一、实验目的

1.熟悉网络爬虫的概念

2.熟悉网络爬虫的实现方式和流程

3.掌握一种下载网页的的方式：httpclient，selenium，OkHttp3等

4.熟练掌握网页内容的提取方法：Xpath，Css Selector，正则表达式等

二、实验要求

1.用常用的编程语言实现（Python/java/c#）

2.爬取《人民网-社会法制-频道首页》的所有内容，频道入口链接http://society.people.com.cn/GB/index.html;爬取时过滤掉外链和其它频道的链接。

3.需要提取的内容为：新闻的标题、内容、链接、发布时间、新闻图片（需下载原图）

4.提取的内容以文本文件的形式存储,每个链接的内容单独存储为一个文件（图片单独存储在一个文件夹中），文件用网页链接命名。

5.注意爬取每个链接的时间间隔，完成爬取后及时关闭程序。

三、高级内容

1. 链接去重

2. 分布式数据存储

3. 搜索引擎

4. 反爬技术