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

December 3, 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.

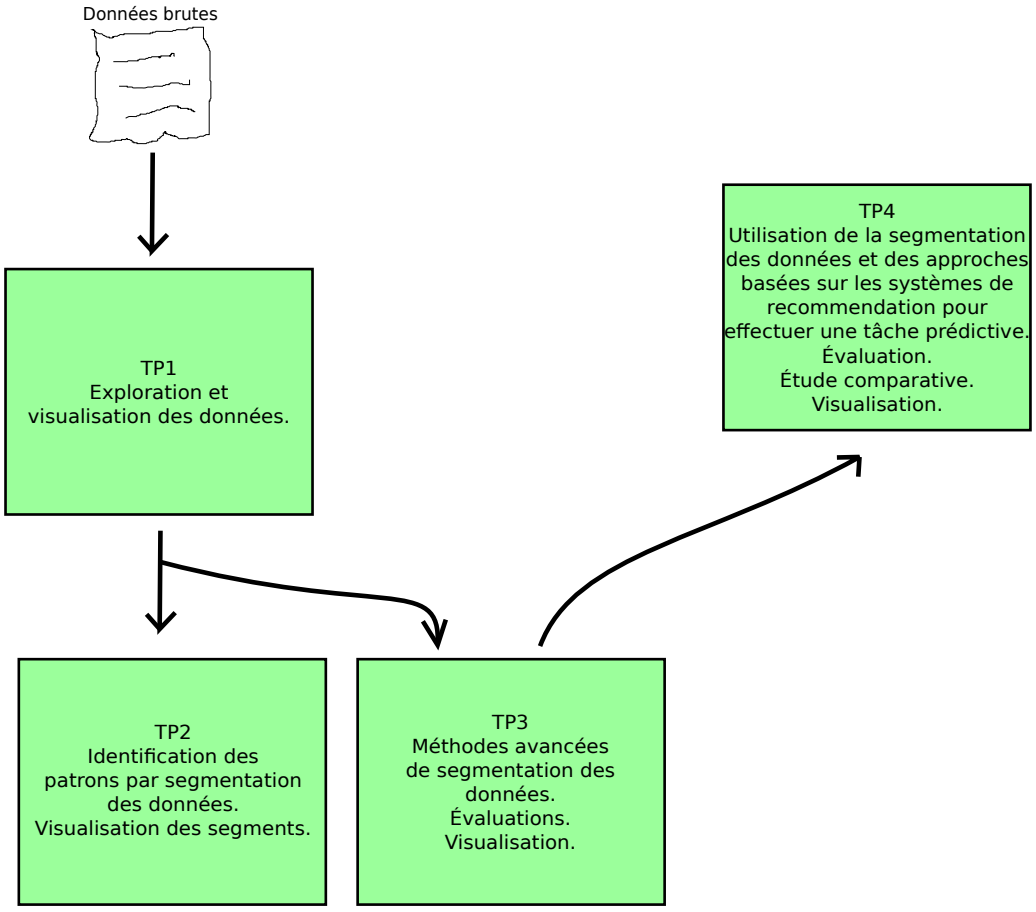

Figure 4: Étapes de réalisation du projet.

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/)

3 TP2 & 3: Segmentation des données et visualisation.

Jusqu'ici vous avez fait une exploration des données, une analyse descriptive, vous permettant d'avoir une idée des données que vous avez traitées. Gardant en esprit que l'on veut au final effectuer une analyse prédictive, on voudrait savoir si le jeu de données étudié ne présente pas des sous-structures pertinentes qui pourraient éventuellement nous aider dans notre tâche de prédiction finale. Pour ce faire, on voudrait segmenter nos données suivant différentes stratégies de clustering. Cependant, vue la taille des données, il serait difficile de les charger en mémoire et rouler un algorithme quelconque de segmentation.

