SAE-15 : Traiter les données

synopsis :

La société **Huile de Code SARL** (http://huiledecode.org) qui gère le site Internet de la société **ControlTower** (http://controltower.fr), ce dernier réalisant des ventes de disques vyniles et dont les clients sont internationaux, vous demande de réaliser un logiciel en Python 3.

Huile de Code vous demande de **réaliser des statistiques** à partir des données de connexion au site Internet. Pour ce faire, la société **HDC** vous **met à disposition un fichier de log du serveur Apache 2** sur lequel est hébergé le site Internet de **ControlTower**.

Vous pouver télécharger le fichier de log à l'adresse Internet suivante :

http://eismall.otf.cloud/iut/controltower_access.zip

Travail à réaliser :

1 Analyse du fichier de log apache 2

Vous devrez dans un premier temps réaliser une analyse des données du fichier de log, pour comprendre sa structure. Vous devrez donc effectuer des recherches sur Internet pour savoir quelles données sont stockées dans ce fichier.

Vous pourrez faire un tableau permettant de voir rapidement à quoi correspondent les informations qui sont stockées dans ce fichier.

2 Quelles statistiques peut-on faire sur ces données ?

- Connaître l'adresse IP et la géolocaliser via un API en ligne : '165.225.76.120' via https://ip-api.com (vous chercherez comment utiliser ce service depuis un programme Phyton)
- Afficher une courbe du nombre de requêtes en fonction de la date de chaque ligne : "[09/Nov/2021:11:36:21', '+0100]"
- Connaître et dénombrer les systèmes d'exploitation qui se connectent sur le serveur : "(Windows', 'NT', '10.0;', 'Win64;', 'x64)"
- Connaître et dénombrer les navigateurs Internet qui se connectent sur le serveur : "Chrome/87.0.4280.141"
- Connaître le nombre d'erreurs de réponse en fonction du code de retour HTTP : "200", "404", "500" …
- Ainsi que toutes statistiques que vous jugerez pertinentes

3 Liste des fonctionnalités à coder avec Python3

- Fonction "**parsing(file)**" : avec en paramètre « **file** », le nom du fichier au format chaîne de caractères pour la lecture des lignes du fichier et stockage des données dans une structure facilement exploitable pour exporter des formats comme CVS, JSON, XML. Il peut être intéressant de supprimer les adresses IP identiques, en doublon. En effet, lorsqu'un navigateur se connecte dans les logs on retrouve tous les clicks de l'utilisateur lors de son parcours sur le site, et donc plusieurs lignes avec la même adresse IP.
- Fonction "**getIP_infos(ip)**" : permettant de connaître des informations sur l'adresse IP via le site : https://ip-api.com
- Fonction "**exportToCSVFile(liste)**" : permettant d'exporter l'ensemble des données au format CSV (chercher des informations sur ce format)
- Fonction "**exportToJSONFile(liste)**" : permettant d'exporter l'ensemble des données au format JSON (chercher des informations sur ce format)
- Fonction "**exportToXMLFile(liste)**" : permettant d'exporter l'ensemble des données au format XML (chercher des informations sur ce format)

NB : vous direz quel format vous semble le mieux adapté pour réaliser des graphiques et des statistiques selon vous.

4 Evaluation de votre travail : contrôle avec le module R107

Le deuxième contrôle R107 intégrera l'évaluation de se module.