北 京 林 业 大 学

2022 学年—2023 学年第 2 学期 Python 应用 实验任务书

专业名称：计算机、大数据、物联网、信息　　实验学时：2

课程名称：Python 应用　　任课教师：王春玲

实验题目：实验 3 爬取中国工程院院士信息

实验环境：Python、PyCharm 等

实验目的：

1. 熟练使用标准库 urllib 读取网页内容。
2. 熟练使用正则表达式提取文本中感兴趣的信息。
3. 熟练使用内置函数 open()创建文本文件和二进制文件。
4. 熟悉 HTML 语法以及常见的 HTML 标签。

实验内容：

爬取中国工程院网页，把每位院士的简介保存为本地文本文件，并把每位院士的照片保存为本地图片，文本文件和图片文件都以院士的姓名为主文件名。

实验步骤如下：

（1）使用 Google Chrome 或其他浏览器打开下面的网址，然后在页面上右击，在弹出的菜单中选择“查看网页源代码”。

http://www.cae.cn/cae/html/main/col48/column_48_1.html

（2）分析网页源代码，确定每位院士的姓名和链接所在的 HTML 标签，为后面编写正则表达式做准备，如图 1 所示。

```
<li class="name_list"><a href="/cae/html/main/colys/63775817.html" target="_blank">曹喜滨</a></li>
<li class="name_list"><a href="/cae/html/main/colys/35791989.html" target="_blank">陈学东</a></li>
<li class="name_list"><a href="/cae/html/main/colys/01567139.html" target="_blank">邓宗全</a></li>
<li class="name_list"><a href="/cae/html/main/colys/25235806.html" target="_blank">丁荣军</a></li>
<li class="name_list"><a href="/cae/html/main/colys/46604755.html" target="_blank">董春鹏</a></li>
<li class="name_list"><a href="/cae/html/main/colys/42793697.html" target="_blank">樊会涛</a></li>
<li class="name_list"><a href="/cae/html/main/colys/15623030.html" target="_blank">冯培德</a></li>
<li class="name_list"><a href="/cae/html/main/colys/48072917.html" target="_blank">冯煜芳</a></li>
<li class="name_list"><a href="/cae/html/main/colys/71311121.html" target="_blank">甘晓华</a></li>
<li class="name_list"><a href="/cae/html/main/colys/15801057.html" target="_blank">高金吉</a></li>
```

图 1 每位院士的链接

(3)使用浏览器打开任意一位院士的链接，然后查看并分析网页源代码，确定简介信息和照片所在的 HTML 标签，为后面编写正则表达式做准备，如图 2 所示。

图 2 院士个人简介和照片

(4)编写代码，爬取信息并创建本地文件。

实验要求：

1. 须独立完成算法设计及代码编写，**严禁抄袭**。
2. 按照“实验报告书模板”要求撰写实验报告。

实验提交：

1. 实验完成后，将实验报告命名为：**班级_学号_姓名_实验 3.pdf**（注意文件名不要写错，特别是下划线和其中各项内容的顺序）。
2. 提交地址：“ftp://211.71.149.53/王春玲/课程作业/Python 应用”文件夹中的相应班级**实验 3** 文件夹中。
3. **上传截止时间 4 月 17 日** 00:00:00。如果实验报告在时间戳之后提交，则本次实验成绩按 60%折算。

参考书目：

[1]江红、余青松编著.Python 程序设计与算法基础教程（第 2 版）.北京：清华大学出版社，2019.

[2]Barry 等编著《Head First Python(中文版)》（第二版）.北京：中国电力出版社，2017.