Subject Description:

In this project, we will be utilizing Natural Language Processing (NLP) techniques to extract precise information from a given text. Our task involves working with documents that contain valuable information about a company's contracts and its customers.

The primary objective of this project is to develop a system that can automatically extract specific pieces of information from these documents. The information we are interested in extracting includes:

- Last Name and First Name of the managers
- Birth date
- CIN number or residence card
- Address

By implementing these NLP techniques, we aim to automate the process of extracting specific information from the company's documents. This will not only save time but also improve the accuracy and efficiency of the information extraction process. Throughout the project, we will explore various NLP libraries, algorithms, and methodologies to achieve our goal of precise information extraction.

here is an example of given text :

Les parts sociales sont indivisibles a l'égard de la société qui ne reconnaît qu'un seul propriétaire pour chacune d'elles. Les copropriétaires indivis sont tenus de désigner l'un d'entre eux pour les représenter auprés de la société ; 4 défaut d'entente, il appartient a la partie la plus diligente de faire désigner par justice un mandataire chargé de les représenter. L'usufruitier représente valablement le nu-propriétaire a l'égard de la société.

TITRE TROISIEME
ADMINISTRATION ET CONTROLE DE LA SOCIETE
ARTICLE 15 : NOMINATION, DUREE ET POUVOIRS DE LA GERANCE

La société est administrée par un ou plusieurs gérants, personnes physiques, associés ou non.
La société sera gérée par :

Mr. KHALID BEN TAGUAR
Nationalité Marocaine
Né le 06.09.1988 à Rabat
Demeurant à Quartier Al Inbiat CP 12000 Temara
Carte de séjour N° KA 214701M

Mr. MOSTAPHA FOUDSI
Nationalité Marocaine
Né le 06.09.1992 à Agadir
Demeurant à 3 AvAhmed Elmansour Eddahbi Cité Dakhla CP 80000 Agadir
CIN SY137932

So as result we expected to have:

KHALID BEN TAGUAR
06.09.1988
Quartier AI Inbiat CP 1200 Temara
KA214701M

MOSTAPHA FOUDSI
06.09.1992
3 AvAhmed Elmansour Eddahbi Cité Dakhla CP 80000 Agadir
SY137932

Objectif:

Your mission is to develop a model based on Natural Language Processing (NLP) that will be able to extract this information from text. To do this, you must use NLP techniques, in particular **Named Entity Recognition** (**NER**).

Data Description:

We have provided you with a structured dataset including the following:

- Training data
- Validation Data
- Test Data

Data in Dataset is labeled according to the following schema:

- For the first and last Names of the managers: labeled with "PERSON"
- For CIN numbers or residence cards: labeled with "cin"
- For Birth date : labeled with "date"
- For addresses: labeled with "LOC"
- For text parts other than names, CIN numbers or residence cards and addresses: labeled with "O"

Here is an example of labeled Data:

```
CIN O
JP302392 B-cin

Né O
le O
28.09.1962 B-date
à O
Boujad O

Mme O
OUSMANE B-PERSON
EL I-PERSON
AISSAOUI I-PERSON

Démeurant O
à O
: O
Quartier B-LOC
Douar I-LOC
Jdid I-LOC
Fes I-LOC

Carte O
de O
séjour O
N° O
VK B-cin
291255 I-cin
```

Explanation of the above image:

1. The **sentences** in the corpus are separated by an empty line.
2. Each row(line) has tow columns. The first column is the word, the second column denotes the BIO-annotated NER tag.
3. B- (Begin): This tag is used to mark the beginning of a named entity. It is assigned to the first token of a named entity.
4. I- (Inside): This tag is used to mark the subsequent tokens within a named entity. It is assigned to all tokens that come after the first token of a named entity.

Here's an example to illustrate how these tags are used in NER data:
Sentence: "Mr. KHALID BEN TAGUAR . Né le 30.07.1950’"
Tokenized: ["Mr.", "KHALID ", "BEN ", "TAGUAR ", ".","Né","le","30.07.1950"]
NER Tags: ["O", "B-PERSON", "I-PERSON", "I-PERSON", "O", "O", "O","B-date"]

In this example, "KHALID BEN TAGUAR" is recognized as a named entity of type PERSON.

Therefore, the first token "KHALID" is labeled as B-PERSON (beginning of person name), and the second token "BEN" is labeled as I-PERSON (inside person name) and so on.....

By using the B- and I- tags in NER data, it becomes possible to identify the boundaries of named entities and distinguish them from other tokens in the text.

Note that if you use a model that uses another labeling schema, you can customize the form of labeling (it's easy with a script) while keeping the same label names ("PERSON", "cin", "date" , "LOC", "O").

Evaluation Measure:

Your task is to develop an NER model using the provided data and deliver the following:

- A model file (.pt) that represents the best-performing model, achieving the highest results.
- A notebook (.ipynb) containing your code sequence and the results obtained on the test set. The notebook should display the Precision, Recall, and F1-score metrics, as depicted in the image below.

```
Results:
- F-score (micro) 1.0
- F-score (macro) 1.0
- Accuracy 1.0

By class:
              precision    recall  f1-score   support

      PERSON     1.0000    1.0000    1.0000        30
         LOC     1.0000    1.0000    1.0000        30
         cin     1.0000    1.0000    1.0000        30
        date     1.0000    1.0000    1.0000        30

   micro avg     1.0000    1.0000    1.0000       120
   macro avg     1.0000    1.0000    1.0000       120
weighted avg     1.0000    1.0000    1.0000       120
 samples avg     1.0000    1.0000    1.0000       120
```

References:

https://towardsdatascience.com/benchmark-ner-algorithm-d4ab01b2d4c3
https://blog.vsoftconsulting.com/blog/understanding-named-entity-recognition-pre-trained-models
https://www.analyticsvidhya.com/blog/2021/11/a-beginners-introduction-to-ner-named-entity-recognition/
https://paperswithcode.com/task/named-entity-recognition-ner
https://medium.com/@b.terryjack/nlp-pretrained-named-entity-recognition-7caa5cd28d7b
https://www.analyticsvidhya.com/blog/2021/06/nlp-application-named-entity-recognition-ner-in-python-with-spacy/
https://www.analyticsvidhya.com/blog/2021/06/nlp-application-named-entity-recognition-ner-in-python-with-spacy/
https://towardsdatascience.com/named-entity-recognition-with-bert-in-pytorch-a454405e0b6a