Named Entity Recognition

About the Dataset

File ner_dataset.csv maps sentence number (Sentence #) to first word (Word) of respective sentence. Then each word is mapped to its respective POS tag (POS) and NER tag (Tag). Ignore the POS tag column for scope of this case-study. For every sentence just consider the word and its NER tag.

Sample mapping of a sentence:

Today	*O*
Micheal	*B-PER*
Jackson	*I-PER*
and	*O*
Mark	*B-PER*
ate	*O*
lasagna	*O*
at	*O*
New	*B-geo*
Delhi	*I-geo*
.	*O*

Sequence tagging scheme: IOB2

- I : inside : word is inside a chunk
- O : outside : word belongs to no chunk
- B : beginning : word is the beginning of a chunk

Columns:

- **Sentences #** : sentence number
- **Word** : word to be classified
- **POS** : POS tags for respective word
- **Tag** : NER tags for respective word

Probable tasks (Below pointers are for direction purpose only):

- divide the dataset into 3 parts:
  - train
  - validation
  - test (at least 20%)
- Identify the matrices for evaluating model's performance.
- Pre-process the data such that words of each sentence is mapped to their respective NER tags.
- Develop a baseline model which takes a sentence (list of words) as input and predicts NER tag for each word in that sentence.
- Identify the short comings of the baseline model.
- Develop a new model which overcome the shortcomings of baseline model.
- Identify future scope to further optimise the model.

System design tasks (Below pointers are for direction purpose only):

- Design system architecture to deploy ML Model in production
- How do you perform canary build?
- What should be the strategy for ML Model Monitoring?
- How do you perform load and stress testing?
- How do you track, monitor and audit ML training?
- Design framework for continuous delivery and automation of machine learning tasks.

Deliverables:

- Jupyter notebook (or equivalent) showcasing your work
- Powerpoint presentation clearly explaining the approach and findings.
- System design architecture (if applicable) and explanations.