Tech blog

LABELS

aicte (3)
aieee (2)
aiims (1)
aipvt (1)
apdept (1)
appsc (9)
bankclerks (28)
blackberry (3)
cardio (1)
cdac (1)
cinema (4)
compitetiveexams (7)
computer (1)
cricket (4)
CSE (1)
dotnet (2)
dotnet code snippets (1)
dsc (1)
eamcet (5)
education (267)
ekadasi (1)
Engineering (2)
Epass (2)
Facinating facts (4)
govtjobs (94)
greetings (30)
hallticketts (1)
hardware (1)
holi (1)
home (54)
icet (4)
ignca (1)
iim (1)
INDEPENDENCE DAY (1)
IT (1)
javaprojects (4)
kavali (4)
majorprojects (11)
MCA (1)
miniprojects (11)
mobiles (17)
n (1)
news (288)
newyeargreetings (10)
nokia (3)
ntpc (1)
Nutch (1)
pan (1)
phpprojects (1)
placements (80)
projects (18)

CRAWLING THE WEBSITE WITH NUTCH AND INTEGRATING WITH APACHE SOLR

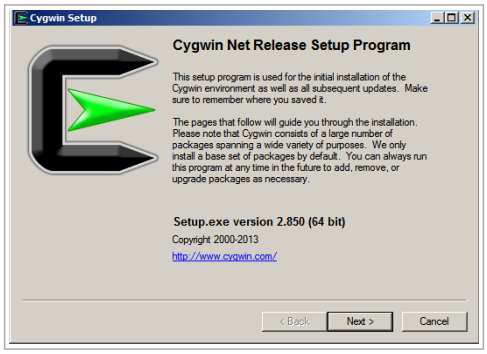
18 July 2014

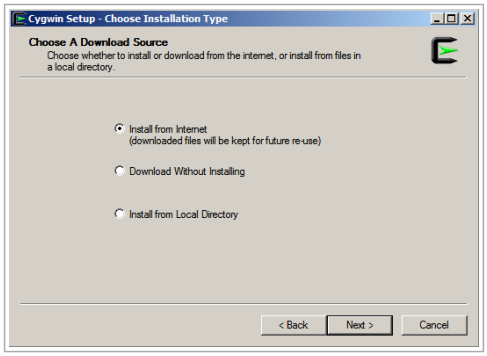
Crawling the Website with Nutch and Integrating with Apache Solr

If you want run the NUTCH commands then you have download the "CYGWIN" from the following url "**https://www.cygwin.com/install.html**". And you need to follow the steps to install the Cygwin.

Steps to Install Cygwin : -

a. Download the Cygwin
b. Double click on the installer





SUBSCRIBE VIA EMAIL

Enter your email address:

Subscribe

Delivered by FeedBurner

286 listeners
BY FEEDBURNER

Subscribe in a reader
Career & Job Blogs - BlogCatalog Blog Directory

FOLLOW ME

Career & Job Blogs - BlogCatalog Blog Directory

PAGES

Home
Exam details
About me
IT careers
Recruitments
Hot Results

BLOGROLL

vurooz
iworlkavali
D B A HELP .....
bujjigadi site
AP TREASURY
jnturesultportal
jntu

MY BLOG LIST

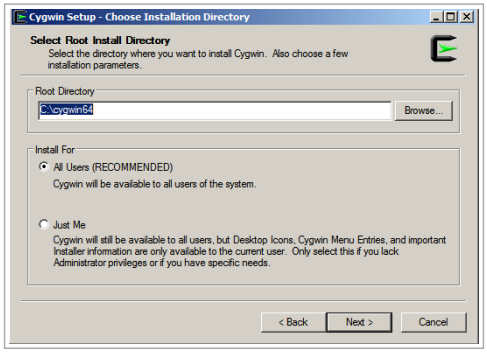
HD JNTU WORLD
-

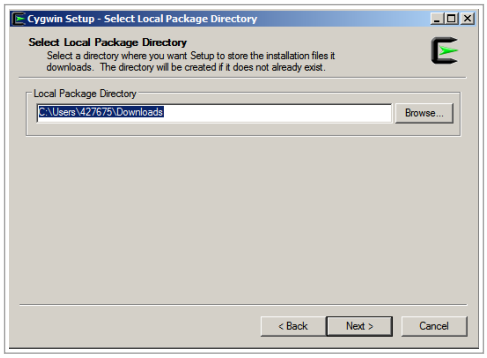
CODE SNIPPET

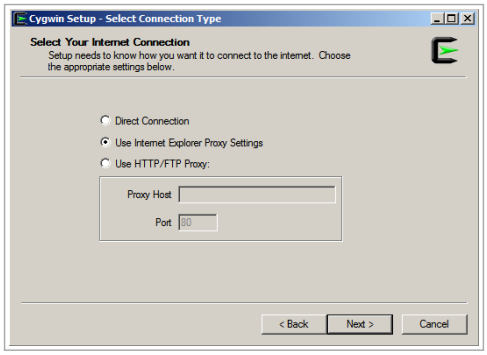
JAVA LAB PROGRAMS

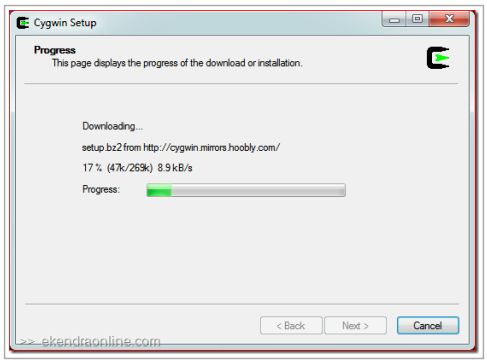
ARCHIVES

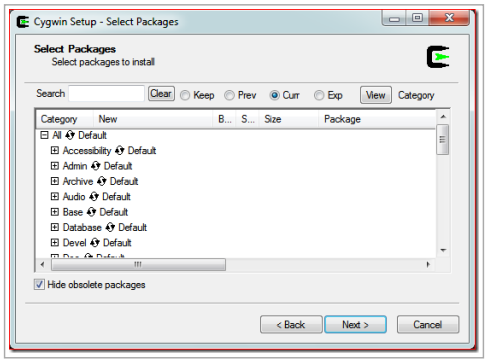
July (1)

