# Digital Assignment – 1

# NATURAL LANGUAGE PROCESSING

# CORPUS CREATION

**TEAM MEMBER - 1**

**NAME:** OMPRAKASH

**REG NO:** 21BCE1950

**TEAM MEMBER - 2**

**NAME :** MOHAMED RIYAAS

**REG NO:** 21BCE5828

## TITLE:
## NAMED ENTITY RECOGNITION (NER) MODEL:

## Introduction:

This project aims to develop a robust Named Entity Recognition (NER) model using traditional and advanced techniques to accurately identify and categorize entities like individuals, organizations, and locations in diverse textual contexts. NER is a fundamental Natural Language Processing (NLP) task that provides a structured understanding of unstructured text, enabling applications such as document summarization, sentiment analysis, and question answering. The project will explore various NER methods, including dictionary-based, rule-based, and machine learning-based approaches, to create a system that can effectively extract and classify key information from textual data. By leveraging the power of NER, this project seeks to unlock the potential of unstructured data and contribute to the advancement of information extraction and knowledge enrichment techniques.

**Example Workflow:** **Text**

**Input:**

"Apple Inc. was founded by Steve Jobs in Cupertino, California."

**NER Model Output:**

Identified entities and their types:

**Organization:** Apple Inc.

**Person:** Steve Jobs

**Location:** Cupertino, California

**Application:** The extracted entities can be used for various applications such as information extraction, question answering, sentiment analysis, and more.

Named Entity Recognition plays a crucial role in making sense of unstructured text and is a fundamental component in various natural language processing applications. The accuracy and performance of NER models depend on the quality and diversity of the training data, as well as the choice of the underlying model architecture.

## TOOLS/TECHNIQUES USED:

**SpaCy:** SpaCy is an open-source natural language processing (NLP) library designed for efficient and fast processing of natural language text. Developed by Explosion AI, SpaCy is written in Python and is widely used for various NLP tasks, including tokenization, part-of-speech tagging, named entity recognition, syntactic parsing, and more. It is known for its simplicity, speed, and state-of-the-art capabilities.

Here are **key features and components of spaCy:**

## 1. Tokenization:

spaCy's tokenization efficiently breaks down a text into individual tokens (words or subwords), taking into account language-specific rules.

## 2. Part-of-Speech (POS) Tagging:

It provides part-of-speech tagging, assigning grammatical categories (e.g., noun, verb, adjective) to each token.

## 3. Named Entity Recognition (NER):

spaCy includes pre-trained models for named entity recognition, allowing users to identify and classify entities such as persons, organizations, locations, etc., in the text.

## 4. Dependency Parsing:

Dependency parsing in spaCy analyzes the syntactic structure of sentences, determining the relationships between words and their dependencies.

## 5. Lemmatization:

Lemmatization is the process of reducing words to their base or root form. spaCy provides lemmatization capabilities.

## 6. Word Embeddings:

spaCy uses pre-trained word embeddings to represent words as vectors in a continuous vector space. This allows the model to capture semantic relationships between words.

7. **Rule-Based Matching:**

In addition to statistical models, spaCy allows for rule-based matching to identify patterns in the text using custom-defined rules.

8. **Entity Linking:** spaCy supports entity linking, which associates named entities in text with entries in a knowledge base.

9. **Multilingual Support:** spaCy supports multiple languages and offers pre-trained models for various languages.

10. **Customization and Training:**

Users can train custom models or fine-tune existing models for specific tasks using their own annotated datasets.

11. **Performance:** spaCy is known for its speed and efficiency. It is designed to be fast and is optimized for production use.

12. **Community and Documentation:**

spaCy has an active community, and its documentation is comprehensive, making it easy for users to get started and find information.

## RESULT:

The project of Named Entity Recognition (NER) using the BERT model has yielded promising results. By leveraging BERT's advanced contextual embeddings and fine tuning techniques, we have achieved state-of-the-art performance in identifying and classifying named entities in text.

## CONCLUSION:

In conclusion, a Named Entity Recognition (NER) model project holds significant importance in the realm of natural language processing. By effectively identifying and categorizing named entities within text, this project contributes to enhancing the understanding and organization of unstructured data. The successful implementation of an NER model involves meticulous data annotation, thorough preprocessing, and the utilization of appropriate machine learning or deep learning techniques.


## KEYWORDS:

Named Entity Recognition (NER),Natural Language Processing (NLP),Machine Learning,DataAnnotation,Entity Classification,SpaCy.


This corpus summarizes the research focus, methodology, results, and significance of the study, highlighting the key techniques and findings.