网络爬虫实战项目说明

WS00

嵩天

www.python123.org

预期目的

开源实战项目

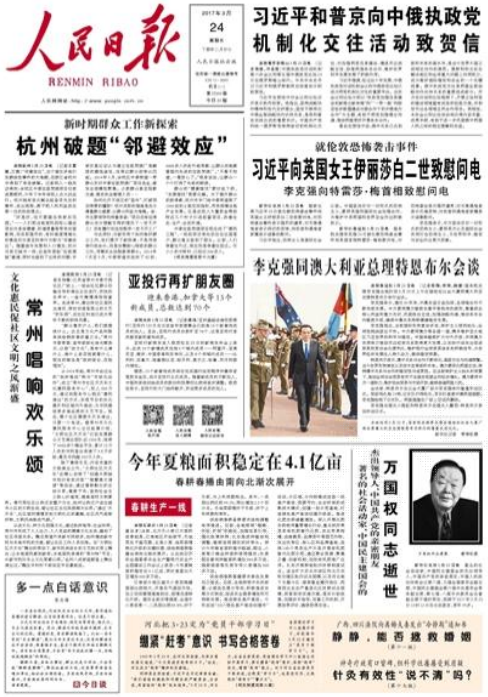
- 用合理时间实践网络爬虫技术，培养工程实践能力
- 搭建具备商用潜力的7x24运行系统
- 构建针对开源项目的协同开发及学习社区
- 探索围绕在线开放课程的开源项目组织和管理方式

实战项目简介

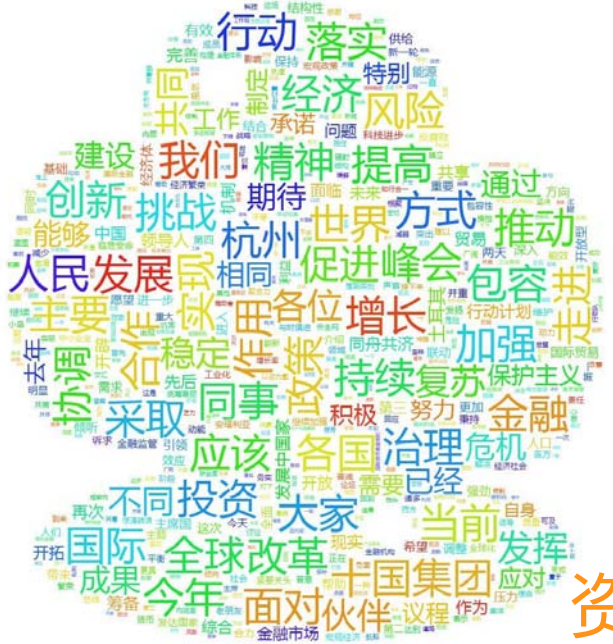
- 项目名称：纸媒资讯爬取与展示
- 项目路线：资讯爬取→数据清洗→资讯展示
- 项目组织方式：自愿报名参与，自选爬取对象，负责人协调

实战项目需求

纸媒资讯爬取与展示



报纸资讯



资讯词云

实战项目需求

爬取部分

- 通过纸媒官方网站爬取正式发行的报纸所刊载资讯，要求：
  - 所爬取资讯必须与所发行的纸质内容一致，不多尽量不少
  - 爬取文本为主，图片为辅（图片可选）
  - 以文章为爬取单元，区分：标题、作者、正文及版面排序（排序可选）
  - 所有爬取的资讯以发行纸媒的编号（发行日期）存储和组织
  - 不限于中国境内发行的报纸，鼓励外文报纸资讯的爬取

实战项目需求

资讯处理部分

- 所有资讯至少分为如下7个类别：
  - 国际新闻、国内新闻、娱乐、体育、经济、广告和其他
- 更多类别可以在开发中逐步细分
- 每个类别以版面顺序为基础区分重要资讯和一般资讯（可选）

实战项目需求

展示部分

- 以词云和列表方式展示所查询的资讯信息
- 以词云方式展示的查询方式包括但不限于：
  - 按照日期（含时间段）、类别、文章题目、作者、全文、纸媒名称
  - 根据含图片资讯和非图片资讯（可选）
- 展示采用Web形式提供

技术要求

纸媒资讯爬取与展示

- 爬虫采用Scrapy框架，7x24自动运行，单个纸媒爬取周期为24小时，CentOS系统
- 各纸媒为单独爬虫模块，采用适当的反爬技术，考虑Robots协议
- 信息存储采用MySQL或MongoDB或文件方式
- 词云展示采用wordcloud等Python第三方库
- 全系统采用Python语言，Web采用HTML5/CSS/JS，不要求使用框架

人员组织和管理

学生主导、自由报名、贡献受益

- 老师仅负责总体技术要求和设计评价，不参与实际系统开发和答疑
- 征集2位同学担任总负责人，分别负责：人员分工与进度管理、系统架构设计与维护
- 所有参与者，每人最多负责1份纸媒爬取，要求完成从爬取、分类到展示的完整功能
- 参与者也可以选择开发系统内其他辅助功能，如数据库、HTML展示等
- 完成单一纸媒全套功能或贡献超过>200行可用代码的参与者为“实际贡献者”
- 项目所有代码对实际贡献者开源，暂不考虑全开源
- 采用CC协议，知识产权归创作者所有，老师可根据参与情况出具纸质参与证明
- 项目未来发展由老师与所有实际贡献者共同决策

项目联络

- 项目以QQ/微信群等线上平台为主要联络方式
- 由项目总负责人确定时间进度
- 老师视项目推进需求，在适当时候通过直播或线下方式给予技术支持或指导
- 老师负担系统运行服务器、数据存储、域名等费用（云服务）

简单说

如果你对管理本项目有兴趣，且认为个人能力和时间可胜任“总负责人”，发送邮件至songtian@bit.edu.cn，附简历，说明胜任理由和可担任期限。

如果你对项目开发有兴趣，请选择一家正式发行的纸媒官网作为爬取和展示对象，在总负责人到位前可以先开始，待人选确定，向总负责人提交个人信息和爬取网站信息，开展团队合作。为保证爬取质量，每一纸媒仅限一位代码贡献者。

该项目所构建系统将为后续自然语言理解、新闻传播规律分析、深度舆论挖掘、信息预警等众多应用提供基础性数据。