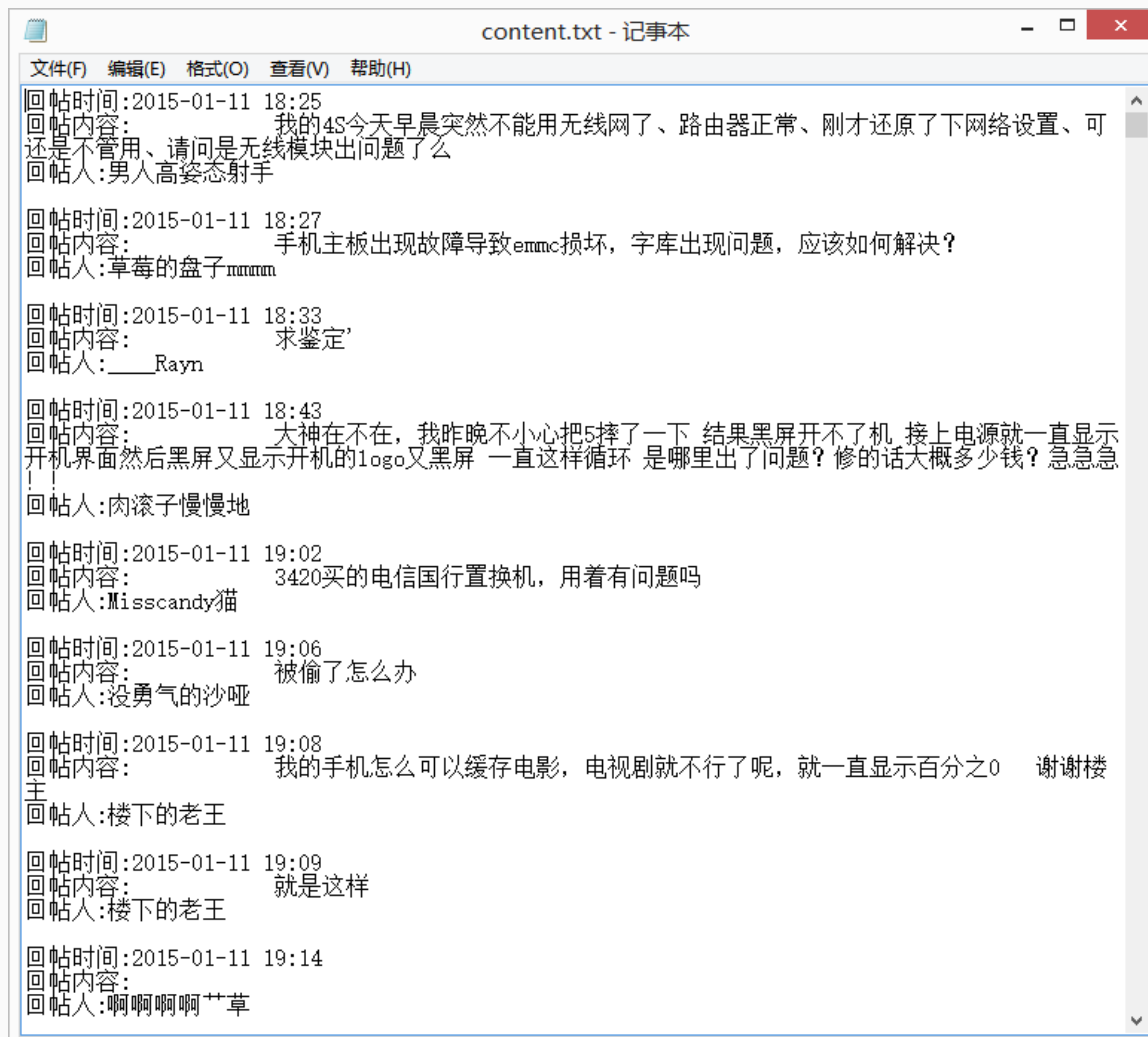


极客学院
jikexueyuan.com

XPath与多线程爬虫

XPath与多线程爬虫—效果展示



```
content.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
回帖时间:2015-01-11 18:25
回帖内容: 我的4S今天早晨突然不能用无线网了、路由器正常、刚才还原了下网络设置、可
还是不管用、请问是无线模块出问题了么
回帖人:男人高姿态射手

回帖时间:2015-01-11 18:27
回帖内容: 手机主板出现故障导致emmc损坏, 字库出现问题, 应该如何解决?
回帖人:草莓的盘子mmmm

回帖时间:2015-01-11 18:33
回帖内容: 求鉴定'
回帖人:___Rayn

回帖时间:2015-01-11 18:43
回帖内容: 大神在不在, 我昨晚不小心把5摔了一下 结果黑屏开不了机 接上电源就一直显示
开机界面然后黑屏又显示开机的logo又黑屏 一直这样循环 哪里出了问题? 修的话大概多少钱? 急急急
!
回帖人:肉滚子慢慢地

回帖时间:2015-01-11 19:02
回帖内容: 3420买的电信国行置换机, 用着有问题吗
回帖人:Misscandy猫

回帖时间:2015-01-11 19:06
回帖内容: 被偷了怎么办
回帖人:没勇气的沙哑

回帖时间:2015-01-11 19:08
回帖内容: 我的手机怎么可以缓存电影, 电视剧就不行了呢, 就一直显示百分之0 谢谢楼
主
回帖人:楼下的老王

回帖时间:2015-01-11 19:09
回帖内容: 就是这样
回帖人:楼下的老王

回帖时间:2015-01-11 19:14
回帖内容:
回帖人:啊啊啊啊**草
```

XPath与多线程爬虫— 课程概要

- 神器XPath的介绍与配置
- 神器XPath的使用
- 神器XPath的特殊用法
- Python并行化介绍与演示
- 实战——百度贴吧爬虫



神器XPath的介绍与配置

神器XPath的介绍与配置

- XPath是什么
- 如何安装使用XPath

神器XPath的介绍与配置— XPath是什么

- XPath 是一门语言
- XPath可以在XML文档中查找信息
- XPath支持HTML
- XPath通过元素和属性进行导航
- XPath可以用来提取信息
- XPath比正则表达式厉害
- XPath比正则表达式简单

神器XPath的介绍与配置— 如何安装使用XPath

- 安装lxml库
- `from lxml import etree`
- `Selector = etree.HTML(网页源代码)`
- `Selector.xpath(一段神奇的符号)`



