網際網路資訊檢索 HW2

- 姓名：陳丕祐
- 學號：403410061

前言

查看下列的搜尋引擎核心，並且與elastic search進行比較。

- mongodb
- sphinx
- solr

在本次報告之中，我選擇solr與elastic進行比較。

solr

solr是Apache所開發的搜尋引擎，與elastic相同，皆是建構在Lucene之上。故在使用solr前，必須架設好java的環境。

環境與版本

- OS: ubuntu 16.04
- solr: 7.3
- java: 9.04

安裝

從Apache官網下載最新版本的solr。

解壓縮

```
tar -xvf solr-7.3.0.tgz
```

啟動

```
./bin/solr start -e cloud
```

此後會是一些設定項目，主要的內容是要啟動多少個node；使用的port；第一個collection的名稱與設定。

由於solr有一個很棒的dashboard可以使用，不需要在terminal使用簡陋的RESTAPI更新。

我利用這個dashboard新增

- Collections
- Collections 的 field

當然，也可以用terminal的工具

```
./bin/solr create -c [collections name] -s [share number] -rf [copy number]
```

> 要記得新增一個 copy Field，這樣才可以正常進行全文檢索

更新資料

```
./bin/post -c [collections name] [json file or other file]
```

刪除某個collection全部的資料

```
./bin/solr delete -c [collections name] -d "<delete><query>*:*</query></delete>"
```

查詢

```
http://localhost:8983/solr/news/select?hl.fl=[where to highlight]&hl=on&q=[what to
query]&rows=[how many result you want]&start=[from where to start]
```

- hl.fl：選擇那一個field你要進行highlight
- hl：啟動highlight與否
- q：你要進行查詢的字串
- rows：你要得到幾筆結果
- start：從第幾筆開始

斷詞

斷詞的部份我使用jieba加上來自於萌典的詞彙。

我將斷詞的工作從搜尋引擎中切開來。整個流程會變為下圖

```
search word -> jieba text segmentation -> query in solr
```

這樣可以減少在安裝solr+jieba的負擔，就我稍微研究的結果，似乎需要一些java程式的撰寫才能成功。總之我打算放棄這個方法。此外，這樣做的好處就是，不需要 search engine core 支援某一個斷詞服務，反正我們都是在外面就斷好詞，更可以方便的測試各個斷詞的威力如何。

新聞資料

結構

```
@
@Gais_REC:
@url:http://travel.ettoday.net/article/718757.htm
@MainTextMD5:892FF0D0A82D5CA1DB70A320611A9BE9
@UntagMD5:E9D666D79C1A4D3337720C920F2E5A5F
@SiteCode:LvYHeMlIgi
```

```
@UrlCode:ACC79899BE4BEBDDC236B50C3979023A
@title:修杰楷訓練愛女自己吃飯　咘咘完食「拍手鼓掌」萌翻！ | ETtoday 東森旅遊雲 | ETtoday旅遊新聞(影劇)
@Size:89230
@keyword:東森,記者,藝人,賈靜雯,育兒生活
@image_links:http://static.ettoday.net/images/1853/d1853989.jpg
@Fetchtime:2017/01/10 23:15:09
@post_time:2016/06/17 00:00:00
@Ref:http://travel.ettoday.net/article/718839.htm
@BodyMD5:818B0989AB17758DA4E17604D1B47052
@Lang:utf-8
@IP:219.85.79.132
@body: 修杰楷訓練愛女自己吃飯　咘咘完食「拍手鼓掌」萌翻！ | ETtoday 東森旅遊雲 | ETtoday旅遊新聞(影劇)
2016年06月17日 22:18
記者黃庠棻／綜合報導 藝人修杰楷出道13年，2015年5月和大9歲的賈靜雯結婚，同年生下一女咘咘，夫妻倆常常會在臉書分享育兒生活，每次都會吸引大批網友迴響，前不久才在新北市政府服替代役的他近日放假，回到家中陪伴女兒，17日晚間又貼出一段訓練咘咘自己吃飯的影片，可愛的模樣造成粉絲熱烈討論。 ▲賈靜雯和修杰楷常會在臉書分享育兒生活。（圖／翻攝自修杰楷臉書） 修杰楷17日貼出一段咘咘吃飯的影片，表示自己開啟了課，要訓練女兒「吃東西就是要自己來」，只見咘咘坐在嬰兒用座椅，靠著自己的力量，抓著碗裡的食物往嘴塞，雖然動作還有些生澀、笨拙，但不用爸媽餵食，成功吃到東西的模樣也讓許多網友感到相當感動，紛紛大讚「咘咘會自己吃飯啦！」 ▲修杰楷貼出訓練咘咘自己吃飯的影片。（圖／翻攝自修杰楷臉書） 不僅如此，咘咘在連續兩次成功靠自身力量吃到飯之後，竟然伸出肉嘟嘟的雙手「拍手鼓掌」，就像自我鼓勵一樣，逗趣的舉動讓大批粉絲不僅笑成一片，也紛紛直呼「要被萌翻了啦！」該則影片也憑著她的高人氣，才貼出短短1小時就吸引超過4萬個人按讚。 ▲咘咘成功吃完飯後，竟然自己拍手鼓勵，可愛的模樣引起網友討論。（圖／翻攝自修杰楷臉書）
@
```

由於我的電腦記憶體不太夠，所以我採用將資料轉換為40個不同的json檔。

> rec2json.py 在處理這一件事

轉換為json檔所使用的時間。

```
python3 main.py  294.61s user 39.65s system 73% cpu 7:36.95 total
```

比較

配備規格

Cpu：Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz

Ram：8GB

Disk：240GB SSD

將資料存入

	solr	elastic
store speed	01:06:41	00:45:58
store size	23GB	13GB

執行1000筆query

- elastic search

```
python3 test.py  2.35s user 0.40s system 0% cpu 7:12.85 total
```

- solr with jieba

```
python3 test.py  13.88s user 0.77s system 4% cpu 5:56.52 total
```

- solr without jieba

```
python3 test.py  9.84s user 0.49s system 4% cpu 3:47.63 total
```

個人感想

這架設的過程，我認為solr在於文件友善的程度遠超於elastic。尤其是solr的quickstart章節，友善且有用，解決了許多我所遭遇的問題。

反而是elastic最喜歡的做的事情就是只給出一個短小的範例，然後一句有看沒有懂的話加以解釋。真的是很不友善，只好依靠強大的社群(stackoverflow)尋求問題的解答。