Université de Sherbrooke

IFT 599 / IFT 799 - Science de données

TP2 : Clustering et visualisation des données

Automne 2024

Enseignants

	Courriel	Local	Téléphone
Shengrui Wang	shengrui.wang@usherbrooke.ca	D4-1018-1	+1 819 821-8000 x62022

Faculté des Sciences,
Département d'informatique

November 13, 2024

Sommaire

Dans le cadre de ce travail pratique (TP) est mis à la disposition des personnes étudiantes un jeu de données. Il est question ici, à partir de ce jeu de données, de mettre en exergue les conceptes de regroupement (*clustering*) vus aux thèmes 4 et 5. Plus précisément, ce TP consiste à chercher des combinaisons de plusieurs concepts et techniques pour comprendre et visualiser la séparation des données. À travers ce projet, vous allez acquérir une bonne capacité d'analyse et maîtriser quelques techniques de clustering.

Contents

1 Jeu de données et énoncé du problème

1.1 Jeu de données

De 2020 à 2021, la pandémie de COVID-19 a sévi dans de nombreux pays, causant des ravages. En raison de sa facilité de transmission, de nombreux gouvernements ont mis en place des mesures restrictives pour contenir la propagation rapide de la pandémie. En réponse à ces mesures, on a observé une augmentation significative de l'utilisation de plateformes de communication virtuelle, notamment les médias sociaux. Pendant cette période, de nombreux internautes ont exprimé leurs opinions et commentaires sur les décisions prises par les gouvernements pour lutter contre la pandémie. Sur Twitter, une analyse a été menée pour étudier les commentaires publiés par les internautes pendant cette période. À partir de ces commentaires, des caractéristiques ont été extraites pour fournir un aperçu des attitudes des internautes à l'égard de la situation. Il s'agit ici de :

- la valence (*valence_intensity*) qui fait référence à la dimension émotionnelle de l'internaute,
- la peur (*fear_intensity*),
- la colère (*anger_intensity*),
- la joie (*happiness_intensity*),
- la tristesse (*sadness_intensity*).

Les caractéristiques ici sont des valeurs numériques. À chaque vecteur de caractéristiques (valence, peur, colère, joie, tritesse) est associé un sentiment qui pourrait prendre l'une des valeurs $\{-1, 0, 1\}$, -1 pour signifier un sentiment négatif, 0 pour un sentiment neutre et 1 pour un sentiment positif.

Les données sont disponibles dans le répertoire travaux pratiques/TP2 du Teams du cours.

1.2 Énoncé du problème

On voudrait, à partir des cinq(5) caractéristiques extraites (hors mis les sentiments associés), rechercher les tendances que présentent les différents internautes. En d'autres termes, on voudrait savoir si les caractéristiques permettent de retrouver (ou bien *caractériser*) les trois tendances relatives aux sentiments que pourrait présenter un internaute. Pour ce faire, on segmente nos données suivant différentes stratégies de clustering.

2 Travail à faire

1. Dans un premier temps, on vous demande de faire une analyse comparative des cinq caractéristiques (valence, peur, colère, joie, tritesse). Plus précisément, vous analysez comment les distributions de ces caractéristiques se présentent en paire. La figure ci-dessous donne un exemple illustratif de representation conjointe de deux distributions suivant les caractéristiques de joie et colère. On vous demande de le

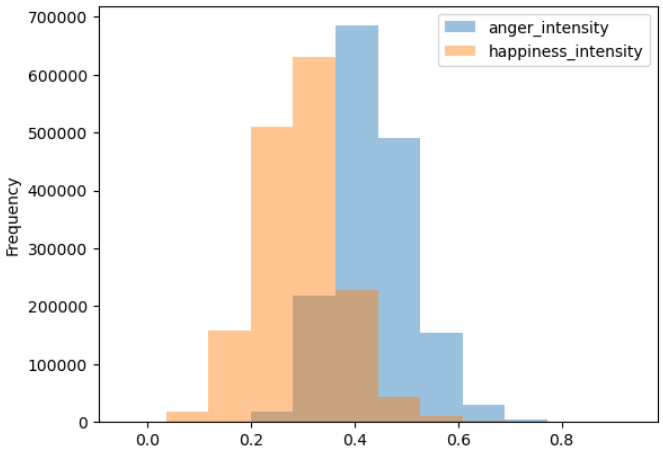

Figure 1: Exemple de representation conjointe de deux distributions.

faire pour toutes les cinq (05) caractéristiques (en somme vous devez avoir un total de dix (10) figures). Pour chacune des figures apportez une interprétation vis-à-vis des sentiments. En guise d'illustration, dans la Figure 1, une interprétation pourrait être : *du faite que les deux distributions sont très peu superposées, cela démontre que ces deux caractéristiques sont assez antagonistes et pourraient être exploitées pour reconnaître un sentiment négatif, positif ou neutre.*

Attention: la figure prise en exemple n'est pas complète. Assurez-vous d'avoir des figures plus lisibles que cet exemple. Tous les axes doivent être bien nommés y compris les titres des figures.

