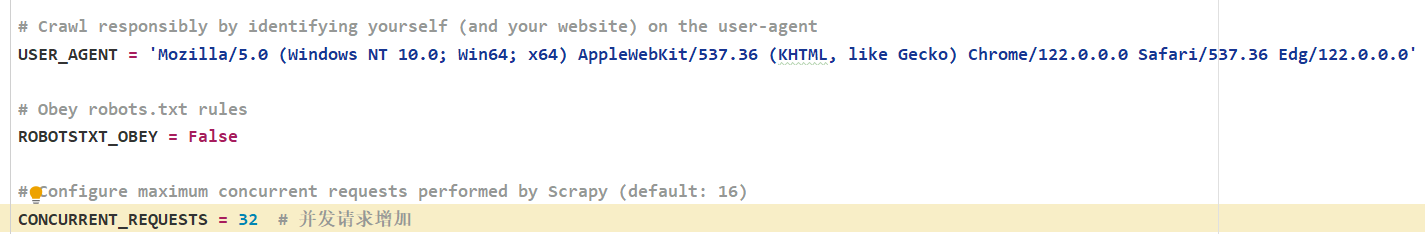


图 15 setting文件修改内容

1. 编写douban.py文件

导入CsvItemExporter以将结果存入douban\_movies.csv文件。

查找网页源代码，构造所需要的XPath获取电影的标题、导演和评分信息。

构造下一页XPath，用于跳转到下一页继续爬取，直到下一页不存在。



图 16 编写douban.py文件内容

1. 构造run.py文件

导入os库，用于删除已存在的douban\_movies.csv文件，以让周期爬取更新数据。

导入CrawlerProcess和get\_project\_settings，用于创建爬虫进程。

导入time库，用于每24小时重新爬取。



图 17 构造run.py文件内容

1. 运行run.py文件，得到结果

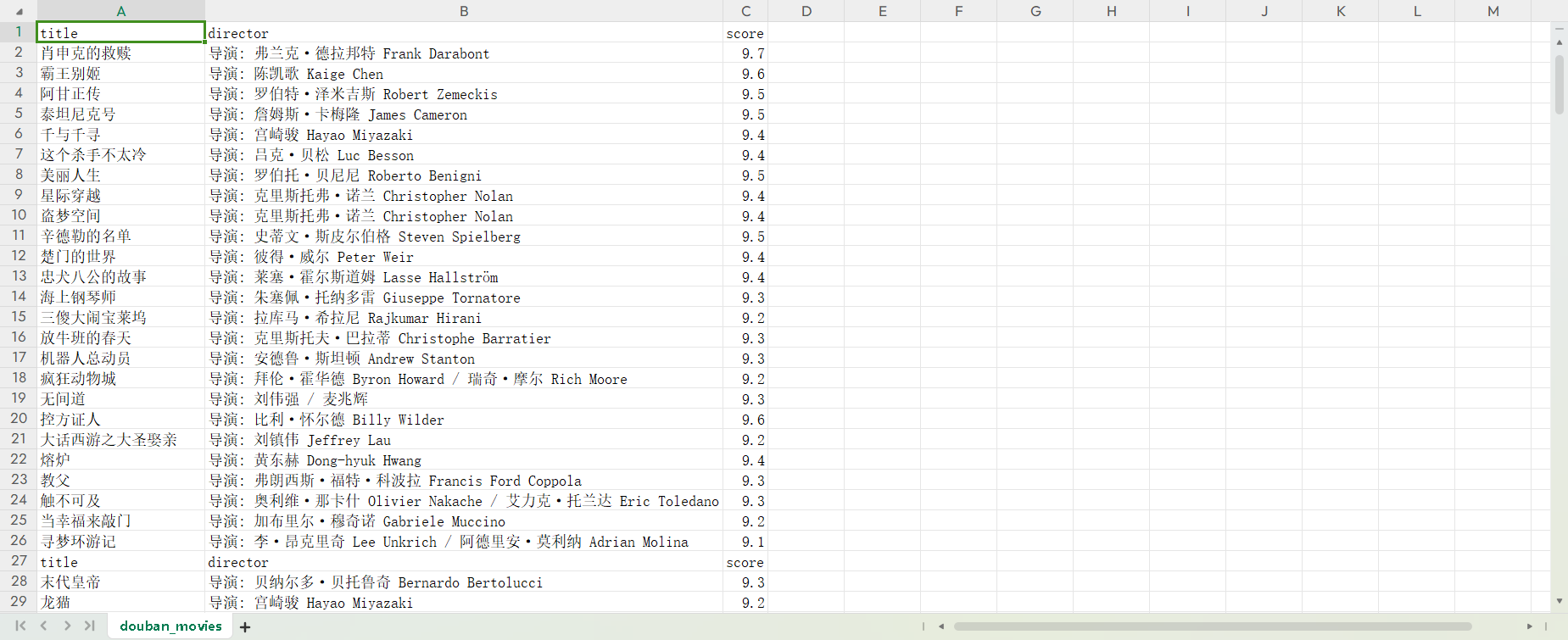


图 18 爬取结果

**五、问题和解决方法**

1. XPath使用困难，与目的有相违背。解决方法：网上学习XPath的构造使用方法，构造出自己所需要的XPath。

2. CSV文件导入异常。解决方法：采用CsvItemExporter，实现结果存入csv文件。

3. 无法周期爬取。解决方法：不采用命令行运行爬虫程序，采用CrawlerProcess创建爬虫进程，以实现周期爬取。

**六、实验总结**

结合复杂舆情数据采集，谈一下此次实验体会，并进行实验总结。

通过本次实验，我学会了如何使用Scrapy爬虫工具进行数据采集。在复杂舆情数据采集中，每个网站的格式和数据都有自己的特点，在对不同的网站数据进行爬取时，有着诸多困难，每个网站都需要按照网站的特点编写程序。在不断的编程中，我对爬虫技术也越来越熟练。同时，周期爬取、多页面爬取等要求，以及网站的反爬虫阻挡也能促进我不断学习新的知识与技术，为今后的数据采集工作打下的坚实基础。

**七、附件**

提交可执行的原程序压缩包，随电子版实验报告提交，文件名称为学号+姓名+实验一。