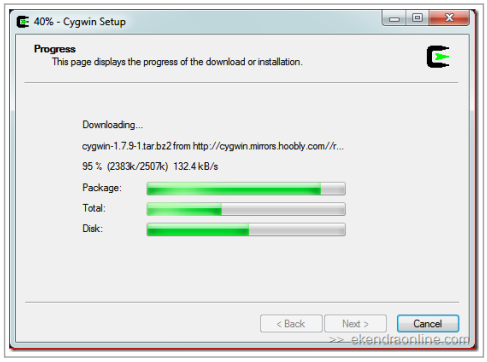


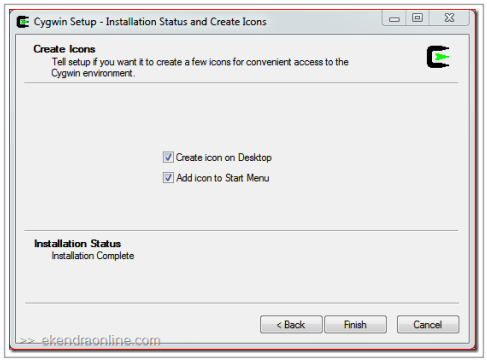






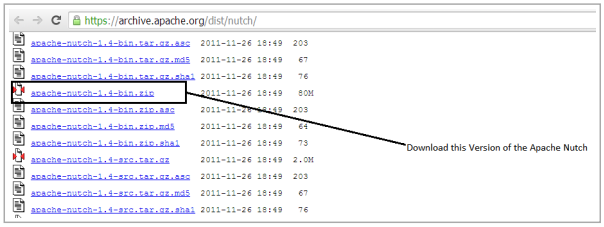




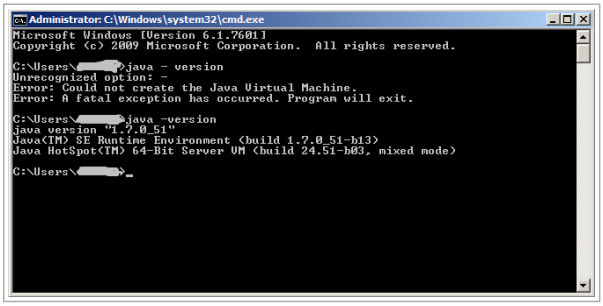


Steps to Install the "Apache Nutch" & "Apache Solr":

1. Download the "**apache-nutch-1.4-bin**" from the following url **https://archive.apache.org/dist/nutch/**
2. From the above url download the "apache-nutch-1.4-bin.zip".



3. First check the Java version and it should be greater than the "Version 1.7" and can be find by using this command with in the command prompt "**C:\>java –version**"

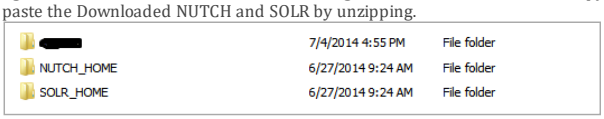


4. If it is greater than the "Version 1.7" it is ok or else if it is less than the "Version 1.7" then install the java version above 1.7 and after installing set the Environment variable. After check again and then proceed to download the Apache Solr.
5.

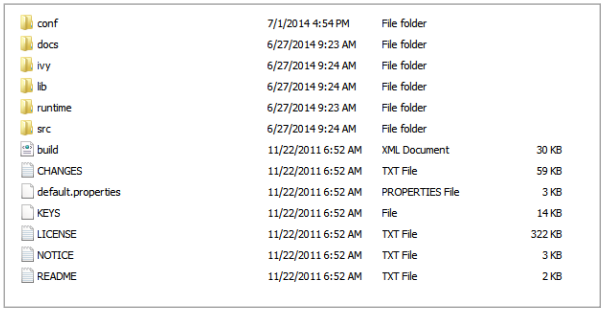
Sl.no	Type Of System	Format to be Downloaded
1	**Linux/Unix/OSX systems**	**.tgz**
2	**Microsoft Windows systems**	**.zip**

6. **Download** the **Apache solr from** this link "**http://lucene.apache.org/solr/**" or "**http://www.apache.org/dyn/closer.cgi/lucene/solr/4.8.1**"

Then go to the CYGWIN Installed folder and then go to the folder HOME in that copy and paste the Downloaded NUTCH and SOLR by unzipping.

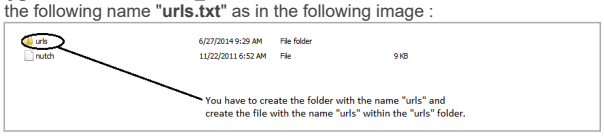


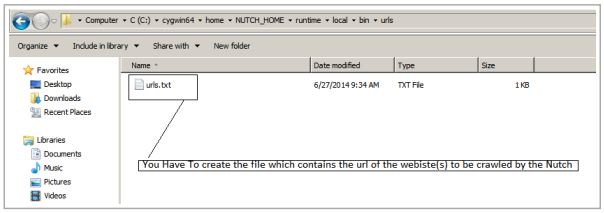
Create the folder "NUTCH_HOME" and copy all of the files as in the below:



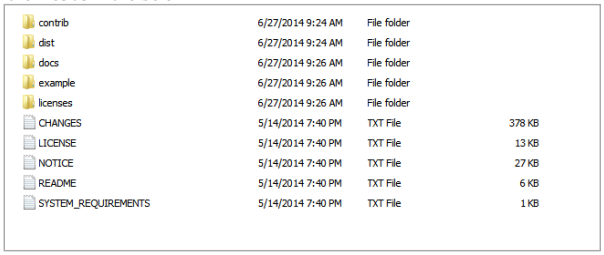
conf	7/1/2014 4:54 PM	File folder	
docs	6/27/2014 9:23 AM	File folder	
ivy	6/27/2014 9:24 AM	File folder	
lib	6/27/2014 9:24 AM	File folder	
runtime	6/27/2014 9:23 AM	File folder	
src	6/27/2014 9:24 AM	File folder	
build	11/22/2011 6:52 AM	XML Document	30 KB
CHANGES	11/22/2011 6:52 AM	TXT File	59 KB
default.properties	11/22/2011 6:52 AM	PROPERTIES File	3 KB
KEYS	11/22/2011 6:52 AM	File	14 KB
LICENSE	11/22/2011 6:52 AM	TXT File	322 KB
NOTICE	11/22/2011 6:52 AM	TXT File	3 KB
README	11/22/2011 6:52 AM	TXT File	2 KB

