

西工大教务处成绩爬虫

高丽

2019.03完成

- [西工大教务处成绩爬虫](#)
 - [功能介绍](#)
 - [所需环境](#)
 - [框架思路](#)
 - [具体实现、问题与解决](#)
 - [登录](#)
 - [爬取](#)
 - [应对反爬虫](#)
 - [结果图](#)

功能介绍

1. 模拟登录教务系统
2. 按用户需求爬取相应成绩
3. 一个简单的应对反爬虫的策略

所需环境

- 需要requests、bs4、lxml、pandas、openpyxl
- 注意，需将所有库都更新到最新，否则部分函数无法使用：

```
pip list # 查看已安装的所有的依赖包
pip list --outdated -- format==columns # 像表格一样列出所有已安装的依赖包的当前版本和可升级版本
# 升级所有依赖包含如下两个命令
pip install pip-review --user # 先安装pip-review函数
pip-review --local --interactive # 成功升级所有的依赖包
```

框架思路

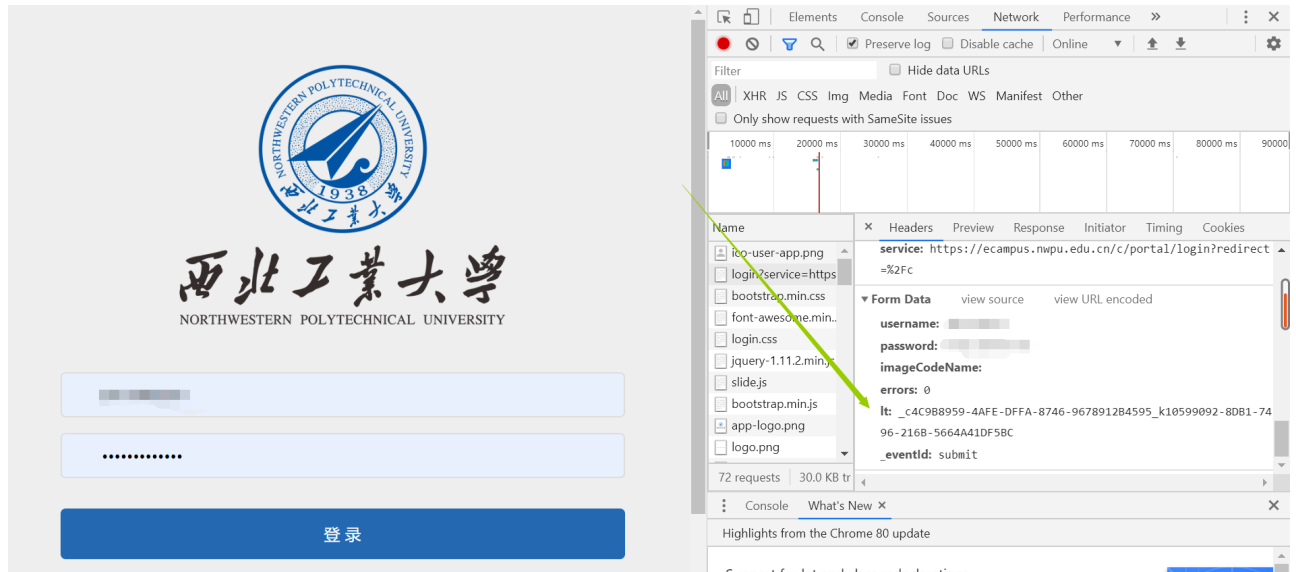
1. 首先爬取登录页面的表单中的其他信息，记录下来，输入学号、密码，提交所有表单信息，随机选择User-agent模拟浏览器登录；
2. 添加session，使得能够访问同一个网站的不同页面；
3. 代码本身通用，支持爬取所有学期，不过考虑到2017年入学，之前没有成绩，故只给出从“2017-2018秋”至今所有学期的选项，可以查看单个学期，也可以查看所有学期。输入选项，按需求爬取，保存至
`output/***.xlsx`

具体实现、问题与解决

登录

- 问题

一开始决定从翱翔门户登录再跳转至教务系统并进入成绩模块，但是发现怎么都无法成功，于是查询表单发现：



此处绿色箭头表示有一个随机数 `lt` 以及其他信息，随机数每次提交表单时会变化。其他信息是固定值。

同时，还发现学号和密码都是明文表示，也没有验证码和其他特殊信息。

所以应对策略是先请求网页将除了学号、密码以外的所有表单信息全部记录下来，接着连同学号和密码一起提交，即可登录。

成功登录后还是遇到了另外一个问题：即使利用session也无法在翱翔门户跳转到教务系统

- 解决

放弃从翱翔门户进入教务系统，直接找到教务系统的特定登录url:

