Build and evaluate a NER (Named Entitiy Recognition) system using NER libraries

In [1]:
```
# !python -m spacy download en_core_web_sm
```

In [2]:
```
import spacy
from sklearn.metrics import precision_score, recall_score, f1_score
```

In [3]:
```
nlp = spacy.load('en_core_web_sm')
```

In [4]:
```
# text = "Tony Stark is the CEO of Stark Industries based in New York."
```

In [5]:
```
def extract_entities(text):
    doc = nlp(text)
    return [(ent.text, ent.label_) for ent in doc.ents], doc

def evaluate_entities(predicted_entities, true_entities):
    # Lowercase everything for fair comparison
    true_entities = [entity.lower() for entity in true_entities]
    predicted_entities = [entity.lower() for entity in predicted_entities]

    # Binary labels
    y_true = [1] * len(true_entities)
    y_pred = [1 if entity in predicted_entities else 0 for entity in true_entities]

    # Calculate metrics
    precision = precision_score(y_true, y_pred, zero_division=0)
    recall = recall_score(y_true, y_pred, zero_division=0)
    f1 = f1_score(y_true, y_pred, zero_division=0)
    return precision, recall, f1
```

In [6]:
```
text = "Tony Stark is the CEO of Stark Industries based in New York."
true_entities = ["Tony Stark", "Stark Industries", "New York"]
```

In [7]:
```
predicted_entities, doc = extract_entities(text)
```

In [8]:
```
print("\nNamed Entities:")
for entity in predicted_entities:
    print(f"{entity[0]} ({entity[1]})")
```

```
Named Entities:
Tony Stark (PERSON)
Stark Industries (ORG)
New York (GPE)
```

In [9]:
```
# Evaluate the extracted entities
predicted_entity_texts = [ent[0] for ent in predicted_entities]
precision, recall, f1 = evaluate_entities(predicted_entity_texts, true_entities)
```

In [10]:
```
# Output evaluation metrics
print("\n--------------------------")
print("Evaluation Metrics:")
print(f"Precision: {precision:.2f}")
print(f"Recall: {recall:.2f}")
print(f"F1 Score: {f1:.2f}")
print("--------------------------")
```

```
---------------------------
Evaluation Metrics:
Precision: 1.00
Recall: 1.00
F1 Score: 1.00
---------------------------
```

In [11]:
```python
from spacy import displacy

displacy.render(doc, style="ent")
```

Tony Stark PERSON is the CEO of Stark Industries ORG based in New York GPE .

In [12]:
```python
# Tokenization and BoW
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import CountVectorizer
import nltk
nltk.download('punkt')

tokens = word_tokenize(text)
print("Tokens:", tokens)

vectorizer = CountVectorizer()
X = vectorizer.fit_transform([text])
print("Vocabulary:", vectorizer.get_feature_names_out())
print("BoW Matrix:", X.toarray())
```

```
Tokens: ['Tony', 'Stark', 'is', 'the', 'CEO', 'of', 'Stark', 'Industries', 'based', 'in', 'Ne
w', 'York', '.']
Vocabulary: ['based' 'ceo' 'in' 'industries' 'is' 'new' 'of' 'stark' 'the' 'tony'
 'york']
BoW Matrix: [[1 1 1 1 1 1 1 2 1 1 1]]
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\ASUS\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```