Create the Folder with name of "**urls**" in the following path "**C:\cygwin64\home\NUTCH_HOME\runtime\local\bin**" and also **create the file** with the following name "**urls.txt**" as in the following image :



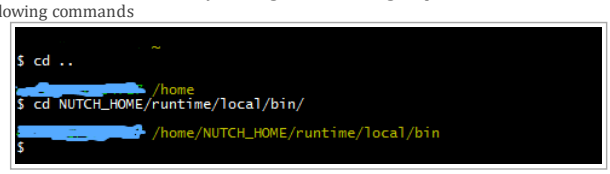


and in the same way we need to create the folder for the solr as "SOLR_HOME" and copy the files as in the below :

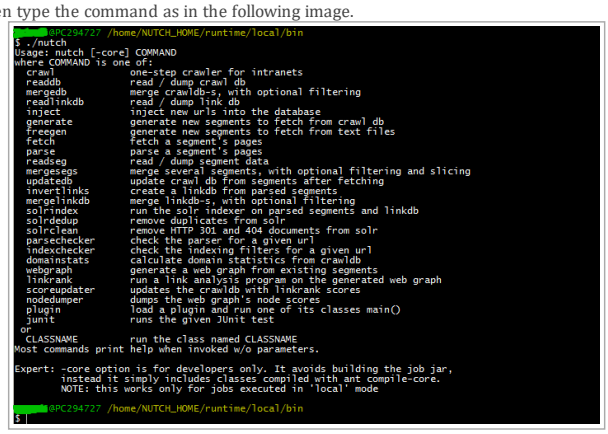


contrib	6/27/2014 9:24 AM	File folder	
dist	6/27/2014 9:24 AM	File folder	
docs	6/27/2014 9:26 AM	File folder	
example	6/27/2014 9:26 AM	File folder	
licenses	6/27/2014 9:26 AM	File folder	
CHANGES	5/14/2014 7:40 PM	TXT File	378 KB
LICENSE	5/14/2014 7:40 PM	TXT File	13 KB
NOTICE	5/14/2014 7:40 PM	TXT File	27 KB
README	5/14/2014 7:40 PM	TXT File	6 KB
SYSTEM_REQUIREMENTS	5/14/2014 7:40 PM	TXT File	1 KB

Then we can get the CYGWIN console shortcut on to the desktop and then double click on the CGYWIN shortcut then you will get the following output and also run the following commands



And if you want to check whether the Cygwin is able to run the command "NUTCH" then type the command as in the following image.



And do the following steps to run the Nutch Commands

1. Go to the folder where your "nutch" file is exists. And create the folder with the name "urls".
2. Create the file with in the folder name "urls" with the name "urls.txt" having the content like which site you need to crawl by using nutch command. For example, I need to crawl the "Geometrixx" website then I need to mention in the **urls.txt** is as follows
   http://localhost:4503/content/geometrixx-outdoors/en.html

First Add your "**agent name**" in the value field of the "**http.agent.name**" property in **conf/nutch-site.xml**, for example:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
    <name>http.agent.name</name>
    <value>Geometrixx Media Crawler</value>
    <property>
        <name>plugin.includes</name>
        <value>protocol-http|parse-(html|tika)|index-(basic|anchor)|indexer-solr</value>
    </property>
</configuration>
```

Then Run the following Command to crawl the website "**./nutch crawl urls -dir MyPaging -depth 3**"

Output:

```
427675@PC294727 /home/apache-nutch-1.4-bin/apache-nutch-1.4-bin/runtime/local/bin
$ ./nutch crawl urls -dir MyPaging -depth 3
cygpath: can't convert empty path
solrUrl is not set, indexing will be skipped...
crawl started in: MyPaging
rootUrlDir = urls
threads = 10
depth = 3
solrUrl=null
Injector: starting at 2014-06-10 17:15:13
Injector: crawlDb: MyPaging/crawldb
Injector: urlDir: urls
Injector: Converting injected urls to crawl db entries.
Injector: Merging injected urls into crawl db.
Injector: finished at 2014-06-10 17:15:22, elapsed: 00:00:08
Generator: starting at 2014-06-10 17:15:22
Generator: Selecting best-scoring urls due for fetch.
Generator: filtering: true
Generator: normalizing: true
Generator: jobtracker is 'local', generating exactly one partition.
Generator: Partitioning selected urls for politeness.
Generator: segment: MyPaging/segments/20140610171527
Generator: finished at 2014-06-10 17:15:28, elapsed: 00:00:06
Fetcher: Your 'http.agent.name' value should be listed first in 'http.robots.agents' property.
Fetcher: starting at 2014-06-10 17:15:28
Fetcher: segment: MyPaging/segments/20140610171527
Using queue mode : byHost
Fetcher: threads: 10
Fetcher: time-out divisor: 2
QueueFeeder finished: total 1 records + hit by time limit :0
Using queue mode : byHost ...
Fetcher: throughput threshold: -1
Fetcher: throughput threshold retries: 5
fetching http://localhost:4503/content/geometrixx-outdoors/en.html/
-finishing thread FetcherThread, activeThreads=9 ...
-finishing thread FetcherThread, activeThreads=1
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0...
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2014-06-10 17:15:36, elapsed: 00:00:08
ParseSegment: starting at 2014-06-10 17:15:36
ParseSegment: segment: MyPaging/segments/20140610171527
Parsing: http://localhost:4503/content/geometrixx-outdoors/en.html/
ParseSegment: finished at 2014-06-10 17:15:40, elapsed: 00:00:03
CrawlDb update: starting at 2014-06-10 17:15:40
CrawlDb update: db: MyPaging/crawldb
CrawlDb update: segments: [MyPaging/segments/20140610171527]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: true
CrawlDb update: URL filtering: true
CrawlDb update: 404 purging: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2014-06-10 17:15:41, elapsed: 00:00:01
Generator: starting at 2014-06-10 17:15:41
Generator: Selecting best-scoring urls due for fetch.
Generator: filtering: true
Generator: normalizing: true
Generator: jobtracker is 'local', generating exactly one partition.
Generator: Partitioning selected urls for politeness.
Generator: segment: MyPaging/segments/20140610171543
Generator: finished at 2014-06-10 17:15:45, elapsed: 00:00:03
Fetcher: Your 'http.agent.name' value should be listed first in 'http.robots.agents' property.
Fetcher: starting at 2014-06-10 17:15:45
Fetcher: segment: MyPaging/segments/20140610171543
Using queue mode : byHost
Fetcher: threads: 10
Fetcher: time-out divisor: 2
QueueFeeder finished: total 18 records + hit by time limit :0
Using queue mode : byHost
Using queue mode : byHost
fetching http://localhost:4503/content/geometrixx-outdoors/en/toolbar/about-us.html
Using queue mode : byHost...
Fetcher: throughput threshold: -1
Fetcher: throughput threshold retries: 5
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=17 ...
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en.html/
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=16 ...
```

