**PREDICTION DU NAF :**

*Objectif* : Prédire le NAF (2 premiers caractères) de l’établissement du recensé pour limiter la périmétrie lors de la fusion avec le fichier SIRUS.  
Echantillon d’analyse :

70% des données 2017, hors départements 44 et 13.

Chaque activité déclarée ACTET\_X est décomposée en mots unitaires.

On calcule ensuite un TF IDF par mot / activité calculée (ACTET\_C).

***Scripts : 1 - Apprentissage\_NAF\_Mot\_TFIDF.R***

***Fichier résultat : Rec.NAF.csv***

Analyse des performances :

Sur les départements 13 et 44 (données 2017) n’ayant pas servi à l’analyse précédente, on restitue les cinq premiers NAF les plus probables en fonction de l’activité déclarée ACTET\_X.

On obtient les résultats suivants :

78% de correspondance : l’un au moins un des 5 NAF prédits (2 premières lettres) est égal au NAF final (2 premières lettres) issu de ACTET\_C.

***Script : 2 - Construction filtre NAF.R***

**Correspondance avec Sirus :**

*Objectif* : Joindre les données du RP avec celle de Sirus sur la commune déclarée et les communes adjacentes, puis filtrer les entreprises sur les 5 codes NAF prédits (ensemble des Siret des communes alentours sur les 5 codes NAF correspondants). Un calcul de distance de Levenshtein combinant les calculs sur les adresses et les raisons sociales est appliqué pour déterminer la correspondance la plus probable (1 seul Siret proposé).

Analyse des performances :

Sur l’échantillon de 884 lignes du RP de départ, 769 ont trouvé une correspondance sur la commune. En analysant le résultat de la jointure RP/Sirus filtrée sur les 5 codes NAF prédits, on peut retrouver 446 Siret corrects soit 50% de l’échantillon. Enfin, la prédiction du Siret le plus probable basée sur une combinaison de calculs de distance de Levenshtein sur les adresses et les raisons sociales permet trouver 247 Siret soit 32% de l’échantillon initial (voir table datafinal\_export.csv)

***Script : 3-Final -- filtres plus distances adresses et RS - version adjacentes\_V2.R***

**WEBSCRAPPING :**

***Script : scraping.py***

Cf. support hackathon2018\_pole\_emploi.PPT