Course Name:	**Data Analysis Laboratory (216H03L501)**	**Semester:**	**V**
Date of Performance:	**27 / 10 / 2025**	**DIV/ Batch No:**	**HDA2**
Student Name:	**Aaryan Sharma**	**Roll No:**	**16010123012**

TITLE : NLP on clinical data (Finding data from Literature)

AIM: Implement NLP on clinical data

Expected Outcome of Experiment:

Books/ Journals/ Websites referred:
Sample case Study
https://towardsdatascience.com/clinical-named-entity-recognition-using-spacy-5ae9c002e86f

Theory:
Named entity recognition (NER) is a natural language processing (NLP) method that extracts information from text. NER involves detecting and categorizing important information in text known as named entities. Named entities refer to the key subjects of a piece of text, such as names, locations, companies, events and products, as well as themes, topics, times, monetary values and percentages.

NER is also referred to as entity extraction, chunking and identification. It's used in many fields in artificial intelligence (AI), including machine learning (ML), deep learning and neural networks. NER is a key component of NLP systems, such as chatbots, sentiment analysis tools and search engines. It's used in healthcare, finance, human resources (HR), customer support, higher education and social media analysis.

The purpose of NER
NER identifies, categorizes and extracts the most important pieces of information from unstructured text without requiring time-consuming human analysis. It's particularly useful for quickly extracting key information from large amounts of data because it automates the extraction process.

As NER models improve their ability to correctly identify important information, they are helping improve AI systems in general. These systems are enhancing AI language comprehension capabilities in areas such as summarization and translation systems and the ability of AI systems to analyze text.
NER uses algorithms that function based on grammar, statistical NLP models and predictive models. These algorithms are trained on data sets that people label with predefined named entity categories, such as people, locations, organizations, expressions, percentages and monetary values. Categories are identified with

abbreviations; for example, LOC is used for location, PER for persons and ORG for organizations.

Dataset:
https://www.mtsamples.com
https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions

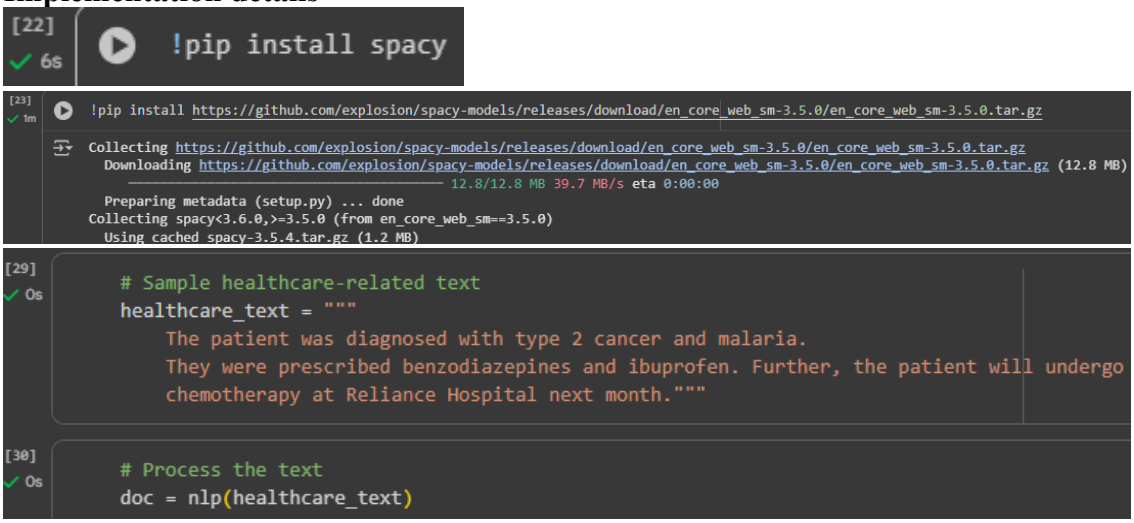
Language: C/C++/Java/Python/other – Choice of student

Steps: 1. Download dataset (Given).
2. Pre-process the dataset if needed.
3. Use spaCy library or any other similar one
4. Visualize and Display result

List out the library used with justification

- spaCy: Used for performing Named Entity Recognition (NER) and text preprocessing. It provides pre-trained NLP pipelines and models that support biomedical and clinical domains through extensions like scispaCy.
- pandas: Used for reading, cleaning, and organizing text data
- matplotlib: Used for visualization of entity frequency or category distribution in the extracted data.