```
fetching http://localhost:4503/content/geometrixx-
outdoors/en/equipment/hiking/interlaken-trek.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=15…
fetching http://localhost:4503/content/geometrixx-outdoors/en/equipment.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=14 …
fetch of http://localhost:4503/content/geometrixx-outdoors/en/equipment.html failed with:
java.net.SocketTimeoutException: Read timed out
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=14…
fetching http://localhost:4503/content/geometrixx-outdoors/en/support.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=13…
fetching http://localhost:4503/content/geometrixx-outdoors/en/toolbar/privacy-
policy.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=12 …
fetching http://localhost:4503/content/geometrixx-outdoors/en/user/cart.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=11 …
fetching http://localhost:4503/content/geometrixx-outdoors/en/community.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=10 …
fetching http://localhost:4503/content/geometrixx-outdoors/en/seasonal.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=9
fetching http://localhost:4503/content/geometrixx-
outdoors/en/equipment/hiking/cuzco.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=8
fetching http://localhost:4503/content/geometrixx-outdoors/en.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=7
fetching http://localhost:4503/content/geometrixx-outdoors/en/company.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=6
fetching http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-
blog.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=5
fetching http://localhost:4503/content/geometrixx-outdoors/en/women.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=4
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400860319
  now           = 1402400860687
  0. http://localhost:4503/content/geometrixx-outdoors/en/toolbar/terms-of-use.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  3. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400866180
  now           = 1402400861688
  0. http://localhost:4503/content/geometrixx-outdoors/en/toolbar/terms-of-use.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  3. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400866180
  now           = 1402400862688
  0. http://localhost:4503/content/geometrixx-outdoors/en/toolbar/terms-of-use.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  3. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400866180
  now           = 1402400863688
  0. http://localhost:4503/content/geometrixx-outdoors/en/toolbar/terms-of-use.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  3. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400866180
  now           = 1402400864688
  0. http://localhost:4503/content/geometrixx-outdoors/en/toolbar/terms-of-use.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  3. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400866180
```

```
  now         = 1402400865689
  0. http://localhost:4503/content/geometrixx-outdoors/en/toolbar/terms-of-use.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  3. http://localhost:4503/content/geometrixx-outdoors/en/men.html
fetching http://localhost:4503/content/geometrixx-outdoors/en/toolbar/terms-of-use.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400871237
  now           = 1402400866689
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400871237
  now           = 1402400867689
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400871237
  now           = 1402400868689
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400871237
  now           = 1402400869689
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400871237
  now           = 1402400870689
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking/nunavut-
fleece.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/men.html
fetching http://localhost:4503/content/geometrixx-
outdoors/en/equipment/hiking/nunavut-fleece.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400871237
  now           = 1402400871689
  0. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400876912
  now           = 1402400872689
  0. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400876912
  now           = 1402400873690
  0. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
```

```
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400876912
  now           = 1402400874690
  0. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400876912
  now           = 1402400875690
  0. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400876912
  now           = 1402400876690
  0. http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-nomad.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/men.html
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-
nomad.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400882200
  now           = 1402400877690
  0. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400882200
  now           = 1402400878691
  0. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400882200
  now           = 1402400879692
  0. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400882200
  now           = 1402400880692
  0. http://localhost:4503/content/geometrixx-outdoors/en/men.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402400882200
  now           = 1402400881692
  0. http://localhost:4503/content/geometrixx-outdoors/en/men.html
fetching http://localhost:4503/content/geometrixx-outdoors/en/men.html
-finishing thread FetcherThread, activeThreads=9 ...
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2014-06-10 17:18:04, elapsed: 00:02:19
ParseSegment: starting at 2014-06-10 17:18:04
ParseSegment: segment: MyPaging/segments/20140610171543
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en.html/
Parsing: http://localhost:4503/content/geometrixx-outdoors/en.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/community.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/company.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-
blog.html
Parsing: http://localhost:4503/content/geometrixx-
outdoors/en/equipment/hiking/cuzco.html
Parsing: http://localhost:4503/content/geometrixx-
outdoors/en/equipment/hiking/interlaken-trek.html
Parsing: http://localhost:4503/content/geometrixx-
outdoors/en/equipment/hiking/nunavut-fleece.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/pants/fulani-
nomad.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/toolbar/about-us.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/toolbar/privacy-
policy.html
```

