

## نصب و راه اندازی Nutch و Solr

- ۱- ابتدا nutch 1.4 bin را دانلود کنید و در پوشه Home در Cygwin قرار دهید.
- ۲- به مسیر `cygwin64\home\apache-nutch-1.4-bin\runtime\local\bin` Seeds به یک پوشه به نام Seeds ایجاد کنید.
- ۳- در داخل آن پوشه فایل متنی به نام `urls.txt` ایجاد کرده و در داخل آن Url های خود را بنویسید. هر url در یک خط جدا(به عنوان مثال <http://um.ac.ir/>)
- ۴- به پوشه `cygwin64\home\apache-nutch-1.4-bin\runtime\local\conf` nutch- default.xml در فایل nutch- به دنبال کد زیر بگردید:

```
<name>http.agent.name</name>  
<value> </value>
```

و آن را به صورت زیر تغییر دهید.

```
<name>http.agent.name</name>  
<value>nutch spider</value>
```

ذخیره کرده و فایل را ببندید.

- ۵- سپس در همان پوشه فایل `regex-urlfilter` را باز کرده و در انتهای آن به جای + خط زیر را تایپ کنید:  
`^+http://([a-z0-9\-\A-Z]*\.)*um.ac.ir/([a-z0-9\-\A-Z]*\v)*`

۶- تنظیمات nutch تمام شد

۷- برای اجرای nutch در Cygwin به مسیر زیر بروید:

```
home\apache-nutch-1.4-bin\runtime\local\bin
```

۸- برای کراول دستور زیر را اجرا کنید

```
./nutch crawl seeds -dir crawl -depth 2 -topN 5
```

به جای اعداد ۲ و ۵ هر عددی می تواند قرار گیرد. اگر عمق را زیاد کنید باعث می شود خزش با عمق بیشتری در لینک ها انجام شود. TopN نیز نشان دهنده اینکه در هر صفحه چند تا لینک ها را وارد شود.

اگر تصویر زیر ظاهر شد، کراول با موفقیت انجام شده است.

```

/home/apache-nutch-1.4-bin/runtime/local/bin
* queue: http://nutch.apache.org
maxThreads = 1
inProgress = 0
crawlDelay = 5000
minCrawlDelay = 0
nextFetchTime = 1479336152494
now = 1479336150584
0. http://nutch.apache.org/bot.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://nutch.apache.org
maxThreads = 1
inProgress = 0
crawlDelay = 5000
minCrawlDelay = 0
nextFetchTime = 1479336152494
now = 1479336151585
0. http://nutch.apache.org/bot.html
Fetching http://nutch.apache.org/bot.html
-Finishing thread FetcherThread, activeThreads=8
-Finishing thread FetcherThread, activeThreads=6
-Finishing thread FetcherThread, activeThreads=8
-Finishing thread FetcherThread, activeThreads=6
-Finishing thread FetcherThread, activeThreads=5
-Finishing thread FetcherThread, activeThreads=4
-Finishing thread FetcherThread, activeThreads=2
-Finishing thread FetcherThread, activeThreads=3
-Finishing thread FetcherThread, activeThreads=1
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0
-Finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: Finished at 2016-11-17 02:12:35, elapsed: 00:00:29
ParseSegment: starting at 2016-11-17 02:12:35
ParseSegment: segment: crawl/segments/20161117021204
Parsing: http://nutch.apache.org/bot.html
Parsing: http://nutch.apache.org/credits.html
Parsing: http://nutch.apache.org/downloads.html
Parsing: http://nutch.apache.org/index.html
Parsing: http://nutch.apache.org/mailling_lists.html
ParseSegment: finished at 2016-11-17 02:12:37, elapsed: 00:00:01
CrawlDb update: starting at 2016-11-17 02:12:37
CrawlDb update: db: crawl/crawlDb
CrawlDb update: segments: [crawl/segments/20161117021204]
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 2016-11-17 02:12:38, elapsed: 00:00:01
LinkDb: starting at 2016-11-17 02:12:38
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: adding segment: File:/D:/cygwin64/home/apache-nutch-1.4-bin/runtime/loca
l/bin/crawl/segments/20161117021153
LinkDb: adding segment: File:/D:/cygwin64/home/apache-nutch-1.4-bin/runtime/loca
l/bin/crawl/segments/20161117021204
LinkDb: finished at 2016-11-17 02:12:39, elapsed: 00:00:01
crawl finished: crawl

```

ارسال نتایج nutch به solr

۱- ابتدا solr 3.4 bin را دانلود کنید

۲- سپس از پوشه زیر فایل schema.xml را کپی کرده

cygwin64\home\apache-nutch-1.4-bin\runtime\local\conf

و در مسیر زیر قرار دهید.

apache-solr-3.4.0\example\solr\conf

توجه کنید که باید جایگزین نسخه موجود شود. (Replace)

۳- حال فایل Schema.xml جایگزین شده را باز کرده و تغییرات زیر را اعمال کنید.