1. Jeu de données ABR:
   (a) Indépendemment de l'aspect temporel, on voudrait exploiter les caractéristiques statistiques incluant le nombre de personnes ayant voté, la moyenne, l'écart-type, et la médiane des scores, le nombre de fois qu'un livre a été apprécié (le nombre de fois le livre a obtenu des scores 4 ou 5), le nombre de fois que le livre n'a pas été apprécié (obtenant des scores 1 ou 2) et le nombre de fois que le choix a été neutre en obtenant le score 3 que vous avez extraites dans

votre TP1 pour segmenter les données. Sélectionner de manière aléatoire une quantité (p) suffisamment grande (selon la capacité de votre machine en terme of la quantité de mémoire) de livres.
1) Construire la matrice de données $X \in R^{p \times c}$ avec c étant le nombre de caractérisitques statistiques extraites.
2) Effectuer une segmentation basée sur les k-moyennes avec $k = 3$ en utilisant dans un premier temps la distance euclidienne et ensuite la similarité cosinus pour comparer les objets.
3) En utilisant l'analyse en composante principale, projectez vos segments obtenus suivant les deux premiers composantes principales. De manière visuelle, les segments sont-ils différents avec la distance euclidienne par rapport aux segments obtenus avec la similarité cosinus? Faire une interprétation de vos résultats.
4) Effectuer une segmentation par le clustering spectral sur deux dimensions en prenant $k = 3$ avec pour distances: la distance euclidienne ensuite la similarité cosinus. Pour ce faire, utilisez d'abord la distance euclidienne pour construire une matrice de similitude M, puis décomposer cette matrice en $M = P\sqrt{D}(P\sqrt{D})^T$ avec D soit la matrice diagonale dont les éléments diagonaux sont des valeurs propres du M organisées en ordre décroissante. Utiliser les deux premiers colones de la matrice $P\sqrt{D}$ pour représenter chaque "livre" et faire une clustering par k-moyenne. Ensuite répéter le même avec similitude cosinus. Au sens visuel, les résultats obtenus sont-ils différents des résultats obtenus dans en 3)?
5) Utilisez les métriques suivantes pour évaluer la qualité de votre segmentation (dans le cas des k-moyennes et le cas spectral) lorsque vous utilisez dans un premier temps une distance euclidienne et ensuite dans un deuxième temps la similarité cosinus: la silhouette et l'information mutuelle. Vous pourrez vous référer respectivement à ces endroits: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`, et `https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation`. Reportez vos résultats dans un seul tableau comme suit et faire une interprétation des résultats:

	distance euclidienne		similarité cosinus	
Approche	Silhouet.	Info. Mut.	Silhouet.	Info. Mut.
K-moyennes				
Spectrale				

(b) On voudrait à présent tenir compte de la totalité des données. Pour ce faire, on voudrait joindre aux différents segments déjà identifiés des nouvelles valeurs.
1) Selon vous quel serait le risque de prendre aléatoirement un sous-ensemble de données pour effectuer les tâches a-1) à a-5) ?
2) En procédant par une sélection stratifiée, on voudrait s'assurer que toutes les catégories soient représentées dans notre sous-ensemble. Pour chacune des trois catégories $1-2$, 3 et $4-5$, sélectionnez aléatoirement des quantités p_1, p_2, p_3, p_4 et p_5 de données tels que $p1 = p2 = p3 = p4 = p5$ et

$p = p1 + p2 + p3 + p4 + p5$ et reconstruire votre matrice X.
3) Refaire les étapes a-2) à a-5). Faite une comparaison des résultats reportés dans votre tableau à ceux reportés dans le tableau obtenu en a-5).
4) À supposer qu'on associe à chacune des catégories $1-2$, 3 et $4-5$ les étiquettes respectives l_1, l_2 et l_3. On voudrait retrouver les étiquettes des données restantes (celles qui n'ont pas été prises en considération lors de la segmentation). Quelle stratégie pensez-vous utiliser, expliquer ?

2. Jeu de données CovT : Refaire le même travail demandé dans le jeu de données ABR.
3. Jeu de données NYCT :
   (a) Dans le TP1, vous avez identifié trois grandes catégories de conduite à savoir les conduites *express*, *régulière* et *lente*. Selon le quart de la journée le plus actif, on décide de l'échantillonner à une fréquence de 30 minutes et observer la vitesse moyenne du conducteur à chaque intervalle de temps dans ledit quart. Sélectionner un nombre p suffisamment grand (facilement gérable par la taille de la mémoire de votre machine) de conducteurs.
   1) Construire une matrice $V \in R^{p\times 12}$ similaire à la table suivante pour le quart de Q_2 par exemple :

	T_1	T_2	T_3	...	T_{12}
	06:00 - 06:30	06:30 - 07:00	07:00 - 07:30	...	11:30 - 12:00
C_1	$v_{1,1}$	$v_{1,2}$	$v_{1,2}$	...	$v_{1,12}$
C_2	$v_{2,1}$	$v_{2,2}$	$v_{2,3}$	...	$v_{2,12}$
...	...	...	...	...	...
C_p	$v_{p,1}$	$v_{p,2}$	$v_{p,3}$	...	$v_{p,12}$

