Hadoop整合應用-Crawlzilla

@Taiwan Hadoop Ecosystem Workshop 2013

楊順發
https://github.com/shunfa
國家高速網路與計算中心

關於我-楊順發(shunfa)

- **現職:**
  - 國家高速網路與計算中心-助理研究員
  - 業餘Android App開發者

- **領域:**
  - 巨量資料分析

- 專案:
  - Crawlzilla-建立私有搜尋引擎工具

- GitHub & Contant
  - https://github.com/shunfa
  - shunfa@gmail.com

先别管Crawlzilla了

你聽過

"Search Engine" 嗎？

Outline

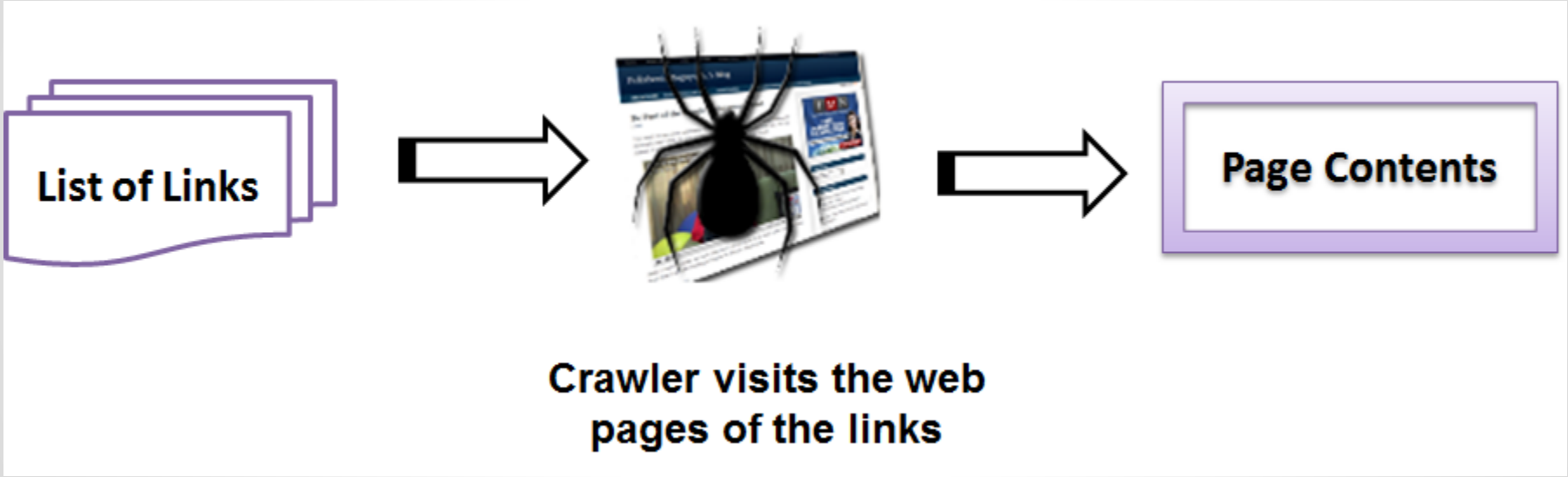
- 初探搜尋引擎運作原理
- 搜尋引擎整合專案介紹-Nutch與Hadoop
- What is Crawlzilla?(using v1.5)
- What is Crawlzilla?(using v2.1)
- Demo

Search Engine

- 三個主要階段
  - 列出爬取清單
  - 網路爬蟲針對清單進行爬取、並建立索引庫 (Index Pool)
  - 搜尋:使用者下搜尋指令時, 系統從索引庫回傳搜尋結果

