# Text mining
# Projet final

Décembre-2023

## Présentation

L'objectif de ce projet est de de développer des outils permettant de répondre à la question 'De quoi parle-t-on dans la presse ?'. Pour ce faire, nous disposons d'un fichier contenant plus d'un million de titres de dépêches publiées entre 2003 et 2021 [1]. Les parties I, II et III sont à faire en *R*. La partie IV est à faire en Python.

## Partie I : de quoi parlait-on en $\cdots$ ?

Dans cette partie, il vous est demandé d'écrire **un code R** permettant d'analyser les dépêches publiées pendant une période donnée, définie par une date de début et une date de fin, pour repérer les mots les plus fréquents (et par conséquent les sujets d'actualité).

Pour tester votre code, vous pouver considérer des périodes de quelques mois commençant respectivement en février 2003 et en décembre 2019.

## Partie II : Associer des classes à des dépêches

Dans cette partie, il vous est demandé de choisir 3 sujets présents dans le fichier de données (e.g. géopolitique, économie et sport, ou guerre d'Irak, covid et jeux olympiques), construire un dataset avec des dépêches appartenant à ces sujets, répartir le dataset en ensemble d'apprentissage et ensemble de test puis entraîner un classifieur bayésien et le tester.

1. Ce fichier a été récupéré sur le site *Kaggle*

# Partie III : Définir des clusters de dépêches

Dans cette partie, il vous est demandé d'utiliser l'algorithme de clustering *Frequent Term Sets* vu en cours, éventuellement adapté à vos besoins. Rappelons que cet algorithme définit un cluster par les mots qui sont présents dans les documents qui le constituent. Rappelons aussi que cet algorithme ne construit pas des clusters disjoints mais des clusters dont l'intersection (overlapping) est aussi réduite que possible. Vous devez donc sélectionner un sous ensemble de dépêches, former des ensembles de mots fréquents puis appliquer l'algorithme de clustering et analyser ses résultats.

# Partie IV : Une étude plus globale des dépêches

Dans cette partie, notre objectif est d'étudier les thèmes présents dans les dépêches et leur évolution dans le temps.

1. Faites des statistiques descriptives sur la fréquence des mots (maximum, moyenne, variance, médiane et autres quantiles, $\cdots$).
2. Peut-on distinguer deux ou plusieurs catégories de mots fréquents ?
3. On souhaite étudier l'évolution de la présence d'un mot donné dans les dépêches. Prenez quelques exemples de mots et donnez cette évolution. On donnera une représentation graphique de cette évolution. Comment se traduisent les catégories de la question précédente en termes d'évolution de la présence du mot dans les dépêches ?
4. En utilisant une méthode de votre choix et en vous servant éventuellement de ce que vous avez fait dans les questions précédentes étudiez les thèmes (topics) présents dans les dépêches ainsi que leur évolution dans le temps.