作业-爬虫及数据库

TA. Hui Xu

Dept. Computer Science & Engineering,

Shanghai Jiaotong University

hnxxjyt@sjtu.edu.cn

DDL: Apr. 19, 23:59

百度百科数据抓取及解析

- 任务目标

抓取百度百科科学类别中，任一科学分类词条库（如数理科学）内的所有内容。并解析每一个词条的相关信息，如介绍等，具体抽取哪些内容不做限制。并将信息存放到自己本地的mysql数据库中，通过python接口实现对数据库正常的查询（在IDE中查询对应数据库内指定条目的信息，如三角函数公式的介绍）。

作业提交形式

- 每小组1份报告，报告语言要求为全英文，报告内容包含但不限于：
  - 本次作业重难点
  - 本次作业完成思路
  - 本次作业中碰到的问题以及如何解决
- 每小组将爬取的信息构造成字典并dump成json格式
- 所有内容打包为一个压缩包，上传至ftp

作业提交

- 提交方式:上传ftp
- 地址:ftp://public.sjtu.edu.cn
- 用户名:hnxxjyt
- 密码:public
- 打开方式：文件资源管理器/filezilla等软件
- 报告上传：/upload/爬虫+数据库
- 报告命名格式：组号_version
- 如：1_v1.pdf