

Naïve SE

李程浩 180110304

Catalogue

1. Introduction.....	1
2. How to build.....	1
3. Structure.....	2
4. Evaluation.....	4
5. Project Catalogue.....	6

1. Introduction

This full name of this project is called “Naïve Search Engine” which focuses on search related information by queries given by users. Users are provided with a simple website and when they enter the queries they want to search, and the website returns 10 most relevant websites that satisfy the queries. The websites are limited in HITSZ and contents are all in Chinese.

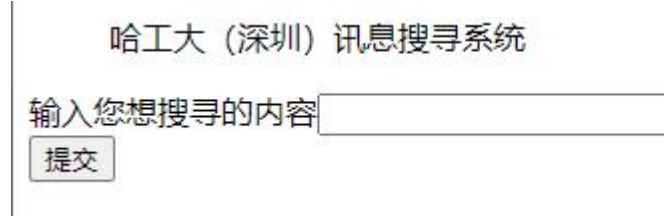
The whole project uses library including Lucene, Requests, BeautifulSoup, Tomcat, IKAnalyzer, and Servlet. At first, we use Request (Python) to build crawlers and then split the useful contents to be stored. After that, we use Lucene (JAVA) to perform indexing and querying. To support Chinese indexing and querying, we use IKAnalyzer (JAVA) to help split Chinese words. Then, to achieve website searching user interface, we use Tomcat and Servlet (JAVA) to create a virtual Internet server which enables user to access websites in Browser.

This catalogue of this project is discussed in the end of this report. Moreover, I want to declare that the final website only shows little part of the whole project. I actually do a lot more in my codes, but I didn’t enables them in the website that users finally face. The reason is that we don’t want to focus too much on the UI, and function is all that matters to me.

2. How to build

As we have built a virtual Internet server (VIS) to support website user interface, we need to start VIS first. First, we enter *NaiveSE_LiChenghao\Tomcat\bin\startup.bat* to start VIS. To successfully achieve this, you have to make sure that you have correctly set environment variable JRE_HOME, JAVA_HOME and

CATALINA_HOME (NaiveSE_LiChenghao\Tomcat). If you have met some problems, you can refer to NaiveSE_LiChenghao\Tomcat\BUILDING.txt. After successfully start VIS, enter <http://localhost:8080/hello.htm> in Browser, and you can see website like this.



Enter the query and click “提交”, you can get 10 most relevant documents. If you want to search again, click” 返回” to return. You have to successfully set documents path to get this.

[返回](#)

[党建引领疫情防控??哈工大 \(深圳\) 以常态化疫情防控举措迎接全面复工复产 - 校园动态 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

新冠肺炎疫情爆发以来, 哈工大 (深圳) 根据党中央的要求, 在疫情期间坚持不懈抓好学习教育, 创新开展组织生活形式, 开展各类抗疫主题党日活动, 发挥党员模范带头作用, 将党建工作与学校教育教学中心工作紧密结合, 力...

[深圳市新型冠状病毒肺炎疫情防控指挥部办公室关于进一步调整疫情防控措施的通告 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

当前, 我市本地疫情基本平息, 全市各区均为低风险地区。但随着全球疫情的大流行, 境外输入风险防范压力极大, 做好疫情“外防输入、内防反弹”仍是当前工作重心。根据我市实际情况, 现就疫情防控措施调整通告如下: 一...

[习近平: 协同推进新冠肺炎防控科技攻关 为打赢疫情防控阻击战提供科技支撑 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

习近平在北京考察新冠肺炎防控科技攻关工作时强调协同推进新冠肺炎防控科技攻关? 为打赢疫情防控阻击战提供科技支撑我国是一个有着14亿多人口的大国, 防范化解重大疫情和重大突发公共卫生风险, 始终是我们须臾不...

[习近平主持中共中央政治局会议 研究新冠肺炎疫情防控工作 部署统筹做好疫情防控和经济社会发展工作 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

■ 在以习近平同志为核心的党中央坚强领导下, 经过全党全军全国各族人民团结奋战, 目前疫情蔓延势头得到初步遏制, 防控工作取得阶段性成效, 全国新增确诊病例数和疑似病例数总体呈下降趋势, 治愈出院人数较快增长, ...

[全面提高依法防控依法治理能力 为疫情防控提供有力法治保障 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

习近平主持召开中央全面依法治国委员会第三次会议强调全面提高依法防控依法治理能力为疫情防控提供有力法治保障李克强栗战书王沪宁出席■要在党中央集中统一领导下, 始终把人民群众生命安全和身体健康放在第一位, ...

