法律声明

口本淉件包括：渍示文稿，示例，代码，题库，视频和声 音等，小象羊院拥有完全知识产权的权利；只限于善意 争习者在本淉程使用，不得在淉程范围外向任何笫三方 散播。任何其他人或机构不得盗版、复制、仿造其中的 创意，我们将保留一切通过法律手段追究违反者的权利。

回•.•

口淉程详情请咨询

* 微信公众号：大数据分析挖掘
* 新浪微博： ChinaHadoop



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn



零基础Python入门

－－梁斌

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

第九讲





AIR QUALITY INDEX I

空气质量指数计算6.0

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

案例描述

* 为了能有效地提取并利用网络信息并工作提高效率，出现了网络爬虫
* 利用网络爬虫实时获取城市的空气质量
* 高效地解析和处理HTML , beautifulsoup4

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

网页解析

* 结构化解析
* DOM (Document Object Model) , 树形结构

|  |  |  |
| --- | --- | --- |
|  | DOCUMENT |  |
|  | I |  |
|  | 《html> |  |
|  | ROOT ELEMENT |  |
| 《head》 |  | 《body》 |
| ELEMENT |  | ELEMENT |
| 《title》 |  | 《div》 |
| ELEMENT |  | ELEMENT |
| I |  |  |
| "MY TITLE" |  |  |
| TEXT | 《h1> | 《p> |
|  | ELEMENT | ELEMENT |
|  |  | “” |
|  |  | ELEMENT |

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

BeautifulSoup解析网页

BeautifulSoup

* 用于解析HTML或XML
* pip install beautifulsoup4
* import bs4
* 步骤

1 . 创建Bea utifu I Soup对象

1. 查询节点

find , 找到第—个满足条件的节点 find\_all, 找到所有满足条件的节点



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

BeautifulSoup解析网页

创建对象

* 创建BeautifulSoup对象
* bs = BeautifulSoup( url,

html\_parser, 指定解析器

enoding 指定编码格式（确保和网页编码格式—致）

）

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

BeautifulSoup解析网页

查找节点

* <a href='a.html'class='a—link'>next page</a>
* 可按节点类型属性或内容访问
* 按类型查找节点

• bs.find—all('a')

• 按属性查找节点

* bs.find—all('a', href='a.html')
* bs.find—all('a', href='a.html', string='next page')
* bs.find—all('a', class—='a—link') • 注意：是class—

• 或者bs.find—all('a', {'class':'a—link'})

互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

Next?

• 获取所有城市的AQI



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

疑问

口问题答疑： <http://www.xxwenda.com/>

·可邀清老师或者其他人回答问题

小象问答邀请@Robin\_TY回答问题



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn

联系我们

小象学院：互联网新技术在线教育领航者 －微信公众号：小象

－亲Ii讳良很女博： ChinaHadoop



互联网新技术在线教育领航者

E山蛊学陲

ChinaHadoop.cn