作业1：简单爬虫的实现

作业目标： 本次作业要求学习、理解爬虫技术并进行简单地应用。

作业任务：

（1）利用爬虫爬取学院官网老师的简历信息（http://cs.hitsz.edu.cn/szll/qzjs.htm），可以使用例如scrapy的爬虫框架；

（2）将爬取的信息进行清理（也可以在爬取的时候就定义好匹配规则）， 然后将爬取的数据存入一个合理设计表项的数据库，对应简历中的各项信息：例如姓名、任职、电话等等；

（3）作业需要提交代码、截图以及一个word文档，内容包括3部分：

1. 爬虫的基本原理以及使用的基本方法（如果使用框架，简述参数和各项module的功能）
2. 实现的简单流程：包括爬虫工具的使用、数据库设计、xpath/css路径匹配等等；
3. 简单的心得体会

（4）将以上内容打包为“张三-18011xxxx-第一次作业.zip”，以添加附件的方式于2021.5.30 20:00前提交至webxxcl@163.com邮箱，邮件主题为张三-18011xxxx-第一次作业”；

中国

王轩

教授　院长

目前就职　计算机科学与技术学院
学科　计算机科学与技术
研究方向　人工智能
主页地址　http://faculty.hitsz.edu.cn/wangxuan

个人简历

个人简介 Personal Profile

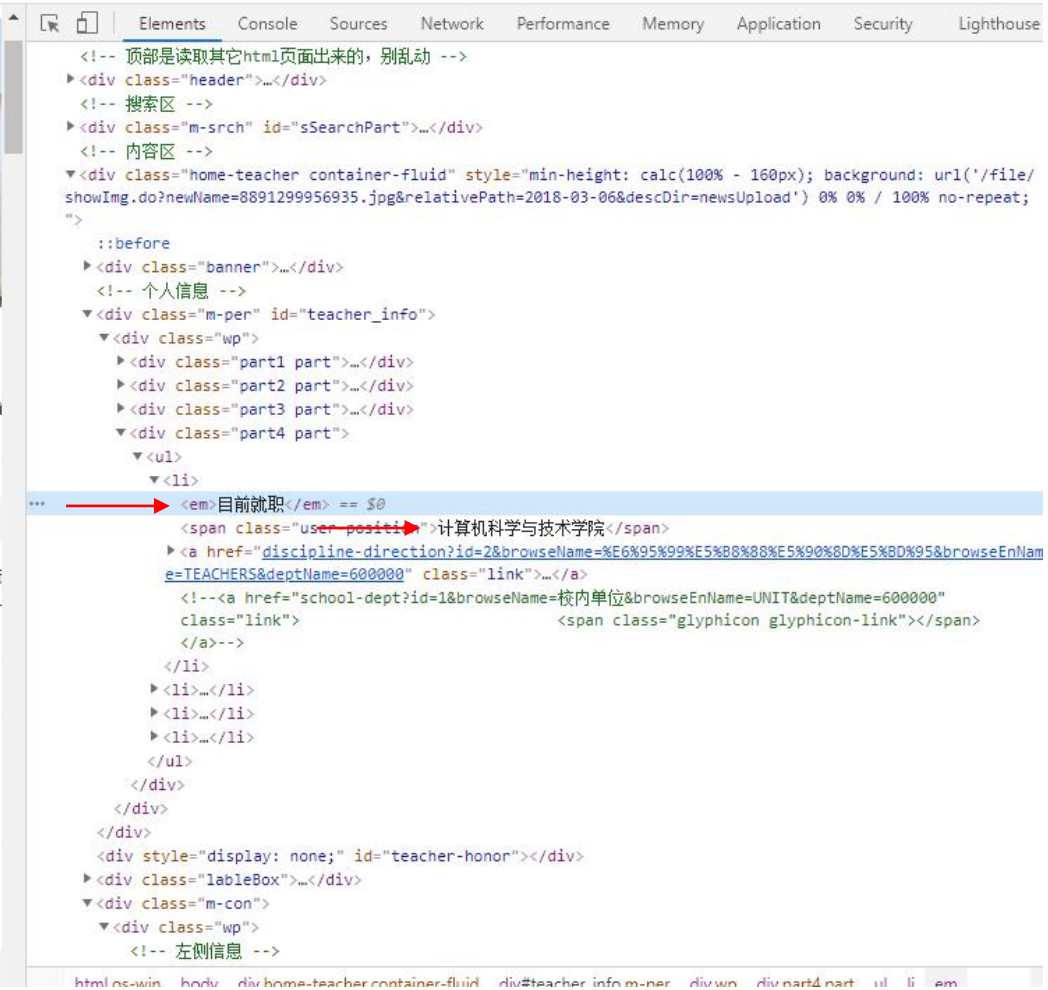
王轩教授现任哈工大计算学部副主任，哈工大（深圳）计算机科学与技术学院执行院长，是深圳市计算机学会理事长，深圳市人工智能学会副理事长，广东省人工智能与机器人学会监事长，广东省计算机学会副理事长，深圳互联网多媒体应用技术工程实验室主任，鹏城实验室AI赋能项目负责人。研究领域包括人工智能和网络空间安全，主持或参与国家科技重大专项项目、国家重点研发计划项目、国家自然科学(重点)基金项目、国家863计划项目、军委科技委项目、总装备部重点项目、广东省科技计划项目以及来自华为、中兴、微软等企业项目50余项，获得教育部一等奖、航天部二等奖、省科学技术奖项发明类一等奖、深圳市科技创新奖、深圳市科技进步奖各一项。获得国家发明专利22项，发表学术论文(SCI /EI)检索150余篇，专著3部。在人工智能的智能人机交互方向，王轩教授是微软拼音(Microsoft PY)主要发明人之一，提出的最少元素中文语句级智能输入技术是远东地区信息处理的首创性解决方案，比尔·盖茨认为有效地解决了国际上大字符集语言的计算机输入瓶颈问题，分别授权给美国微软、日本佳能等，用户数亿计，极大地推动了中文信息处理技术发展。该项成果获中国软件博览会金奖，教育部科技进步一等奖，省级科学技术奖一等奖。在人工智能博弈决策方向，王轩教授是有

