项目题目：

爬取技术类博客。将技术类博客进行汇总分类，如java，python，hadoop，spark等。java目录又分为算法，多线程，分布式等。

项目简介：

以CSDN为例，对CSDN相关技术博客进行汇总分类，并将分类结果存放与数据库中。用户可检索相关结果，或以图表形式查看各类技术的统计报告。

分类依据：博客标题，标签、博客分类等。

存储数据：博客标题、分类、阅读次数、发表时间、博客链接等。

项目技术（暂拟）：

爬虫框架：webMagic

全文检索工具：Lucene

图表工具：Echarts

编程语言：Java

数据库：MySQL