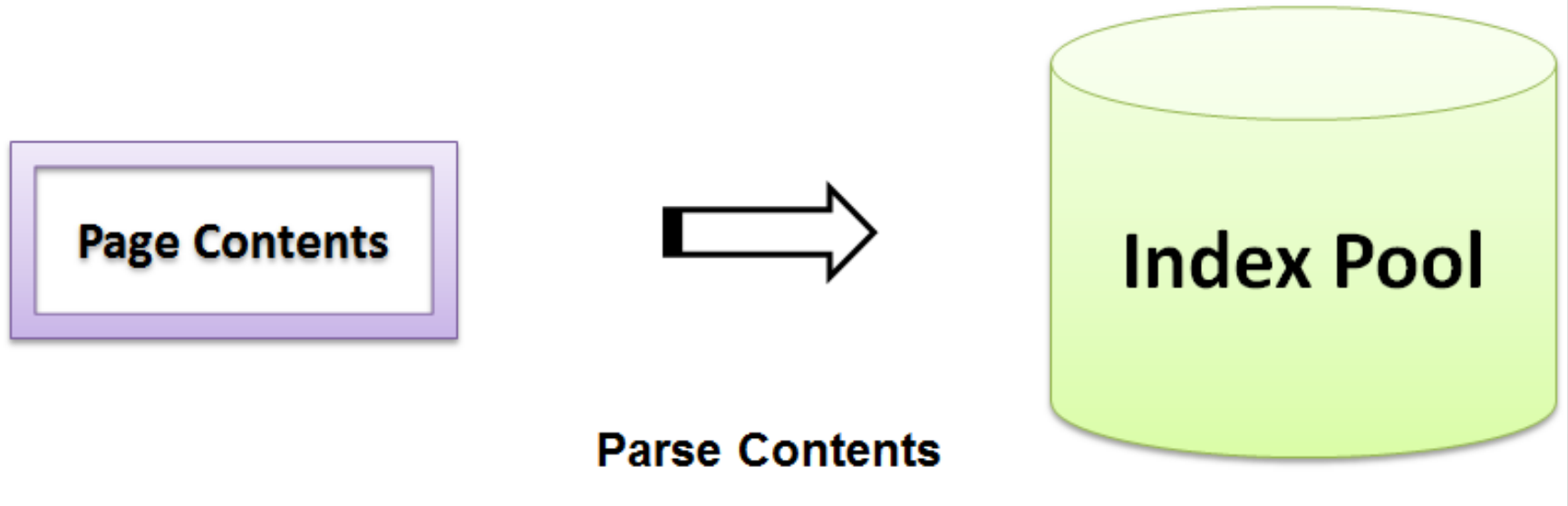
圖解搜尋引擎原理

- Crawling the Web



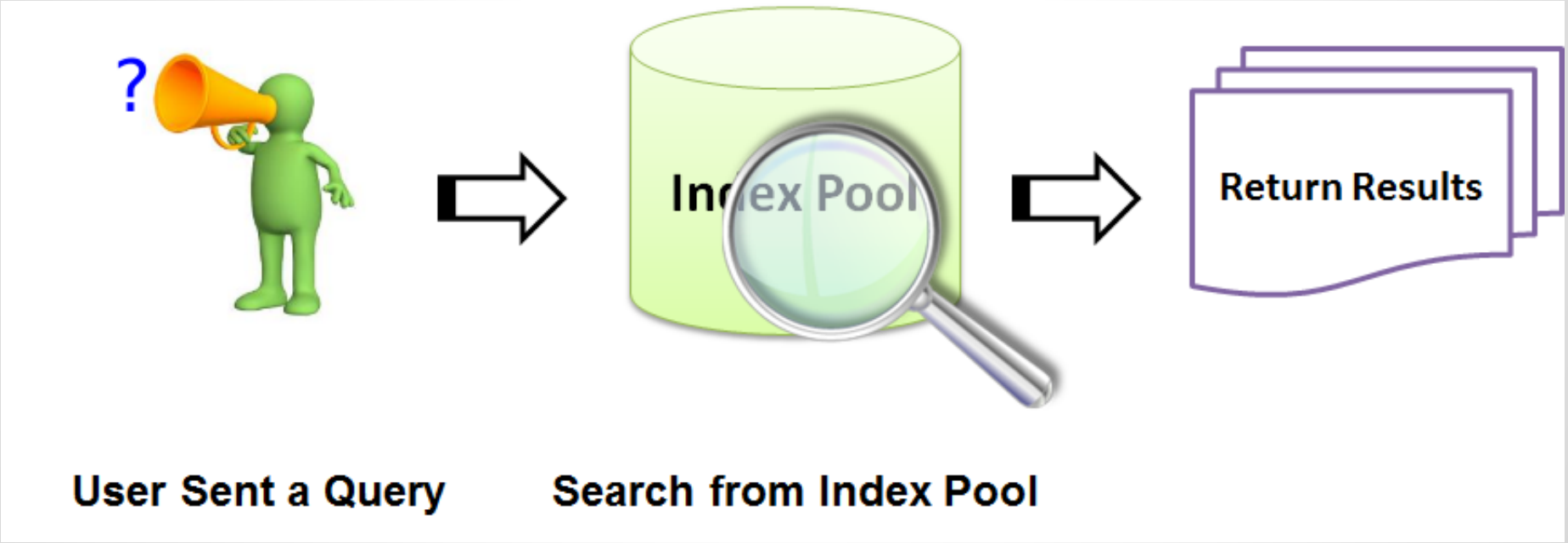
圖解搜尋引擎原理

- Building the Index Pool



圖解搜尋引擎原理

- Serving Queries



Google Search Engine的關鍵技術

- SOSP 2003 : “The Google File System”
- OSDI 2004 : “MapReduce : Simplifed Data Processing on Large Cluster”
- OSDI 2006 : “Bigtable: A Distributed Storage System for Structured Data”

