# COMP9313 Assignment3
# z5014567, Senlin Deng

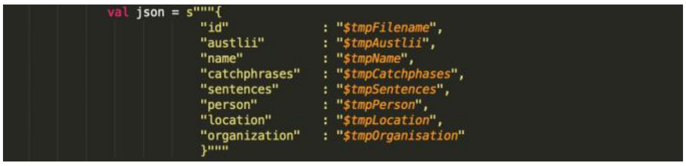
## Introduction

This assignment is built based on Spark-Submit. Elasticsearch technology is used to store legal case reports as documents in customised schema. To enrich the raw data, Named Entity Recognition API is called to get the characteristic of a word. And the entities will be further stored as attributes of an indexed document. All Elasticsearch and coreNLP calls are based on HTTP requests.

## Elasticsearch

Elasticsearch allows clients to retrieve information from big data storage. This engine helps users to do search in two scenarios, which are queried based on general terms and specific entities. The schema of the documented indices is designed to match the query engine.

To make the engine match the query requirements, each legal case reports in one XML is stored as one specific document in Elasticsearch. For each XML file, sentences and catchphrases are stored as attributes of documents in Elasticsearch. The structure of the documents shows as follow.



```
val json = s"""{
        "id"             : "$tmpFilename",
        "austlii"        : "$tmpAustlii",
        "name"           : "$tmpName",
        "catchphrases"   : "$tmpCatchphases",
        "sentences"      : "$tmpSentences",
        "person"         : "$tmpPerson",
        "location"       : "$tmpLocation",
        "organization"   : "$tmpOrganisation"
        }"""
```

All of the documents are stored in one single index called *legal_idx* and in one single type called *cases*. In each document, some of the attributes are stored without any pre-processing (e.g. id, URL and name). Other attributes are stored as a list a JSON string (e.g. person, location and organisation). Those data are trained using NLP server and words are given specific characteristic with Named Entity Recognition API.

## NLP

To enable the search engine support the query in the second scenario, NLP server is used to enrich the legal reports. To recognized whether a given term is a mention of one of the entity types of interest (i.e. person, location and organisation), Core NLP server helped us to identity the entities. The connections are build using HTTP requests.

## Query and Technology

There are two types of query in this assignment. One is based on entity type and another is based general query.

Example based on entity type:
```
$ curl -X GET "http://localhost:9200/legal_idx/cases/_search?pretty&q=location:Melbourne"
```

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

External packages are used to support the characteristic. *Scalaj* for the HTTP requests and *play json* to help deal with json files.

## How to complie:

```
$ spark-submit --packages org.scalaj:scalaj-http_2.11:2.4.2,com.typesafe.play:play-json_2.11:2.7.3 --class "CaseIndex" --master local[2] target/scala-2.11/caseindex_2.11-1.0.jar cases_test
```