```
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/toolbar/terms-of-use.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/user/cart.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women.html
ParseSegment: finished at 2014-06-10 17:18:06, elapsed: 00:00:01
CrawlDb update: starting at 2014-06-10 17:18:06
CrawlDb update: db: MyPaging/crawldb
CrawlDb update: segments: [MyPaging/segments/20140610171543]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: true
CrawlDb update: URL filtering: true
CrawlDb update: 404 purging: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2014-06-10 17:18:07, elapsed: 00:00:01
Generator: starting at 2014-06-10 17:18:07
Generator: Selecting best-scoring urls due for fetch.
Generator: filtering: true
Generator: normalizing: true
Generator: jobtracker is 'local', generating exactly one partition.
Generator: Partitioning selected urls for politeness.
Generator: segment: MyPaging/segments/20140610171809
Generator: finished at 2014-06-10 17:18:10, elapsed: 00:00:03
Fetcher: Your 'http.agent.name' value should be listed first in 'http.robots.agents' property.
Fetcher: starting at 2014-06-10 17:18:10
Fetcher: segment: MyPaging/segments/20140610171809
Using queue mode : byHost
Fetcher: threads: 10
Fetcher: time-out divisor: 2
Using queue mode : byHost
QueueFeeder finished: total 97 records + hit by time limit :0
Using queue mode : byHost
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/men/pants/fulani-
nomad.html
Using queue mode : byHost...
Fetcher: throughput threshold: -1
Fetcher: throughput threshold retries: 5
fetch of http://localhost:4503/content/geometrixx-outdoors-mobile/en/men/pants/fulani-
nomad.html failed with: Http code=500, url=http://localhost:4503/content/geometrixx-
outdoors-mobile/en/men/pants/fulani-nomad.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=96
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__if0q-i_would_liketoknow.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=95
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/shirts.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=94
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__wjda-do_the_blackcombglo.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=93
fetching http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=92
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/abidjan-water.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=91
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/toolbar/about-
us.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=90
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/lagos-mini-longboard.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=89
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/coats/edmonton-
winter.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=88
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/brazzaville.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=87
/nutch solrindex-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=87
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/apparel/scarves/sherbrooke-winter.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=86
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/men.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=85
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/tuareg-summer.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=84
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/pants.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=83
fetching http://localhost:4503/content/geometrixx-outdoors-
mobile/en/equipment/hiking/interlaken-trek.html
fetch of http://localhost:4503/content/geometrixx-outdoors-
mobile/en/equipment/hiking/interlaken-trek.html failed with: Http code=500,
url=http://localhost:4503/content/geometrixx-outdoors-
mobile/en/equipment/hiking/interlaken-trek.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=82
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__skq7-if_i_buy_thewhistle.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=81
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=80
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/user/cart.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=79
fetching http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-
blog/2011/11/layer_it_on.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=78
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/apparel/hats/baffin-snow.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=77
fetching http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-
blog/2011/12/summer_training.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=76
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/mont-tremblant.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=75
fetching http://localhost:4503/content/geometrixx-outdoors/en/company/our-story.html
```

```
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=74
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/tacna.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=73
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__0yic-what_do_i_doifine.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=72
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/saskatoon-parka.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=71
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__pley-is_the_lagosminilo.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=70
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/shirts/palau-
summer.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=69
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/mombassa-runners.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=68
fetching http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=67
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/interlaken-trek.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=66
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/shirts/maui-
marine.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=65
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/fernie-snow.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=64
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/shirts/tupai-
summer.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=63
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/pants.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=62
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/davos-trek.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=61
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/shirts/bambara-
cargo.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=60
fetching http://localhost:4503/content/geometrixx-outdoors/en/community/surfing.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=59
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/coats.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=58
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/nairobi-runners.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=57
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/equipment.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=56
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/kawartha-snow.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=55
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/coats.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=54
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/shorts/jola-
summer.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=53
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/shorts.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=52
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/shorts.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=51
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/marka-sport.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=50
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/shorts/tuareg-
summer.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=49
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/nunavut-fleece.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=48
fetching http://localhost:4503/content/geometrixx-outdoors-
mobile/en/equipment/hiking/cuzco.html
fetch of http://localhost:4503/content/geometrixx-outdoors-
mobile/en/equipment/hiking/cuzco.html failed with: Http code=500,
url=http://localhost:4503/content/geometrixx-outdoors-
mobile/en/equipment/hiking/cuzco.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=47
fetching http://localhost:4503/content/geometrixx-outdoors/en/community/running.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=46
fetching http://localhost:4503/content/geometrixx-outdoors/en/community/winter-
sports.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=45
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/apparel/scarves/halifax-winter.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=44
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/shorts/tahiti-
summer.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=43
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/blackcomb-snow.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=42
fetching http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=41
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__inai-i_would_am_intereste.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=40
fetching http://localhost:4503/content/geometrixx-outdoors-
mobile/en/equipment/hiking/nunavut-fleece.html
fetch of http://localhost:4503/content/geometrixx-outdoors-
mobile/en/equipment/hiking/nunavut-fleece.html failed with: Http code=500,
```

```
url=http://localhost:4503/content/geometrixx-outdoors-
mobile/en/equipment/hiking/nunavut-fleece.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=39
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/shirts/bora-
bora.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=38
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/whistler-snow.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=37
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/longirod-trek.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=36
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/bora-bora.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=35
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/calgary-winter.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=34
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/tobermory-snow.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=33
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/fiji-sport.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=32
fetching http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-
blog/2012/01/going_for_gold.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=31
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__ksko-i_am_havingtrouble.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=30
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/shirts.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=29
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/shirts/ashanti-
nomad.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=28
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=27
fetching http://localhost:4503/content/geometrixx-outdoors/en/company/the-team.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=26
fetching http://localhost:4503/content/geometrixx-outdoors/en/community/hiking.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=25
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/toolbar/terms-of-
use.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=24
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/toolbar/privacy-
policy.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=23
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/kamloops-snow.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=22
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/contact.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=21
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/edmonton-winter.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=20
fetching http://localhost:4503/content/geometrixx-outdoors/en/men/shorts/marka-
sport.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=19
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/apparel/hats/montevideo.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=18
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/shorts/fiji-
sport.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=17
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__qwio-is_there_a_waterproo.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=16
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/women.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=15
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/fulani-nomad.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=14
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/apparel/hats/montreal-snow.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=13
fetching http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-
blog/2012/02/yes_i_ski_like_agi.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=12
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/cajamara.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=11
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__zabn-i_would_liketoknow.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=10
fetching http://localhost:4503/content/geometrixx-outdoors/en/user/checkout.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=9
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/coats/saskatoon-
parka.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=8
fetching http://localhost:4503/content/geometrixx-
outdoors/en/support.topic.html/forum__6qqb-does_anyoneknowif.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=7
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/kelowna-snow.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=6
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/summer/equipment/cuzco.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=5
fetching http://localhost:4503/content/geometrixx-outdoors-mobile/en/seasonal.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=4
* queue: http://localhost
```