   Chaque valeur $v_{i,j}$, $1 \leq i \leq p$, $1 \leq j \leq 12$ de cette matrice représente la vitesse moyenne du chauffeur C_i à l'intervalle de temps T_j.
   2) Effectuer une segmentation basée sur les k-moyennes avec $k = 3$ en utilisant dans un premier temps la distance euclidienne et ensuite la similarité cosinus pour comparer les conducteurs. Dans chacun des cas, peut-on clairement dissocier les différents types de conducteurs (*express*, *régulier* et *lent*) ?

   Vu que nous avons également trois catégories comme dans le jeu de données ABR, refaire pour le cas des conducteurs toutes les étapes $1.(a)-3)$ à $1.(a)-5)$ ensuite les étapes $1.(b)-1)$ à $1.(b)-4)$.
4. Jeu de données ELF:
   (a) Tout comme dans le jeu de données NYCT, nous avons quatre quarts dans une journée Q_1, Q_2, Q_3 et Q_4. Suivant le quart de la journée suivant lequel on a les plus hauts pics de consommation, on vous demande de regarder la consommation électrique aux 30 minutes. Sélectionner un nombre p suffisamment grand de consommateurs.
   1) Construire une matrice $E \in R^{p\times 12}$ similaire à la table suivante pour le quart de Q_2 par exemple :

	T_1	T_2	T_3	...	T_{12}
	06:00 - 06:30	06:30 - 07:00	07:00 - 07:30	...	11:30 - 12:00
C_1	$e_{1,1}$	$e_{1,2}$	$e_{1,2}$	...	$e_{1,12}$
C_2	$e_{2,1}$	$e_{2,2}$	$e_{2,3}$	...	$e_{2,12}$
...	...	...	...	...	...
C_p	$e_{p,1}$	$e_{p,2}$	$e_{p,3}$	...	$e_{p,12}$

Chaque valeur $e_{i,j}$, $1 \leq i \leq p$, $1 \leq j \leq 12$ de cette matrice représente la consommation totale du client C_i à l'intervalle de temps T_j.
2) Effectuer une segmentation basée sur les k-moyennes avec $k = 3$ en utilisant dans un premier temps la distance euclidienne et ensuite la similarité cosinus pour comparer les conducteurs. Dans chacun des cas, peut-on clairement dissocier les différents types de clients (*à faible consommation*, *à consommation normale* et *à consommation haute*) ?

Vu que nous avons également trois catégories comme dans le jeu de données ABR, refaire pour le cas des clients toutes les étapes $1.(a) - 3)$ à $1.(a) - 5)$ ensuite les étapes $1.(b) - 1)$ à $1.(b) - 4)$.

Les données du TP2_3 se trouvent dans le répertoire public du cours. La remise du TP2_3 est due au mardi 29 novembre et doit être effectuée par le système "turnin" du Département d'informatique (https://turnin.dinf.usherbrooke.ca/)

4 TP4: Prédiction basée sur un principe de collaboration; évaluation hors-ligne.

Étant données un jeu données pour lequel nous avons des entités qui pourraient prendre différentes valeurs. On entend par *collaboration*, le fait qu'il y aurait des similitudes entres certaines entités. Dans le cadre de ce TP, on voudrait tirer profit de la connaissance des collaborations entre entités pour effectuer une tâche prédictive.

1. Jeu de données ABR:
   Dans ce jeu de données, vous travaillerez uniquement avec un sous-échantillon prélevé de votre ensemble de données.
   (a) De manière stratifié, sélectionnez p ($p \geq 1.000$) livres y compris tous les n ($n \geq 2.000$) *reviewers* qui ont participé aux *ratings* de ces livres. Vous devez avoir en fin de compte un dataframe similaire à celui ci-dessous:

$Fr_1 =$

	$livre_1$	$livre_2$	$livre_3$	...	$livre_p$
$reviewer_1$	$r_{1,1}$	$r_{1,2}$	$r_{1,3}$	...	$r_{1,p}$
$reviewer_2$	$r_{2,1}$	$r_{2,2}$	$r_{2,3}$	...	$r_{2,p}$
...	...	...	...	...	...
$reviewer_n$	$r_{n,1}$	$r_{n,2}$	$r_{n,3}$	...	$r_{n,p}$

   où $r_{i,j} \in \{1, 2, 3, 4, 5\}$, $1 \leq i \leq n$, $1 \leq j \leq p$, est le *rating* que le $reviewer_i$ a donné au $livre_j$.

