IFT 599 / IFT 799 - Science de données

Guide des travaux pratiques

Automne 2022

Enseignants

	Courriel	Local	Téléphone
Shengrui Wang	shengrui.wang@usherbrooke.ca	D4-1018-1	+1 819 821-8000 x62022
Etienne G. Tajeuna	etienne.gael.tajeuna@usherbrooke.ca		+1 819 821-8000 x

Faculté des Sciences,
Département d'informatique

September 19, 2022

Sommaire

Dans le cadre des travaux pratiques exigés par le cours IFT 599 / IFT 799, quatre (04) jeux de données sont mis à la disposition des personnes étudiantes. Il est question ici, à partir d'un jeu de données de son choix, que la personne étudiante puisse finaliser un projet au complet. Quelque soit la nature des données choisies, ledit projet s'effectuera en quatre (04) étapes ($i = 1, \ldots, 4$) dont chacune constituera la réalisation d'un TPi. Après finalisation d'un TPi sur un jeu de données précis, la personne étudiante est libre de changer son choix de jeu données pour le prochain TP$(i + 1)$. Toutefois, nous ne recommendons pas cela. Il est préférable, une fois avoir fait le choix de son jeu de données, d'aller jusqu'au bout du projet avec le même jeu de données. Pour finir, la personne étudiante sera libre d'utiliser le langage de programmation qui lui sied le mieux. En pratiques, les langages Python et R sont assez fournis pour réaliser des tâches en sciences de données.

Contents

1 Jeux de données et présentation du projet à réaliser

1.1 Jeux de données

1. Amazon book reviews (ABR) (`http://jmcauley.ucsd.edu/data/amazon/links.html`).
   Ce jeu de données est constitué des revues effectuées par des clients sur différents livres. Chaque revue est présentée sous forme d'un dictionnaire comme suit:


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


Figure 1: Amazon review sample

2. Canada and USA COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes (CovT) (`https://www.openicpsr.org/openicpsr/project/120321/version/V12/view`).
   Ce jeu de données est constitué des sentiments que présentent des internautes des États-Unies d'Amérique et du Canada de Tweeter vis-à-vis de la Covid-19. À chaque instant, un sentiment est évalué par cinq (05) critères d'émotions. Les sentiments sont catégorisés par *neutre*, *positif* et *négatif* tandis que les critères d'émotions sont données par les variables *valence*, *peur*, *joie*, *colère* et *tristesse*.


```
> db.Sentiment_Tweets.findOne()
{
        "_id" : ObjectId("62591a947012ae68e09536b3"),
        "user_id" : "1319491585",
        "tweet_timestamp" : ISODate("2020-01-27T16:44:36Z"),
        "keyword" : "wuhan",
        "country/region" : "Malaysia",
        "valence_intensity" : 0.336,
        "fear_intensity" : 0.575,
        "anger_intensity" : 0.505,
        "happiness_intensity" : 0.184,
        "sadness_intensity" : 0.507,
        "sentiment" : "negative",
        "emotion" : "fear"
}
```


Figure 2: Sentiment record sample.

3. New-York City taxi trips (NYCT) (`https://data.cityofnewyork.us/Transportation/2018-Yellow-Taxi-Trip-Data/t29m-gskq`).

Figure 3: New York city trip record sample.

Ce jeu de données contient les trajets effectués par les taximen à New York City. Pour chacun des trajets effectué par un taximan, nous avons le nombre de passager(s), les positions géographiques (latitude et longitude) des points de départ et d'arrivée du trajet, le temps mis dans le trajet et la distance parcourue en secondes.

4. Electricity Load Forecasting (ELD) (`http://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014`).
   Dans ce jeu de données (sous forme tabulaire) nous avons les consommations électriques des clients sur une période allant de l'année 2011 à l'année 2014. Les valeurs enregistrées sont en KiloWatt (KW) au 15 minutes. Chaque colonne du jeu de données représente un client.

1.2 Scénario général du projet

Quelque soit l'ensemble des données, l'objectif final est de réaliser une tâche prédictive en utilisant une approche basée sur les méthodes de systèmes de recommendation et des algorithmes de clustering.

- Dans le jeu de données ABR, il est question de faire de la classification des opinions (indirectement, la classification de textes). On veut savoir, à partir des revues effectuées par un internaute sur des livres donnés, quelle serait l'opinion générée par l'internaute sur un autre (ou nouveau) livre. Cette opinion correspond à un certain dégré de satisfaction variant entre 1 et 5. On suppose que plus le dégré de satisfaction de l'internaute tend vers 5 plus on a des chances que le livre en question soit acheté par l'internaute.
- Dans CovT, on veut savoir, à partir de l'historique des sentiments présentés par un internaute, quel serait son prochain sentiment. Il est important de noter que, tous les internautes ne sont pas toujours actifs sur la toile. De ce fait, à certaines dates on a aucune connaissance sur le sentiment de certains internautes.