```
 maxThreads    = 1
 inProgress    = 1
 crawlDelay    = 5000
 minCrawlDelay = 0
 nextFetchTime = 1402401420994
 now           = 1402401421494
 0. http://localhost:4503/content/geometrixx-outdoors/en/women/coats/calgary-
winter.html
 1. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
 2. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
 3. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
 maxThreads    = 1
 inProgress    = 0
 crawlDelay    = 5000
 minCrawlDelay = 0
 nextFetchTime = 1402401427122
 now           = 1402401422494
 0. http://localhost:4503/content/geometrixx-outdoors/en/women/coats/calgary-
winter.html
 1. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
 2. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
 3. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
 maxThreads    = 1
 inProgress    = 0
 crawlDelay    = 5000
 minCrawlDelay = 0
 nextFetchTime = 1402401427122
 now           = 1402401423494
 0. http://localhost:4503/content/geometrixx-outdoors/en/women/coats/calgary-
winter.html
 1. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
 2. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
 3. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
 maxThreads    = 1
 inProgress    = 0
 crawlDelay    = 5000
 minCrawlDelay = 0
 nextFetchTime = 1402401427122
 now           = 1402401424494
 0. http://localhost:4503/content/geometrixx-outdoors/en/women/coats/calgary-
winter.html
 1. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
 2. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
 3. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
 maxThreads    = 1
 inProgress    = 0
 crawlDelay    = 5000
 minCrawlDelay = 0
 nextFetchTime = 1402401427122
 now           = 1402401425494
 0. http://localhost:4503/content/geometrixx-outdoors/en/women/coats/calgary-
winter.html
 1. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
 2. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
 3. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://localhost
 maxThreads    = 1
 inProgress    = 0
 crawlDelay    = 5000
 minCrawlDelay = 0
 nextFetchTime = 1402401427122
 now           = 1402401426494
 0. http://localhost:4503/content/geometrixx-outdoors/en/women/coats/calgary-
winter.html
 1. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
 2. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
 3. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/coats/calgary-
winter.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
 maxThreads    = 1
 inProgress    = 0
 crawlDelay    = 5000
 minCrawlDelay = 0
 nextFetchTime = 1402401432270
 now           = 1402401427494
 0. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
 1. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
```

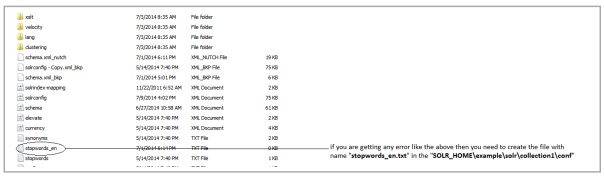
```
  2. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401432270
  now           = 1402401428495
  0. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
  1. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401432270
  now           = 1402401429495
  0. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
  1. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401432270
  now           = 1402401430495
  0. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
  1. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=3
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401432270
  now           = 1402401431495
  0. http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
  1. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  2. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
fetching http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-
fashion.html
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 1
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401432270
  now           = 1402401432495
  0. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401437504
  now           = 1402401433495
  0. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401437504
  now           = 1402401434495
  0. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401437504
  now           = 1402401435495
  0. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
```