Implementation details



[31] ✓ 0s

```
# Print named entities
print("Entities in the text:")
for ent in doc.ents:
    print(f"{ent.text} - {ent.label_}")
```

```
Entities in the text:
2 - CARDINAL
Reliance Hospital - ORG
next month - DATE
```

[32] ✓ 0s

```
# Print explanations of entity labels
print("\nEntity Label Explanations:")
for ent in doc.ents:
    print(f"{ent.text}: {spacy.explain(ent.label_)}")
```

```
Entity Label Explanations:
2: Numerals that do not fall under another type
Reliance Hospital: Companies, agencies, institutions, etc.
next month: Absolute or relative dates or periods
```

[38] ✓ 0s

```
# Define custom entity labels for diseases and drugs
disease_labels = ["type 2 cancer", "malaria"]  # Add more disease terms as needed
drug_labels = ["benzodiazepines", "ibuprofen"]  # Add more drug names as needed
```

[39] ✓ 0s

```
# Iterate through the text and identify entities based on the custom labels
entities = []
for label in disease_labels:
  if label in healthcare_text:
    entities.append((label, "DISEASE"))
for label in drug_labels:
  if label in healthcare_text:
    entities.append((label, "DRUG"))
```

[40] ✓ 0s

```
# Print the identified entities
print("Entities in the text:")
for entity, label in entities:
    print(f"{entity} - {label}")
```

```
Entities in the text:
type 2 cancer - DISEASE
malaria - DISEASE
benzodiazepines - DRUG
ibuprofen - DRUG
```

```
[44]
✓ 0s
import pandas as pd
entity_df = pd.DataFrame(entities, columns=['Entity', 'Label'])
display(entity_df)
```

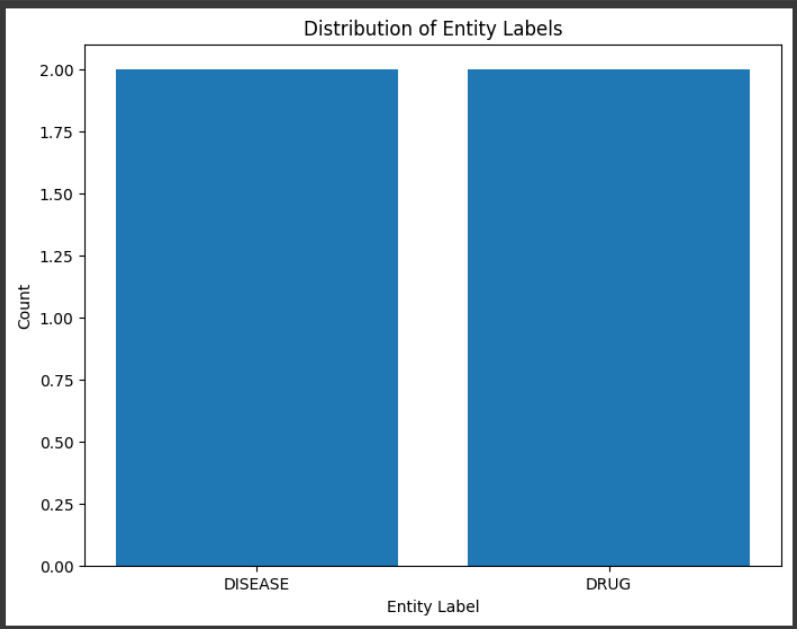
	Entity	Label
0	type 2 cancer	DISEASE
1	malaria	DISEASE
2	benzodiazepines	DRUG
3	ibuprofen	DRUG

```
[45]
✓ 0s
label_counts = entity_df['Label'].value_counts()
display(label_counts)
```

Label	count
DISEASE	2
DRUG	2

dtype: int64

```
[46]
import matplotlib.pyplot as plt
plt.figure(figsize=(8, 6))
plt.bar(label_counts.index, label_counts.values)
plt.title("Distribution of Entity Labels")
plt.xlabel("Entity Label")
plt.ylabel("Count")
plt.show()
```



Conclusion (Interpretation of result):

The experiment successfully demonstrates the use of Named Entity Recognition (NER) for extracting clinical entities from unstructured medical text. Using spaCy, we efficiently identified medical terms such as diseases, drugs, and treatments from clinical notes. This automation reduces manual data extraction effort and provides structured insights from large volumes of clinical documents. Such NLP-based techniques are highly valuable in clinical data analysis, medical research, and healthcare informatics.