REPUBLIQUE DE COTE D'IVOIRE

Union – Discipline – Travail

****

MINISTERE DE L'ÉCONOMIE, DU PLAN ET DU DEVELOPPEMENT

****

ECOLE NATIONALE SUPERIEURE DE STATISTIQUE ET D'ECONOMIE APPLIQUEE

Travail de groupe sur :

Mise en place d'un module d'anonymisation des données

Réalisé par :

- GNABRO Mathy Aristide
- MONSO Akrebé Jean-Christ
- SANOKO Awa

Élèves ingénieurs statisticiens économistes en troisième année

Supervisé par :

Mme. KADIA Essan,

M. ZEGBOLOU Gbizié

Enseignants à l'ENSEA

Juin 2024

Tableau 1

SOMMAIRE

INTRODUCTION

Dans un monde de plus en plus digitalisé, les entreprises collectent et possèdent des quantités croissantes de données sur leurs clients. Ces informations sont cruciales pour la gestion des relations clients (CRM[1]) et pour offrir des services personnalisés. Cependant, la possession de ces données implique une grande responsabilité : les entreprises doivent utiliser ces informations de manière éthique et conforme aux règles, normes et lois en vigueur, notamment pour respecter les droits des clients en matière d'image, de notoriété et de vie privée.

La sécurité des données devient donc un enjeu majeur, tant pour la confiance des clients que pour la conformité réglementaire. En protégeant les informations sensibles, les entreprises peuvent éviter des violations de données qui pourraient nuire à leur réputation et entraîner des sanctions légales.

L'anonymisation des données est une méthode clé pour assurer cette sécurité. En rendant les données personnelles non identifiables, elle permet leur utilisation tout en protégeant la vie privée des individus. L'anonymisation repose sur plusieurs techniques comme le masquage des données, la pseudonymisation, le hachage, la substitution, la généralisation, la randomisation, la cryptographie, et bien d'autres.

Ces techniques doivent être soigneusement choisies et mises en œuvre en fonction des types de données et des objectifs d'anonymisation. Par exemple, le masquage est souvent utilisé pour protéger des numéros de carte de crédit, tandis que la pseudonymisation peut être utile pour remplacer des noms.

Dans le cadre de ce projet, nous développerons une **bibliothèque Python** capable de réaliser l'anonymisation des données. Notre solution sera en mesure de traiter différents formats de données, allant des fichiers texte (.txt) aux phrases et mots isolés. Le pipeline de notre choix inclura l'extraction des données, leur anonymisation et leur stockage.

Ce projet représente une opportunité d'appliquer les connaissances théoriques et pratiques acquises durant notre formation, tout en contribuant à la création de solutions innovantes pour la protection des données. Nous veillerons à ce que notre bibliothèque soit conforme aux réglementations en vigueur, telles que le RGPD, en garantissant un traitement sécurisé des données et en minimisant les risques liés à la manipulation des informations sensibles.

NB : Nous tenons à préciser que notre proposition pourrait être limitée par le fait que nous ne sommes pas encore des experts en matière de sécurité des données. Ce projet est avant tout académique, et les personnes qui s'engagent à utiliser ces méthodes doivent en tenir compte. Nous ne saurions être tenus responsables des conséquences potentielles de l'utilisation de cette bibliothèque dans des contextes professionnels ou sensibles.

[1] CRM : Customer Relationship Management,

I-RAPPEL SUR LES TECHNIQUES D'ANONYMISATION

1- La Randomisation

La randomisation est une technique essentielle d'anonymisation des données, qui remplace des valeurs spécifiques par des valeurs générées de façon aléatoire. Cela rend impossible de retracer l'origine des informations.

Elle protège efficacement la vie privée en ajoutant de l'imprévisibilité, conserve les propriétés statistiques des données mais peut être difficile à mettre en œuvre pour les données complexes, peut rendre les données moins utiles pour certaines analyses.

Exemple : Dans une étude médicale, l'âge d'une personne peut être modifié par un nombre aléatoire dans une certaine plage.

2- La Généralisation

La généralisation consiste à remplacer des données spécifiques par des valeurs plus générales mais moins précises. Cela permet de préserver l'utilité des données tout en réduisant le risque de réidentification.

Elle protège la vie privée en rendant les données moins précises, conserve une partie de l'utilité des données tandis qu'elle peut réduire la précision des analyses ou des recherches.

Exemple : Remplacer les âges précis par des tranches d'âge tel que 18-25 ans, 26-35 ans

3- Le Chiffrement

Le chiffrement rend les données illisibles sans la clé de déchiffrement correspondante, offrant ainsi une protection forte de la vie privée.