[疫情防控, 哈工大 \(深圳\) 在行动 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

【哈工大 (深圳) 宣】(常溪/文、图)新型冠状病毒感染疫情引发公众的高度关注, 为全面贯彻落实党中央国务院和省委省政府的部署要求, 哈工大 (深圳) 全力做好疫情防控工作。1月23日, 哈工大 (深圳) 常务副校长葛...

[习近平主持召开中共中央政治局常务委员会会议 分析国内外新冠肺炎疫情防控形势 研究部署完善常态化疫情防控举措 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

新华社北京4月29日电?中共中央政治局常务委员会4月29日召开会议, 分析国内外新冠肺炎疫情防控形势, 研究部署完善常态化疫情防控举措, 研究确定支持湖北富经济社会发展一揽子政策。中共中央总书记习近平主持会...

[统筹推进疫情防控与学校发展 应对疫情工作领导小组会议召开 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

哈工大报讯 (刘培善/文) 2月27日, 应对疫情工作领导小组会议在行政楼626会议室召开。会议传达学习中央最新文件精神和相关要求, 听取各单位对照检查落实情况汇报, 并分析当前疫情形势, 研究部署近期防控重点工...

If you want to rebuild the whole project, you have to compile the java files in NaiveSE_LiChenghao\Code to get class files and copy class files to NaiveSE_LiChenghao\Tomcat\webapps\ROOT\WEB-INF\classes. Of course, we have to use relevant jar packages to perform which needs to be copied to NaiveSE_LiChenghao\Tomcat\webapps\ROOT\WEB-INF\bin.

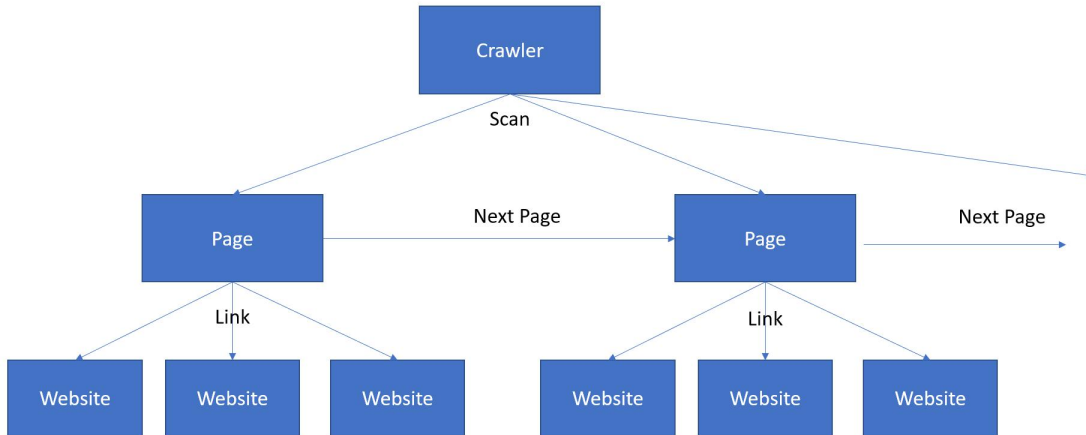
Concerning crawlers and indexing, you have to pip install Request, and do the crawling to get documents which are then used to create indexes by Lucene.

3. Structure

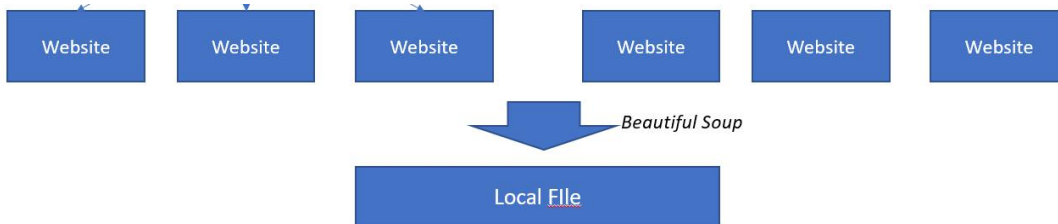
a. Crawler

According to the structure of the website, we use linear crawler structure to scan the website page by page. For each page, we go deep into the website to get the title, url and content of the website. And then, we separately store this information for the benefit of indexing. For politeness, we set time delay when crawling.

