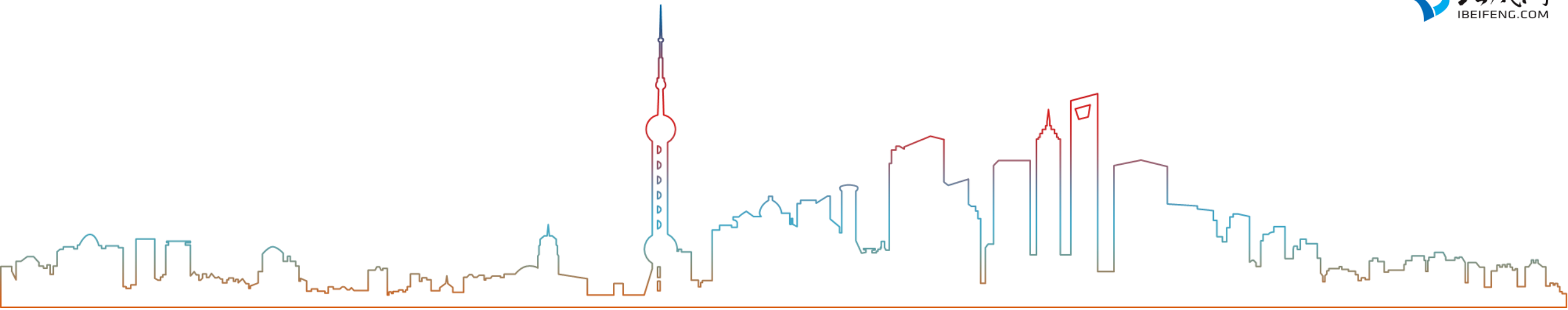
跟我学系列——走进Scrapy爬虫

浅谈Python爬虫

上海育创网络科技股份有限公司

主讲人：子沐



导 入

大数据时代，越来越多的人发现了数据的价值，所以为了能产生利益，那么就必须要得到更多的数据，如此才能创造价值。所以我们才要使用工具来快速获取数据，那么爬虫是你做好的选择。

