Simple Search Engine

简介

本项目实现了一个简单的搜索引擎，下面简单介绍要点，详细技术要点参考`技术文档`

- 功能要点： 查询网页、异步爬虫、添加文档、缓存快速删除文档、搜索文档、其他数据维护功能
- 技术要点：
    1. 项目包含 web、crawler、engine 三大模块，全部由 scala 编写，使用akka提供Actor并发模型支持
    2. web模块：由 Spring Boot 后端和静态 HTML 前端（原生ES6）组成
    3. crawler模块：基于Actor模型的异步爬取和异步添加/更新文档、定期搜集、种子搜索+维护url列表策略
    4. engine模块：Actor并发与Future异步方式、缓存快速删除文档设计、缓存索引表和url库、文件读写锁
    5. 其他模块：定时爬取任务，定时缓存持久化，分词功能
- 搜索引擎过程要点：
    1. web搜集：异步爬虫、定期搜集、种子搜索+维护url列表策略
    2. 预处理：分词、正则提取url、正则提取正文
    3. 维护文件：网页库（文档信息、文档正文）、索引表（文档id、内容hash、网页库文件偏移量）、倒排表（词、文档id、文档位置列表）、文档计数表（总文档数、各文档词数）、url库（url、正文hash）
    4. 查询服务：包含信息列表三要素（标题、URL、摘要）、字符串切份生成摘要
    5. 查询排序：BM25排序算法

用户手册

演示

本项目提供简单的DEMO，以便跳过部署和运行从而直接使用：
http://39.106.185.26:8080/

配置

修改`Config.scala`相关项目，如：文件存储位置、超时时间、hash范围等

编译

使用Maven管理，springboot插件编译打jar包：`mvn clean package`

部署

建议使用`systemd`：

1. 编写`sse.service`，这里提供简单例子：

   ```
   [Unit]
   Description=Simple search engine

   [Service]
   Type=simple
   User=root
   ExecStart=/usr/bin/java -jar /opt/sse/simple-search-engine.jar --
   server.port=8080

   [Install]
   WantedBy=multi-user.target
   ```

