# Python实验报告8

## 实验十： **网络爬虫**

## 实验目的：学习网络爬虫编写，获取数据

## 实验要求：

1. 能运用request库和beautifulsoup4库访问URL并解析获取的HTML

2. 能向百度等搜索引擎自动提交关键词并获取返回结果

## 实验内容：

1. 程序练习题10.1 修改教材实例代码20.1中的25-27行即可，只输出给定的省份的学校。要求输出江西省和北京市的高校

二、程序练习题10.2 实现USNEWS美国大学排名的爬虫，并打印结果。

提示：

1、美国大学排名网站：<https://www.usnews.com/best-colleges/rankings/national-universities>

2、上述网站只列出了排名前10的高校，当鼠标往下翻滚时才会加载后10个高校，仔细分析html源代码，找到显示后10个高校的网址。

三、分析百度图片搜索返回结果的HTML代码，编写爬虫抓取图片并下载形成专题图片库。

提示：

1、下载网页上的图片代码如下函数：imgUrl是图片的网址；destUrl是存储在本地的地址；fname是图片保存时的名字，默认值为空，图片保存时按网络上图片的名字保存

2、注意分析百度图片的搜索结果，图片的结果是保存在json格式中，关键字为thumbURL所对应的值就是图片的网址。

3、百度图片返回的搜索结果只有30幅图像，只有当鼠标下移才能看到后面30张图像，依此类推。分析HTML，找到相关网址下载后30张图像。

4、选择一个自己喜欢的明星，下载TA的90张照片。

## 实验思考：

网络爬虫应用一般分为两个步骤：1.通过网络链接获取网页内容；2.对获得的网页内容进行处理。这两个步骤分别使用不同的函数库：request和beautifulsoup4.

采用pip或pip3安装requests库和beautifulsoup4库

pip3 install requests

pip3 install beautifulsoup4

网络爬虫，相对于前面实验的基础理论知识，其更加贴近生活和实际应用。例如可以爬取试卷信息、获取网络的数据。

通过实验，我认为在进行网络爬虫操作之前，必须先得学会读取HTML语言，能根据参数找到对应标签，返回列表类型。