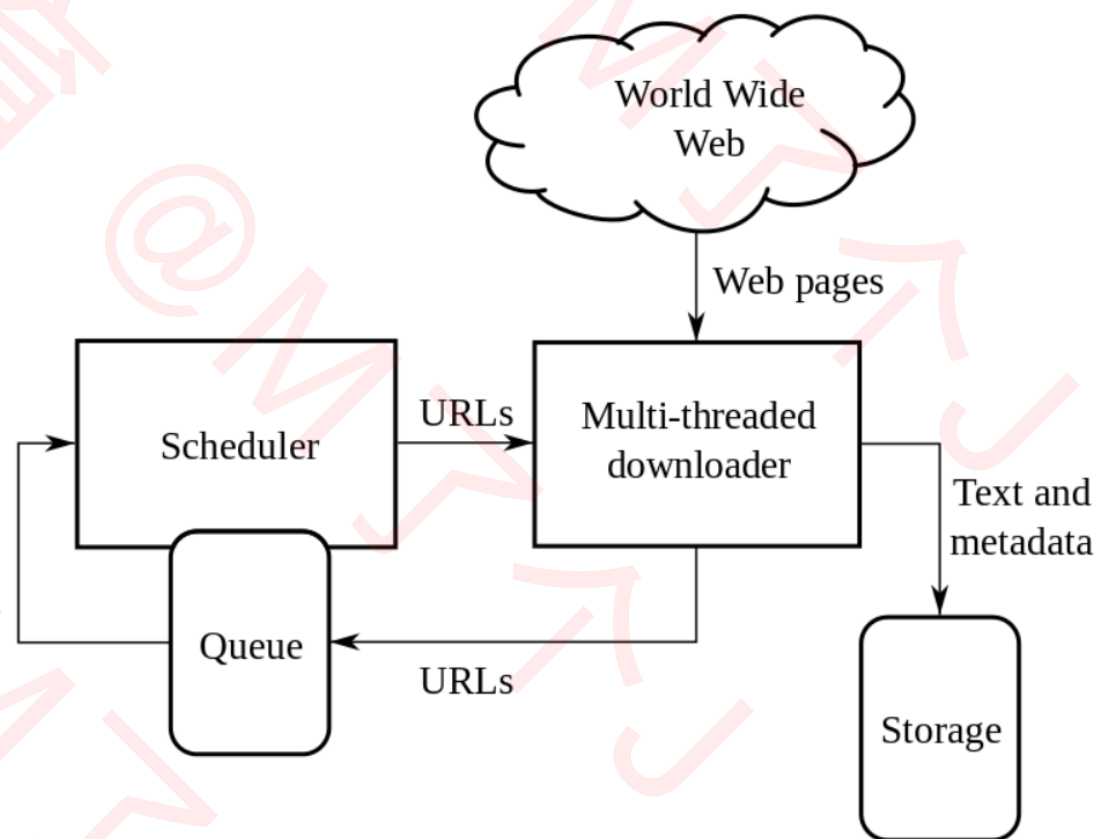
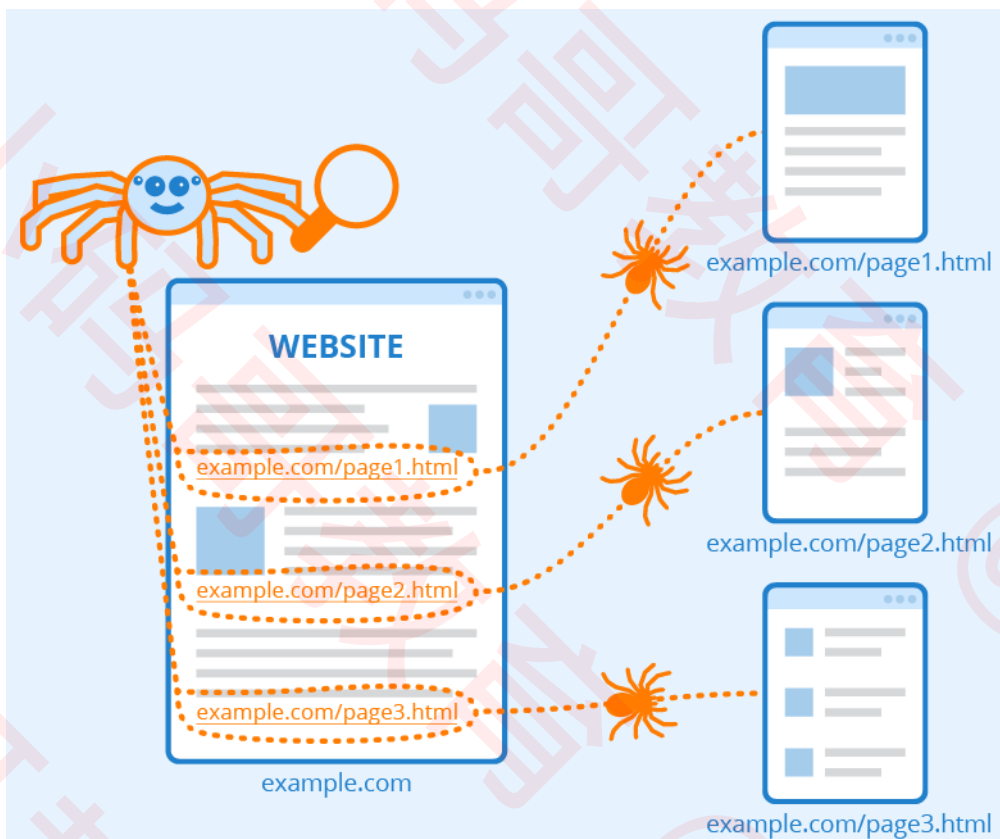
