سیاوش حسن پور

921042006

Apache Solr & Apache Lucene

Lucene:

lucene چیست :

لوسین یک کتابخانه بازیابی اطلاعات(Information Retrieval) متن باز است که در سال ۱۹۹۹ توسط Doug cutting نوشته شده است. با استفاده از آن می‌توان اسناد را نمایه سازی کرد سپس به روی آنها جستجو کرد. Lucene جزو پروژه‌های Apache و تحت لیسانس ASF است. موتورهای جستجو نیز اطلاعات را نمایه گذاری می‌کنند و جستجو را بر روی نمایه‌ها انجام می‌دهند.

Index :

نمایه وسیله‌ای است برای هدایت منظم به یک متن، محتوا، مجموعه‌ای از مدارک و یا هر گونه اطلاعات ضبط شده‌ای که به شکل معموﻻً الفبایی مرتب شده باشد و با استفاده از شیوه خاص و نظام مند

ارجاعی مشخص، موقعیت و محل هر مطلب را در مرحله بازیابی اطلاعات نشان دهد. واحد اصلی نمایه سند است هر سند شامل یک یا چند ستون (Field) است. ساختار نمایه تولید شده توسط لوسین نمایه معکوس است. این ساختمان داده هسته تمام موتورهای جستجو موجود است. ابتدا سند ساخته می‌شود سپس ستون‌ها ساخته و به سند اضافه می‌شوند.

من در زبان 2.7python و در سیستم عامل 14.04 Ubuntu نوشتم .

در ابتدا باید java و ant و lvy که توضیحات نصب بصورت کامل در این [سایت](http://www.sixarm.com/about/java-install-openjdk-ant-ivy-on-ubuntu-linux.html) امده است.

سپس برای این که کتابخانه ی lucene را در python بتوانیم استفاده کنیم باید نصب کنیم (sudo apt-get install pylucene) (برای python 2)/

توجه : دانلود lucene و solr از سایت خود apache .

توضیح کد:

Import lucene که کتابخانه lucene را import میکند.

INDEX\_DIR آدرس جایی که می خواهیم index ها را ذخیره کنیم.

FILES\_TO\_INDEX مکان داده ی ما که می خواهیم index کنیم.

() lucene.initVM: Initialize lucene and JVM.

Analyzerنحوه تحلیل ستون‌ها را نشان می‌دهد

analyzer = lucene.StandardAnalyzer(lucene.Version.LUCENE\_CURRENT)

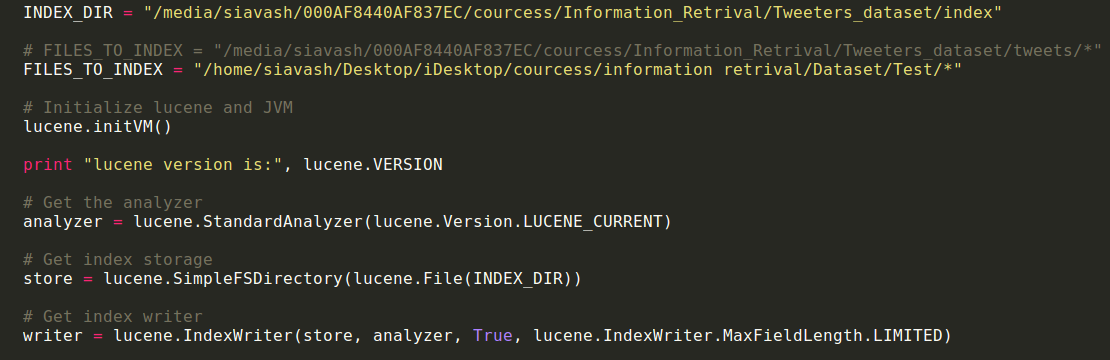
مکان ذخیره ی index را باز میکنیم

store = lucene.SimpleFSDirectory(lucene.File(INDEX\_DIR))

