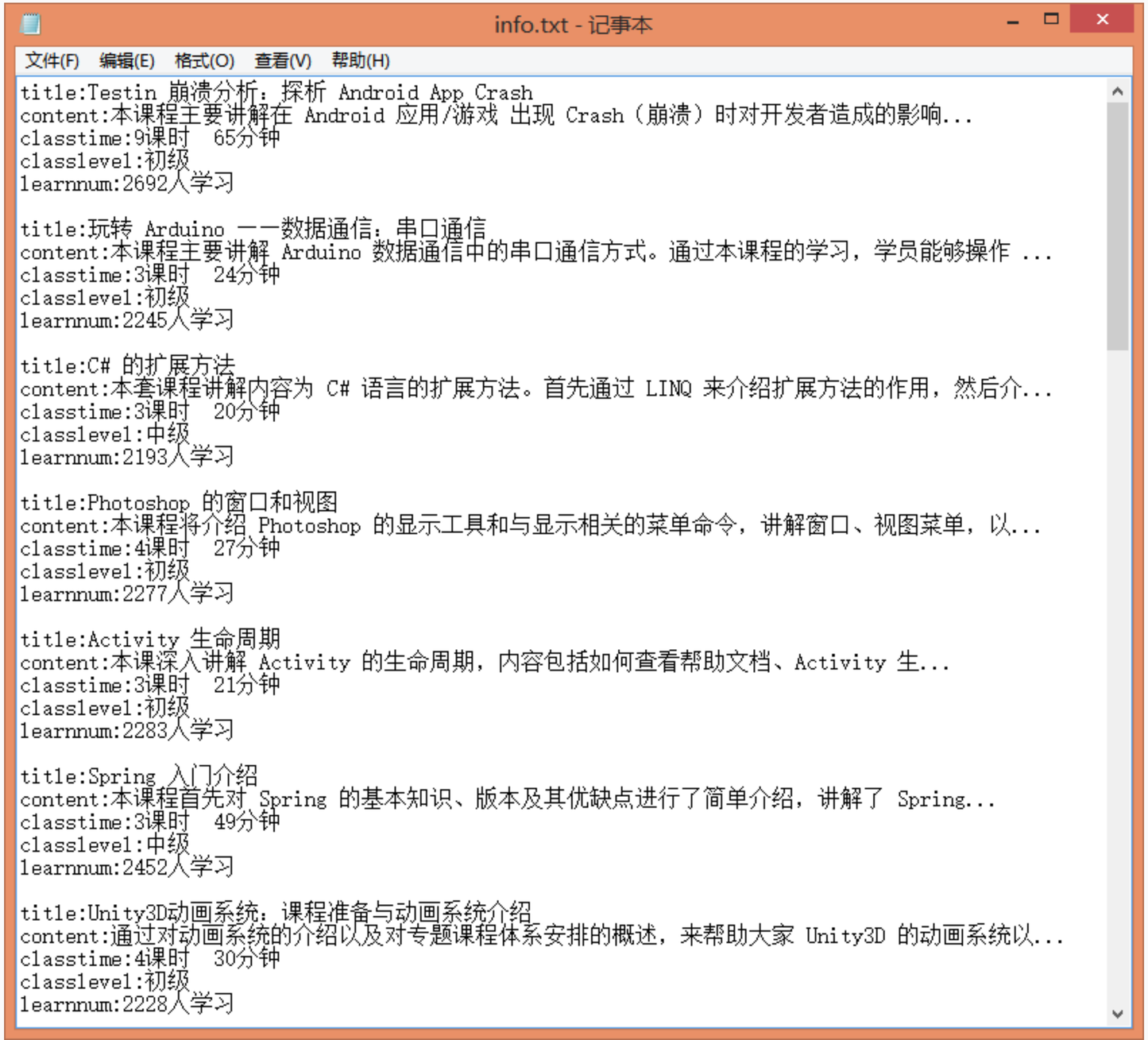
极客学院
jikexueyuan.com

单线程爬虫

单线程爬虫— 效果展示


info.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
title:Testin 崩溃分析：探析 Android App Crash
content:本课程主要讲解在 Android 应用/游戏 出现 Crash（崩溃）时对开发者造成的影响...
classtime:9课时 65分钟
classlevel:初级
learnnum:2692人学习
title:玩转 Arduino ——数据通信：串口通信
content:本课程主要讲解 Arduino 数据通信中的串口通信方式。通过本课程的学习，学员能够操作 ...
classtime:3课时 24分钟
classlevel:初级
learnnum:2245人学习
title:C# 的扩展方法
content:本套课程讲解内容为 C# 语言的扩展方法。首先通过 LINQ 来介绍扩展方法的作用，然后介...
classtime:3课时 20分钟
classlevel:中级
learnnum:2193人学习
title:Photoshop 的窗口和视图
content:本课程将介绍 Photoshop 的显示工具和与显示相关的菜单命令，讲解窗口、视图菜单，以...
classtime:4课时 27分钟
classlevel:初级
learnnum:2277人学习
title:Activity 生命周期
content:本课深入讲解 Activity 的生命周期，内容包括如何查看帮助文档、Activity 生...
classtime:3课时 21分钟
classlevel:初级
learnnum:2283人学习
title:Spring 入门介绍
content:本课程首先对 Spring 的基本知识、版本及其优缺点进行了简单介绍，讲解了 Spring...
classtime:3课时 49分钟
classlevel:中级
learnnum:2452人学习
title:Unity3D动画系统：课程准备与动画系统介绍
content:通过对动画系统的介绍以及对专题课程体系安排的概述，来帮助大家 Unity3D 的动画系统以...
classtime:4课时 30分钟
classlevel:初级
learnnum:2228人学习

单线程爬虫 — **课程概要**

- Requests介绍与安装
- 第一个网页爬虫
- 向网页提交数据
- 实战——极客学院课程爬虫

Requests介绍与安装

Requests介绍与安装

- Requests介绍
- Requests安装
- 第三方库安装技巧

Requests介绍与安装— **Requests介绍**

- Requests：HTTP for Humans
- 完美替代Python的urllib2模块
- 更多的自动化
- 更友好的用户体验
- 更完善的功能

Requests介绍与安装— Requests介绍

```python
#!/usr/bin/env python
# -*- coding: utf-8 -*-

import requests

