



INF1473 -Entreposage et prospection des données

Projet P2 : Clustering

Automne 2024

Enseignant

	Courriel	Local
Etienne Tajeuna	etiennegael.tajeuna@uqo.ca	A2222

DÉPARTEMENT D'INFORMATIQUE ET D'INGÉNIERIE (DII)

November 16, 2024

Sommaire

Dans le cadre de ce travail pratique (TP) est mis à la disposition des personnes étudiantes un (01) jeu de données. Il est question ici, à partir de ce jeu de données de mettre en exergue les concepts de *clustering* vu sur les thèmes portant sur le clustering des données. 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
1.1	Jeu de données	1
1.2	Énoncé du problème	1
2	Travail à faire	1
3	Remise du TP	3

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 et pour chaque vecteur de caractéristiques (valence, peur, colère, joie, tristesse) 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 ici (travaillez uniquement avec les données correspondant au Canada) : https://drive.google.com/file/d/1vrzc05Pa05TWMI159wsjS9phDovHeLM_/view?usp=sharing.

1.2 Énoncé du problème

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

2 Travail à faire

1. Dans un premier temps on voudrait faire une investigation sur les caractéristiques extraites (valence, peur, colère, joie, tristesse). Pour cela on vous demande dans un premier temps de regarder de manière conjointe comment les distributions de ces caractéristiques se présentent. La figure ci-dessous donne un exemple illustratif de

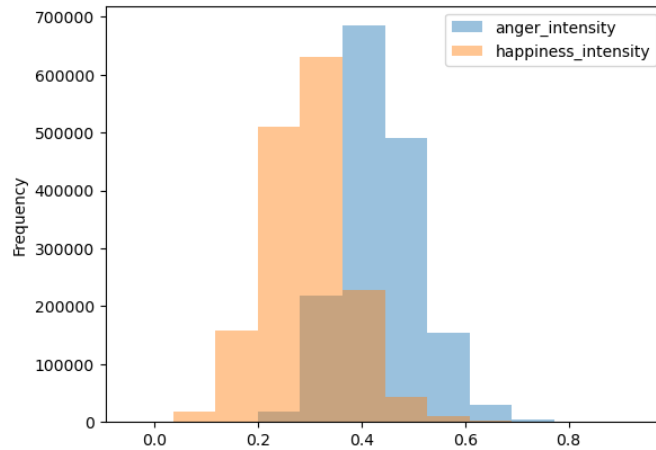


Figure 1: Exemple de representation conjointe de deux distributions.

representation conjointe de deux distributions suivant les caractéristiques de joie et colère. On vous demande de le 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 pourrait être exploité pour reconnaître un sentiment négatif, positif ou neutre.*

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

2. La question précédente vous a permis de regarder de manière conjointe (avec deux caractéristiques) si l'on pourrait tirer des conclusions sur les sentiments des internautes. Ignorant le fait qu'il existe trois classes pertinentes, on vous demande d'exploiter l'algorithme de K-means pour naturellement trouver les différents groupes d'internautes. Évidemment le problème est le choix du nombre 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.

Suivant la projection UMAP (en deux dimensions), présentez le nuage des points des neuf (09) clustering. Il est question ici de colorier vos points suivant les différents clusters.

Quelle interprétation tirez-vous de ces représentations graphiques?

3. Le critère visuel n'est généralement pas suffisant pour tirer une conclusion sur un clustering effectué. De manière numérique, on vous demande d'exploiter le critère d'Overlap utilisé dans le TP1 et le score silhouette (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html) pour evaluer vos résultats de clustering. Dans un tableau, rapportez les résultats de l'Overlap et de la silhouette puis faites une interprétation de vos résultats. Pour le cas particulier de $K = 3$, faite une nouvelle évaluation pour savoir si les clusters retrouvés correspondent effectivement aux internautes présentant un

sentiment négatif (-1), neutre (0) ou positif (1). 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) pour cette évaluation. Tirez une conclusion sur vos résultats.

- Après avoir testé votre clustering en utilisant le K-means, on vous demande d'utiliser une approche hiérarchique. Sans fixer le nombre de clusters, fixer différentes valeurs de seuil pour naturellement retrouver le nombre de K .

Présentez les différents dendrogrammes correspondant.

En utilisant les critères d'Overlap et silhouette, évaluer les clusters obtenus.

Dans vos différents seuils, y-a-t-il un vous permettant de trouver naturellement trois groupes d'internautes?

Attention: dans l'interprétation de vos résultats il est important de préciser le type de lien que vous avez utilisé pour votre clustering hiérarchique. Par ailleurs vous devez travailler avec la distance euclidienne.

- En fixant le nombre de nombre de cluster à 3 dans votre clustering hiérarchique, , faite une nouvelle évaluation pour savoir si les clusters retrouvés correspondent effectivement aux internautes présentant un sentiment négatif (-1), neutre (0) ou positif (1). 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) pour cette évaluation. Tirez une conclusion sur vos résultats.

3 Remise du TP

- Vous devez respecter vos groupes de travail initiaux;
- La date de remise du projet est le 26 novembre 2024 23h59, aucun projet 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), votre code bien expliqué.
- N'oubliez pas d'identifier les membres du groupe de travail. Le fichier soumis doit porter le nom INF1473-P2-EquipeN, où N est les numéro/lettre du groupe. La remise doit être faite par moodle.