COMP9313 Assignment-3

Amritesh Singh
z5211987

Documentation

The following process was followed to complete the assignment:

- **build.sbt** was created in the project directory with project name given as **CaseIndex**. Library dependencies for Apache Spark and Scalaj-HTTP were also included in the file.
- The file **CaseIndex.scala** was created inside the project directory at the path **/src/main/scala/**. This file contains the main code.
- The project was built by executing the commands **sbt package** and **spark-submit --packages "org.scalaj:scalaj-http_2.11:2.4.2" --class "CaseIndex" --master local[2] JAR_FILE INPUT_DIR**.
- The **CoreNLP** server is run on **http://localhost:9000**.
- The **Elasticsearch** server is run on **http://localhost:9200**.

The main code has the following workflow:

- A new Spark configuration is created and the app name is set to **CaseIndex**. A Scala Spark Context variable is set.
- Initially, the directory containing the input files, i.e., a list of legal case reports in XML format, is loaded and the file names are also stored as keys.
- For each case report, the contents of the file string are mapped to the corresponding XML object.
- The text content of each XML object is subsequently sent to the CoreNLP server to curate the data through the **Named Entity Recognition (NER) API** of CoreNLP.
- The **generate_et()** function is called for each file string that has been enriched. This function generates sets for each of the desired entity types, i.e., location, organization, and person, by loading the XML object from the file string and iterating over the **NER** tag contents of the file string. Whenever there is a match for one of the entity types, i.e., if the text content of the key matches one of **LOCATION**, **ORGANIZATION**, or **PERSON**, then the text content, which corresponds to the value index of the **word** tag contents of the file string, is added to the corresponding entity type set. Finally, each set is converted to String format by combining all set elements into a single string, with each entry separated by spaces in the string, and is returned by the function as a tuple that consists of 3 elements, i.e., **(location, organization, person)**.
- Subsequently, an Elasticsearch index is created with the name **legal_idx** and a mapping is created for this index, having entity types **(file_name, file_content, location, organization, person)**, by sending PUT requests to the Elasticsearch server. For each case report, an Elasticsearch document is created by sending a PUT request to the Elasticsearch server with the data in the mapped format.
- Thus finally, the curated data is indexed on the Elasticsearch server.

OUTPUT EXAMPLES

- **General Term**


```
s":400}z5211987@vx3:/tmp_amd/reed/export/reed/3/z5211987$ curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=(criminal%20AND%20law)"
{
  "took" : 32,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 2,
    "max_score" : 1.0285228,
    "hits" : [
      {
        "_index" : "legal_idx",
        "_type" : "cases",
        "_id" : "06_717",
        "_score" : 1.0285228,
        "_source" : {
          "file_name" : "06_717",
          "file_content" : "<?xml version=\"1.0\"?> <case> <name>Tower Software Engineering Pty Limited; Pendant Software Pty Limited v Harwood [2006] FCA 717 (
au/au/cases/cth/FCA/2006/717.html</AustLII> <catchphrases> <catchphrase >interlocutory application to restrain defendant from taking any further step in a proce
e> <catchphrase >interlocutory application for orders to dismiss or stay court proceeding</catchphrase> <catchphrase >pre-emptive rights regime under company c
r</catchphrase> <catchphrase >whether it is within discretion of directors to refuse to register acceptance of shares for reason that may prevent a higher bid</
```


- **Location**


```
z5211987@vx3:/tmp_amd/reed/export/reed/3/z5211987$ curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=location:Ruski"
{
  "took" : 5,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 1,
    "max_score" : 0.5565415,
    "hits" : [
      {
        "_index" : "legal_idx",
        "_type" : "cases",
        "_id" : "06_14",
        "_score" : 0.5565415,
        "_source" : {
          "file_name" : "06_14",
          "file_content" : "<?xml version=\"1.0\"?> <case> <name>S.P.I. Spirits (Cyprus) Ltd v Diageo Australia Ltd [2006] FCA 14 (25 January 2006)</na
14.html</AustLII> <catchphrases> <catchphrase >separate decision of questions</catchphrase> <catchphrase >where application related to distribution of
leading conduct, false representations and breach of contract</catchphrase> <catchphrase >where cross-claim related to ownership of trademarks used in
 not involved in application</catchphrase> <catchphrase >where no relief sought against respondent in cross-claim</catchphrase> <catchphrase >whether
```


- **Organization**


```
z5211987@vx3:/tmp_amd/reed/export/reed/3/z5211987$ curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=organization:Prestige"
{
  "took" : 14,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 1,
    "max_score" : 0.2876821,
    "hits" : [
      {
        "_index" : "legal_idx",
        "_type" : "cases",
        "_id" : "06_11",
        "_score" : 0.2876821,
        "_source" : {
          "file_name" : "06_11",
          "file_content" : "<?xml version=\"1.0\"?> <case> <name>Australian Liquor, Hospitality and Miscellaneous Workers Union v Prestige Property Serv
>http://www.austlii.edu.au/au/cases/cth/FCA/2006/11.html</AustLII> <catchphrases> <catchphrase >award</catchphrase> <catchphrase >breach</catchphrase>
ase> <catchphrase >rates of pay for employees of contractors cleaning premises of state government departments and instrumentalities</catchphrase> <cat
ent instrumentality</catchphrase> <catchphrase >ordinary meaning of instrumentality</catchphrase> <catchphrase >whether meaning affected by history of
rase >transition from state to federal regulation of contract cleaning industry</catchphrase> <catchphrase >whether common understanding</catchphrase>
```


- **Person**


```
z5211987@vx3:/tmp_amd/reed/export/reed/3/z5211987$ curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=person:Lina"
{
  "took" : 12,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 1,
    "max_score" : 0.2876821,
    "hits" : [
      {
        "_index" : "legal_idx",
        "_type" : "cases",
        "_id" : "06_13",
        "_score" : 0.2876821,
        "_source" : {
          "file_name" : "06_13",
          "file_content" : "<?xml version=\"1.0\"?> <case> <name>Skymaker Holdings Pty Ltd v Jadjet Pty Ltd [2006] FCA 13 (20 January 2006)</name>
/AustLII> <catchphrases> <catchphrase >interlocutory mandatory injunction</catchphrase> <catchphrase >injunctions</catchphrase> </catchphrases> <se
rlocutory relief in the form of a mandatory injunction.</sentence> <sentence id=\"s1\">It arises in the following circumstances.</sentence> <senten
claim pursuant to ss 52 , 75B , 80 , 82 and 87 of the Trade Practices Act 1974 (Cth) ('the Act') and for breach of contract relating to the purchas
```