Nutch

- Open Source Search Engine Project
- 始於2002

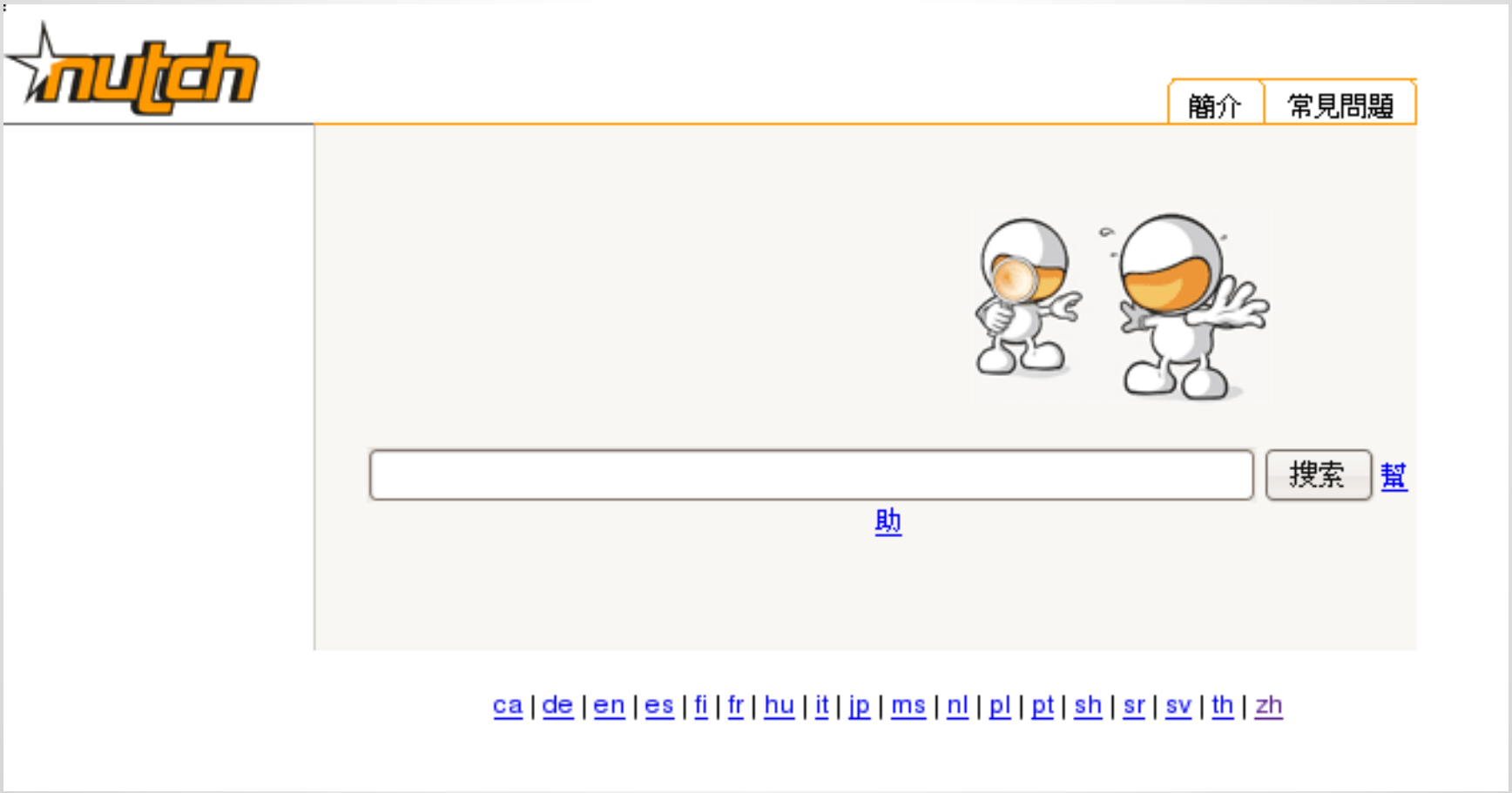
Nutch

- In June, 2003, a successful 100-million-page demonstration system was developed.
- Nutch project has also implemented a MapReduce facility and a distributed file system.
- The two facilities have been spun out into their own subproject, called Hadoop.

