

## Projet BDA :

Le projet de web scraping que nous avons réalisé visait à extraire des informations sur les formateurs à partir d'un fichier CSV contenant des données de formations. En utilisant Selenium pour le scraping web, nous avons nettoyé les données extraites et les avons intégrées dans une base de données Oracle XE. Ce document détaille les étapes du projet, les technologies utilisées, et les défis rencontrés durant le processus.

### 1. Objectifs du Projet

Le principal objectif de ce projet était de créer une base de données de formateurs, enrichie par des informations trouvées sur Internet. Plus précisément, nous avons :

- Extrait des données de formateurs à partir de plusieurs sites web.
- Nettoyé et structuré ces données pour une utilisation optimale.
- Stocké les informations dans une base de données Oracle XE.
- Développé une interface graphique avec Streamlit pour faciliter la recherche des formations par formateurs.

### 2. Technologies Utilisées

Le projet a été réalisé en utilisant plusieurs technologies et bibliothèques Python, notamment :

- **Selenium** : Pour l'automatisation du web scraping. Cette bibliothèque nous a permis d'interagir avec les sites web et d'extraire les données de manière dynamique.
- **Pandas** : Utilisé pour le traitement et le nettoyage des données. Pandas a facilité la manipulation des données tabulaires et leur préparation pour l'insertion dans la base de données.
- **cx\_Oracle** : Pour se connecter à la base de données Oracle XE. Cette bibliothèque a permis d'effectuer des opérations CRUD (Créer, Lire, Mettre à jour, Supprimer) sur la base de données.
- **Streamlit** : Pour le développement de l'interface graphique. Streamlit a été choisi pour sa simplicité et sa capacité à créer rapidement des applications web interactives.

## 3. Étapes du Projet

### 3.1. Préparation des Données

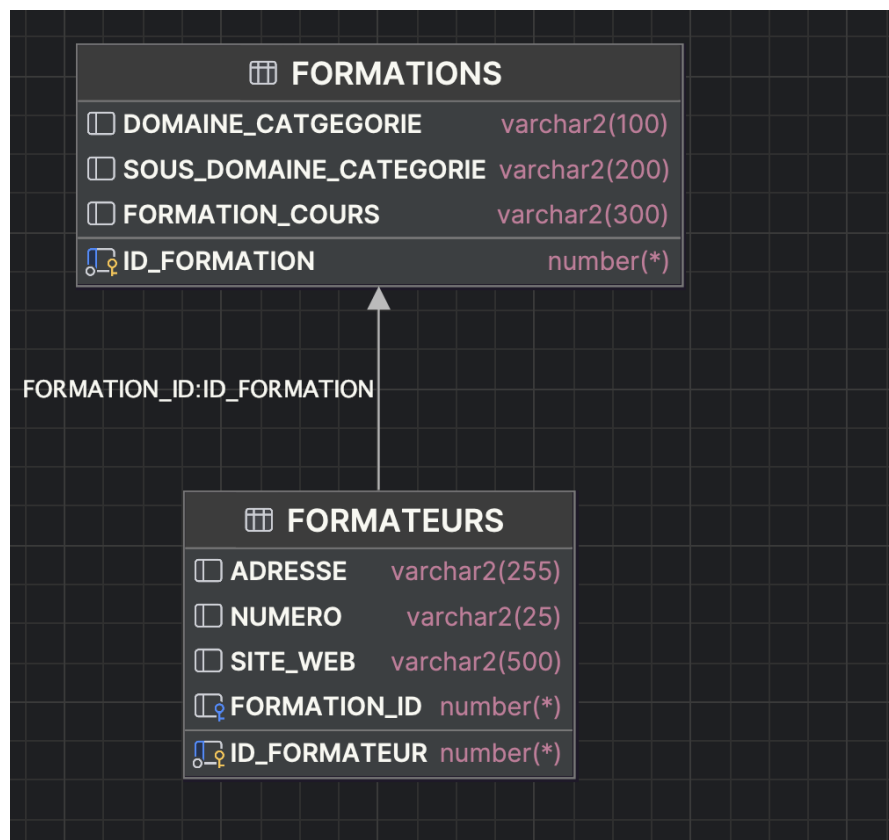
La première étape consistait à insérer le fichier CSV contenant les données de formations dans la table FORMATIONS de la base de données Oracle XE. Ce fichier servait de point de départ pour l'extraction des formateurs.

### 3.2. Web Scraping

Utilisant Selenium, nous avons conçu un script pour parcourir les sites web et extraire les informations des formateurs associés aux formations. Ce processus a nécessité l'identification des éléments HTML pertinents pour accéder aux données souhaitées. Le scraping a été exécuté sur plusieurs heures, prenant environ 4 heures pour extraire les informations de 1000 formations, soit 5 à 7 formateurs par formation.

### 3.3. Intégration dans la Base de Données

Les données nettoyées ont ensuite été insérées dans la table formateurs de la base de données Oracle XE via la bibliothèque cx\_Oracle. Cela a permis d'enrichir notre base de données avec des informations fiables sur les formateurs.