Il est important de noter que dans le dataframe Fr_1 on pourrait avoir des cases vides dû au fait que certains *reviewers* n'auraient pas fait des *ratings* pour certains livres.
Vous devez donc vous assurez que, lors de la sélection de votre sous-échantillon, le nombre de cases vides ne soit pas statistiquement significatif (le nombre de cases vides par colonne doit être négligeable par rapport au nombre de cases non-vides).
Effectuer un *heatmap* plot vous permettant d'apprécier de manière visuelle le sous-ensemble de données que vous avez sélectionné.

(b) Maintenant que vous avez votre jeu de données Fr_1, on vous demande d'extraire 3 sous-ensembles de données. Construire une fonction *DataSelection()* qui vous permettra d'exécuter les instructions ci-dessous :
   i. Pour chacune des lignes i de Fr_1, masquer/cacher de manière aléatoire quelques *ratings* effectués par le $reviewer_i$. Vous obtiendrez un nouveau dataframe Fr_2 où il y aurait plus de cases vides que celui de Fr_1.
   ii. Repétez l'étape 1(b)i trois (03) fois de suite et retourner 3 sous-ensembles d'entrainement $Train\_Sets = \{Fr^1_{2,train}, Fr^2_{2,train}, Fr^3_{2,train}\}$.

(c) À l'aide du coefficient de Pearson Correlation, pour chacun des sous-ensembles $Fr^k_{2,train}$, $1 \leq k \leq 3$:
   i. Calculer les matrices de similitudes $B^k \in R^{p\times p}$ des livres et les matrices de similitudes $U^k \in R^{n\times n}$ des reviewers. Faire un *heatmap* plot vous permettant d'apprécier la similitude entre les livres et les reviewers.
   ii. Pour chacun des $livre_j$ dans $Fr^k_{2,train}$, extraire les ensembles $\boldsymbol{I}^k_{\sim j}$ des livres les plus similaires au $livre_j$ tel que $|\boldsymbol{I}^k_{\sim j}| = 3$.
   iii. Pour chacun des $reviewer_i$ dans $Fr^k_{2,train}$, extraire les ensembles $\boldsymbol{U}^k_{\sim i}$ des reviewers les plus similaires au $reviewer_i$ tel que $|\boldsymbol{U}^k_{\sim i}| = 8$.

(d) On suppose que le *rating* $r_{i,j}$ donné par un $reviewer_i$ sur un $livre_j$ peut en tout temps être calculé par la fonction $\mathcal{R}(livre_j|reviewer_i, \Omega)$ donnée comme suit,

$$\mathcal{R}(livre_j|reviewer_i, \Omega) = \left\| (\omega_1, \omega_2, \omega_3, \omega_4, \omega_5) \cdot \sum_{livre_{j^*} \in \boldsymbol{I}^k_{\sim j}} Z^k_* \right\|$$

avec $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$, l'ensemble des paramètres.
$Z^k_* \in R^{5\times 8}$ est une matrice dont les éléments $z^k_{i^*,j^*}$ sont définis comme suit:

$$z^k_{i^*,j^*} = \begin{cases} r_{i^*,j^*} \textit{ si le } reviewer_{i^*} \textit{ a une appréciation au } livre_{j^*} \\ 0 \textit{ sinon} \end{cases}$$

On voudrait minimiser l'erreur qu'on commet lorsqu'on génère les *ratings* avec $\mathcal{R}()$. Pour ce faire on se donne la fonction d'estimation suivante:

$$\mathcal{E}val(\mathcal{R}(livre_j|reviewer_i, \Omega), r_{i,j}) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} |\mathcal{R}(livre_j|reviewer_i, \Omega) - r_{i,j}|$$

i. Calculer le gradient de $\mathcal{E}val(\Omega)$
ii. Pour chacun des sous-ensembles $Fr^k_{2,train}$, suivant la méthode de descente des gradients estimer les paramètres $\hat{\Omega}^k$.
iii. À partir de la valeur de $\hat{\Omega}^k$, calculer les *ratings* cachés en utilisant la fonction de $\mathcal{R}()$
iv. En utilisant l'erreur quadratique RMSE (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html#sklearn.metrics.mean_squared_error`) et l'erreur absolue moyen MAE (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html#sklearn.metrics.mean_absolute_error`) construire le dataframe ci-dessous:

	RMSE	MAE
Test-1		
Test-2		
Test-3		

Calculer la moyenne de l'erreur commise ainsi que l'écart-type. Faire un diagramme en bâton les illustrant et faire une interprétation.

