Rapport de Nettoyage des Données - Projet de Scraping Github

Date de réalisation: 19/06/23 - 23/06/23

Introduction :

Le présent rapport décrit le processus de nettoyage des données collectées à partir de Github. L'objectif de ce processus était de garantir la qualité et la fiabilité des données extraites afin de les rendre prêtes à être utilisées pour des analyses ultérieures. Ce rapport détaille les problèmes identifiés, les techniques de nettoyage appliquées et les résultats obtenus.

Problèmes Identifiés :

Lors de l'analyse initiale des données collectées, les problèmes suivants ont été identifiés :
Valeurs Manquantes : Les colonnes 'description' et 'language' contenaient des valeurs manquantes, ce qui aurait pu affecter l'intégrité des données et la validité des analyses ultérieures.
Valeurs Extrêmes : Des valeurs extrêmes ont été détectées dans certaines variables, ces valeurs extrêmes ont été visualisées.
Caractères Spéciaux et Emojis : La colonne 'description' contenait des caractères spéciaux et des emojis, ce qui pourrait rendre l'analyse du texte plus difficile et introduire du bruit dans les données.
Langues Multiples dans la Colonne de Description : La colonne 'description' contenait des descriptions dans différentes langues, ce qui pourrait nécessiter une traduction pour une analyse homogène des données.

Techniques de Nettoyage Appliquées :

Pour résoudre les problèmes identifiés, les techniques de nettoyage suivantes ont été appliquées :
Gestion des Valeurs Manquantes : Les valeurs manquantes dans la colonne 'description' ont été remplacées par la phrase "There is no description" à l'aide de la méthode fillna() de pandas.
Les lignes avec les valeurs manquantes dans la colonne 'language' ont été supprimées à l'aide de la méthode dropna() de pandas
Traitement des Valeurs Extrêmes : Aucune solution spécifique n'a été appliquée pour les traiter. Cependant, en identifiant les valeurs extrêmes, les analystes peuvent prendre des décisions éclairées sur la manière de les gérer en fonction du contexte et des objectifs de leur analyse ultérieure.
Suppression des Caractères Spéciaux et des Emojis : Une fonction "clean_description" a été utilisée pour nettoyer le texte de la colonne 'description'. Cette fonction utilise des expressions régulières pour supprimer les caractères spéciaux, les emojis et les caractères non alphanumériques. Ainsi, les descriptions sont rendues plus lisibles et cohérentes.
Traduction des Descriptions : Pour les descriptions dans des langues autres que l'anglais, on a détecté la langue de chaque description et les traduire en anglais. Cela permet d'homogénéiser les données et de faciliter leur compréhension et leur analyse.

Résultats Obtenu :

Après l'application des techniques de nettoyage, les résultats suivants ont été obtenus :
-Les valeurs manquantes dans la colonne 'description' ont été remplacées par la phrase "There is no description". Les lignes contenant des valeurs manquantes dans la colonne 'language' ont été supprimées, garantissant que seules les entrées avec une langue spécifiée sont conservées.
-Les caractères spéciaux, les emojis et les caractères non alphanumériques ont été supprimés des descriptions, rendant le texte plus lisible et cohérent.
-Les descriptions dans différentes langues ont été traduites en anglais, permettant une analyse homogène des données.

Code de nettoyage :

Le processus de nettoyage des données a été effectué en utilisant Python. Le code pour le nettoyage des données, ainsi que les détails de sa mise en œuvre, peuvent être consultés sur mon dépôt GitHub. Ce

code comprend les étapes nécessaires pour nettoyer les données collectées à partir de GitHub. Il aborde les problèmes tels que les valeurs manquantes, la traduction...
Vous pouvez accéder au code pour le nettoyage des données sur mon dépôt GitHub à Faissal-00/Cleaning_up_scraping_data_on_Github

Conclusion :

Le processus de nettoyage des données collectées à partir de Github a permis de résoudre les problèmes identifiés et de rendre les données prêtes à être utilisées pour des analyses ultérieures. Les techniques de nettoyage appliquées ont permis de garantir la qualité et la fiabilité des données extraites. Le jeu de données nettoyé peut maintenant être utilisé en toute confiance pour des analyses approfondies et des modèles.