- Dans NYCT, on voudrait prédire le revenu horaire que fera un taximan. On part sur la base fictive que le gain effectué par un taximan sur un trajet est calculé par,

$$gain(trip) = (duration \times 0.011\$ \times \#passengers) - (distance \times 0.016\$) \quad (1)$$

où *duration* corresponds au temps mis sur le trajet en secondes, *distance* est la distance parcourue dans le trajet en miles et $\#passengers$ est le nombre de passager(s) présents dans le taxi pendant le trajet. La valeur de 0.011$ corresponds au coût facturé à un passager à chaque seconde. La valeur de 0.016$ corresponds au coût du carburant à l'unité de distance parcourue.
- Dans le jeu de données ELD, suivant les tendances de consommation électriques hebdomadaires, on voudrait identifier les profils d'utilisation les plus récurrents et voir comment est-ce que les clients changeraient leurs habitudes d'usage d'électricité dans un but de prédire leurs prochaines habitudes.

Sommairement, suivant le choix du jeu de données de la personne étudiante, le projet suivra les étapes élaborées par le schéma ci-dessous.

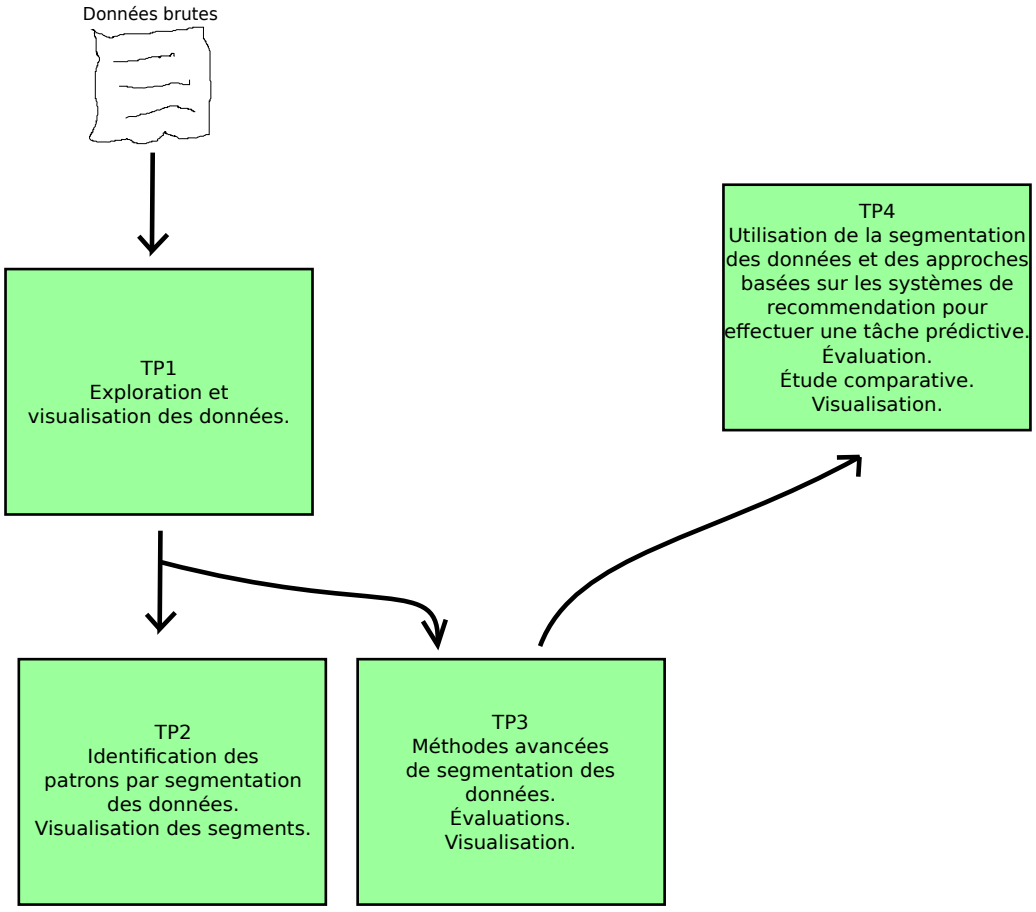

Figure 4: Étapes de réalisation du projet.

