

第三课 网络数据的获取与表示

第四节 网页解析模块 (1)

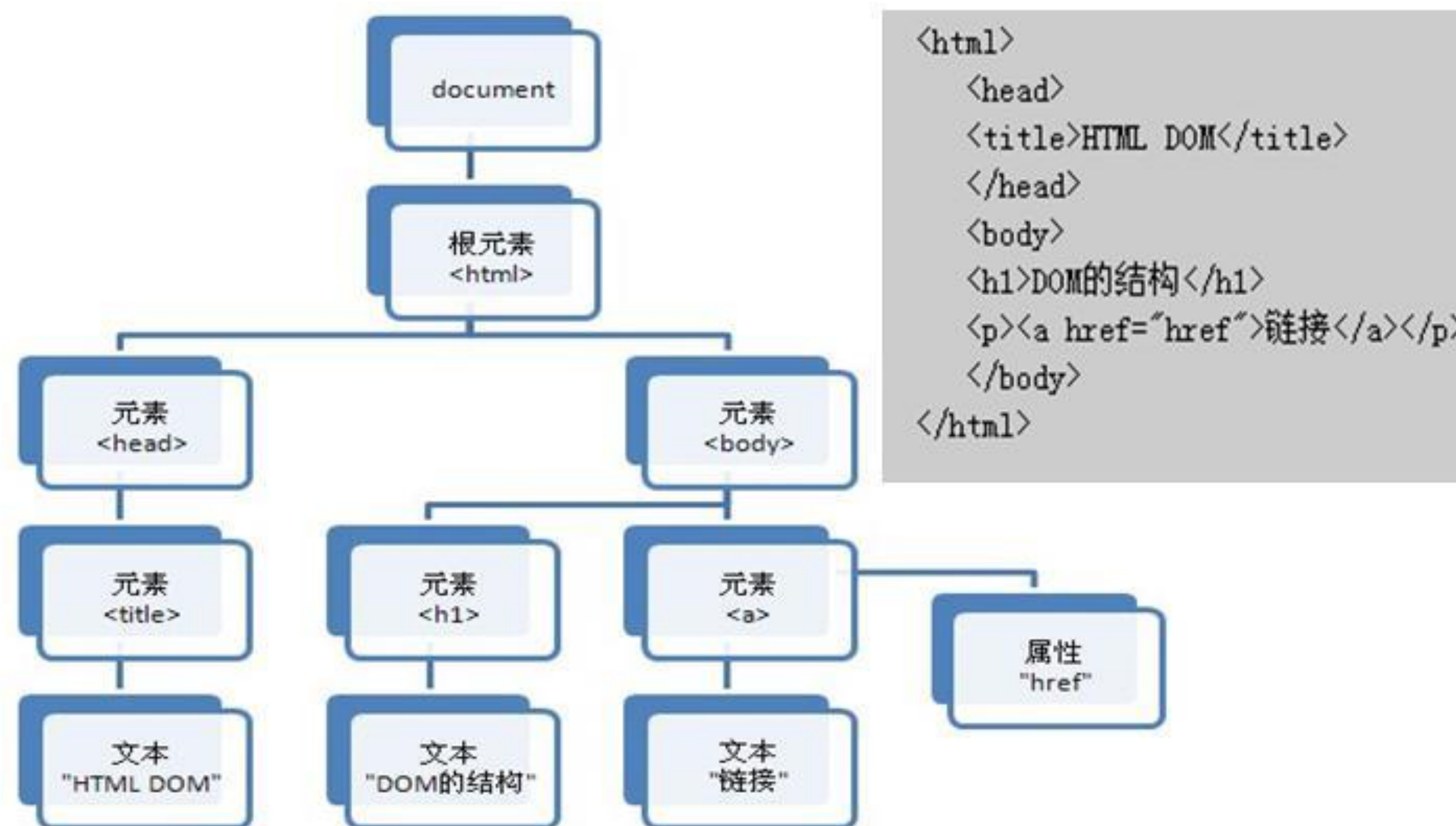
主 讲: Robin

网页解析模块

- 从已下载的网页中解析所需的内容
- 实现方式
 - 字符串匹配：如正则表达式
 - html.parser：Python自带的解析html的工具
 - BeautifulSoup：结构化的网页解析

