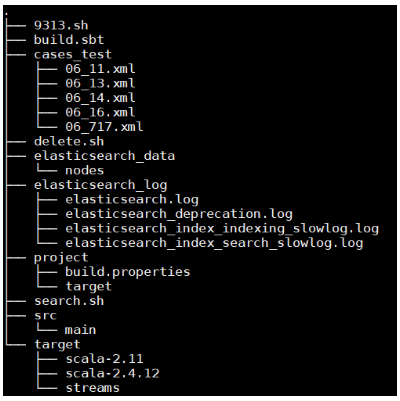
COMP9313 Assignment3 Data Curation + Indexing with ElasticSearch Report

Name: RAN BAI   z-id:z5187292

1. My work directory

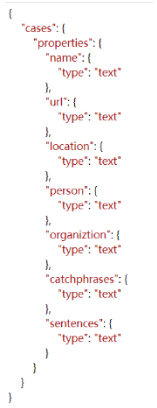


```
.
├── 9313.sh
├── build.sbt
├── cases_test
│   ├── 06_11.xml
│   ├── 06_13.xml
│   ├── 06_14.xml
│   ├── 06_16.xml
│   └── 06_717.xml
├── delete.sh
├── elasticsearch_data
│   └── nodes
├── elasticsearch_log
│   ├── elasticsearch.log
│   ├── elasticsearch_deprecation.log
│   ├── elasticsearch_index_indexing_slowlog.log
│   └── elasticsearch_index_search_slowlog.log
├── project
│   ├── build.properties
│   └── target
├── search.sh
├── src
│   └── main
└── target
    ├── scala-2.11
    ├── scala-2.4.12
    └── streams
```

2. Spark-submit command

spark-submit --class "CaseIndex" --packages "org.scalaj:scalaj-http_2.11:2.4.2" –master local[2] ./target/scala-2.11/caseindex_2.11-1.0.jar cases_test/

3. Index design



```
{
  "cases": {
    "properties": {
      "name": {
        "type": "text"
      },
      "url": {
        "type": "text"
      },
      "location": {
        "type": "text"
      },
      "person": {
        "type": "text"
      },
      "organiztion": {
        "type": "text"
      },
      "catchphrases": {
        "type": "text"
      },
      "sentences": {
        "type": "text"
      }
    }
  }
}
```

My mapping has 7 fields, as shown below

(1) name: This is the title of the case.

(2) url: The original source of the legal report

(3) location: The list of location entities found in catchphrases or sentences with the help of Stanford corenlp server.

(4) person: The list of person entities found in catchphrases or sentences with the help of Stanford corenlp server.

(5) organization: The list of organization entities found in catchphrases or sentences with the help of Stanford corenlp server.

(6) catchphrases: These are short sentences that summarize the case

(7) sentences: The sentences contained in the legal case report.

4. Solution Implementation

(1) Firstly, send http request to elasticSearch server to create a new index(named legal_idx) and mapping(named cases).

(2) Secondly, get directory path from argument, and load all .xml file from it one by one. Then, parsing out the contents of the sentences and catchphrases in the xml file, and send them to Stanford corenlp server to find name entities. In my solution, I only focus on person, location and organization. But in fact we can get more identification types, such as Date, Number.

(3) Finally, build an Json string and send it with http request to elasticSearch server to create an new document for each .xml file.

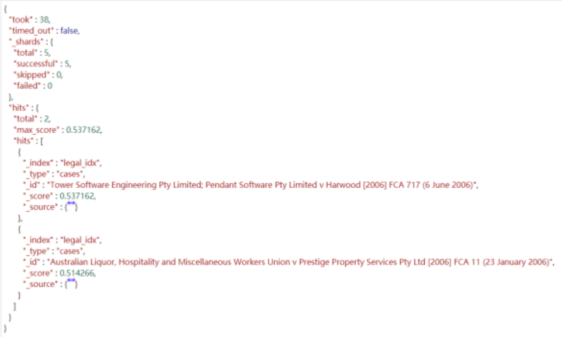
5. Queries example

(1) General term search

Command:

curl -X GET http://localhost:9200/legal_idx/cases/_search?pretty&q=(criminal%20AND%20law)

Result:

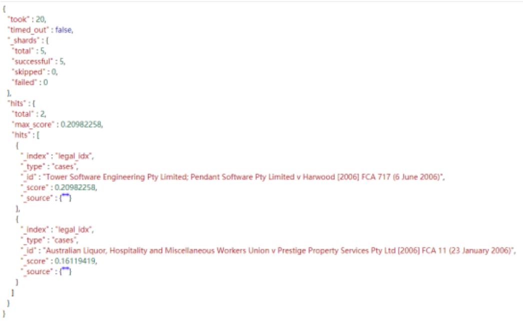


```
{
 "took" : 38,
 "timed_out" : false,
 "_shards" : {
  "total" : 5,
  "successful" : 5,
  "skipped" : 0,
  "failed" : 0
 },
 "hits" : {
  "total" : 2,
  "max_score" : 0.537162,
  "hits" : [
   {
    "_index" : "legal_idx",
    "_type" : "cases",
    "_id" : "Tower Software Engineering Pty Limited; Pendant Software Pty Limited v Harwood [2006] FCA 717 (6 June 2006)",
    "_score" : 0.537162,
    "_source" : {↔}
   },
   {
    "_index" : "legal_idx",
    "_type" : "cases",
    "_id" : "Australian Liquor, Hospitality and Miscellaneous Workers Union v Prestige Property Services Pty Ltd [2006] FCA 11 (23 January 2006)",
    "_score" : 0.514266,
    "_source" : {↔}
   }
  ]
 }
}
```

