Antoine Boutet – Département Informatique, INSA Lyon
Mathieu Cunche – Département Télécommunications, INSA Lyon
Benjamin Nguyen – Département Informatique, INSA Centre Val de Loire

Défi d'Anonymisation et Réidentification de Données

Contexte : Le projet est organisé sous forme de compétition entre équipes de l'INSA CVL et de l'INSA de Lyon. L'objectif est double et se divise en deux parties : 1) anonymiser des données tabulaires (de type « ticket de caisse ») et 2) essayer de réidentifier un maximum de données des autres équipes. Bien entendu, il s'agit de continuer d'être capable d'obtenir des informations à partir des données anonymes. Il s'agit donc de mettre en place des techniques d'anonymisation qui vont préserver tant que possible la qualité et la précision des données, soit en paramétrant des modèles existants, soit en développant vos propres modèles. Dans un second temps, vous mettrez au point vos propres modèles d'attaque, en utilisant une analyse manuelle, des techniques statistiques, des heuristiques, etc.

Exigences : Le traitement des données personnelles est très contraint par le RGPD (Règlement Général sur la Protection des Données). Toutefois, une donnée anonyme n'est plus protégée par le règlement, et peut être traitée et stockée sans limite. La définition de l'anonymat est :

Pour déterminer si une personne physique est identifiable, il convient de prendre en considération l'ensemble des moyens raisonnablement susceptibles d'être utilisés par le responsable du traitement ou par toute autre personne pour identifier la personne physique directement ou indirectement, tels que le ciblage. Pour établir si des moyens sont raisonnablement susceptibles d'être utilisés pour identifier une personne physique, il convient de prendre en considération l'ensemble des facteurs objectifs, tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte des technologies disponibles au moment du traitement et de l'évolution de celles-ci. Il n'y a dès lors pas lieu d'appliquer les principes relatifs à la protection des données aux informations anonymes, à savoir les informations ne concernant pas une personne physique identifiée ou identifiable, ni aux données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable. Le présent règlement ne s'applique, par conséquent, pas au traitement de telles informations anonymes, y compris à des fins statistiques ou de recherche. (considérant 26 du RGPD)

L'objectif de ce projet est donc de proposer la meilleure anonymisation possible, tout en conservant la meilleure utilité possible des données.

Compétences : Vous allez mettre en œuvre des compétences de gestion et d'analyse de données, ainsi que des techniques d'anonymisation sur des données tabulaires.

Déroulement du projet : Le projet se déroule en deux phases : création d'un (ou plusieurs) datasets anonymisés et désanonymisation des données anonymisées des autres équipes. La compétition se déroulera sur une plateforme spécifique mise en place par l'organisation d'une compétition à venir (crowdAI). Les règles complètes du concours sont données sur le Moodle du séminaire. Vous devez former des groupes de 5/6 personnes.

Ressources : projet git - https://gitlab.crowdai.org/Drayer34/DARC

Notation : Votre note finale ne dépendra pas uniquement de votre résultat au concours.

Planning :

- 09/10 : Début du challenge
- 26/10 : Suivi de projet
- 09/11 : Suivi de projet
- 04/12 : Coup d'envoi de la phase d'attaques
- 13/12 : Annonce des résultats