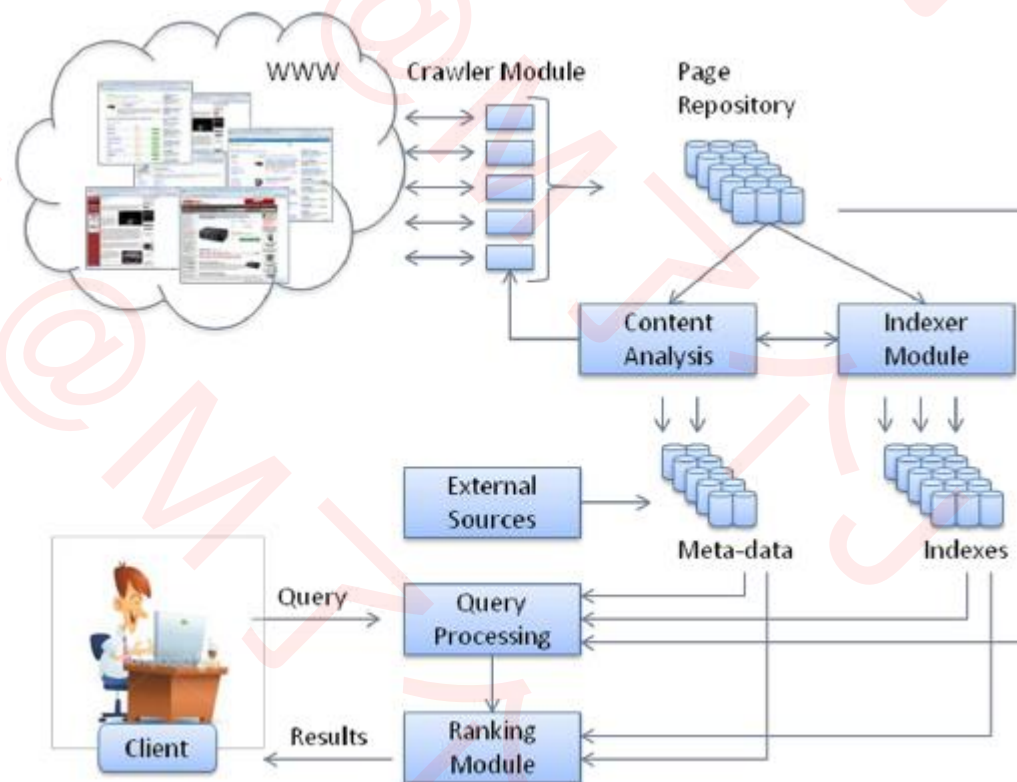
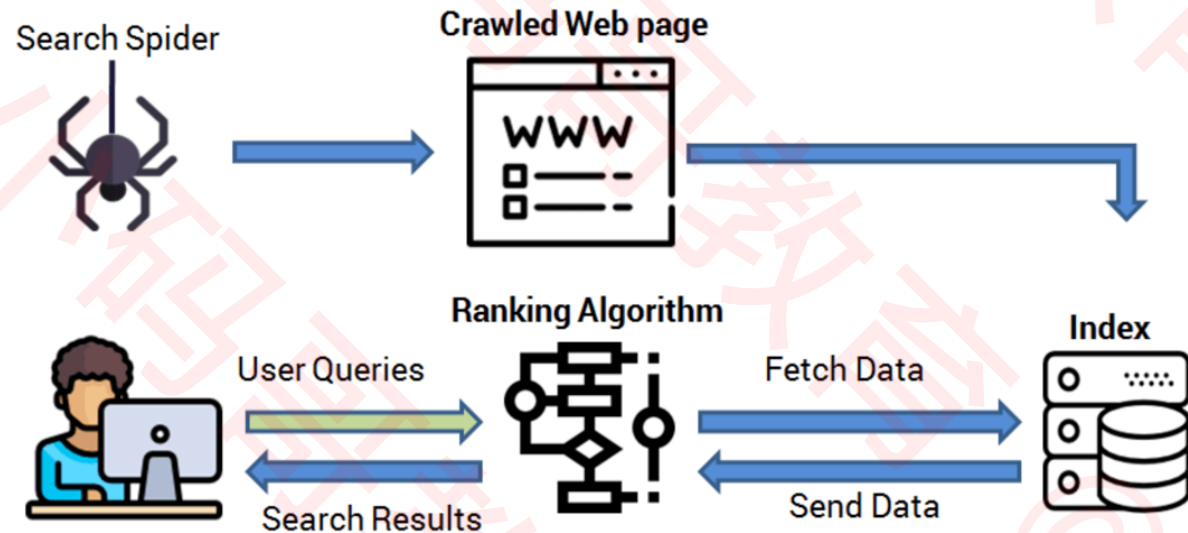


