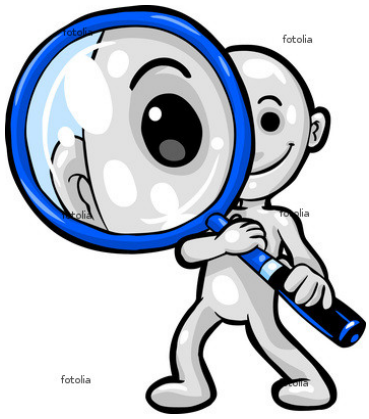
README for JoySearcher

A small project for tasting cloud computing



Introduction

How to clean the data?

How to build the index?

How to search with keyword?

The project— JoySearcher— is a keywords based Information Retrieval tools with hadoop. It is also a standard hadoop program.

In the following sections, I will give you a brief instruction of running such a light-weight internet keyword based Information Retrieval tools—JoySearcher— in pseudo distributed mode.

This project contains the source code of JoyCrawler 0.1.1. Now it has been updated to JoyCrawler 0.2(http://code.google.com/p/joycrawler/)

Getting Started

Prerequisites: JRE 6, Cygwin if Windows (PATH variable must be configured), Hadoop 0.20.1 above, (Hadoop-0.19.* may not work).

1. (Optional) Before running JoySearcher, please specify one parameter in configuration file conf/Joycrawler-site.xml:

```
<property> <name>org.joy.crawler.regEx</name>
<value>http://.*YOUR_DOMAIN_HERE.*</value> </property>
```

This property will set the host/website domain you want to crawl on. NOTE: any regular expression is acceptable for this property

The default set is "http://(\w+\.)*comp.nus.edu.sg/.*"It means only the web pages come from the website of School of computing can be crawled.

This step is optional; you *DO NOT* need to setup it.

2. (Optional) Add URL seeds in "seeds.txt" in program's folder, one URL per line.

The default set is "www.comp.nus.edu.sg" .It means the crawler will crawl the page starting from the homepage of SoC.

This step is optional; you DONOT need to setup it.

3. You MUST

1) Copy nekohtml-customized.jar, xercesImpl.jar, xercesMinimal.jar, and xml-apis.jar to hadoop's lib folder. This four files can be found in *JoySearcher/Resource/hadoop-lib*

2) Copy Joycrawler-site.xml in hadoop's conf folder; This file can be found in *JoySearcher/Resource/hadoop-conf*

3) Copy JoySearcher.jar SearcherDriverCall.jar and seeds.txt to hadoop's top-level folder. Those three files can be found in *JoySearcher/Resource/hadoop-root.*

Then start your hadoop (restart if already running). All this file can be found in the folder : *JoySearcher/Resource*. The figure 1 show the director of them.

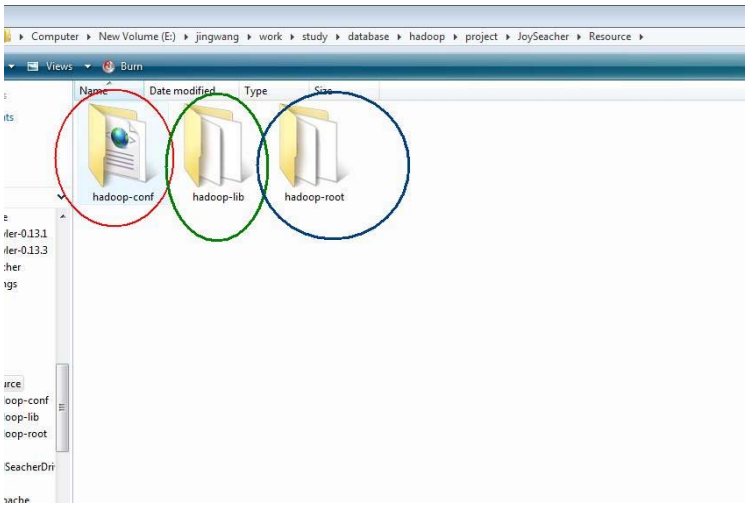


Figure 1

Make JoySearcher Work

1. **Prepare the data**
   Run in command line:

bin/hadoop jar SearcherDriverCall.jar CallSearcherDriver.SearcherDriverCall

As show in figure 2

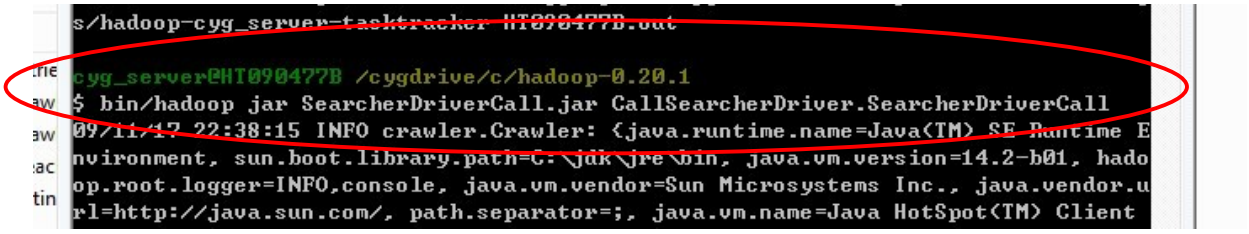


Figure 2

This process may be long. You should have patience. The whole process includes crawling the page (downloading, paring, merging, optimizing, filtering, etc), cleaning the data and building the index. It costs 6-9mins. When the process ends, you can see the hints as Figure 3.