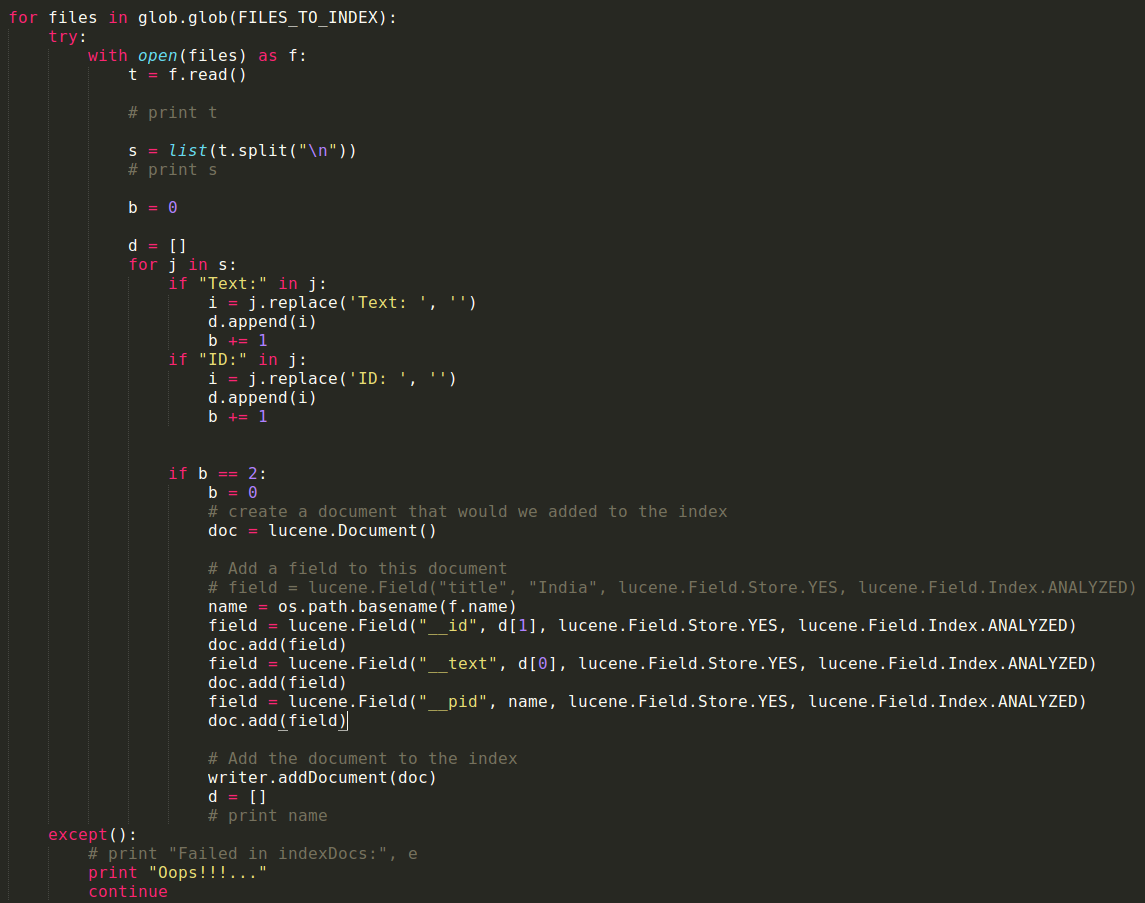
IndexWriter برای ایجاد نمایه از شیء استفاده می‌شود. شیء IndexWriter در زمان ایجاد دو شیء Analyzer و Directory را دریافت م ی کند. شیء Directory پوشه محل ذخیره‌سازی نمایه را مشخص می‌کند شیء. Analyzer نحوه تحلیل ستون‌ها را نشان می‌دهد متغیر بولین اختیاری بوده و به طور پیش فرض false م ی‌باشد

writer = lucene.IndexWriter(store, analyzer, True, lucene.IndexWriter.MaxFieldLength.LIMITED)

