UNIVERSITE DE MAROUA
FACULTE DES SCIENCES
Département : Mathématiques-Informatique
M2 – Data Sciences
Année académique : 2023-2024

Distribution : Lundi le 06 Mai 2024. Par : M. Touza Isaac

WEB MINING (DSI 553)
FICHE DE TRAVAIL PERSONNEL ENCADRE - TPE

Consignes :

- Le travail se fera par groupe de deux étudiants
- Le travail est à remettre au tard jeudi le 20 mai 2024 à 12h00. Délai de rigueur.
- Les fichiers numériques (script python, fichier csv, rapport TPE en PDF) doivent être compressé dans un fichier et nommé TPE_WebMining_Groupe_NumeroGroupe (Exemple : TPE_WebMining_Groupe_1)
- Le fichier ci-dessus doit être envoyé par mail à l'adresse isaac_touza@outlook.fr au plus tard à la date indiquée plus haut.
- Le code source (fichier .ipynb) doit être commenté

TRAVAIL A FAIRE :

Groupe 1 :

Objectif : Votre tâche est de construire un modèle de classification de texte pour prédire le sentiment des commentaires non étiquetés sur les articles de site de e-commerce.

Données d'apprentissage :

Vous disposez d'un ensemble de données contenant des commentaires d'utilisateurs sur un site de commerce électronique présent à l'adresse suivante :

https://docs.google.com/spreadsheets/d/1QShriueBxfPp2Jjyhk7rvBAehVdJNelkH2jWaBS2guQ/edit?usp=sharing

Chaque commentaire est étiqueté comme positif (+) ou négatif (-) en fonction du sentiment exprimé.

Tâches à réaliser :

Il vous est demandé d'utiliser l'ensemble de données d'entraînement pour construire votre modèle en répondant aux questions suivantes :

1. Télécharger les données d'apprentissage ci-dessus.
2. Réaliser le prétraitement de ces données
3. Subdiviser l'ensembles des données en données d'entrainement (80%) et en données de test (20%)
4. Construire un modèle de classification en utilisant l'algorithme de votre choix (Naïve Bayes ou Knn).
5. Générer la matrice de confusion du modèle de classification construit
6. Imprimer les rapports de la classification puis évaluer la performance du modèle utilisé.

7. Utiliser votre modèle entraîné pour prédire le sentiment des commentaires non étiquetés suivants :
   - "Ce produit est incroyable !"
   - "Je suis déçu de mon achat."
   - "Le service client est exceptionnel."
   - " Vous êtes le meilleur

Groupe 2 :

Objectif : le but de ce travail d'étudier le classement de l'équipe national du Cameroun au championnat africain et mondial.

Données d'apprentissage : Les classements de l'équipe du Cameroun aux championnats passés se trouvant à l'adresse https://fr.wikipedia.org/wiki/%C3%89quipe_du_Cameroun_de_football

Tâches à réaliser :

1. Démarrer le logiciel JupyterLab puis importer les modules suivants :
   - urllib
   - bs4
   - pandas
   - request
   - os
   - requests
2. Accédez au site mentionné plus haut puis scraper les informations sur le classement FIFA de l'équipe du Cameroun puis les stocker dans une base des données et l'exporter sous format Excel.
3. Construire deux modèles de classification l'un bayésien et l'autre basé sur l'algorithme KNN en utilisant les données de la question 2.
4. Générer la matrice de confusion de chacun de vos classifieur puis l'exportez sous format image.
5. Imprimer le rapport des classifications puis évaluer les performances des modèles utilisés.
6. Interroger vos deux modèles de classification pour prédire l'année où le Cameroun occupera les classements ci-dessous :

Classement mondiale	Classement africain	Année	
		KNN	Naïf bayésien
1	1		
10	1		
3	20		
3	6		

7. Lequel de ces deux modèles est meilleur ? justifiez

Groupe 3 :

Objectif : En utilisant les données fournies sur les patients, les symptômes observés et les maladies, l'objectif de ce travail consiste à créer un modèle de classification capable de prédire la maladie en fonction des symptômes observés.

Données d'apprentissage : Les données relatives aux maladies que vous utiliseriez dans ce travail se trouvent à l'adresse : https://drive.google.com/file/d/1Ji5UGyZT0_SocmsDA7sqL_H9W8IN-I-q/view?usp=drive_link

Tâches à réaliser :

1. Importer les données mises à votre disposition dans votre programme.
2. Divisez les données en un ensemble d'entraînement et un ensemble de test. Par exemple, vous pouvez utiliser 80% des données pour l'entraînement et 20% pour les tests.
3. Préparez les données en les transformant en un format adapté à l'apprentissage automatique. Vous pouvez utiliser la représentation en vecteurs, où chaque symptôme observé correspond à une caractéristique binaire (0 pour absent, 1 pour présent). Par exemple, si vous avez 5 symptômes possibles, vous pouvez représenter chaque patient par un vecteur de longueur 5.
4. Créer un modèle de classification en utilisant l'algorithme KNN
5. Entraînez le modèle en utilisant l'ensemble d'entraînement. Ajustez les paramètres de l'algorithme si nécessaire.
6. Évaluez les performances du modèle en utilisant l'ensemble de test. Calculez des mesures telles que la précision, le rappel et la F-mesure pour évaluer la qualité des prédictions.
7. Utiliser votre modèle pour prédire la maladie sur de nouvelles données non étiquetées suivants :

Symptômes	**Maladie prédite**
Maux de tête, frissons, maux de ventre	
Fièvre, vertige	
Douleur abdominale, fatigue, vomissements	

Groupe 4 :

Objectif : À partir d'un ensemble de courriels, l'objectif est de développer un modèle de classification capable de prédire si un courriel est un spam ou non-spam (ham) en fonction de son contenu.

Données d'apprentissage : Vous disposez d'un jeu de données contenant des courriels ainsi que leur étiquette correspondante (spam ou non-spam) se trouvent à l'adresse :

https://drive.google.com/file/d/1kQgjY2TH-WFJUxYpiN56LBEZwkzNMTt4/view?usp=drive_link

Tâches à réaliser :

1. Importez les données de courriels et leurs étiquettes (spam ou non-spam) dans votre programme.
2. Divisez les données en un ensemble d'entraînement et un ensemble de test (par exemple, 80% pour l'entraînement et 20% pour les tests).
3. Prétraitez les données en les nettoyant et en les transformant en une représentation numérique adaptée à l'apprentissage automatique
4. Créez un modèle de classification de texte en utilisant un algorithme tel que Naive Bayes, SVM ou Réseaux de Neurones.
5. Entraînez le modèle en utilisant l'ensemble d'entraînement et ajustez les hyperparamètres si nécessaire.
6. Évaluez les performances du modèle en utilisant l'ensemble de test. Calculez des mesures telles que la précision, le rappel et le score F1 pour évaluer la qualité des prédictions.
7. Utilisez votre modèle pour prédire si de nouveaux courriels suivants sont des spams ou non-spams.
   - Recevez nos vœux les plus chaleureux pour votre anniversaire de mariage.
   - Inscrivez-vous dès maintenant pour recevoir nos offres par SMS.
   - Ne manquez pas notre promotion de rentrée.
   - Vous avez gagné de l'argent.