Enoncé TP04

Partie 01 : Prétraitement de texte

Dans la première partie, vous devez commencer par télécharger un dataset textuel et puis ajouter une description de votre dataset texte (le dataset doit contenir 150 lignes minimum). Comme exemple, voila un datset du site web kaggle.com :
https://www.kaggle.com/datasets/jensenbaxter/10dataset-text-document-classification

Exemple de dataset (avant transformation CSV ou ARFF) :

No	Texte	Classe
01	Lufthansa flies back to profit…	business
.	Sundance to honour foreign films	entertainment
.	Grilled Shrimp Tacos with Avocado-Corn Salsa	food
.	I need help in creating my 4x4 perspective matrix.	Graphics
.	…	…
150	US economy still growing says Fed	Business

Sur votre du dataset csv ou arff (créer comme le fichier ReutersCorn-train.arff du dossier data de weka), vous allez appliquer la fonction ”StringToWordVector”, qui est une étape de prétraitement permettant de transformer l'attribut texte en un vecteur de mots qui représente le texte sous forme de:
1- Modèle booléen
2- Modèle tf
3- Modèle tf-idf
4- etc.
https://www.youtube.com/watch?v=IY29uC4uem8
https://www.youtube.com/watch?v=jSZ9jQy1sfE
https://www.youtube.com/watch?v=4TPJ7WI-XqA
exemple de sortie :

No	Je	Suis	Un	Etudiant	I	M	A	Student	Das	Auto	Mi	Amo	Espagne	Classe
01	1	1	1	1	0	0	0	0	0	0	0	0	0	Fr
.	0	0	0	0	1	1	1	1	0	0	0	0	0	EN
	0	0	0	0	0	0	0	0	1	1	0	0	0	Ger
150	0	0	0	0	0	0	0	0	0	0	1	1	1	Sp

Table: exemple de représentation dans le modèle booléen

Les modèles sont vus dans le document du cours et précisés dans l’énoncé. Vous devriez prouvez que vous avez bien lu le cours et cela en ajoutant une explication de chaque méthode sous forme de résumés rédiger par vous-même non pas copier (très bonne idée pour prouver que vous avez compris)

Après transformation, vous allez essayer d’appliquer les 5 algorithmes de classification (knn, id3, c4.5, 1rule et nb) avec les 3 méthodes d’évaluation (percentage split, cross validation et leave one out) sur chacune des méthodes de prétraitement pour en trouver qu’elle est la meilleure.

Une conclusion expliquera comme toujours la meilleure méthode d’évaluation et le meilleur algorithme pour chaque méthode de prétraitement. Et une conclusion générale qui résume le tout.

La partie pratique comprend l'implémentation d'une interface de choix de prétraitement de texte (Booleen, TF, TF-IDF, …etc.) suivi par l'application d'un algorithme de data Mining sur

votre dataset textuel. Il doit nous ramener le code source commenté au maximum pour consultation le jeudi 05 Mai 2022

Partie 02 : Méthodes Ensemblistes

Le dernier TP portera sur le dernier chapitre : « les méthodes ensemblistes ».
Il s'agit donc d'expérimenter l'approche combinaison de modèles.
Il s'agit de vérifier que l'étudiant a compris leurs fonctionnements et leurs paramètres.
Les méthodes concernées sont :

1. Bagging

https://www.youtube.com/watch?v=4B8V5IMT6GU

2. Boosting

https://www.youtube.com/watch?v=CqEkVnpOJ1A

3. Voting

https://youtu.be/WJZN4eatdeM

4. Stacking

https://www.youtube.com/watch?v=062w-dGDRr0

5. Random forest

https://www.youtube.com/watch?v=_LbnNTz4-mI

La vidéo suivante permet de mieux comprendre les étapes à suivre :
https://youtu.be/WJZN4eatdeM
https://www.youtube.com/watch?v=Zw4w8l8f5_4

Sur un dataset benchmark (reconnu) ramené d'internet, de votre choix, vous allez expérimenter des méthodes ensemblistes.
Des comparaisons de ces méthodes au meilleur algorithme simple (C4.5 par exemple) pour le Dataset en question seront effectuées.
N'oubliez pas de faire varier les paramètres.
L'étudiant doit aussi les comparer entre eux.
Exemple : Le Bagging de C4.5 ou le Boosting de C4.5 est plus intéressant ?
Bien sûr la réponse changera selon le Dataset.
L'étudiant peut comparer plusieurs combinaisons et expliquer ou justifier les résultats obtenus.

Puisque c'est le TP final, les étudiants peuvent aussi intégrer les notions acquises du TP1 au TP3 (discrétisation, traitement des valeurs manquantes, méthodes d'évaluation, sélection d'attributs, nettoyage de données, transformation, etc.), dans l'objectif d'améliorer leurs notes finales de TP.

Pour la partie implémentation, l'étudiant doit implémenter lui-même une méthode ensembliste ou faire appel à deux méthodes ensemblistes déjà implémentés dans WEKA. Il doit nous ramener le code source commenté au maximum pour consultation le jeudi 05 Mai 2022.

Délai de remise du TP4 (partie 1 et 2) est le Jeudi 05 Mai 2022 à 23h59. L'essentiel est de faire le meilleur TP possible.