Nutch 與 Hadoop

- inject url
- crawldb JOBNAME /crawldb
- generate: select from JOBNAME /crawldb
- generate: partition JOBNAME /segments/20101126161446
- fetch JOBNAME /segments/20101126161446
- crawldb JOBNAME /crawldb
- linkdb JOBNAME /linkdb
- index-lucene JOBNAME /indexes
- dedup 1: urls by time
- dedup 2: content by hash
- dedup 3: delete from index(es)

Using Nutch


nutch
簡介
常見問題
搜索
幫
助
ca | de | en | es | fi | fr | hu | it | jp | ms | nl | pl | pt | sh | sr | sv | th | zh

Nutch 安裝步驟 - Step1. Hadoop單機安裝

- 安裝Hadoop

```
~$ cd /opt
/opt$ sudo wget http://ftp.twaren.
net/Unix/Web/apache/hadoop/core/hadoop-x.x.x/hadoop-x.x.x.
tar.gz
/opt$ sudo tar zxvf hadoop-x.x.x.tar.gz
/opt$ sudo mv hadoop-x.x.x/ hadoop
/opt$ sudo chown -R hadooper:hadooper hadoop
/opt$ cd hadoop/
/opt/hadoop$ gedit conf/hadoop-env.sh
```

Nutch 安裝步驟 - Step1. Hadoop單機安裝

- 修改 conf/hadoop-env.sh

```
export JAVA_HOME=/usr/lib/jvm/java-6-sun
export HADOOP_HOME=/opt/hadoop
export HADOOP_CONF_DIR=/opt/hadoop/conf
export HADOOP_LOG_DIR=/tmp/hadoop/logs
export HADOOP_PID_DIR=/tmp/hadoop/pid
```

Nutch 安裝步驟 - Step1. Hadoop單機安裝

- 修改 conf/hadoop-site.xml

```
<configuration>
  <property>
      <name>fs.default.name</name>
      <value>hdfs://localhost:9000/</value>
      <description> </description>
  </property>
  <property>
      <name>mapred.job.tracker</name>
      <value>localhost:9001</value>
      <description>  </description>
  </property>
  <property>
      <name>hadoop.tmp.dir</name>
      <value>/tmp/hadoop/hadoop-${user.name}</value>
      <description> </description>
  </property>
</configuration>
```

Nutch 安裝步驟 - Step1. Hadoop單機安裝

- HDFS格式化並啟動Hadoop

```
/opt/hadoop$ bin/hadoop namenode -format
/opt/hadoop$ bin/start-all.sh
```

Nutch 安裝步驟 - Step2. 安裝Nutch

- 下載 nutch 並解壓縮
- 配置hadoop,nutch目錄結構
- 複製函式庫檔
- 編輯設定檔
  - hadoop-env.sh
  - nutch-site.xml
  - crawl-urlfilter.txt

Nutch 執行-爬取網站

- 編輯url清單
- 上傳清單到HDFS
- 執行nutch crawl
  - $ bin/nutch crawl urls -dir search -threads 2 -depth 3 -topN 100000

Nutch 執行-設定Query UI(Tomcat)

- 下載tomcat
- 解壓縮
- tomcat server設定
  - 修改 /opt/tomcat/conf/server.xml 以修正中文亂碼問題
- 下載crawl結果
  - $ cd /opt/nutch
  - $ bin/hadoop dfs -get search /opt/search
- 設定nutch的搜尋引擎頁面到tomcat
- 設定搜尋引擎內容的來源路徑
  - $ gedit /opt/tomcat/webapps/ROOT/WEB-INF/classes/nutch-site.xml
- 啟動tomcat

叢集版使用方式

Nutch 安裝步驟 - Step1. Hadoop叢集安裝

前言

清除所有在實做一作過的環境

step 0. 設定機器的ip & hostname 資訊

step 1. 設定兩台機器登入免密碼

step 2. 安裝java

step 3. 下載安裝Hadoop到"主機一"