Il Protège efficacement les données, même en cas de divulgation non autorisée.

Il peut être aussi complexe à mettre en œuvre, peut ralentir l'accès aux données.

Exemple : Chiffrer les données sensibles stockées dans une base de données à l'aide d'un algorithme de chiffrement (césar)

4- Le Hachage :

Le hachage consiste à encoder les données en une chaîne de caractères de longueur fixe, unique pour chaque entrée. Contrairement au chiffrement, une fois les données hachées, elles ne peuvent pas être inversées pour retrouver les données originales. Le hachage protège les informations

sensibles et permet de vérifier l'intégrité des données. Il est utilisé pour stocker des mots de passe, vérifier l'intégrité des données et anonymiser les informations.

Le principal avantage du hachage est qu'il protège la confidentialité des données et assure l'intégrité. Cependant, il présente un inconvénient : les collisions, c'est-à-dire lorsque deux données différentes produisent le même hachage, peuvent poser des problèmes de fiabilité

Exemple : Hacher les adresses e-mail des utilisateurs pour stocker des identifiants uniques dans une base de données.

5- Masquage :

Le masquage consiste à remplacer les données sensibles par un caractère de masquage, généralement un astérisque (*). Cela permet de protéger des informations telles que les numéros de carte de crédit ou de sécurité sociale.

Exemple : On affiche les six premiers chiffres d'une carte de crédit dans les enregistrements, en masquant les autres.

6- Pseudonymisation :

La pseudonymisation consiste à remplacer les données identifiables par des pseudonymes ou des identifiants uniques. Plutôt que d'utiliser des informations réelles comme les noms, on représente les individus par des codes générés aléatoirement. Ces identifiants sont ensuite utilisés pour le traitement des données, ce qui rend cette technique utile en recherche médicale ou en gestion de la clientèle.

Exemple : le nom réel "Jean Dupont", un identifiant unique comme "Patient_12345" est attribué.

II-PRESENTATION DE LA BIBLIOTHEQUE PYTHON

Notre bibliothèque Python permet d'implémenter différentes techniques d'anonymisation, vues en classe. Nous nous évertuerons à les présenter dans des cas pratiques d'utilisation. Cette section comprendra donc deux parties. Une partie donnons la liste des packages nécessaires au bon fonctionnement de la bibliothèque et une partie présentant le fonctionnement de la bibliothèque.

1- Les fichiers de base

Les packages Python nécessaires au bon fonctionnement de la bibliothèque et leurs fonctions sont présentés dans un doc string accompagnant la bibliothèque. Nous les présenterons donc de façon succincte dans les paragraphes suivants.

- **cryptocode :** ce module est utilisé pour procéder au hachage des données qui lui seront fournies en argument.
- **re :** ce module python permet de détecter des schémas de mot dans nos *inputs*. Dans notre cas, nous l'utiliserons pour retrouver les dates de formats : J/M/A, J-M-A, J.M.A, M.J.A, M-J-A, M/J/A. Soit les formats francophones et anglosaxonne de datation.
- **faker :** De ce module nous exploiterons la classe Faker. Cette classe sera ensuite paramétrée en français et nous permettra de substituer des noms entrés en argument, dans une liste, accompagnant le corpus, par des noms fictifs générés par Faker.
- **numpy :** Ce module nous a servi à générer des nombres de manière aléatoire.
- **datetime :** ce package est utilisé pour la généralisation et la substitution de dates trouvées dans le corpus de mots, donné en argument. Ces dates sont les dates respectant les formats susmentionnés (J/M/A, J-M-A, J.M.A, M.J.A, M-J-A, M/J/A)
- **ipywidgets** : Ce module permet de rendre la bibliothèque interactive afin de procéder à plusieurs anonymisations successivement. Il n'est pas utilisé dans la bibliothèque mais nous l'avons utilisé pour permettre un cas pratique dans le notebook « Script.ipynb »

Les modules qui ne sont pas automatiquement préinstallées par Python que nous utilisons sont : **faker, numpy, *ipywidgets*** et **cryptocode**.

2- Fonctionnement du module

a. Présentation des variables

La fonction d'anonymisation est composée de quatre variables clés :

- « **input_inf** » : C'est une variable de type string. Elle est censée qui contient l'information à anonymiser ou le chemin d'accès menant à un fichier texte à anonymiser.
- « **txt** » : C'est une variable de type booléenne qui prend deux valeurs *True* ou *False*. *True* permet de signifier que le string entré en argument est un chemin d'accès vers un fichier texte et *False* pour dire qu'il s'agit d'un simple texte à anonymiser.

