Assignment 1

Assignment: Building a Named Entity Recognition (NER) System for Financial Text

Assignment Overview

The financial industry relies heavily on timely and accurate information from various unstructured text sources like news articles, earnings reports, and analyst opinions. Named Entity Recognition (NER) plays a crucial role in extracting meaningful entities such as company names, stock symbols, events, and key financial figures. Your task is to design and implement a domain-specific NER system tailored for financial text.

Assignment Objectives

1. Understand the challenges and opportunities in applying NER to financial data.
2. Explore methods to preprocess financial text and identify domain-specific entities.
3. Build, evaluate, and deploy an NER system to extract meaningful insights from financial documents.
4. Present your findings and demonstrate the system's capabilities with real-world examples.

Deliverables

1. **Problem Analysis (10% of grade):**

   - Write a brief overview of why NER is important for financial text.
   - Identify key challenges in implementing NER in the financial domain.
   - Define the scope of entities to be extracted (e.g., companies, events, financial figures).

2. **Data Collection and Preprocessing (20% of grade):**

   - Collect a dataset of financial text from at least two sources (e.g., news websites, financial reports, or public datasets like SEC filings).
   - Preprocess the data to handle:

     - Tokenization
     - Removal of irrelevant characters or noise
     - Standardization of financial abbreviations (e.g., "USD" for U.S. dollars).

   - Document your preprocessing steps and rationale.

3. **Model Development (40% of grade):**

   - Use a pre-trained NER model (e.g., SpaCy, BERT, or FinBERT) as a baseline.
   - Fine-tune or customize the model with annotated financial text data.
   - Define and train your model to recognize domain-specific entities such as:

- Companies
- Stock symbols
- Financial events (e.g., mergers, bankruptcies, IPOs)
- Key figures (e.g., revenue, stock price changes).
  - Evaluate the model using metrics like precision, recall, and F1-score.

4. **Visualization and Deployment (20% of grade):**

   - Create a visualization to showcase the extracted entities in the context of the text.
   - Build a simple interactive tool (using Streamlit, Flask, or similar frameworks) where users can input text and see the extracted entities.
   - Document how the system could be deployed in a real-world setting, such as integration into an API or a dashboard.

5. **Presentation and Documentation (10% of grade):**

   - Prepare a report summarizing your methodology, challenges, results, and system performance.
   - Include screenshots or examples of the system in action.
   - Present your findings in a clear and concise manner (e.g., as a slide deck or recorded presentation).

Key Evaluation Criteria

- **Innovation:** How well did you tailor the system to handle the challenges of financial text?
- **Accuracy:** The performance of your NER system based on evaluation metrics.
- **Practicality:** How useful and scalable is your solution for real-world applications?
- **Presentation:** Clarity and professionalism in reporting and demonstrating your solution.

Optional Advanced Challenges

- Implement **entity linking** to associate recognized entities with financial databases (e.g., linking "Apple" to its stock ticker AAPL).
- Include **sentiment analysis** to evaluate the tone of financial text in relation to recognized entities.
- Extend the system to handle **multi-lingual financial text**.

Submission Instructions

- Submit the following by **22nd January**:
  1. **Codebase:** Upload your complete codebase to a GitHub repository. Include a `README.md` file with clear instructions to run the project.
  2. **Report/Slide Deck:** Attach a detailed PDF report or slide deck summarizing your project (methodology, results, and findings).
  3. **Interactive Tool (Optional):** If applicable, provide a link or executable file for your interactive tool.

- Email your submission to **ojhagaurav36@gmail.com** with the subject line:
  "NER for Financial Text Submission - [Your Name]"

Email Contents

In your email, include:

1. A brief overview of your project.

2. Links to the GitHub repository and any hosted tools (if applicable).

3. Attachments for your report/slide deck (if not hosted in the GitHub repository).