**Assignment No: - 4**

**Name :- Isha Ghorpade**

**Enrollnment :- 22420020**

**Roll :- 381066**

---

**Problem Statement**

Build a Named Entity Recognition (NER) system for extracting entities from real-world text such as news articles or social media data. And measure its accuracy, precision, recall, and F1 score.

**Objectives**

- To understand the concept of Named Entity Recognition (NER)
- To extract entities such as PERSON, ORGANIZATION, LOCATION, DATE, etc.
- To implement NER using a pretrained NLP model
- To evaluate the NER system using standard performance metrics
- To analyze the effectiveness of NER on real-world text

---

**Software & Hardware Requirements**

**Software**

- Operating System: Windows / Linux / macOS
- Programming Language: Python 3.x
- Development Tool: Jupyter Notebook / Google Colab
- Libraries Used: spaCy, scikit-learn

**Hardware**

- Computer / Laptop
- Processor: Intel / AMD or equivalent
- Minimum 4 GB RAM
- Internet connection (for model download)

---

**Theory**

**Named Entity Recognition (NER)**

NER is a Natural Language Processing (NLP) technique used to identify and classify named entities in text into predefined categories such as:

- PERSON
- ORGANIZATION (ORG)
- LOCATION (GPE)
- DATE
- PRODUCT

**NER is widely used in:**

- News analysis
- Information extraction
- Question answering
- Social media monitoring

**Working Principle of NER**

NER systems work by analyzing:

- Word patterns
- Contextual information
- Grammar and sentence structure

Modern NER systems use machine learning and deep learning models trained on large annotated datasets. Pretrained models like those provided by spaCy can recognize entities without manual rule creation.

---

**Real-World Text for NER**

NER systems are commonly applied to:

- **News articles**, which contain formal and structured language
- **Social media text**, which may include hashtags, abbreviations, emojis, and informal expressions

NER performance is generally higher on news text and slightly lower on social media text due to informal language.

---

**Evaluation of NER Systems**

To measure the effectiveness of an NER system, standard evaluation metrics are used:

**Accuracy**

Accuracy measures the overall correctness of entity predictions.

$$Accuracy = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

---

**Precision**

Precision indicates how many of the predicted entities are actually correct.

$$Precision = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

---

**Recall**

Recall measures how many actual entities were successfully identified by the system.

$$Recall = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

---

**F1-Score**

F1-score provides a balance between precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

---

**Token-Level Evaluation**

NER evaluation is usually performed at the **token level**, where:

- Each word is assigned an entity label
- Words not belonging to any entity are labeled as **O (Outside)**

Token-level evaluation allows precise measurement of entity detection performance.

---

**Advantages of Using Pretrained NER Models**

- No need for manual feature engineering
- High accuracy on real-world text
- Faster development and deployment
- Supports multiple entity categories

---

**Limitations of NER**

- Performance drops for informal or noisy text
- May fail for unseen or rare entity names
- Domain-specific entities may require custom training

---

**Sample Output (Typical)**

```
Accuracy: 0.90

              precision    recall  f1-score

         GPE       0.89      0.93      0.91
         ORG       0.86      0.82      0.84
      PERSON       0.92      0.95      0.93
     PRODUCT       0.85      0.80      0.82
           O       0.93      0.91      0.92

   macro avg       0.89      0.89      0.89
```

**Interpretation of Results**

- High Precision → Model predicts entities accurately
- High Recall → Most actual entities are detected
- F1-Score balances precision and recall
- spaCy performs well on news-style text
- Performance may slightly drop for informal social media text

---

**Conclusion**

Named Entity Recognition is a powerful NLP technique used to extract meaningful entities from unstructured text. By applying NER to news and social media data and evaluating it using accuracy, precision, recall, and F1-score, we can measure its effectiveness. Pretrained NER models provide efficient and accurate solutions for real-world information extraction tasks.