ElasticSearch 附件內容搜尋 with Java

陳宗霆

起源

Base: JSPWiki

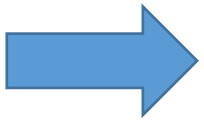
法務需求

- 希望搜尋公文、函示等文件內容
  (政府發布法條異動通常都是doc, pdf)

想找某個關鍵字的相關法令，都得一個 個點開檔案ctrl+F確認

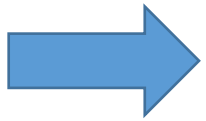
Idea

檔案



AAAAA……
Where am I
….
….
….

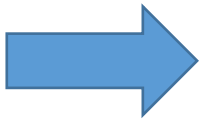
字串



elasticsearch

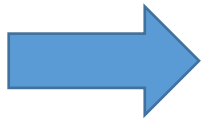
JSPWiki

檔案



AAAAA……
Where am I
….
….
….

字串



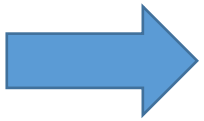
elasticsearch

透過Java poi,
PDF Reader 等
套件分別處理
各種類檔案

用Lucene而非
ElasticSearch

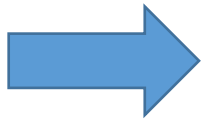
Idea

檔案



AAAAA……
Where am I
….
….
….

字串



改用Tika !!!

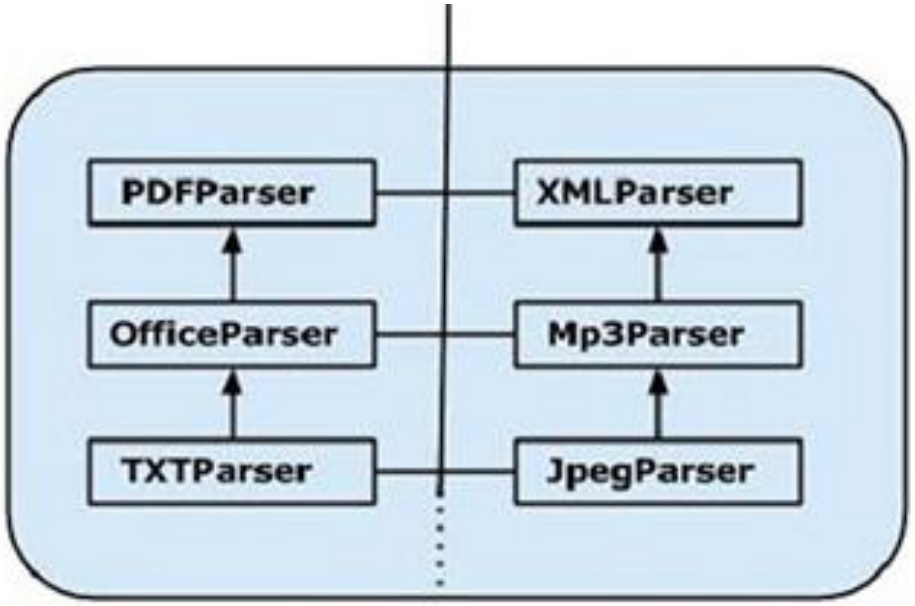
Apache Tika

功能: 自動檢視檔案類型，若有支援則解析出檔案資訊、內容

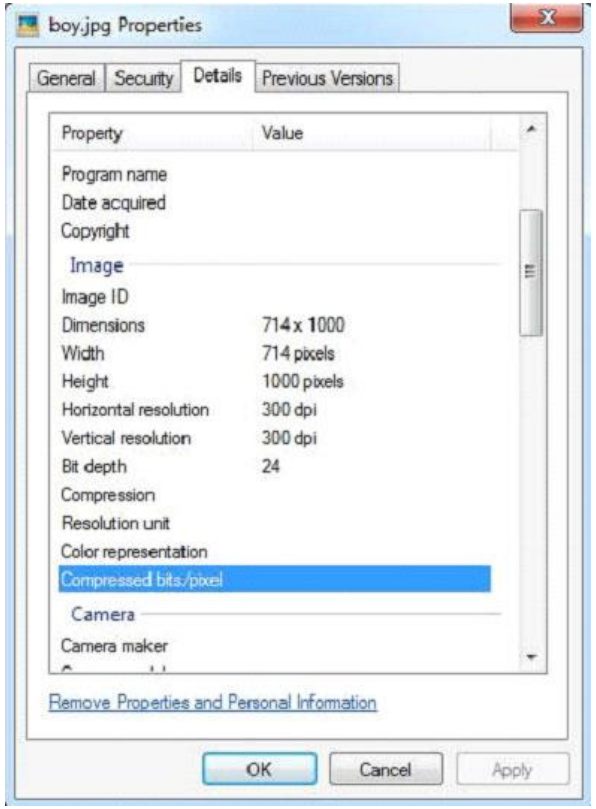
Input: File, FileStream

Output: String

Java Based



Apache Tika




```
Contents of the document:



Metadata of the document:
IPTC-NAA record: 92 bytes binary data
Number of Components: 3
Image Height: 1000 pixels
Resolution Units: inch
Data Precision: 8 bits
tiff:BitsPerSample: 8
Compression Type: Baseline
Component 1: Y component: Quantization table 0, Sampling factors 1 horiz/1 vert
Component 2: Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
tiff:ImageLength: 1000
Component 3: Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert
X Resolution: 300 dots
tiff:ImageWidth: 714
Application Record Version: 4
Image Width: 714 pixels
Original Transmission Reference:
 53616c7465645f5fd22a84941585d89cc735d889c9d5ac58a01faf2c92ee3c6f9bcb38359bbe1eef
Y Resolution: 300 dots
```

Elastic Search Plugin

- https://www.elastic.co/guide/en/elasticsearch/plugins/master/ingest-attachment.html

- Take a rest & install it !!!

- Input: Base64 of File

Idea

檔案

234gfdst34fse
wbgdfgtDSFfd
SDAF

Base64

AAAAA……
Where am I
….
….
….

字串

ingest attachment plugin

elasticsearch

Let’s start practice~~

Environment

- Java 1.7 or higher
- UTF-8
- Library (given pom.xml)

- ElasticSearch 5.6.10 or higher
- UTF-8

Method

- Create Pipeline
- Put File to ElasticSearch
- Query
- Delete

- ……