1.9 实践项目—爬取学生信息

深圳信息职业技术学院
Shenzhen Institute Of Information Technology
教师：黄锐军

目 录

COMPANY

PART ONE

项目简介

项目简介

设计一个 Web服务器server.py，它读取students.txt文件中的学生数据，以表格的形式呈现在网页上，其中students.txt的格式如下：

No,Name,Gender,Age
1001,张三,男,20
1002,李四,女,19
1003,王五,男,21

第一行是学生表格的标题，有学号No、姓名Name、性别Gender、年龄Age，每个学生占一行，各个数据之间用逗号分开。

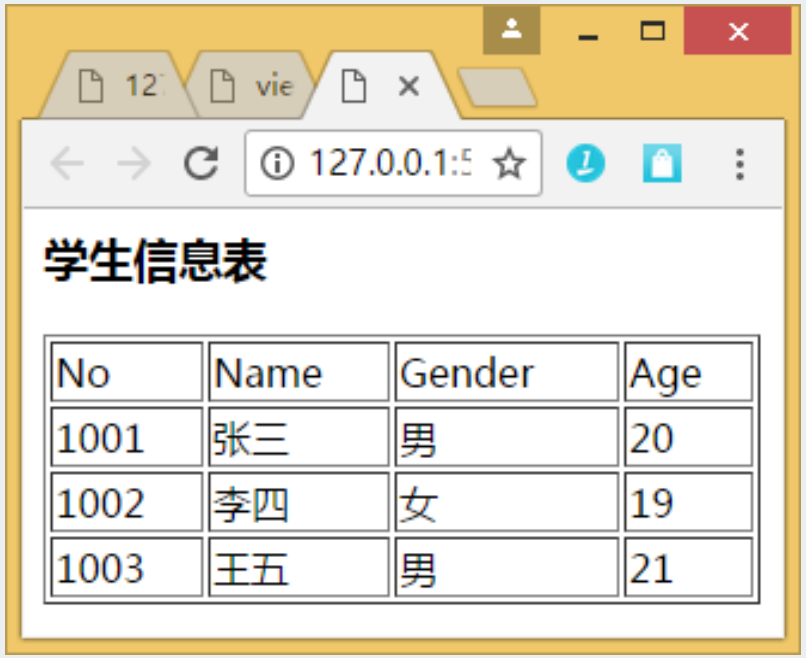
设计一个客户端的爬虫程序，它从这个网页上爬行学生的这些信息，存储到数据库中。学生数据库可以使用Sqlite数据库students.db。

PART TWO

服务器程序

服务器程序

服务器程序首先读取同一个目录下的students.txt文件，然后组成一张<table>的HTML表格用网页的形式呈现，效果如图1-8-1。



```
from flask import Flask,request
import os

app=Flask(__name__)

@app.route("/")
def show():
    if os.path.exists("students.txt"):
        st="<h3>学生信息表</h3>"
        st=st+"<table border='1' width='300'>"
        fobj=open("students.txt","rt",encoding="utf-8")
        while True:
            #读取一行，去除行尾部"\n"换行符号
            s=fobj.readline().strip("\n")
            #如果读到文件尾部就退出
            if s=="":
```

```
            break
        #按逗号拆分开
        s=s.split(",")
        st=st+"<tr>"
        #把各个数据组织在<td>...</td>的单元中
        for i in range(len(s)):
            st=st+"<td>"+s[i]+"</td>"
        #完成一行
        st=st+"</tr>"
    fobj.close()
    st=st+"</table>"
    return st

if __name__=="__main__":
    app.run()
```

PART Three

客户端程序

客户端程序

客户端程序访问http://127.0.0.1:5000/的网址，从中下载其HTML网页，这个网页的结果如下：

<h3>学生信息表</h3><table border='1' width='300'><tr><td>No</td><td>Name</td><td>Gender</td><td>Age</td></tr><tr><td>1001</td><td>张三</td><td>男</td><td>20</td></tr><tr><td>1002</td><td>李四</td><td>女</td><td>19</td></tr><tr><td>1003</td><td>王五</td><td>男</td><td>21</td></tr></table>

序要从这个HTML网页爬取数据，只要分解出第一行：

<tr><td>No</td><td>Name</td><td>Gender</td><td>Age</td></tr>

再次分解这一行的<td>...</td>数据，就知道这个表有哪些标题字段，这个表目前有No、Name、Gender、Age字段。

接下来再次分解出下一行<tr>...</tr>：

<tr><td>1001</td><td>张三</td><td>男</td><td>20>/td></tr>

再次分解这一行的<td>...</td>数据，得到No、Name、Gender、Age的数据依次是"1001"、"张三"、"男"、"20"，把这一行的数据写入对应的数据库即可。

要分解出<tr>...</tr>只要使用r"<tr>"与r"</tr>"的正则表达式即可，先用r"<tr>"匹配HTML代码，得到第一个<tr>的位置，再使用r"</tr>"匹配HTML字符串，得到第一个</tr>的位置，取出<tr>...</tr>的数据部分，再次使用r"<td>"与r"</td>"的正则表达式分解<td>...</td>的数据。

```
import urllib.request
import re

try:
    resp=urllib.request.urlopen("http://127.0.0.1:5000")
    data=resp.read()
    html=data.decode()
    m=re.search(r"<tr>",html)
    n = re.search(r"</tr>", html)
    while m!=None and n!=None:
        row=html[m.end():n.start()]
        a=re.search(r"<td>",row)
        b = re.search(r"</td>", row)
        while a!=None and b!=None:
```

```
            row=html[m.end():n.start()]
            a=re.search(r"<td>",row)
            b = re.search(r"</td>", row)
            while a!=None and b!=None:
                s=row[a.end():b.start()]
                print(s,end=" ")
                row=row[b.end():]
                a=re.search(r"<td>",row)
                b = re.search(r"</td>", row)
            print()
            html=html[n.end():]
            m=re.search(r"<tr>",html)
            n = re.search(r"</tr>", html)
    except Exception as e:
        print(e)
```

THANK YOU