



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）。

三角函数公式 编辑 讨论 39

本词条由“科普中国”科学百科词条编写与应用工作项目 审核。

三角函数是数学中属于初等函数中的超越函数的函数。它们的本质是任何角的集合与一个比值的集合的变量之间的映射。通常的三角函数是在平面直角坐标系中定义的。其定义域为整个实数域。另一种定义是在直角三角形中，但并不完全。现代数学把它们描述成无穷数列的极限和微分方程的解，将其定义扩展到复数系。

三角函数公式看似很多、很复杂，但只要掌握了三角函数的本质及内部规律，就会发现三角函数各个公式之间有强大的联系。而掌握三角函数的内部规律及本质也是学好三角函数的关键所在。

首页	分类	秒懂百科	特色百科	用户	权威合作
历史上的今天	艺术 科学	懂啦	数字博物馆	蝌蚪团	合作模式
百科冷知识	自然 文化	秒懂星课堂	非遗百科	燃梦计划	常见问题
图解百科	地理 生活	秒懂大师说	百度数说	百科任务	联系方式
	社会 人物	秒懂看瓦特	城市百科	百科商城	
	经济 体育	秒懂五千年	恐龙百科		
	历史	秒懂全视界	多肉百科		

 健康医疗 风热感冒、肺结核、急性心肌梗死 查看详情 >>	 航空航天 飞行力学、超高速、光子卫星 查看详情 >>	 天文学 大爆炸宇宙论、日食、太阳系外行星 查看详情 >>	 环境生态 二十四节气、雾霾、沙尘暴 查看详情 >>
 农业科学 转基因食物、羊奶、漂白粉 查看详情 >>	 生命科学 拉布拉多猎犬、英国短毛猫、接吻鱼 查看详情 >>	 数理科学 微积分、黄金分割点、拓扑学 查看详情 >>	 心理学 同性恋、女性更年期综合征、心灵感应 查看详情 >>
 信息科学 图灵测试、增强现实、八进制 查看详情 >>	 工程技术 X射线、原子、原油 查看详情 >>	 化学 正辛基三氯硅烷、五氟苯胺、2-硝基苯酚 查看详情 >>	 地球科学 近震震级、地球放射性、化学岩 查看详情 >>
 其他 沸水堆、石墨矿、高岭土 查看详情 >>			



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