(2) Entity search (location, person, organization)

Command:

curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=person:John"Result:

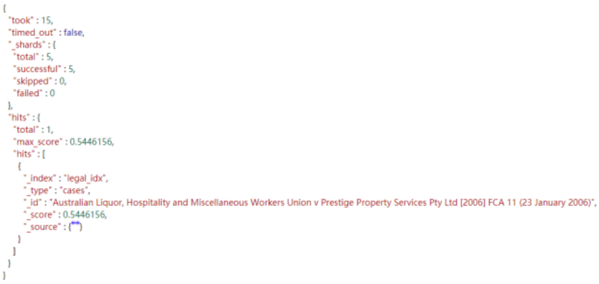


```
{
 "took" : 20,
 "timed_out" : false,
 "_shards" : {
  "total" : 5,
  "successful" : 5,
  "skipped" : 0,
  "failed" : 0
 },
 "hits" : {
  "total" : 2,
  "max_score" : 0.20982258,
  "hits" : [
   {
    "_index" : "legal_idx",
    "_type" : "cases",
    "_id" : "Tower Software Engineering Pty Limited; Pendant Software Pty Limited v Harwood [2006] FCA 717 (6 June 2006)",
    "_score" : 0.20982258,
    "_source" : {↔}
   },
   {
    "_index" : "legal_idx",
    "_type" : "cases",
    "_id" : "Australian Liquor, Hospitality and Miscellaneous Workers Union v Prestige Property Services Pty Ltd [2006] FCA 11 (23 January 2006)",
    "_score" : 0.16119419,
    "_source" : {↔}
   }
  ]
 }
}
```

Command:

curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=location:Melbourne"

Result:



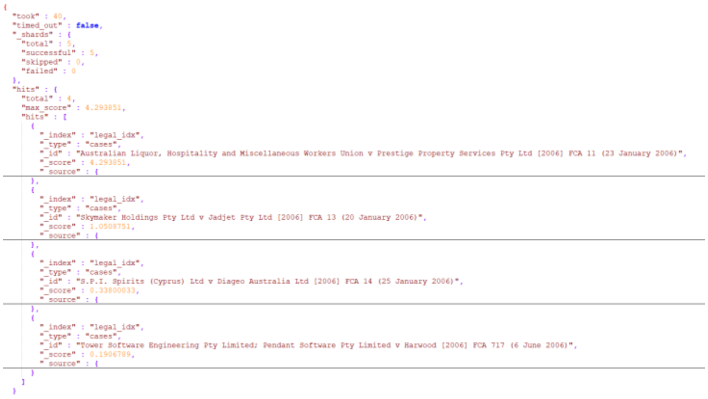
```
{
 "took" : 15,
 "timed_out" : false,
 "_shards" : {
  "total" : 5,
  "successful" : 5,
  "skipped" : 0,
  "failed" : 0
 },
 "hits" : {
  "total" : 1,
  "max_score" : 0.5446156,
  "hits" : [
   {
    "_index" : "legal_idx",
    "_type" : "cases",
    "_id" : "Australian Liquor, Hospitality and Miscellaneous Workers Union v Prestige Property Services Pty Ltd [2006] FCA 11 (23 January 2006)",
    "_score" : 0.5446156,
    "_source" : {↔}
   }
  ]
 }
}
```

Command:

curl -X GET

"http://localhost:9200/legal_idx/cases/_search?pretty&q=organization:State%20Bank%20New%20South%20Wales"

Result:



```
{
  "took" : 40,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 4,
    "max_score" : 4.293851,
    "hits" : [
      {
        "_index" : "legal_idx",
        "_type" : "cases",
        "_id" : "Australian Liquor, Hospitality and Miscellaneous Workers Union v Prestige Property Services Pty Ltd [2006] FCA 11 (23 January 2006)",
        "_score" : 4.293851,
        "_source" : {
      },
      {
        "_index" : "legal_idx",
        "_type" : "cases",
        "_id" : "Skymaker Holdings Pty Ltd v Jadjet Pty Ltd [2006] FCA 13 (20 January 2006)",
        "_score" : 1.0508751,
        "_source" : {
      },
      {
        "_index" : "legal_idx",
        "_type" : "cases",
        "_id" : "S.P.I. Spirits (Cyprus) Ltd v Diageo Australia Ltd [2006] FCA 14 (25 January 2006)",
        "_score" : 0.33800033,
        "_source" : {
      },
      {
        "_index" : "legal_idx",
        "_type" : "cases",
        "_id" : "Tower Software Engineering Pty Limited; Pendant Software Pty Limited v Harwood [2006] FCA 717 (6 June 2006)",
        "_score" : 0.1906789,
        "_source" : {
      }
    ]
  }
```