ACTIVITY#14

ANSWER#1

a. Named Entity Relationship (NER) is a subtask of natural language processing (NLP) that involves identifying and classifying named entities in text, such as names of persons, organizations, locations, dates, and other specific categories.

b. The most common named entity tags used in NER are:

PERSON: Individual people, including names.

ORGANIZATION: Companies, institutions, and organizations.

LOCATION: Places, such as cities, countries, and landmarks.

DATE: Specific dates or date ranges.

TIME: Specific times or time ranges.

MONEY: Monetary values.

PERCENT: Percentage values.

QUANTITY: Measurements or quantities.

ORDINAL: Words representing ranks or positions.

MISCELLANEOUS: Other named entities that do not fall into the above categories.

c. The main task of named entity recognition is to identify and classify named entities in text. It involves automatically detecting and classifying relevant entities, such as person names, organization names, and locations. The goal is to extract meaningful information from unstructured text and enable further analysis or information retrieval.

d. NER is challenging for several reasons:

Ambiguity: Many words in natural language can have multiple meanings, and determining whether a word is a named entity or a regular word can be ambiguous.

Contextual variations: Named entities can have different forms and variations, such as abbreviations, acronyms, misspellings, or different word orders, making it difficult to identify them accurately.

New entities: NER models need to handle new or unseen named entities that were not present in the training data, requiring generalization capabilities.

Domain-specific challenges: NER performance can vary across different domains due to domain-specific terminology or language patterns.

Language complexity: Different languages have unique challenges in named entity recognition, including morphological variations, word order differences, or lack of explicit entity markers.

Co-reference resolution: Resolving pronouns or referring expressions to their corresponding named entities is a complex task that often requires additional language understanding.

ANSWER#2

a. To turn the structured problem of named entity recognition into a sequence problem with one label per word, we can use BIO or IOB encoding. Each word in the input text is assigned a label indicating whether it is the beginning of a named entity (B), inside a named entity (I), or not part of a named entity (O). This way, we can treat NER as a sequence labeling task, similar to part-of-speech (POS) tagging, where each word is assigned a specific label.

b. Named entity recognition of the given text:

Jane Villanueva: PERSON

United: ORGANIZATION

United Airlines Holding: ORGANIZATION

Chicago: LOCATION

ANSWER#3

BIO tagging, also known as IOB tagging, is a labeling scheme commonly used in named entity recognition (NER). It assigns labels to each word in a sequence to indicate whether it is the beginning (B), inside (I), or outside (O) of a named entity.

Here's how BIO tagging works:

B: Indicates the beginning of a named entity. It is used for the first word of a multi-word entity or a single-word entity.

I: Indicates that a word is inside a named entity. It is used for subsequent words within a multi-word entity.

O: Indicates that a word is outside any named entity.

For example, let's consider the sentence: "Samia Jabbar works as a freelancer on Fiverr and Upwork"

Using BIO tagging, the named entities would be labeled as follows:

Samia: B-PERSON

Jabbar: I-PERSON

works: O

as: O

a: O

freelancer: O

on: O

Fiverr: B-ORGANIZATION

and: O

Upwork: B-ORGANIZATION

In this case, "Samia Jabbar" is the named entity representing a person, so "Samia" is labeled as B-PERSON (beginning of a person entity), and "Jabbar" is labeled as I-PERSON (inside a person entity). The named entities "Fiverr" and "Upwork" represent organizations, so they are labeled as B-ORGANIZATION (beginning of an organization entity). The other words in the sentence are labeled as O (outside any named entity).

ANSWER#4

Jane - B-PERSON

Villanueva - I-PERSON

of - O

United - B-ORGANIZATION

, - O

a - O

unit - O

of - O

United - B-ORGANIZATION

Airlines - I-ORGANIZATION

Holding - I-ORGANIZATION

, - O

said - O

the - O

fare - O

applies - O

to - O

the - O

Chicago - B-LOCATION

route – O

Using BIO tagging, the named entities in the given text are:

PERSON: Jane Villanueva

ORGANIZATION: United, United Airlines Holding

LOCATION: Chicago

ANSWER#5

The standard algorithms commonly used for named entity recognition (NER) include:

Rule-based approaches: These algorithms use handcrafted rules and patterns to identify named entities based on specific linguistic patterns, regular expressions, or dictionaries. Rule-based systems are relatively simple but require manual effort to define the rules.

Hidden Markov Models (HMMs): HMMs are probabilistic models that can be used for sequence labeling tasks like NER. They assign probabilities to different sequences of labels based on observed words. HMMs assume a Markov property where the current label depends only on the previous label.

Conditional Random Fields (CRFs): CRFs are another type of probabilistic model used for sequence labeling tasks. They take into account both observed words and contextual features to assign labels. CRFs can capture complex dependencies among labels and perform well in NER tasks.

