



## Exercice de Data Science - NLP

### Objectif de l'exercice

L'objectif de l'exercice consiste à évaluer votre capacité à résoudre une problématique de Data Science. Différents éléments seront évalués :

- La qualité du code
- La capacité à implémenter une approche descriptive et prédictive
- La capacité d'analyser les données à disposition
- Votre capacité de restitution

### Quelques détails logistiques

- Les données sont accessibles via le lien suivant : <https://data.world/crowdfunder/brands-and-product-emotions>
- Vous avez 48h au moment de la réception de l'exercice pour envoyer votre code au recruteur
- Il est attendu lors de l'entretien une présentation de 20 minutes qui synthétise et présente votre travail ainsi que vos résultats. Vous n'avez pas de contrainte temporelle concernant la présentation finale
- N'hésitez pas à poser vos questions aux recruteurs si vous avez des doutes

## Enoncé :

Cet exercice est purement fictif. Votre client est un vendeur de tablettes et de téléphones qui souhaite identifier les produits les plus appréciés sur le marché afin de pouvoir les mettre en avant.

En scrappant les données sur Twitter, il a été capable de construire un dataset et d'annoter l'ensemble des tweets avec un avis positifs et négatifs et souhaite automatiser la détection des sentiments associés à chaque tweet.

Il fait appel à vous afin que vous puissiez l'aider à comprendre les informations présentes dans le dataset et pour construire le moteur d'analyse de sentiments. Après un atelier de brainstorming, vous avez identifié un ensemble de questions à répondre.

- Est-ce qu'il est possible de compléter la colonne « emotion\_in\_tweet\_is\_directed\_at » afin d'identifier plus de produits/marques ?
- Quelles sont les marques ou les produits les plus appréciés ?
- Quelles sont les marques ou les produits les moins appréciés ?
- Est-ce qu'il est possible de segmenter les tweets en famille en fonction des termes utilisés ?
- Quels sont les termes les plus récurrents ayant le plus d'impact sur la détection du sentiment ?

Une fois les questions répondues, il vous demande d'implémenter le moteur de détection des sentiments associés aux tweets. De nature méfiante, il souhaite pouvoir choisir parmi 3 modèles de Machine Learning /Deep Learning différents.

Un aspect optionnel mais qu'il apprécierait serait de comprendre pourquoi l'algorithme se trompe quand il classe mal les tweets.

Bon courage dans la réalisation de votre projet