Abstract

In this project, we implement Named Entity Recognition (NER) using SpaCy, a robust NLP library. NER is crucial for Information Retrieval, enabling the extraction of entities such as names, locations, organizations, and monetary values from unstructured text. The project begins with the installation of SpaCy and the application of its pre-trained English model (en\_core\_web\_sm). We demonstrate entity recognition and visualization using SpaCy's ner pipeline and Doc object properties.

We cover methods for custom entity recognition by adding new entities not predefined in the SpaCy model through the Span class. Visualization techniques using the displacy library are illustrated, including customization options for enhancing visual representations of entities.

Advanced topics include accessing entity annotations, utilizing the IOB scheme for token-level annotations, and employing the PhraseMatcher for tagging multiple occurrences of specific phrases. Practical examples highlight the identification of geopolitical entities, monetary values, and organizational names, equipping users with the skills to customize NER for specific applications.