به جای دستور

```

<filter class="solr.
EnglishPorterFilterFactory" protected="protwords.txt"/>

```

دستور زیر را قرار دهید.

```

<!-- <filter class="solr.
EnglishPorterFilterFactory" protected="protwords.txt"/> -->

```

سپس بعد از خط <field name="id" ... /> (احتمالا خط ۶۹-۷۰)

دستور زیر را اضافه کنید:

```
<field name="_version_" type="long" indexed="true" stored="true"/>
```

۴- تنظیمات Solr نیز تمام شد

۵- برای اجرای Solr در Cmd به پوشه زیر بروید:

apache-solr-3.4.0\example

۶- دستور زیر را اجرا کنید:

```
java -jar start.jar
```

بعد از مدت کوتاهی اجرا شده و برای مشاهده اجرا، مرورگر خود را باز کرده و به آدرس زیر بروید:

<http://localhost:8983/solr/>



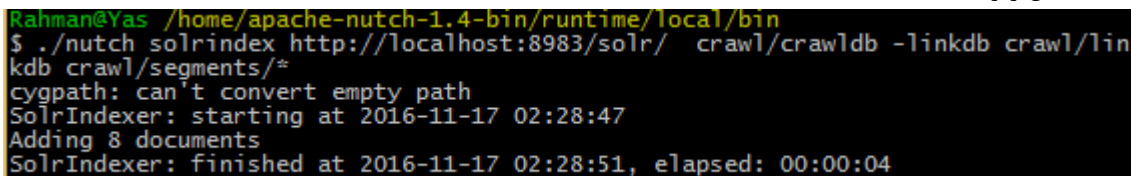
**Welcome to Solr!**

[Solr Admin](#)

۷- برای ارسال نتایج کراول شده توسط nutch دستوز زیر را در Cygwin وارد کنید.

```
./nutch solrindex http://localhost:8983/solr/ crawl/crawldb -linkdb crawl/linkdb  
crawl/segments/*
```

مانند شکل زیر:



```
Rahman@Yas /home/apache-nutch-1.4-bin/runtime/local/bin  
$ ./nutch solrindex http://localhost:8983/solr/ crawl/crawldb -linkdb crawl/linkdb crawl/segments/*  
cygpath: can't convert empty path  
SolrIndexer: starting at 2016-11-17 02:28:47  
Adding 8 documents  
SolrIndexer: finished at 2016-11-17 02:28:51, elapsed: 00:00:04
```

۸- حال Solr را باز کرده و جستجوی مورد نظر خود را انجام دهید.

۹- پایان

توضیحات اضافه:

۱۰- برای مشاهده تمام صفحات کراول شده نیز می توانید از دستورات زیر در Cygwin استفاده کنید:

```
./nutch mergesegs crawl/merged crawl/segments/*
```

`./nutch readseg -dump crawl/merged/* dumpedContent`

تصویر زیر:

```
Rahman@Yas /home/apache-nutch-1.4-bin/runtime/local/bin
$ ./nutch mergesegs crawl/merged crawl/segments/*
cygpath: can't convert empty path
Merging 3 segments to crawl/merged/20161117025249
SegmentMerger: adding crawl/segments/20161117021153
SegmentMerger: adding crawl/segments/20161117021204
SegmentMerger: adding crawl/segments/20161117021442
SegmentMerger: using segment data from: content crawl_generate crawl_fetch crawl
_parse parse_data parse_text

Rahman@Yas /home/apache-nutch-1.4-bin/runtime/local/bin
$ ./nutch readseg -dump crawl/merged/* dumpedContent
cygpath: can't convert empty path
SegmentReader: dump segment: crawl/merged/20161117025249
SegmentReader: done
```

که بعد از اجرای آن در پوشه `dumpedContent` می‌توانید `html` صفحات را مشاهده کنید

موفق باشید

رحمان جلایر: [Rahman\\_jalayer@yahoo.com](mailto:Rahman_jalayer@yahoo.com)