NLP ASSIGNMENT

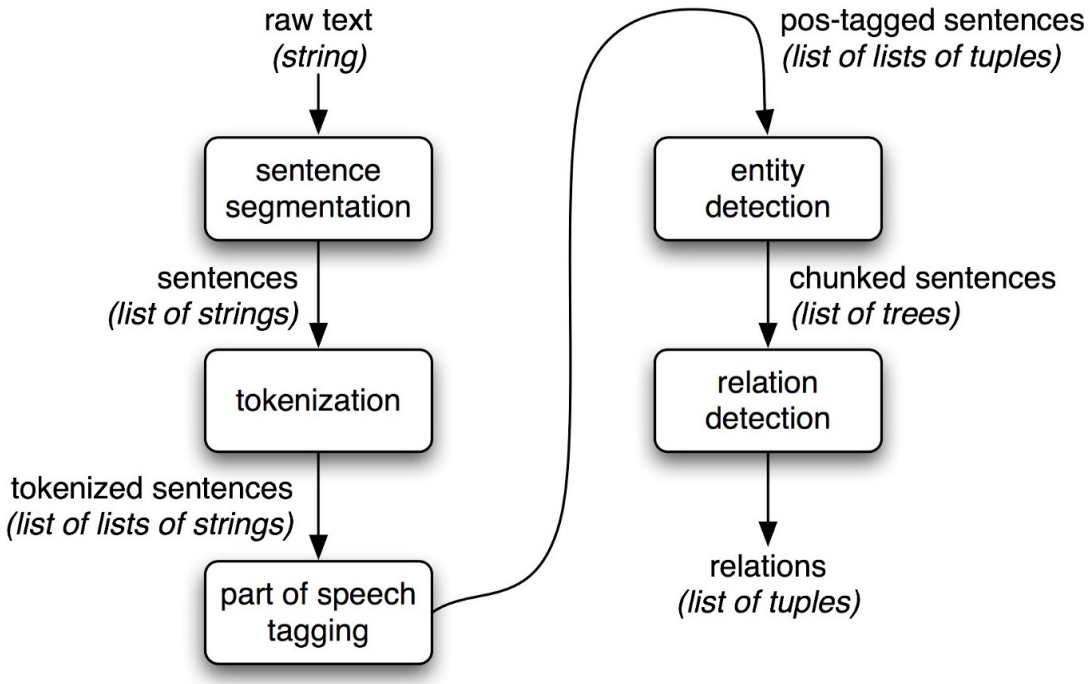
Submitted By: Anisha Das
Roll no:- UMDS20001
M.Tech DSA.

Named entity recognition is a process where the named entity gets identified and linked to its class. As we know that any given raw text data consists of various kinds of words like some of them are stopwords, part of speech words likewise there can be various kind words that can be presented in a text file which can be segregated as named entities. These words do not represent any feeling but they can represent the relationship between two sentences or two words.
Eg;
“Rahul sold his Maruti 800 at rupees 50000 in 2015”

And the named entity recognition system will give results as

“Rahul(person) sold his Maruti 800 (car/object) at rupees 50000 (price) in 2015 (time)”

Here in the sentence, we can see the recognition process of a NER model by classifying the words into the name of the person, car, prize and time.



NER deals with extracting the real-world entity from the text such as a person, an organisation, or an event. Named Entity Recognition is also simply known as entity identification, entity chunking, and entity extraction. They are quite similar to POS(part-of-speech) tags.

Program Link: https://github.com/anishadas5/NER-NLP

```
#Implementation of NER using NLTK
#Let's start with the importing library.
import nltk
from nltk.tokenize import word_tokenize
from nltk.tag import pos_tag

#NLTK provides some already tagged sentences, we can check it using the treebank package.
nltk.download('maxent_ne_chunker')
nltk.download('words')
nltk.download('treebank')
sent = nltk.corpus.treebank.tagged_sents()
print(nltk.ne_chunk(sent[0]))

#Information Extraction
raw_text="""The Board of Control for Cricket in India (BCCI) is the governing body for cricket in India and is under the jurisdiction of Ministry of Youth Affairs and Sports, Government of India.[2] The board was formed in December 1928 as a society, registered under the Tamil Nadu Societies Registration Act. It is a consortium of state cricket associations and the state associations select their representatives who in turn elect the BCCI Chief. Its headquarters are in Wankhede Stadium, Mumbai. Grant Govan was its first president and Anthony De Mello its first secretary. With the surge of cricket in India, BCCI was criticised for its monopolistic practices and has suffered from corruption allegations. The Supreme Court on 30 January 2017 nominated a four-member panel Committee of Administrators:- Vinod Rai, Ramachandra Guha, Vikaram Limaye and Diana Edulji to look after the administration of the BCCI in order to implement Lodha Committee reforms.Vinod Rai, ex-CAG of India heads the four members panel to look after the administrative duties of the board until the fresh elections are called.Presently, Sourav Ganguly is the president of BCCI.On 9 August 2019, the BCCI agreed to adhere to the anti-doping mechanisms governed by the National Anti-Doping Agency. Sunil Joshi, former Indian cricket team spinner was named as Chairman of the national selection panel by the Cricket Advisory Committee (CAC) of BCCI replacing MSK prasad in that role."""

#Before extracting the named entity we need to tokenize the sentence and give them part of the speech tag to the tokenized words.

nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
raw_words= word_tokenize(raw_text)
tags=pos_tag(raw_words)
```

```
#Now we'll implement noun phrase chunking to identify named entities using a regular expression consisting of rules that indicate how sentences should be chunked.
#Our chunk pattern consists of one rule, that a noun phrase, NP, should be formed whenever the chunker finds an optional determiner, DT, followed by any number of adjectives, JJ, and then a noun, NN.

nltk.download('maxent_ne_chunker')
nltk.download('words')
ne = nltk.ne_chunk(tags, binary=True)
print(ne)

#For better understanding, we can use the IOB tagging format. This format provides tags similar to the pos tagging but gives clarification about the position and the entity of the words.
# Here the IOB Tagging system contains tags of the form:

#B-{CHUNK_TYPE} – for the word in the Beginning chunk
#I-{CHUNK_TYPE} – for words Inside the chunk
#O – Outside any chunk

from nltk.chunk import tree2conlltags
iob = tree2conlltags(ne)
iob
```

NER is used extensively in biomedical data for gene identification, DNA identification, and also the identification of drug names and disease names. These experiments use CRFs with features engineered for their domain data.

NE Type	Examples
ORGANIZATION	Georgia-Pacific Corp., WHO
PERSON	Eddy Bonte, President Obama
LOCATION	Murray River, Mount Everest
DATE	June, 2008-06-29
TIME	two fifty a.m, 1:30 p.m.
MONEY	175 million Canadian Dollars, GBP 10.40
PERCENT	twenty pct, 18.75 %
FACILITY	Washington Monument, Stonehenge
GPE	South-East Asia, Midlothian
