NER EVALUATION STEPS

1. Data preparation

This step can be further divided into two processes

- **Data exploration**
  Understanding the characteristics of data, format, and quality to get better insights.
- **Data pre-processing**
  It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next phase.
  - **Steps**
    1. Read the news_sample_ner text file and convert it to HTML format to extract the needed information.
    2. Clean text by removing html tags and extract each paragraph separately with its entities and tags.
    3. The output consists of three lists, list of paragraphs which contains all clean paragraphs, list of paragraph words "splitting each paragraph into list of words", and list of tags which contains the ground truth for each word in specific paragraph.

- **Libraries: BeautifulSoup , html2text, NLTK**

2. Entities Extraction

Extracting entities using two approaches

- **Statistical approach**
  Using **Stanford Named Entity Recognition** model as a statistical model which uses the Conditional Random Field (CRF) model.
  Source https://nlp.stanford.edu/software/CRF-NER.html

- **Transformer Based**
  Using Spacy-Transformer package which provides spaCy model pipelines that wrap Hugging Face's transformers package.
  Source https://spacy.io/universe/project/spacy-transformers

3. Evaluation

Evaluating two models using eval4ner library which is a Python toolkit of MUC-5 evaluation metrics for evaluating Named Entity Recognition (NER) results.
Source https://github.com/cyk1337/eval4ner

- **Stanford NER Evaluation**

Mode	Precision	Recall	F1 Score
strict	0.5681	0.6184	0.5846
exact	0.5760	0.6238	0.5905
partial	0.6535	0.7679	0.6885
type	0.7231	0.9064	0.7804

- **Transformer Based Evaluation**

Mode	Precision	Recall	F1 Score
strict	0.7688	0.7561	0.7599
exact	0.8092	0.7933	0.7981
partial	0.8765	0.8608	0.8654
type	0.9034	0.8910	0.8945