

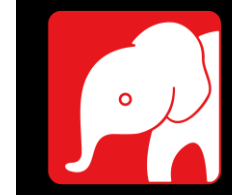
第三课 网络数据的获取与表示

第二节 URL 管理模块

主 讲: Robin

URL管理模块

- 防止重复爬取或循环指向
- 功能：
 - 可添加新的URL
 - 管理已爬取的URL和未爬取的URL
 - 获取待爬取的URL
- 实现方式
 - Python的set数据结构, why?
 - 数据库中的数据表, how?



只为遇见明天更优秀的你！