M2 Statistiques et Econométrie
Introduction au langage Python
2018 – 2019

Ce document constitue le sujet du projet que vous aurez à réaliser pour valider le module d' *Introduction à Python*. Le projet est à réaliser seul et plusieurs aspects seront considérés dans l'évaluation :

- **Correction du code.** Celui-ci devra bien entendu réaliser ce qui est demandé.
- **Présence de commentaires.** Le code devra être abondamment commenté. Notamment, chaque fonction devra être précédée de commentaires pour détailler la nature des éventuelles entrées/sorties, l'objectif de la fonction et son fonctionnement.
- **Propreté du code.** Le code fourni devra être le plus propre possible en évitant des noms de variables ou de fonctions non explicites.
- **Non utilisation de librairies.** Ce cours étant une introduction au langage Python, il vous est demandé de ne pas utiliser de librairies qui faciliteraient le travail (par exemple Pandas pour le chargement et la manipulation des données). Mon postulat est que pour utiliser une librairie, il faut d'abord maîtriser les concepts élémentaires du langage. Et c'est ce point qui est évalué ici. Toute utilisation de librairie (même si le code répond aux questions) sera fortement sanctionné.

Jeu de données. Les données sont téléchargeables sur l'espace Moodle de ce cours. Les données manipulées dans ce projet sont des données de logs associés à un site internet pendant une période de 2 jours. Chaque ligne du fichier représente la demande d'accès d'un hôte (visiteur) à une resources du site en question (une page, une image, etc.). Nous disposons des informations suivantes concernant cette demande : l'heure précise, le

code retourné par le serveur suite à cette demande (par exemple le code 404 indique que la ressource demandée n'existe pas alors que le code 200 indique que la demande a bien été traitée) et la taille de la réponse fournie par le serveur (donc la taille de la ressource quand le code retourné est 200). Pour plus d'informations sur ce jeu de données, vous pouvez consulter la page `http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html`.

Instructions. Vous devez écrire du code python pour répondre aux questions suivantes :

1. Combien de requêtes d'accès (de lignes) sont de type `GET` ?
2. Fournissez les statistiques de nombre d'accès par code retourné par le serveur (ie., le nombre d'accès avec un code 200, le nombre d'accès avec le code 404, ...).
3. Combien de requêtes ont été effectuées par heure et par jour ?
4. Combien d'utilisateurs uniques ont accédé au site par heure ?
5. Pour la journée du 30, combien de requêtes de type `GET` ont cherché à accéder à un fichier contenu dans le répertoire `docs` ?
6. Combien de requêtes concernent des fichiers de type image ?
7. Qui a le plus accédé au site durant cette période ?
8. Quelle est la taille moyenne des images demandées par les utilisateurs ? Attention, si un utilisateur souhaite accéder à une image mais que le code retourné est 400, alors la taille de la ressource (la dernière valeur de la ligne) n'indique pas la taille de l'image.

Remarques complémentaires.

- La date de remise du projet sera annoncée sur Moodle et un dépôt sera mis en place à cet effet.
- Un forum va être ouvert pour que vous puissiez poser des questions sur le projet. Il s'agira du seul mode de communication autorisa pour le projet et les posts ne devront pas faire apparaître de code. Tout morceau de code sur le forum sera effacé et l'auteur du post sera sanctionné par un point en moins.