step 4. 設定 hadoop-env.sh

step 5. 設定 hadoop-site.xml

step 6. 設定masters及slaves

step 7. Hadoop_Home內的資料複製到其他主機上

step 8. 格式化HDFS

step 9. 啟動Hadoop

step 10. 停止hadoop

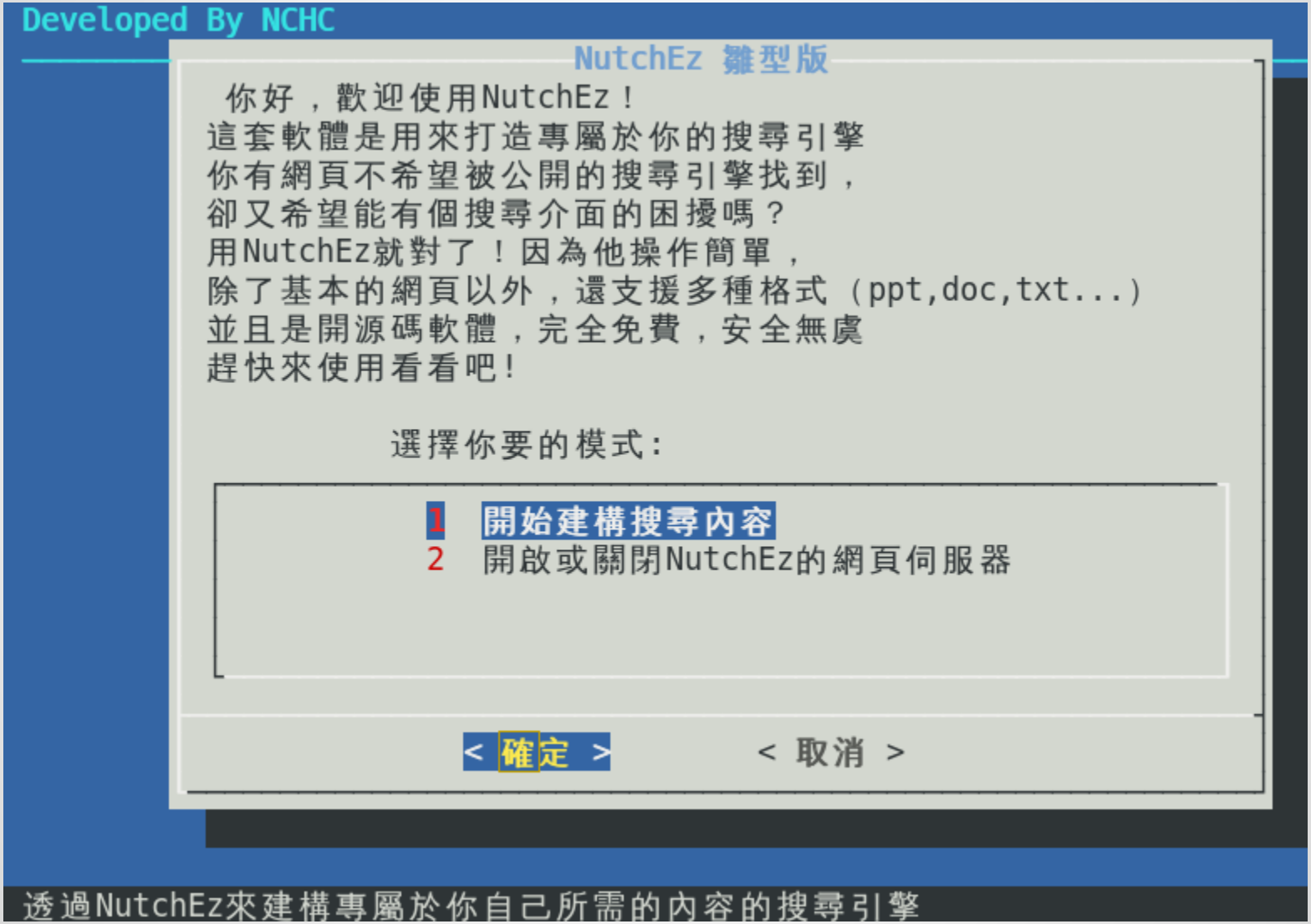
練習

And then...

- 開始安裝 Nutch
- 啟動服務
- 開始建立索引庫
- 建立連結
- 上線服務


OH~BOY !

