UNIVERSITE DE MAROUA
Faculté de sciences
Master / Niveau II

DEPARTEMENT DE MATHS-INFO
Parcours: Data Science
Année : 2023 – 2024

WEB MINING (DSI 553)
FICHE DE TRAVAUX DIRIGES

Par : Pr Kaladzavi/ M. Touza Isaac

Partie 1 : Web Content Mining

Exercice 1 : Questions du cours

1. Définir les notions suivantes : data mining, Web Mining, classification, clustering.
2. Enumérer quelques tâches du data mining
3. Décrire deux objectifs du Web Mining
4. Décrire le processus du Web Mining
5. Enumérer un exemple d'algorithme permettant de réaliser chacune des tâches suivantes : La classification, La prévision, la Recherche des règles d'association et la segmentation
6. Donner la différence entre la classification supervisée et la classification non supervisée.
7. Présenter la différence entre les méthodes descriptives et les méthodes prédictives.
8. Décrire le principe de fonctionnement de l'algorithme KNN
9. Expliquer comment doit se faire le choix de la valeur de k lors de l'exécution de l'algorithme KNN avec un jeu des données quelconque.
10. Expliquer pourquoi la valeur de k dans l'algorithme KNN doit être impair.
11. Présenter deux inconvénients de l'algorithme KNN.
12. Quels sont les différentes sources de données collectées ?
13. Donner quelques opérations de prétraitement des données textuelles.
14. Quelles bibliothèques Python sont utilisées pour le prétraitement des données ?
15. Qu'est-ce qu'une transaction dans le contexte des règles d'association ?
16. Quels sont les concepts de base des règles d'association ?
17. Quels sont les algorithmes couramment utilisés pour les règles d'association ?
18. Enumérer quelques méthodes de pondération ou de calcul des fréquences ?

Exercice 2 : Vous disposez d'un ensemble de données représentant les symptômes de patients atteints de différentes maladies. Chaque patient est décrit par les symptômes qu'il présente. Utilisez l'algorithme Apriori pour trouver les ensembles de symptômes fréquents et générer des règles d'association basées sur l'ensemble de données donné.

Patient 1 : {Fièvre, Maux de tête, Fatigue}
Patient 2 : {Maux de tête, Toux, Frissons, Fatigue}
Patient 3 : {Fièvre, Toux, Douleurs musculaires}
Patient 4 : {Fièvre, Maux de tête, Toux, Frissons}
Patient 5 : {Toux, Frissons, Fatigue}

1. Donner une représentation binaire de ces données.
2. Utilisez l'algorithme Apriori pour trouver tous les ensembles fréquents avec un support minimum de 2 patients.
3. En utilisant les ensembles fréquents trouvés à l'étape précédente, générez toutes les règles d'association possibles avec une confiance minimale de 60%.

4. Interprétez les règles d'association intéressantes trouvées en termes de symptômes de maladies.

Exercice 3 : Le tableau suivant contient des données sur des individus d'une population décrite selon deux attributs : attribut 1 et attribut 2. La classe d'un individu peut être : C1, ou C2, ... ou C6.

Tableau des données

N°	Attribut 1	Attribut 2	Classe
1	1	2	C1
2	2	6	C1
3	2	5	C2
4	2	1	C3
5	4	2	C5
6	5	6	C4
7	6	5	C3
8	6	1	C6

1. Représentez sur le plan les données du tableau précédent (On prendre l'attribut 1 en abscisse et l'attribut 2 en ordonnée).
2. On veut classer un nouvel individu U ayant comme attributs (1, 4) en utilisant la méthode KNN. Quelle sera la classe de U si on choisit k=3. Justifiez. (Utiliser la distance euclidienne)
3. On utilise maintenant la variante de KNN qui utilise la distance $1/d^2$ (inverse de la distance au carré) pour calculer les voisins. Quelle sera la classe de U avec k=3 ?. Justifiez.

Exercice 4 : Soit l'ensemble D des entiers suivants :

D= { 2, 5, 8, 10, 11, 18, 20 }

On veut répartir les données de **D** en trois (3) clusters, en utilisant l'algorithme K-means. La distance ***d*** entre deux nombres a et b est calculée ainsi :

d(a , b) = |a - b| (la valeur absolue de a moins b)

1. Appliquez l'algorithme Kmeans en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11. Montrez toutes les étapes de votre calcul.
2. Donnez le résultat final et précisez le nombre d'itérations qui ont été nécessaires.
3. Peut-on avoir un nombre d'itérations inférieur pour ce problème ? Discutez.

