数据分析及实践实验二实验报告

PB18061443 江昊霖

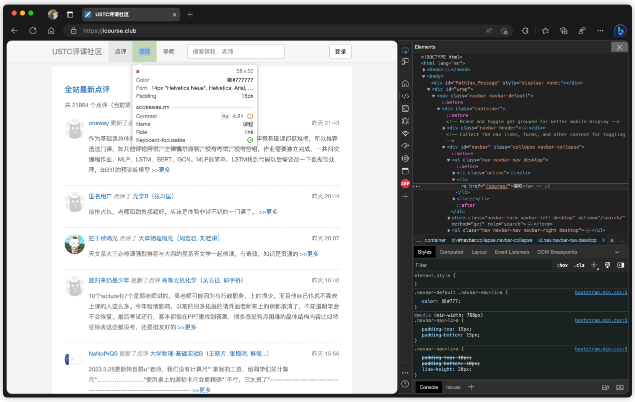
2023 年 3 月 29 日

1 实验目的

给定评课社区网站，需要设计一个网站遍历策略，爬取至少 200 个课程的详细信息，记录于 json 格式的文件中.

2 实验步骤

由给定的 url 进入课程列表网站爬取课程信息。



a 58 × 50
Color #777777
Font 14px "Helvetica Neue", Helvetica, Arial, ...
Padding 15px
ACCESSIBILITY
Contrast Aa 4.21
Name 课程
Role link
Keyboard-focusable

全站最新点评

共 21864 个点评（当前第

oneway 更新了点 昨天 21:43

作为基础课总体来 毕竟基础课都挺难搞，所以推荐选这门课，如其他评论所说，上课偶尔点名，没有考试，没有分组，作业需要独立完成，一共四次编程作业，MLP、LSTM、BERT、GCN。MLP很简单。LSTM找到代码以后需要改一下数据预处理，BERT的预训练模型 >>更多

匿名用户 点评了 光学B（张斗国） 昨天 20:44

前排占坑，老师和助教都超好，应该是体验非常不错的一门课了。 >>更多

把千秋喝光 点评了 天体物理概论（周宏岩, 刘桂琳） 昨天 20:07

天文系大三必修课强烈推荐与大四的星系天文学一起修读，有奇效，知识是贯通的 >>更多

愿归来仍是少年 更新了点评 高等无机化学（吴长征, 郭宇桥） 昨天 16:40

10个lecture有7个是郭老师讲的，吴老师可能因为有行政职务，上的很少，而且他自己也说不喜欢上课的人这么多。今年疫情影响，以前的很多拓展的请外面老师来上的课都取消了，不知道明年会不会恢复。最后考试还行，基本都能在PPT里找到答案，很多感觉有点困难的晶体结构内容比如特征标表这些都没考，还是挺友好的 >>更多

NaNofNQS 更新了点评 大学物理-基础实验B（王晓方, 张增明, 蔡俊...） 昨天 15:56

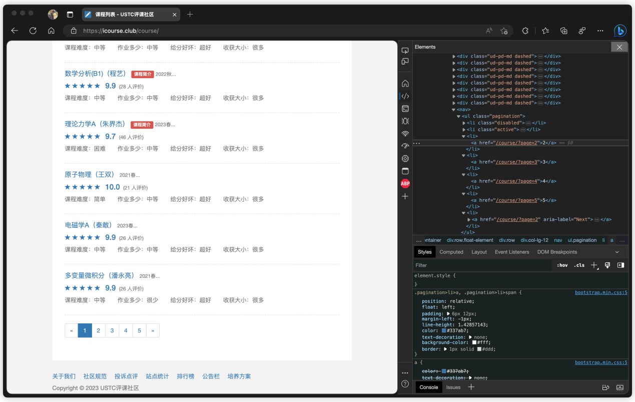
2023.3.28更新转自群u"老师，我们没有计算尺""拿我的工资，给同学们买计算尺"......................."使用桌上的游标卡尺会更精确""不行，它太贵了"-- >>更多

```
<!DOCTYPE html>
<html lang="en">
 ▶<head>⋯</head>
 ▼<body>
    <div id="MathJax_Message" style="display: none;"></div>
  ▼<div id="wrap">
    ▼<nav class="navbar navbar-default">
       ::before
     ▼<div class="container">
         ::before
         <!-- Brand and toggle get grouped for better mobile display -->
       ▶<div class="navbar-header">⋯</div>
         <!-- Collect the nav links, forms, and other content for toggling
         -->
       ▼<div id="navbar" class="collapse navbar-collapse">
           ::before
         ▼<ul class="nav navbar-nav desktop">
             ::before
           ▶<li class="active">⋯</li>
           ▼<li>
...            <a href="/course/">课程</a> == $0
             </li>
           ▶<li>⋯</li>
             ::after
           </ul>
         ▶<form class="navbar-form navbar-left desktop" action="/search/"
          method="get" role="search">⋯</form>
         ▶<ul class="nav navbar-nav navbar-right desktop">⋯</ul>
```

