Data-Mining

I- Introduction au Data-Mining
II- Text mining
III- Web mining
IV- KNIME

Dr Ghazi Aziz

I.Introduction au

Data-Mining

Data-Mining : introduction

Data-mining ≡ Fouille de données

Regroupe un ensemble de techniques et d'outils de la Statistique, l'Informatique et la Science de l'information

A évolué vers la science des donnéesMachine Learning, Data-Mining Big Data (explosion des données) Formalismes de stockage et de traitement distribués des données(NoSQL, Hadoop, MapReduce, Spark ...)

2/30

Historique des techniques de statistique et de data mining

1875 Régression linéaire de Francis Galton.
1896 Formule du coefficient de corrélation de Karl Pearson2.
1900 Distribution du X^2 de Karl Pearson.
1936 Analyse discriminante de Fischer et Mahalanobis
1941 Analyse factorielle des correspondances de Guttman
1943 Réseaux de neurones de Mac Culloch et Pitts
1944 Régression logistique de Joseph Berkson
1958 Perceptron de Rosenblatt
1962 Analyse des correspondances de J.-P. Benzécri
1962 Régression logistique de J. Cornfield
1964 Arbre de décision AID de J.-P. Sonquist et J.-A. Morgan
1965 Méthode des centres mobiles de E. W. Forgy
1967 Méthode des k means (k moyennes) de Mac Queen
1971 Méthode des nuées dynamiques de Diday
1972 Modèle linéaire généralisé de Nelder et Wedderburn
1975 Algorithme génétique de Holland
1977 Méthode de classement DISQUAL de Gilbert Saporta
1980 Arbre de décision CHAID de KASS
1983 Régression PLS de Herman et Svante Wold
1984 Arbre CART de Breichman, Friedman, Olshen, Stone
1986 Perceptron multicouches de Rumelhart et Mac Clelland
1989 Réseaux de T. Kohonen (cartes auto-adaptatives)
1990 Apparition du concept de Data Mining
1993 Arbre C4.5 de J. Ross Quinlan
1996 Bagging (Breiman) et boosting (Freund-Shapire)
1998 Support vector machine de Vladimir Vapnik
2001 Régression logistique PLS de Tenenhaus

QUESTIONS : Exemples

- Combien de clients ont acheté tel produit pendant telle période ? ⟹ SQL
- A-t-on vendu plus d'un tel produit cette année que l'année dernière? ⟹ OLAP
- Quel est le profil des clients ?
- Quels autres produits les intéresseront ?
- Quand seront-ils intéressés ?

⟹ Data Mining

Data-Mining : définition

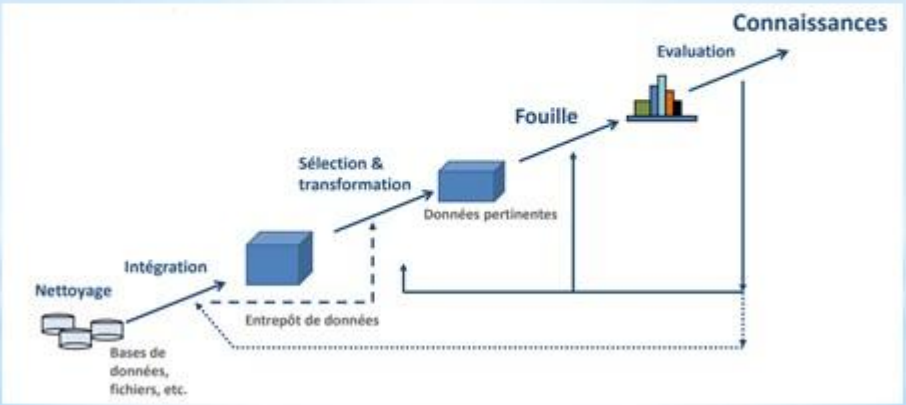
Définition 1

Le data-mining est un processus de découverte de règle, relations, corrélations et/ou dépendances à travers une grande quantité de données, grâce à des méthodes statistiques, mathématiques et de reconnaissances de formes.

Définition 2

Le data-mining est un processus d'extractions automatique d'informations predictives à partir de grandes bases de données.

Définition : Processus ou méthode qui extrait des connaissances « intéressantes » ou des motifs (patterns) à partir d'une grande quantité de données



Data-Mining : les raisons du développement

Données

Big Data : augmentation sans cesse de donnéesgénérées

- Twitter : 50M de tweets /jour (=7 téraoctets)
- Facebook : 10 téraoctets /jour
- Youtube : 50h de vidéos uploadées /minute
- 2.9 million de mail /seconde

Puissance de calcul

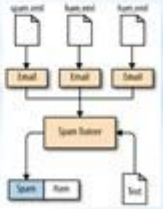
- Loi de Moore
- Calcul massivement distribué

Création de valeur ajoutée

- Intérêt : du produit aux clients.
- Extraction de connaissances des big data

Exemples d'applications

- Entreprise et Relation Clients : création de profils clients, ciblage de clients potentiels et nouveaux marchés
- Finances : minimisation de risques financiers
- Bioinformatique : analyse du génome, mise au point de médicaments
- Internet : spam, e-commerce, détection d'intrusion, recherche d'informations ...
- Sécurité
- ...



Exemples d'applications : E-commerce

Targeting (ciblage)

Stocker les séquences de clicks des visiteurs, analyser les caractéristiques des acheteurs

Faire du "targeting" lors de la visite d'un client potentiel

Systèmes de recommandation

Opportunité : les clients notent les produits ! Comment tirer profit de ces donnéespour proposer des produits à un autre client ?

Solutions : technique dit de filtrage collaboratif pour regrouper les clients ayant les mêmes "goûts".

Exemples d'applications : Analyse des risques

Détection de fraudes pour les assurances

- Analyse des déclarations des assurés par un expert afin d'identifier les cas de fraudes.
- Applications de méthodes statistiques pour identifier les déclarations fortement corrélées à la fraude.

Prêt Bancaire

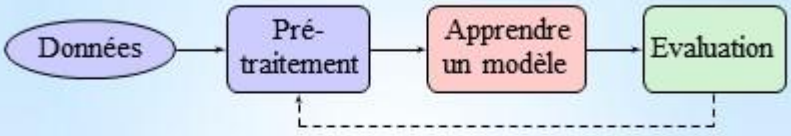
- Objectif des banques : réduire le risque des prêts bancaires.
- Créer un modèle à partir de caractérisques des clients pour discriminer les clients à risque des autres.

Exemples d'applications : Commerce

Opinion mining

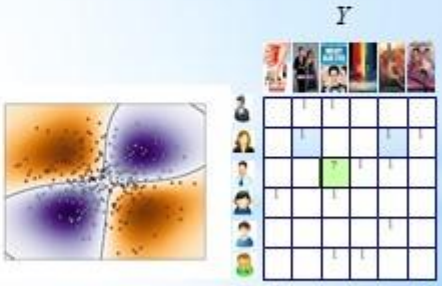
Exemple : analyser l'opinion des usagers sur les produits d'une entreprise à travers les commentaires sur les réseaux sociaux et les blogs

Mise en oeuvre d'un projet de DM



Principales étapes

- Collecte de données
- Pré-traitement
- Analyse statistique
- Identifier le problème de DM
- Apprendre le modèle mathématique
- Évaluer ses capacités



Ensemble de données

Donnéesd'un problème de DM

- Les informations sont des exemples avec des attributs
- On dispose généralement d'un ensemble de N données

Attributs

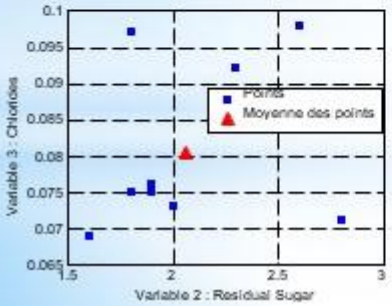
- Un attribut est un descripteur d'une entité. On l'appelle également variable, ou caractéristique

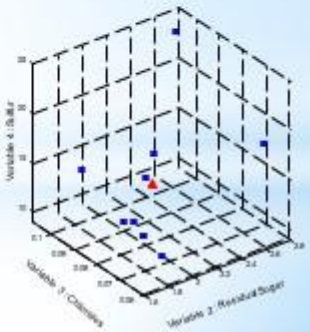
Exemple

- C'est une entité caractérisant un objet ; il est constitué d'attributs.
- Synonymes : point, vecteur (souvent dans R^d)

Données: illustration

Points x \ Variables	citric acid	residual sugar	chlorides	sulfur dioxide
1	0	1.9	0.076	11
2	0	2.6	0.098	25
3	0.04	2.3	0.092	15
Point x $\in R^4$	0.56	1.9	0.075	17
5	0	1.9	0.076	11
6	0	1.8	0.075	13
7	0.06	1.6	0.069	15
8	0.02	2	0.073	9
9	0.36	2.8	0.071	17
10	0.08	1.8	0.097	15





Type de données

Capteurs → variables quantitatives, qualitatives, ordinales
Texte → Chaîne de caractères
Parole → Séries temporelles
Images → données2D
Videos → données2D + temps
Réseaux → Graphes
Flux → Logs, coupons. . .
Etiquettes → information d'évaluation

Big Data (volume, vélocité, variété)

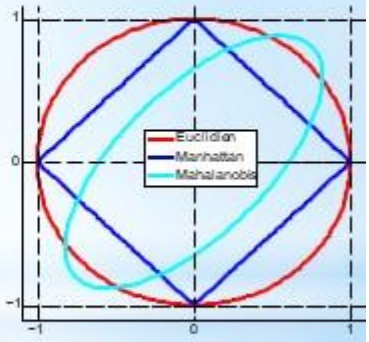
Flot "continu" de données
→ Pre-traitement des données(nettoyage, normalisation, codage . . .)
→ Représentation : des donnéesaux vecteurs

Donnéeset Métriques

Les algorithmes nécessitent une notion de similarité dans l'espace X des données. La similarité est traduite par la notion de distance.

- distance euclidienne : $x, z \in \mathbb{R}^d$, on a
  $d(x,z) = \|x - z\|_2 = \sqrt{\sum_{j=1}^{d}(x_j - z_j)^2} = \sqrt{(x-z)^\top (x-z)}$
- distance de manhattan
  $d(x,z) = \|x - z\|_1 = \sum_{j=1}^{d} |(x_j - z_j)|$
- distance de mahalanobis
  $d(x,z) = \sqrt{(x-z)^\top \Sigma^{-1} (x-z)}$

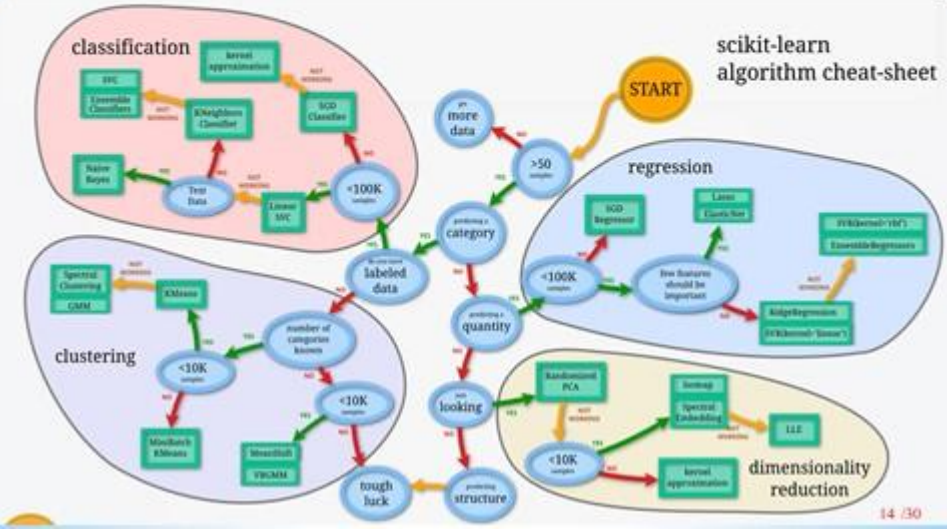
  $\Sigma \in \mathbb{R}^{d \times d}$: matrice carrée définie positive



Caractérisation des méthodes de Data-Mining

Types d'apprentissage

- Apprentissage supervisé
- Apprentissage non-supervisé
- Apprentissage semi-supervisé



Récapitulatif

1. Donnez une définition au DM
2. Donnez qq applications du DM
3. Donnez le schéma de mise en œuvre d'un projet DM
4. Donnez qq type de donnéesutilisés en DM
5. Donnez la formule de calcul de la distance euclidienne ? De Manhattan ? Et Mahalanobis?
6. Donnez l'objectif du ML supervisé et donnez qq algorithmes
7. Donnez l'objectif du ML non supervisé et donnez qq algorithmes
8. Donnez l'objectif du ML semi supervisé et donnez qq algorithmes

Tâches du datamining

1. Classification
2. Estimation
3. Prédiction
4. Segmentation (clustering)
5. Association
6. Description

Tâches du Data Mining : Exemples

Classification : Déterminer grade en fonction de l'âge, l'ancienneté, le salaire et les affectations.

Estimation : (sur Variables continues)

Estimer le salaire en fonction de l'âge, ancienneté et affectations

Prédiction : Prédire quelle sera la prochaine affectation d'un militaire.

Association: Déterminer des règles de type :

Le militaire qui est sergent entre 25 et 30 ans sera lieutenant colonel entre 45 et 50 ans (fiable à n %).

Segmentation (ou clustering) :

Segmenter les militaires en fonction de leurs parcours (carrière) et affectations.

Description : Indicateurs statistiques traditionnels :

Age moyen, pourcentage de femmes, salaire moyen

Les types de modèles

En plus des différents types de référentiels de donnéeset d'informations sur lesquels l'exploitation peut être effectuée, les types de modèles peuvent être également exploités. Il existe un certain nombre de fonctionnalités d'exploration de données:

- La caractérisation et discrimination
- L'exploration de modèles fréquents, d'associations et de corrélations
- Classification et régression
- Analyse de regroupement
- Analyse des valeurs aberrantes

1. Description du concept : Caractérisation et discrimination

Les entrées de donnéespeuvent être associées à des classes ou à des concepts. De telles descriptions d'une classe ou d'un concept sont appelés des descriptions de classe/concept.
Ces descriptions peuvent être dérivées en utilisant :

- La caractérisation des données: en résumant les donnéesde la classe étudiée appelée souvent la classe cible en termes généraux
- La discrimination des données: par comparaison de la classe cible avec une ou un ensemble de classes comparatives appelées souvent classes
- Hybride : à la fois la caractérisation et la discrimination des données.

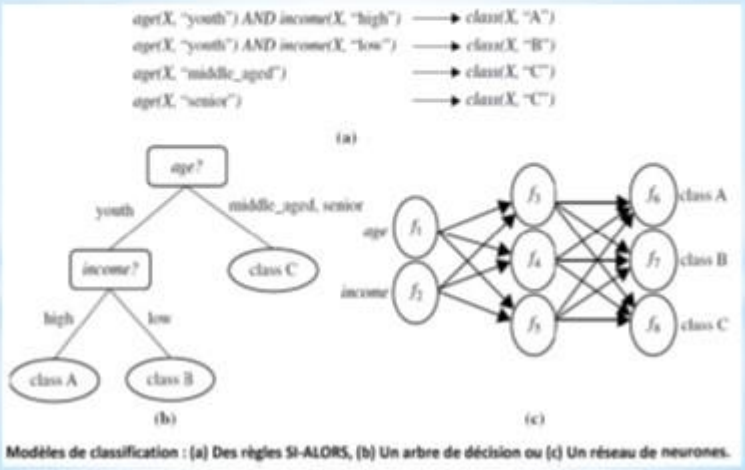
2. Extraction de modèles fréquents, d'associations et de corrélations

Les modèles fréquents sont des modèles qui se produisent fréquemment dans les données. Il existe de nombreux types de modèles fréquents :

- Des ensembles d'éléments fréquents : fait généralement référence à un ensemble d'éléments qui apparaissent souvent ensemble dans un ensemble de donnéestransactionnelles, par exemple, le lait et le pain, qui sont fréquemment achetés ensemble dans les épiceries par de nombreux clients.
- Des sous-séquences fréquentes : également appelées modèles séquentiels telle que le modèle selon lequel les clients ont tendance à acheter d'abord un ordinateur portable, suivi d'un appareil photo numérique, puis d'une carte mémoire, est un modèle séquentiel (fréquent)
- Des sous-structures fréquentes : peut faire référence à différentes formes structurelles (par exemple, des graphiques, des arbres ...) qui peuvent être combinéesavec des ensembles d'éléments ou des sous-séquences. Si une sous-structure se produit fréquemment, on l'appelle un modèle structuré (fréquent).

3. Classification et régression pour l'analyse prédictive

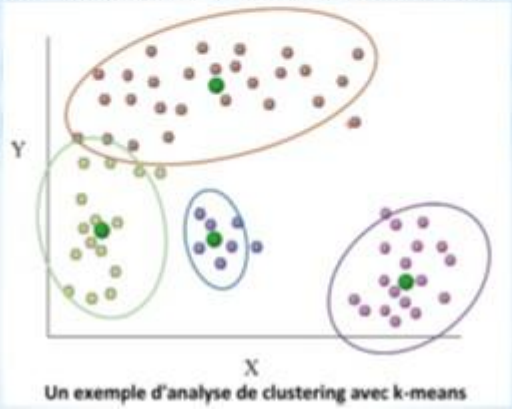
La classification est le processus de recherche d'un modèle (ou d'une fonction) qui décrit et distingue des classes de donnéesou des concepts. Le modèle dérivé peut être représenté sous diverses formes, telles que des règles de classification (c'est-à-dire des règles SI-ALORS), des arbres de décision, des formules mathématiques ou des réseaux neuronaux (Figure). Un test sur une valeur d'attribut de chaque branche représente un résultat du test et les feuilles de l'arbre représentent des classes ou des distributions de classes.



Modèles de classification : (a) Des règles SI-ALORS, (b) Un arbre de décision ou (c) Un réseau de neurones.

4. Analyse de cluster

Contrairement à la classification et à la régression, qui analysent des ensembles de données(d'apprentissage) étiquetés par classe, le clustering analyse les objets de donnéessans consulter les étiquettes de classe. Dans de nombreux cas, les donnéesétiquetées par classe peuvent tout simplement ne pas exister au début. Le clustering peut être utilisé pour générer des étiquettes de classe pour un groupe de données. Les objets sont regroupés selon le principe de maximisation de la similarité intraclasse et de minimisation de la similarité interclasse.