2. Jeu de données CovT:
Dans ce jeu de données, vous travaillerez uniquement avec un sous-échantillon prélevé de votre ensemble de données.
(a) Sachant qu'il y a des utilisateurs inactifs dans ce jeu de données. On vous demande de regarder vos données de manière bi-hebdomadaire et ainsi sélectionner le sentiment dominant d'un utilisateur aux deux semaines. On vous demande de construire le nouveau dataframe:

$Fr_1 =$

	bw_1	bw_2	bw_3	...	bw_p
u_1	$s_{1,1}$	$s_{1,2}$	$s_{1,3}$	...	$s_{1,p}$
u_2	$s_{2,1}$	$s_{2,2}$	$s_{2,3}$	...	$s_{2,p}$
...	...	...	...	...	...
u_n	$s_{n,1}$	$s_{n,2}$	$s_{n,3}$	...	$s_{n,p}$

où $s_{i,j} \in \{1, 2, 3\}$, $1 \leq i \leq n$, $1 \leq j \leq p$, est le sentiment le plus fréquent de l'utilisateur u_i durant les deux semaines bw_j. La valeur 1 est prise comme le sentiment négatif, 2 sentiment neutre et 3 sentiment positif.
Il est important de noter que dans le dataframe Fr_1 on pourrait avoir des cases vides dû au fait que certains utilisateurs n'auraient pas été actif sur la plateforme Twitter au courant de certaines semaines.
Vous devez donc vous assurez que, lors de la sélection de votre sous-échantillon, le nombre de cases vides ne soit pas statistiquement significatif (le nombre de cases vides par ligne doit être négligeable par rapport au nombre de cases non-vides de cette ligne).
Effectuer un *heatmap* plot vous permettant d'apprécier de manière visuelle le sous-ensemble de données que vous avez sélectionné.
(b) On voudrait se baser sur le principe de filtrage collaboratif pour prédire le sentiment de certains utilisateurs lors des 10 dernières *bi-semaines* (les 10

dernières colonnes de votre dataframe). On vous demande de cacher les sentiments des utilisateurs (actifs) lors de ces dernières semaines et obtenir un nouvel ensemble de données Fr_2.

(c) En vous référant uniquement aux colonnes avant les 10 dernières et en utilisant le coefficient de Pearson Correlation,

i. Calculer les matrices de similitudes $B \in R^{(p-10)\times(p-10)}$ entre les *bi-semaines* et $U \in R^{n\times n}$ des utilisateurs.

ii. Faire un *heatmap* plot vous permettant d'apprécier la similitude entre les semaines et les utilisateurs.

iii. Pour chacune des périodes bw_j, $1 \leq j \leq p-10$ extraire l'ensemble $\boldsymbol{I}_{\sim j}$ des périodes les plus similaires à bw_j tel que $|\boldsymbol{I}_{\sim j}| = 10$.

iv. Pour chacun des utilisateurs u_i, extraire l'ensemble $\boldsymbol{U}_{\sim i}$ des utilisateurs les plus similaires à l'utilisateur u_i tel que $|\boldsymbol{U}_{\sim i}| = 20$.

(d) On suppose que le sentiment $s_{i,j}$ donné par un utilisateur u_i durant une période bw_j peut en tout temps être calculé par la fonction $\mathcal{R}(bw_j|u_i, \Omega)$ donnée comme suit,

$$\mathcal{R}(bw_j|u_i, \Omega) = \left\| (\omega_1, \omega_2, \omega_3) \cdot \sum_{bw_{j^*} \in \boldsymbol{I}_{\sim j}} Z_* \right\|$$

avec $\Omega = \{\omega_1, \omega_2, \omega_3\}$, l'ensemble des paramètres.
$Z_* \in R^{3\times 20}$ est une matrice dont les éléments z_{i^*,j^*} sont définis comme suit:

$$z_{i^*,j^*} = \begin{cases} s_{i^*,j^*} \textit{ si l'utilisateur } u_{i^*} \textit{ a un sentiment durant la période } bw_{j^*} \\ 0 \textit{ sinon} \end{cases}$$