2 TP1 : Visualisation des données

Pour le TP1 (première phase du projet), l'équipe étudiante, dépendemment du choix de son jeu de données, devra mener une étude explorative et faire ressortir des informations statistiques pertinentes et utiles pour la suite du projet. Il est donc question de choisir et d'appliquer des methodes d'analyse et de tracer des graphiques afin de faire parler les données. Pour ce TP, les exigences sont décrites ci-dessous. **La description pour le jeu de données ABR est plus détaillée, les descriptions pour les autres jeux de données est une adaptation abrégée de celle pour ABR**.

1. Jeu de données ABR:
   (a) Indépendemment de la durée du temps durant lequel chaque livre est évalué, on s'intéresse à la question de savoir si chaque livre est apprécié ou pas par les gens. En supposant que les scores 5 et 4 signifient que le livre est très apprécié par la personne, le score 3 plus ou moins apprécié, et les scores 2 et 1 pas apprécié, quels sont les livres les plus (ou les moins) appréciés? Ou encore entre deux livres quelconques, lequel est plus apprécié. Pour répondre à ces questions, vous devez construire une matrice de données (de scores) contenant 5 lignes et autant de colonnes qu'il y a des livres. Chaque élément de la matrice va correspondre au nombre obtenu pour chaque score s ($s = 1, 2, 3, 4, ou, 5$). Vous appliquez des mesures suivantes au moins une fois chaque, la somme totale, la moyenne ou la moyenne pondérée, l'écart-type, la médiane (et des quartiles), le max, le min pour répondre aux questons suivantes :
   1) Quelle est la moyenne de score de chaque livre (ou est-ce que le livre est en général apprécié ou pas) ?
   2) Quels sont le(s) livre(s) le(s) mieux apprécié(s) et le(s) moins apprécié(s)?
   3) Quels sont le 1er quart des livres les plus approciés ?
   4) Entre deux livres, lequel est mieux apprécié?
   5) Est-ce que l'utilisation des comparaisons de scores moyennes est toujours une bonne façon de faire pour répondre à ces questions? Sinon, quels sont les alternatives?.
   6) Faire un diagramme en moustaches (*box plot*) affichant les tendances (les étendus) statistiques pour chacune des scores 1, 2, 3, 4 et 5 et interpéter la figure.
   Il est à noter qu'en général, "répondre à une question" ici signifie que vous proposez et implantez une méthode dans votre code pour générer des données pouvant répondre à la question et expliquer bièvement votre méthode dans le rapport du TP. Il ne s'agit pas toujours de fournir des données/résults de calcul dans le rapport pour répondre à la question surtout quand la réponse complète inclue beaucoup de valeurs (vous pouvez certainement donner LE livre le plus apprécié dans votre rapport, mais ne devez pas lister le quart des livres les mieux appréciés).
   (b) À partir de la représentation pour chaque livre construite dans la Section (a), c'est-à-dire la matrice des scores,
   1) faire une analyse en composantes principales pour représenter chaque livre par la projection sur deux premières composantes principales et afficher le

nuage de points représentant les livres sur un plan.
2) En utilisant une des alternatives que vous avez décrits en Section (a)-5), colorier votre nuage de points projetés en trois groupes distincts. Un groupe représentant les livres les moins appréciés, un autre les livres plus-ou-moins appréciés et un dernier groupe représentant les livres les plus appréciés. À titre illustratif, vous pouvez par exemple supposer qu'un livre apprécié a une moyenne d'appréciation strictement supérieure à 3.5, un livre plus-ou-moins apprécié a une moyenne d'appréciation comprise entre [2.5, 3.5] et un livre non-apprécié a une moyenne d'appréciation strictement inférieure à 2.5. Dessiner l'histogramme ou le diagramme en bâtons illustrant la proportion des livres appréciés, plus-ou-moins appréciés et non-appréciés.
3) Dessiner le triangle dont chaque côté relie deux centres des groupes. Selon vous, ces centres permettraient-ils de respectivement représenter chacun des groupes de livres?

(c) Sachant que les livres ne sont pas tous évalués en même temps, on aimerait savoir comment des tendances statistiques mesurant les opinions évoluent mensuellement. Pour une durée annuelle, refaire les questions 1) 2) et 3) de la tâche demandée à la Section (a) à chaque mois. On s'attend ici à avoir 12 matrices, vous n'êtes pas obligés de faire les diagrammes en moustaches dans ce cas figure. Pour chaque mois, projeter chacune des matrices dans le même espace vectoriel trouvé à la Section (b). On s'attend ici à avoir, pour chaque livre, 12 vecteurs à 2 dimensions. Analyser visuellement et commenter s'il y a des sous-structures qui se developpent dans le temps.