```
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401437504
  now           = 1402401436495
  0. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401437504
  now           = 1402401437495
  0. http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
  1. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
fetching http://localhost:4503/content/geometrixx-
outdoors/en/seasonal/winter/equipment/banff-snow.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401442672
  now           = 1402401438496
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401442672
  now           = 1402401439496
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401442672
  now           = 1402401440496
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401442672
  now           = 1402401441496
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://localhost
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1402401442672
  now           = 1402401442496
  0. http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
fetching http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
-finishing thread FetcherThread, activeThreads=9
-finishing thread FetcherThread, activeThreads=7
-finishing thread FetcherThread, activeThreads=7
-finishing thread FetcherThread, activeThreads=5
-finishing thread FetcherThread, activeThreads=5
-finishing thread FetcherThread, activeThreads=4
-finishing thread FetcherThread, activeThreads=3
-finishing thread FetcherThread, activeThreads=2
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2014-06-10 17:27:27, elapsed: 00:09:16
ParseSegment: starting at 2014-06-10 17:27:27
ParseSegment: segment: MyPaging/segments/20140610171809
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en.html
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en/contact.html
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en/equipment.html
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en/men.html
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en/seasonal.html
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en/toolbar/about-
us.html
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en/toolbar/privacy-
policy.html
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en/toolbar/terms-of-
use.html
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en/user/cart.html
Parsing: http://localhost:4503/content/geometrixx-outdoors-mobile/en/women.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/community/hiking.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/community/running.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/community/surfing.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/community/winter-
sports.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/company/our-story.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/company/the-team.html
```

Parsing: http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-blog/2011/11/layer_it_on.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-blog/2011/12/summer_training.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-blog/2012/01/going_for_gold.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/company/unlimited-blog/2012/02/yes_i_ski_like_agi.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/equipment/hiking.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/equipment/skiing.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/coats.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/coats/edmonton-winter.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/pants.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/shirts.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/shirts/ashanti-nomad.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/shirts/bambara-cargo.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/shorts.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/shorts/jola-summer.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/shorts/marka-sport.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/men/shorts/tuareg-summer.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/apparel/hats/montevideo.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/abidjan-water.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/bora-bora.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/brazzaville.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/cajamara.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/cuzco.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/davos-trek.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/fiji-sport.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/fulani-nomad.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/interlaken-trek.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/lagos-mini-longboard.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/longirod-trek.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/marka-sport.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/mombassa-runners.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/nairobi-runners.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/nunavut-fleece.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/tacna.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/summer/equipment/tuareg-summer.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/apparel/hats/baffin-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/apparel/hats/montreal-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/apparel/scarves/halifax-winter.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/apparel/scarves/sherbrooke-winter.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/banff-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/blackcomb-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/calgary-winter.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/edmonton-winter.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/fernie-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/kamloops-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/kawartha-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/kelowna-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/mont-tremblant.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/saskatoon-parka.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/tobermory-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/seasonal/winter/equipment/whistler-snow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__0yic-what_do_i_doifine.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__6qqb-does_anyoneknowif.html

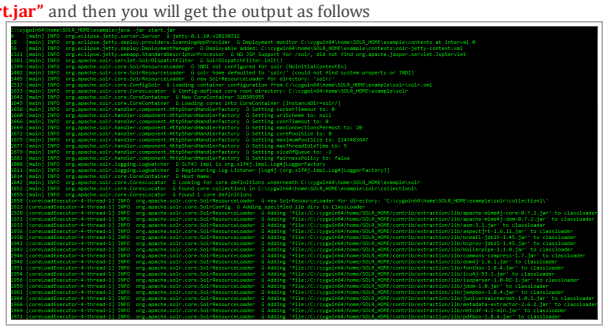
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__if0q-i_would_liketoknow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__inai-i_would_am_intereste.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__ksko-i_am_havingtrouble.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__pley-is_the_lagosminilo.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__qwio-is_there_a_waterproo.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__skq7-if_i_buy_thewhistle.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__wjda-do_the_blackcombglo.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/support.topic.html/forum__zabn-i_would_liketoknow.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/user/checkout.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/coats.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/coats/calgary-winter.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/coats/saskatoon-parka.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/pants.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/pants/tonga-fashion.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/shirts.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/shirts/bora-bora.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/shirts/maui-marine.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/shirts/palau-summer.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/shirts/tupai-summer.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/shorts.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/shorts/fiji-sport.html
Parsing: http://localhost:4503/content/geometrixx-outdoors/en/women/shorts/tahiti-summer.html
ParseSegment: finished at 2014-06-10 17:27:29, elapsed: 00:00:02
CrawlDb update: starting at 2014-06-10 17:27:29
CrawlDb update: db: MyPaging/crawldb
CrawlDb update: segments: [MyPaging/segments/20140610171809]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: true
CrawlDb update: URL filtering: true
CrawlDb update: 404 purging: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2014-06-10 17:27:35, elapsed: 00:00:05
LinkDb: starting at 2014-06-10 17:27:35
LinkDb: linkdb: MyPaging/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: file:/C:/cygwin64/home/apache-nutch-1.4-bin/apache-nutch-1.4-bin/runtime/local/bin/MyPaging/segments/20140610171527
LinkDb: adding segment: file:/C:/cygwin64/home/apache-nutch-1.4-bin/apache-nutch-1.4-bin/runtime/local/bin/MyPaging/segments/20140610171543
LinkDb: adding segment: file:/C:/cygwin64/home/apache-nutch-1.4-bin/apache-nutch-1.4-bin/runtime/local/bin/MyPaging/segments/20140610171809
LinkDb: finished at 2014-06-10 17:27:37, elapsed: 00:00:01
crawl finished: MyPaging

Then you have to create the file **C:\cygwin64\home\SOLR_HOME\example\solr\collection1\conf**

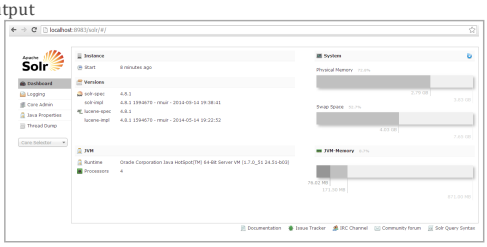


Steps To Start the "Apache Solr":

1. After Downloading and Extracting the Apache Solr then go to the command prompt and type the following one **"C:\cygwin64\home\SOLR_HOME\example>java –jar start.jar"** and then you will get the output as follows



2. Then open the browser and type this **http://localhost:8983/solr/** then you will the following output



If you want to integrate the Nutch data with Solr then just make the following changes in the Solr Folder

1. Edit the file "**solr-config.xml**" from the following path " **C:\cygwin64\home\SOLR_HOME\example\solr\collection1\conf**" with following code

```
<requestHandler name="/nutch" class="solr.SearchHandler" >
     <lst name="defaults">
          <str name="defType">dismax</str>
          <str name="echoParams">explicit</str>
          <str name="tie">0.01</str>
          <str name="qf">content^0.5 anchor^1.0 title^1.2</str>
          <str name="pf">content^0.5 anchor^1.5 title^1.2 site^1.5</str>
          <str name="fl">url</str>
          <str name="mm">2&lt;-1 5&lt;-2 6&lt;90%</str>
          <str name="ps">100</str>
          <str name="hl">true</str>
          <str name="q.alt">*:*</str>
          <str name="hl.fl">title url content</str>
          <str name="f.title.hl.fragsize">0</str>
          <str name="f.title.hl.alternateField">title</str>
          <str name="f.url.hl.fragsize">0</str>
          <str name="f.url.hl.alternateField">url</str>
          <str name="f.content.hl.fragmenter">regex</str>
     </lst>
</requestHandler>
```

2. Edit the file "schema.xml" from the same path as in the above "**C:\cygwin64\home\SOLR_HOME\example\solr\collection1\conf**" with the following code

```
<field name="_version_" type="long" indexed="true" stored="true"/>
<field name="_root_" type="string" indexed="true" stored="false"/>
<field name="id" type="string" indexed="true" stored="true" required="true" multiValued="false" />