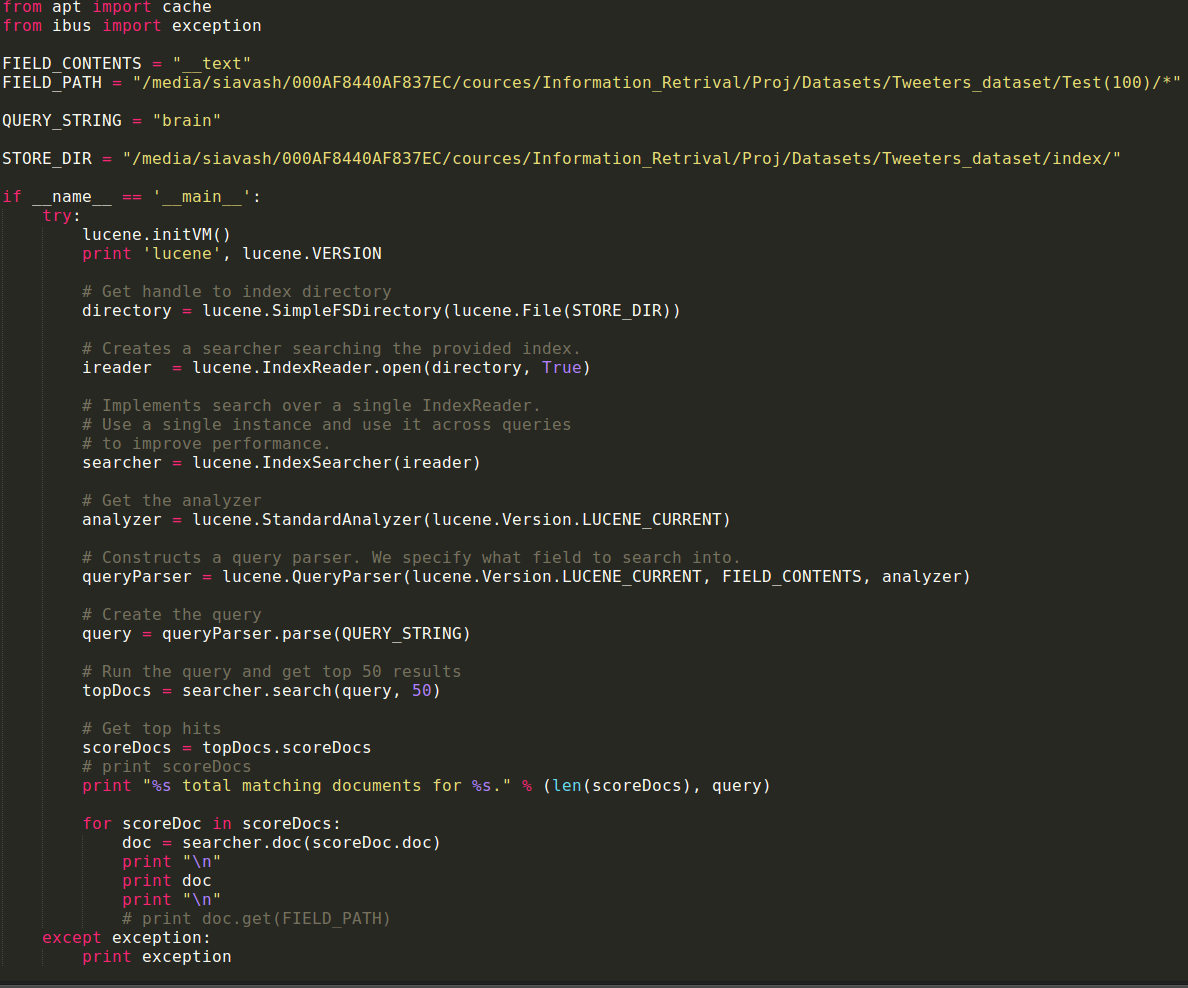
بصورت کلی :



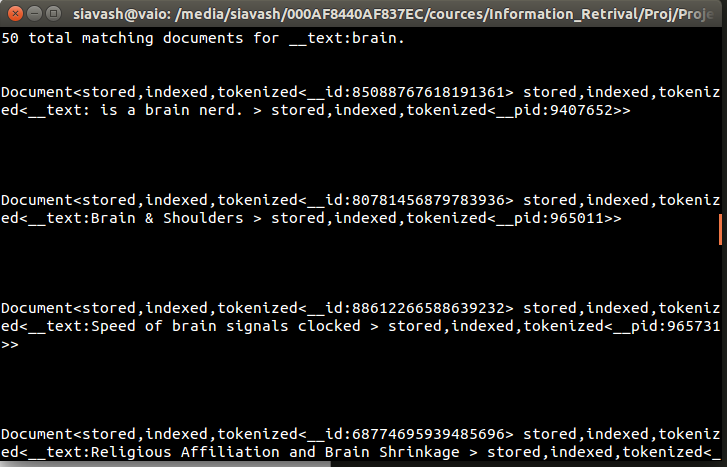
حال با توجه به گفته های بالا و با توجه به این که هر file ما شامل چندین tweet هست و در هر tweet فقط فیلد های text و id مورد نیازهست پس باید هر فایل را split کرده و فقط text , id, pid را به عنوان فیلد بهdoc داده و در آخر این doc را به writer میدهیم تا index کند.



