

README for Joycrawler



A spider program for hadoop applications

<http://joycrawler.googlecode.com>

Introduction

The project Joycrawler is a spider program for hadoop applications. It is also a standard hadoop program. Like the most popular spiders, it downloads the internet pages on given websites, hosts or even WWW. Meanwhile, a pagerank calculation program is integrated and this may very helpful for some web analysis applications.

In the following sections, we will give you a brief instruction of running such a light-weight internet web spider in both stand-alone and pseudo distributed mode.

Getting Started

Stand-Alone

Prerequisites: JRE 5, Apache Ant, Cygwin if Windows, Linux not required.

1. Download latest Joycrawler release on <http://joycrawler.googlecode.com/>, and unpack it.
2. Before running Joycrawler, please specify one parameter in configuration file in **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 you want to crawl on.

NOTE: any regular expression is acceptable for this property

3. Add URL seeds in "seeds.txt" in program's folder, one URL per line.
4. Run in command line: "ant run", and you may see the following output:

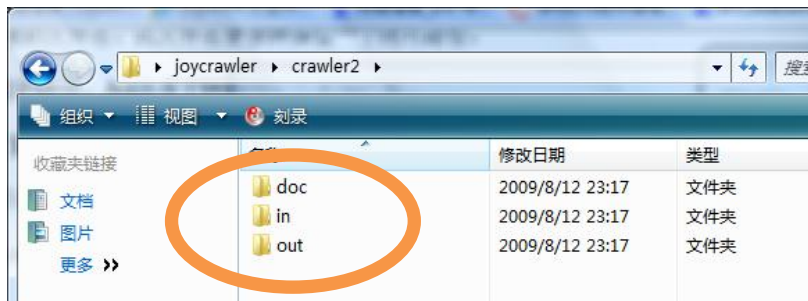
```

C:\cygdrive\c\Users\Administrator\Desktop\joycrawler
[Java] parsing http://youth.suda.edu.cn/fawenvview.aspx?id=520
[Java] parsing http://youth.suda.edu.cn/fawenvview.aspx?id=510
[Java] parsing http://youth.suda.edu.cn/fawenvview.aspx?id=529
[Java] parsing http://youth.suda.edu.cn/fawenvview.aspx?id=514
[Java] parsing http://youth.suda.edu.cn/fawenvview.aspx?id=516
[Java]
[Java] start filtering...
[Java] start downloading...
[Java] downloading http://youth.suda.edu.cn/newsview.aspx?Article_id=6805
[Java] succeed http://youth.suda.edu.cn/newsview.aspx?Article_id=6805
230k
[Java]
[Java] start parsing...
[Java] parsing http://youth.suda.edu.cn/newsview.aspx?Article_id=6805
[Java]
[Java] start filtering...
[Java] Crawling finished!
[Java]
[Java] start merging...
[Java] MergeDoc Done!
[Java] MergeLinks Done!
[Java] Normalizing...
[Java] Start Ranking...
[Java] Round0

```

After fetching all the pages or we reached the 15th iteration of fetching, the program will exit.

5. You can check the “crawler” folder for downloaded documents and links.



“Doc” folder is your webpage repository, and “out” folder contains all the link structure of the website you crawl. Both *doc* and *out* folders are standard hadoop output format. You can use SequenceFileInputFormat to read them into your application.

The Doc folder’s data pair is <Text, DocumentWritable>.

The Out folder’s data pair is <Text, OutlinksWritable>.

6. Pagerank result is in the *Ranking* folder, which is a plain text file, containing all the URL’s rank score and links structure one record per line.

```

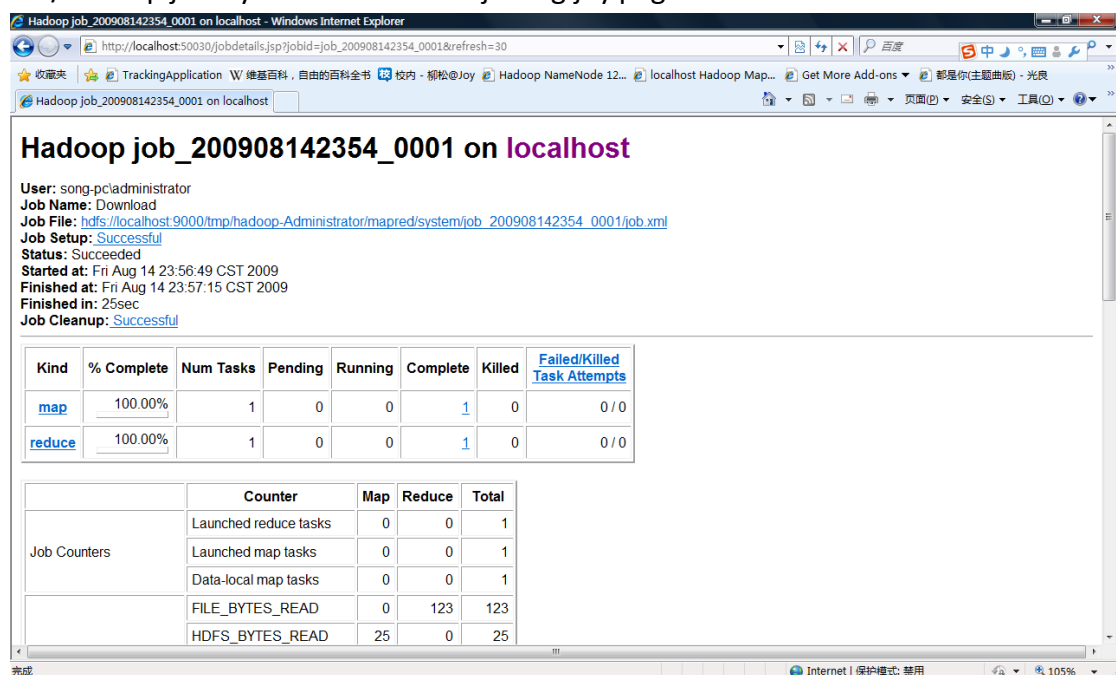
http://youth.suda.edu.cn/ 1.0966473698539565 http://youth.suda.edu.cn/tw_yxfc.
http://youth.suda.edu.cn/?page=0 1.3888358641909522 http://youth.suda.edu.cn/
http://youth.suda.edu.cn/download.aspx 0.8080326038099659 http://youth.suda.edu
http://youth.suda.edu.cn/fawenvview.aspx?id=510 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=511 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=512 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=513 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=514 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=515 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=516 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=517 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=518 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=519 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=520 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=521 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=522 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=523 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=524 0.5042575232603769
http://youth.suda.edu.cn/fawenvview.aspx?id=525 0.5042575232603769 http://youth.

```

Pseudo Distributed Mode

Before launching the Joycrawler in distributed mode, you must copy lib/nekohtml-customized.jar, lib/xercesImpl.jar, lib/xercesMinimal.jar, lib/xml-apis.jar to hadoop's lib folder, and conf/Joycrawler-site.xml in hadoop's conf folder, and then start your hadoop (restart if already running).

1. Copy Joycrawler-*VERSION*.jar and seeds.txt to hadoop's top-level folder.
2. Run in command line:
bin/hadoop jar Joycrawler-*VERSION*.jar org.joy.crawler.Crawler seeds.txt 15
3. Run in command line (pagerank computing):
bin/hadoop jar Joycrawler-*VERSION*.jar org.joy.pagerank.RankDriver



Contact

Song Liu, School of Computer Science, Soochow University, China

Email: (lamfeeling@126.com)

MSN: lamfeeling@126.com