COMP9313 - 2019T2 Assignment #2 (Data Curation + Indexing with ElasticSearch)

z5147810 Meng Lingxu

Environment: CSE Lab environment, Spark-shell, Scala
Packages: org.apache.spark, spark-core -> 2.4.3, scalaj-http -> 2.4.1, scalatestplus-play -> 4.0.3
Working period: 2019.07.24 - 2019.07.31

1. Run Command

Build scala project package:

```
sbt package
```

Run Scala SBT-packaged jar file:

```
spark-submit --class "CaseIndex" --packages
org.scalaj:scalaj-http_2.11:2.4.1,org.scalatestplus.play:scalatestplus-play_2.11:4.0.3 --master local
JAR_FILE FULL_PATH_OF_DIRECTORY_WITH_CASE_FILES
```

2. Init Index and Set Mapping

Elasticsearch needs a mapping to set index, obviously all fields in the mapping could be set as text type:
"{"cases":{"properties":{"id":{"type":"text"},"name":{"type":"text"},"url":{"type":"text"},"catchphrase":{"type":"text"},"sentence":{"type":"text"},"person":{"type":"text"},"location":{"type":"text"},"organization":{"type":"text"}}}}"""

id, name url catchphrase, sentences are parsed from XML file, describing file attributes.
person, location, organization are entity types supporting search queries generating by NLP.

3. Parsing XML

First, read the second user input argument as input dir, put all files in a list and start to traverse a list the file list.
Parse XML using scala.xml.XML. Extract name, url, catchphrase and sentence for NLP.

4. NLP Analysing

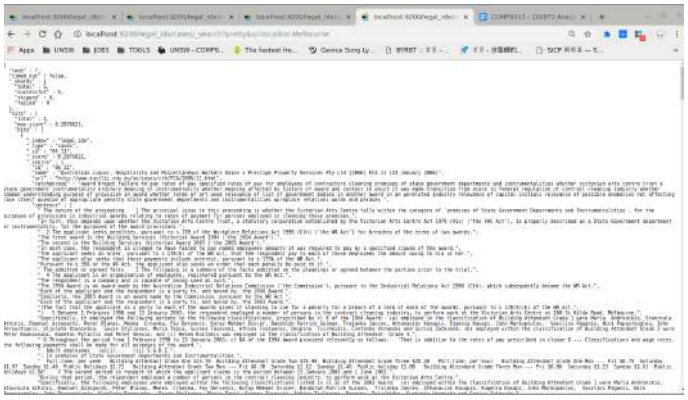
Post each sentence to localhost 9000 port (coreNLP) for analysing. Three sets are initialed for storing unique entity types (person, location, organization).
Then post the result to elasticsearch database.
play.api.libs.json._ is imported for parsing coreNLP result.

4. Query Example

Example: Query based on general terms
curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=(criminal%20AND%20law)"

Example: Queries based on entity type

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



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

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