实验报告十（19/12/18）

姓名：白宇 学号：118010100327 专业班级：18无非3班

1. 实验名称：网络爬虫  
   二、实验目的：学习网络爬虫编写，获取数据  
   三、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
2. 实验内容
3. 程序练习题10.1 参考实例20，实现按照省份输出中国大学排名的功能。  
   提示：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  
   2. 程序练习题10.2 参考实例20，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  
   3. 程序练习题10.6 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  
   def downloadImageFile(imgUrl, destUrl, fname=''):  
       local\_filename = imgUrl.split('/')[-1]  
       print('Download Image File={}'.format(local\_filename))  
       try:  
           r = requests.get(imgUrl, stream=True)  
           r.raise\_for\_status()  
           if len(fname) == 0:  
               fname = local\_filename  
           print('fname={}'.format(fname))  
           with open(destUrl + "/" + fname, 'wb') as f:  
               for chunk in r.iter\_content(chunk\_size=1024):  
                   if chunk:  
                       f.write(chunk)  
                       f.flush()  
               f.close()  
           return r.status\_code  
       except:  
           return r.status\_code  
     
   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。
4. 实验小结

通过本次实验介绍了运用request库和beautifulsoup4库访问URL并解析获取的HTML，学习网络爬虫可以获取更多的数据源进行分析，将这些数据源进行采集，去掉无关数据。向百度等搜索引擎自动提交关键词并获取返回结果加以利用。