於是有了NutchEZ



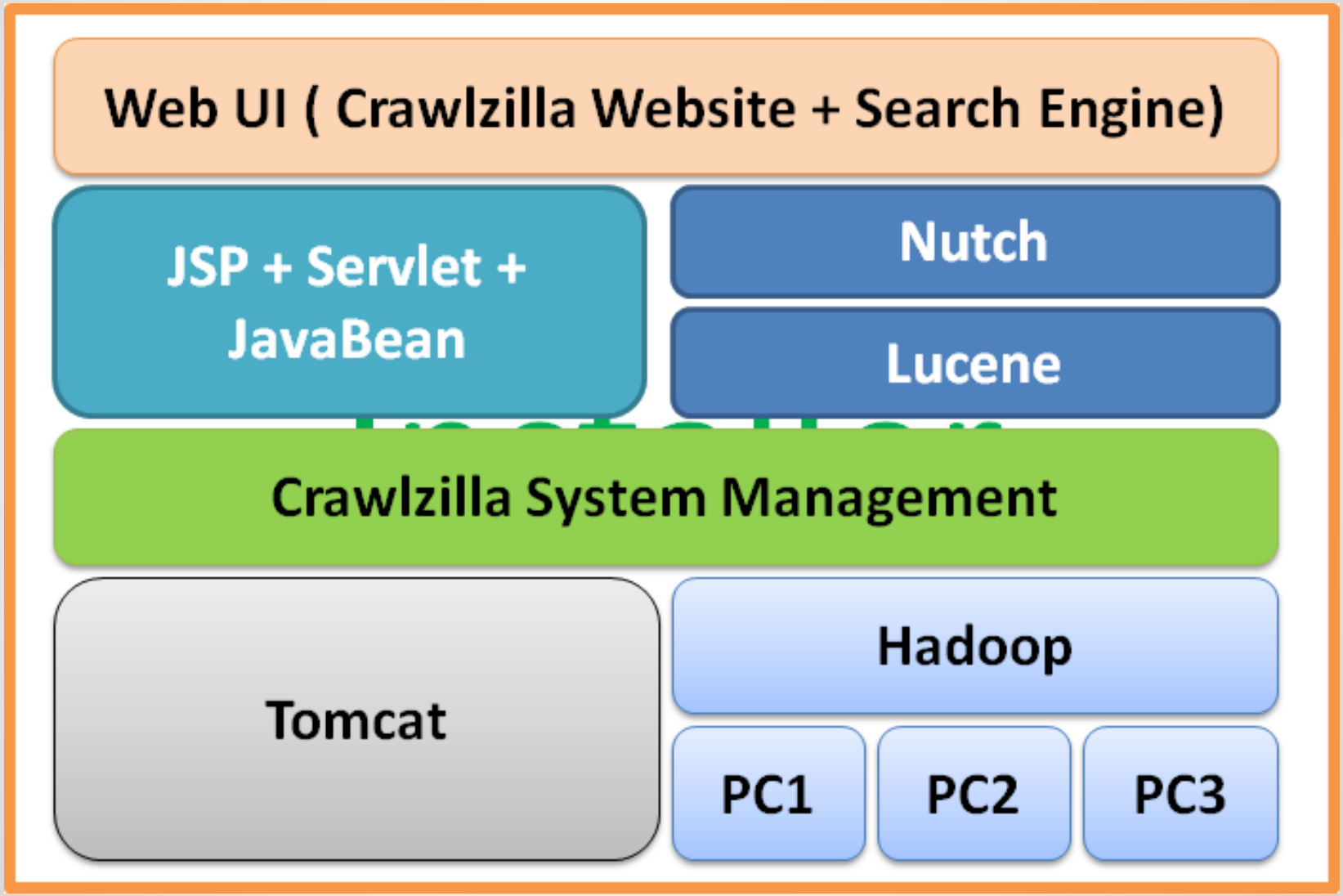
NutchEZ 功能

- 快速安裝Nutch
- 文字模式輸入url list
- 單一搜尋引擎
- 很簡單好用的**單機版**Nutch

NutchEZ 更名為 Crawlzilla

- 快速整合Hadoop，包括架設叢集
- 友善的管理介面
- 大家都會用的瀏覽器使用模式
- 可以知道搜尋引擎的內容
- 多搜尋引擎
- 中文分詞
- .....

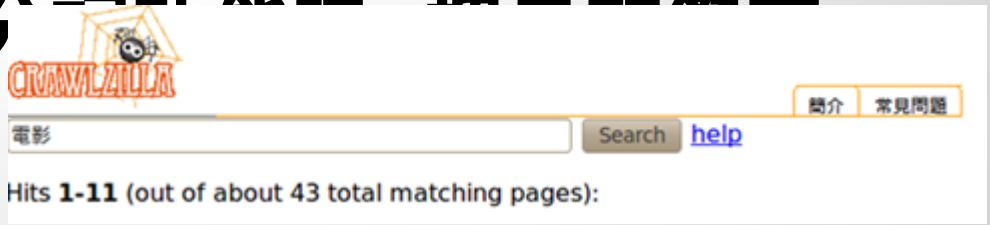
How Crawlzilla Works(V1.5)?