Crawler



The crawler is built based on Requests, Python Library. Requests help get the content as it is represented without modified. To help analyze the website, we use another library called BeautifulSoup to get constructed information in HTML file of the website. And then, we collect these information, and store them as local files.



b. Indexing

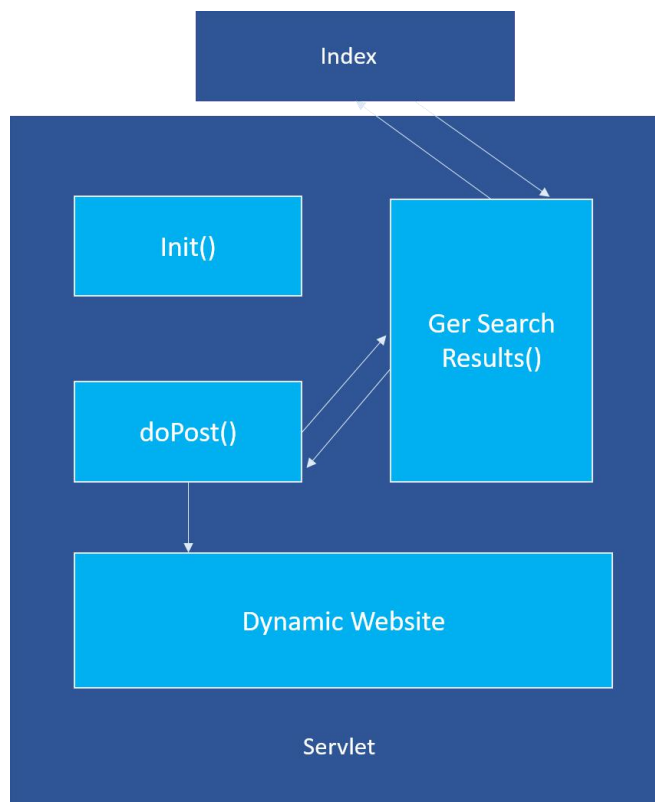
As we have already get local files from websites, we use Lucene to perform indexing and IKAnalyzer to perform Chinese split. First, we create directory which stores index and create an Object of IndexWrite based on Object Directory. And then, we read files in disk and create file objects for each of them to get IDs, urls, titles and contents. Finally, we create fields for each of them and write them in the index files. Lucene will help us do the indexing with contents we give with assistance of IKAnalyzer.

c. VIS

To create dynamic website which displays contents based on users' input, we use *Servlet* to communicate between html websites and JAVA programs. *Servlet* is a "class" which operates on the server applied for JAVA. In this project, we use *Servlet* to send and receive data from Web Server through HTTP.

When we receive request from other application to start the server, servlet starts with function *init()*. If the application sends a request, the servlet will operate function *doGet()* and *doPost()* based on the way applications send data (we choose *doPost()*). In this project, when the servlet receives the request, it will immediately analyze the query in Lucene. When Lucene finished querying, servlet will construct html-style string to construct html website shown to the user.

What needs to be mentioned is that this process still needs the operation of Lucene to do querying and *IKAnalyzer* to split the Chinese query. The main process is hidden behind the servlet communication. The user firstly face a static website called *hello.htm* and then when user inputs the query and click the button, the website leads the user to another website created by the *servlet*.



4. Evaluation

a. Precision@5

In this section, we will calculate the precision@5 to evaluate the performance of returned websites provided by the *Naïve SE*. We randomly selected 20 different queries

related to HITSZ and manually defines the relevance between the websites and the queries. Due to limits of space, we won't list all the results of returned websites. Instead, we only list queries and precision count.

Queries	Precision Count	P.S
党建	5	
社团活动	4	
奖学金	4	only 4 returns
化学	1	linked to "自动化学院"
疫情防控	5	
百年校庆	5	
社会实践	5	
新年晚会	4	only 4 returns
话剧	2	only 2 returns
论文发表	5	
副书记	5	
计算机科学与技术学院	5	
暑期实践	5	
讲座	5	
心理健康	5	
开学	4	
考试安排	5	
如何去哈工大深圳	0	
哈工大深圳教学好吗?	2	
哈工大深圳校历	0	

Average Precision = 3.8 per 5 top returned web

b. Response time

This project has no response time calculator because it doesn't actually reflect the response time. The in-program calculator only knows how much time it spends on processing data, but doesn't know the response time within response time to VIS and so on. So we use chrome to calculate the ultimate response time beginning when the button is pressed and ending when the whole related websites are displayed. We use 5 queries to perform so.

