Support de TP Data - mars 2024 – Intervenant : Luiz MIRANDA

Contexte

L'étude de la Qualité de l'Air Intérieur (QAI) est une préoccupation de santé publique, qui est notamment prise en compte dans le 4e plan national santé environnement. En effet, en calculant le temps passé au bureau, en classe ou à la maison, nous restons plus de 80 % de notre vie à l'intérieur. Les efforts récents pour améliorer l'efficacité énergétique des bâtiments font que le renouvellement de l'air dans les environnements intérieurs est limité et, par conséquent, la pollution générée dans ce type d'environnement reste concentrée, ce qui fait que la concentration des polluants est cinq fois plus élevée à l'intérieur qu'en extérieur.

Les sources de pollution dans l'environnement intérieur sont multiples, mais les plus communs sont les émissions relatives à des travaux et des rénovations dans l'environnement (peintures, sol, isolations, …), à la présence de certains objets, d'animaux ou des émissions relatives à des activités humaines quotidiennes. Il est à noter que dès que les émissions causées par des travaux et rénovation sont stabilisées, la source principale de pollution dans l'environnement intérieur est l'être humain.

Des activités, comme la cuisson et le ménage, sont sources de plusieurs types de polluants, comme des particules fines (particulate matter en anglais, ou PM) et des composants organiques volatiles (COV), qui causent ou intensifient certaines maladies respiratoires comme l'asthme et des allergies. Quand les niveaux de pollution sont élevés pendant une longue période cela peut causer un ensemble de symptômes aigus (maux de tête, irritation des yeux, du nez ou de la gorge, toux sèche, peau sèche ou qui démange, vertiges et nausées, difficultés de concentration, fatigue et sensibilité aux odeurs) qui constituent le syndrome du bâtiment malsain (sick building syndrome en anglais, ou SBS). Ainsi, reconnaitre qu'une activité polluante est en cours, devient une étape essentielle pour développer des systèmes de ventilation automatiques et pour prévenir l'usager des dangers liés à cette activité.

L'expérience

C'est dans ce contexte que nous avons fait l'expérience qui a généré les données utilisées pour ce TP. Cette expérience a été réalisée dans une salle de 13 m² (47 m³) avec des conditions environnementales (température et humidité) non contrôlées où 21 types de capteurs du type oxide métallique (MOX) ont été utilisés pour monitorer la QAI. Les capteurs sont distribués en 5 modules d'acquisition présentés dans le Tableau 2. Dans cette salle, 10 types d'activités quotidiennes ont été réalisées, elles sont listées dans le Tableau 1.

Tableau 1. Liste des activitésréalisées dans la salle

La QA dans le labo MOSS	Activité	Tag	Durée	#
Activité statique	1 personne (travailler au pc)	AS1	30 min	1
Repas/Cuisson	Cuissond'un œuf avec huile	Oeuf	10 min (cuire puis manger)	2
Activité humide type SdB	Séchage de linge	SdB	30 min	3
Ménage	Aspirateur (particules)	Asp	10 min	4
	Nettoyage de sol (Liquide vaisselle + javel dilués)	Nett	15 min	5
Sport	Sport baisse intensité ou similaire	Saber	15 min	6
Bougies parfumées	Identiques	Bougie	15 min allumé/ 15 min étain	7
Aération extérieure	Ouvrir la fenêtre	Aeration	10 min	8
Bricolage/Activité manuelle	Pistolet à colle chaude	BricoP	15 min	9
	Couper/Poncer du bois	BricoC	10 min	10

Les capteurs ont enregistré les émissions provenant de ces activités dans une base de données, et un calendrier a été rempli avec les moments de début et de fin de chaque activité (fichier activites.xlsx), constituant ainsi les labels de la base. En d'autres termes, nous savons que les mesures des capteurs entre le début et la fin d'une activité proviennent de cette activité (fichier « activites.xlsx »). L'ordre d'exécution des activités a été préétabli par le calendrier et elle a été construite à partir d'une chaine de Markov pour représenter la routine d'une maison dans les probabilités de transition entre des activités. Cette info peut être utilisée pour enrichir vos algorithmes. La matrice de transition de ces activités est présentée dans le fichier activites.xlsx.

