ASSIGNMENT 1.3

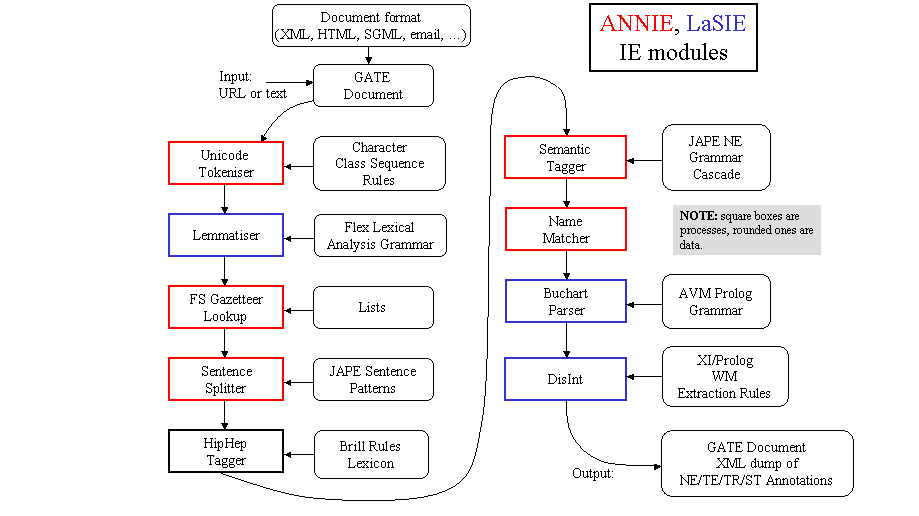
A) Get to know the information extraction pipeline of GATE

➢ Read the reference for ANNIE

(https://gate.ac.uk/sale/tao/splitch6.html, Sections 6.1-6.8)

➢ Briefly summarize the purpose of each component.

Solutions:-



ANNIE components :-

Tokenizer :-

* The tokeniser splits the text into very simple tokens such as numbers, punctuation and words of diﬀerent types.

Gazetteer Lookup:-

The gazetteer lists used are plain text ﬁles, with one entry per line. Each list represents a set of names, such as names of cities, organisations, days of the week, etc.

The role of the gazetteer is to identify entity names in the text based on lists

Sentence Splitter:-

The sentence splitter is a cascade of ﬁnite-state transducers which segments the text into sentences.

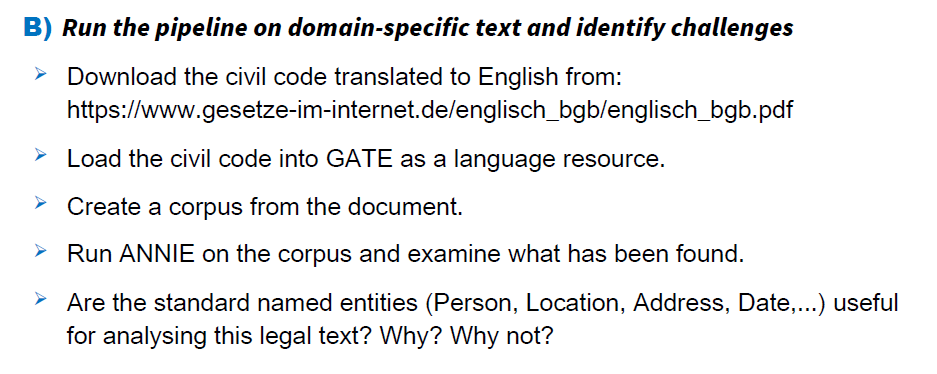
Each sentence is annotated with the type ‘Sentence’. Each sentence break (such as a full stop) is also given a ‘Split’ annotation.

Semantic Tagger :-

It contains rules which act on annotations assigned in earlier phases, in order to produce outputs of annotated entities. Ex: Person, Date , Address etc.

Name Matcher/OrthoMatcher :-

The Orthomatcher module adds identity relations between named entities found by the semantic tagger, in order to perform coreference. It does not ﬁnd new named entities as such, but it may assign a type to an unclassiﬁed proper name (an Unknown annotation), using the type of a matching name.



Solution:-

Standard Named Entities such as Person, Location, Address, Date may not be useful in the context Legal Text.

Legal Text mostly do not contain such entities. Even if so, they are provided in the following context:-

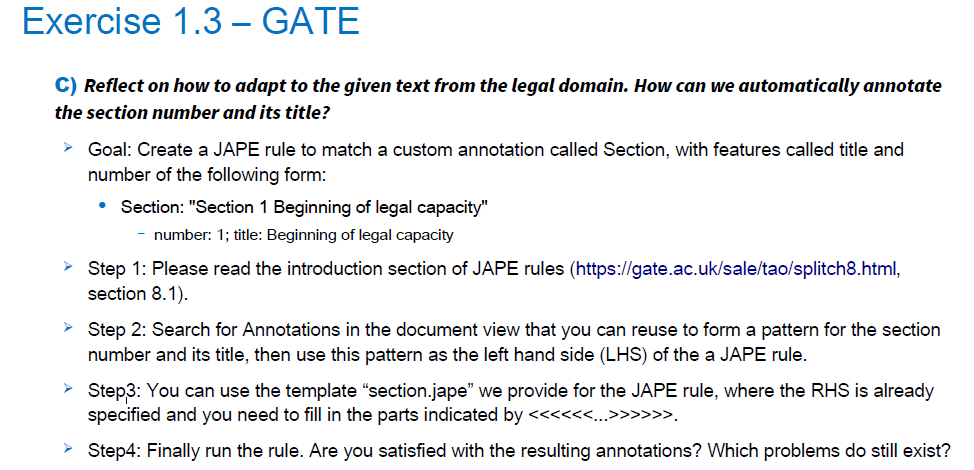
Person’s Name – As part of Reference to a Judgement- Wagner Vs State.

Date – Can also be as part of the Circular, Order, Act or Rule passed in a court of law or in Parliament- Date act as a Timestamp for such Documents

Location- May indicate Jurisdiction, Political Maps and Offices or Ministries.

Although these Entities help us identifying the above-mentioned details, their interpretation is different from the traditional sense.

A domain specific classification of these entities will help us understand them better.



My Jape code is able to identify only the Section and Number.

The Resulting rule does not identify the Subsection mentioned as 31a.