Un exemple d'analyse de clustering avec k-means

5. Analyse des valeurs aberrantes

Un ensemble de donnéespeut contenir des objets qui ne sont pas conformes au comportement général ou au modèle des données. Ces objets de donnéessont des valeurs aberrantes.
De nombreuses méthodes d'exploration de donnéesrejettent les valeurs aberrantes comme du bruit ou des exceptions.
Cependant, dans certaines applications (par exemple, la détection de fraude), les événements rares peuvent être plus intéressants que ceux qui se produisent plus régulièrement.
L'analyse des donnéesaberrantes est appelée analyse des valeurs aberrantes ou extraction d'anomalies.
Les valeurs aberrantes peuvent être détectées à l'aide de tests statistiques qui supposent un modèle de distribution ou de probabilité pour les données, ou à l'aide de mesures de distance où les objets éloignés de tout autre cluster sont considérés comme des valeurs aberrantes

Exercice 1 : K-Means

Soit l'ensemble D des entiers suivants : D= { 2, 5, 8, 10, 11, 18, 20 } On veut répartir les donnéesde D en trois (3) clusters, en utilisant l'algorithme Kmeans. La distance d entre deux nombres a et b est calculée ainsi : d(a , b) = |a - b| (la valeur absolue de a moins b) Travail à faire :
1/ Appliquez Kmeans en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11. Montrez toutes les étapes de calcul. -Premier clusters : C1={ 2, 5, 8} C2={10} C3={11, 18, 20} -
2/ Donnez le résultat final et précisez le nombre d'itérations qui ont été nécessaires.

Correction Exercice 1 :

1)

- Itération 1 :

Mise à jour des clusters : C1={ 2, 5, 8} C2={10} C3={11, 18, 20}
R-estimation des centres de gravité : μ1= (2+5+8)/3 μ2=10/1 μ3=(11+18+20)/3
=> μ1=5 μ2=10 μ3=16.33

- Itération 2 :

Mise à jour des clusters : C1={ 2, 5} C2={8, 10, 11} C3={18, 20}
R-estimation des centres de gravité : μ1= (2+5)/2 μ2=(8+10+11)/3 μ3=(18+20)/2
=>μ1=3.5 μ2=9.66 μ3=19

- Itération 3

Mise à jour des clusters : C1={ 2, 5} C2={8, 10, 11} C3={18, 20}
R- estimation des centres de gravité : μ1= (2+5)/2 μ2=(8+10+11)/3 μ3=(18+20)/2
μ1=3.5 μ2=9.66 μ3=19 Stabilité : Les centres de gravité n'ont pas changé. L'algorithme s'arrête
2) Les clusters résultats : C1={ 2, 5} C2={8, 10, 11} C3={18, 20}
Nombre d'itérations = 3

Exercice 2 : Table de décision

Le tableau suivant contient des donnéessur 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 donnéesdu 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 donnéesd'apprentissage. Les lignes restantes (de 13 à 16) sont utilisées comme donnéesde tests.

1/ Utiliser les donnéesdes lignes de 1 à 12 pour construire l'arbre en utilisant l'algorithme ID3. Montrez toutes les étapes de calcul. Dessinez l'arbre final.

	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

Entropie et gain d'information

- S contient s tuples de classe C pour i = {1, m} S contient s_i tuples de classe Ci pour i = {1..m}
- L'information mesure la quantité d'information nécessaire pour classer un objet classer un objet :

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

- Entropie d'un attribut A ayant pour valeurs $\{a_1, \ldots, a_v\}$:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$$

- Gain d'information en utilisant l'attribut A :

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

Correction :

1) On remarque que sur les 12 lignes des donnéesd'apprentissage, 8 correspondent à la classe "Admis" et 4 à la classe "Non admis". L'entropie de l'ensemble S (à la racine de l'arbre) est donc égale à :

*Entropie (S)= - (8/12) * Log2 (8/12) - (4/12) * Log2 (4/12) = 0.92*

Pour connaitre quel attribut on doit choisir comme test au niveau de la racine de l'arbre, il faut calculer le gain d'entropie sur chacun des attributs : "Doublant", "Série" et "Mention".

Calcul du gain d'entropie sur l'attribut "Doublant" :

*Gain(S, Doublant) = Entropie(S) - 7/12 * Entropie(SOui) - 5/12 * Entropie(SNon)*

*avec Entropie(SOui)= -3/7*Log2(3/7) - 4/7*Log2(4/7)*

*et Entropie(SNon)= -5/5*log2(5/5) Gain(S, Doublant) =0.34*

Calcul du gain d'entropie sur l'attribut "Série" :

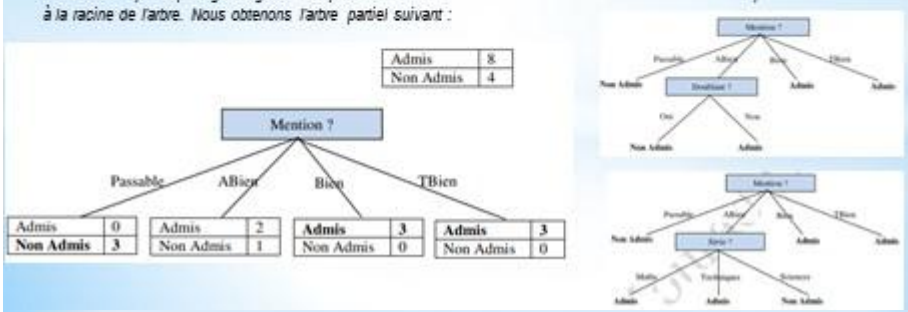
*Gain(S, Série) = Entropie(S) - 4/12 * Entropie(SMaths) - 4/12 * Entropie(STechniques) - 4/12 * Entropie(SSciences) Gain(S, Série) = 0.04*

Calcul du gain d'entropie sur l'attribut "Mention" :

*Gain(S, Série) = Entropie(S) - 3/12 * Entropie(SPassable) - 3/12 * Entropie(SABien) - 3/12 * Entropie(SBien) - 3/12 * Entropie(STBien)*

*Gain(S, Mention) = **0.69***

On constate que le plus grand gain d'entropie est obtenu sur l'attribut "Mention". C'est donc cet attribut qui est choisi comme test à la racine de l'arbre. Nous obtenons l'arbre partiel suivant :



Caractérisation des méthodes : apprentissage supervisé

Objectif

A partir des données $\{(x_i, y_i) \in X \times Y, i = \cdots, N\}$, estimer les dépendances entre X et Y.

On parle d'apprentissage supervisé car les y_i permettent de guider le processus d'estimation.

Exemples

Estimer les liens entre habitudes alimentaires et risque d'infarctus. x_i : attributs concernant le régime d'un patient, y_i sa catégorie (risque, pas risque).

Applications : détection de fraude, diagnostic médical ...

Techniques

k-plus proches voisins, SVM, régression logistique, arbre de décision ...

Caractérisation des méthodes : Apprentissage non-supervisé

Objectifs

Seules les données $\{x_i \in X, i = \cdots, N\}$ sont disponibles. On cherche à décrire comment les donnéessont organisées et en extraire des sous-ensemble homogènes.

Exemples

Catégoriser les clients d'un supermarché. x_i représente un individu (adresse, âge, habitudes de courses ...)

Applications : identification de segments de marchés, catégorisation de documents similaires, segmentation d'images biomédicales ...

Techniques

Classification hiérarchique, Carte de Kohonen, K-means, extractions de règles ...

Caractérisation des méthodes : apprentissage semi-supervisé

Objectifs

Objectifs : parmi les données, seulement un petit nombre ont un label i.e $\{(x_1, y_1), \cdots, (x_n, y_n), x_{n+1}, \cdots, N\}$. L'objectif est le même que pour l'apprentissage supervisé mais on aimerait tirer profit des donnéessans étiquette.

Exemples

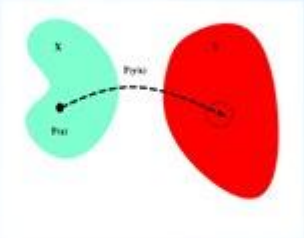
Exemple : pour la discrimination de pages Web, le nombre d'exemples peut être très grand mais leur associer un label (ou étiquette) est coûteux.

Techniques

Méthodes bayésiennes, Self training, Co-Training, SVM ...

Apprentissage supervisé : les concepts

Soit deux ensembles X et Y munis d'une loi de probabilité jointe p(X, Y).



Objectifs : On cherche une fonction $f : X \rightarrow Y$ qui à X associe f(X) qui permet d'estimer la valeur y associée à x. f appartient à un espace H appelé espace d'hypothèses.

Exemple de H : ensemble des fonctions polynomiales

Apprentissage supervisé : les concepts

On introduit une notion de coût L(Y, f(X)) qui permet d'évaluer la pertinence de la prédiction de f, et de pénaliser les erreurs.

L'objectif est donc de choisir la fonction f qui minimise

$$R(f) = E_{X,Y}[L(Y, f(X))] \quad \text{(Esperance)}$$

où R est appelé le risque moyen ou erreur de généralisation. Il est également noté EPE(f) pour expected prediction error

Apprentissage supervisé : les concepts

Exemples de fonction coût et de risque moyen associé.

Coût quadratique (moindres carrés)

$$L(Y, f(X)) = (Y - f(X))^2$$

$$R(f) = E[(Y - f(X))^2] = \int (y - f(x))^2 p(x, y) dxdy$$

Coût (moindres valeurs absolues)

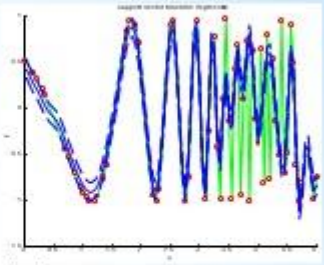
$$L(Y, f(X)) = |Y - f(X)|$$

$$R(f) = E[|Y - f(X)|] = \int |y - f(x)| p(x, y) dxdy$$

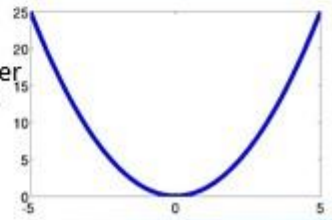
Apprentissage supervisé : les concepts

Régression

On parle de régression quand Y est un sous-espace de $\mathbb{R}^d$.
Fonction de coût typique :
quadratique $(y - f(x))^2$



NB : La régression en apprentissage supervisé est une méthode qui permet de prédire une valeur numérique à partir de donnéesd'entrée. Par exemple, on peut utiliser la régression pour estimer le prix d'une maison en fonction de sa surface, de son nombre de pièces, de son quartier, etc. La régression cherche à trouver une fonction qui minimise l'écart entre les valeurs prédites et les valeurs réelles

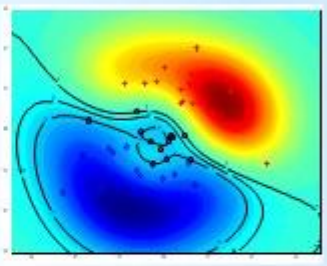


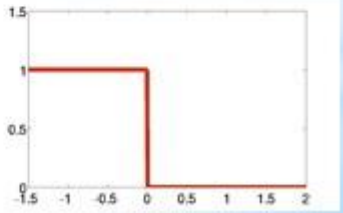
Apprentissage supervisé : les concepts

Discrimination (classification)

si Y est un ensemble discret non-ordonné, (par exemple $\{-1, 1\}$), on parle de discrimination ou classification.

La fonction de coût la plus utilisée est : $\Theta(-yf(x))$ où Θ est la fonction échelon.





Apprentissage supervisé : les concepts

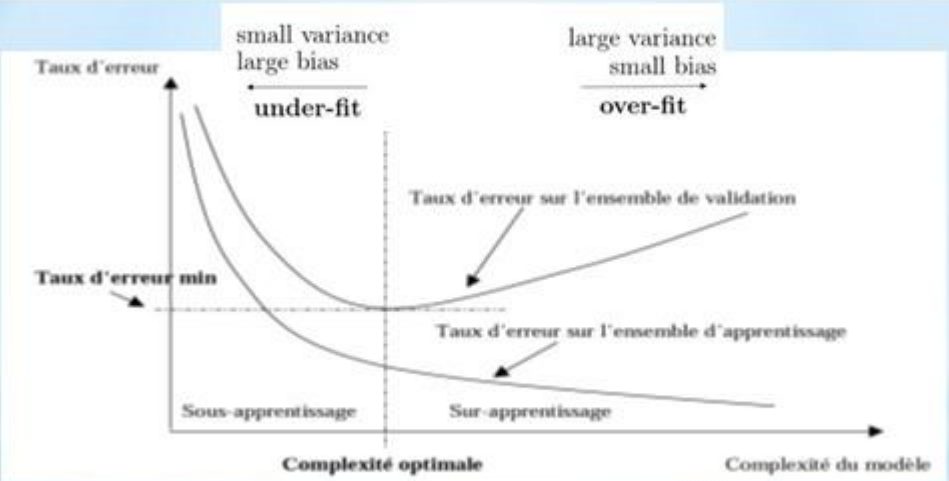
- En pratique, on a un ensemble de données $\{(x_i, y_i) \in X \times Y\}_{i=1}^{N}$ appelé ensemble d'apprentissage obtenu par échantillonnage indépendant de $p(X, Y)$ que l'on ne connaît pas.

On cherche une fonction f, appartenant à H qui minimise le risque empirique :

$$R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$$

- Le risque empirique ne permet pas d'évaluer la pertinence d'un modèle car il est possible de choisir f de sorte que le risque empirique soit nul mais que l'erreur en généralisation soit élevée. On parle alors de sur-apprentissage

Illustration du sur-apprentissage et sous-apprentissage



Il s'agit de trouver un compromis entre la fiabilité du modèle sur l'ensemble d'apprentissage et la généralisation du modèle : trouver le juste milieu entre sous et sur -apprentissage.

Sélection de modèles

Problématique

On cherche une fonction f qui minimise un risque empirique donné. On suppose que f appartient à une classe de fonctions paramétrées par α. Comment choisir α pour que f minimise le risque empirique et généralise bien ?

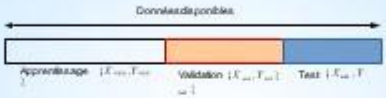
Exemple : On cherche un polynôme de degré α qui minimise un risque $R_{emp}(f_\alpha) = \sum_{i=1}^{N}(y_i - f_\alpha(x_i))^2$.

Objectifs :

- proposer une méthode d'estimation d'un modèle afin de choisir (approximativement) le meilleur modèle appartenant à l'espace hypothèses.
- une fois le modèle choisi, calculer son erreur de généralisation.

Sélection de modèles : approche classique

Cas idéal : les données D_N abondent (N est très grand)



- Découper aléatoirement $D_N = D_{app} \cup D_{val} \cup D_{test}$
- Apprendre chaque modèle possible sur D_{app}
- Evaluer sa performance en généralisation sur D_{val}

$$R_{val} = \frac{1}{N_{val}} \sum_{i \in D_{val}} L(y_i, f(x_i))$$

- Sélectionner le modèle qui donne la meilleure performance sur D_{val}
- Tester le modèle retenu sur D_{test}

Remarque

D_{test} n'est utilisé qu'une seule fois !

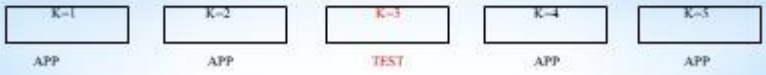
Sélection de modèles : Validation Croisée

Cas moins favorable : les données D_N sont modestes (N est petit)

Estimation de l'erreur de généralisation par rééchantillonnage.

Principe

- Séparer les N données en K ensembles de part égales.
- Pour chaque $k = 1, \cdots, K$, apprendre un modèle en utilisant les $K - 1$ autres ensemble de données et évaluer le modèle sur la k-ième partie.
- Moyenner les K estimations de l'erreur obtenues pour avoir l'erreur de validation croisée.



Sélection de modèles : Validation Croisée (2)

Détails :

$$R_{CV} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k} L(y_i^k, f^{-k}(x_i^k))$$

où f^{-k} est le modèle f appris sur l'ensemble des donnéessauf la k-ième partie.

Propriétés : Si K = N, CV est approximativement un estimateur sans biais de l'erreur en généralisation. L'inconvénient est qu'il faut apprendre N − 1 modèles.

typiquement, on choisit K = 5 ou K = 10 pour un bon compromis entre le biais et la variance de l'estimateur.

Conclusions

Pour bien mener un projet de DM

- Identifier et énoncer clairement les besoins.
- Créer ou obtenir des données représentatives du problème
- Identifier le contexte de l'apprentissage
- Analyser et réduire la dimension des données
- Choisir un algorithme et/ou un espace d'hypothèses.
- Choisir un modèle en appliquant l'algorithme aux données prétraitées.
- Valider les performances de la méthode.

Etapes de Traitement des données:

- Nettoyage des données
- Transformation des données
- Sélection des données
- Réduction des données

Techniques d'Extraction de connaissances

- Techniques descriptives :
- Techniques prédictives :

Nettoyage des données

Objectif :

Supprimer les données bruitées ou non pertinentes.

Questions :

- Que faire si certaines données sont manquantes ?
  - Certains clients n'ont pas donné leur adresse.
- Toutes les données sont-elles fiables (problèmes d'inconsistance) ?
  - Un même article appartient à différentes catégories (dans des magasins différents).
  - Le prix d'un même article est très supérieur à la normale dans un magasin donné.
- Que faire si certaines données sont numériques dans le cas où la technique d'extraction ne peut manipuler que des données symboliques ?

Données manquantes

Solutions :

- Ne pas tenir compte des tuples contenant des données manquantes (valeurs nulles).
- Remplir manuellement les champs non remplis.
- Utiliser les valeurs connues :
  - Remplacer un salaire manquant par le salaire médian des clients.
  - Prédire les valeurs manquantes, en le déduisant d'autres paramètres (salaire à partir de l'âge et de la profession).

Données bruitées

Plusieurs solutions : lissage, segmentation, régression linéaire.

Techniques de lissage (data smoothing) :

1. Trier les différentes valeurs de l'attribut considéré.
   {4, 8, 15, 21, 21, 24, 25, 28, 34}
2. Partitionner l'ensemble résultat.
   {{4, 8, 15}, {21, 21, 24}, {25, 28, 34}}
3. Remplacer les valeurs initiales par de nouvelles valeurs en fonction du partitionnement réalisé :
   - par la valeur moyenne des regroupements réalisés
     {9, 22, 29}
   - par les min et max des regroupements réalisés.
     {{4, 4, 15}, {21, 21, 24}, {25, 25, 34}}

Implique une perte de précision ou d'information.

Données bruitées

Techniques de segmentation (clustering) :

- Les valeurs similaires sont placées dans une même classe.
- On ne tient pas compte des valeurs isolées (dans une classe comportant trop peu d'éléments).

Techniques de régression linéaire :

- Hypothèse : un attribut Y dépend linéairement d'un attribut X.
  - Années d'expérience X et salaire Y.
- Trouver les coefficients a et b tels que $Y = aX + b$.
- Remplacer les valeurs de Y par celles prédites.

Données bruitées : régression linéaire

Données de départ :
Un ensemble de couples (X_i, Y_i).

Détermination des coefficients :

- Soient $\overline{X}$ et $\overline{Y}$ les valeurs moyennes des attributs X et Y.
- $a = \frac{cov(x,y)}{V(x)}$.
- $b = \overline{Y} - a\overline{X}$.