Exercice 5 : Le tableau suivant contient des données sur les résultats obtenus par des étudiants de Tronc Commun (première année à l'Université). Chaque étudiant est décrit par 3 attributs : Est-il doublant ou non, la série du Baccalauréat obtenu et la mention. Les étudiants sont répartis en deux classes : Admis et Non Admis. On veut construire un arbre de décision à partir des données du tableau, pour rendre compte des éléments qui influent sur les résultats des étudiants en Tronc Commun. Les lignes de 1 à 12 sont utilisées comme données d'apprentissage. Les lignes restantes (de 13 à 16) sont utilisées comme données de tests.

	Doublant	**Série**	**Mention**	**Classe**
1	Non	Maths	ABien	Admis
2	Non	Techniques	ABien	Admis
3	Oui	Sciences	ABien	Non Admis
4	Oui	Sciences	Bien	Admis
5	Non	Maths	Bien	Admis
6	Non	Techniques	Bien	Admis
7	Oui	Sciences	Passable	Non Admis
8	Oui	Maths	Passable	Non Admis
9	Oui	Techniques	Passable	Non Admis
10	Oui	Maths	TBien	Admis
11	Oui	Techniques	TBien	Admis

12	Non	Sciences	TBien	Admis
13	Oui	Maths	Bien	Admis
14	Non	Sciences	ABien	Non Admis
15	Non	Maths	TBien	Admis
16	Non	Maths	Passable	Non Admis

1. Définir arbre de décision
2. On souhaite utiliser l'algorithme ID3 pour construire l'arbre de décision en utilisant les données d'apprentissage de ce tableau.
   2.1. Montrez toutes les étapes de calcul pour la construction de deux arbres de décisions différents puis Dessinez-les.
   2.2. Quels sont les résultats de test de l'arbre obtenu sur les données des lignes de 13 à 16 ? Ces résultats correspondent-ils à ceux du tableau ci-dessus ?
   2.3. Déduire le taux d'erreur et de succès de chacun d'arbre de décision construit.
   2.4. Déduire l'arbre à choisir puis justifiez.
3. Présenter deux limites de l'algorithme ID3.
4. Pour améliorer l'algorithme ID3, l'on a mis sur pied l'algorithme C4.5.
   4.1. Présenter deux extensions de l'algorithme ID3 introduit par C4.5
   4.2. Donner l'intérêt d'élaguer un arbre de décision.

Exercice 6 : On veut apprendre un modèle permettant de déterminer si un client est intéressé à acheter un certain produit (Oui ou Non), en fonction de son sexe (Homme ou Femme), son âge (< 18,$18 - 35$ ou > 35), son état civil (Célibataire ou Marie), et son revenu (Faible, Moyen ou Elevé).

Soit l'échantillon suivant d'exemples d'entraınement :

ID	Sexe	Âge	État civil	Revenu	Achat
1	Homme	$18 - 35$	Marié	Moyen	Non
2	Homme	< 18	Célibataire	Faible	Non
3	Homme	> 35	Marié	Élevé	Oui
4	Femme	< 18	Célibataire	Moyen	Non
5	Homme	$18 - 35$	Célibataire	Moyen	Non
6	Femme	$18 - 35$	Célibataire	Élevé	Oui
7	Femme	$18 - 35$	Marié	Faible	Non
8	Homme	$18 - 35$	Marié	Élevé	Oui
9	Homme	> 35	Célibataire	Faible	Oui
10	Femme	< 18	Célibataire	Moyen	Non
11	Femme	> 35	Célibataire	Moyen	Oui
12	Femme	> 35	Marié	Élevé	Oui
13	Homme	$18 - 35$	Célibataire	Faible	Non
14	Femme	$18 - 35$	Marié	Moyen	Oui

Construisez l'arbre de décision résultant de ces exemples en utilisant l'algorithme ID3. On suppose qu'on arrête la subdivision uniquement lorsque les nœuds sont purs (entropie de 0).

