**Projet 3A**

Profiling web users

FOURKA Bilal

* Description du projet :

L’analyse des profils utilisateurs web est un problème fondamental pour l'exploration du Web et l'analyse des réseaux sociaux. Ses tâches de base comprennent l'extraction de l’information de base, le Mining des préférences des utilisateurs et inférer leurs données démographiques. L’objectif de ce projet est de construire une chaine, en utilisant un système big data, capable de nous classifier les profils selon les données collectées sur les utilisateurs, les détecter et proposer des styles de publicités et d’affichage adaptés à chaque profil.

* La semaine du 18 octobre et avant :

J’ai commencé ce projet par la découverte des solutions Big Data offertes par les grandes boites des services cloud en faisant un benchmark. Ainsi, le choix c’était d’utiliser l’environnement google cloud vu la grande communauté qui existe sur le web.

Ensuite, il fallait chercher une conception de données où on aura les informations sur les clicks et les pages vu par l’utilisateur. Pour cela j’ai pu trouver une compétition kaggle ou il y a beaucoup de données. Lien de la compétition :

<https://www.kaggle.com/c/outbrain-click-prediction>

L’ensemble de données de cette compétition contient un échantillon de pages vues et de clics d’utilisateurs, comme observé sur plusieurs sites d’éditeurs aux États-Unis entre le 14-juin-2016 et le 28-juin-2016 (14 jours).

Concernant la partie google cloud, j’ai pu crée un cluster de machine virtuelle et lancer une communication SSH. En réalité, je devais passer deux semaines à résoudre le problème de connexion SSH depuis mon pc. Il s’est avéré qu’il y avait plus de de réglage à faire sur le cluster et j’ai recommencé pour construire un cluster où je peux exécuter des commandes via SSH à distance. En plus j’ai pu lancer un script python ou j’utilise des fichiers qui existent dans un Bucket (Espace de stockage) que j’avais créé.

Pour la prochaine semaine, je compte construire un modèle qui lit les données des tables de la compétition kaggle et fait du preprocessing afin de les utiliser dans un algorithme de machine learning.