Données inconsistantes

Données inconsistantes dans une base de données :

- Contraintes d'intégrités ou dépendances fonctionnelles non respectées.
- Exemples :
  - La contrainte I_ID→I_CATEGORY n'est pas respectée au moment de l'intégration des données.
  - Unicité de clés non respectée.

Intégration des données

Objectif :

Regrouper les données provenant de différentes sources.

→ Problématique typique lors de la construction d'entrepôts de données.

Exemple :

Un attribut nommé C_ID dans la BD de Paris peut très bien se nommer CUST_ID dans la BD de Londres.

→ Utilisation de méta-données (XML) pour la mise en correspondance.

Transformation des données

- Lissage de données : utilisation de techniques de régression.
- Normalisation des données : normaliser certains attributs numériques afin qu'ils varient entre 0 et 1.
  - Pour ne pas privilégier les attributs ayant les plus grands domaines de variation (salaire/âge).
- Agrégation des données : opérations OLAP (On-Line Analytical Processing) permettant une analyse multidimensionnelle sur les BD volumineuses afin de mettre en évidence une analyse particulière des données.
  - Calculer les niveaux de ventes réalisées de tel produit par mois plutôt que par jour.
- Généralisation des données : remplacer les données finies par des données de plus haut niveau.
  - Remplacer les adresses precises des clients par leur code postal.
  - Remplacer l'âge des clients par « jeune », « adulte », « sénior ».

Discrétisation des connaissances

Répartition des valeurs des attributs :

À chaque étape, on cherche à découper l'intervalle de variation des données en K intervalles comportant le même nombre de valeurs.

On divise C_AGE= [0, 100] en $A_1 = [0, 20]$ et $A_2 = [20, 100]$ si 50 % des clients ont moins de 20 ans.

Entropie et classification à priori des données :

On cherche à caractériser les individus achetant les différents types de lait (entier, demi-écrémé, écrémé).

Facilité à appréhender le découpage obtenu :

On veut obtenir des intervalles du type [−12.5, 0] plutôt que [−12.536, 0.0005].

Discrétisation basée sur l'entropie (1/2)

Entropie d'un ensemble de données S :

Définition :

- S est découpé en k classes $C_1, \ldots, C_k$.
- $Ent(S) = \sum p_i . \log(p_i)$ avec $p_i = \frac{|C_i|}{|S|}$.

Propriétés :

- Ent(S) est maximale (égale à 0) si les données sont réparties dans une seule et même classe.
- Ent(S) est minimale si les données sont uniformément réparties dans toutes les classes.

Discrétisation basée sur l'entropie (2/2)

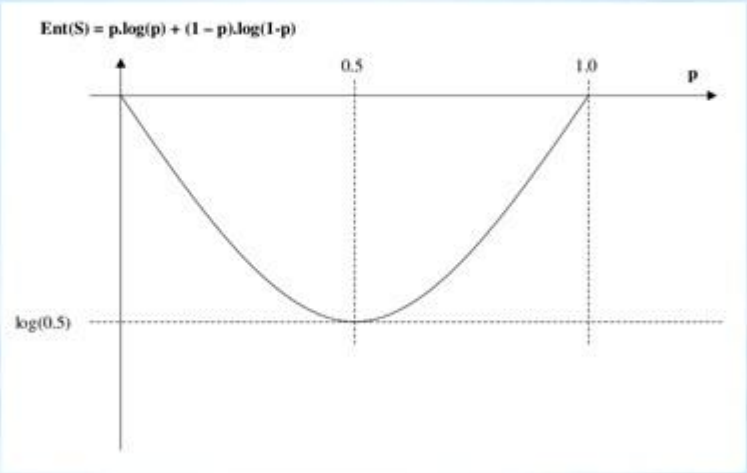
Méthode :

- Découper S = [a, b] en S_1 = [a, c] et S_2 = [c, b].
- Maximiser le gain d'information

$$I(S, c) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) - Ent(S).$$

- Arrêt du découpage si le gain devient insuffisant, quel que soit c.

Variation de l'entropie



Sélection des données

Objectif :

Garder uniquement les données pertinentes pour l'étude réaliser.

Exemple :

- Doit-on s'intéresser à toutes les catégories de produits de vente ?
- Doit-on s'intéresser aux ventes réalisées il y a plus d'un an ?

Réduction des données

Réduction en ligne par échantillonnage :

- Pour des raisons de performance.
- Du fait de la complexité importante des algorithmes d'extraction.
- Plusieurs méthodes : échantillonnage aléatoire (avec ou sans remise), échantillonnage par clustering/segmentation.

Réduction en colonne par suppression des attributs redondants :

- Cas triviaux (âge et date de naissance).
- Via une analyse des corrélation entre attributs :

$$corr_{A,B} = \frac{P(A \wedge B)}{P(A).P(B)} = \frac{P(B/A)}{P(B)}$$

- Indépendance : $corr_{A,B} = 1$ si $P(B/A) = P(B)$.
- Corrélation positive : $corr_{A,B} > 1$ si $P(B/A) > P(B)$.

Réduction des données
Matrice de contingence

Exemple de matrice de contingence :

	Avec pain	Sans pain	Total
Avec beurre	4.000	3.500	7.500
Sans beurre	2.000	500	2.500
Total	6.000	4.000	10.000

Analyse de corrélation :

- $P(Beurre) = \frac{7.500}{10.000} = 0.75$ et $P(Pain) = 0.6$.
- $P(Beurre \wedge Pain) = \frac{4.000}{10.000} = 0.4$.
- $corr_{Pain,Beurre} = \frac{0.4}{0.75 \times 0.6} = 0.89 < 1$
  → Indique une corrélation négative.

NB: En logique classique « ¬ » (non), « ∨ » (ou), et ∧ (et)

Réduction des données

Qualité de la corrélation

Coefficient de corrélation :

$$r_{A,B} = \frac{\sum(A_i - \bar{A})(B_i - \bar{B})}{\sigma_A \cdot \sigma_B}$$

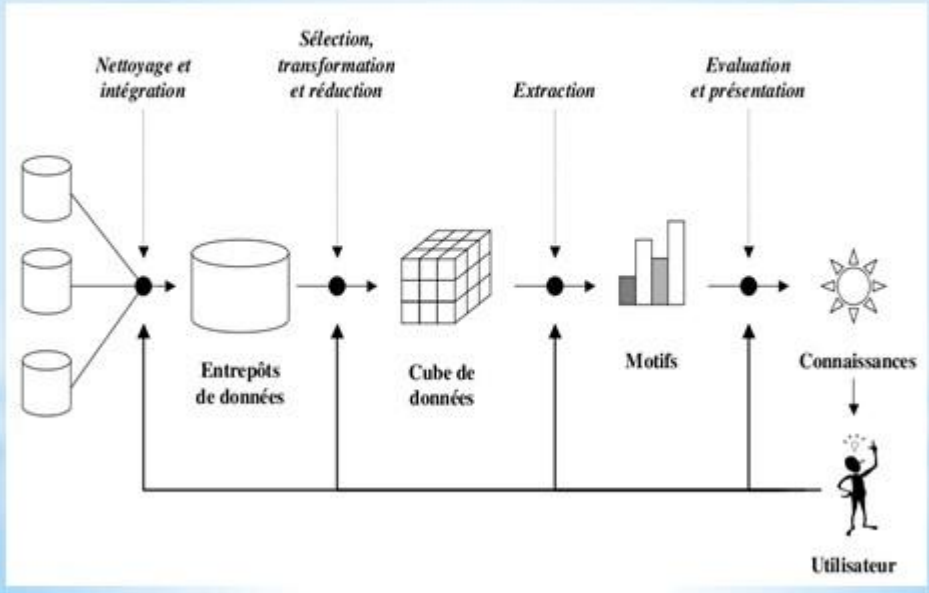
avec $\sigma_X = \sqrt{\sum(X_i - \bar{X})^2}$.

Signification :

Plus $r_{A,B}$ s'éloigne de zèro, meilleure est la corrélation :

- $r_{A,B} = +1$: corrélation positive parfaite.
- $r_{A,B} = -1$: corrélation négative parfaite.
- $r_{A,B} = 0$: absence totale de corrélation.

Extraction de connaissances



Extraction de connaissances (1/2)

Techniques descriptives :

- Visent à mettre en évidence des informations présentes, mais cachées dans les gros volumes de données.
  Cas de la segmentation de la clientèle, de la recherche d'association de produits sur les tickets de caisse.
- Permettent de réduire, de résumer et de synthétiser les données.
- Pas de variable « cible » à prédire.

Exemples :

- Techniques de segmentation/clustering : nuées dynamiques, segmentation hiérarchique, réseaux de neurones.
- Extraction de règles d'association.

Extraction de connaissances (2/2)

Techniques prédictives :

- Visent à extrapoler de nouvelles informations à partir des informations présentes.
  Cas général du scoring (impayés, attrition, crédit).
- Permettent d'« expliquer » les données.
- Il existe une variable « cible » à prédire.

Exemples :

- Classification/discrimination (variable cible qualitative) :
  - analyse discriminante ;
  - arbres de classification ;
  - réseaux neuronaux multi-couches.
- Prédiction (variable cible quantitative) :
  - régression linéaire (simple et multiple) ;
  - arbres de régression.

Présentation des connaissances

Problème :
Comment représenter/visualiser les connaissances extraites ?

Formules logiques :

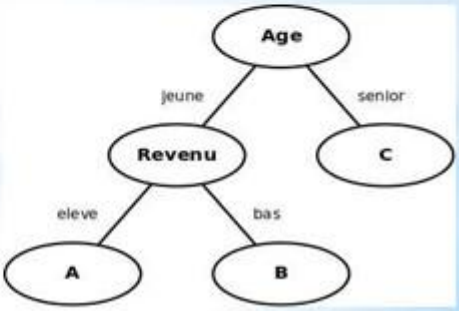
- Age(X, 'jeune'), Revenu(X, 'eleve') → class(X, 'A') [1.402].
- Age(X, 'jeune'), Revenu(X, 'bas') → class(X, 'B') [1.038].
- Age(X, 'senior') → class(X, 'C') [2.160].

Présentation des connaissances

Tableau :

Age	Revenu	Class	Count
jeune	eleve	A	1.402
jeune	bas	B	1.038
senior	eleve	C	786
senior	bas	C	1.374

Arbre de décision :



Les logiciels de data mining

Il existe de nombreux logiciels de statistiques et de data mining.

- Parmi les gros logiciels, on peut citer :
  - Clementine de SPSS. Clementine est la solution de data mining la plus vendue dans le monde.
  - Entreprise Miner de SAS.
  - Statistica Data Miner de StatSoft
  - XL Miner (data mining sous excel)
  - ORACLE, comme d'autres SGBD, fournit des outils de data mining
- Parmi les logiciels gratuits, on peut citer :
  - TANAGRA, logiciel de data mining gratuit pour l'enseignement et la recherche.
  - ORANGE, logiciel libre d'apprentissage et de data mining.
  - WEKA, logiciel libre d'apprentissage et de data mining.

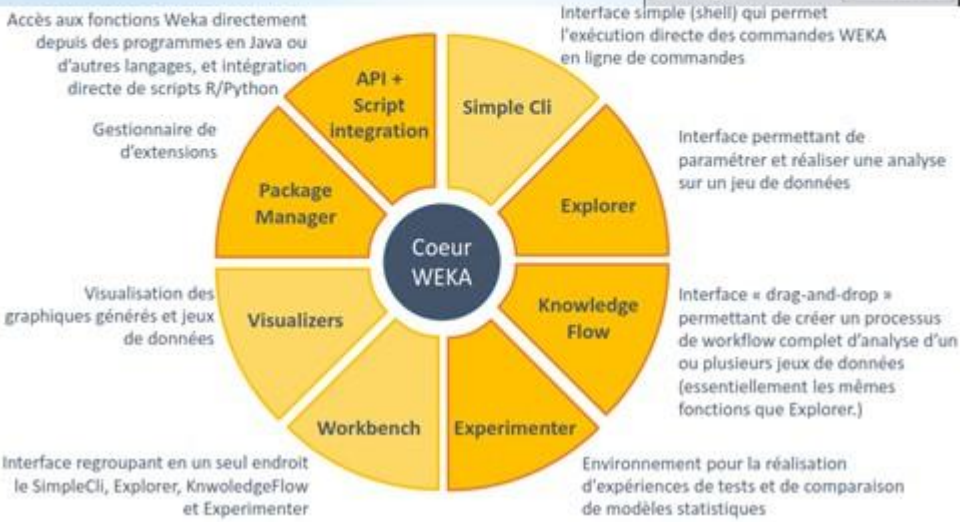
WEKA

- Weka 3: Data Software Mining en Java
- Weka est une collection d'algorithmes d'apprentissage machine pour les tâches d'exploration de données. Les algorithmes peuvent être soit appliquées directement à un ensemble de donnéesou appelés à partir de votre propre code Java.
- Weka contient des outils pour les donnéesde pré-traitement, la classification, la régression, le regroupement, les règles d'association, et la visualisation.
- Il est également bien adapté pour le développement de nouveaux programmes d'apprentissage machine.
- Site web: http://www.cs.waikato.ac.nz/ml/weka/

TP WEKA : pour le Dataset IRIS.CSV déterminer l'algorithme d'apprentissage qui a un bon score d'évaluation

Fonctionnalités et mode opératoire

Interfaces utilisateur Weka



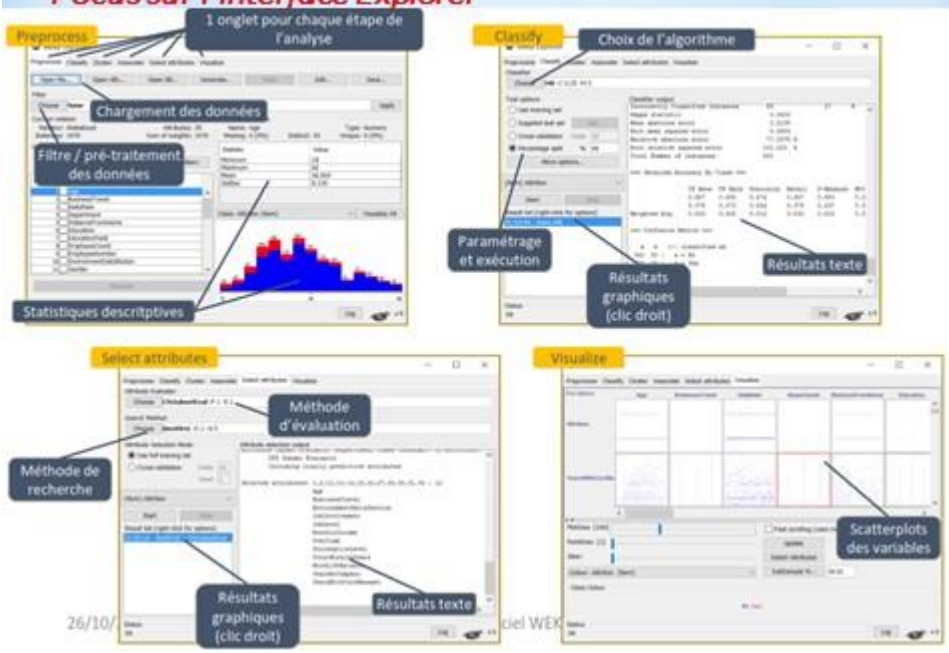
Fonctionnalités et mode opératoire

Principales fonctionnalités de traitement des données

Preprocessing	Import, inspection et préparation/filtre des données
Classification	Mise en œuvre des différents algorithmes de classification
Clustering	Accès aux techniques de clustering comme l'algorithme de k-means
Association	Accès aux apprentissages par règles d'association qui essaient d'identifier toutes les relations importantes entre les variables
Feature (attribute) selection	Choix des variables les plus pertinentes et prometteuses
Visualization	Affichage graphique scatterplot, arbres

Fonctionnalités et mode opératoire

Focus sur l'interface Explorer



Récapitulatif

1. Donnez la signification des fonctions L et R en DM
2. Quelles sont les taches de data mining ?
3. Donnez des exemples de taches : Estimation, prédiction, association
4. Quels sont les types de modèles peuvant être exploités par le DM ?
5. Quelle est la différence entre régression et classification en apprentissage supervisé ?
6. Quelle est la formule de calcul de l'erreur en validation croisée ?, combien de modèle faut-il apprendre si K=N?
7. Citez les étapes pour bien mener un projet de DM
8. Quelles sont les étapes de Traitement des données avant éxtraction de connaissaces ?
9. Quelles les techniques utilisées pour l'extraction des connaissances ?
10. Citez qq logiciels propriétaires du DM
11. Citez qq logiciels gratuits du DM
12. Donnez qq caractéristiques de WEKA ?
13. Donnez les principales fonctionnalités de traitement de donnéessur WEKA

III. Web-Mining

Introduction au Web mining

Le Web Mining s'est développé à la fin des années1990. Ce domaine consiste à utiliser l'ensemble des techniques du Data Mining afin de développer des approches et des outils, permettant d'extraire des informations pertinentes à partir d'un large volume de donnéesdu web (documents, traces d'interactions, structure des pages, liens visités, parcours dans le site..)

Basé sur le type de donnéesà étudier dans le domaine du Web Mining, trois grands axes de recherches ont été envisagés portant sur le Web Content Mining (mining using HTML pages), le Web Structure Mining(mining using URL links) et le Web Usage Mining (mining using clicks, visits et logs).

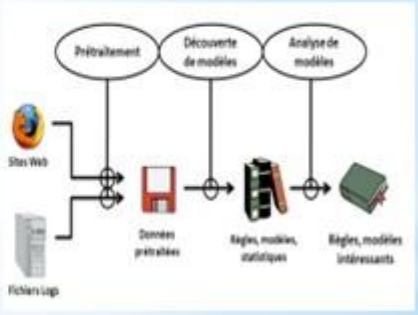
* Processus du Web mining

Le processus du Web Mining se déroule en trois étapes :

1. Collecte des donnéessur l'utilisateur,

2. Utilisation de ces donnéesà des fins de Personnalisation.

3. Présentation à l'utilisateur d'un contenu ciblé.



Processusdu WebMining

Les bases de données(BD) comparées à la recherche d'information (RI)

Format des données:

BD : donnéesstructurées. Sémantique claire, basée sur un modèle formel

RI : non-structurées - texte libre

Requêtes :

BD : formelles (par exemple, SQL)

RI : souvent - langues naturelles (recherche de mots-clé)

Résultats :

BD : résultat exact

RI : trouver les informations les plus pertinents pour une requête donnée

Le fonctionnement d'un moteur de recherche. Etapes :

L'exploration ou crawl faite par un robot d'indexation qui parcours récursivement tous les hyperliens qu'il trouve en récupérant les ressources intéressantes. L'exploration est lancée depuis une ressource pivot, comme une page d'annuaire web.

L'indexation des ressources récupérées extraire les mots considérés comme significatifs du corpus de documents `a explorer. Les mots extraits sont enregistrés dans une BD. Un dictionnaire inverse ou l'index TF-IDF permet de retrouver rapidement dans quel document se situe un terme significatif donné.