`http://us.nwpu.edu.cn/eams/login.action` 然后继续观察Form Data，发现这次登录的表单信息更加简单，省略了随机数，仅有 `username`、`password`、`encodedPassword` (这里为空)、`session_locale` (设置中英文的选项，中文用“zh-CN”表示)

然后顺利登录。

爬取

成绩部分的url非常有规律，可以直接指定参数 `semesterId` 以访问不同学期的成绩单。

- 问题

成绩部分的代码虽然并没有用js或其他机制，是可以直接爬取的，但还是遇到了问题——每个学期的成绩表格格式不一样，例如有的学期有实验成绩，而有的学期没有；

并且，同一学期不同人的表格也不一样，有的人有补考成绩，有的人没有。所以不能用简单的索引来读取。

再加上，这些选项之间在属性和标签上没有任何差异，bs无法通过find函数读取特定的某一项以保存。

我的成绩 所有学期成绩

学年学期: 2019-2020学年秋学期 切换学期

学年学期	课程代码	课程序号	课程名称	课程类别	学分	平时成绩	期中成绩	实验成绩	期末成绩	总评成绩	最终	绩点
2019-2020 秋	U09M13001	U09M13001.01	嵌入式系统及应用 (双)	专业选修课程	4	100	100	91	92	94	94	3.9
2019-2020 秋	U09M11129	U09M11129.01	计算机网络	学科基础课程	3.5	97			96	96	96	4
2019-2020 秋	U09M11131	U09M11131.01	密码学	专业核心课程	3	97			90	93	93	3.9
2019-2020 秋	U31G71001T	U31G71001T.03	体育1 (瑜伽)	体育	1				89	89	89	3.8
2019-2020 秋	U30L11006	U30L11006.01	戏剧鉴赏	综合素养	2	95			96	96	96	4
2019-2020 秋	U09M11182	U09M11182.01	软件安全	专业核心课程	2	82		95		86	86	3.6
2019-2020 秋	U09M11125	U09M11125.01	计算机操作系统	学科基础课程	3.5	90			87	88	88	3.7
2019-2020 秋	U09P61004	U09P61004.01	计算机网络实训	集中实践环节	2			99		99	99	4
2019-2020 秋	U33L11016	U33L11016.01	音乐图像电影文件编辑处理	综合素养	2				98	98	98	4

• 解决

直接利用pandas的DataFrame数据结构读取会非常方便，可以实现项名和值之间的映射。

首先利用字典将“课程序号”“课程代码”“课程名称”等这些全部读取，接着逐行读取数据（每个‘tr’属性），最后转化为DataFrame格式。这样代码对于任何格式的表格都可以直接读取。

并且pandas只需调用to_excel函数即可输入到xlsx文件，非常方便。

应对反爬虫

实际上并没有发现教务系统的登录和成绩单有明显的反爬虫的机制（登录上没有随机数，复杂程度与翱翔门户比也差了很多，数据也没有被隐藏）；此外，也没有看到robots协议。

但是为了以防万一，还是写了一个非常简单的应对反爬虫的策略——python的默认User-agent会暴露自己爬虫的身份，所以这里构造UA池，每次随机选用百度、谷歌、safari、Maxthon浏览器中的一个User-agent访问，这样可以认为是一个小网络中多个用户同时访问一个页面，即不会被认为是爬虫。

```
ua_list=[#ua池
'Mozilla/5.0 (compatible; Baiduspider/2.0;
+http://www.baidu.com/search/spider.html)',
#百度
'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/50.0.2661.94 Safari/537.36',
#谷歌
'User-Agent:Mozilla/5.0 (Windows; U; Windows NT 6.1; en-us) AppleWebKit/534.50
(KHTML, like Gecko) Version/5.1 Safari/534.50',
# Safari
'Mozilla/4.0(compatible;MSIE7.0;WindowsNT5.1;Maxthon2.0)'
# (Maxthon)
]
```

结果图

1. 输入学号、密码，开始界面：

```
请输入学号/工号：
请输入密码：
模拟浏览器，随机选择User-agent: Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com/search/spider.html)
登录中.....

=====成功登录西北工业大学教务系统=====

目前可以查询的学期有——
1、2017-2018年秋季学期
2、2017-2018年春季学期
3、2018-2019年秋季学期
4、2018-2019年春季学期
5、2019-2020年秋季学期
6、查询以上所有学期
输入对应序号查询该学期的成绩：1
```

