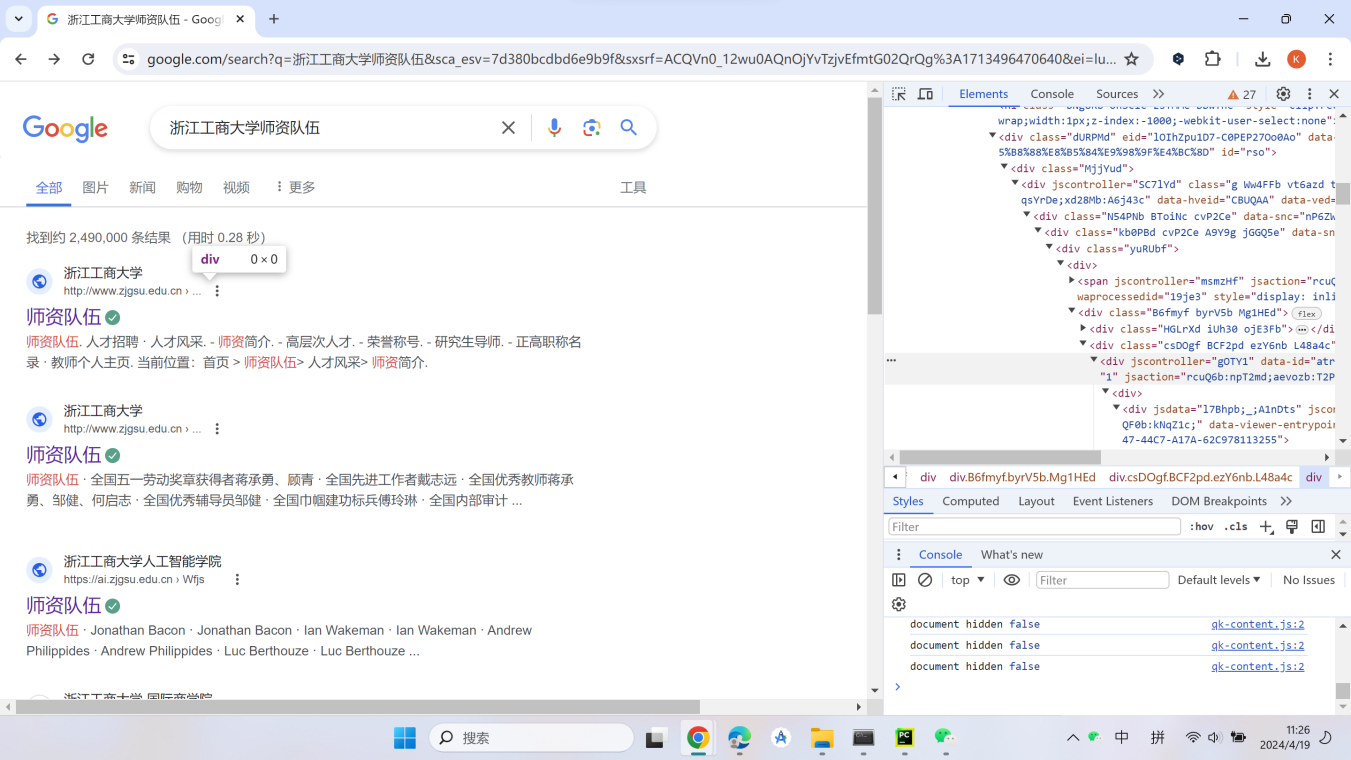
0419

采用webdriver来驱动浏览器模仿用户行为

计划爬取跳转链接到师资队伍界面



三个jscontroller对应的三个控件，定位到跳转链接



配置Xpath

获得路径//h3/a//@href

<https://baike.baidu.com/item/%E6%B5%99%E6%B1%9F%E5%B7%A5%E5%95%86%E5%A4%A7%E5%AD%A6>