2. Ignorant l'information concernant les trois classes pertinentes (correspondant aux trois sentiments), on vous demande d'exploiter l'algorithme de K-means pour trouver les différents groupes d'internautes. Évidemment le nombre de groupes optimal n'est pas nécessairement 3. Afin de déterminer le nombre optimal de K, pour chacune des valeurs de $K = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ on vous demande de rouler votre K-means et évaluer la qualité des résultats obtenues selon un indice de validity. Deux indices sont à tester :
   - le critère d'Overlap utilisé dans le TP1 ;
   - le score Silhouette (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`)

   Dans un tableau, rapportez les résultats de l'Overlap et de la Silhouette, puis déterminez le nombre optimal de clusters selon chacun des deux indices.
   Finalement, pour $K = 3$ et $K =$ chacune des valeurs optimales déterminées par le critère d'Overlap et la Silhouette, faite une nouvelle évaluation pour savoir si les clusters retrouvés correspondent effectivement aux internautes présentant un sentiment négatif (-1), neutre (0) ou positif (1). Pour cette évaluation, on vous demande d'utiliser les critères de : Precision / Recall et F1-score (`https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html`). Tirez une conclusion sur vos résultats.
   Amendement : Comme il est fort probable que vous obtenez une valeur optimale de 2 ou 3 pour le nombre de clusters K, vous pouvez compléter l'évaluation uniquement pour $K = 3$. En option comme j'ai suggéré en classe, vous pourriez tester votre

programme de déterminaton du nombre optimal de clusters en utilisant des sous-ensembles d'attributs, e.g. 4 attributs au lieu de 5. Il y a probablement une chance d'arriver à un nombre optimal plus grand que 3. Si c'est le cas, rapportez le résultat concernant la détermination de cette valeur optimale de K, mais ne rapportez pas de résultats d'évaluation de Precision / Recall et F1-score pour cette valeur (de K) trouvée.

3. Après avoir testé votre clustering en utilisant le K-means, on vous demande d'utiliser une approche hierarchique. Vous avez la liberté de choisir entre une approche agglomérative ou une approche divisive. Pour déterminer le nombre de clusters, fixer un niveau de "coupe" soit intuitivement soit par une méthode automatique (de votre choix). ~~Si vous choisissez de fixer un niveau de "coupe" de façon intuitive, faite au moins deux choix qui vous semblent intéressants.~~ Si vous avez une bonne raison de fixer un niveau de "coupe" vous donnant $K > 3$, alors ~~Ensuite,~~ faite une nouvelle évaluation des clusters obtenus en fonction de Precision / Recall et F1-score comme dans la sous-question précédente. ~~Évitez que le nombre de clusters soit 3 car ce scénario fait partie de la sous question suivante.~~
~~Présentez les différents dendrogrammes correspondant. Dans vos différents seuils, y-a-t-il un vous permettant de trouver naturellement trois groupes d'internautes?~~
Présentez (une partie de) votre dendrogramme (structure hierarchique de vos clusters).
Attention: dans l'interprétation de vos résultats il est important de préciser le type de lien que vous avez utilisé ou la façon d'effectuer la division pour votre clustering hierarchique. Par ailleur vous devez travaillez avec la distance euclidienne.
4. En fixant le nombre de nombre de cluster à 3 dans votre clustering hierarchique, faite une nouvelle évaluation pour savoir si les clusters retrouvés correspondent effectivement aux internautes présentant un sentiment négatif (-1), neutre (0) ou positif (1) en utilisant les critères de Precision / Recall et F1-score. Tirez une conclusion sur vos résultats incluant par exemple une comparaison avec les résultats de clustering par K-means que vous avez obtenus à l'étape 2.

3 Remise du TP

- Le nombre maximal des membres d'une équipe est 3;
- La date de remise du TP est le vendredi 22 novembre 2024 à 23h59, aucun TP ne sera accepté après cette date;
- Soignez votre rapport, une pénalité de 5% pourrait être appliquée pour un rapport mal rédigé;
- Les fichiers à soumettre sont le rapport (en Word ou pdf) et l'ensemble de vos programmes. Ne pas soumettre les données!
- N'oubliez pas d'identifier les membres du groupe de travail. Indiquez les noms et cips (ou matricules) des membres du groupe dans chacun des fichiers que vous soumettez. La remise doit être faite par http://turnin.dinf.usherbrooke.ca