... container div#navbar.collapse.navbar-collapse ul.nav.navbar-nav.desktop li a ...
Styles Computed Layout Event Listeners DOM Breakpoints
Filter :hov .cls

```
element.style {
}
.navbar-default .navbar-nav>li>a {                    bootstrap.min.css:5
   color: #777;
}
@media (min-width: 768px)
.navbar-nav>li>a {                                    bootstrap.min.css:5
   padding-top: 15px;
   padding-bottom: 15px;
}
.navbar-nav>li>a {                                    bootstrap.min.css:5
   ~~padding-top: 10px;~~
   ~~padding-bottom: 10px;~~
   line-height: 20px;
}
```

通过观察，课程列表的页面规律为：`https://icourse.club/course/?page={pagenumber}`，每页为 10 个课程。



课程列表 - USTC评课社区
https://icourse.club/course/

课程难度：中等　作业多少：中等　给分好坏：超好　收获大小：很多

数学分析(B1)（程艺） 课程简介 2022秋...
★★★★★ 9.9 (28 人评价)
课程难度：中等　作业多少：中等　给分好坏：超好　收获大小：很多

理论力学A（朱界杰） 课程简介 2023春...
★★★★★ 9.7 (46 人评价)
课程难度：困难　作业多少：中等　给分好坏：超好　收获大小：很多

原子物理（王双） 2021春...
★★★★★ 10.0 (21 人评价)
课程难度：简单　作业多少：中等　给分好坏：超好　收获大小：很多

电磁学A（秦敢） 2023春...
★★★★★ 9.9 (26 人评价)
课程难度：中等　作业多少：中等　给分好坏：超好　收获大小：很多

多变量微积分（潘永亮） 2021春...
★★★★★ 9.9 (26 人评价)
课程难度：中等　作业多少：很少　给分好坏：超好　收获大小：很多

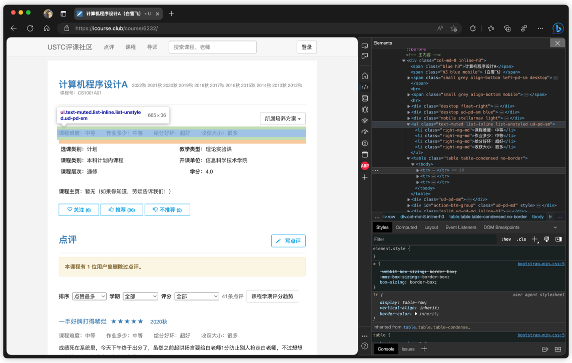
« 1 2 3 4 5 »
关于我们 社区规范 投诉点评 站点统计 排行榜 公告栏 培养方案
Copyright © 2023 USTC评课社区
Elements

```
          ▶<div class="ud-pd-md dashed">⋯</div>
          ▶<div class="ud-pd-md dashed">⋯</div>
          ▶<div class="ud-pd-md dashed">⋯</div>
          ▶<div class="ud-pd-md dashed">⋯</div>
          ▶<div class="ud-pd-md dashed">⋯</div>
          ▶<div class="ud-pd-md dashed">⋯</div>
          ▶<div class="ud-pd-md dashed">⋯</div>
          ▶<div class="ud-pd-md dashed">⋯</div>
          ▼<nav>
            ▼<ul class="pagination">
              ▶<li class="disabled">⋯</li>
              ▶<li class="active">⋯</li>
              ▼<li>
...              <a href="/course/?page=2">2</a> == $0
               </li>
              ▼<li>
                 <a href="/course/?page=3">3</a>
               </li>
              ▼<li>
                 <a href="/course/?page=4">4</a>
               </li>
              ▼<li>
                 <a href="/course/?page=5">5</a>
               </li>
              ▼<li>
                ▶<a href="/course/?page=2" aria-label="Next">⋯</a>
               </li>
             </ul>
```

