REN

TRANSFORMING TEXT INTO KNOWLEDGE

BY TEAM 32:-

SUJEETH REDDY K- SE21UECM027

JOSHITHA KORRAPATI-SE21UECM025

INTRODUCTION

Named Entity Recognition (NER) is a subfield of natural language processing (NLP) that focuses on identifying and classifying named entities in text. These entities are specific terms that represent real-world objects such as people, organizations, locations, dates, times, and other categories like quantities and so on. It is one of the applcation of Information Extraction

PROJECT OBJECTIVES

The objectives will often center around improving the accuracy and efficiency of the entity recognition process while also addressing specific application needs.

01. We are trying to develop a model that can perform well even with limited labelled data.

02. we are developing the model to scale to large text corpora and high-throughput data processing.

PIPENLINING

01

iMPORTING NLTK LIBRARY

TEXT PROCESSING INVOLVES;-

1. TOKENIZATON
2. PREPROCESSING
3. NORMALIZATION

02

ASSIGN PARTS OF SPEECH AND TAG THE TOKENS.
USE REGEXPPARSER FOR THE RESULT BY NLTK LIBRARY ONLY

03

IMPORT THE SPACY LIBRARY AND TREE BUILDER
MODEL AND TRAIN THE DATA WITH THE HELP Of TREE BUILDER
.USE THE TRAINED MODEL TO IDENTIFY THE ENTITIES

04

DEFINE THE ARTICLE FOR THE NER THROUGH URL.
USE BEAUTIFULSOUP LIBRARY FOR HTML/XML TEXT.LABEL AND COUNTER THE TOKENS

MODEL ARCHITECTURE

spaCy is a popular and modern NLP library that offers efficient and state-of-the-art NER models as part of its built-in NLP pipeline. It is based on machine learning techniques like neural networks.

spaCy integrates NER as part of its overall NLP pipeline, which includes tokenization, part-of-speech tagging, and dependency parsing. This integration allows the NER component to benefit from other NLP tasks and share context.

We even use NLTK library for the NER but NLTK cannot give precise Name recognition.

KEY COMPONENTS

1.NER is typically treated as a sequence labeling task, where each token in a sentence is assigned a label.

2. Most NER models use cross-entropy loss for training.

3. Pre-trained embeddings are often used as input

TIMELINE

4th MARCH

DATA COLLECTION
AND
PREPROCESSING

15TH APRIL

TRAINING OUR DATA

4TH MAY

WORK ON OUR
CODE

16TH MAY

TEST THE FINAL
CODE AND
MAKE CHANGEES.
IMPPORT
MAIN DATASET

RESULTS

We are done upto the work of NLTK library.Now, we are focusing on the machine learning techniques from spacy library. After the source code is completed. We are deciding to make this as a chatbot or as a model itself.

Thank you very much!