# 网络爬虫

- 网络爬虫 (Web Crawler) , 也叫做网络蜘蛛 (Web Spider)
- 模拟人类使用浏览器操作页面的行为, 对页面进行相关的操作
- 常用爬虫工具: Python的[Scrapy](#)框架



# 网络爬虫 — 搜索引擎



# 网络爬虫 — 简易实例

■ 可以使用Java的一个小框架[Jsoup](https://jsoup.org/packages/jsoup-1.13.1.jar)爬一些简单的数据

□ jar包

✓ <https://jsoup.org/packages/jsoup-1.13.1.jar>

✓ <https://mirror.bit.edu.cn/apache/commons/io/binaries/commons-io-2.8.0-bin.zip>

□ 爬取目标: <https://ext.se.360.cn/webstore/category>

```
String url = "https://ext.se.360.cn/webstore/category";
Elements apps = Jsoup.connect(url).get().select(".appwrap");
for (Element app : apps) {
    String img = app.selectFirst("img").attr("src");
    String name = app.selectFirst("h3").text();
    String intro = app.selectFirst(".intro").text();
    System.out.println(name + "_" + intro + "_" + img);

    String filepath = "F:/imgs/" + name + ".jpg";
    FileUtils.copyURLToFile(new URL(img), new File(filepath));
}
```

# 网络爬虫 — robots.txt

- robots.txt是存放于网站根目录下的文本文件，比如<https://www.baidu.com/robots.txt>
- 用来告诉爬虫：哪些内容是不应被爬取的，哪些是可以被爬取的
- 因为一些系统中的URL是大小写敏感的，所以robots.txt的文件名应统一为小写
- 它并不是一个规范，而只是约定俗成的，所以并不能保证网站的隐私
- 只能防君子，不能防小人
- 无法阻止不讲“武德”的年轻爬虫爬取隐私信息

# 网络爬虫 — robots.txt

允许所有的爬虫:

```
User-agent: *  
Disallow:
```

仅允许特定的爬虫: (name\_spider是爬虫名字)

```
User-agent: name_spider  
Allow:
```

仅禁止爬虫访问特定目录:

```
User-agent: name_spider  
Disallow: /private/
```

另一写法

```
User-agent: *  
Allow: /
```

拦截所有的爬虫:

```
User-agent: *  
Disallow: /
```

禁止所有爬虫访问特定目录:

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /images/  
Disallow: /tmp/  
Disallow: /private/
```

禁止所有爬虫访问特定文件类型

```
User-agent: *  
Disallow: /*.php$  
Disallow: /*.js$  
Disallow: /*.inc$  
Disallow: /*.css$
```