r = requests.get('https://api.github.com', auth=('user', 'pass'))

print r.status_code
print r.headers['content-type']

# ------
# 200
# 'application/json'
```

```python

urllib2.install_opener(opener)

handler = urllib2.urlopen(req)

print handler.getcode()
print handler.headers.getheader('content-type')

# ------
# 200
# 'application/json'
```

Requests介绍与安装— **Requests安装**

- Windows：pip install requests
- Linux：sudo pip install requests

Requests介绍与安装— **第三方库安装技巧**

- 少用easy_install 因为只能安装不能卸载
- 多用pip方式安装
- 撞墙了怎么办？请戳->

http://www.lfd.uci.edu/~gohlke/pythonlibs/

第一个网页爬虫

第一个网页爬虫

- Requests获取网页源代码
- Requests与正则表达式

第一个网页爬虫— Requests获取网页源代码

- 直接获取源代码
- 修改http头获取源代码

Requests介绍与安装— **Requests与正则表达式**

使用Requests获取网页源代码，再使用正则表达式匹配出感兴趣的内容，这是单线程简单爬虫的基本原理。

向网页提交数据

向网页提交数据

- Get与Post介绍
- 分析目标网站
- Requests的表单提交

向网页提交数据— Get与Post介绍

- Get是从服务器上获取数据
- Post是向服务器传送数据
- Get通过构造url中的参数来实现功能
- Post将数据放在header提交数据

向网页提交数据— **分析目标网站**

- 网站地址：https://www.crowdfunder.com/browse/deals
- 分析工具：Chrome-审核元素-Network

向网页提交数据— **Requests表单提交**

- 核心方法：request.post
- 核心步骤：构造表单-提交表单-获取返回信息

实战——极客学院课程爬虫

实战——极客学院课程爬虫

- 目标网站：http://www.jikexueyuan.com/course/
- 目标内容：课程名称，课程介绍，课程时间，课程等级，学习人数
- 涉及知识：

  Requests获取网页

  re.sub换页

  正则表达式匹配内容

单线程爬虫

在本次课程中我们学习了使用Requests获取网页源代码，通过本次课程，你应该要掌握以下知识：

- Requests获取网页源代码
- 修改Http头绕过简单的反扒虫机制
- 向网页提交内容

学习完本课以后，你就算是定向爬虫入门了。如果想提高，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