La recherche un algorithme identifie dans le corpus documentaire (en utilisant l'index), les documents qui correspondent le mieux aux mots contenus dans la requête. Les résultats des recherches sont donnés par ordre de pertinence supposée.

Contenu - GoogleBot

Google calcule pour chaque page un score de crawl => niveaux de crawl de GoogleBot :

couche de base : la plupart des pages du web. Elles sont crawlées avec une fréquence liée à la fréquence de mise à jour du contenu ainsi qu'à leur PageRank.

couche quotidienne : un petit nombre de pages crawlées de façon quotidienne.

couche temps réel : un nombre très petit de pages crawlées avec une fréquence de l'ordre de la minute, l'heure,... (ex. l'actualité)

Les caractéristiques clé de Google :

Utilisation des indexes basés sur le contenu

Indexes peuvent être partitionnés sur plusieurs machines ou bien les données sont dupliquées sur les machines et les requêtes sont distribuées parmi celles-ci

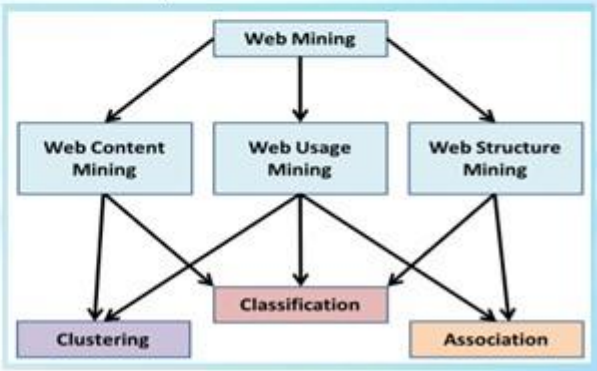
PageRank ordonner les documents trouvés, en utilisant les liens entre pages comme indice de pertinence

Qu'est ce que le Web Mining ?

Le Web Mining a pour objectif l'extraction de connaissances à partir de la structure formée par les liens entre pages/objets/communautés, du contenu des pages Web ou de données d'usages (typiquement des logs de serveurs web).
Bien que cette discipline se base fortement sur le Data Mining, on ne saurait réduire cette discipline à la seule application d'algorithmes existants dans ce contexte nouveau. Ceci est dû aux particularités des donnéesdu Web.
De nouvelles tâches ont été mises au point au cours de la dernière décennie. Une classification des approches du Web Mining fait désormais consensus et est basées sur les types de données manipulées par les approches. Nous décrivons ces 3 catégories, ci-dessous.

1– Web structure mining.

La fouille de structures issues du Web permet l'extraction de connaissances à partir des liens qu'il existe entre les pages Web ou toutes autres entités reliées entre elles. Les liens hypertexts représentent en quelques sortes la structure du Web. Typiquement, l'identification de pages importantes, est très utilisée par les moteurs de recherche. Ces approches permettent également de découvrir des communautés d'utilisateurs qui partager des intérêts communs. Il est important de noter que les approches de Data Mining classiques n'abordent pas cette thématique car les donnéestraitées ne contiennent pas de liens.



2— Web content mining.

La fouille de contenus de pages Web permet l'extraction de motifs à partir des contenus des pages Web. Par exemple, il est possible de regrouper des pages Web selon certaines thématiques communes comme le feraient des techniques de Data Mining plus classiques.
Il existe toutefois des spécificités associées au donnéesdu Web. Par exemple, la fouille des avis d'internautes sur des produits, e.g., musiques, films, voyages, permet de découvrir automatiquement l'opinion des consommateurs face à un produit pour accroître la qualité des systèmes de recommandation

3 — Web usage mining.

La fouille de donnéesd'utilisation du Web fait référence à l'analyse automatique des logs stockés sur les serveurs Web. De nombreux algorithmes peuvent être appliqués sur ce type de donnéescar celles-ci sont structurées, rendant possible l'utilisation de techniques standards. Une des principales difficultés lorsque l'on souhaite considérer de telles donnéesest leur prétraitement. En effet plusieurs problématiques de recherche sont apparues quant à l'identification des sessions utilisateurs ou encore la suppression du bruit.

Web structure mining

Les premiers moteurs de recherche retrouvaient les documents pertinents seulement en fonction de la similarité du contenu avec la requête de l'utilisateur.
Depuis le milieu des années1990, il est devenu de plus en plus évident que cette simple similarité n'était pas suffisante pour au moins deux raisons principales :

- D'abord, l'augmentation rapide du nombre de pages Web a rendu très difficile l'ordonnancement des documents retrouvés. Par exemple, la requête « classification technique » retourne quelques 190000 résultats. Classer ces documents seulement en fonction de la similarité de leur contenu pour ne retenir que 30 ou 40 documents à afficher dans la première page des documents est très difficile.
- Ensuite, il est très facile de « falsifier » le contenu d'une page de telle sorte à la rendre similaire à des très nombreuses requêtes. En effet, un concepteur de pages Web peut facilement ajouter artificiellement pléthore de mots-clés stratégiques pour augmenter la similarité de ses pages avec des requêtes.

Web structure mining

- La prise en compte de liens hypertextes est alors apparue comme la solution. En effet, à la différence des documents textuels « classiques » qui sont souvent considérés comme indépendants, les pages Web sont inter-connectées grâce aux hyperliens qui représentent une source d'information importante. Ces liens peuvent être internes, i.e., permettant la définition de la structure du site Web, ou externes, i.e., et implicitement indique à l'utilisateur que le site ciblé est une source d'information valide. Ces liens externes doivent alors de toute évidence être considérés par les moteurs de recherche lors de la recherche d'information.
- Au cours de la période 1997-1998, deux des plus influents algorithmes de recherche d'information basés sur l'influence des liens ont été proposés : PageRank et HITS. Ces deux algorithmes ont originellement été mis au point pour l'étude des réseaux sociaux. Ils exploitent tous les deux la structure du Web formée par les liens afin de définir « le prestige » ou « l'autorité » d'un site Web.

Web structure mining

1 PageRank

Développé dans l'université de Stanford par Brin et Page (Page et al., 1998), puis intégré dans le moteur de recherche Google, PageRank est un algorithme d'analyse de structure des liens entre pages très puissant pour mesurer la peupularité d'une page Web (Zhang et al., 2006).

A l'inverse de HITS, PageRank calcul une seule mesure de qualité (on dit aussi popularité ou prestige) pour une page au moment du crawl. Cette mesure et alors combinée avec un autre score similaire à ceux utilisés dans les systèmes de recherche d'information (SRI) lors de l'interrogation. Il est réputé être plus efficace, car la mesure de popularité d'une page est précalculée indépendamment de la requête. Actuellement, il permet de capturer la notion d'importance d'une page en se basant sur les informations des liens.

Ainsi, l'intérêt d'une page A dans le résultat est fonction du nombre de pages qui pointent vers elle.

De plus, et afin d'éviter les liens vains, qui peuvent être de nos jours générés facilement de manière artificielle, cet algorithme étend l'idée des citation, en prenant en compte non seulement le nombre de liens référençant une page mais aussi l'importance et le poids des pages pointant vers elle (Zhang et al., 2006).

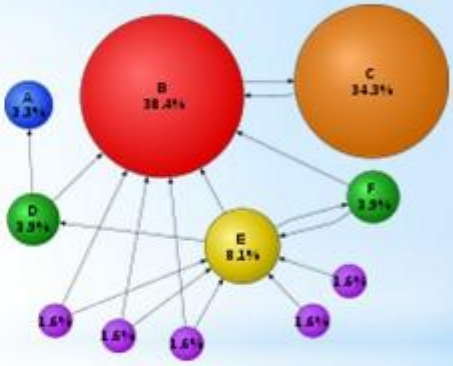
Web structure mining

1 PageRank

L'expression suivante donne la formule de calcul du score r(v), d'une page v utilisé dans PageRank, α étant une constante de normalisation, pa et ch deux fonctions représentant les parents et les fils d'une page (Baldi et al., 2003) :

$$r(v) = \alpha \sum_{w \in \mathrm{pa}[v]} \frac{r(w)}{|\mathrm{ch}[w]|}$$

Remarquons que chaque parent w contribue par une quantité proportionnelle à son score r (w) et inversement proportionnelle à son degré de sortie.



Le *PageRank* d'une page a tendance à être d'autant plus élevé que la somme des *PagesRanks* des pages qui pointent vers elle est élevée.

Web structure mining

2 HITS

L'algorithme HITS (Hyperlink Induced Topic Search) proposé par Kleinberg (Kleinberg, 1999) s'appuie sur un principe simple : toutes les pages n'ont pas la même importance, et ne jouent pas le même rôle. Certaines sont importantes ayant un contenu informatif (authoritative ou de référence), et d'autres sont des pages hubs contenant des liens vers des pages de référence.

Contrairement à PageRank, HITS démarre par un graphe d'une requête qu'il soumis à un SRI standard (les systèmes de recherche d'information), pour collecter un premier ensemble de pages liées au sujet recherché (les nœuds du graphe) appelé ensemble racine (root set) noté R. Ce dernier est utilisé pour obtenir un deuxième ensemble étendu (expanded set) noté E, par l'ajout de tous les nœuds fils, et un nombre fixe de nœuds parents de l'ensemble racine. Vu le nombre important de nœuds pouvant être impliqués, et afin d'alléger l'algorithme, les arrêtes représentant les liens qui relient les nœuds du même serveur seront éliminés. Ils sont seulement considérés comme des liens de navigation.

Web usage mining

- Si le WCM et le WSM utilisent les donnéesWeb primaires ou réelles, le Web Usage Mining (WUM) opère sur les donnéessecondaires de celui-ci, générées par l'interaction des utilisateurs avec les sites Web (Kosala et al., 2000).
- Avec le succès et la popularité du Web, de grandes quantités de donnéessont générées chaque jour par les serveurs Web, en enregistrant dans des fichiers de type log (journaux) les traces de navigation, appelée aussi clickstream (flux continu de clics), (Kimball et al., 2000), englobant les parcours et les actions, effectuées par les utilisateurs lors de leurs visites.
- Le WUM est donc la découverte et l'extraction de connaissances à partir des donnéesd'interaction des utilisateurs avec le Web, sous forme de modèles et de motifs de navigation, en utilisant les techniques du data mining (Cooley et al., 1997).
- De par la nature distribuée du Web, la journalisation (logging) dans le cadre du WUM peut être effectuée dans trois niveaux : les journaux de traces au niveau des serveurs Web, les journaux consignés par les serveurs proxy et les traces de navigation coté client (Cooley, 2000).

Web content mining

le Web contient une masse importante de contenu non structuré, i.e., des donnéestextuelles. Analyser ces textes peut fournir un grand nombre d'informations qui peuvent se révéler utiles pour un décideur. Dans ce cours, nous nous focalisons sur une tâche bien particulière : la fouille d'opinions. Cette tâche n'est pas seulement difficile, car elle implique l'utilisation de techniques issues du traitement automatique de la langue naturelle (TALN)

Des techniques ont été développées pour analyser cette formidable source d'informations. nous nous faciliterons sur 3 tâches particulières :

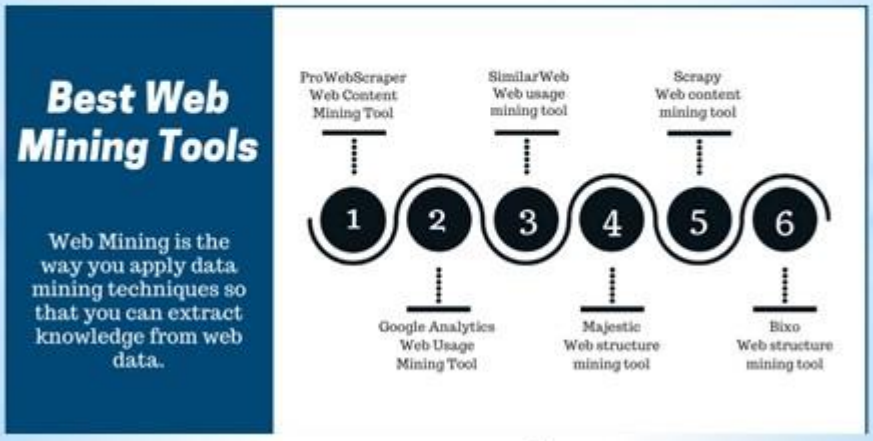
1— Classification de sentiments. La fouille d'opinion est traitée comme un problème de classification. Ces approches fonctionnent généralement au niveau d'un document et permettent de dire si une opinion positive ou négative a été formulée sur ce document

Web content mining

2 — La fouille d'opinions basée sur les caractéristiques. Ces approches fonctionnent à un niveau de granularité un peu plus fin : la phrase généralement. Elles permettent de découvrir les raisons qui poussent un internaute à aimer ou pas un produit ou un service.

3— La détection de spam. Il est très important de pouvoir détecter les faux avis, i.e., des avis malhonnêtes, de ceux qui sont réellement formulés par les internautes. La détection de ces fausses opinions est un problème difficile car les auteurs sont rarement correctement identifiés.

Les meilleures outils de WEB MINING



Google Analytics (Web Usage Mining Tool)

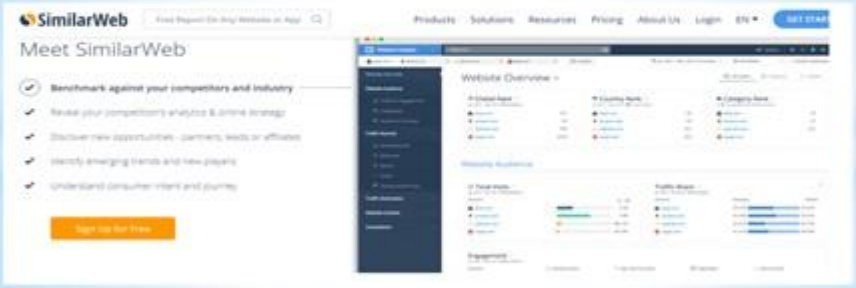
* Google Analytics est considéré comme l'un des meilleurs outils d'analyse des activités. Il peut suivre et signaler le site Web. Il peut suivre et signaler le trafic du site Web.

Caracteristiques:

- Analyse du rendement de la publicité et de la campagne
- Analyse et mise à l'essai du site Web
- Intégration facile avec le produit de Google comme, Adsense, Adwords, Google Display Network, Google Tag Manager, etc

SimilarWeb (Web usage mining tool)

SimilarWeb est un puissant outil de business intelligence. Il offre des renseignements sur le trafic et le marketing pour n'importe quel site Web.

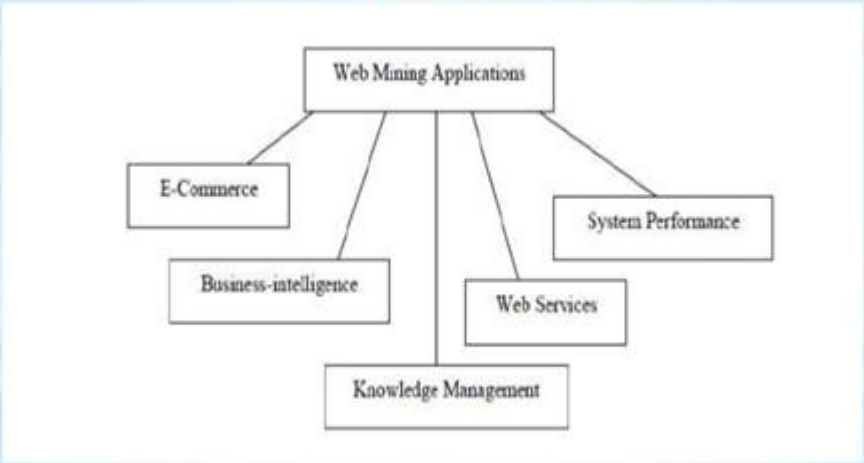


Caractéristiques

* Indicateurs de trafic et d'engagement
* Optimisation des moteurs de recherche et mots clés PPC
* Intérêts du public

Exemple d'analyse Similarweb de site web : Menara.ma

*Exemple d'applications



Récapitulatif

1. Donnez les étapes de fonctionnement d'un moteur de recherche
2. Quels sont les niveaux de crawl de GoogleBot pour calculer le score de crawl d'une page web
3. Qu'est ce que le PageRank sur google search
4. Donnez la formule de calcul du score r(v) de PageRank
5. Quel est le principe de l'algorithme HITS
6. Quel est l'objectif principale du web mining
7. Donnez le schéma de la structure web mining
8. Qu'est ce que le Web content mining
9. Qu'est ce que le Web usage mining
10. Quels sont les niveaux de journalisation (logging) dans le cadre du WUM
11. Quelles sont les 3 tâches particulières du TALN dans le WCM ?
12. Donnez les meilleurs applications du Web mining

IV - Text Mining

Définition de text mining :

Le Text Mining, également appelé fouille de textes ou extraction de donnéesà partir de textes, est un ensemble de méthodes, de techniques et d'outils pour exploiter les documents non structurés que sont les textes écrits, comme les fichiers bureautiques de type word, les emails, les documents de présentation de type PowerPoint...etc. Pour extraire du sens de documents non structurés, le Text Mining s'appuie sur des techniques d'analyse linguistique. Le Text Mining est utilisé pour classer des documents, réaliser des résumés de synthèse automatique ou encore pour assister la veille stratégique ou technologique selon des pistes de recherches prédéfinies. Schématiquement, on peut énoncer :

TEXT MINING = LINGUISTIQUE + DATA MINING

Les tâches de Text Mining :

Le Text Mining cherche des réponses aux questions difficiles ou impossibles à résoudre avec les seuls moteurs de recherche.
Des exemples de tels services incluent :

- Résumer des documents qui décrivent une consommation du produit dans certaines régions ;
- Etudier des réclamations des clients, raisons des changements de comportements de consommation, analyse de l'image de l'entreprise, ...
- Faire la gestion de la relation client : orienter mes mails clients reçus sur le site vers les services adéquats et les aider à répondre le plus rapidement et correctement possible ;
- Connaître les réseaux relationnels des personnes ou entreprises.

Chacune de ces tâches sera un cas particulier du schéma général de la figure ci-dessous

Schéma général d'une tache du Text Mining

Processus de Text Mining :

Les étapes nécessaires pour effectuer le processus de text mining sont :

- L'acquisition : Source de donnéestelle que : corpus textuels, bibliothèques électroniques, Web...etc ;
- Le filtrage : Sélection des mots les plus pertinents (techniques de sélection d'attribut) ;
- Le nettoyage des données: Segmentation du texte, élimination des mots vides, lemmatisation;
- L'identification des mots pertinents : Analyse statistique (n-gram), analyse sémantique, analyse syntaxique ou structurelle (extraction d'attribut) ;
- L'extraction des connaissances : Application de l'un des algorithmes de la fouille de textes

Applications du Text Mining

L'importance de la fouille de textes (Text Mining) ne cesse d'évoluer d'un jour à l'autre. Plusieurs domaines vitaux exploitent les techniques et les outils du Text Mining pour trouver l'information pertinente fouillée dans des quantités énormes de textes de différentes formes, parmi ces domaines on peut citer:

- La recherche d'information.
- Les applications biomédicales.
- Le filtrage des communications.
- Les applications de sécurité.
- La gestion des connaissances.
- L'Analyse du sentiment.