2. 复制jar包到对应位置
3. `systemctl daemon-reload`重新加载

运行

`systemctl start sse`

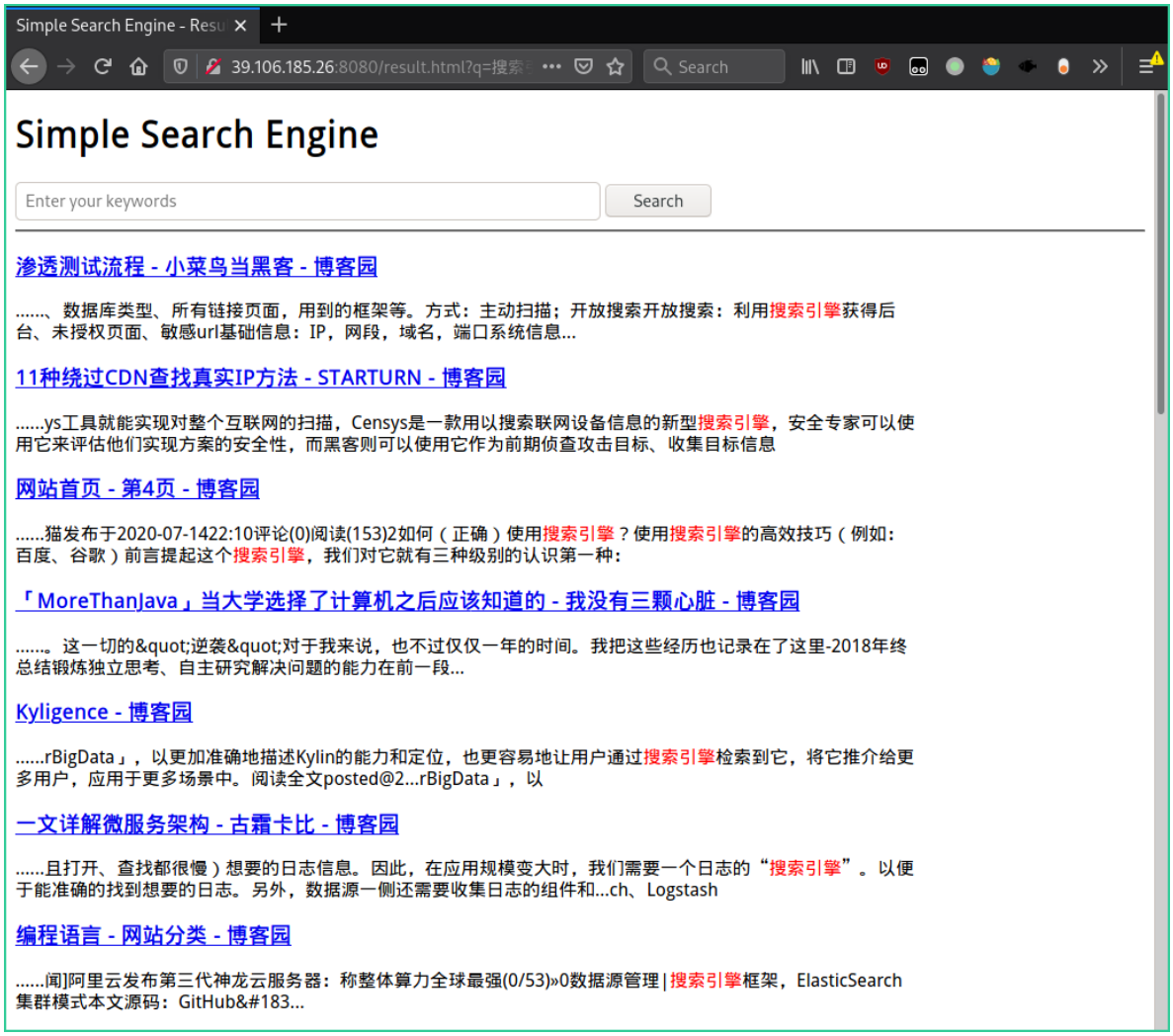
使用

访问8080端口即可使用

使用截图

DEMO配置中设置了只爬取博客园的策略，所以搜索结果都是博客园

PC端


Simple Search Engine - Resu
39.106.185.26:8080/result.html?q=搜索
Search
Simple Search Engine
Enter your keywords
Search
渗透测试流程 - 小菜鸟当黑客 - 博客园
……、数据库类型、所有链接页面，用到的框架等。方式：主动扫描；开放搜索开放搜索：利用搜索引擎获得后台、未授权页面、敏感url基础信息：IP，网段，域名，端口系统信息…
11种绕过CDN查找真实IP方法 - STARTURN - 博客园
……ys工具就能实现对整个互联网的扫描，Censys是一款用以搜索联网设备信息的新型搜索引擎，安全专家可以使用它来评估他们实现方案的安全性，而黑客则可以使用它作为前期侦查攻击目标、收集目标信息
网站首页 - 第4页 - 博客园
……猫发布于2020-07-1422:10评论(0)阅读(153)2如何（正确）使用搜索引擎？使用搜索引擎的高效技巧（例如：百度、谷歌）前言提起这个搜索引擎，我们对它就有三种级别的认识第一种：
「MoreThanJava」当大学选择了计算机之后应该知道的 - 我没有三颗心脏 - 博客园
……。这一切的"逆袭"对于我来说，也不过仅仅一年的时间。我把这些经历也记录在了这里-2018年终总结锻炼独立思考、自主研究解决问题的能力在前一段…
Kyligence - 博客园
……rBigData」，以更加准确地描述Kylin的能力和定位，也更容易地让用户通过搜索引擎检索到它，将它推介给更多用户，应用于更多场景中。阅读全文posted@2…rBigData」，以
一文详解微服务架构 - 古霜卡比 - 博客园
……且打开、查找都很慢）想要的日志信息。因此，在应用规模变大时，我们需要一个日志的“搜索引擎”。以便于能准确的找到想要的日志。另外，数据源一侧还需要收集日志的组件和…ch、Logstash
编程语言 - 网站分类 - 博客园
……闻]阿里云发布第三代神龙云服务器：称整体算力全球最强(0/53)»0数据源管理|搜索引擎框架，ElasticSearch集群模式本文源码：GitHub&#183…

移动端

Simple Search Engine - Result

Simple Search Engine

Enter your keywords Search

待解决高分问题_博问_博客园

.....源于数据库查询的数据集合。suqq小白回答(0)浏览(38)4天前1回答数100vue移动端项目中pdfjs-dist插件报错报错信息有下面这两条Uncaugh...<'pd

博问_程序员问答社区，解决您的IT难题_博客园

.....据，导致爬虫爬不到，渲染的页面就不对了，我看了下network，也没看到发出过请vue axios遗失的美好灬回答(0)浏览(24)1周前3回答数5webuplo...cd~user1进入个

所有评论 - 博客园

.....经很开心了(●ˇ∀ˇ●)倾叶子佮评论于2020-07-1522:41Re:记录下vue中引用echarts出现"TypeError:Cannotrea...步加载完可以不h祝h评

精华区 - 第4页 - 博客园

.....时下热门的任何一种技术选型都有成熟的

方案，比如：vue + vue-i18nang
ular + angular-tran...

伍华聪 - 博客园

.....见的图标，不过数量不是很多，应该是30