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 رفته و 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 رفته و در فایل nutch-default.xml

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]*\/)*

۶- تنظیمات nutch تمام شد

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

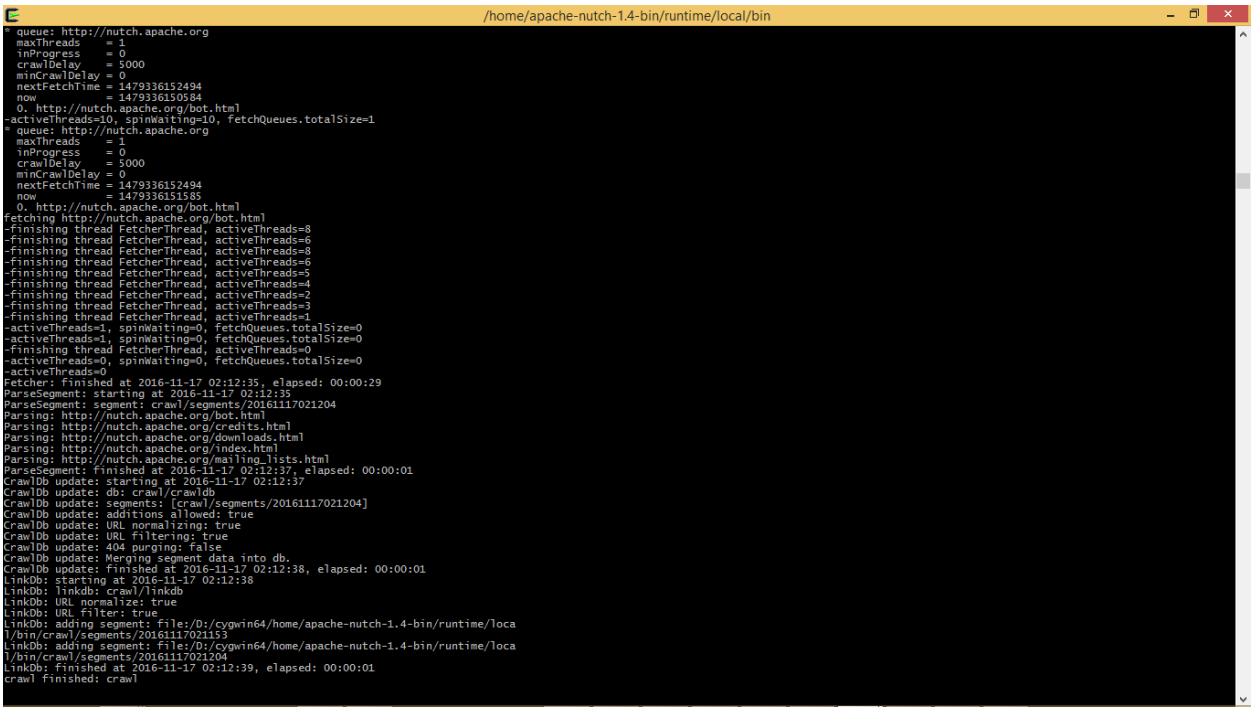
home\apache-nutch-1.4-bin\runtime\local\bin

۸- برای کراول دستور زیر را اجرا کنید

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

به جای اعداد ۲ و۵ 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.




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 بعدا 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
```

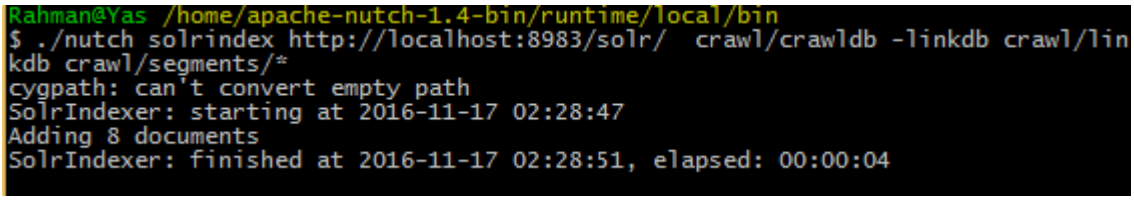
بعد از مدت کوتاهی اجرا شده و برای مشاهده اجرا، مروگر خود را باز کرده و به آدرس زیر بروید:

http://localhost:8983/solr/


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

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

مانند شکل زیر:




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

۹- پایان

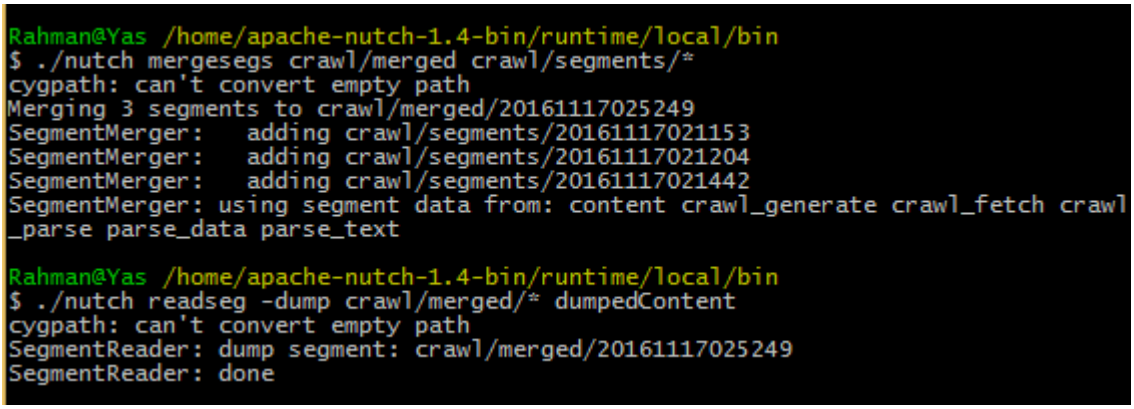
توضیحات اضافه:

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:




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

موفق باشید

رحمان جلایر: Rahman_jalayer@yahoo.com