### 3.4. Recherche Avancée

Pour améliorer l'expérience utilisateur, nous avons implémenté des algorithmes de recherche avancée. Ces algorithmes, basés sur les enseignements acquis pendant les tps, ont permis aux utilisateurs de trouver rapidement des formations spécifiques associées à des formateurs particuliers.

```
create PROCEDURE rechercher_formation_avancee(
    p_chaine VARCHAR2,
    cat VARCHAR2,
    p_result OUT SYS_REFCURSOR,
    touscat NUMBER -- 1 for TRUE, 0 for FALSE
) AS
    v_pattern VARCHAR2(4000);
BEGIN
    v_pattern := REPLACE(p_chaine, ' ', '|');

    OPEN p_result FOR
        SELECT ID_FORMATION,
               DOMAINE_CATEGORIE,
               SOUS_DOMAINE_CATEGORIE,
               FORMATION_COURS,
               UTL_MATCH.edit_distance_similarity(UPPER(FORMATION_COURS), UPPER(p_chaine)) AS
edit_distance,
               UTL_MATCH.jaro_winkler_similarity(UPPER(FORMATION_COURS), UPPER(p_chaine)) AS
jaro_winkler,
               CASE
                   WHEN LOWER(FORMATION_COURS) = LOWER(p_chaine) THEN 3
                   WHEN LOWER(FORMATION_COURS) LIKE '%' || LOWER(p_chaine) || '%' THEN 2
                   WHEN REGEXP_LIKE(LOWER(FORMATION_COURS), LOWER(v_pattern)) THEN 1
                   ELSE 0
               END
pertinence
        FROM formations
        WHERE (
            (LOWER(FORMATION_COURS) LIKE '%' || LOWER(p_chaine) || '%' OR
              LOWER(DOMAINE_CATEGORIE) LIKE '%' || LOWER(p_chaine) || '%' OR
              LOWER(SOUS_DOMAINE_CATEGORIE) LIKE '%' || LOWER(p_chaine) || '%')
            OR
            (REGEXP_LIKE(LOWER(FORMATION_COURS), LOWER(v_pattern)) OR
              REGEXP_LIKE(LOWER(DOMAINE_CATEGORIE), LOWER(v_pattern)) OR
              REGEXP_LIKE(LOWER(SOUS_DOMAINE_CATEGORIE), LOWER(v_pattern)))
        )
        AND (DOMAINE_CATEGORIE = cat OR touscat = 1)
        AND (DOMAINE_CATEGORIE = cat OR 1 = touscat)
        ORDER BY pertinence DESC,
               CASE
                   WHEN edit_distance <= 2 THEN edit_distance
                   ELSE NULL
               END NULLS LAST,
               jaro_winkler DESC,
               ID_FORMATION ASC
        FETCH FIRST 10 ROWS ONLY;
END;
```


### 3.5. Développement de l'Interface Graphique

Finalement, une interface graphique a été développée avec Streamlit. Cette application permet aux utilisateurs de rechercher des formations en utilisant différents critères,

tels que le nom du formateur ou le type de formation. L'interface est intuitive et responsive, offrant une expérience utilisateur fluide.

## SeFormer

Categories


Tous les categories  Commercial - Ventes Communication Comptabilité - Fiscalité Achats Assistant(e) Banque Bureautique - PAO/CAO Changement Coaching

Contrôle de gestion Création d'entreprise Direction de l'entreprise Efficacité professionnelle Finance - Trésorerie **Droit des affaires** Droit social Développement personnel International

Logistique - Supply chain Management Formation de formateurs Formations en anglais Gestion du temps IA - Digital Immo Informatique SI Innovation créativité Marketing

RSE - Dev. Durable Relation client Organisation - Audit Paie/Admin. Production - Lean Projet QVT Qualité-Santé-Sécurité-Env Secteur public Soft skills Ressources humaines

Travail à distance Web

PowerPoint - Débutant - Logiciels bureautique 

**Belformation PowerPoint - Débutant - Logiciels bureautique**

📍 46 Rue du Faubourg Saint-Martin, 75010 Paris

[Visitez le site web](#)

**Ellipse Centre de Formation - Location de Salles - Coworking PowerPoint - Débutant - Logiciels bureautique**

📍 8 Cité Joly, 75011 Paris

## 4. Difficultés Rencontrées

L'un des principaux défis rencontrés lors de ce projet a été le temps nécessaire pour le scraping. Le processus a pris un temps considérable (environ 4 heures pour 1000 formations), ce qui a nécessité une gestion efficace des ressources et une planification minutieuse pour s'assurer que le projet respecte les délais impartis. De plus, nous avons dû faire face à des problèmes liés à l'accessibilité de certaines données sur les sites web, ce qui a nécessité des ajustements dans notre approche de scraping.

Executed at 2024.10.30 04:01:33 in 3h 54m 26s