<div class="paraTitle\_IjGV2 level-2\_PLGkw MARK\_MODULE" data-index="2-2" data-tag="header" data-uuid="go7upo7kw3" data-level="2"><div class="anchorList\_Z6LJi"><a name="师资力量"></a><a name="2-2"></a></div><h3 class="xh-highlight">师资力量</h3></div>

<div class="para\_TBR9F content\_BGWoU MARK\_MODULE" data-tag="paragraph" data-uuid="go7uxyxhsw" data-idx="1-8"><span class="text\_fofZk" data-text="true">截至2023年10月，学校有教职工总计4794人，其中专任教师2197人。专任教师中，具有正高级职称的1219人，具有副高级职称的840人，具有中级职称的126人，其中包括中国科学院院士30人，中国工程院院士4人，国家级领军人才近400人次，国家级青年人才400余人次，国家科技重大专项、“973计划”、“863计划”、科技创新2030—重大项目、国家重点研发计划等重大项目首席科学家135人次，国务院学位委员会学科评议组成员22人，国家级教学名师11人，“百千万人才工程”国家级人选38人。</span>

**//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::div[1]/span**

百度百科中定位到师资介绍，下一步定位表格

更新的getSZ\_url中爬取百度百科的网页链接

**第一版可行xpath**

**//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::div[1]/following-sibling::div[1]//tr/td**

**第一版可行xpath**

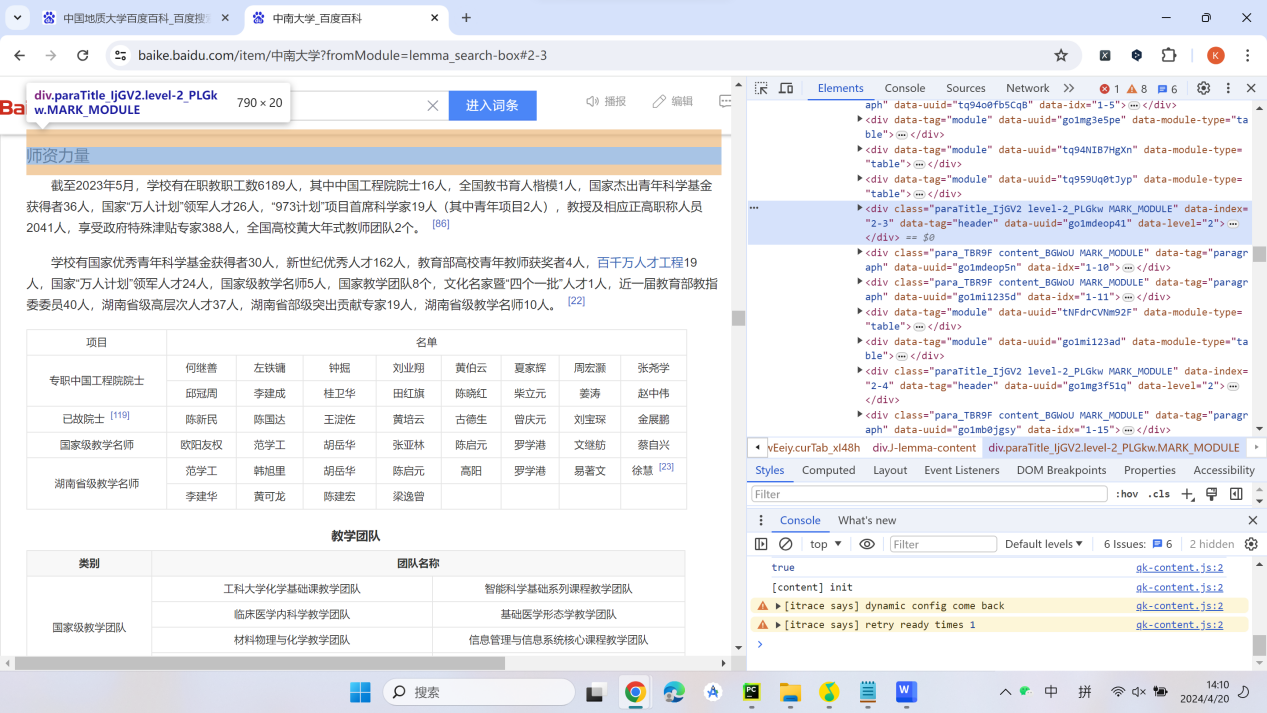
目前有用的三个文件：

Get\_universit\_name:从软科2023api爬取学校名录

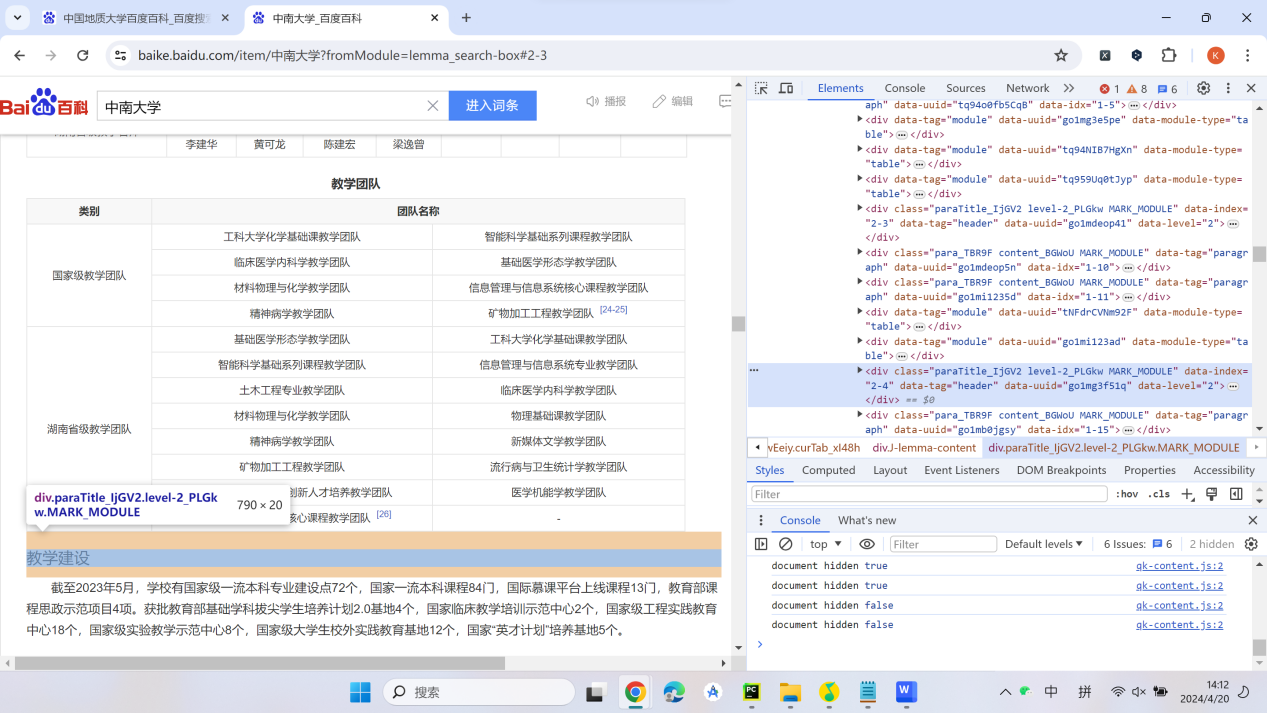
Get\_SZurl:获取学校百科的link

Baidu\_spider:从百度百科抓取师资数据

出现目录不一而导致的爬取数据不完整的情况，下一步预计做出范式，即爬取整个师资队伍下的内容：



定位相邻class中间的所有文本



<div class="paraTitle\_IjGV2 level-2\_PLGkw MARK\_MODULE" data-index="2-4" data-tag="header" data-uuid="tKK4d9wvfpDl" data-level="2"><div class="anchorList\_Z6LJi"><a name="师资力量"></a><a name="2-4"></a></div><h3>师资力量</h3></div>

