Détection des doublons

Dans ce mini-projet, vous disposez de deux fichiers CSV contenant des informations sur des produits qui sont en vente sur deux plateformes web, et qui sont caractérisés par des attributs différents.

- Le fichier ./Data/Company1.csv : contient des informations sur les produits vendus sur la première plate-forme web, et
- Le fichier ./Data/Company2.csv contient des informations sur les produits vendus dans une deuxième plate-forme.

Les deux fichiers utilisent des identificateurs et des attributs différents pour caractériser les produits.

L'objectif du mini-projet est de développer un programme python capable d'identifier les produits qui sont vendus sur les deux plates-formes, en d'autres termes, détecter les doublons.

Afin de pouvoir estimer l'efficacité de votre programme en termes de précision et de rappel, le fichier ./Data/Ground_truth_mappings.csv spécifie les doublons existants.

A titre d'exemple, vous trouverez le notebook (sous répertoire ./ExempleDuCours(Restaurants)) que nous avons vu pendant le cours pour la détection des enregistrements qui représentent le même restaurant. Ce fichier est à titre d'information uniquement. Il s'agit de données différentes, et les règles et distances à utiliser pour identifier les doublons se tissent donc différemment.

Vous trouverez également dans le répertoire le notebook que nous avons vu dans le cours et qui présente les principales fonctions de la bibliothèque NLTK de python que vous pouvez utiliser pour calculer les distances.

J'ai également mis dans le répertoire un sous-répertoire /SampleData qui contient un petit échantillon de données, que vous pouvez utiliser pour débugger votre programme avant de l'exécuter sur les données complètes du sous-répertoire ./Data. Cela peut vous faire gagner du temps.

Vous pouvez travailler par binôme, et vous avez 8 semaines pour renvoyer le TP, c'est-à-dire avant le 16 avril.

Vous avez le choix entre :
- Renvoyer un notebook jupyter documenté (éventuellement sur colab), ou
- Un fichier Python + Un fichier PDF documentant ce que vous avez fait et vos conclusions.

Dans les deux cas, votre mail devra avoir comme objet « [IASD02] Nom1+NOM2 » où Nom1 et Nom2 représentent les noms des membres du binôme.