Exercice 7: Considérons l'ensemble des données d'entrainement ci-dessous contenant des documents (d) auquel on a associé leur classes (A ou B).

d	***c***	***d***	***c***
aa	*A*	*ba*	*A*
ab	*A*	*bb*	*B*

1. Calculer P(A), P(B), P(a|A), P(b|A), P(a|B), P(b|B). Utiliser la technique de Laplace smoothing pour le calcul de probabilités.
2. Classer chacune des nouvelles données ci-dessous, supprimer tous les mots inconnus.

Documents:
aaba
a
bbba
bccbba
bbbb

Exercice 8 : Vous avez un ensemble de documents texte que vous souhaitez prétraiter avant de les utiliser dans une tâche de classification de texte. Appliquez les étapes de prétraitement des données textuelles suivantes à un document donné :

Document 1: "Le chat noir se promène dans le jardin."
Document 2 : " Le chat saute sur les arbres et attrape des oiseaux"
Document 3 : " Le chien aboie et le chat miaule."

1. Donner la représentation à 4 grammes de l'ensemble du document 1.
2. Diviser le document en tokens individuels.
3. Supprimer les mots vides (stop words) français
4. Normaliser les tokens obtenus.
5. Appliquer l'algorithme de stemming de Porter pour réduire les mots à leur racine.
6. Calculez la fréquence de chaque terme dans l'ensemble du corpus constitué des documents 1, 2 et 3.

Partie II : Le Web structure Mining

Exercice 9 : Questions du cours

1. Définir les termes et expressions suivantes :
    - Web-graph
    - Directed Path
    - Lien
    - In-degree
    - Out-degree
    - Diamètre
    - Tubes
2. Enumérer 04 structures du web
3. A l'aide d'un schéma, représenter la forme en nœud papillon du Web (préciser ses différente parties).
4. Enumérer deux algorithmes d'analyse de la structure du Web

Exercice 10 : Soit les adresses urls suivants :

Adresse 1 : http://www.univ-maroua.cm/
Adresse 2 : /fr/service/centre-medico-social.html

1. Décomposer chacun de ces liens en donnant ses différentes parties
2. Identifier le type de chacun de ces liens

3. Déduire la différence entre le lien interne et le lien externe.

Exercice 11 : Considérons un site Web constitué de 04 pages A, B, C et D. Il existe un lien entre la page A vers B et C, un lien de la page D vers A et C. Il existe également un lien entre les pages B et C vers A.

1. Représenter le web-graph de ce site, en considérant les pages Web comme des nœuds.
2. Calculer PageRank(A), PageRank(B), PageRank(C) et PageRank(D).

NB : Commencez avec la valeur initiale de PageRank à 0 et compléter jusqu'à trois itérations.

Exercice 11 : Considérons un site web contenant 5 pages, représentées par les identifiants de page suivants : P1, P2, P3, P4 et P5. Les liens entre ces pages sont les suivants :

- P1 contient des liens vers P2 et P3.
- P2 contient un lien vers P4.
- P3 contient un lien vers P4.
- P4 contient des liens vers P2 et P5.
- P5 contient un lien vers P3.

1. Dessinez un graphe où chaque page est représentée par un nœud, et les liens entre les pages sont représentés par des arêtes. Utilisez les informations fournies pour connecter les nœuds correspondant aux pages concernées.
2. Construisez la matrice d'adjacence A en utilisant la représentation du graphe. La matrice d'adjacence est une matrice carrée où chaque élément A[i, j] indique s'il existe un lien de la page i à la page j (1 pour un lien existant, 0 sinon).
3. Utilisez une méthode de clustering, pour regrouper les pages en fonction de leurs liens. Choisissez le nombre de clusters à 2.

Partie III : Le Web Usage Mining

Exercice 13 : Questions du cours

1. Définir : Web Usage Mining , ressource web, requête , épisode, serveur web , page web, vue d'une page Web, visite, fichier log
2. Donner les objectifs du Web Usage Mining
3. Énumérer les types des fichiers logs
4. Citer trois endroits ou peuvent etre stockés un fichier logs
5. Énumérer la composition d'un fichier log selon le format CLF
6. Citer deux autres formats des fichiers logs
7. Quelles sont les valeurs de types de requêtes dans un fichier log
8. Donner les valeurs du champ « statut » dans un fichier log puis donner leur signification.
9. Décomposer la ligne du fichier log donné ci-dessous **: "129.0.210.220 - - [24/Jul/2022 :09 :05 :00 +0200] "GET /img/portfolio/6.jpg HTTP/2.0" 301 333 "https ://profsinfocmr.org/" "Mozilla/5.0 (Linux ; Android 8.1.0 ; TECNO CF7k) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/103.0.0.0 Mobile Safari/537.36"**
10. Enumérer les opérations à réaliser pour un prétraitement des fichiers logs
11. Présenter les trois méthodes pour identifier les épisodes

12. Identifier toutes les épisodes se trouvant dans la phrase suivante : **« Pendant une session sur un site web, un utilisateur a vérifié son adresse e-mail, a téléchargé deux images, a changé sa photo de profil et s'est déconnecté du site** ».