Techniques liées à la fouille de textes :

La fouille de textes s'apparente à d'autres domaines avec qui elle est très complémentaire :

- le traitement automatique des langues (TAL)
- la recherche documentaire (RI)
- l'extraction de l'information (EI).

1-Le traitement automatique des langues « TAL» :

Depuis une quinzaine d'années, avec la généralisation de l'outil informatique et d'Internet, les applications du TAL au sens large du terme se multiplient dans les disciplines philologiques. Le TAL est une discipline à la frontière de la linguistique et de l'informatique, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain.

2. La recherche d'information « RI»

La recherche d'information « RI » s'intéresse aux documents dans leur globalité et aux thèmes qu'ils abordent, pour comparer les documents et détecter des typologies. Elle cherche à détecter tous les thèmes présents.

3. L'extraction d'information « EI » :

L'extraction d'information « EI » recherche des informations précises dans les documents, sans les comparer, en tenant compte de l'ordre et de la proximité des mots pour discriminer des énoncés différents ayant des mots_clés identiques.

L'extraction d'information consiste en l'alimentation d'une base de donnéesstructurée à partir de donnéesexprimées en langage naturel. Il s'agit de détecter dans le texte en langage naturel les mots correspondant à chaque champ de la base de données.

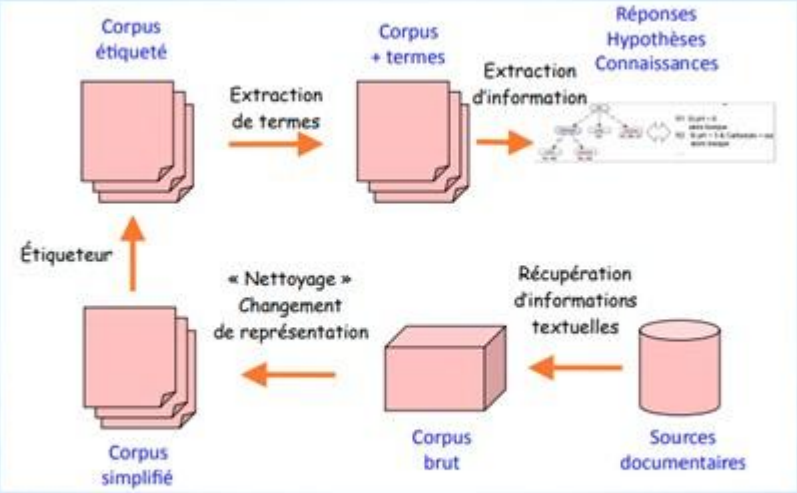
Méthodes utilisées pour la fouille de textes

La fouille de textes fait appel à l'analyse de donnéesqui se caractérise par deux grandes familles de méthodes :

▫ Les méthodes de classification qui produisent un regroupement d'objets ou d'individus décrits par un certain nombre de variables ou de caractères (classification de type nuées dynamiques, Classification Ascendante Hiérarchique - CAH -).

▫ Les méthodes factorielles qui font essentiellement des représentations graphiques caractérisant les liens entre les différents critères (Analyse en Composantes Principales, Analyse Factorielle des Correspondances, Analyse des Correspondances Multiples).

Processus de Text Mining



- Un corpus est un ensemble de texte éventuellement annotés

Text mining TAL : Quelques applications

- Traduction automatique
- Filtrage / Classification
  - Sentiment favorable ou pas …
- Recherche d'information
- Extraction d'information
- Résumé automatique
- Question / réponses (interfaces en langage naturel)
- …

Récapitulatif

1. A quoi sert le Text Mining
2. Donnez qq exemples de services texte traiter par le TM
3. Donnez le processus des étapes successifs de TM
4. Donnez qq exemples de domaines d'exploitation du TM
5. Que signifie le traitement automatique du langage TAL
6. Que signifie la recherche documentaire (RI)
7. Que signifie l'extraction de l'information (EI).
8. Quelles sont les Méthodes techniques utilisées pour la fouille de textes
9. Donnez qq applications du Text mining

L'outil KNIME

Présentation KNIME

Konstanz Information Miner, connu sous l'acronyme KNIME, est un logiciel de data mining développé en 2004 à l'université de Konstanz en Allemagne. C'est un logiciel open source sous la licence GPL(General Public License), écrit en langage JAVA sous la plate forme Éclipse SDK. KNIME est une plate forme modulaire pour la conception et l'exécution de flux de travail(workflows) en utilisant des composants prédéfinis appelés nœuds. Contrairement à beaucoup de logiciels du domaine, qui offrent une version allégée gratuite et une autre version complète(professionnelle) payante, la plateforme de KNIME est complète, il n'y a pas de restrictions par rapport au volume de données, ainsi les projets qui manipulent de grands volumes de donnéesdépendent entièrement des ressources de l'utilisateur(taille disque, RAM ...) et non pas de la plate forme KNIME

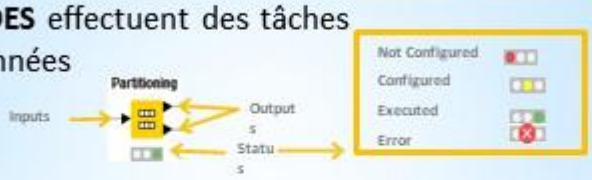
Présentation KNIME

Open for Innovation
KNIME

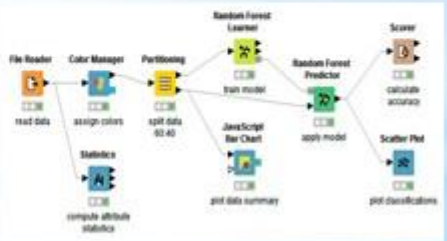
- Un outil pour l'analyse des donnéesanalytiques, manipulation, visualisation, et reporting
- Basé sur le graphical programming paradigm
- Fournis divers possibilités d'extensions:
  - Text Mining
  - Network Mining
  - Cheminformatics
  - Weka, H2O, etc.

*Visual KNIME Workflows

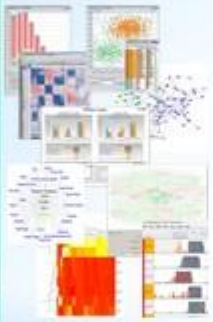
Les NOEUDES effectuent des tâches sur des données



Nodes are combined to create **WORKFLOWS**



*Visualization



- Interactive Visualizations
- JavaScript-based nodes
  - Scatter Plot, Box Plot, Line Plot
  - Networks, ROC Curve, Decision Tree
  - Adding more with each release!
- Misc
  - Tag cloud, open street map, molecules
- Script-based visualizations
  - R, Python

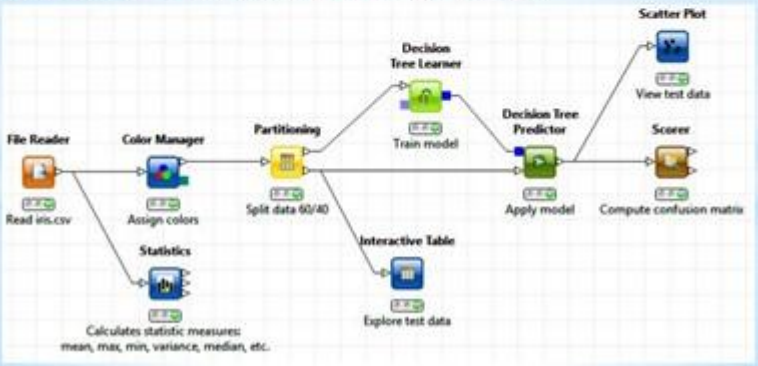
* Over 1500 native and embedded nodes included:

Le Nœud :

Le nœud est une entité qui exécute une tâche bien définie. L'algorithme de racinisation "Porter" est un bel exemple de nœud KNIME.

Le flux de travail (Workflows) :

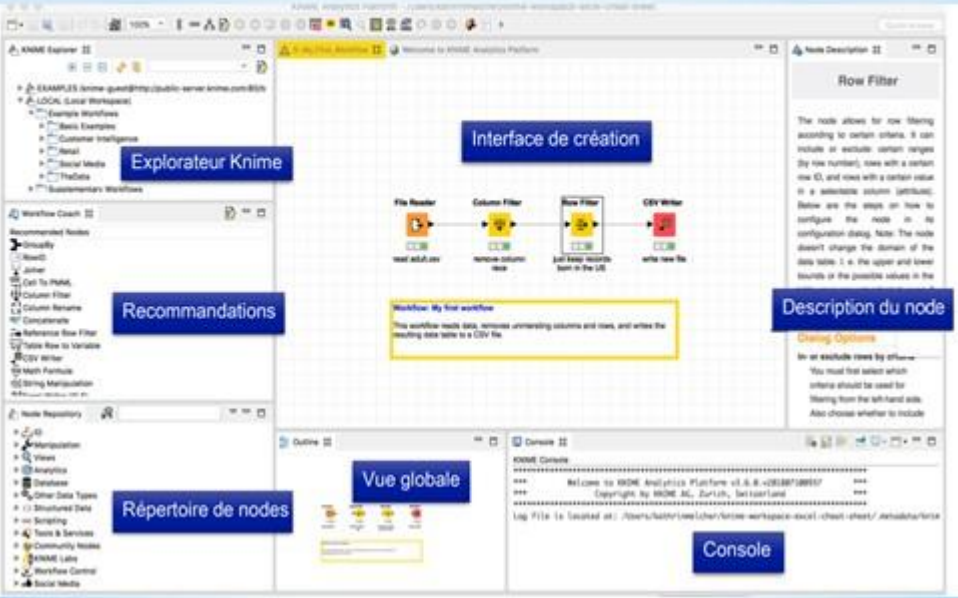
Le Workflows est une succession logique de nœuds .



*Data Access

- Databases
  - MySQL, MS SQL Server, PostgreSQL
  - any JDBC (Oracle, DB2, ...)
- Files
  - CSV, txt
  - Excel, Word, PDF
  - SAS, SPSS
  - XML, JSON
  - PMML
  - Images, texts, networks, chem
- Web, Cloud
  - REST, Web services
  - Twitter, Google
- Big Data
  - Spark
  - Impala
  - Vertica
  - HDFS support, Hive, In-database processing

*The KNIME Workbench



La méthodologie:

1 Le corpus :

Plusieurs corpus de documents texte sont disponibles sur internet, à l'image de Reuters-21578, 20Newsgroups, Ohsumed, DEFT'05 et bien d'autres. Certains sont entièrement gratuits et disponibles à 100%, d'autres nécessitent plus au moins des étapes d'authentifications ainsi que des renseignements personnels et professionnels (tel que l'organisme de recherche par exemple), tandis que d'autres exigent un coût pour certaines fonctionnalités.

Exemple le corpus Ohsumed. C'est un ensemble de 23 176 documents, extrait de la base MEDLINE (base d'informations médicales en ligne), il est constitué principalement d'extraits de journaux médicaux rédigés en langue Anglaise. Les documents sont divisé en 23 catégories distinctes, et chaque catégorie est représentée par un répertoire contenant tous les documents à qui cette dernière est assignée

2 Le prétraitement :

Dans cette section nous aborderons les différentes manipulations que nous avons effectuées sur nos documents d'entrée, nous détaillerons le fonctionnement de chaque nœud ainsi que son paramétrage et surtout les motivations de son choix. Notons que cette première étape sert seulement à la préparation de nos documents, afin qu'il soit possible d'extraire les descripteurs et ainsi calculer leur pondérations.

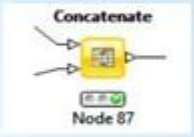
- Analyseur des fichiers plats :

Ce nœud sert à lire les fichiers plats. Nous insérons notre ensemble de documents au travers d'un champ de chargement. Chaque catégorie est insérée avec un seul nœud. Une fois le paramétrage effectué, nous avons qu'à cliquer sur le bouton exécute pour lancer l'analyse et lire l'ensemble de nos documents. Le résultat fourni est sous forme d'un tableau où chaque enregistrement représente un seul document.

- Filtre d'enregistrements :

Ce nœud sert à filtrer les enregistrements, nous pouvons les sélectionner selon leurs numéros ou la valeur de l'un de leurs attributs. .

- Concaténation : Ce nœud sert à regrouper toute les entrées en un seul document.



- Suppression des signes de ponctuation: Ce nœud sert à éliminer tous les signes de ponctuation, en l'exécutant, nous aurons une table dans laquelle nous pouvons comparer le document original et celui traité.

- Filtre de nombres : Sert à éliminer tous les nombres.

- Convertisseur : Sert à ramener toutes les lettres minuscules vers la majuscule ou vis versa.

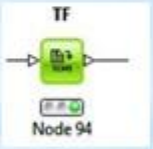
- Filtre de mots vides : Sert à éliminer les mots vides

- Nœud Porter : C'est un nœud qui utilise le plus célèbre des algorithmes de racinisation.

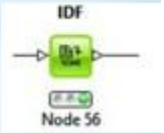
3 Les pondérations : La deuxième étape de la préparation de nos documents est le calcul des pondérations. Il y a, certes, différentes pondérations dans le domaine du texte mining, nous rencontrons souvent de nouvelles manières de les calculer à travers notamment les articles scientifiques et les livres de data mining.

- Sac à mots : Sert à grouper tous les descripteurs qui apparaissent au moins une fois dans tout le corpus.

- Fréquences des termes (TF) : se limite généralement sur la fréquence d'un mot-clé sur le document.

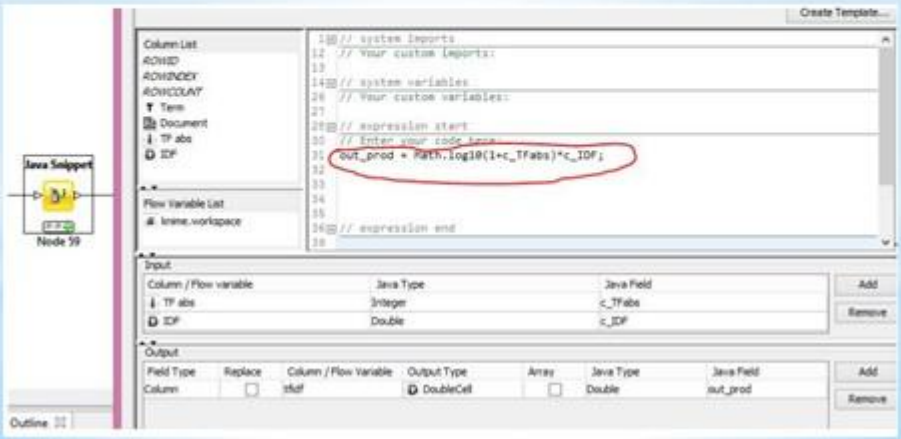


- Fréquences inverses du document (IDF) : est une formule qui réduit la valeur des mots-clés les plus fréquents et augmente la valeur des termes et des phrases uniques ou moins fréquents dans un contenu

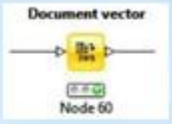


- Java Snippet :

C'est l'un des nœuds les plus importants de Knime. l'une des plus intéressante caractéristique que nous offre Knime était la possibilité d'écrire du code JAVA et de combiner des équations mathématiques. C'est avec le nœud Java Snippet que nous pouvons le faire. exemple combiner la pondération TF-IDF.



- Document vector : Notre vecteur d'entrée, comportant tous les poids des descripteurs, est désormais prêt à l'emploi, même si, il nous reste encore quelques nœuds qui vont servir à l'ajout et à la suppression de certaines colonnes et notamment pour les besoin de la visualisation.

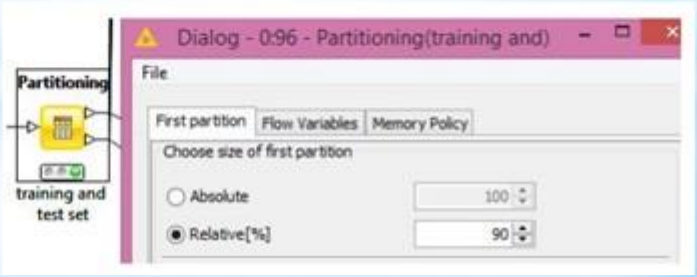


- Ajout de la classe : Maintenant que tout le prétraitement est fait et notre document vecteur est fin prêt, il nous reste qu'introduire la colonne "Documents class" qui correspond à la catégorie de chaque document.

- Nœud couleur : Nous utilisons ce nœud pour des fin de visualisation, ainsi, nous attribuons à chaque catégorie une couleur bien distincte.

- Filtre de colonne : Dans notre cas nous l'avons utilisé pour éliminer la colonne qui comporte notre texte sur lequel nous avons effectué tout le traitement, il ne reste que les descripteurs ainsi que leurs poids et la catégorie correspondante.

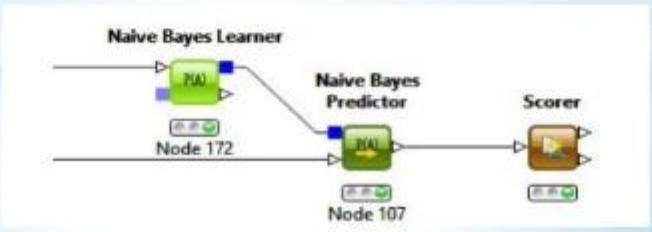
- Partitionnement : C'est la dernière étapes avant l'entrainement des algorithmes. Pour notre cas nous avons préféré partitionner notre ensemble en 90/10, qui signifie que 90% des documents seront dédiés pour l'entrainement et 10% pour le test..



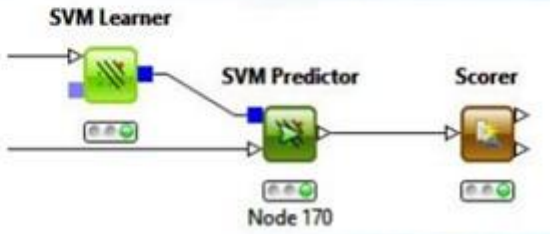
4 Les classificateurs : Une fois le prétraitement, le calcul des pondérations et le partitionnement du document vecteur effectués, nous procédons maintenant à l'entrainement des classificateurs avec l'ensembles d'entrainement, ensuite nous prédisons la classe(catégorie) de chaque document de l'ensemble test. Pour ce faire nous allons utiliser des classificateurs très répandus dans la littérature telque naïf bayes, Machine à vecteurs de support et l'arbre de décision.

4.1 Classificateurs Naïf Bayes : Nous insérons le nœud "Naive Bayes Learner" pour entrainer d'abord le classificateur, mais avant d'exécuter quoi que ce soit nous devons effectuer certains paramétrage.

