Take-Home Exercise: Named Entity Recognition (NER) in News Articles

Objective:

Develop a Named Entity Recognition (NER) system to identify and classify entities such as names of people, organizations, locations, expressions of times, quantities, monetary values, percentages, etc., from a collection of news articles.

Task Description:

1. Data Preparation:
    a. Use a provided dataset of news articles. This dataset should contain a diverse range of articles from various news sources and topics.
    b. Perform necessary data cleaning and text preprocessing.
2. Model Development:
    a. Implement an NER model using any NLP techniques or frameworks you prefer (e.g., CRF, SpaCy, BERT-based models).
    b. Explain the choice of the model and its architecture.
3. Model Evaluation:
    a. Evaluate the model using appropriate metrics (e.g., Precision, Recall, F1-Score at the entity level).
    b. Provide analysis on the model's performance across different entity types.
4. Error Analysis:
    a. Conduct an error analysis to identify the types of errors the model is making.
    b. Suggest potential improvements or strategies to address these errors.
5. Code Quality and Documentation:
    a. Ensure the code is well-organized, documented, and easily runnable.
    b. Include a README file with instructions on how to set up and run the project, and any necessary environment setup.
6. Bonus (Optional):
    a. Extend the model to handle entity disambiguation or linking entities to a knowledge base (e.g., Wikipedia).

b. Explore and implement advanced techniques or models to improve performance.

Sample Data Set:

CNN News Data Set from data.world
https://data.world/opensnippets/cnn-news-dataset

Deliverables:

- Source code.
- A detailed report documenting the approach, evaluation results, and error analysis.
- (If attempted) Additional features or advanced implementations.