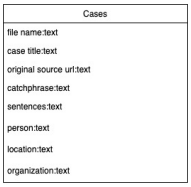
Assignment3 Solution:

Please do not forget to add packages in command.
Example command: spark-submit --class "CaseIndex" --packages 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 FULL_PATH_OF_DIRECTORY_WITH_CASE_FILES

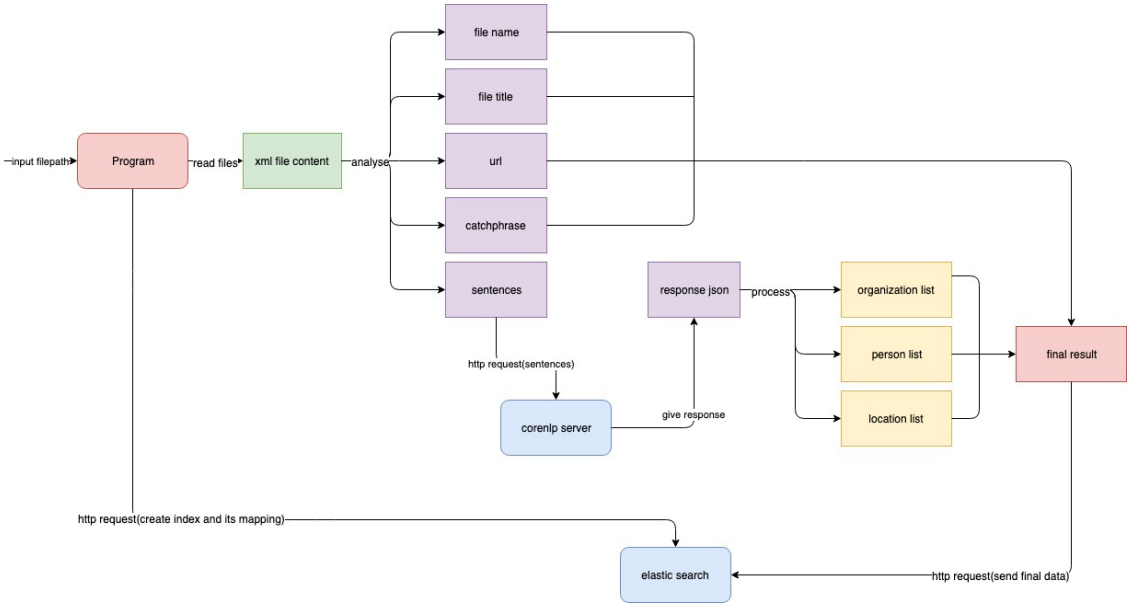
Index Design:

The index design is shown in picture below.



Program Solution:

First, the program should read all the .xml files from the provided file path and make http requests to elastic search server to create an index which is called legal_idx and define its mapping. After that, we should deal with the content in xml file and make it can be processed by coreNLP server. The information that coreNLP server required is in label <sentences>, so we should take them out and combine them to a string, then send a http request that contains the processed data to the corenlp. Also, we should take the content in label <name>, <AustLII> and <catchphrases> for further data processing.

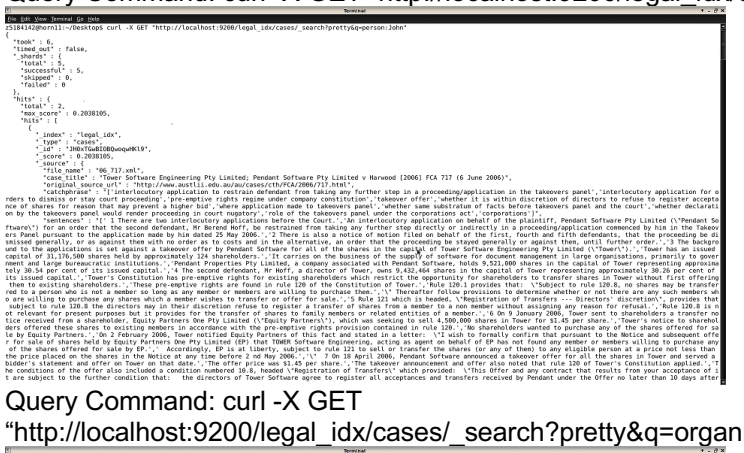


While we send the http request to the coreNLP server, it will give us a response in json format, so we need import packages that contains json parser. In this assignment I use play json package. After parsing the response, we should assort the data that has tags like person, location and organization and add them to lists (person list, location list, organization list). Until now, we have completed all the data processing steps. Then we just need to combine them and send to elastic search.

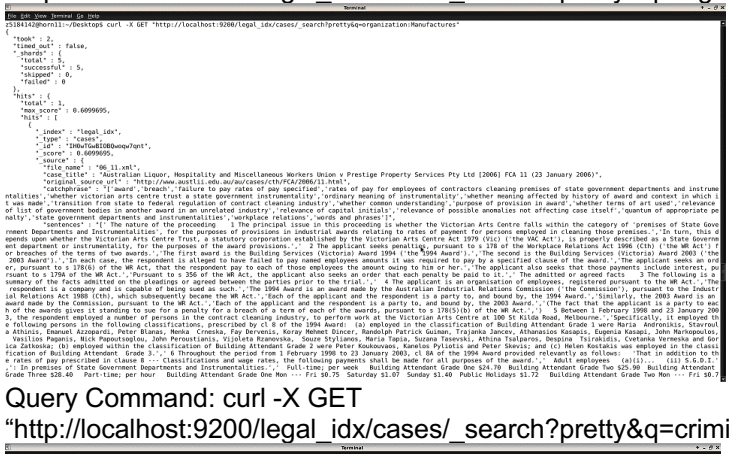
Example Queries Results:

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

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



Query Command: curl -X GET “http://localhost:9200/legal_idx/cases/_search?pretty&q=organization:Manufactures”



Query Command: curl -X GET “http://localhost:9200/legal_idx/cases/_search?pretty&q=criminal%20AND%20Law”