<div class="para\_TBR9F content\_BGWoU MARK\_MODULE" data-tag="paragraph" data-uuid="gntadnv3wx" data-idx="1-28"><span class="text\_fofZk" data-text="true">截至2022年12月，学校现有教职工1723人，</span><span class="text\_fofZk" data-text="true"><a class="innerLink\_QXnyA" href="/item/专任教师/4799208?fromModule=lemma\_inlink" target="\_blank" data-from-module="">专任教师</a></span><span class="text\_fofZk" data-text="true">1007人，其中教授303人，副教授389人，博士研究生指导教师385人。学校有</span><span class="text\_fofZk" data-text="true"><a class="innerLink\_QXnyA" href="/item/中国科学院院士/327194?fromModule=lemma\_inlink" target="\_blank" data-from-module="">中国科学院院士</a></span><span class="text\_fofZk" data-text="true">11人，</span><span class="text\_fofZk" data-text="true"><a class="innerLink\_QXnyA" href="/item/中国工程院院士/5925613?fromModule=lemma\_inlink" target="\_blank" data-from-module="">中国工程院院士</a></span><span class="text\_fofZk" data-text="true">1人，国家杰出青年科学基金获得者16人，国家级教学名师奖获得者1人，</span><span class="text\_fofZk" data-text="true"><a class="innerLink\_QXnyA" href="/item/全国优秀教师/8004091?fromModule=lemma\_inlink" target="\_blank" data-from-module="">全国优秀教师</a></span><span class="text\_fofZk" data-text="true">2人，全国高校黄大年式教学团队2个，国家级教学团队1个，国家优秀青年科学基金获得者19人，北京市教学名师28人，北京市青年教学名师5人。</span><span class="supWrap\_by0iz J-supWrap"><sup data-tag="ref"> [15]<em id="sup-15"></em></sup></span></div>

<div data-tag="module" data-uuid="gntana9ksb" data-module-type="table"><div class="moduleTable\_m98Gc"><table class="tableBox\_ezPLl"><caption>国际高等学校科学院院士</caption><tbody class="tableBody\_YBxHG"><tr><td width="749" align="center" rowspan="1" colspan="1" class="tableTd\_rr4LV" data-uuid="gntana9kvw"><div class="para\_TBR9F table\_y9EPs MARK\_MODULE" data-tag="paragraph" data-uuid="gntana9kyn" data-idx=""><span class="text\_fofZk" data-text="true"><a class="innerLink\_QXnyA" href="/item/赵鹏大/4136828?fromModule=lemma\_inlink" target="\_blank" data-from-module="">赵鹏大</a></span></div></td></tr></tbody></table></div></div>

<div class="paraTitle\_IjGV2 level-2\_PLGkw MARK\_MODULE" data-index="2-5" data-tag="header" data-uuid="u1BCEpIvcwKH" data-level="2"><div class="anchorList\_Z6LJi"><a name="所获荣誉"></a><a name="2-5"></a></div><h3>所获荣誉</h3></div>

0421最新：

利用position定位可以选择所需定位数量：

//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::div[position()<=4]

**原型：**

//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::node()[position()<=count()-count()]

**定位到师资力量下一个同类型标题：**

//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::div[@class='paraTitle\_IjGV2 level-2\_PLGkw MARK\_MODULE'][1]

**定位到师资力量标题：**

//h3[contains(text(), '师资力量')]/ancestor::div[1]

**Count差：**

count(//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::node())-count(//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::div[@class='paraTitle\_IjGV2 level-2\_PLGkw MARK\_MODULE'][1]/following-sibling::node())

**定位到师资力量所有内容：**

//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::node()[position()<count(//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::node())-count(//h3[contains(text(), '师资力量')]/ancestor::div[1]/following-sibling::div[@class='paraTitle\_IjGV2 level-2\_PLGkw MARK\_MODULE'][1]/following-sibling::node())]

**第二版xpath范式**

**由以上公式推导爬取三部分内容：**

1. **文字介绍部分：./span/text()**
2. **表格文字部分：.//td//span/text()**
3. **图片：.//img/@src**