COMP9313 – Bigdata Management

Assignment – 3

Z5200336

The assignment involves creating an Elasticsearch index with the given XML files. The code starts with reading all the files with the ".xml" from the given directory. The names of all files are stored in a list.

The next step is to open the files and read the data as a single string and save it in fileText variable which is then passed to the NLP server. The NLP server parses the data and gives a response with all the data as a JSON.

An Elasticsearch query is sent as PUT request to create the index legal_idx. Then a mapping is created with the following data to be stored in Text format:

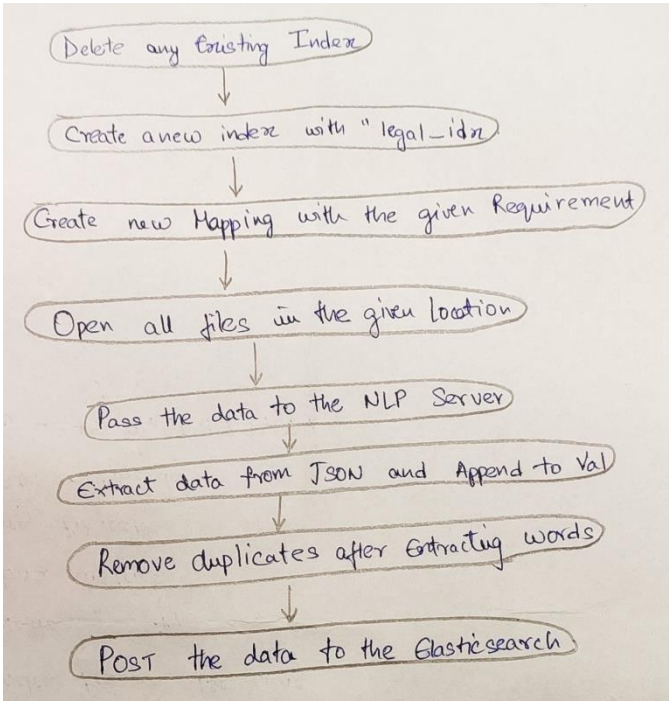
1. Persons: Text - List of all people in a given XML file
2. Location: Text - List of all locations in a given XML file
3. Organization: Text - List of all organizations in a given XML file
4. File: Text - The entire contents of the XML file

This mapping is considered so as to satisfy the index specific query and the general query which is satisfied by the File index.

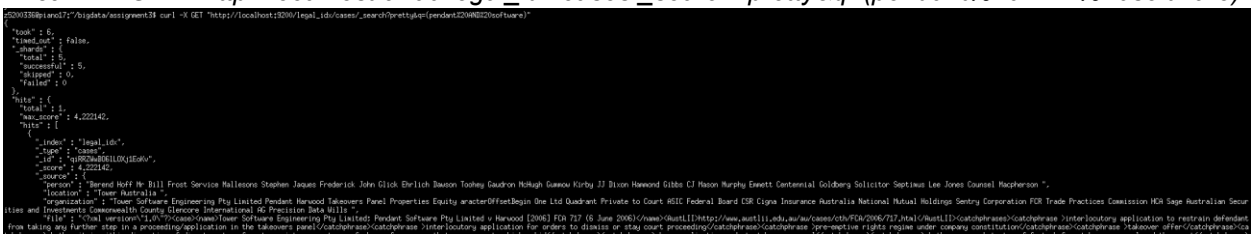
An important step before sending data to the Elasticsearch is cleaning the data of all escape sequences which are not allowed in JSON formatting.

The output JSON object from the NLP is parsed as a string to extract 3 NER values with their words, that is, Location, Person and Organization. After adding these words to their respective variables, the data is then sent to the Elasticsearch index as a POST request.

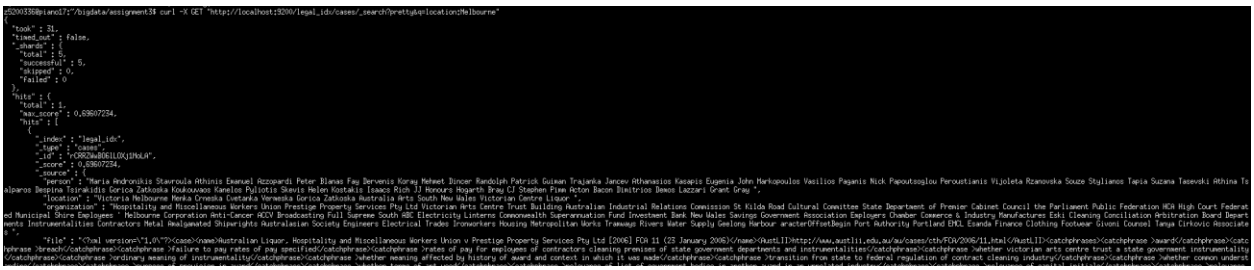
One entry in the Elasticsearch index is done for every file read. All file data are stored under the above mapping and can be viewed by the following specific or general queries.



1. curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=(pendant%20AND%20software)"

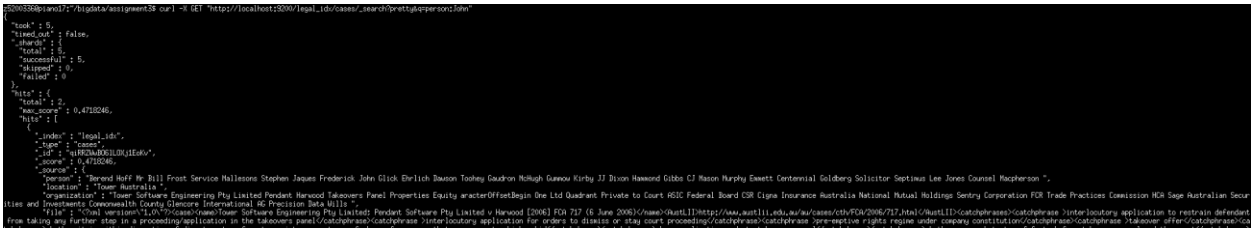


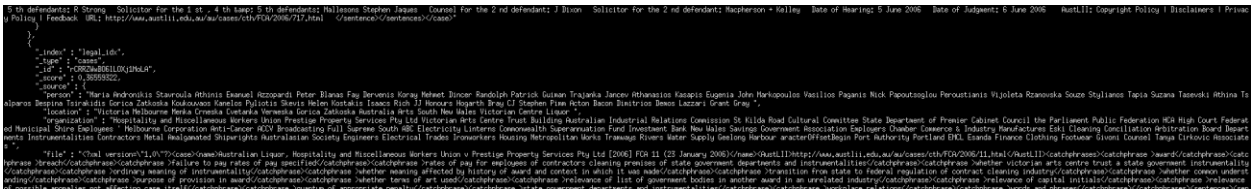
2. curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=location:Melbourne"



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

has 2 entries.





4. curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=organization:Workers%20Union"

has 2 entries