حال برای تست کردن این که درست index میکند یا نه در همان lucene یک search نوشته ام:



که خروجی برای q = brain بصورت زیر است :



Solr :

Apache solr یک موتور جستوجو open source است. که ما داده ها را یا توسط خود solr یا با lucene , index کرده و میتوانیم با انواع query سرچ کنیم.

بعد از دانلود solr و extract کردن ان مراحل زیر را برای راه اندازی solr به ترتیب انجام می دهیم:

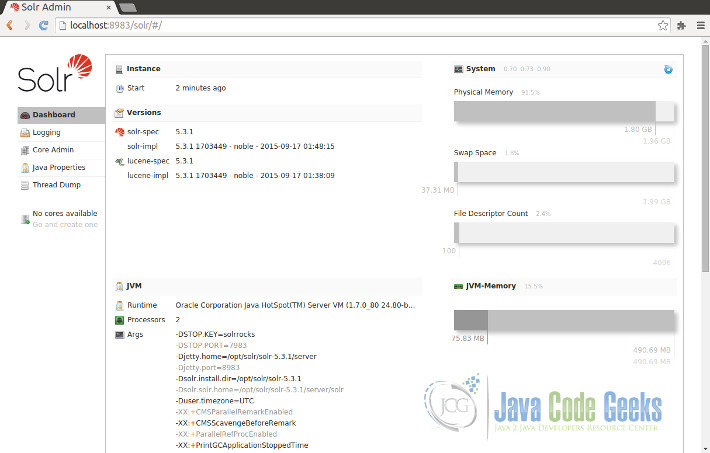
1. ابتدا جاوا 8 نصب را نصب میکنیم
2. sudo apt-get install python-software-properties
3. sudo add-apt-repository ppa:webupd8team/java
4. sudo apt-get update
5. sudo apt-get install oracle-java8-installer
6. نصب solr :
7. cd ~
8. wget <http://apache.mirror1.spango.com/lucene/solr/5.2.1/solr-5.2.1.tgz>
9. tar xzf solr-5.2.1.tgz solr-5.2.1/bin/install\_solr\_service.sh --strip-components=2
10. sudo bash ./install\_solr\_service.sh solr-5.2.1.tgz
11. حال به پوشه ی bin رفته و دستور sudo service solr status را اجرا می کنیم
12. ایجاد core :

* sudo su - solr -c "/opt/solr/bin/solr create -c core\_name -n data\_driven\_schema\_configs"

1. تغییر managed\_schema :

* ابتدا به فایل coreایجاد شده میرویم (/var/solr/core\_name) سپس فایل در managed\_schema تغییرات زیر را در قسمت <uniqueKey>id</uniqueKey> اعمال میکنیم:
* <field name="\_\_id" type="text\_general" indexed="true" stored="true"/>
* <field name="\_\_text" type="text\_general" indexed="true" stored="true"/>
* <field name="\_\_pid" type="text\_general" indexed="true" stored="true"/>
* توجه نام ها بر اساس نام فیلد ها در python تعیین میشود.
* حال باید solr , restart شود : sudo service solr restart

1. حال داده های index شده در lucene را در /var/solr/core\_name/data/index/ کپی کرده و دوباره restart می کنیم.



مراحل config و راه اندازی solr به پایان رسید و حالا میتوانیم سلر را در مرورگر با آدرسه localhost:8983 مشاهده کرد

که core ساخته شده در سمت چپ دیده میشود که وقتی بر روی آن کلیک کرده تعداد index ها دیده میشود .

Query را نیز میتوان در قسمت url بجای (q=hello)داد.

<http://localhost:8983/solr/query?debug=query&q=hello>

**با تشکر**

**حسن پور**