**Projet Détection de mots interdits avec Python**

Le but de ce projet est de détecter la présence de mots interdits dans le texte de la table Osirisk.

Pour davantage d’information, se référer au rapport de stage *2022-23-stage5A-Roig-Lila\_PFE.pdf*, aux présentation PowerPoint ou aux commentaires et description des fonctions implémentées. Il est conseillé de lire le rapport de stage avant de reprendre le projet pour avoir une idée de la démarche globale.

/.\ Nous avons développé ce projet en Python et en Alteryx. **La solution Python ne donne pas des résultats tout à fait corrects**. Cependant, comme Alteryx fonctionne et que c’est la solution Alteryx que nous allons utiliser, je n’ai pas perdu de temps à essayer de trouver l’erreur en Python. On conserve Python uniquement à titre d’information, au cas où une utilisation ultérieure serait requise.

**Mise en route du projet**

Pour faire fonctionner le projet, il vaut avoir au préalable installé Python sur sa machine (voir le OneNote [Data & Décisionnel Notebook](onenote:https://groupebpce.sharepoint.com/sites/DataDcisionnel/SiteAssets/Data%20&%20Décisionnel%20Notebook/) et la partie « A lire pour le prochain alternant »).

Puis créer un nouvel environnement conda spécifique à ce projet, par exemple intitulé *motsInterdits*. Ensuite, installer dans cet environnement *motsInterdits*:

* Jupyter Notebook (voir le OneNote)
* Les librairies listées dans le fichier *requirements.txt* avec la version donnée

L’environnement de travail est prêt.

**Description des répertoires & fichiers**

* Le répertoire data contient :
  + "Osirisk\_export\_20230524 - mots interdits RGPD.xlsx" : table Osirisk contenant le texte dans lequel rechercher les mots interdits. Seules les colonnes «Libellé Incident», « Description », « Réclamation », « Local » , « Client » et « Lieu » de la table Osirisk sont susceptibles de contenir des mots interdits. Ce sont donc ces colonnes que nous allons analyser.
  + "2023 01 20 Dictionnaire contrôle a posteriori.xlsx" et "2023 01 20 Dictionnaire contrôle a priori.xlsx" : listes contenant les mots interdits à rechercher dans le texte de la table Osirisk. Il faut prendre l’union de ces 2 listes pour avoir l’ensemble des mots interdits à rechercher.
  + "prenoms\_France\_1900-2021.csv" : liste de tous les prénoms de France. Utile pour détecter si un mot interdit n’est pas en réalité un prénom (ex : Nègre Véronique).
* Notebook motsInterdits.ipynb : Notebook contenant l’ensemble des codes pour réaliser la détection de mots interdits dans la table Osirisk
* Répertoire utilities : contient le répertoire fr\_core\_news\_sm-3.5.0 qui est le modèle de lemmatiseur français, qu’il a fallu installer à la main (voir requirements.txt ou le notebook motsInterdits.ipynb)
* Répertoire temp\_result : contient des sauvegardes de fichiers au format pickle de Python. Au fur et à mesure de l’exécution du notebook motsInterdits.ipynb, on sauvegarde les résultats des fonctions qui sont très longues à exécuter au format pickle afin de les charger rapidement et de ne pas avoir à les réexécuter systématiquement. Voir le notebook motsInterdits.ipynb pour savoir à quoi correspondent ces fichiers.