Objectifs du TP (en binôme)

L'objectif de ce TP est de vous montrer toute la chaîne de travail d'un Data scientist. La méthodologie d'analyse est divisée en trois étapes : 1- Nettoyage/rangement des données dans une base propre ; 2- Extraction des caractéristiques au sein des données ; et 3- Classification/prédiction et présentation des résultats mesurables. Ces étapes sont représentées par les 4 tâches proposées dans ce TP. **Les tâches 1 et 2 représentent l'étape 1, la tâche 3 représente l'étape 2, et la tâche 4 représente l'étape 3.**

Tableau 2. Systèmes capteurs et leurscaractéristiques

Module	Temps d'échantillonnage	# de capteurs mox	# de copies	Noms des fichiers
Libelium New	4 s	6	2	mod1.txt / mod2.txt
PODs	10 s	4	3	POD 200085.csv POD 200086.csv POD 200088.csv
Piano Thick	10 s	9	1	IMT_Thick.csv
Piano Thin	10 s	10	1	IMT_Thin.csv
Piano PICO	10 s	3	1	IMT_PICO.csv

Les modules Libelium New, PODs et Piano PICO permettent également de mesurer la température et l'humidité. Les systèmes PODs sont capables de mesurer des PM et du CO_2.

Les tâches :

1- Pendant des périodes d'expériences nous avons interrompu l'acquisition des données pour les sauvegarder. Par conséquent, nous avons des fichiers correspondent à des périodes différentes. La première tâche de ce TP est de grouper tous les fichiers qui ont le même nom dans un seul fichier organisé chronologiquement et sans échantillon dupliqué. Les fichiers sont disponibles dans ce lien : https://partage.imt.fr/index.php/s/ComZJoFowmkP5w9

2- Chaque module d'acquisition a son propre temps d'échantillonnage. La deuxième tâche de ce TP est de synchroniser les données provenant de chaque module en utilisant le même temps d'échantillonnage et créer une base de données avec toutes les données disponibles dans un seul fichier .csv. Autrement dit, vous devez fusionner tous les fichiers en un seul qui contient toutes les colonnes des autres fichiers et sans lignes marqués comme NaN ou NULL. Ce fichier final sera nommé comme vous voulez et il sera la base des données sur laquelle la troisième tâche de ce TP sera faite.

3- La troisième tâche de ce TP est de trouver la signature moyenne de chaque activité. Par signature, on considère la série temporelle (avec tous les capteurs) qui représente chaque activité en moyenne.

4- La quatrième (et dernière) tâche de ce TP est de proposer une méthodologie pour faire la reconnaissance des activités en utilisant des données dans la base. Toutes les propositions de techniques doivent être justifiées. Vous devez également exécuter la méthodologie proposée et mesurer sa performance de reconnaissance d'activité.

Notation

Vous devez produire un rapport (par binôme), entre 2 et 5 pages, où vous allez décrire les solutions que vous avez développé pour chaque tâche du TP. Le rapport doit être clair et bien expliqué avec des tableaux et des figures, si vous le jugez nécessaire. Au-delà des explications, vous devez rendre disponible les codes (en Python) utilisés dans votre page GitHub (ou le dépositoire de votre préférence) dont le lien doit être dans le rapport.

La note sera composée de la qualité du rapport et de la réussite de chaque tâche. L'évaluation de réussite des tâches sera faite avec le code que vous ajouterez sur le dépositoire. Le code doit être net et bien commenté pour rendre possible l'évaluation.

Le rapport doit être cohérent avec les codes, c'est-à-dire, les solutions décrits dans le rapport doivent être les mêmes que celles implémentées dans les codes.

La composition de votre note suit le barème au-dessous :

- **Note du rapport (4 / 20) :**
  - Qualité du texte, style et soins ;
  - Cohérence rapport / code ;
  - Justifications des solutions ;
  - Description des problèmes trouvés et des solutions proposées/implémentées ;
- **Réussite de la tâche 1 (3 / 20) :**
  - Obtention d'un fichier unique par système capteur où chaque fichier a tous les échantillons du système ;
  - Les échantillons dans les fichiers unique sont en ordre chronologique ;
  - Il n'y a pas des échantillons dupliqués ou erronés dans les fichiers uniques ;
- **Réussite de la tâche 2 (5 / 20) :**
  - Les 8 fichiers uniques sont devenus un seul fichier avec tous les données présentes dans les 8 fichiers uniques ;
  - Les données sont tous avec le même temps d'échantillonnage (10 s) ;
  - Les échantillons du système « Libellium new » qui sont entre des échantillons des autres systèmes doivent être calculés en moyenne et mises dans le même instant de temps que les autres systèmes ;
  - Les colonnes des modules marquées avec « aqi » ou « iaq » doivent être supprimées ;
- **Réussite de la tâche 3 (3 / 20) :**
  - La segmentation de chaque activité est faite ;
  - La série temporelle moyenne de chaque activité est présentée ;
- **Réussite de la tâche 4 (6 / 20) :**
  - Description et justificatif de la méthodologie proposée ;
  - Exécution et présentation des résultats provenant de la méthodologie proposée ;
    - Exécution de ce qui a été proposé ;
    - Mesure de performance de la méthodologie.