On voudrait minimiser l'erreur qu'on commet lorsqu'on génère les sentiments avec $\mathcal{R}()$. Pour ce faire on se donne la fonction d'estimation suivante:

$$\mathcal{E}val(\mathcal{R}(bw_j|u_i, \Omega), r_{i,j}) = \frac{1}{n(p-10)} \sum_{i=1}^{n} \sum_{j=1}^{p} |\mathcal{R}(bw_j|u_i, \Omega) - r_{i,j}|$$

i. Calculer le gradient de $\mathcal{E}val(\Omega)$

ii. Suivant la méthode de descente des gradients estimer les paramètres $\hat{\Omega}$.

iii. À partir de la valeur de $\hat{\Omega}$, prédire les sentiments des utilisateurs dont vous avez caché les valeurs lors des 10 dernières périodes en utilisant la fonction $\mathcal{R}()$

iv. En utilisant l'erreur quadratique RMSE (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html#sklearn.metrics.mean_squared_error`) et l'erreur absolue moyen MAE (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html#sklearn.metrics.mean_absolute_error`) construire le dataframe ci-dessous:

	RMSE	std-RMSE	MAE	std-MAE
Test-1				
Test-2				
...	...	...		
Test-10				

Où Test-1 est l'erreur moyen lorsqu'on prédit une période (de deux semaines) à l'avance, Test-2 lorsqu'on prédit deux périodes à l'avance, ..., Test-10 l'erreur qu'on prédit dix périodes à l'avance. std-RMSE (resp. std-MAE) est l'écart-type sur l'erreur. Faire un diagramme en bâton les illustrant et faire une interprétation.

3. Jeu de données NYCT:
   On voudrait prédire le mode de conduite des conducteurs aux heures. Dans les TPs précédents suivant un critère vous aviez été capable de déterminer trois catégories de conduites à savoir *lente, régulière et express*. Dans ce Tp on codera la catégorie lente par la valeur 1, régulière la valeur 2 et express la valeur 3. Pour une période allant de 05:00 à 20:00 sur toutes les dates, contruire le dataframe:

$Fr_1 =$

	T_1	T_2	T_3	...	T_p
c_1	$m_{1,1}$	$m_{1,2}$	$m_{1,3}$	...	$m_{1,p}$
c_2	$m_{2,1}$	$m_{2,2}$	$m_{2,3}$	...	$m_{2,p}$
...	...	...	...	...	...
c_n	$m_{n,1}$	$m_{n,2}$	$m_{n,3}$	...	$m_{n,p}$

   où $m_{i,j} \in \{1, 2, 3\}$, $1 \leq i \leq n$, $1 \leq j \leq p$, est le mode de conduite dominant du conducteur c_i durant la période T_j.
   Il est important de noter que dans le dataframe Fr_1 on pourrait avoir des cases vides dû au fait que certains conducteurs n'auraient pas été actif durant certaines périodes.
   Vous devez donc vous assurez que, lors de la sélection de votre sous-échantillon, le nombre de cases vides ne soit pas statistiquement significatif (le nombre de cases vides par ligne doit être négligeable par rapport au nombre de cases non-vides de cette ligne).
   Effectuer un *heatmap* plot vous permettant d'apprécier de manière visuelle le sous-ensemble de données que vous avez sélectionné.
   Refaire toutes les étapes (b) à (d) dans le jeu de données CovT pour votre jeu de données.

4. Jeu de données ELF:

   Tout comme dans le jeu de données NYCT, ici vous avez trois catégories de consommateurs. *faibles, modérés* et *hautes* consommations dont on se propose de coder par les valeurs 1, 2 et 3 respectivement. On voudrait uniquement observer les consommations des clients entre 05h00 et 20h00.
   Appliquer le même travail que dans le jeu de données NYCT dans un but de prédire le type de consommation des clients lors de la dernière journée entre 05h00 et 20h00.

Il est important de noter qu'ici, vous aurez à prédire les 15 dernières heures (et non les 10 dernières périodes comme dans le jeu de données NYCT).

Les données du TP4 se trouvent dans le répertoire public du cours. La remise du TP4 est due au vendredi 23 décembre 2022 et doit être effectuée par le système "turnin" du Département d'informatique (https://turnin.dinf.usherbrooke.ca/).