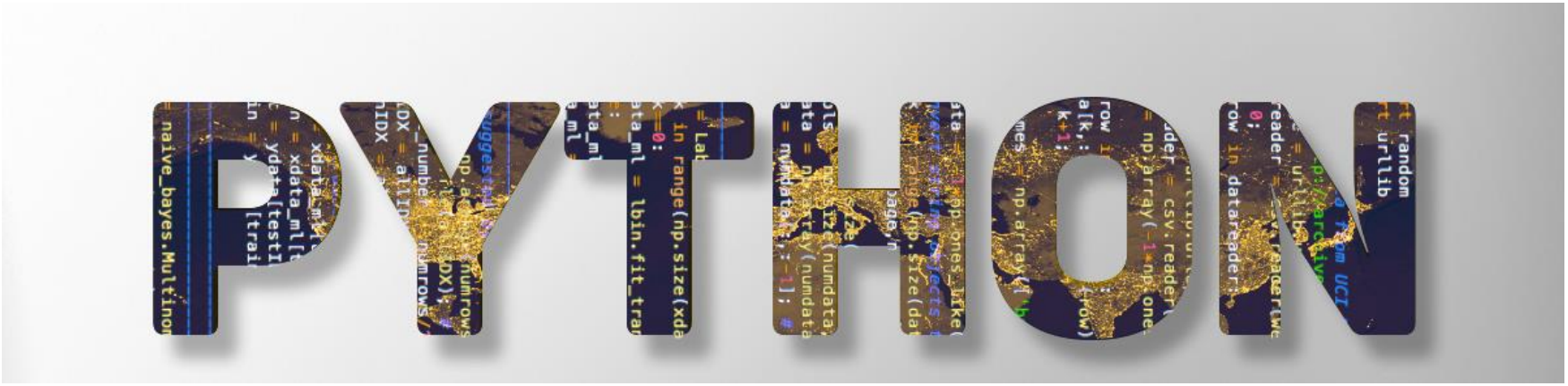
法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。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。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop



零基础Python入门

--梁斌

第九讲

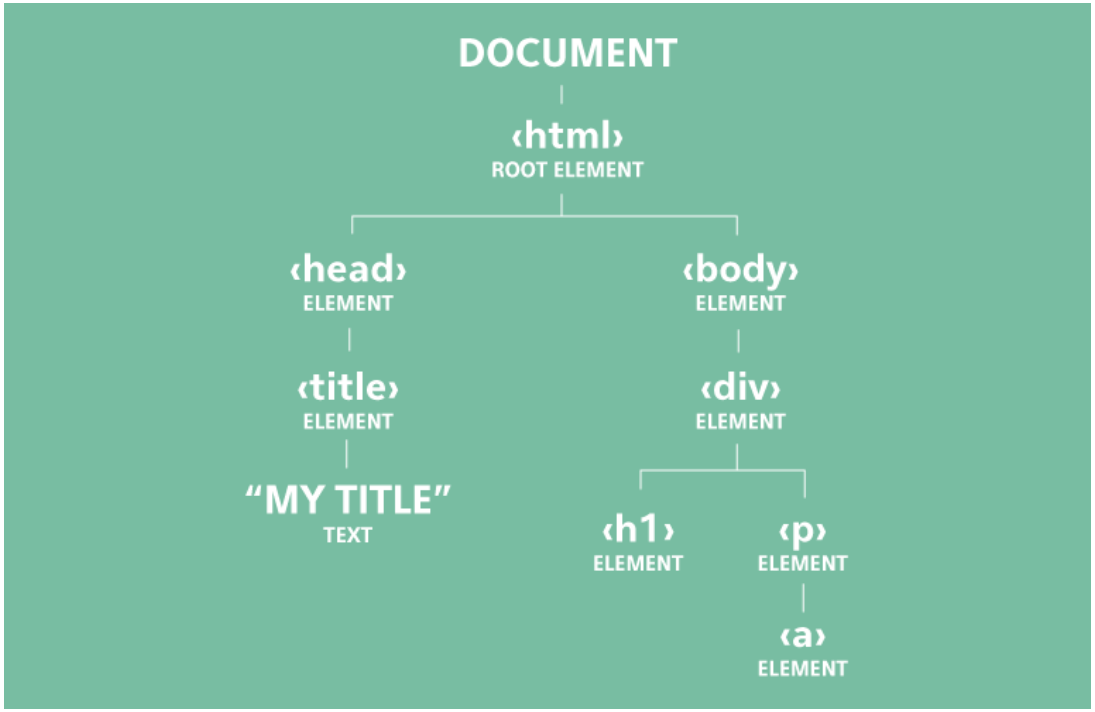
空气质量指数计算 6.0

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

网页解析

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



BeautifulSoup解析网页

BeautifulSoup

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

  1. 创建BeautifulSoup对象
  2. 查询节点

     find，找到第一个满足条件的节点

     find_all, 找到所有满足条件的节点

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一致）

)

BeautifulSoup解析网页

查找节点

- next page
- 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'})

Next?

- 获取所有城市的AQI

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

小象问答邀请 @Robin_TY 回答问题

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
- 新浪微博：ChinaHadoop