2. 选择2017-2018秋学期：

```
输入对应序号查询该学期的成绩：1

查询2017-2018年秋季学期成绩结果：





学年学期  课程代码  课程序号  课程名称  课程类别  学分  平时成绩  期中成绩  实验成绩  期末成绩  总评成绩  最终  绩点
0  2017-2018 秋  UOCL11013  逻辑学导论  人文素养类课程  1  93  88  79
1  2017-2018 秋  UOCL11056  创新创业训练  创新创业类课程  0.5
2  2017-2018 秋  U10621019  U10621019.01  C++程序设计I实验  计算机类基础课程  2
3  2017-2018 秋  U13611012  U13611012.07  思想道德修养与法律基础  思想政治理论课程  3
4  2017-2018 秋  U34611002  U34611002.14  军事理论  军事课程1  2
5  2017-2018 秋  U03111001  U03111001.01  航海概论  三航概论  0.5
6  2017-2018 秋  U31671001F  U31671001F.07  体育1（羽毛球）  体育  1
7  2017-2018 秋  U34611003  U34611003.03  大学生职业生涯规划  职业规划与发展课程  0.5
8  2017-2018 秋  U16612038  U16612038.24  大学英语（D）  大学英语公共课组  2
9  2017-2018 秋  U13611001  U13611001.02  中国近现代史纲要  思想政治理论课程  2
10 2017-2018 秋  U34611001  U34611001.08  大学生心理健康教育  心理成长与个人发展课程  0.5
11 2017-2018 秋  U11611022  U11611022.02  高等数学（上）  非专业数学类课程  5.5
12 2017-2018 秋  U34641001  军事技能训练  军事课程1  1
13 2017-2018 秋  U01111001  U01111001.06  航空概论  三航概论  0.5
14 2017-2018 秋  U10611018  U10611018.01  C++程序设计I  计算机类基础课程  3
15 2017-2018 秋  UOCL11018  艺术鉴赏  艺术素养类课程  1

=====成绩单以excel文件格式保存至output文件夹中=====

是否继续？(y or n) [ ]
```

并选择多个选项查看；

3. 显示文件已经输出：

	2017-2018年春季学期成绩单.xlsx	2020-03-14 17:41	Microsoft Excel 工...	7 KB
	2017-2018年秋季学期成绩单.xlsx	2020-03-15 1:06	Microsoft Excel 工...	7 KB
	2018-2019年秋季学期成绩单.xlsx	2020-03-14 16:48	Microsoft Excel 工...	6 KB
	所有学期成绩单.xlsx	2020-03-14 17:40	Microsoft Excel 工...	14 KB

4. 在output文件夹中查看其中一个excel文件：

	学年学期	课程代码	课程序号	课程名称	课程类别	学分	平时成绩	期中成绩	期末成绩	总评成绩	最终	绩点
0	2017-2018	U08M1105	U08M1105	电路基础	学科基础	3.5	100		85	88	88	3.7
1	2017-2018	U13L1100	U13L1100	中外文学	综合素养	2			87	87	87	3.7
2	2017-2018	U11G2305	U11G2305	大学物理	自然科学	1.5	87		87	87	87	3.7
3	2017-2018	U16G1203	U16G1203	大学英语	大学英语	2	85	88	77	80	80	3.3
4	2017-2018	U11G1102	U11G1102	高等数学	非专业数	6			90	90	90	3.8
5	2017-2018	UOCL11030		从爱因斯坦	综合素养	1				P	P	0
6	2017-2018	U09G6100	U09G6100	新生研讨	职业规划	1	P			P	P	0
7	2017-2018	UOCL11020		人人爱设	艺术素养	0.5				P	P	0
8	2017-2018	U11G2304	U11G2304	大学物理	自然科学	3.5	90		72	76	76	2.9
9	2017-2018	U13G1101	U13G1101	形势与政	思想政治	2	75		88	84	84	3.5
10	2017-2018	U11G1102	U11G1102	线性代数	非专业数	2.5			87	87	87	3.7
11	2017-2018	U08M2106	U08M2106	电路基础	学科基础	1			89	89	89	3.8
12	2017-2018	U31G7100	U31G7100	体育1（	体育	1			84	84	84	3.5