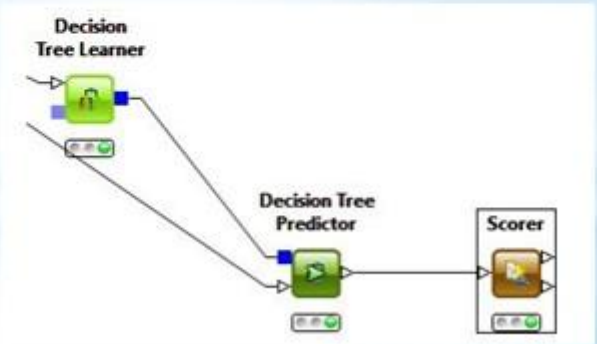
Default probability : C'est la probabilité apriori, que nous ajoutons à chaque descripteur pour éviter le cas où un descripteur n'apparait pas dans un document. le mieux et que cette dernière soit la plus petite valeur possible. Désormais nous pouvons exécuter notre "nœud Naive Bayes Learner". Une fois l'apprentissage du classificateur est effectué avec les données d'apprentissage, nous insérons le nœud "Naive Bayes Predictor" pour tester la fiabilité de ce dernier mais cette fois avec les données test.



4.2 Machine à vecteurs de support (SVM) : Nous procédons de la même manière, décrite précédemment (classificateur de bayes), pour la machine à vecteurs de support. Seulement cette dernière nécessite un peu plus de paramétrage. Overlapping penality : C'est la pénalité de chevauchement, elle est utilisée particulièrement lorsque les donnéesne sont pas linéairement séparables. Nous devons choisir un noyau parmi les trois noyau offert par Knime (Polynomial, HyperTangent, RBF) et leurs paramètres respectifs. par exemple nous choisissons le Polynomial avec : Power : 0,1 ; Bias : 0,1; Gamma :0,1. Nous exécutons et nous passons à "SVM Predictor". La non disponibilité d'une SVM linéaire sur KNIME nous a reconduit à une autre implémentation de cette dernière disponible sur WEKA, puisque KNIME nous permet cette interaction.



4.3 Arbre de décision (AD) : De même pour l'arbre de décision nous effectuons les paramétrages suivants: Quality measure : Gini index (Knime utilise dans ce cas l'algorithme CART).
Une fois le paramétrage effectué nous passons à l'exécution, et voilà que notre classificateur est entrainé, maintenant nous allons juste insérer le nœud "Decision Tree Predictor" pour vérifier l'efficacité de notre classificateur avec les données test. Ce dernier ne nécessite pas de paramétrage particulier.



*Additional Resources

KNIME pages (https://www.knime.com)

- SOLUTIONS for example workflows
- RESOURCES/LEARNING HUB https://www.knime.com/learning-hub
- RESOURCES/NODE GUIDE https://www.knime.com/nodeguide
- Book WILL THEY BLEND https://www.knime.com/knimepress/will-they-blend

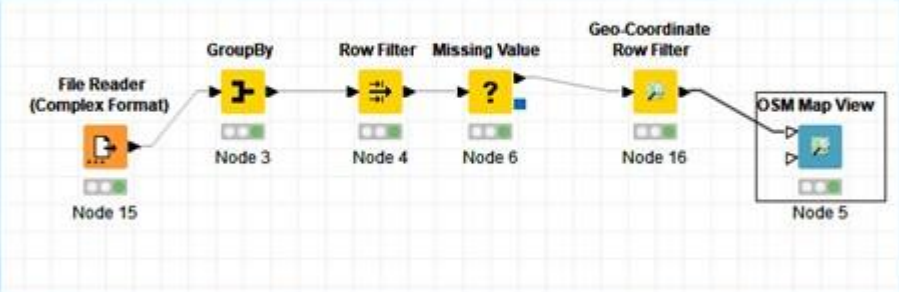
KNIME Tech pages

- FORUM for questions and answers https://forum.knime.com
- DOCUMENTATION for docs, FAQ, changelogs, ...
- COMMUNITY CONTRIBUTIONS for dev instructions and third party nodes

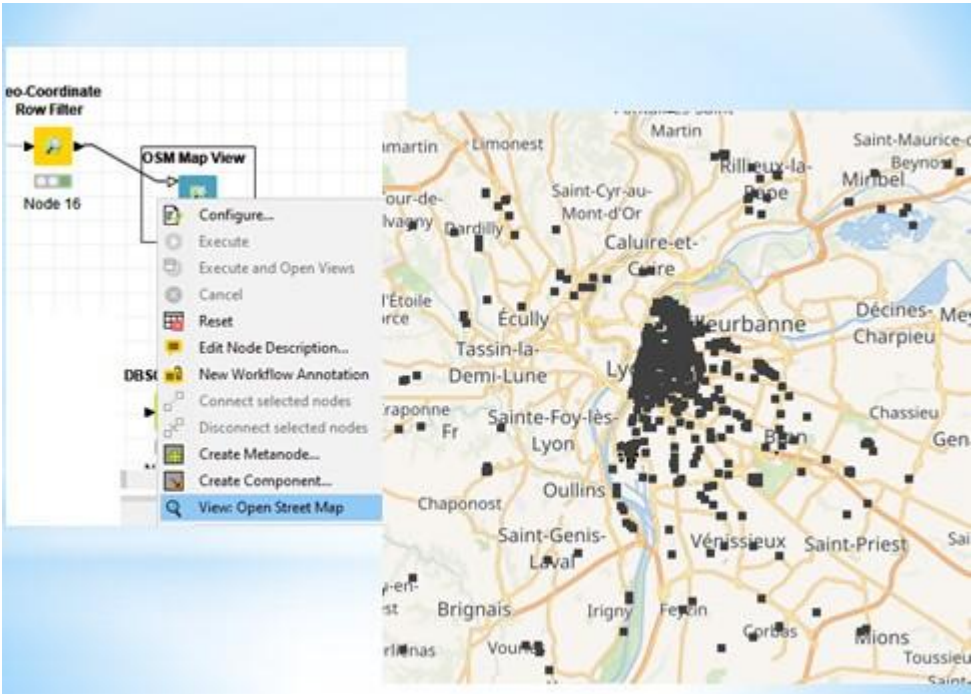
KNIME TV on YouTube https://www.youtube.com/user/KNIMETV

TP1 : Détecter les points d'intérêt des touristes grâces aux photos prises avec GPS

L'objectif de cet exercice est de trouver de maniée automatique des points d'intérêts intéressants dans la ville de Lyon, définis par une activités forte de prise de photos.



Les données sont disponibles sur le lien suivant :
http://liris.cnrs.fr/~mplantev/ENS/DMTP/flickr_data.csv



* TP3 : Démarche de modélisation prédictive sous Knime - Régression Logistique.

* Démarche :

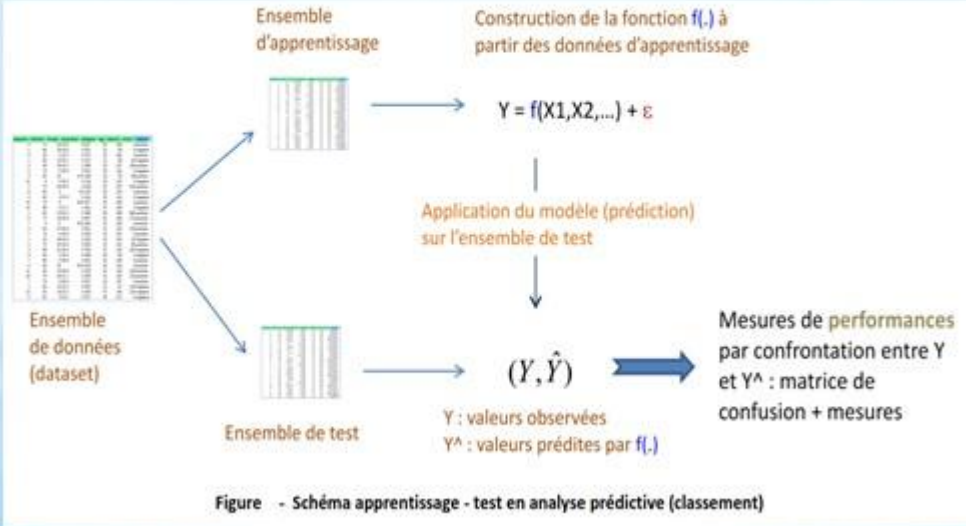


Figure - Schéma apprentissage - test en analyse prédictive (classement)

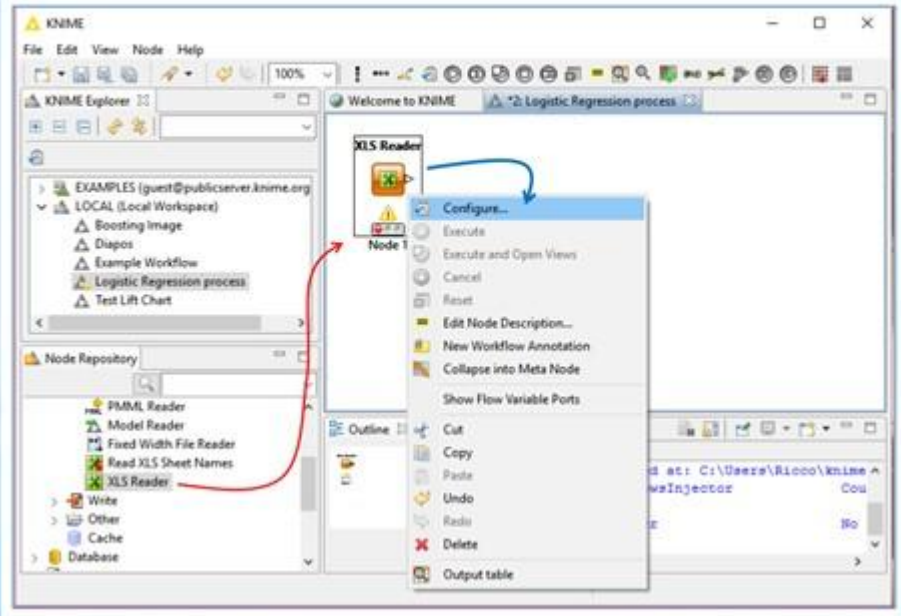
*Données:

Le fichier « Pima Indian Diabetes » décrit un ensemble de personnes amér-indiennes de sexe féminin souffrant ou non du diabète. Il a le mérite d'être parfaitement identifié par les data miners et est relativement propre même si, par ailleurs, certains éléments interrogent. Des individus par exemple présentent un BMI (*body mass index - indice de masse corporelle) nul*.

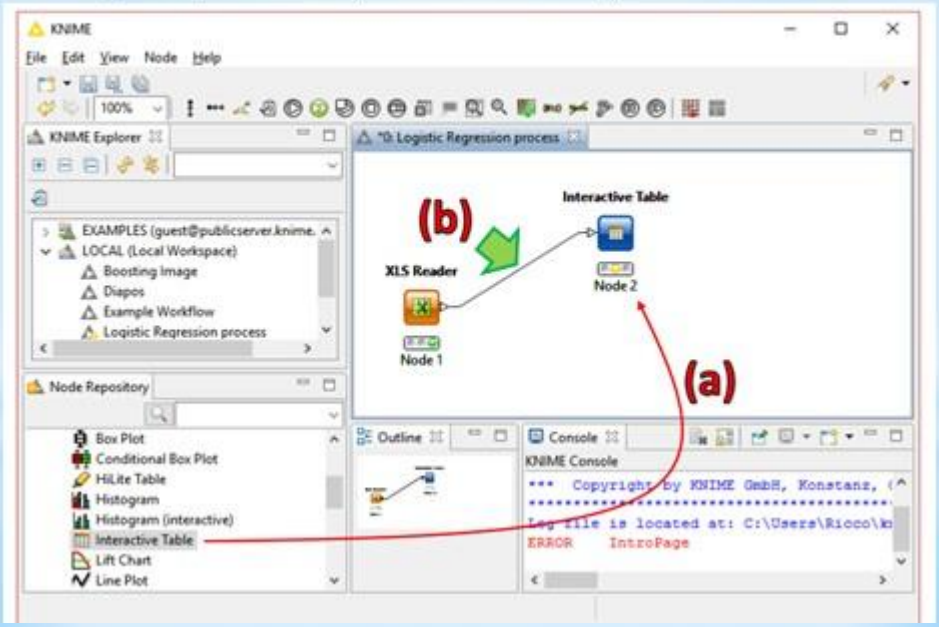
* Nous disposons de 768 observations :

pregnant	diastolic	triceps	bodymass	pedigree	age	plasma	serum	diabete
6	72	35	33.6	0.627	50	148	0	positive
1	66	29	26.6	0.351	31	85	0	negative
8	64	0	23.3	0.672	32	183	0	positive
1	66	23	28.1	0.167	21	89	94	negative
0	40	35	43.1	2.288	33	137	168	positive

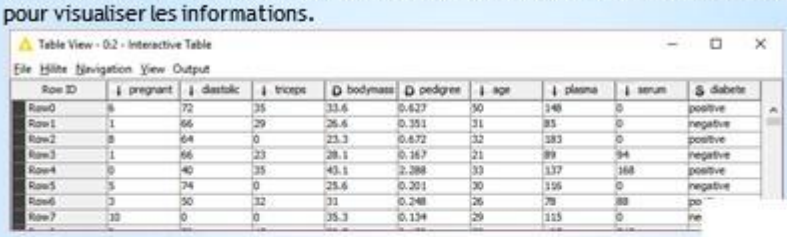
*Apprentissage - Test sous Knime



* Pour visualiser le contenu du jeu de données, nous utilisons le composant INTERACTIVE TABLE. Nous l'insérons dans l'espace de travail (a) et nous lui relions (et non pas l'inverse) l'outil XLS Reader (b)



Il n'y a pas de paramètres à spécifier. Nous cliquons sur le menu EXECUTE and OPEN VIEWS pour visualiser les informations.

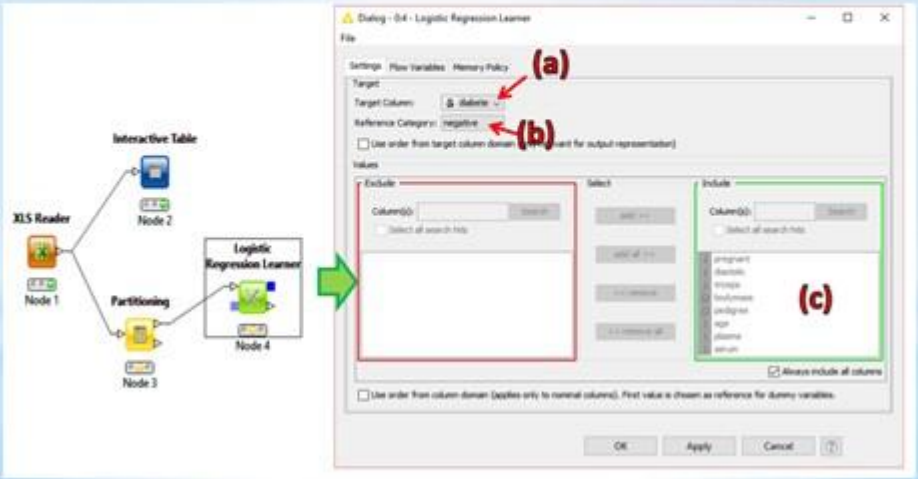


Partitionnement en apprentissage et test

Nous exploitons l'outil PARTITIONING (branche Data Manipulation / Row / Transform dans le Node Repository) pour scinder aléatoirement la base en échantillons d'apprentissage (568 observations) et de test (200). Nous lui connectons la source XLSReader.

*Construction du modèle :

Nous ajoutons le composant LOGISTIC REGRESSION LEARNER (Statistics / Regression) dans le workflow. Nous lui connectons la sortie « training set » de PARTITIONING.



Les coefficients estimés sont affichés dans une nouvelle fenêtre après que l'on ait cliqué sur le menu EXECUTE and OPEN VIEWS. Nous disposons des coefficients estimés, de leurs écarts-type, des z-score pour les tests individuels de significativité, et de la p-value dudit test. La log-vraisemblance du modèle est égale à :

Log-Likelihood = -267.8158.

Logistic Regression Result View - 04 - Logistic Regressi...

File

Statistics on Logistic Regression

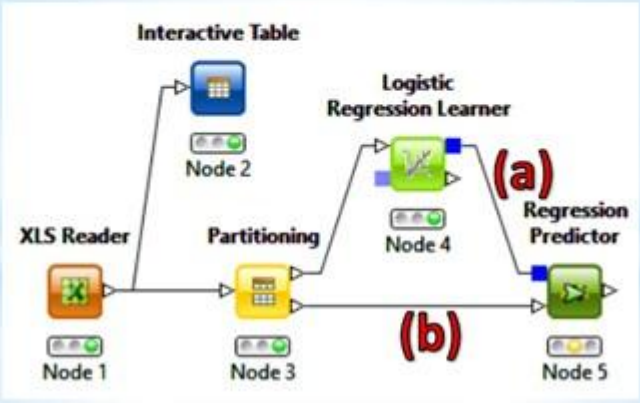
Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
positive	pregnant	0,1345	0,0368	3,6547	0,0003
	diastolic	-0,0096	0,0066	-1,4509	0,1468
	triceps	-0,0006	0,0079	-0,0815	0,935
	bodymass	0,0806	0,0174	4,6387	3,51E-6
	pedigree	1,3438	0,3637	3,6947	0,0002
	age	0,014	0,0106	1,3183	0,1874
	plasma	0,0322	0,0041	7,8723	3,44E-15
	serum	-0,0005	0,001	-0,5398	0,5893
	Constant	-8,2354	0,8271	-9,9567	0.0

Log-likelihood = -267,8158
Number of iterations = 10

Figure - Paramètres estimés de la régression logistique

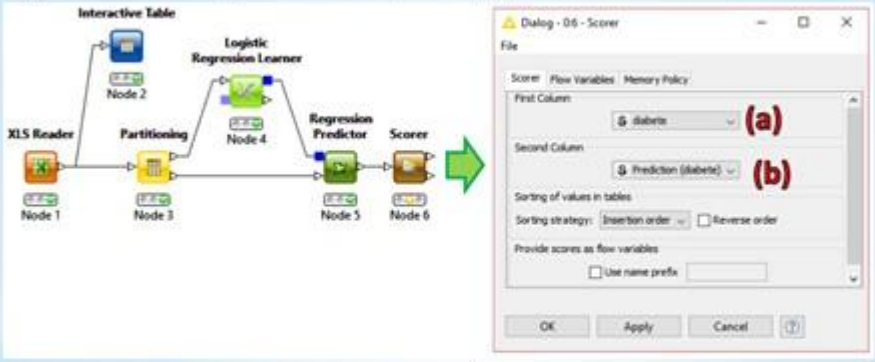
Prédiction sur l'échantillon test :

Nous devons appliquer le modèle sur l'échantillon test pour obtenir les prédictions. Le composant REGRESSION PREDICTOR (Statistics / Regression) prend en entrée (a) le classifieur élaboré par le LEARNER et (b) l'échantillon test issu de PARTITIONING. Nous avons la configuration suivante.