中文分詞

- 索引資料庫會以中文字詞為基本單位建立索引
- 搜尋引擎無中文分詞功能時，搜尋關鍵字 - 電影

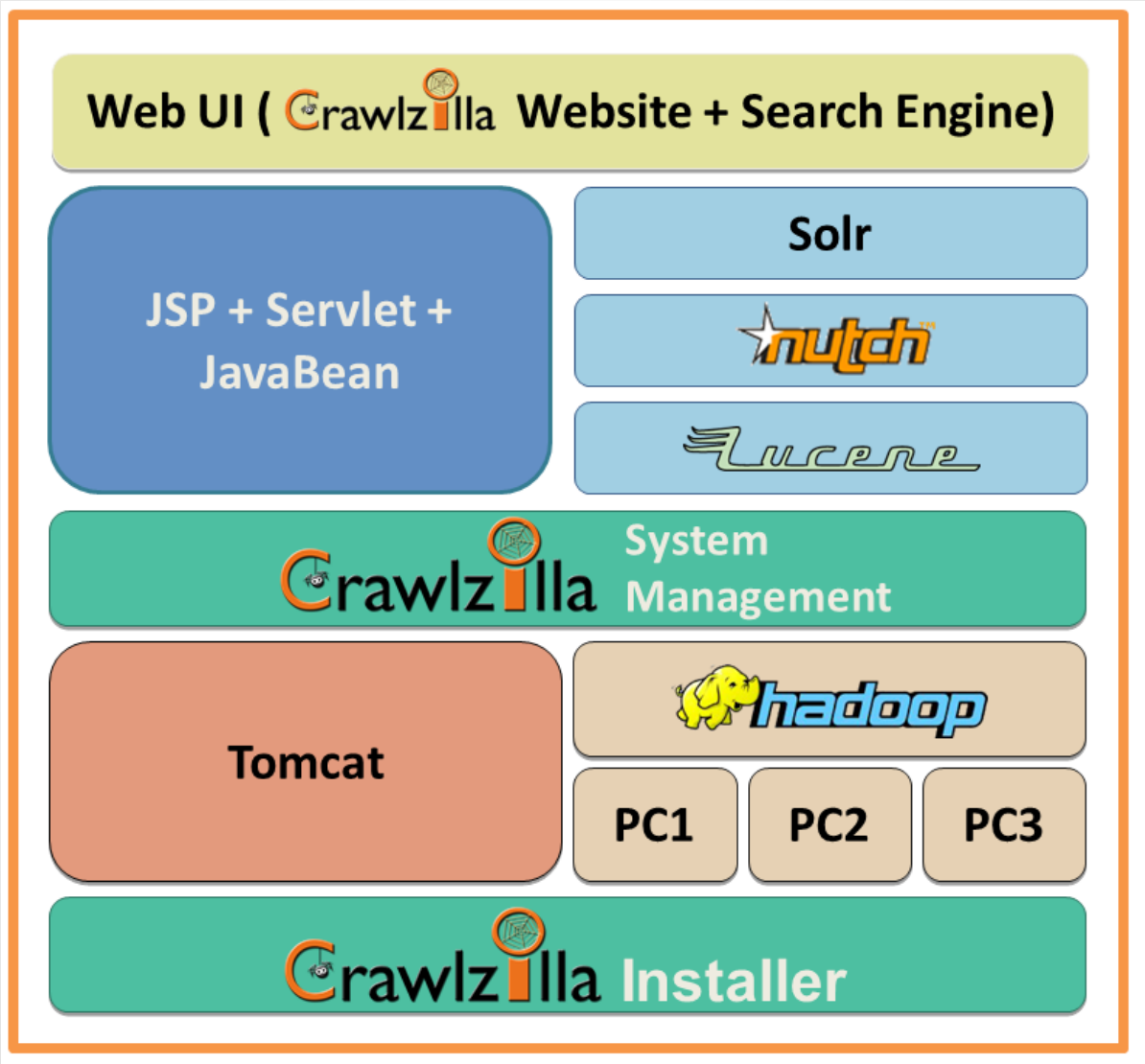
- 760 筆搜尋結果
- 搜尋引擎加入中文分



電影

Demo
(http://demo.crawlzilla.info/)

How Crawlzilla Works(V2.1)?



Crawlzilla 三版本比較表

	0.3	1.x	2.1
Nutch版本	v1.0	v1.2	v1.6
多人版本	X	O	X
檢索工具	Lucene	Lucene	Lucene&Solr
多索引庫	O	O	O
叢集架設	O	O	X(需手動)

Crawlzilla V2.1 與 Solr的整合

- 特色
  - Nutch v1.6
  - 輕量
  - Solr有提供相當完整的索引庫工具
- 缺點
  - 資料量太大時無法爬取成功，需搭配叢集
  - 官方已無提供Query UI，可搭配ajax-solr服用

