PkuJava32课程项目

1501210718王乙闲

1501210890 程 权

1501210381林雄辉

1、项目题目：

技术类博客数据的获取与分类

2、介绍：

爬取技术类博客，将博客内容以.txt形式存入文件系统数据库，再对博客内容进行汇总分类，如java、python、hadoop等，java目录又分为算法，多线程，分布式等。

3、所涉及的技术（还不确定具体技术有哪些）：

技术类博客数据的获取，通过java语言实现，采用宽度优先搜索策略，利用HTTP请求、Html Parser、Jsoup、User-Agent判别等解决“反爬虫”问题和实现博客内容的爬取；

技术类博客数据分类，主要在分布式平台hadoop/spart上完成，利用机器学习的分类算法对文章进行有指导的分类，然后预测新文章的类别，自动给用户提供新文章的分类标签。