Queries	Response time(ms)
化学	18
话剧	8
社团活动	9

哈工大深圳	12
计算机科学与技术学院	18
生物科学	14

5. Running example

哈工大（深圳）讯息搜寻系统

输入您想搜寻的内容

- Start

哈工大（深圳）讯息搜寻系统

输入您想搜寻的内容

- Enter query
- After clicked “提交”

[返回](#)

[安实任哈工大党委常务副书记，吴松全任哈工大党委副书记、副校长，甄良任哈工大党委常委、副校长，孙雪任哈工大党委常委 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

哈工大报讯 (吉星/文) 日前，工业和信息化部党组决定，安实同志任哈尔滨工业大学党委常务副书记（正局级），吴松全同志任哈尔滨工业大学党委副书记、副校长，甄良同志任哈尔滨工业大学党委常委、副校长，孙雪同志任...

[哈工大 \(深圳\) 举办《初心的力量》哈工大“八百壮士”先进事迹展 - 综合新闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

【哈工大 (深圳) 宣】(向碧霞/文、图) 10月23日，《初心的力量》哈工大“八百壮士”先进事迹展在哈工大 (深圳) T3教学楼大厅展出。该展览展出了部分哈工大“八百壮士”的事迹，在全体师生中弘扬伟大的爱国主...

[太空返回-哈工大一校区三区土壤在哈工大 \(深圳\) 培育“问天树” - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

【哈工大 (深圳) 宣】(常溪/文 学校办公室/图) 躬耕一百年，育才新世纪。7月16日，太空返回?哈工大一校区三区土壤培育“问天树”深圳校区仪式在校训石旁举行。“问天树”共有三株，分别种植在哈尔滨、威海和深...

[哈工大党委常委会专题学习习近平总书记致哈工大建校100周年的贺信 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

哈工大报讯 (商艳凯/文) 6月7日晚，哈工大党委常委会在行政楼G26会议室就习近平总书记致哈尔滨工业大学建校100周年的贺信进行专题学习和集中研讨，研究部署全校下一步学习宣传和贯彻落实总书记贺信精神。哈...

[吴德林、甄良在哈工大 \(深圳\) 首届本科毕业生代表座谈会上深情寄语：牢记哈工大人的责任与担当 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

【哈工大 (深圳) 宣】(向碧霞/文、图) “感恩”“难忘”“成长”“不舍”.....在6月18日举行的哈工大 (深圳) 首届本科毕业生代表座谈会上，这些高频词表达了毕业生的共同心声。离别之际，毕业生畅叙本科4年的成...

[情系母校！哈工大土木工程深圳校友会向哈工大教育发展基金会捐赠50万元 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

【哈工大 (深圳) 宣】(谢梁晖/文、图) 殷殷母校情，浓浓校友心。6月3日，哈工大教育发展基金会捐赠纪念奖牌颁发仪式在深圳校区举行，哈工大土木工程深圳校友会向哈工大教育发展基金会捐赠50万元，用于学校校园...

[哈工大威海：第六届哈工大一校区三地土木工程学科建设研讨会在威海校区召开 - 校园动态 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

10月12日上午，由威海校区承办的第六届哈工大一校区三地土木工程学科建设研讨会在主楼二号会议室召开，副校长张文从出席会议并致辞，教务处姜永远处长、科技处徐龙军处长，哈工大土木学院吕大刚副院长、王玉银副院长...

[【纪念五四运动100周年】土木与环境工程学院17级土木工程硕士团支部开展“学习哈工大先贤，传承哈工大精神”主题活动 - 校园动态 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

哈工大 (深圳) 土木与环境工程学院宣 (刘美娜/文? 刘美娜、冯芝文/图) 为纪念五四运动100周年，培养同学们的爱国情操、发扬青年的奋斗精神，17级土木工程硕士团支部部分代表在T5305开展了“学习哈工大...

[哈工大纪念建校100周年直播预告 - 校区要闻 - 新闻中心 - 哈尔滨工业大学 \(深圳\)](#)

“ 一百年砥砺前行，一百年岁月如歌。2020年6月7日，哈尔滨工业大学将迎来建校百年纪念日。不管你我身在何处，定将有同

- Enter the hyperlink
- Or clicked “返回” button

哈工大 (深圳) 讯息搜寻系统

输入您想搜寻的内容

6. Project Catalogue

- Code: Python and JAVA codes and jar packages.
- Documents: local resources including indexes, titles, urls, contents.
- Tomcat: VIS environment
- ppt.pptx: Structure of the whole project.
- Readme.md : Brief introduction.
- Report.docx: Current file.

7. Attention

This project is private and doesn't allow any kinds of sharing.