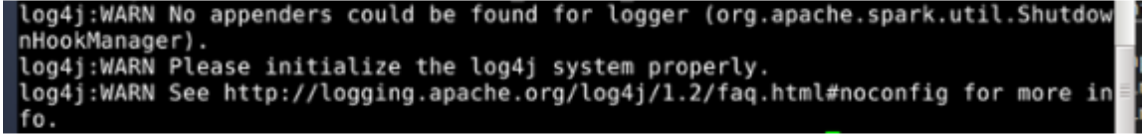
Comp9313 Assignment3 Solution
Name: PEIGUO GUAN
zID: z5143964

Step of test:
Firstly, I use **sbt console** to run the code, and there is not error occurred. And in order to distinguish from others' /legal_idx/cases/, I use /legal_idxgg11/cases/ to check my result.

And then I use spark-submit to run my code, and change my index and again, although I got this error below(I still do not why and the solution in the forum cannot solve my problem), I still got the same result as expected and run successfully.

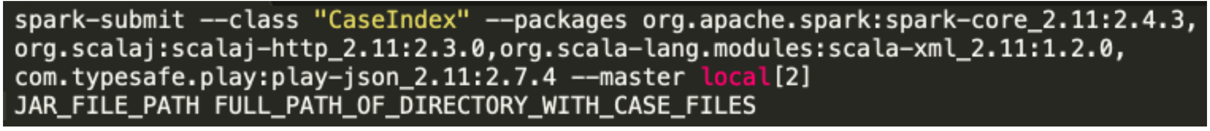


```
log4j:WARN No appenders could be found for logger (org.apache.spark.util.Shutdow
nHookManager).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
```

The explanation of my code and logic:

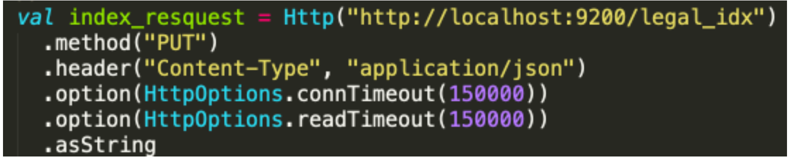
Firstly, get read the file and get all the case test file path by function: ***getAllFiles***, then setting ***index_request*** and **mapping_request** to elasticsearch. I use the function ***processAllFile*** function to process all the next steps: 1.load xml file and separate each part 2.pick the sentences part and send to the corenlp to parse sentence and return ***person***, ***location***, and ***organization*** 3.reconstruct, filter, get tokens… to different ***_list***, and build a ***es_post_data***, PUT es_post_data to the elasticsearch.

Running part:
Spark-submid command:



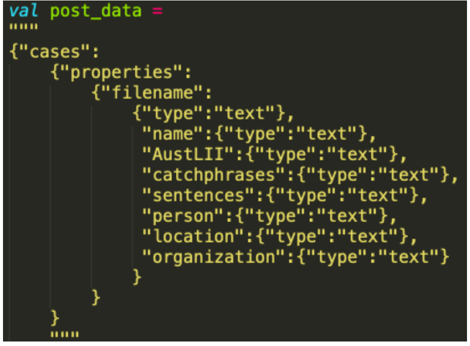
```
spark-submit --class "CaseIndex" --packages org.apache.spark:spark-core_2.11:2.4.3,
org.scalaj:scalaj-http_2.11:2.3.0,org.scala-lang.modules:scala-xml_2.11:1.2.0,
com.typesafe.play:play-json_2.11:2.7.4 --master local[2]
JAR_FILE_PATH FULL_PATH_OF_DIRECTORY_WITH_CASE_FILES
```

my index design:



```
val index_resquest = Http("http://localhost:9200/legal_idx")
  .method("PUT")
  .header("Content-Type", "application/json")
  .option(HttpOptions.connTimeout(150000))
  .option(HttpOptions.readTimeout(150000))
  .asString
```

my mapping:



```
val post_data =
"""
{"cases":
    {"properties":
        {"filename":
            {"type":"text"},
             "name":{"type":"text"},
             "AustLII":{"type":"text"},
             "catchphrases":{"type":"text"},
             "sentences":{"type":"text"},
             "person":{"type":"text"},
             "location":{"type":"text"},
             "organization":{"type":"text"}
            }
        }
    }
    """
```

(please ignore the \n Part in post data)
There are 8 parts in my mapping:

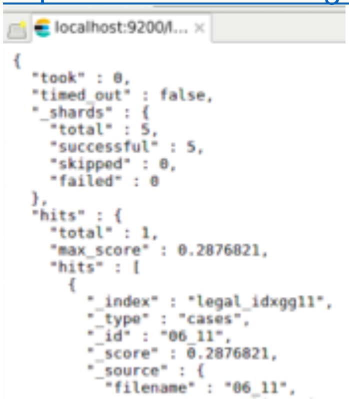
1. filename: is for the file name in case_test

2. name: is the name in file
3. AustLII: is for the AustLII in file
4. catchphrases: for the Catchphrases in test file
5. sentences: for the sentences in test file
6. person: for the person after processing by corenlp
7. location: for the location after processing by corenlp
8. organization: for the organization after processing by corenlp

Result of example queries:(I create my own legal_idxgg11 to distinguish from others in real test)

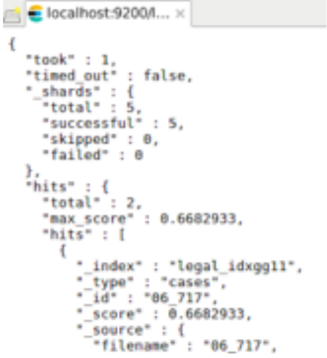
1. queries based on entity type:

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



```
{
  "took" : 0,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 1,
    "max_score" : 0.2876821,
    "hits" : [
      {
        "_index" : "legal_idxgg11",
        "_type" : "cases",
        "_id" : "06_11",
        "_score" : 0.2876821,
        "_source" : {
          "filename" : "06_11",
```

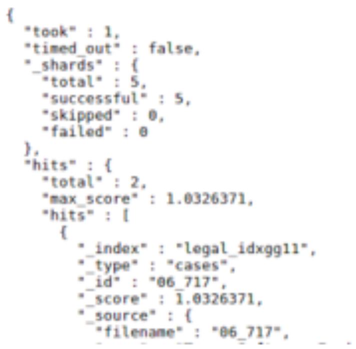
curl -X GET http://localhost:9200/legal_idx/cases/_search?pretty&q=person:John



```
{
  "took" : 1,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 2,
    "max_score" : 0.6682933,
    "hits" : [
      {
        "_index" : "legal_idxgg11",
        "_type" : "cases",
        "_id" : "06_717",
        "_score" : 0.6682933,
        "_source" : {
          "filename" : "06_717",
```

2. query based on general term:

curl -X GET http://localhost:9200/legal_idxgg11/cases/_search?pretty&q=(criminal AND law)
(which is actually: (criminal%20AND%20law))



```
{
  "took" : 1,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 2,
    "max_score" : 1.0326371,
    "hits" : [
      {
        "_index" : "legal_idxgg11",
        "_type" : "cases",
        "_id" : "06_717",
        "_score" : 1.0326371,
        "_source" : {
          "filename" : "06_717",
```