Sécurité Appliquée-PVP TD2

Jean-François COUCHOT
couchot[arobase]femto-st[point]fr

8 novembre 2021

1 Extensions du k-anonymat

QID			SA	
Gender	**ZIP Code**	**Age**	**Disease**	**Salary**
Male	400071	35	bronchitis	10k
Male	400182	37	pneumonia	11k
Male	400095	39	stomach cancer	12k
Female	440672	54	gastritis	12k
Female	440123	58	Flu	15k
Male	440893	54	bronchitis	16k
Male	400022	41	gastric ulcer	16k
Male	400135	46	gastritis	17k
Female	400182	44	stomach cancer	18k

TABLE 1 – Données originales

Gender	Zip Code	Age	Disease	Salary
*	400***	[30, 40[	bronchitis	10k
*	400***	[30,40[	pneumonia	11k
*	400***	[30, 40[	stomach cancer	12k
*	400***	[40, 50[	stomach cancer	18k
*	400***	[40, 50[	gastric ulcer	16k
*	400***	[40, 50[	gastritis	17k
*	440***	[50, 60[	gastritis	12k
*	440***	[50, 60[	Flu	15k
*	440***	[50, 60[	bronchitis	16k

TABLE 2 – Une version 3-diverse

Exercice 1.1 (l-diversité). *La table 1 reprend un jeu de donné minime issu de* [1]

1. *Quelles stratégies de généralisation proposez-vous ?*
2. *On considère la proposition donnée dans le tableau 2.*
   (a) *Est-elle 3-anonyme ?*
   (b) *Pourquoi est-elle* 3*-diverse ?*
   (c) *Quelle est sa valeur de Loss ?*

Exercice 1.2 (t-proximité). *On reprend le même jeu de donné que précédemment.*

1. *Calculer l'indice de proximité de la proposition donnée.*
2. *Existerait-il une autre proposition (différente de la suppression de tous les QID) qui aurait un indice t de proximité plus faible ? Dans ce cas, quel serait sa valeur de Loss ?*

2 Distances entre bases de données, bases voisines

Dans cette section, on nomme D_1 la base de données donnée dans le tableau 1 dont on a oublié tous les QID. Il ne reste donc plus que deux colonnes, Disease et Salary.

Exercice 2.1 (Insertion d'une ligne). *Soit D_2 la base obtenue en ajoutant à D_1 la ligne suivante.*

Gender	*Zip Code*	*Age*	*Disease*	*Salary*
Female	*400022*	*38*	*gastric ulcer*	*16k*

1. *Exprimer les bases D_1 comme un ensemble de paires.*

1. Elabd, E., Abdulkader, H., & Mubark, A. (2015). L–diversity-based semantic anonymaztion for data publishing. IJ Information Technology and Computer Science (IJITCS), 10, 1-7.

2. *Comment pourrions-nous exprimer la base D_1 comme un vecteur ? Le faire.*
3. *Exprimer la base D_2 comme un vecteur.*
4. *Calculer $\|D_1\|_1$, $\|D_2\|_1$ et $\|D_1 - D_2\|_1$. Lorsque la distance entre deux base vaut l'unité, on dit que ces deux bases sont voisines.*
5. *Dans ces deux bases, on voudrait compter le nombre de personnes qui ont des soucis gastriques (c.-à-d. qui ont l'une des maladies suivantes : stomach cancer, gastris ou gastric ulcer) et dont le salaire est inférieur à 15k.*
   (a) *Exprimer cette requête S à l'aide d'un vecteur.*
   (b) *A l'aide d'un produit scalaire entre 2 vecteurs, répondre à cette question pour D_1 et D_2.*
6. *Calculer $\Delta_{Q_S} = |Q_S(D_1) - Q_S(D_2)|$.*
7. *La quantité Δ_{Q_S} dépend-elle des valeurs de la ligne ajoutée.*

Exercice 2.2 (Modification d'une ligne). *Soit D_3 la base obtenue en modifiant dans D_1 la ligne*

Gender	*Zip Code*	*Age*	*Disease*	*Salary*
Female	*440672*	*54*	*gastris*	*12k*

par

Gender	*Zip Code*	*Age*	*Disease*	*Salary*
Female	*440672*	*54*	*gastric ulcer*	*12k*

1. *Calculer $\|D_3\|_1$ et $\|D_1 - D_3\|_1$. Comparer cette dernière valeur à celle obtenue dans l'exercice précédent. D_3 et D_1 sont-elles voisines ?*
2. *Calculer $|Q_S(D_1) - Q_S(D_3)|$.*