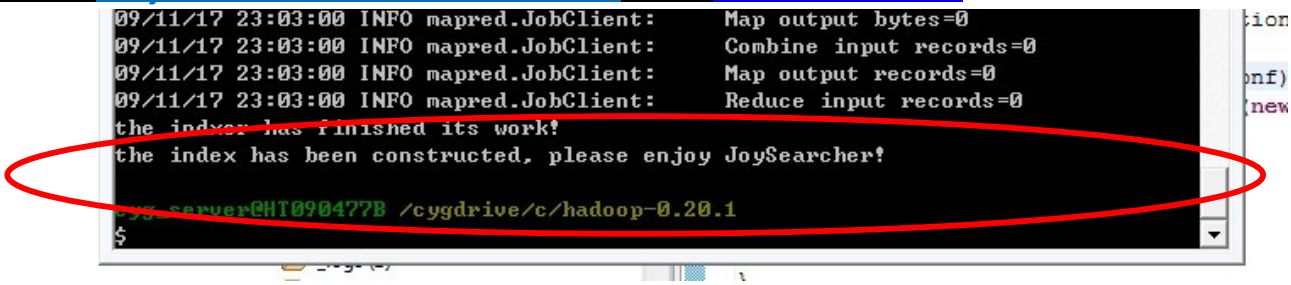


Figure 3

2. **Search with keywords**

Run in command line:

bin/hadoop jar JoySeacher.jar SearcherDriver.JoySearcher [–full/-regex][keywords]

the*[–full/-regex]* represents the recognize pattern. If you input *–full*, the JoySearcher will search the keywords exactly. If you input *–regex* , the JoySearcher will search the keywords fuzzily.

For *][keywords*], you can input one or more keywords.

The result is in *the [root]/ crawler/SearchResult.* There will be a file, named "part-r-0000" (maybe other names). After opening it, you can get the result.

3. **Example**

Run in command line as show in figure 4:

bin/hadoop jar JoySearcher.jar SearcherDriver.JoySearcher -full 2006

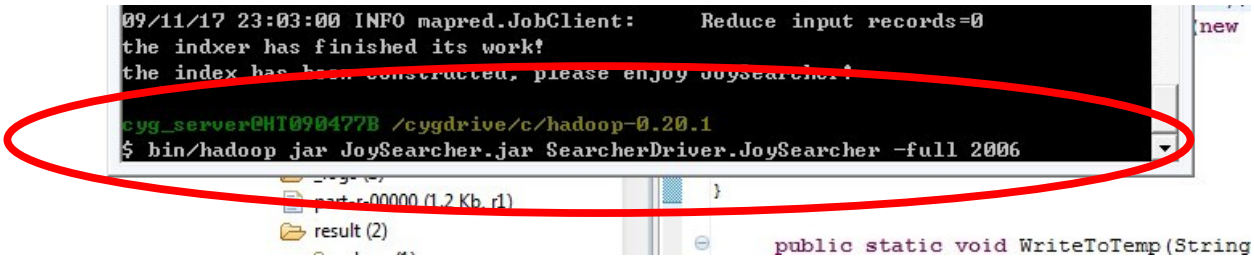


Figure 4

After show that the "*JoySearcher has finished its works",* Run in command line as show in figure 5:

*bin/hadoop fs –cat crawler/searchResult/**

Figure 5 show command line and the result

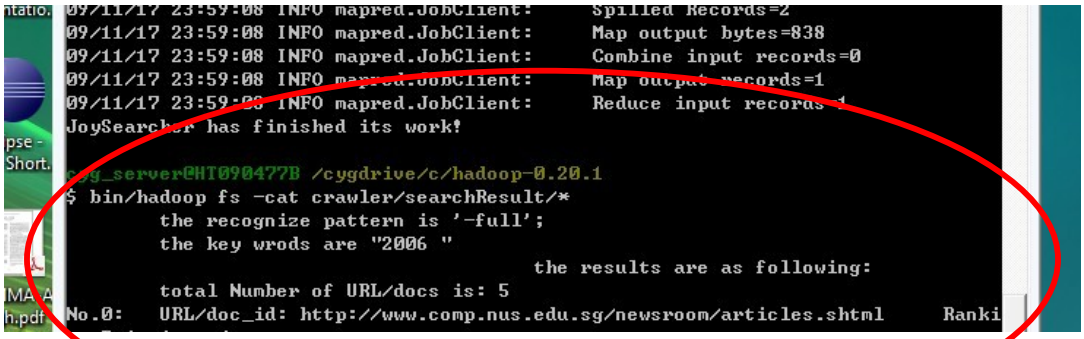


Figure 5

Contact

Zhou Jingbo; Email: letbefool@gmail.com