联系方式

电话

邮箱

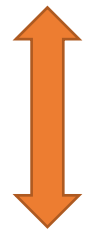
地址



Fields | Indexes | Foreign Keys | Triggers | Options | Comment | SQL Preview

Name	Type	Length	Decimals	Not Null	Virtual	Key	Comment
name	varchar	255	0	☑	☐	🔑	
title	varchar	255	0	☐	☐		
phone	varchar	255	0	☐	☐		
fax	varchar	255	0	☐	☐		
email	varchar	255	0	☐	☐		
field	varchar	255	0	☐	☐		

```
▼ <div class="teacher-box">
  ▼ <dl>
      <dt>任职: </dt>
      <dd>中国工程院院士</dd>
    </dl>
  ▼ <dl>
      <dt>电话: </dt>
      <dd> +86-755-</dd>
    </dl>
  ▼ <dl>
      <dt>传真: </dt>
      <dd> +86-755-</dd>
    </dl>
  ▼ <dl>
      <dt>Email: </dt>
    ▼ <dd>
        <a href="mailto:
        fangbx@cae.cn">fangbx@cae.cn</a>
      </dd>
    </dl>
  ▼ <dl>
      <dt>研究方向: </dt>
      <dd>信息网络与信息安全</dd>
    </dl>
```



name	title	phone	fax	email	field
WILLIAM A.GRUVER	教授	604-291-4339		gruver@cs.sfu.ca	自动控制、机器人技术。
丁宇新	副教授	86-755-26032193	86-755-26033608	yxding@hit.edu.cn	网络安全,人工智能,编译。
何震宇	教授，博导	+86-755-26033060	+86-755-26033060	zhenyuhe@hit.edu.cn	人工智能，机器学习，计算机视
刘川意	副教授，硕士生导师	+86-755-	+86-755-	liuchuanyi@hits.edu.cn	云计算与云安全，大规模存储系
刘洋	助理教授	+86-755-	+86-755-	liu.yang@hit.edu.cn	数据安全与隐私保护
卢光明	副院长、教授、博导	+86-755-26032458	+86-755-26032461	luguangm@hit.edu.cn	图像处理、模式识别、生物特征
唐琳琳	高级讲师，硕士生导师，博士	0755-86540396	+86-755-	lltang@hit.edu.cn	图像识别、信息安全、云计算负
堵宏伟	副教授 博导	+86-755-26619401	+86-755-26619401	hwdu@hit.edu.cn	无线多跳网络（无线 ad-hoc, 传
夏文	副教授 硕士生导师	+86-755-	+86-755-	xiawen@hit.edu.cn OR wx.hust@(	云存储、云计算、去重压缩
廖清	副教授	0755-86134382		liaoqing@hit.edu.cn	数据挖掘、人工智能、信息安全
张加佳	副研究员	15989359103		zhangjiajia@hit.edu.cn	机器博弈决策、金融博弈决策、
张春慨	副教授	+86-755-26032545	+86-755-26032461	ckzhang@hit.edu.cn	研究方向为流数据挖掘、网络信
张晓峰	博士、副教授、硕士生导师	+86-755-	+86-755-	zhangxiaofeng@hit.edu.cn	数据挖掘、机器学习、人工智能
张正	助理教授			zhengzhang@hit.edu.cn	Machine Learning, Computer
张海军	教授，博导	+86-755-26033086	+86-755-26033008	hjzhang@hit.edu.cn; aarhzhang@	多媒体数据挖掘及大数据分析应
徐增林	国家青年特聘专家、计算机学院			xuzenglin@hit.edu.cn	致力于解决涉及现代大数据分析
徐晓飞	教授	+86-451-8641 8566	+86-451-8641 8566	xiaofei@hit.edu.cn	计算机集成制造CIMS，数据库，