神器XPath的使用

神器XPath的使用

- XPath与HTML结构
- 获取网页元素的Xpath
- 应用XPath提取内容

神器XPath的使用— XPath与HTML结构

- 树状结构
- 逐层展开
- 逐层定位
- 寻找独立节点

神器XPath的使用— 获取网页元素的XPath

- 手动分析法
- Chrome生成法

神器XPath的使用—应用XPath提取内容

- //定位根节点
- /往下层寻找
- 提取文本内容: /text()
- 提取属性内容: /@xxxx



神器XPath的特殊用法

神器XPath的特殊用法

- 以相同的字符开头
- 标签套标签

神器XPath的特殊用法— 以相同的字符开头

- starts-with(@属性名称, 属性字符相同部分)

```
<div id="test-1">需要的内容1</div>
```

```
<div id="test-2">需要的内容2</div>
```

```
<div id="testfault">需要的内容3</div>
```

神器XPath的特殊用法— 标签套标签

- string(.)

```
<div id="class3">美女，  
    <font color=red>你的微信是多少？ </font>  
</div>
```




Python并行化介绍与演示

Python并行化介绍与演示

- Python并行化介绍
- Map的使用

Python并行化介绍与演示— Python并行化介绍

- 多个线程同时处理任务
- 高效
- 快速

Python并行化介绍与演示— **map**的使用

- map 函数一手包办了序列操作、参数传递和结果保存等一系列的操作。
- `from multiprocessing.dummy import Pool`
- `pool = Pool(4)`
- `results = pool.map(爬取函数, 网址列表)`

实战——百度贴吧爬虫

实战——百度贴吧爬虫

- 目标网站: <http://tieba.baidu.com/p/3522395718>
- 目标内容: 跟帖用户名, 跟帖内容, 跟帖时间
- 涉及知识:

Requests获取网页

XPath提取内容

map实现多线程爬虫

XPath与多线程爬虫

在本次课程中我们学习了使用XPath提取网页内容与多线程爬虫，通过本次课程，你应该要掌握以下知识：

- 使用Xpath提取网页内容
- 使用map实现多线程爬虫

学习完本课以后，你制作的爬虫效率将大大提升。如果想继续提高，你可以继续在极客学院学习《Python定向爬虫入门》课程。

极客学院

jikexueyuan.com

中国最大的IT职业在线教育平台