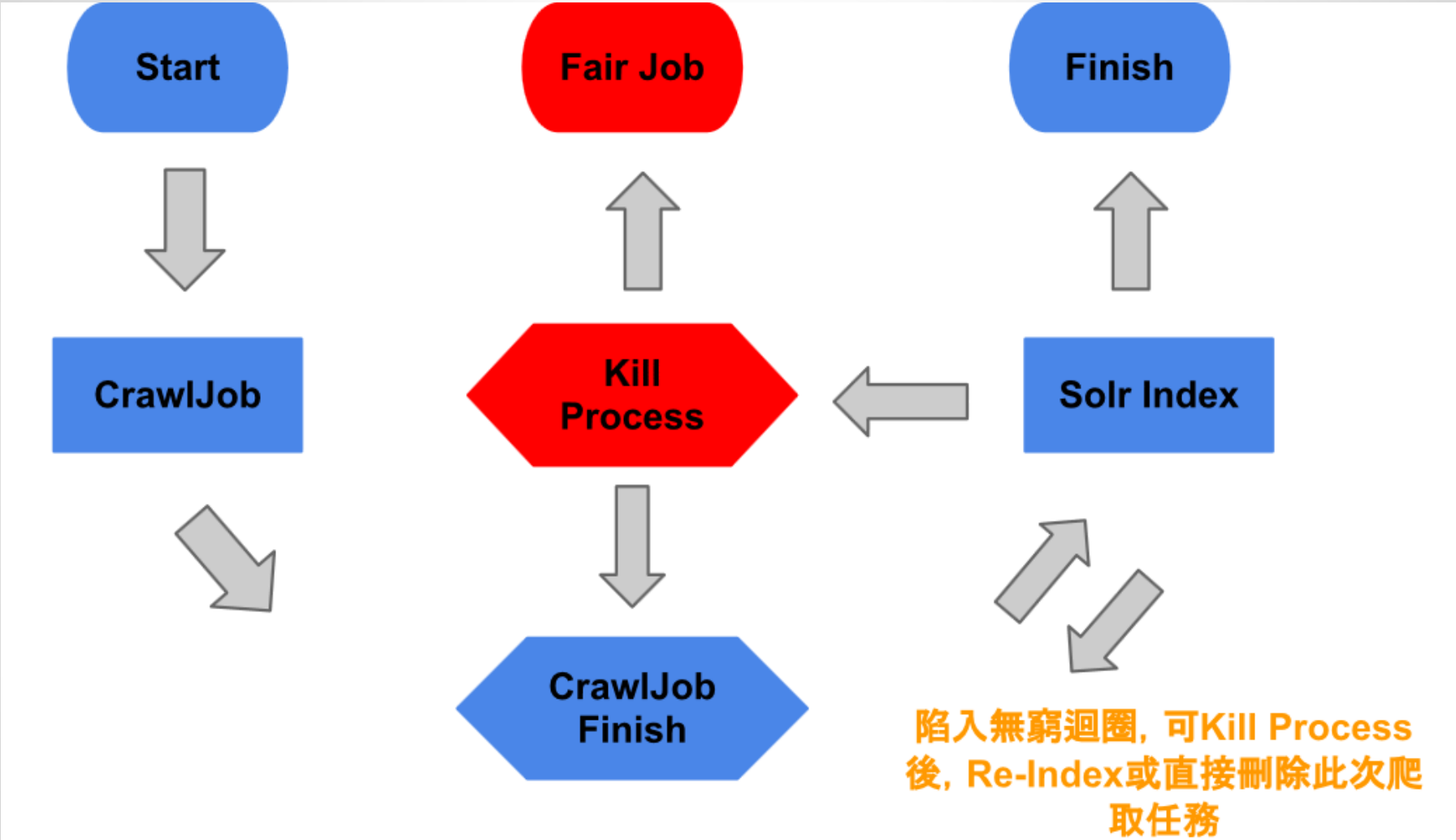
Nutch 1.6 Local (Standalone) Mode

- 不必安裝Hadoop
- 設定以下
  - nutch-site.xml
  - regex-urlfilter.txt
- 執行
  - bin/nutch crawl urls -dir crawl -depth 3 -topN 5
  -

Nutch 1.6 Local (Standalone) Mode

- 配置Solr
  - 將Nutch的配置複製到Solr中
  - 啟動Solr
  - 將Nutch 的 index 匯入至 Solr
  - 修改Solr設定
  - ...more

問題:Crawl Status



判斷Job是否可以修復

- Step1：檢查crawlDB資料夾是否存在DB Name
- Step2：檢查solr.xml及solr資料夾是否存在DB Name，均存在則可執行
  - Step3.a，若不存在則必須執行
  - Step3.b，砍掉從練
- Step3.a：執行reindex程序
- Step3.b：若存在以下資料，則刪除
  - crawlDB/DB_Name
  - solr.xml
  - solr folder

多重索引庫

- Solr設定
  - 新增solr設定檔，直接複製預設即可（collection1）
  - 設定solr.xml
  - 新增
    - <core schema="schema.xml" instanceDir="NCHC_EN_SP_0517/" name="NCHC_EN_SP_0517" config="solrconfig.xml" dataDir="data"/>

Stop Words

- 很多文字可以使文章更流暢，但對搜尋引擎來說試沒有意義的
  - Ex: and, but, then ...

- 建立SolrIndex前可先設定stopword
  - 路徑：$SOLR_HOME/example/solr/IDBName/conf/stopword.txt
  - 格式：一行一個字

Stop Words

- 效果
  - 未設定

10 /7618 Top-Terms:

227	中
203	768
	1024
200	3
183	2013
181	7
179	5776085
177	for
175	4
174	文

Stop Words

- 效果
  - 加入stopword

10 /8726 Top-Terms:

266	中
233	2013
209	文
206	computing
	tel
204	3
202	886
201	fax
	national
199	center

Crawlzilla V2.1-Beta Demo

Feature works

- 中文詞庫
- 更多的通訊協定(ftp, smb...etc)
- Search UI Design
- 結合搜尋引擎的應用

Questions?

- 特別感謝
  - 技術顧問
    - 王耀聰
  - Crawlzilla原創群
    - 陳威宇
    - 郭文傑
    - 楊順發