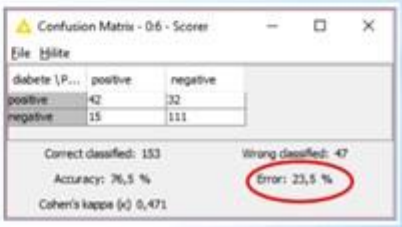


Matrice de confusion et indicateurs :

Le composant SCORER (Mining / Scoring) complète notre dispositif. Il sert à produire la matrice de confusion à partir des valeurs observées et prédites de DIABETE c.-à-d. nous mettons en opposition DIABETE (a) et PREDICTION(DIABETE) (b).



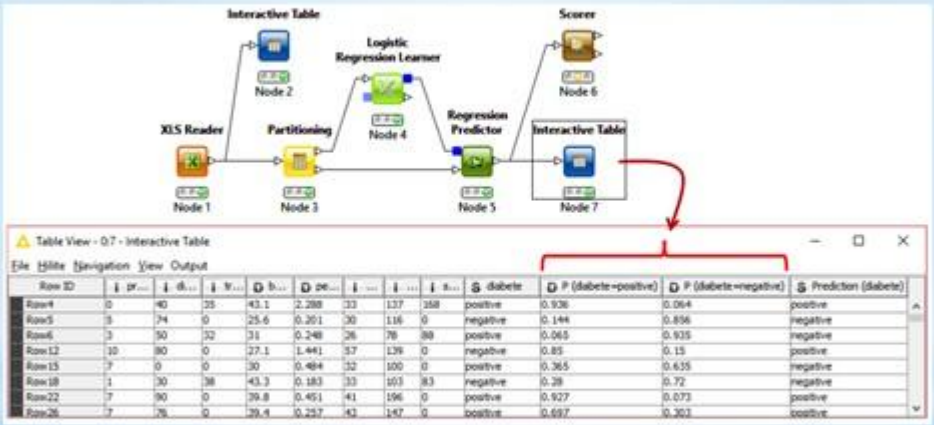
Un clic sur EXECUTE and OPEN VIEWS produit une nouvelle fenêtre de visualisation. Nous distinguons la matrice de confusion. Le taux d'erreur en test est de 23.5%.



Scoring et courbe de gain :

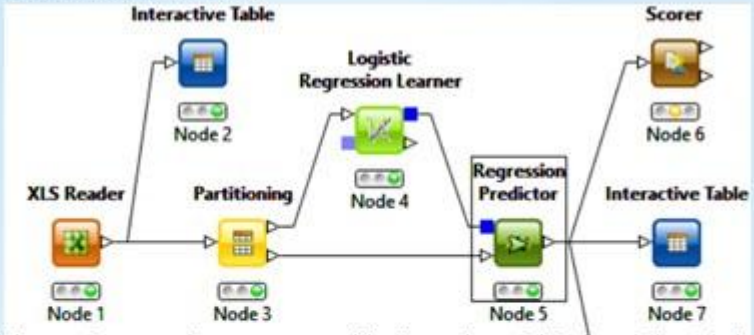
Dans le ciblage (ciblage marketing notamment), l'objectif n'est pas tant de prédire la classe d'appartenance que de quantifier le degré de positivité. A la sortie, nous devons pouvoir annoncer, pour un nombre de personnes sollicitées dans une campagne commerciale par exemple, combien achèteraient réellement le produit si l'on sélectionnait les individus qui présentent les scores les plus élevés. Nous devons modifier le paramétrage de REGRESSION PREDICTOR afin qu'il produise les probabilités d'appartenance aux classes.

Lançons les calculs et ajoutons un INTERACTIVE TABLE pour voir ce que donne tout ça.



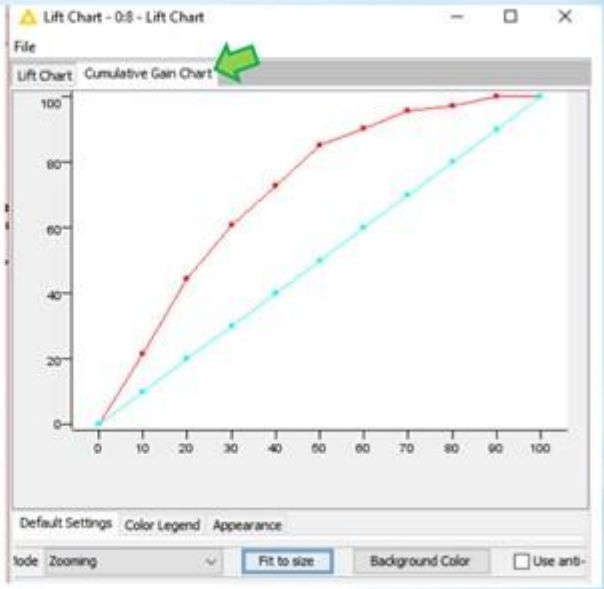
Deux nouvelles colonnes apparaissent [P(diabete = positive) et P(diabete = negative)]. Le premier nous intéresse en particulier puisqu'il correspond au « score » de positivité des individus que nous exploiterons par la suite

Il ne nous reste plus qu'à introduire le composant LIFT CHART (Data Views) et lui connecter le REGRESSION PREDICTOR.



La boîte de dialogue de paramétrage permet d'indiquer la variable à prédire (a), la modalité cible (b) et la colonne de score (c).

Avec EXECUTE and OPEN VIEWS, nous accédons à la fenêtre de visualisation. Nous activons l'onglet « Cumulative Gain Chart ».



*Exposé Emmanuel
L'analyse des sentiments

* L'analyse des sentiments est définie comme le processus d'utilisation de techniques telles que la NLP, les ressources lexicales, la linguistique et ML/DL pour extraire des informations subjectives et liées à l'opinion comme les émotions, l'attitude, l'humeur, modalité, etc. et essayez de les utiliser pour calculer la polarité exprimée par un document texte.
* Le terme « Fouille d'Opinions » est utilisé pour évoquer le traitement automatiques des opinions, des sentiments et de la subjectivité dans les textes.
* Ce domaine est connu sous les noms de *opinion mining*, *sentiment analysis*, ou encore *subjectivity analysis* et est souvent associé à un problème de classification sur des textes évaluatifs.

*Définition des Termes

* **Opinion** : C'est l'expression des sentiments d'une personne envers une entité, (un sujet, un thème, un évènement, etc.), étant subjective et opposée à l'expression de faits.
* **Lexiques** : sont des dictionnaires ou des vocabulaires spécialement construits pour être utilisé pour l'analyse des sentiments et calculer les sentiments sans utiliser des techniques de supervision.
* **Corpus** : est un ensemble de textes, écrits et/ou transcriptions de discours ou de signes, construits dans un but précis et généralement stockés dans une base de donnéesinformatique.
* Un corpus peut être assez petit, par exemple, ne contenant que 50 000 mots de texte, ou très grand, contenant plusieurs millions de mots.

*Définition des Termes

* **Tokenisation/segmentation** : C'est le processus de décomposition ou de rupture de donnéestextuelles en des composants significatifs plus petits appelés jetons.

- Tokenisation de phrases est le processus de décomposition d'un corpus de texte en phrases qui agissent comme le premier niveau de jetons dont est composé le corpus.
- Tokenisation des mots est le processus de décomposition/segmentation des phrases dans les mots qui les constituent.

*Définition des Termes

* **POS tagging** (pour Part-Of-Speech tagging) : déterminer la catégorisation des mots comme étant un verbe, un adjectif, un nom, un adverbe, etc..
* **Lemmatisation** : la récupération de la racine du mot dans le cas de verbe conjugué ou de mots au pluriel, etc...
* **StopWords** : des mots qui n'ont pas d'intérêt au niveau grammatical ou dans le cas de l'opinion mining, non porteur d'opinion. Comme : le, la, à, de, pour, par, elle, il, nous, dans, etc...

*Extraction des Caractéristiques: Bag of Words (Sac de Mots)

* Le modèle du sac de mots est l'une des techniques les plus simples mais les plus puissantes pour extraire des caractéristiques de documents texte.
* Le principe de ce modèle est de convertir des documents textes en vecteurs de telle sorte que chaque document soit converti en un vecteur qui représente la fréquence de tous les mots distincts qui sont présents dans l'espace vectoriel pour ce document spécifique.
* Une chose intéressante est que nous pouvons même créer le même modèle pour l'occurrences des mots individuels ainsi que des occurrences pour les n-grams, ce qui en ferait un modèle de sac de mots n-grams de telle sorte que la fréquence des n-grams distincts dans chaque document soit également pris en considération.

*Extraction des Caractéristiques: TF-IDF

* Mathématiquement, idf peut être représentée par:

$$idf(t) = 1 + \log \frac{C}{1 + df(t)}$$

* Où :
  - idf(t) représente l'idf pour le terme t,
  - C représente le nombre total de documents dans le corpus,
  - et df(t) représente la fréquence du nombre de documents dans lequel le terme t est présent

*Analyse de la subjectivité

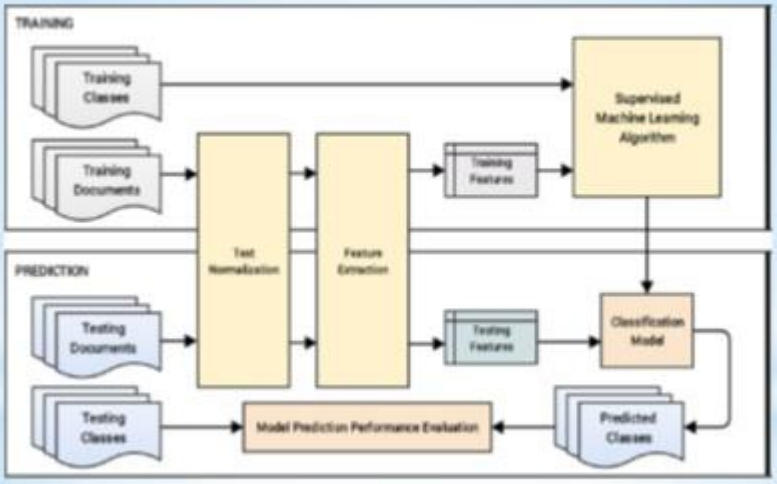
* La subjectivité quantifie la quantité d'opinions personnelles et d'informations factuelles contenues dans le texte.
* La subjectivité plus élevée signifie que le texte contient une opinion personnelle plutôt que des informations factuelles.
* La subjectivité se situe entre [0, 1] où 0.0 est très objectif et 1.0 est très subjectif.

```
In [12]:
1 from textblob import TextBlob
2
3 testimonial = TextBlob("Textblob is amazingly simple to use. What great fun!")
4
5 testimonial.sentiment

Out[12]: Sentiment(polarity=0.39166666666666666, subjectivity=0.4357142857142857)
```

*Les processus de création d'un modèle d'analyse du sentiment

* Préparation des ensembles de donnéesd'entraînement et de test.
* Normalisation du texte.
* Extraction de caractéristiques.
* Entraînement/Apprentissage du modèle.
* Prédiction et évaluation du modèle.
* Déploiement du modèle



*Difficultés de la fouille d'opinions et de l'analyse de sentiments

* difficulté due à l'ambiguïté des mots.
* difficulté due à la structuration de la phrase.
* difficulté due au context.
* difficulté due au vocabulaire qu'on utilise pour exprimer une opinion.
* difficulté due à l'emploi d'une thématique.
* difficulté due au langage qu'utilisent les internautes pour s'exprimer.
* difficulté de déterminer un lexique adapté à l'analyse de l'ensemble des textes d'opinion.

*Exposé Yiboe
Parlons du TAL (Traitement Automatique du Langage)

* Encore appelé NLP pour Natural Language Processing ou Traitement Numérique du Langage, le TAL est une discipline qui porte essentiellement sur la compréhension, la manipulation et la génération du langage naturel par les machines.
* Ainsi, le *NLP* est réellement à l'interface entre la science informatique et la linguistique. Il porte donc sur la capacité de la machine à interagir directement avec l'humain.
* Par langage naturel, on entend le langage utilisé par les humains dans leur communication de tous les jours par opposition aux langages artificiels comme les langages de programmation ou les notations mathématiques.

*Quelles sont les applications du Traitement Automatique du Langage ?

Le NLP est terme assez générique qui recouvre un champ d'application très vaste. Voici les applications les plus populaires :

- Traduction automatique

  Le développement d'algorithmes de traduction automatique a réellement révolutionné la manière dont les textes sont traduits aujourd'hui. Des applications, telles que Google Translator, sont capables de traduire des textes entiers sans aucune intervention humaine.
  Le langage naturel étant par nature ambigu et variable, ces applications ne reposent pas sur un travail de remplacement mot à mot, mais nécessitent une véritable analyse et modélisation de texte, connue sous le nom de Traduction automatique statistique (Statistical Machine Translation en anglais).

*Quelles sont les applications du Traitement Automatique du Langage ?

- Sentiment analysis

Aussi connue sous le nom de « *Opinion Mining* », l'analyse des sentiments consiste à identifier les informations subjectives d'un texte pour extraire l'opinion de l'auteur.

À titre exemple, lorsqu'une marque lance un nouveau produit, elle peut exploiter les commentaires recueillis sur les réseaux sociaux pour identifier le sentiment positif ou négatif globalement partagé par les clients.

De manière générale, l'analyse des sentiments, bien plus efficace que les méthodes classiques comme les sondages, permet de mesurer le niveau de satisfaction des clients vis-à-vis des produits ou services fournis par une entreprise ou un organisme.

En effet, une partie croissante des consommateurs partage aujourd'hui fréquemment leurs opinions sur les réseaux sociaux. Ainsi, la recherche de textes négatifs et l'identification des principales plaintes permettent d'améliorer les produits, d'adapter la publicité et de réduire le niveau d'insatisfaction des clients.

*Quelles sont les applications du Traitement Automatique du Langage ?

- Marketing

Les spécialistes du marketing utilisent également le NLP pour rechercher des personnes étant susceptibles d'effectuer un achat.

Ils s'appuient pour cela sur le comportement des internautes sur les sites, les réseaux sociaux et les requêtes aux moteurs de recherche. C'est grâce à ce type d'analyse que Google génère un profit non négligeable en proposant la bonne publicité aux bons internautes. Chaque fois qu'un visiteur clique sur une annonce, l'annonceur reverse jusqu'à 50 dollars !

De manière plus générale, les méthodes de NLP peuvent être exploitées pour dresser un portrait riche et complet du marché existant, des clients, des problèmes, de la concurrence et du potentiel de croissance des nouveaux produits et services de l'entreprise.

Les sources de données brutes pour cette analyse comprennent les journaux de ventes, les enquêtes et les médias sociaux ...

*Quelles sont les applications du Traitement Automatique du Langage ?

- Chatbots

Les **méthodes NLP** sont au cœur du fonctionnement des Chatbots actuels.

Bien que ces systèmes ne soient pas totalement parfaits, ils peuvent aujourd'hui facilement gérer des tâches standards telles renseigner des clients sur des produits ou services, répondre à leurs questions, etc.

Ils sont utilisés par plusieurs canaux, dont l'internet, les applications et les plateformes de messagerie. L'ouverture de la plateforme **Facebook Messenger** aux chatbots en 2016 a contribué à leur développement.

*Quelles sont les applications du Traitement Automatique du Langage ?

Autres domaines d'application :

- Classification de texte : cela consiste à **attribuer un ensemble de catégories prédéfinies à un texte donné**. Les classificateurs de texte peuvent être utilisés pour organiser, structurer et catégoriser à ensemble de textes.
- Reconnaissance de caractères : Cela permet d'**extraire**, à partir de la reconnaissance des caractères, les **principales informations** des reçus, des factures, des chèques, des **documents** de facturation légaux, etc.
- Correction automatique : la plupart des éditeurs de texte sont aujourd'hui muni d'un correcteur orthographique qui permet de **vérifier si le texte contient des fautes d'orthographe**.
- Résumé automatique : les **méthodes NLP** sont également utilisées pour produire des **résumés courts, précis et fluides** d'un document texte plus long

*Quels types de méthodes sont employés pour traiter automatiquement le langage ?

En Traitement Automatique du Langage, deux méthodes peuvent être employées, soit de façon indépendante ou hybride : les méthodes statistiques et linguistiques.

- TAL - Méthodes statistiques:

  Le principe de ces méthodes est de faire de la langue un objet mathématique. Nous appliquons alors différentes formules probabilistes et statistiques aux données de la langue (mots, phrase, ...). On va par exemple, s'intéresser à la fréquence d'apparition des mots dans un texte pour en déduire les thèmes principaux de celui-ci.

  L'avantage de ces méthodes est qu'elles sont rapides à implémenter et les résultats sont très bons. Le seul inconvénient est qu'elles ne sont efficaces que sur des corpus de données de très grandes tailles.

*Quels types de méthodes sont employés pour traiter automatiquement le langage ?

- **TAL - Méthodes linguistiques :**

  Ces méthodes vont utiliser tous les différents niveaux d'analyse utilisés en linguistique comme la phonétique, la syntaxe ou la sémantique.

  Les méthodes linguistiques ont l'avantage d'être très performantes mais elles sont très coûteuses en temps. En effet, constituer des lexiques ou des grammaires pour décrire de façon la plus exhaustive possible une langue peut prendre des années.

*Principe d'application du TAL

- La phase de prétraitement (du texte aux données)

Supposons que vous vouliez être capable de déterminer si un mail est un spam ou non, uniquement à partir de son contenu. À cette fin, il est indispensable de transformer les donnéesbrutes (le texte du mail) en des donnéesexploitables.

Parmi les principales étapes, on retrouve :

- Nettoyage : Variable selon la source des données, cette phase consiste à réaliser des tâches telles que la **suppression d'urls, d'emoji**, etc.
- Normalisation des données :
  - Tokenisation, (segmentation) ou découpage du texte en plusieurs pièces appelés *tokens*.
    Exemple : « Vous trouverez en pièce jointe le document en question » ; « Vous », « trouverez », « en pièce jointe », « le document », « en question ».

*Principe d'application du TAL

- Stemming : un même mot peut se retrouver sous différentes formes en fonction du genre (masculin féminin), du nombre (singulier, pluriel), la personne (moi, toi, eux…) etc.
  Le **stemming** désigne généralement le processus heuristique brut qui consiste à découper la fin des mots dans afin de ne conserver que la racine du mot.
  Exemple : « trouverez » -> « trouv »
- Lemmatisation : cela consiste à réaliser la même tâche mais en utilisant un vocabulaire et une analyse fine de la construction des mots. La **lemmatisation** permet donc de supprimer uniquement les terminaisons inflexibles et donc à isoler la forme canonique du mot, connue sous le nom de lemme. *Exemple : « trouverez » -> trouvez*

*Principe d'application du TAL

* La phase d'apprentissage (des donnéesau modèle)

De manière globale, on peut distinguer **3 principales approches NLP** : les méthodes basées sur des règles, modèles classiques de Machine Learning et modèles de Deep Learning.