课程介绍 - 课程体系

1

什么是爬虫

知道爬虫的定义。

2

为什么学习爬虫

理解学习Python爬虫的原因

本章任务

01 本章开始学习Python爬虫

02 需要同学们理解为什么使用爬虫，它的概念、特点。

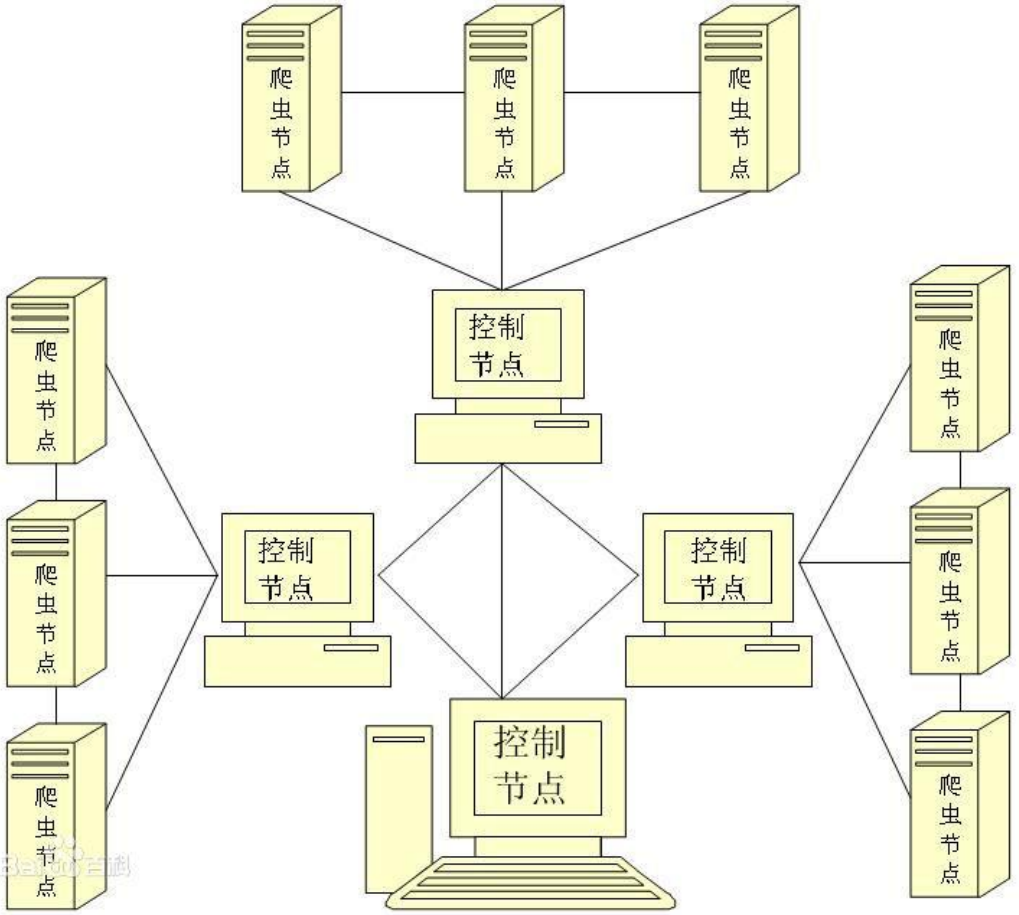
03 最主要的是需要大家学会如何使用基础的爬虫库。

01 什么是爬虫

本章内容

爬虫的由来

- 随着网络的迅速发展，万维网成为大量信息的载体，如何有效地提取并利用这些信息成为一个巨大的挑战。
- 如何获取这些数据



本章内容

爬虫的定义

- 网络爬虫，即Web Spider，是一个很形象的名字。把互联网比喻成一个蜘蛛网，那么Spider就是在网上爬来爬去的蜘蛛。网络蜘蛛是通过网页的链接地址来寻找网页的。

- 如果把整个互联网当成一个网站，那么网络蜘蛛就可以用这个原理把互联网上所有的网页都抓取下来。这样看来，网络爬虫就是一个爬行程序，一个抓取网页的程序。网络爬虫的基本操作是抓取网页。

本章内容

爬虫的工作过程

- 想象你是一只蜘蛛，现在你被放到了互联“网”上。
- 那么，你需要把所有的网页都看一遍。怎么办呢？
- 没问题呀，你就随便从某个地方开始。

本章内容

爬虫存在的意义

- 爬取数据
- 分析并推送
- 资源批量下载
- 数据监控
- 社会计算方面的统计和预测
- 机器学习
- 网络爬虫
- 建立机器翻译的语料库
- 搭建大数据的数据库

本章内容

爬虫怎么抓取网页数据

- 网页三大特征
- 爬虫的设计思路

通用爬虫与聚焦爬虫的区别

- 通用爬虫：目标、流程（爬取网页 - 存储数据 - 内容处理 - 提供检索/排名服务）、遵循Robots协议、缺点
- 聚焦爬虫，是"面向特定主题需求"的一种网络爬虫程序，它与通用搜索引擎爬虫的区别在于：***聚焦爬虫在实施网页抓取时会对内容进行处理筛选，尽量保证只抓取与需求相关的网页信息。***

02

为什么要学习Python爬虫

本章内容

Python语言的流行程度

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

Mar 2018	Mar 2017	Change	Programming Language	Ratings	Change
1	1		Java	14.941%	-1.44%
2	2		C	12.760%	+5.02%
3	**3**		**C++**	**6.452%**	**+1.27%**
4	5	↑	Python	5.869%	+1.95%
5	4	↓	C#	5.067%	+0.66%
6	6		Visual Basic .NET	4.085%	+0.91%
7	7		PHP	4.010%	+1.00%
8	8		JavaScript	3.916%	+1.25%
9	12	↑	Ruby	2.744%	+0.49%
10	-	⇈	SQL	2.686%	+2.69%
11	11		Perl	2.233%	-0.03%
12	10	↓	Swift	2.143%	-0.13%
13	9	⇊	Delphi/Object Pascal	1.792%	-0.75%
14	16	↑	Objective-C	1.774%	-0.22%
15	15		Visual Basic	1.741%	-0.27%
16	13	↓	Assembly language	1.707%	-0.53%
17	17		Go	1.444%	-0.54%
18	18		MATLAB	1.408%	-0.45%
19	19		PL/SQL	1.327%	-0.34%
20	14	⇊	R	1.128%	-0.89%

本章内容

各类爬虫框架比较

- 如果是定向爬取几个页面，做一些简单的页面解析，爬取效率不是核心要求，那么用什么语言差异不大。
- 如果是定向爬取，且主要目标是解析js动态生成的内容
- 如果爬虫是涉及大规模网站爬取，效率、扩展性、可维护性等是必须考虑的因素时候

本章内容

为什么Python适合写爬虫

- 抓取网页本身的接口
- 对数据库的操作能力（mysql、mongo等）
- 爬取效率
- 代码量
- 对页面解析能力
- 网页抓取后的处理（去重，过滤，清洗，保存）

本章内容

爬虫 - 反爬虫 - 反反爬虫 之间的斗争

- 爬虫做到最后头疼的不是复杂的页面，也不是晦涩的数据，归根结底是人
- 数据的价值

THANKS !

上海育创网络科技股份有限公司

主讲人：子沐老师