2. Jeu de données CovT.

(a) Dans ce jeu de données, l'émotion d'une personne à un moment donné est représentée par un vecteur de 5 attributs désignant la joie, la colère, la valence, la peur et la tristesse. À chacun des 5 attributs joints ensemble, on a un sentiment qui pourrait être négatif, positif ou neutre. Indépendemment du temps, construire les matrices de données pour répondre aux questions suivantes :
1) Quelle est la moyenne et la matrice de covariance des émotions de chacun des trois (classes de) sentiments : *negative*, *neutral* et *positive* ?
2) Afficher les cartes de chaleur (*heatmap*) correspondant aux matrices de covariances des trois sentiments. Faites une interprétation de la correlation des émotions vis-à-vis de chacun des sentiments.
3) Quelles sont les 10 personnes qui ont des sentiments les plus negatifs ?
4) Pour cet ensemble de données, quelle combinaison des mesures statistiques permet-elle de mieux décrire chacune des 5 émotions : moyenne + écartype ou médiane + IQR ? Vous pouvez répondre à cette question en examinant l'ensemble des données tous sentiments confondus ou en examinant les données appartenant à chacune des classes de sentiment séparément.
5) Entre deux personnes, comment allez-vous déterminer qui est plus positive ?
6) Faire un diagramme en moustaches (*box plot*) affichant les sentiments *neg-*

ative, *neutral* et *positive* et interpéter la figure.

(b) À partir de la matrice des données construite,
1) faire une analyse en composantes principales selon les 5 attributs d'émotions et afficher le nuage de points représentant le mieux possible les sentiments selon deux composantes principales. Attention, il ne s'agit pas nécessairement des deux premières composantes principales. Vous devez essayer des combinaisons de deux composantes afin de trouver une bonne (pour ne pas dire la meilleure) combinaison permettant de bien séparer visuellement les personnes exprimant de différents sentiments.
2) Colorier le nuage de points suivant trois couleurs distinctes et examiner si ces groupes se séparent les uns des autres. Tracez des segments de droites reliant les centres des trois groupes.
3) En considérant que les groupes se séparent bien dans votre plan, comment selon vous, vous pourriez déterminer qu'une personne est plus négative qu'une autre ? De même comment pourriez-vous déterminer qu'une personne est plus positive qu'une autre ?

(c) Sachant que les utilisateurs pourraient avoir des sentiments qui varient dans le temps, on aimerait savoir comment les statistiques associées évoluent mensuellement. Pour une durée annuelle, refaire (a) - 1), (a) - 2), (a) - 3) de la tâche demandée en point (a) à chaque mois. Pour chaque mois, effectuer les anaylses demandée au point (b) à l'exception de recalculer les composantes principales. Utilisez les même axes de projection obtenus en (b) tout au long de cette analyse temporelle.

3. Jeu de données NYCT:

(a) Dans ce jeu de données, on a les voyages effectués par les taxis dans la ville de NYCT sur une durée de 4 mois (de septembre à décembre 2013). Pour chaque voyage effectué par un taxi, on s'intéresse aux informations liées à la durée du parcours, le nombre de passagers et la distance parcourue par le taxi. On voudrait étudier l'activité du transport en taxis dépendemment de la période de la journée. En supposant qu'une journée est divisible en quatre quarts: nuit (Q1: de 00h00 à 05h59), matin (Q2: de 06h00 à 11h59), après-midi (Q3: de 12h00 à 17h59) et soir (Q4: de 18h00 à 23h59).
1) Construire une matrice dont chaque ligne reportera les informations telles que le nombre de passagers transporté, la durée des trajets et la distance des trajets.
2) Faire un classement des périodes les plus actives au moins actives de la journée. Expliquez votre critère de classement. Faire un diagramme en bâton illustrant le classement suivant votre critère.
3) Suivant le quart de la journée, durant quelle période dans la journée les taxis roulent plus rapidement (il est question ici de déterminer la vitesse moyenne par voyage et faire un classement) ?
4) Suivant le quart de la journée le plus actif, déterminez la vitesse moyenne par voyage effectuée par chaque taxi. Subdivisez les vitesses en trois classes

distinctes. On supposera que les classes représenteront les groupes de taxis *express*, *réguliers* et *lents*.