网页解析模块

- 结构化网页解析
- DOM (Document Object Model) , 文档对象模型



网页解析模块

- BeautifulSoup使用步骤

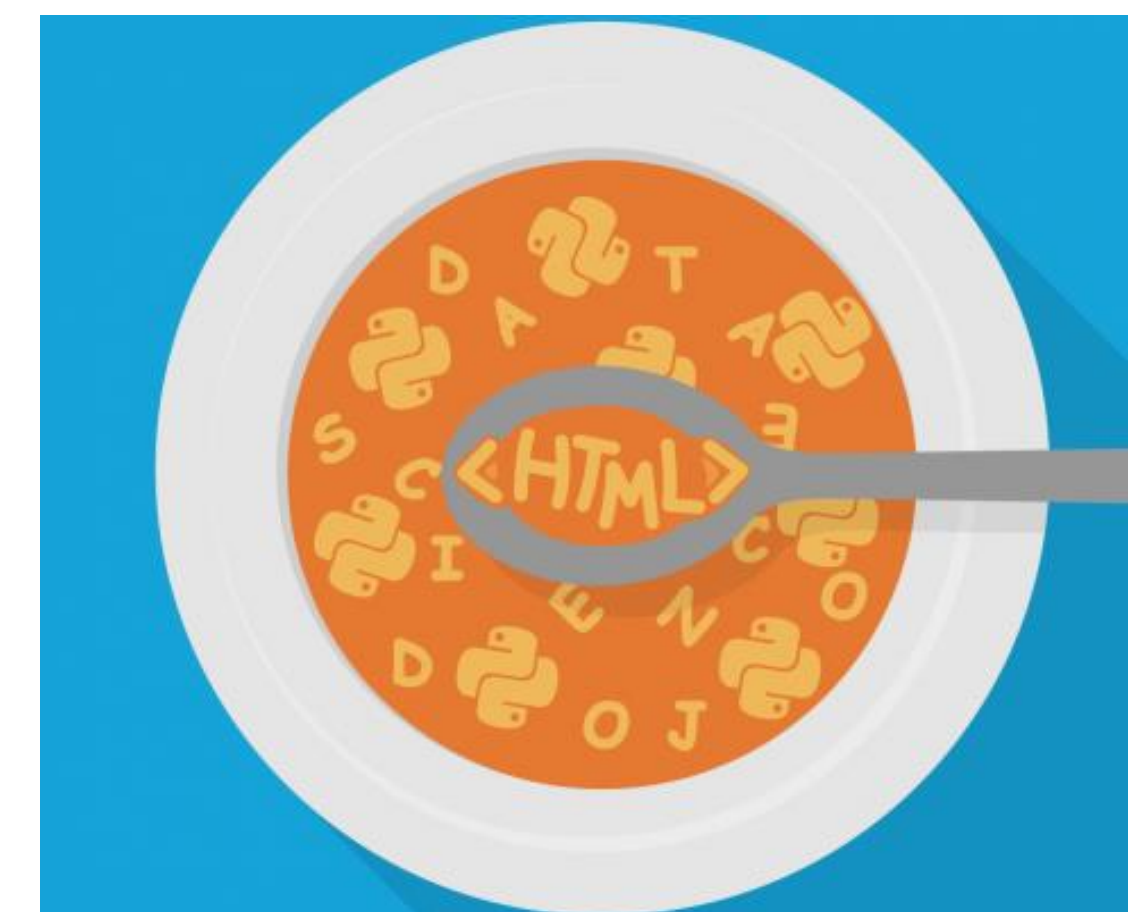
1. 创建BeautifulSoup对象

```
BeautifulSoup(  
    url,  
    html_parser, 指定解析器  
    encoding      指定编码格式（确保和网页编码格式一致）  
)
```

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

3. 获取节点信息

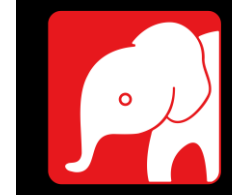


网页解析模块

- 查询节点
- 可按节点类型或属性查找

如：next page

- 按类型：find_all('a')
- 按属性：
 - find_all('a', href='test.html')
 - find_all('a', href='test.html', string='next page')
 - find_all('a', class_='test_link')



只为遇见明天更优秀的你！