... ntainer div.row.float-element div.row div.col-lg-12 nav ul.pagination li a ...
Styles Computed Layout Event Listeners DOM Breakpoints
Filter :hov .cls

```
element.style {
}
.pagination>li>a, .pagination>li>span {               bootstrap.min.css:5
   position: relative;
   float: left;
   padding: ▶ 6px 12px;
   margin-left: -1px;
   line-height: 1.42857143;
   color: #337ab7;
   text-decoration: ▶ none;
   background-color: #fff;
   border: ▶ 1px solid #ddd;
}
a {                                                   bootstrap.min.css:5
   ~~color: #337ab7;~~
   ~~text-decoration: ▶ none;~~
```

递归进入每个课程页面后爬取信息.



https://icourse.club/course/6232/

USTC评课社区 点评 课程 导师 搜索课程、老师 登录

计算机程序设计A 2022秋 2021秋 2020秋 2019秋 2018秋 2017秋 2016秋 2015秋 2014秋 2013秋 2012秋
课程号：CS1001A01

ul.text-muted.list-inline.list-unstyled.ud-pd-sm 665 × 36

所属培养方案

课程难度：中等 作业多少：中等 给分好坏：超好 收获大小：很多

选课类别：计划 教学类型：理论实验课
课程类别：本科计划内课程 开课单位：信息科学技术学院
课程层次：通修 学分：4.0

课程主页：暂无（如果你知道，劳烦告诉我们！）

关注 (6) 推荐 (36) 不推荐 (2)

点评 写点评

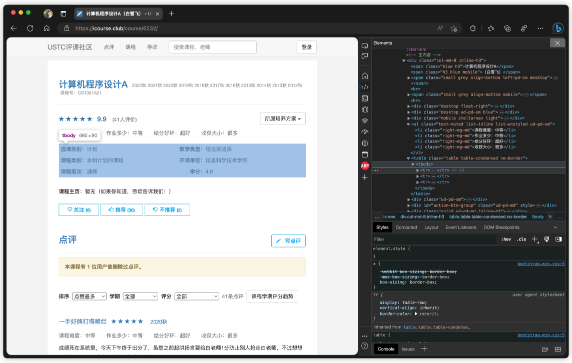
本课程有 1 位用户曾删除过点评。

排序 点赞最多 学期 全部 评分 全部 41条点评 课程学期评分趋势

一手好牌打得稀烂 ★★★★★ 2020秋
课程难度：中等 作业多少：中等 给分好坏：超好 收获大小：很多
成绩死在系统里，今天下午终于出分了，虽然之前起哄扬言要给白老师1分防止别人抢走白老师，不过想想

```
<!-- 主内容 -->
▼<div class="col-md-8 inline-h3">
   <span class="blue h3">计算机程序设计A</span>
   <span class="h3 blue mobile"> (白雪飞) </span>
  ▶<span class="small grey align-bottom left-pd-sm desktop">…
   </span>
   <br>
  ▶<span class="small grey align-bottom mobile">…</span>
   <hr>
  ▶<div class="desktop float-right">…</div>
  ▶<div class="desktop ud-pd-sm blue">…</div>
  ▶<div class="mobile stellarnav light">…</div>
  ▼<ul class="text-muted list-inline list-unstyled ud-pd-sm">
     <li class="right-mg-md">课程难度: 中等</li>
     <li class="right-mg-md">作业多少: 中等</li>
     <li class="right-mg-md">给分好坏: 超好</li>
     <li class="right-mg-md">收获大小: 很多</li>
   </ul>
  ▼<table class="table table-condensed no-border">
    ▼<tbody>
      ▶<tr>…</tr> == $0
      ▶<tr>…</tr>
      ▶<tr>…</tr>
     </tbody>
   </table>
  ▶<div class="ud-pd-sm">…</div>
  ▶<div id="action-btn-group" class="ud-pd-md" style>…</div>
```

```
element.style {
}
* {                                   bootstrap.min.css:5
  -webkit-box-sizing: border-box;
  -moz-box-sizing: border-box;
  box-sizing: border-box;
}
tr {                             user agent stylesheet
  display: table-row;
  vertical-align: inherit;
  border-color: ▶ inherit;
}
Inherited from table.table.table-condense…
table {                               bootstrap.min.css:5
```



https://icourse.club/course/6232/

USTC评课社区 点评 课程 导师 搜索课程、老师 登录

计算机程序设计A 2022秋 2021秋 2020秋 2019秋 2018秋 2017秋 2016秋 2015秋 2014秋 2013秋 2012秋
课程号：CS1001A01