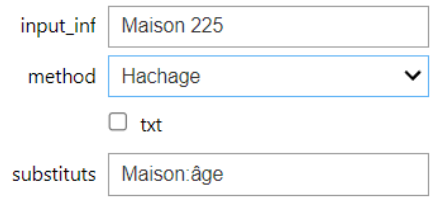
- « **method** » : Elle permet de choisir la technique d'anonymiser à utiliser pour traiter les informations entrées dans la variable « input_inf ».
- « **substituts** » : Cette variable peut être rempli facultativement dépendamment de la technique d'anonymisation que vous souhaitez appliquer aux données.

Toutes ces variables peuvent être utilisées interactivement grâce à la fonction « **interact** » de « **ipywidgets** ».

b. Présentation des méthodes d'anonymisation

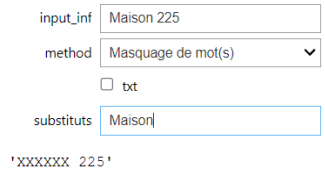
Afin de mieux cerner le rôle de chaque variable et le fonctionnement de la bibliothèque, nous proposons une série d'exemples d'utilisation avec différentes techniques accessibles avec la variable « **method** » dans le notebook « **Script.ipynb** » qui permet d'utiliser la fonction « **cryptage** » de notre bibliothèque « **Anonymisation.py** ». Nous avons en tout huit (08) méthodes d'anonymisation accessibles depuis notre bibliothèque basée sur cinq (05) techniques d'anonymisation. Il s'agit de :

- **"Hachage"** : Il permet de faire du hachage du contenu entré en argument de type "CH-1234". Dans cet exemple nous hachons le string ''Maison 225'' avec cette méthode.

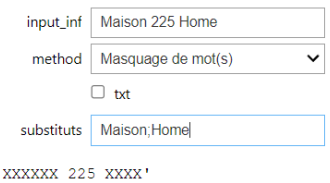



```
'S/+LHDYEvQ==*QF2UeUBBZ6ohs2H3HBZHZg==*F+ixA5uTSXzeenmM6qEZfA==*X3CJBOTN0lsj+BGIeOw3gg=='
```

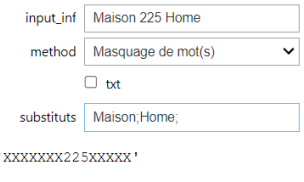
- "**Masquage de mots**" : Il permet de masquer un mot à la fois ou un groupe de mot séparé par la ponctuation « ; ». Dans le premier exemple, nous masquons le mot ''Maison'' du texte input ''Maison 225'' et dans le deuxième nous masquons les mots ''Maison'' et ''Home'' du groupe de mot 'Maison 225 Home''. Il faudra aussi faire attention de ne pas mettre '';'' et d'appuyer sur la touche espace au risque de transformer les espaces entre les mots en ''X'' comme dans le dernier exemple




```
'XXXXXX 225'
```

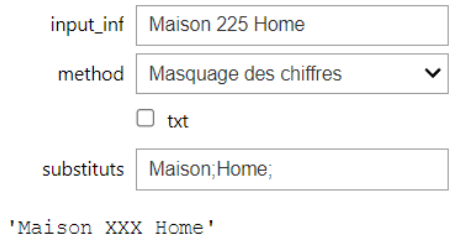



```
'XXXXXX 225 XXXX'
```

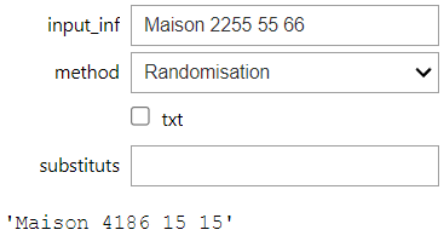



```
'XXXXXXX225XXXXX'
```

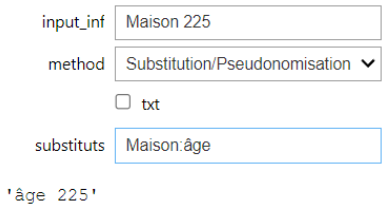
- **"Masquage des chiffres"** : Il permet de substituer TOUS les chiffres du corpus de mots par le caractère ''X''.