<field name="sku" type="text_en_splitting_tight" indexed="true" stored="true" omitNorms="true"/>
<field name="name" type="text_general" indexed="true" stored="true"/>
<field name="manu" type="text_general" indexed="true" stored="true" omitNorms="true"/>
<field name="cat" type="string" indexed="true" stored="true" multiValued="true"/>
<field name="features" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="includes" type="text_general" indexed="true" stored="true" termVectors="true" termPositions="true" termOffsets="true" />
<field name="weight" type="float" indexed="true" stored="true"/>
<field name="price"  type="float" indexed="true" stored="true"/>
<field name="popularity" type="int" indexed="true" stored="true" />
<field name="inStock" type="boolean" indexed="true" stored="true" />
<field name="store" type="location" indexed="true" stored="true"/>
<field name="title" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="subject" type="text_general" indexed="true" stored="true"/>
<field name="description" type="text_general" indexed="true" stored="true"/>
<field name="comments" type="text_general" indexed="true" stored="true"/>
<field name="author" type="text_general" indexed="true" stored="true"/>
<field name="keywords" type="text_general" indexed="true" stored="true"/>
<field name="category" type="text_general" indexed="true" stored="true"/>
<field name="resourcename" type="text_general" indexed="true" stored="true"/>
<field name="url" type="text_general" indexed="true" stored="true"/>
<field name="content_type" type="string" indexed="true" stored="true" multiValued="true"/>
<field name="last_modified" type="date" indexed="true" stored="true"/>
<field name="links" type="string" indexed="true" stored="true" multiValued="true"/>
<field name="content" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="text" type="text_general" indexed="true" stored="false" multiValued="true"/>
<field name="text_rev" type="text_general_rev" indexed="true" stored="false" multiValued="true"/>
<field name="manu_exact" type="string" indexed="true" stored="false"/>
<field name="payloads" type="payloads" indexed="true" stored="true"/>
```

```
<field name="site" type="string" stored="false" indexed="true"/>
<field name="host" type="string" stored="false" indexed="true"/>
```

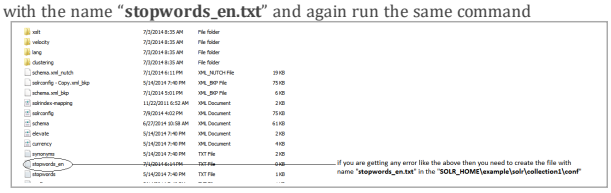
```
<field name="boost" type="string" stored="false" indexed="true"/>
<field name="digest" type="string" stored="false" indexed="true"/>
<field name="tstamp" type="string" stored="false" indexed="true"/>
```

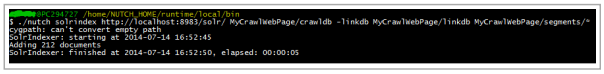
3. And restart the system if at all you are getting the error then run the following command for indexing the nutch data into the solr "**./nutch solrindex http://localhost:8983/solr/ ./MyPaging/crawldb -linkdb ./MyPaging/linkdb ./MyPaging/segments/***"

Output:

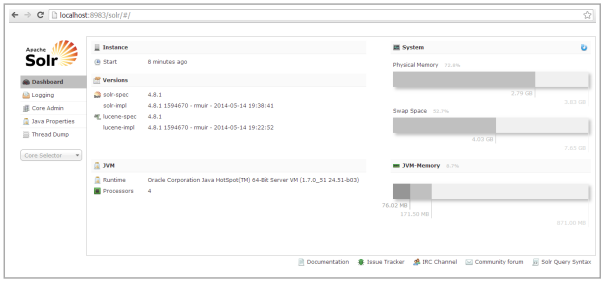
427675@PC294727 /home/apache-nutch-1.4-bin/apache-nutch-1.4-bin/runtime/local/bin
$ ./nutch solrindex http://localhost:8983/solr/ ./MyPaging/crawldb -linkdb ./MyPaging/linkdb ./MyPaging/segments/*
cygpath: can't convert empty path
SolrIndexer: starting at 2014-06-11 09:32:48
Adding 335 documents
java.io.IOException: Job failed!

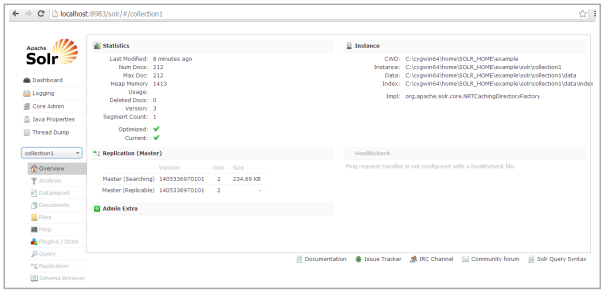
If at all you get any error like this then go to the following path "**C:\cygwin64\home\SOLR_HOME\example\solr\collection1\conf"** and create the file with the name "**stopwords_en.txt**" and again run the same command

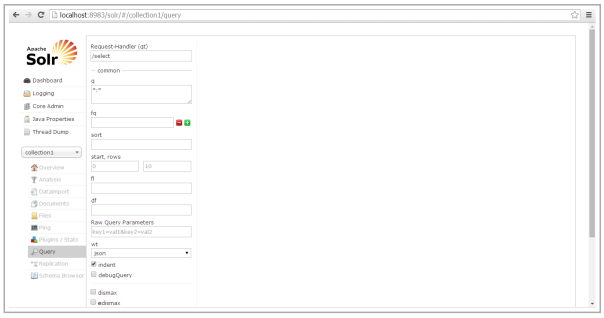


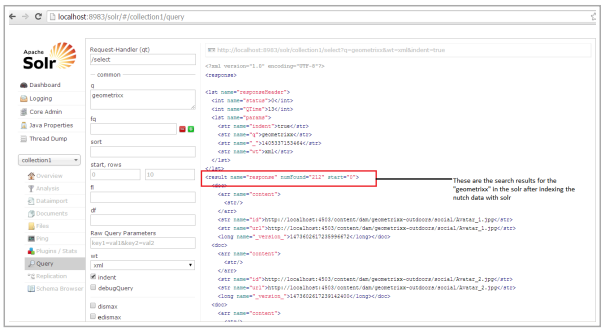


Then go to the url http://localhost:8983/solr









Related Posts : Nutch

Posted by bujjigadu

Labels: Nutch

1 COMMENTS:

bujjigadu said... **July 21, 2014 at 5:44 PM**

NSeq = (Sequence_No < 10) ? ("00000"+NSeq) : ((Sequence_No < 100) ? ("0000"+NSeq) : ((Sequence_No < 1000) ? ("000"+NSeq) : ((Sequence_No < 10000) ? ("00"+NSeq) : ((Sequence_No < 100000) ? ("0"+NSeq) : NSeq))));
System.out.println(NSeq);

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)