Exercice 14 : Considérons l'utilisateur **u** qui a généré la session serveur suivante : v = {u, 16 : 09 : 10, < (A, 16 : 09 : 10), (B, 16 : 09 : 43), (C, 16 : 12 : 02), (A, 18 : 32 : 02), (C, 18 : 33 : 05), (E, 18 : 47 : 12), (C, 18 : 48 : 20), (H, 19 : 15 : 49), (C, 19 : 51 : 32) >}

1. Définir les notions suivantes :
    - Utilisateur
    - Session
2. En considérant le seuil temporel standard $\Delta t = 30$ minutes, déterminer toutes les visites de l'utilisateur u.

Exercice 15 : En se servant de l'extrait du fichier log ci-dessous, identifier les différentes sessions en utilisant la règle h2, prendre le seuil $\beta = 10$.

Time	IP	URL	REF
0 :01	1.2.3.4	A	-
0 :09	1.2.3.4	B	A
0 :19	1.2.3.4	C	A
0 :25	1.2.3.4	E	C
1 :15	1.2.3.4	A	-
1 :26	1.2.3.4	F	C
1 :30	1.2.3.4	B	A
1 :36	1.2.3.4	D	B

Exercice 16 : Vous travaillez pour une entreprise de commerce électronique qui souhaite améliorer l'expérience utilisateur sur son site web. Votre tâche consiste à analyser les données des utilisateurs afin de comprendre le comportement des utilisateurs et d'identifier des opportunités d'optimisation du site. On met donc à votre disposition un ensemble de données contenant des journaux de navigation des utilisateurs. Chaque entrée du journal contient des informations telles que l'adresse IP de l'utilisateur, l'URL visitée, la date et l'heure de la visite, ainsi que d'autres informations pertinentes.

IP	**URL**	**Date et heure**
192.168.0.1	/home	2023-06-01 08:15
192.168.0.2	/products	2023-06-01 09:32
192.168.0.3	/home	2023-06-01 10:05
192.168.0.4	/products	2023-06-01 10:15
192.168.0.1	/cart	2023-06-01 11:25
192.168.0.5	/home	2023-06-01 12:10
192.168.0.2	/products	2023-06-01 12:40
192.168.0.3	/checkout	2023-06-01 13:20
192.168.0.6	/home	2023-06-01 14:05
192.168.0.4	/products	2023-06-01 14:30
192.168.0.5	/products	2023-06-01 15:15
192.168.0.1	/home	2023-06-01 16:00

192.168.0.3	/products	2023-06-01 16:45
192.168.0.6	/cart	2023-06-01 17:20
192.168.0.2	/products	2023-06-01 17:50
192.168.0.4	/home	2023-06-01 18:35
192.168.0.5	/products	2023-06-01 19:10
192.168.0.3	/checkout	2023-06-01 19:40
192.168.0.6	/home	2023-06-01 20:25
192.168.0.1	/products	2023-06-01 20:50

1. En se servant de l'adresse IP, identifiez les utilisateurs de ce site en précisant les ressources consultées et la date de consultation
2. En considérant le seuil temporel Δt = 20 minutes, déterminer pour chaque utilisateur ses visites.
3. Recopier et compléter le tableau ci-dessous

Id	**Utilisateur**	**Pages consultées**
1	192.168.0.1	
2	192.168.0.2	
3	192.168.0.3	
4	192.168.0.4	
5	192.168.0.5	
6	192.168.0.6	

4. Donner la représentation horizontale du tableau de la question 3
5. En utilisant l'algorithme Apriori :
   5.1. Déterminer les pages fréquents, sachant que le support minimal est de 0,7
   5.2. Déduire les règles d'association entre ces pages fréquents avec une confiance minimale de 80%.
   5.3. Donner une interprétation d'une règle de votre choix.