**本爬虫项目一共由三个py文件构成：**



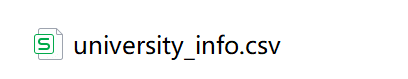
其中get\_user\_agent.py为user\_agent列表，用于导入执行文件中反反爬

test.py不用管，是草稿

1. 运行get\_university\_name

从软科排名api获取2023年度学校排名列表

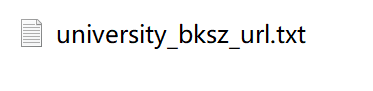
输出文件



1. 运行get\_SZul.py

将列表学校名称转换为url编码并组装成完整的百度百科网页

输出文件



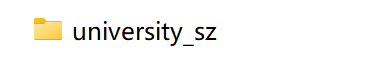
1. 运行BaiDu\_spider.py

访问university\_bksz\_url.txt中的链接

定位到“师资力量”部分



将师资力量标题下的所有信息爬取，输入到文件夹



文件夹下为每个学校设置新的文件夹放置信息



爬取的信息分为三部分：

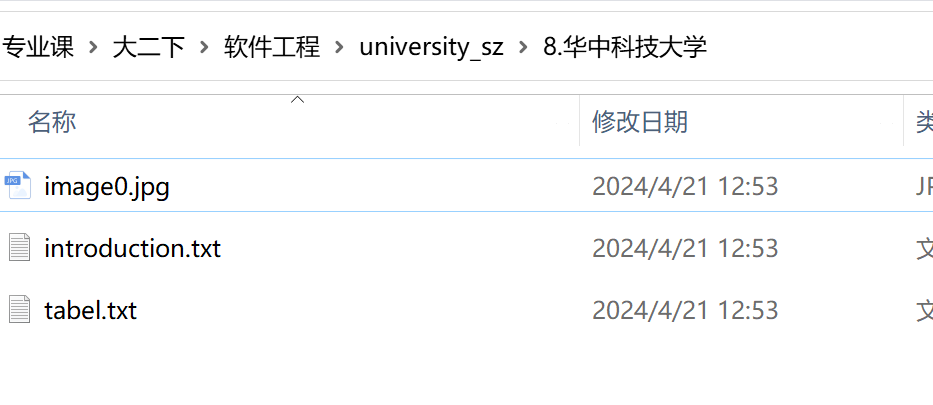
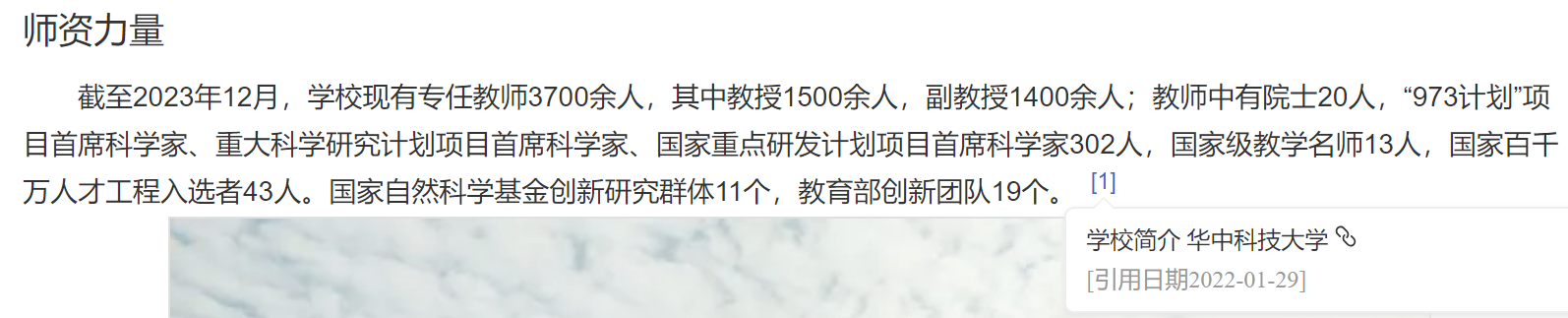


Image.jpg即为绿色框中图片

Introduction.txt为蓝色框中部分，即为对学校简介，来自于学校官网简介

Tabel.txt为紫色框中部分，即为学校各人才或教学团队的信息展示，来自于学校官网信息统计







引用来自官网，数据稍有时滞。但信息可信度高，易于批量抓取。

**爬取信息的格式未完成处理，还在考虑怎么处理，先这样吧。。。**