(b) Suivant le quart de la journée le plus actif, extraire la matrice résumant le nombre de passagers transportés, la durée des trajets et la distance des trajets effectués par un taxi.

1) Faire une analyse en composantes principales. Vous devez essayer des combinaisons de deux composantes afin de trouver une bonne combinaison permettant de bien séparer visuellement les taxis.

2) Colorier le nuage de points par trois couleurs distinctes (représentant la catégorie de vitesse des taxis) et examiner si ces groupes se séparent les uns des autres.

(c) Sachant que les taxis pourraient avoir des parcours différents dans le temps, on aimerait savoir comment ses tendances statistiques évoluent hebdomadairement. Sur 4 mois, refaire la tâche demandée en (a) à chaque semaine. Pour chaque semaine, projeter chacune des matrices dans le même espace vectoriel trouvé en (b).

4. Jeu de données ELF:

(a) Pour ce jeu de données, l'équipe étudiante devra construire la matrice qui fera ressortir les tendances statistiques sur la consommation électrique des différents clients. On voudrait ici, pour chaque client, connaître sa consommation électrique suivant les quarts de nuit (Q1: de 00h00 à 05h59), du matin (Q2: de 06h00 à 11h59), de l'après-midi (Q3: de 12h00 à 17h59) et du soir (Q4: de 18h00 à 23h59).

1)Faire un diagramme en moustaches (*box plot*) affichant les tendances statistiques pour chacun des quarts Q1, Q2, Q3 et Q4.

2) À quelle période de la journée on a les plus hauts pics de consommation électrique? Faire un classement des périodes suivant la totalité des consommation électriques. Illustrer par un diagramme en bâton votre classement.

3) En ignorant les quarts de la journée, déterminer la consommation moyenne effectuée par chaque client. Subdiviser ces consommations moyennes en trois groupes. Un groupe représentant les clients à consommations hautes, normales et faibles. Afficher la distribution des trois classes.

4) Refaire les mêmes calculs que dans la question précédente pour chacun des quarts. Les 10 premiers clients à consommations hautes observés dans la question (a)-3) sont-ils les même à chaque quart de la journée ? Sinon, comment expliquerez-vous le fait que cela pourrait varier par quart de la journée ?

(b) Dans un premier temps en supposant que l'on ignore les quarts de la journée, effectuer une analyse en composante principale sur les consommations électriques des clients.

1) Colorier le nuage de points obtenus en trois couleurs distinctes. Les couleurs représentant chacun des groupes consommations hautes, normales et faibles. Les groupes se séparent-ils des uns des autres ?

2) Dans un deuxième temps, en tenant compte des quarts de la journée, refaire la même opération que dans la question précédente pour chacun des quarts

de la journée. À chaque quart, retrouve-t-on les mêmes groupes que dans la question (b)-1).

(c) Sachant que le profil de consommation électrique d'un client pourrait drastiquement changer dans le temps, on aimerait savoir comment ces tendances statistiques évoluent mensuellement. Pour une durée annuelle, refaire la tâche demandée en (a) à chaque mois (on s'attend ici à avoir 12 matrices, vous n'êtes pas obligés de faire les diagrammes en moustaches dans ce cas figure). Pour chaque mois, projeter chacune des matrices dans le même espace vectoriel trouvé en (b) (on s'attend ici, à avoir, pour chaque client 12 vecteurs à 2 dimensions). Il est à noter que, pour cette question vous pouvez ignorer la notion des quarts dans une journée. Vous devez tout simplement prendre le mois au complet.

Livrable: Au terme de ce TP1, l'équipe étudiante devra retourner deux documents. Le premier document *Rapport-TP1-IFTX-Prenom-Nom-Prenom-Nom.pdf* avec $X \in \{599, 799\}$ (à la place de du *Prenom-Nom*, vous pouvez utiliser votre *cip*) contenant le rapport détaillé du TP1. Ce document devra contenir une explication des méthodes employées/proposées, de même qu'une interprétation des résultats. Toujours dans ce document, la personne étudiante devra expliquer brièvement comment rouler son code. Le deuxième document est le code et doit être nommé comme *Code-TP1-IFTX-Prenom-Nom-Prenom-Nom.extension_code* (avec par exemple *.extension_code = .py* pour un code fait en python ou *.extension_code = .ipynb*) pour un code python fait dans un notebook.

Les données du TP1 se trouvent dans le répertoire public du cours. La remise du TP1 est due au samedi 8 octobre 2022 et doit être effectuée par le système "turnin" du Département d'informatique (https://turnin.dinf.usherbrooke.ca/)