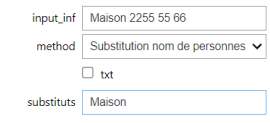
- **"Randomisation"** : Il permet de remplacer TOUS les nombres du corpus par d'autres nombres contenant le même nombre de chiffres. Dans l'exemple ci-dessous, les chiffres 2255, 55 et 66 sont respectivement remplacés par 4186, 15, 15 soit des nombres de quatre (04) chiffres pour les nombres de quatre (04) et ceux deux (02) par des nombres de deux (02) chiffres également.



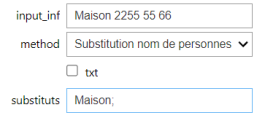
- **"Substitution/Pseudonomisation"** : Cette méthode permet de remplacer un groupe de mot par un autre groupe de mots entrés en argument. En inscrivant le nom du texte à remplacer suivi de deux le remplaçant, on arrive aisément à remplacer l'ancienne valeur (celle de gauche) par la nouvelle valeur (celle de droite).



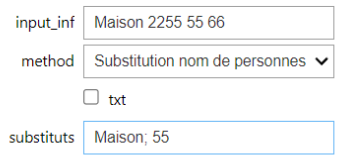
- **"Substitution nom de personnes"** : Elle consiste à remplacer le nom d'une personne ou d'une liste de noms de personnes entré en argument. Elle peut être appliquée pour des noms autres que ceux de personnes ou des chiffres mais ramènera systématiquement le nom d'une personne. Pour cette fonction, il faudra faire attention de ne pas mettre de signe de ponctuation « ; » à la fin de la séquence de noms au risque que la fonction remplace la chaine de caractères vide ‘’’’ par des noms comme on peut le voir dans le deuxième exemple.



'Martin Labbé 2255 55 66'

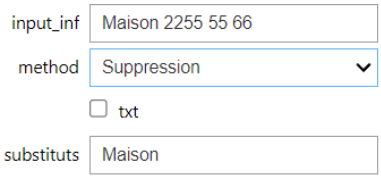


'Hélène GuillouPHélène GuillouaHélène GuilloutHélène GuillourHélène GuillouiHélène GuilloucHélène GuilloukHélène Guillou



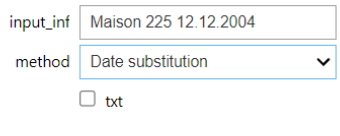
'Jérôme Marchal-Fabre 2255Joséphine Thibault Le Valentin 66'

- **"Suppression"** : Cette méthode permet de supprimer une liste de valeurs entrée en argument. Dans l'exemple ci-dessous, la mot ''Maison'' a été retiré du corpus.

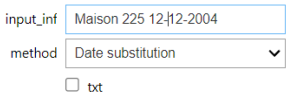


' 2255 55 66'

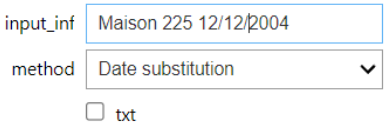
- **"Date substitution"** : Elle consiste à remplacer toutes les dates du document qui respectent un certain format (J/M/A, J-M-A, J.M.A, M.J.A, M-J-A, M/J/A) par d'autres dates du même format générées aléatoirement.



'Maison 225 27.05.2003'

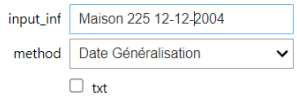


'Maison 225 13-09-2006'

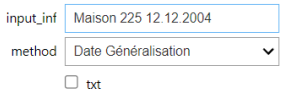


'Maison 225 16/08/2009'

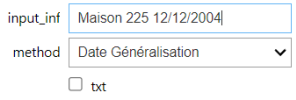
- **"Date Généralisation "** : Elle consiste à remplacer toutes les dates du document qui respectent un certain format (J/M/A, J-M-A, J.M.A, M.J.A, M-J-A, M/J/A) par l'année où ils ont été émis.



'Maison 225 2004'



'Maison 225 2004'



'Maison 225 2004'

Cas des fichiers textes : En cochant, la case txt, la variable « **txt** » prend la valeur *True*. Si la valeur *True* est renseigné, le texte entré dans la varaible « **input_inf** » est considéré comme un chemin d'accès. Il faut que le fichier texte soit trouvé dans le même dossier que le fichier de la bibliothèque pour la procédure d'anonymisation du fichier d'extension .txt fonctionne correctement. On ne peut traiter qu'un seul fichier à la fois.

ANNEXES

Tableau 2 : Membres du groupe

MONSO	GNABRO	SANOKO
Akrebé Jean-Christ	Mathy Aristide	Awa