- Méthodes basées sur des règles

  Les méthodes fondées sur des règles reposent en grande partie sur l'**élaboration de règles spécifiques à un domaine** (par exemple, les expressions régulières). Elles peuvent être utilisées pour résoudre des problèmes simples tels que l'extraction de donnéesstructurées à partir de donnéesnon structurées (par exemple, les pages web).

  *Dans le cas de la **détection de spams**, cela pourrait consister à considérer comme **e-mails indésirables**, ceux qui comportent des **buzz words** tels que « promotion », « offre limitée », etc.*

*Principe d'application du TAL

Néanmoins, ces méthodes simples peuvent être rapidement dépassées par la complexité du langage naturel et s'avérer être inefficaces.

- Modèles classiques de Machine Learning

  Les approches classiques d'apprentissage automatique peuvent être utilisées pour résoudre des problèmes plus difficiles. Contrairement aux méthodes fondées sur des règles prédéfinies, elles reposent sur des méthodes qui portent réellement sur la compréhension du langage.

  Elles exploitent les donnéesobtenues à partir des textes bruts prétraités via une des méthodes décrites en haut par exemple. Elles peuvent également utiliser des donnéesrelatives à la longueur des phrases, à l'occurrence de mots spécifiques, etc. Elles mettent généralement en œuvre un modèle statistique d'apprentissage automatique tels que ceux de Naive Bayes, de Régression Logistique, etc.

*Principe d'application du TAL

- Modèles de Deep Learning

  L'utilisation de modèles d'apprentissage en profondeur pour les problématiques NLP fait l'objet de nombreuses recherches actuellement.

  Ces modèles se généralisent encore mieux que les approches classiques d'apprentissage car ils nécessitent une phase de prétraitement du texte moins sophistiquée : les couches de neurones peuvent être perçues comme des extracteurs automatiques de features.

  Cela permet alors de construire des modèles de bout en bout avec peu de prétraitement des données. Aussi, les capacités d'apprentissage des algorithmes de *Deep Learning* sont généralement plus puissantes que celles de *Machine Learning* classique, ce qui permet d'obtenir de meilleurs scores sur différentes tâches complexes de NLP dures telles que la traduction.

*Les outils de TALN

Pour s'initier au traitement automatique du langage naturel, il existe des **outils et aides pratiques en ligne.** Mais quel est l'outil le plus adapté pour vous ? Cela va dépendre de la langue et de la méthode TALN que vous souhaitez utiliser. Parmi les outils open-source les plus connus, on trouve :

- Le Natural Language Toolkit (NLTK) est un ensemble d'outils TALN en langage Python. L'outil propose un accès à plus de 100 corpus de textes, parmi lesquels des textes en anglais, portugais, polonais, néerlandais, catalan et basque. De plus, le kit peut effectuer le traitement de différents textes, comme l'étiquetage morphosyntaxique, l'arbre syntaxique, la segmentation (tokenization en anglais, ce qui constitue souvent la première étape du TALN) et la synthèse de texte. Le kit d'outils TALN comporte également une introduction à la programmation et une documentation détaillée. Il est ainsi bien adapté aux étudiants, doctorants et chercheurs.

*Les outils de TALN

- Stanford NLP Group Software: l'un des groupes de recherche les plus importants dans le domaine du traitement automatique du langage naturel. De nombreux outils sont proposés. Ils permettent de définir la forme de base des mots (segmentation en unité), la fonction des mots (étiquetage morphosyntaxique) et la structure des phrases (arbre syntaxique). De plus, il existe des outils pour les processus compliqués comme le deep learning pour lequel le contexte de la phrase est pris en compte.

  - Stanford Core NLP présente la plupart des fonctions de base. L'ensemble des programmes du Stanford NLP sont écrits en langage Java et sont disponibles en français, anglais, allemand, espagnol et

*Les outils de TALN

- Visualtext est un ensemble d'outils écrits dans un langage de programmation propre au TALN : le langage NLP++. Ce langage de programmation a surtout été développé pour ce que l'on appelle les analystes Deep Text.

  - Visualtext sert surtout à extraire des informations depuis une grande quantité de textes. Il permet par exemple de résumer des textes longs mais aussi de regrouper des évènements sur un thème précis à partir de plusieurs sites Web et de créer un aperçu. Visualtext peut être utilisé gratuitement pour des fins non-commerciales.

*Les outils de TALN

- Egalement de nombreux autres outils (en fonction de l'orientation du traitement dont vous pourrez trouver une liste exhaustive avec des liens d'explications et d'exemples) comme :
  - spaCy, gensim, OpenNLP, scikit-learn pour l'extraction d'entités nommées,
  - word2vec, glove, BERT (de Google), RoBERTa (de Facebook) pour le plongement lexical,
  - Node2vec, PyTorch-BigGraph (PBG), Oracle Labs PGX: Parallel Graph AnalytiX pour les réseaux et graphes d'information.

Exposé Insafe
Le Web Mining

Le Web Mining s'est développé à la fin des années1990. Ce domaine consiste à utiliser l'ensemble des techniques du Data Mining afin de développer des approches et des outils, permettant d'extraire des informations pertinentes à partir d'un large volume de donnéesdu web (documents, traces d'interactions, structure des pages, liens visités, parcours dans le site..)

Basé sur le type de donnéesà étudier dans le domaine du Web Mining, trois grands axes de recherches ont été envisagés portant sur le Web Content Mining (mining using HTML pages), le Web Structure Mining(mining using URL links) et le Web Usage Mining (mining using clicks, visits et logs).

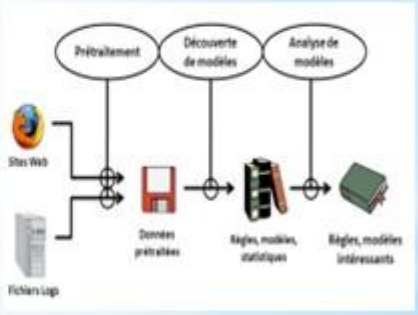
* Processus du Web mining

Le processus du Web Mining se déroule en trois étapes :

1. Collecte des donnéessur l'utilisateur,

2. Utilisation de ces donnéesà des fins de Personnalisation.

3. Présentation à l'utilisateur d'un contenu ciblé.



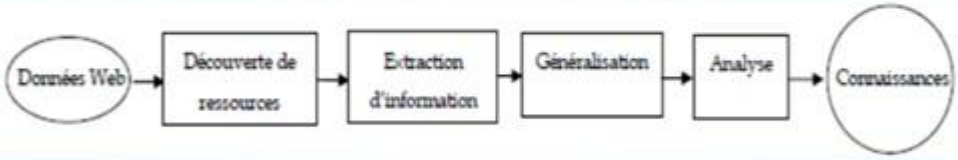
Processusdu Web Mining

* Web Content Mining. Fouille du contenu du web

Le web Contenu Mining est le domaine qui fournit des méthodes permettant **l'extraction, l'organisation, la gestion et la découverte automatique de la quantité énorme d'informations et de ressources disponibles** dans le web.

Le Web Content Mining peut être défini comme le processus d'extraction des informations à partir de différentes sources de donnéesdans le web. Ces sources de donnéespeuvent être **structurées, semi-structurées ou non structurées**

Processus du Web Content Mining



* Les techniques de WCM

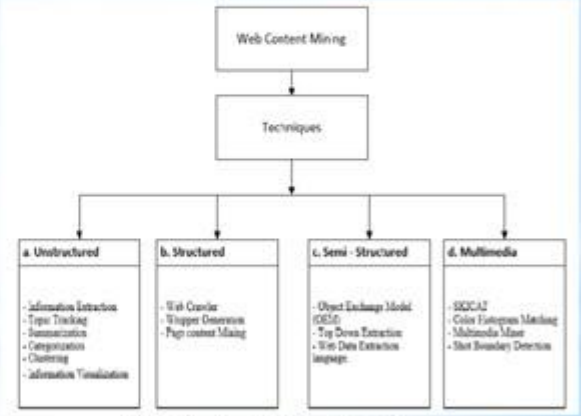
Regroupement

Classification

Identification des associations

Identification des sujets, suivi et analyse de la dérive

Traitement du langage naturel rétieval infromation



Web Content Mining techniques

Le Web Usage Mining (WSM)

- Le Web Usage Mining appelé également Web Log Mining est défini comme étant l'application du processus d'Extraction des Connaissances à partir des Donnéesissues des fichiers Logs http
- se focalise sur l'analyse de la structure des liens entre les pages ou les sites Web, qui constitue une source riche d'information

L'intérêt principal du Web Usage Mining est qu'il fournit des informations sur la manière dont les utilisateurs naviguent réellement sur le site web.

Les méthodes de requêtes WUM

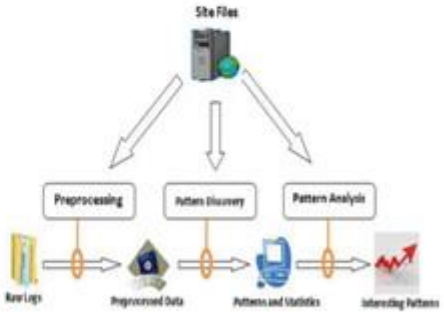
Dans un fichier log, les principales valeurs de types de requêtes sont les méthodes de type : GET, HEAD, PUT, POST, TRACE et OPTIONS

— La méthode *GET* : est une requête d'information. Le serveur traite la demande et renvoie le contenu de l'objet.
— La méthode *HEAD* : est similaire à la méthode GET, cependant le serveur ne retourne que l'en-tête de la ressource demandée sans les données.
— La méthode *PUT* : permet de télécharger un document dont le nom est précisé dans l'URL ou d'effacer un document si le serveur l'autorise.
— La méthode *POST* : est utilisée pour envoyer des donnéesau serveur.
— La méthode *TRACE* : est employée pour le débogage. Le serveur renvoie le contenu qu'il a reçu du client comme réponse. Ceci permet de comprendre ce qui se passe lorsque la requête transite par plusieurs serveurs intermédiaires.
— La méthode *OPTIONS* :permet de demander au serveur les méthodes autorisées pour le document référencé

* Processus du Web Usage Mining

le processus du Web Usage Mining passe par quatre grandes étapes à savoir :

- La collection des données.
- Le pré-traitement des données
- L'extraction d'informations, ou La recherche des motifs intéressants
- L'analyse des motifs extraits.



* Les techniques du WUM

1. Association Rules: C'est la technique la plus utilisée dans WUM. Elle se concentre sur les relations entre les pages Web qui apparaissent fréquemment ensemble dans les sessions des utilisateurs. Les pages consultées ensemble sont toujours regroupées en une seule session serveur

2. Classification: L'objectif principal de La classification dans le WUM est de développer ce genre de profil des utilisateurs/clients qui sont associés à une classe/catégorie particulière.

3. Clustering Il y a principalement 2 types de clusters- le premier est l'Usage cluster et le second est le cluster de pages.

* Les applications du WUM :

TABLE I: TOOLS FOR WEB USAGE MINING CLASSIFIED BY MAJOR APPLICATIONS

Application	Tools
Personalization	WUM – SiteHelper – WebPersonalizerWebSIFT – SETA –Tellim – Oracle9iAS Netmind – Litezia – Web Watcher Krishnapuram – Analog – Accure G2 – Accure Insight 5 – Pilot HitList – SEWeP
System improvement	Rexford – Schechter – AggarwallSpeedTracerWebLogMiner – NetTracker
Business intelligence	SurfAid – Tuzilin – Burchner – SAS Intellivisor ECOMMINER – InterShop – Logisma Business Wbstore
Site design support	Etzioni – Perkowitz – iJADEMiner

183

* Web Structure Mining

Le Web Structure Mining est le domaine qui vise à générer des résumés structurels sur les sites web et les pages web. Il étudie la structure hyperliens du web et caractérise les pages en générant des informations sur la similarité et la relation entre les pages. Ces relations sont examinéesen utilisant l'information véhiculée par les liens hypertextes de chaque page.

184

* Les algorithms d'analyse de WSM

L'analyse de la structure du Web utilise plusieurs algorithmes, dont les plus célèbres sont PagRank (Page et al., 1998) et HITS (Kleinberg, 1999).

PageRank

Développé dans l'université de Stanford par Brin et Page (Page et al., 1998), puis intégré dans le moteur de recherche Google, PageRank est un algorithme d'analyse de structure des liens entre pages très puissant (Zhang et al., 2006).

HITS

L'algorithme HITS (Hyperlink Induced Topic Search) proposé par Kleinberg (Kleinberg, 1999) s'appuie sur un principe simple : toutes les pages n'ont pas la même importance, et ne jouent pas le même rôle. Certaines sont importantes ayant un contenu informatif (authoritative ou de référence), et d'autres sont des pages hubs contenant des liens vers des pages de référence

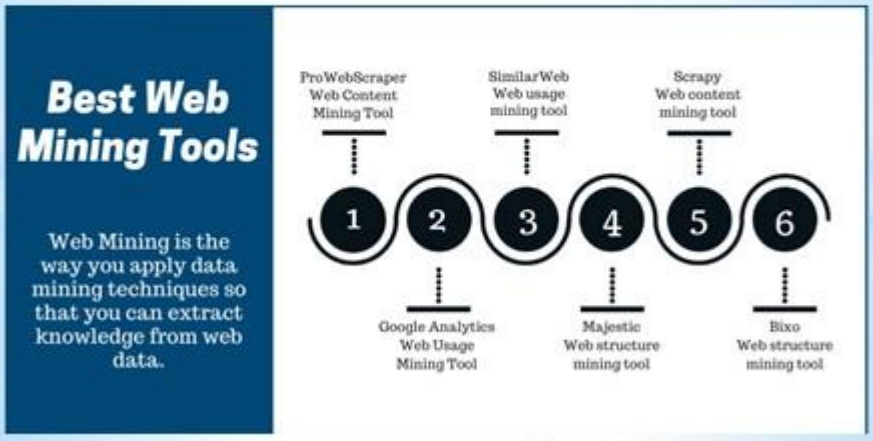
Comparaison des types du WSM

Comparison of Web mining types

	Web content mining		Web structure mining	Web usage mining
	IR view	DB view		
View of data	• Unstructured • Structured	• Semi-structured • Web site as DB	• Link structure	• Interactivity
Main data	• Text documents • Hypertext documents	• Hypertext documents	• Link structure	• Server logs • Browser logs
Representation	• Bag of words, n-gram terms • phrases, concepts or ontology • Relational	• Edge labeled graph • Relational	• Graph	• Relational table • Graph
Method	• Machine learning • Statistical (including NLP)	• Proprietary algorithms • Association rules	• Proprietary algorithms	• Machine learning • Statistical • Association rules
Application categories	• Categorization • Clustering • Finding extract rules • Finding patterns in text	• Finding frequent sub structures • Web site schema discovery	• Categorization • Clustering	• Site construction • Adaptation and management

187

Les meilleures outils de WEB MINING



188

Google Analytics (Web Usage Mining Tool)

* Google Analytics est considéré comme l'un des meilleurs outils d'analyse des activités. Il peut suivre et signaler le site Web. Il peut suivre et signaler le trafic du site Web.

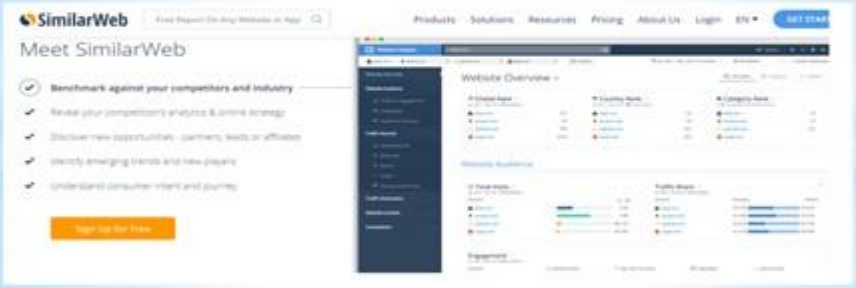
Caracteristiques:

- Analyse du rendement de la publicité et de la campagne
- Analyse et mise à l'essai du site Web
- Intégration facile avec le produit de Google comme, Adsense, Adwords, Google Display Network, Google Tag Manager, etc

189

SimilarWeb (Web usage mining tool)

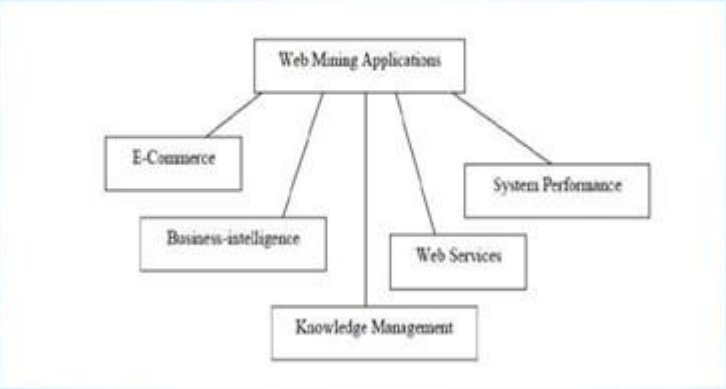
SimilarWeb est un puissant outil de business intelligence. Il offre des renseignements sur le trafic et le marketing pour n'importe quel site Web.



Caractéristiques

* Indicateurs de trafic et d'engagement
* Optimisation des moteurs de recherche et mots clés PPC
* Intérêts du public

*Exemple d'applications



Exposés :

*Web mining + TP

*Fouille d'opinions et analyse des sentiments +TP

*TAL : Traitement Artificiel du langage + TP

DEXPOSES

* Détection de fraudes avec data mining +TP ADANTO Cornelia OK
* Data mining pour une gestion de la relation client réussie +TP Carlin OK
* Data minining et La medcine +TP Anass
* Data mining en ressources humaines +TP MAOULIDA
* Data mining en banque et finance +TP Salif KEITA
* Google Analytics (Web Usage Mining Tool) +TP Kwame Amoako

* Références :

1- C.C. Aggarwal. Data Mining: The Textbook. Springer International Publishing, 2015. • P. Bhatia. Data Mining and Data Warehousing: Principles and Practical Techniques. Cambridge University Press, 2019.

2- M.J.A. Berry and G.S. Linoff. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Wiley, 2004. • M. Bramer. Principles of Data Mining. Undergraduate Topics in Computer Science. Springer London, 2007.

3- M.S. Brown. Data Mining For Dummies. -For dummies. Wiley, 2014.

4- univ-guelma https://dspace.univ-guelma.dz › jspui › bitstream

5-https://moodle.insa-rouen.fr/pluginfile.php/1336/mod_resource/content/4/Parties_1_et_3_DM/IntroDMBeamer.pdf

6- https://cedric.cnam.fr/~saporta/DM.pdf

7-http://bliaudet.free.fr/IMG/pdf/Cours_de_data_mining_3-Modelisation-EPF.pdf

8-https://www.emse.fr/~augusto/enseignement/icm/gis1/UP3-2-Fouille_de_donnees-handout.pdf

8- CLASSIFICATION SUPERVISÉE DE DOCUMENTS ÉTUDE COMPARATIVE, LAHLOU OUCHIHA 2016