Objectif NER tagging
Sur des documents scannés travailler sur le texte brut pour extraire des informations bine découpées, la mettre en bdd puis l'automatiser
A partir de texte d'annonces immobilières => qu'on reconnaisse des caractéristiques claires
BUT : entrainer un modèle de NER tagging a qui on donne une annonce et qui redonne les informations souhaitées
Méthodes possibles : construire un classifier qui balaie tout le texte de l'annonce et qui va sortir la distribution pour les différentes entités
Appartement (AUTRE) à (AUTRE) Courbevoie (VILLE) de 26(SURFACE) m2
Sinon Deep learning pour améliorer en efficacité (si possible)

Pour entrainer ce classifieur, il faut une base d'entrainement. On va créer le jeu d'apprentissage en catégorisant les mots

Utilisation de pandas car on bosse avec des DF (une ligne = un mot) + quels features pour le mots (majuscule ? Composition uniquement de chiffres ? Présence de mot clé (genre charge copra etc etc) ? Etc..)
=> Matrice X de mots et des features associés (une seule matrice pour tous les mots)

Pour Doccano, il faut surligner QUE les chiffres si on a un mot tq « 3 pieces » MAIS si on a F3 il faut surligner l'ensemble. Si on a 15m2 on surligne la totalité et si on a 15 m2 (donc avec un espace) on surligne UNIQUEMENT le 15.

Extraction données doccano : format dl jzon (?) qui faudra coder pour le lire en python

Pour les features, est- il possible que le mot d'après corresponde à un certain type de mot (selectionner le mot précédent pour faire un feature) ? Oui, l'appartenance a une liste de mot donnée est valable + possibilité de considérer la distance de Livenstein (?) En cas de fautes de frappe par exemple