Sujet 2: Optimisation et clustering des campagnes de recrutement sur le web

Xtramile souhaite optimiser le choix des canaux de diffusion des offres d'emploi et proposer le canal qui sera en mesure d'avoir le plus de CV pertinents possibles.

Pour cela 4 phases sont nécessaires:

1. Etude de correlation entre les variables d'entrée: offre d'emploi, catégorie, pays, ville, canal de diffusion, nombre de CVs reçus, nombre de CVs pertinents, nombre de clics, saisonnalité...
2. Suivant les résultats de l'étape une, création d'un clustering permettant de regrouper les offres d'emploi grâce à différents paramètres.
3. Créer une API qui prendra en entrée une offre d'emploi et choisira un cluster qui permettra d'optimiser les campagnes pour choisir le meilleur job board et le budget à allouer sur celui-ci

Compétences aquises à l'issu de ce hackaton:

- Langage python
- Pré-traitement des données textuelles
- Vectorisation des offres d'emploi
- Fouile de données dans une base de données de plusieurs millions de data
- Clustering
- Algorithmes d'apprentissages non supervisés ou semi supervisés (selon le choix du groupe)
- Tests et utilisation de cross validation pour valider les modèles
- Création d'une api
- Manipulation d'une machine virtuelle

Quelles sont les données à disposition?

Les données issues de la base de données. Xtramile utilise un tracker qui permet de récupérer les statistiques en temps réel.

id, title, category, job_group_id, country, cpc, name,
title, 'keywords', 'description', 'job_type', 'status', 'employer',
job, name, enabled, action', 'amount_action', 'limit_cv',
budgetmax, budgetleft, 'uppdate', 'creation'

Comment étudier la correlation entre les variables du dataset et bien les choisir ?

Pandas propose à partir d'un dataframe d'étudier les différentes correlations existantes entre les variables.

```
df['country'].corr(df['action'], method= 'spearman')
```

Dans cet exemple on étudie la corrélation entre l'attribut country et action (qui représente un click ou une conversion). L'ojectif étant d'étudier toutes les corrélations de prendre -1<corr< -0.5 et 0.5<corr< 1. Ces variables feront partis du cluster.

Qu'est ce qu'un clustering?

Le clustering est une des méthodes d'analyse des données. Elle vise à diviser un ensemble de données en différents « paquets » homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité informatique) que l'on définit en introduisant des mesures et classes de distance entre objets.

L'objectif du clustering dans ce projet est de rassembler les offres d'emploi similaires pour différentes variables qui peuvent être la description de

l'offre, le nombre de conversion, de clicks...