DM MA431
Investigation de logs

11 novembre 2022

Introduction

Un serveur web montre des signes de dysfonctionnement. À l'aide de vos connaissances en réseau, en sécurité et en manipulation de la donnée, décrivez l'activité journalisée par ce serveur.

Votre compte-rendu prendra la forme d'un notebook Jupyter. Chaque cellule doit pouvoir s'exécuter sans erreur (l'évaluation se fera sur les résultats d'une ré-exécution). Vous devez utiliser Pandas et/ou Seaborn pour présenter vos résultats, mais vous pouvez utiliser pour certains de vos calculs intermédiaires Numpy, Scipy et Scikit-learn. Les autres modules sont fortement déconseillés.

1 Extraction des données

Les données fournies sont dans un format texte semi-formaté, le format **Apache Access Log**. Dans cette partie, vous devez analyser le fichier pour trouver les champs à extraire, puis employer une expression régulière pour les intégrer à un DataFrame. Vous fournirez pour chaque champ une courte description de celui-ci et commenterez l'extraction.

2 Description des données

Dans cette partie, vous commencerez par décrire la distribution des données de chaque colonne de votre DataFrame. Vous chercherez ensuite des corrélations entre les colonnes de données en les visualisant avec Seaborn. Soignez la lisibilité de vos visualisations (vous pouvez par exemple restreindre l'ensemble des données affichées si cela rend le graphique plus lisible).

N'oubliez pas les titres et les légendes dans vos résultats, cela fait partie de la note finale.

3 Vectorisation des données

Dans cette partie, vous donnerez puis appliquerez une stratégie de vectorisation, c'est-à-dire une stratégie pour transformer toutes vos données en données numériques normalisées.
Vous pouvez utiliser les modules `preprocessing` et `feature_extraction` de Scikit-learn pour y parvenir, ou bien extraire et décompter des informations spécifiques contenues dans vos colonnes : par exemple, le nombre de slashs de la ressource accédée.

4 Réduction de dimensions

Dans cette partie, vous appliquerez un algorithme de réduction de dimensions pour obtenir un jeu de données à deux ou trois colonnes numériques. Vous avez le choix de l'algorithme entre ACP, MDS, t-SNE ou UMAP. Vous donnerez le plan de projection et la variance expliquée du modèle lorsque le modèle le permet. Dans tous les cas vous évaluerez la pertinence de vos résultats.

5 Catégorisation automatique

Dans cette partie, vous appliquerez un algorithme de catégorisation (*clustering*) sur les données réduites. Vous avez le choix de l'algorithme entre DBSCAN, k-Means ou k-NN. Justifiez ce choix, et interprétez vos résultats dans l'optique de décrire des profils de requêtes HTTP.