Documentation des métriques

1 Les pré-traitement des données

1. Tous les records avec le marquage 'DEL' sont supprimés du dataset anonymise avant que celui-ci est passe en argument aux différentes métriques d'utilité. En d'autres termes si vous décidez de supprimer 70% des records en précisant 'DEL' pour l'ID, toutes les métriques d'utilité définies reçoivent en paramètre les 30% des records du dataset anonymise qui restent non supprimés.
2. Les valeurs des colonnes dans ces records supprimés n'est pas important. Il suffit de respecter le type de la colonne ou même de laisser les valeurs initiales.
3. Les différentes métriques ne reçoivent pas exactement le même nombre et type d'arguments :
   - 3.1. Les métriques {POI, Tuile et Meet} doivent recevoir en paramètres :
     - Le dataframe anonymisé (certainement sans les records 'DEL')
     - Le dataset original brut
   - 3.2. Les métriques Date, Hour, Distance doivent recevoir en paramètres :
     - Le dataframe anonymisé (certainement sans les records 'DEL')
     - e dataset original excluant les records qui correspondent aux records 'DEL' dans le dataset anonymisé. (Ce qui fait que les 2 arguments passés en paramètre le dataset original et le dataset anonymisé ont le même nombre de records)
     - Le nombre total de records original (pour pouvoir calculer la moyenne en tenant en compte des records supprimés 'DEL' (la somme de leurs utilités est égal à 0 automatiquement)

2 Les métriques d'utilité

2.1 Date utility

Le but de cette métrique est de calculer l'écart de date pour chaque ligne du fichier anonymisé. Ainsi, on s'assure de l'authenticité de la date à laquelle la position GPS a été relevée. Le score est calculé de la manière suivante :

1. Chaque ligne vaut 1 point
2. 1/7 de points est enlevé par jour d'écart

$$1 - \frac{\textit{Difference nombre de jours}}{7}$$

3. Le score final est la moyenne des points par rapport au nombre de lignes dans le dataset original

2.2 Hour Utility

Le but de cette métrique est de calculer l'écart d'heure pour chaque ligne du fichier anonymisé. Ainsi, on s'assure de l'authenticité à la laquelle la position GPS a été relevée. Une modification de jour n'est pas prise ne compte (déja reperée dans l'utilité de Dates). Une position GPS le mardi à 16h déplacé le mercredi à 16h gardera TOUTE son utilité Le score est calculé de la manière suivante :

1. Chaque ligne vaut 1 point

2. Une fraction de 1/24 est enlevée à chaque heure d'écart

$$1 - \frac{Difference\ d'heures}{24}$$

3. Le score final est la moyenne de points par rapport au nombre de lignes dans le dataset original

2.3 Distance Utility

Cette métrique calcule la distance entre le point GPS original et le nouveau point GPS en utilisant la métrique de Haversine. L'unité utilisé est le km. Le score final est l'inverse de la moyenne de distance calculée pour toutes les lignes. Explication de la distance Haversine

1. Calculer la distance Haversine (en KM) entre les Points GPS dans les dataset original et anonymisé sur chaque ligne.
2. Le score de l'utilité de chaque ligne est l'inverse de la distance calculée (plus la distance est minimale plus l'utilité est important). Au cas où la distance entre les deux points est inferieur à 1 km le score de la ligne est automatiquement égal à 1 pour eviter un score plus grand à 1.
3. Le score final est la somme des scores de toutes les lignes divisé par le nombre de records dans le dataset original (c'est la moyenne par rapport au dataset original)

2.4 POI Utility

Le but de cette métrique est de détecter les points d'intérêts d'un individu. Les points d'intérêts correspondent aux lieux où l'utilisateur a le plus séjourné. Dans ce cadre de ce fichier d'utilité, nous regardons par défaut 3 POI les plus importants pour les 12 semaines du dataset. 1 POI correspond à l'un des trois éléments : Lieu d'habitation (22h à 6h), lieu de travail (9h à 17h) et lieu d'activité (le weekend de 10h à 18h) . Les POI des individus sont calculés par semaine.
L'idée globale de cette utilité est de s'assurer que l'on retrouve bien dans le fichier anonymisé les lieux clé de la vie d'un individu. Le score est calculé à partir de cet algoithme:

1. Arrondir les latitues et longitudes à n points après la virgule ($n = 2$ dans notre cas)
2. Pour chaque semaine:
   2.1. Pour chaque individu:
      i. Trier par ordre croissant de 'DateTime' les points GPS de l'individu
      ii. Pour chaque type de POI (Lieu d'habitation, lieu de travail et lieu d'activité):
         - L'encontre d'un nouveau point GPS est considéré comme le moment d'arrivée a ce point (*arrival_time*)
         - Le dernier point detecté avant le changement de point GPS c'est à dire avant un déplacement est considéré comme le moment de départ (*departure_time*)
         - Le temps passé entre l'*arrival_time* et le *departure_time* pour le même point GPS sans déplacement est calculé selon $departure\_time - arrival\_time$
3. Pour chaque groupe {individu, semaine, point GPS, type de POI}, calculer le temps de séjour en additionnant le temps passé respective
4. Signaler les points POI de chaque individu par semaine et type de POI en choisissant le point GPS pour lequel le temps de séjour est maximal (*temps_sejour_original*)
5. Rechercher le temps de séjour des points POI detectés à l'étape précedente dans le fichier anonymisé en respectant la même logique des étapes 1, 2 et 3 (*temps_sejour_modifié*)
6. Pour chaque groupe {individu, semaine, point GPS, type de POI}: calculer la difference de temps de sejour $diff\_sejour = |temps\_sejour\_original - temps\_sejour\_modifié|$
7. Calculer le score final selon: $1 - \frac{\sum diff\_sejour}{\sum temps\_sejour\_original}$

2.5 Tuile Utility

Nom de la métrique: Déplacements effectués. Le but de cette métrique est de calculer la différence de zone de couverture d'un individu durant les 12 semaines d'étude. L'idée est la suivante : la métrique permet de vérifier que, globalement, la version anonymisée garde les informations de déplacement et de couverture d'un individu. Pour ce faire on mesure le nombre de cellules différentes dans laquelle l'utilisateur a séjourné.
Le score est calculé de la manière suivante :

1. Arrondir les latitues et longitudes à n points après la virgule ($n = 2$ dans notre cas)
2. Pour chaque individu i
   - Calculer le nombre de points GPS distincts dans le fichier original $nb\_cellule\_fichier\_original\_pour\_i$
   - Calculer le nombre de points GPS distincts dans le fichier anonymisé $nb\_cellule\_fichier\_anonyme\_pour\_i$
3. Calculer le score selon
   - Si $nb\_cellule\_fichier\_anonyme\_pour\_i < nb\_cellule\_fichier\_anonyme\_pour\_i$ alors:
   $$score = \frac{nb\_cellule\_fichier\_anonyme\_pour\_i}{nb\_cellule\_fichier\_original\_pour\_i}$$
   - Sinon
   $$score = \frac{nb\_cellule\_fichier\_original\_pour\_i}{nb\_cellule\_fichier\_anonyme\_pour\_i}$$

2.6 Meet Utility

Le but de cette métrique est d'identifier les cellules où circulent le plus d'utilisateurs.

1. Arrondir les latitues et longitudes à n points après la virgule ($n = 2$ dans notre cas)
2. Récupérer toutes les positions uniques et les trier par les plus visitées en terme de nombre d'enregistrements par position.
3. Récupérer les top $n\%$ des positions les plus visitées. (danc notre cas on ucherche les top 10%)
4. Récuperer le nombre de cellules distinctes qui verifie la condition de l'étape 3 ($m = |Top\ n\%\ des\ positions\ distinctes\ |$)
5. Répéter les etapes 1 et 2 pour le fichier anonymisé
6. Récupérer les top m positions visitées dans le fichier anonymisé.
7. Récupérer les positions communes m_{commun} des top $n\%$ du fichier original et les top m du fichier anonymisé
8. Calculer le score selon $\frac{nombre\ de\ points\ détectés\ sur\ les\ m\_commun\ positions\ GPS\ dans\ le\ fichier\ anonymisé}{n\%\ de\ la\ taille\ du\ fichier}$