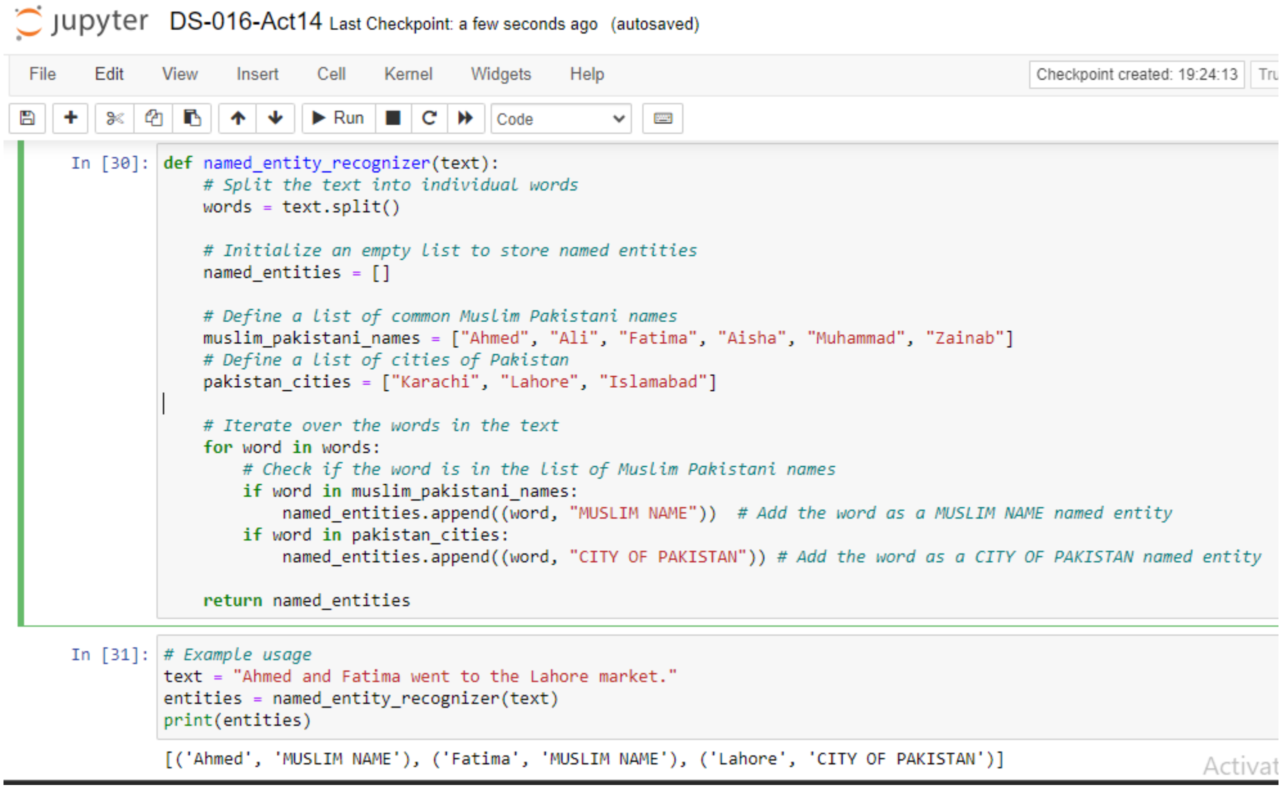
Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM): RNNs, particularly LSTM-based models, have been widely used for NER. These models can capture sequential information and dependencies in text effectively. They take word embeddings as input and predict the corresponding labels for each word.

Transformer-based models: Transformer-based architectures, such as BERT (Bidirectional Encoder Representations from Transformers) and its variants, have achieved state-of-the-art performance in NER. These models use self-attention mechanisms to capture contextual information and have significantly advanced the field of NER.

It's important to note that the choice of algorithm depends on factors like the available labeled data, computational resources, and specific requirements of the NER task.

ANSWER#6

I have already uploaded the jupyter file. Name: Ds-016-Act14


jupyter
DS-016-Act14 Last Checkpoint: a few seconds ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Checkpoint created: 19:24:13
Tru
Run
Code
Activat

In [30]:

```python
def named_entity_recognizer(text):
    # Split the text into individual words
    words = text.split()

    # Initialize an empty list to store named entities
    named_entities = []

    # Define a list of common Muslim Pakistani names
    muslim_pakistani_names = ["Ahmed", "Ali", "Fatima", "Aisha", "Muhammad", "Zainab"]
    # Define a list of cities of Pakistan
    pakistan_cities = ["Karachi", "Lahore", "Islamabad"]

    # Iterate over the words in the text
    for word in words:
        # Check if the word is in the list of Muslim Pakistani names
        if word in muslim_pakistani_names:
            named_entities.append((word, "MUSLIM NAME"))  # Add the word as a MUSLIM NAME named entity
        if word in pakistan_cities:
            named_entities.append((word, "CITY OF PAKISTAN")) # Add the word as a CITY OF PAKISTAN named entity

    return named_entities
```

In [31]:

```python
# Example usage
text = "Ahmed and Fatima went to the Lahore market."
entities = named_entity_recognizer(text)
print(entities)
```

```
[('Ahmed', 'MUSLIM NAME'), ('Fatima', 'MUSLIM NAME'), ('Lahore', 'CITY OF PAKISTAN')]
```