Experiment No. 06

Aim -

Implement Named Entity Recognizer for the given text input.

Theory -

The named entity recognition (NER) is one of the most data preprocessing task. It involves the identification of key information in the text and classification into a set of predefined categories. An entity is basically the thing that is consistently talked about or referred to in the text.

NER is the form of NLP.

At its core, NLP is just a two-step process, below are the two steps that are involved:

- Detecting the entities from the text
- Classifying them into different categories

Some of the categories that are the most important architecture in NER such that:

- Person
- Organization
- Place/ location

Other common tasks include classifying of the following:

- date/time.
- Expression
- Numeral measurement (money, percent, weight, etc)
- E-mail address

Ambiguity in NE

For a person, the category definition is intuitively quite clear, but for computers, there is some ambiguity in classification. Let's look at some ambiguous example:

- *England (Organisation)* won the 2019 world cup vs The 2019 world cup happened in *England(Location)*.
- *Washington(Location)* is the capital of the US vs The first president of the US was *Washington(Person)*

Methods of NER

- One way is to train the model for multi-class classification using different machine learning algorithms, but it requires a lot of labelling. In addition to labelling the model also requires a deep understanding of context to deal with the ambiguity of the sentences. This makes it a challenging task for a simple machine learning algorithm.
- Another way is that Conditional random field that is implemented by both NLP Speech Tagger and NLTK. It is a probabilistic model that can be used to model sequential data

such as words. The CRF can capture a deep understanding of the context of the sentence. In this model, the input $\mathrm{X} = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots, \vec{x}_T\}$

$$\mathrm{p}\left(\mathrm{y}—\mathbf{x}\right) = \frac{1}{z(\vec{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{k} \omega_k f_k \left(y_t, y_{t-1}, \vec{x}_t\right) \right\}$$

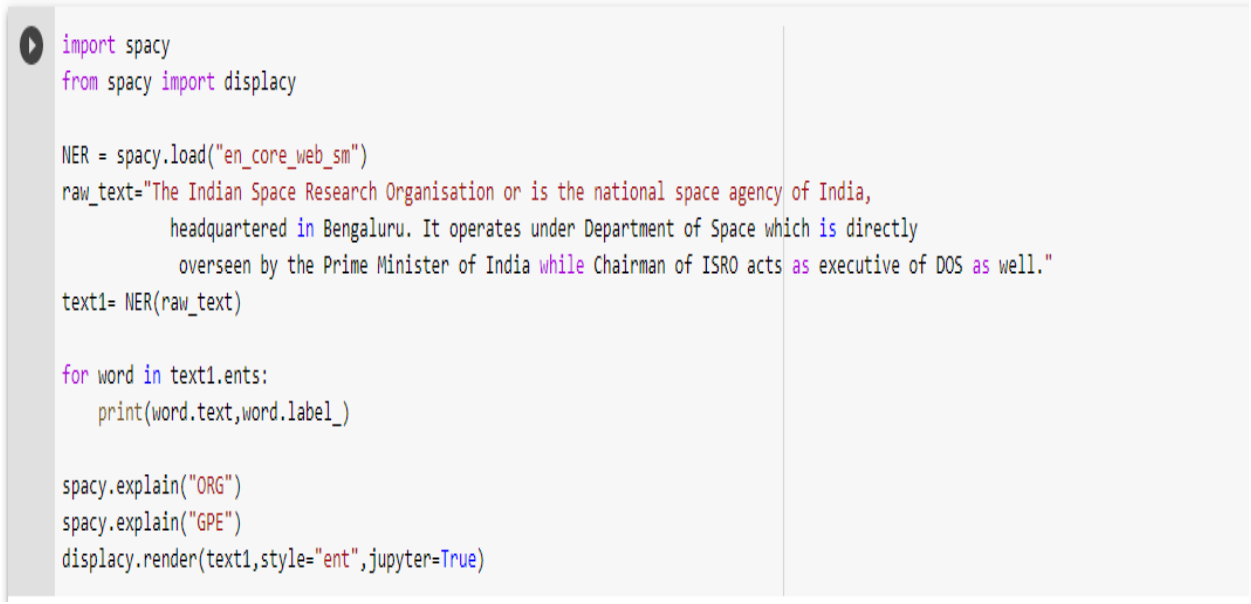
- **Deep Learning Based NER:** deep learning NER is much more accurate than previous method, as it is capable to assemble words. This is due to the fact that it used a method called word embedding, that is capable of understanding the semantic and syntactic relationship between various words. It is also able to learn analyzes topic-specific as well as high level words automatically. This makes deep learning NER applicable for performing multiple tasks. Deep learning can do most of the repetitive work itself, hence researchers for example can use their time more efficiently.

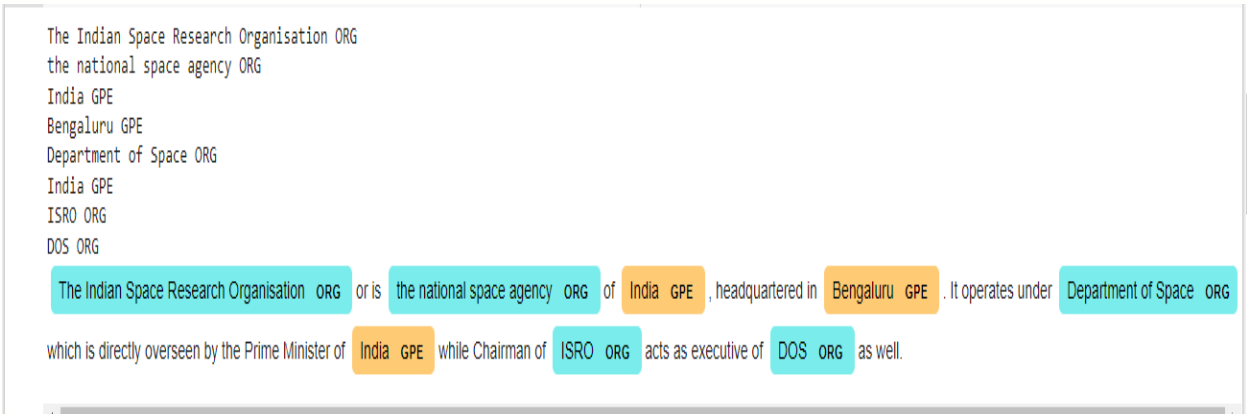
Code -

```
import spacy
from spacy import displacy

NER = spacy.load("en_core_web_sm")
raw_text="The Indian Space Research Organisation or is the national space agency of India,
headquartered in Bengaluru. It operates under Department of Space which is directly overseen by
the Prime Minister of India while Chairman of ISRO acts as executive of DOS as well."
text1= NER(raw_text)
for word in text1.ents:
    print(word.text,word.label_)
spacy.explain("ORG")
spacy.explain("GPE")
displacy.render(text1,style="ent",jupyter=True)
```

Output -





Conclusion -

Hence, the implementation of Named Entity Recognizer for the given text input is performed.