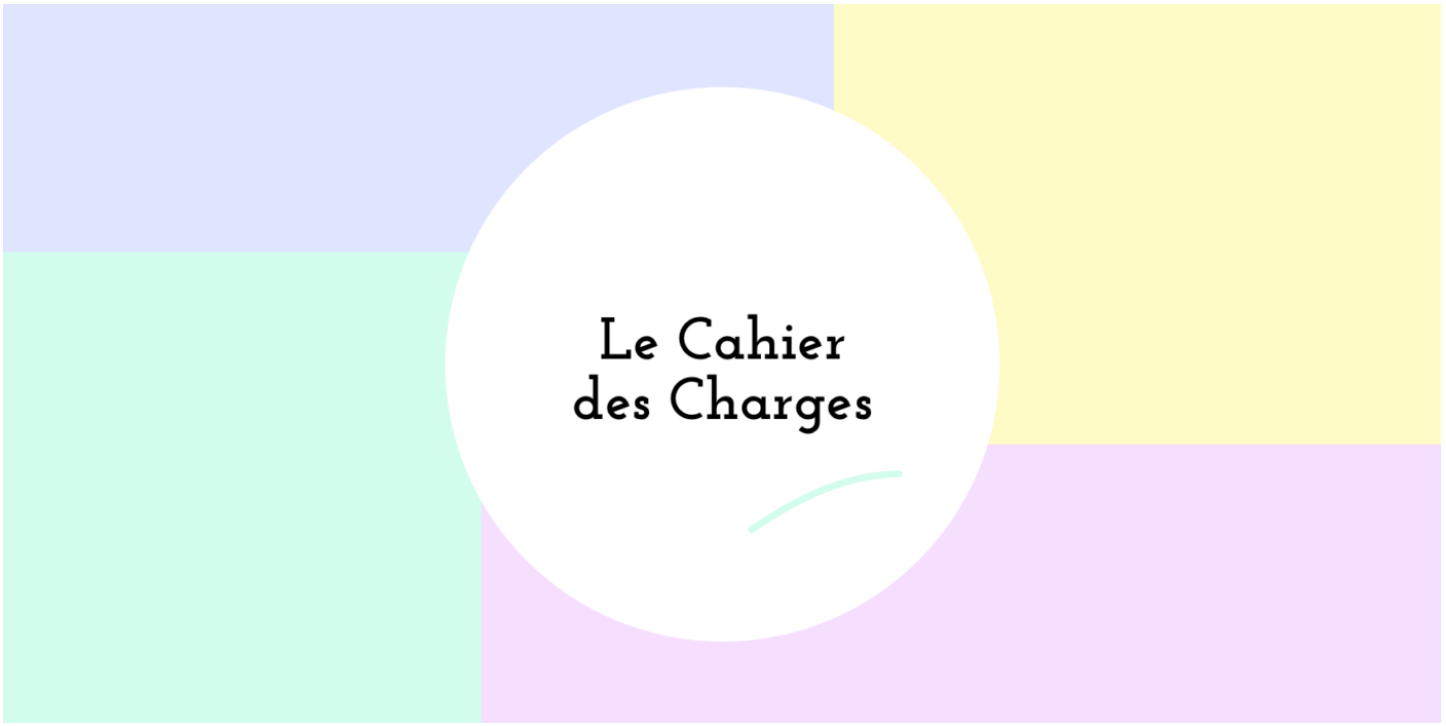

Master 2 EID2
Responsable du cours : Dr M.F Boufarès

Équipe du projet

- BASLAM Ismail
- SAIDI Anis
- TERTAKI Anis
- OZDEMIR Sedanur
- LACHKAR El Hassan
- BOUDJEMA Bilal
- BELKHOUMALI Ahmed Abd El Mouaim

Introduction

Le projet Big Forma vise à simplifier la comparaison des formations en ligne en rassemblant des informations essentielles telles que les tarifs, la qualité du contenu, la durée et les avis des utilisateurs. L'objectif est de créer un site fiable pour aider les étudiants et les professionnels à explorer les nombreuses opportunités d'apprentissage disponibles sur différentes plateformes éducatives.

Objectifs du projet

Collecte des données :

- Implémentation d'un scraper pour extraire les informations pertinentes des sites de formations en ligne.
- Acquisition de données telles que les tarifs, la durée, la qualité du contenu et les avis des utilisateurs.

Centralisation des données :

- Création d'une base de données centralisée pour stocker toutes les informations collectées.
- Structuration de la base de données pour faciliter la gestion et l'analyse ultérieures.

Nettoyage et normalisation des données :

- Nettoyer les données, éliminer les erreurs et standardiser les formats.
- Identification et correction des incohérences dans les données.

Recherche avancée des formations similaires :

- Mise en place d'algorithmes avancés pour détecter les similitudes entre les formations, par exemple en utilisant des abréviations ou d'autres critères pertinents.

Suppression des doublons :

- Identifier et éliminer les doublons dans la base de données.
- Assurer l'intégrité des données en éliminant les entrées redondantes.

Optimisation de la base de données :

- Création d'une table de métadonnées pour gérer la base de données, incluant des informations sur chaque colonne, les dates de collecte, etc.

Outils et technologies utilisées

- Langages de programmation : Python, PL/SQL
- Utilisation des bibliothèques beautifulsoup, Selenium
- Outils et logiciels : VS code, Jupyter Notebook, Excel, Oracle SQL Developer, Git/Github, Trello, SQL Live

Difficultés et contraintes rencontrées dans le projet

Problème de temps :

- La collecte de données a pris du temps.

Problème de sécurité des sites :

- Certains sites ont bloqué l'accès ou ont utilisé des captchas (détection de robots d'extraction de données), nécessitant des solutions pour éviter les interruptions.

Problème de diversité des structures de données :

- Chaque site a sa propre manière d'organiser les données, compliquant la tâche d'unifier tout dans une seule base.

Problème d'homogénéisation des données :

- Les différences de formats et de termes ont rendu difficile la normalisation des données.
- La diversité des langues des sites de formations (français / anglais).

Problème de recherche :

- Les méthodes avancées de recherche n'ont pas toujours donné des résultats aussi précis qu'attendu.
- L'ordre d'affichage des résultats obtenus.
- la recherche par abréviations n’a pas été implémenté

Fonctionnalités du Projet (Tâches Faites)

- L’extraction des données
- La modélisation de la base de données
- La création de la base de données
- Les métadonnées
- La fonction de la recherche avancée
- L'application web de la plateforme
- La documentation du projet (Repo Github)

Note : Chaque tâche mentionnée ci-dessus fera l'objet d'améliorations continues (modification quotidienne).

Tâches restantes

- L'amélioration de la fonction de recherche avancée.
- L'hébergement de l'application web sur le cloud.
- Intégration de la fonctionnalité Real-time data.
- Mettre à jour le scraping de manière consécutive (afin d'extraire plus de données et de propriétés).
- Implémenter le partitionnement des données pour améliorer la complexité de la recherche.
- Créer des champs lexicaux pour les formations et/ou sous-formations et combiner cela avec le partitionnement pour avoir une complexité de recherche optimale.
- Recherche par langue.
- Recherche par abréviation.