# Comparaison de deux méthodes de classification de profils de Twitter

## 1. L'Objectif

L'objectif est d'implémenter sur un même dataset et pour une même problématique une approche non supervisée et une approche supervisée et d'en faire le comparatif.

Pour ce faire il vous a été fourni deux datasets :

- Tweet Worldcup.zip = dataset complet,
- Tweet Worldcup 200Twets.zip = les 200 premiers fichiers pour vos tests (au cas où).

*Mais si vous pouvez travailler directement sur le dataset complet, je vous le conseille.*

Chaque groupe projet implantera deux algorithmes de Machines Learning (un non-supervisé et un autre supervisé) de votre choix (vus en cours ou TD ou autres) pour la détection de profils « suspects » de Twitter.

Ensuite dresser une étude comparative des résultats fourni par chacun des deux algorithmes sur les dimensions les plus pertinentes.

## 2. Les livrables

L1 : rapport de projet
L2 : code documenté (avec un readme sur son utilisation) facile à prendre en main et à utiliser (pourquoi pas intuitif).
L3 : présentation de votre travail au cours d'une soutenance

## 3. Les Grandes étapes (ceci n'est qu'une suggestion)

**Etape 1 : Préparation de vos données**

- Extraire et de prendre en main les données,
- Comprendre les données,
- Comprendre les indicateurs métier qui répondent à la question posée,
- Choisir le format de vos données.

Il est fortement conseillé d'utiliser du mongoDB au vu de la taille et surtout du format de vos fichiers.

**Etape 2 : Data exploration**

- Générer les attributs dérivés du graphe social, du contenu et du comportement général du profil. Par exemple :
  - Nombre de followers, nombre de profils suivi et leur ratio,
  - Fréquence de publication de tweets, nombre moyen d'URL, nombre moyen de hashtag,
  - Nombre moyen de mentions, nombre moyen de retweets, nombre de réponses,
  - Longueur moyen des tweets,
  - Agressivité et Visibilité (de SPOT)
  - Etc.

- Normaliser / réduire le nombre de dimension si cela est utile

- Pré-visualiser vos données sous différents « angles » afin de mieux les comprendre ou faire ressortir des facteurs de compréhension de la problématique
- Identifier les premiers comportements (corrélations par exemple).

**Etape 2 : Formulation de vos deux approches**

Expliquer théoriquement les deux approches choisies et les mettre en œuvre.

**Etape 3 : Spécification de votre POC et expérimentations**

**Etape 4 : Rédiger votre rapport**

## 4. Exemple de modules/fonctionnalités à mettre en place