★★★★★ 9.9 (41人评价) 所属培养方案

tbody 660 × 90

作业多少：中等 给分好坏：超好 收获大小：很多

选课类别：计划 教学类型：理论实验课
课程类别：本科计划内课程 开课单位：信息科学技术学院
课程层次：通修 学分：4.0

课程主页：暂无（如果你知道，劳烦告诉我们！）

关注 (6) 推荐 (36) 不推荐 (2)

点评 写点评

本课程有 1 位用户曾删除过点评。

排序 点赞最多 学期 全部 评分 全部 41条点评 课程学期评分趋势

一手好牌打得稀烂 ★★★★★ 2020秋
课程难度：中等 作业多少：中等 给分好坏：超好 收获大小：很多
成绩死在系统里，今天下午终于出分了，虽然之前起哄扬言要给白老师1分防止别人抢走白老师，不过想想

```
<!-- 主内容 -->
▼<div class="col-md-8 inline-h3">
   <span class="blue h3">计算机程序设计A</span>
   <span class="h3 blue mobile"> (白雪飞) </span>
  ▶<span class="small grey align-bottom left-pd-sm desktop">…
   </span>
   <br>
  ▶<span class="small grey align-bottom mobile">…</span>
   <hr>
  ▶<div class="desktop float-right">…</div>
  ▶<div class="desktop ud-pd-sm blue">…</div>
  ▶<div class="mobile stellarnav light">…</div>
  ▼<ul class="text-muted list-inline list-unstyled ud-pd-sm">
     <li class="right-mg-md">课程难度: 中等</li>
     <li class="right-mg-md">作业多少: 中等</li>
     <li class="right-mg-md">给分好坏: 超好</li>
     <li class="right-mg-md">收获大小: 很多</li>
   </ul>
  ▼<table class="table table-condensed no-border">
    ▼<tbody>
      ▶<tr>…</tr> == $0
      ▶<tr>…</tr>
      ▶<tr>…</tr>
     </tbody>
   </table>
  ▶<div class="ud-pd-sm">…</div>
  ▶<div id="action-btn-group" class="ud-pd-md" style>…</div>
```

```
element.style {
}
* {                                   bootstrap.min.css:5
  -webkit-box-sizing: border-box;
  -moz-box-sizing: border-box;
  box-sizing: border-box;
}
tr {                             user agent stylesheet
  display: table-row;
  vertical-align: inherit;
  border-color: ▶ inherit;
}
Inherited from table.table.table-condense…
table {                               bootstrap.min.css:5
```

3 代码

由于在进行请求时需要请求头，使用 fake_useragent 生成 user-agent.

3.1 获取课程列表网页链接

```python
def courseUrl(url: str) -> str:
    """get course list page suffix from given page

    Args:
        url (str): given url

    Returns:
        str: course list page
    """

    headers = {'User-Agent': UserAgent().random}
    ret = Request(url, headers=headers)
    res = urlopen(ret)
    data = res.read().decode('utf-8')
    soup = bf(data, features="lxml")

    suffix = soup.find('a', string=re.compile('.*?课程。*?'))['href']
    return url + suffix
```

3.2

```python
def getCoursesID(url: str) -> tuple[list, list]:
    """get courses' links and courses' names

    Args:
        url (str): courses list page

    Returns:
        tuple[list, list]: courses' links and names
    """

    headers = {'User-Agent': UserAgent().random}
    ret = Request(url, headers=headers)
    res = urlopen(ret)
    data = res.read().decode('utf-8')
    soup = bf(data, features="lxml")

    links = soup.find_all(class_="px16")
    courses = []

    for index, link in enumerate(links):
        courses.append(link.text)
        links[index] = link['href']

    return links, courses
```

```python
def courseInfo(courseID: str, courseName: str, serial: int) -> list:
    """get courses info

    Args:
        courseID (str): the courses id to course page
        courseName (str): course's name and teacher
        serial (int): serial of course

    Returns:
        info(list): info of the course
    """

    url = URL+courseID

    headers = {'User-Agent': UserAgent().random}
    ret = Request(url, headers=headers)
    res = urlopen(ret)
    data = res.read().decode('utf-8')
    soup = bf(data, features="lxml")

    info = []
    info.append(str(serial))
    info.append(courseName)

    block1 = soup.find_all(class_="right-mg-md", limit=5)
    del (block1[0])
    for name in block1:
        info.append(name.string.split(': ')[-1])

    links = soup.find_all('strong', limit=6)

    for link in links:
        info.append(link.nextSibling)

    return info
```