M3401 - Introduction à l'**apprentissage automatique**

Etude de cas : Classification Automatique de commentaires de séries télévisées

Plan

- Intelligence artificielle
  - Approche symbolique
  - Approche statistique (ou automatique)
- Apprentissage supervisé
  - Définition(s)
  - Méthode et exemple
- Apprentissage non supervisé
  - Définition(s)
  - Méthode et exemple
- Présentation de l'Etude de cas

Plan

- Intelligence artificielle
  - Approche symbolique
  - Approche statistique
- Apprentissage supervisé
  - Définition(s)
  - Méthode et exemple
- Apprentissage non supervisé
  - Définition(s)
  - Méthode et exemple
- Présentation de l’Etude de cas

Intelligence artificielle

- Ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine
  - Précurseur : Alan Turing en 1950 avec « le test de Turing » → si un juge dialogue avec une machine et un humain à l'aide d'un terminal, et ne peut les différencier l'un de l'autre
  - Pionniers : **John McCarthy** et **Marvin Lee Minsky**
  - Malgré les débats fondamentaux qu'elle suscite, l'intelligence artificielle a produit nombre de réalisations spectaculaires, par exemple dans les domaines de la *reconnaissance des formes* ou de *la voix*, de l'*aide à la décision* (le jeu d'échecs en 1997, le jeu de go en 2016 et le poker en 2017) ou de la *robotique*

Intelligence artificielle

- Approche symbolique
  - Basée sur la logique
  - SE = { faits } U { règles } U {moteur d'inférence}

    (on parle de base de faits et de base de règles)

    règle : SI *condition_1* et/ou *condition_2* et/ou … *condition_n* ***prémisse***

    ALORS *action_1 action_2 … action-p* ***conclusion***

    fait : "valeur"

Intelligence artificielle

exemple :

REGLE r1 : SI animal vole ET animal pond des oeufs
ALORS animal est un oiseau

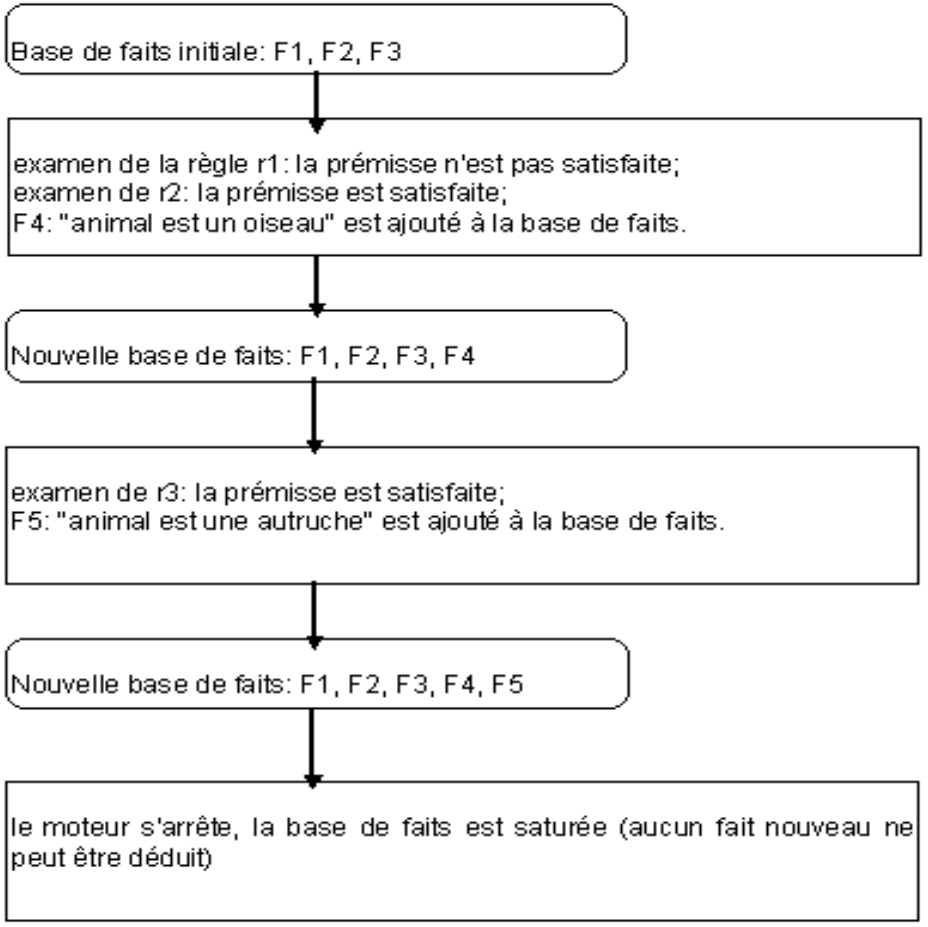
REGLE r2 : SI animal a des plumes
ALORS animal est un oiseau

REGLE r3 : SI animal est un oiseau ET animal a un long cou ET animal a de longues pattes
ALORS animal est une autruche

FAIT F1 : animal a des plumes

FAIT F2 : animal a un long cou

FAIT F3 : animal a de longues pattes

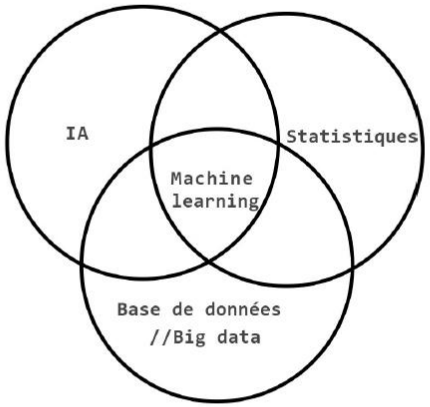


Intelligence artificielle

- Approche symbolique
  - Systèmes Experts
    - DENDRAL (en chimie)
    - MYCIN (en médecine)
    - Hersay II (en compréhension de la parole)
    - Prospector (en géologie)
    - etc.
  - Limites : technologie
    cognition
    très coûteux à mettre en œuvre et à maintenir

Intelligence artificielle

- Approche statistique
  - **Apprentissage automatique** (en anglais ***machine learning***)
    - Evolution des technologies numériques
    - Augmentation exponentielle des volumes de données produites
    - Avancées théoriques



Intelligence artificielle

- Approche statistique

Principe :

  - Trouver des algorithmes qui permettent à un système *d'**adapter** ses analyses et ses comportements*, en se basant sur l'analyse de données empiriques provenant d'une base de données ou de capteurs.
  - La difficulté réside dans le fait que l'ensemble de tous les comportements possibles, compte tenu de toutes les entrées possibles, devient rapidement *trop complexe à décrire* (on parle d'explosion combinatoire).
  - On confie donc à des programmes le soin d'ajuster *un **modèle** pour simplifier cette complexité* et d'utiliser ce modèle de manière opérationnelle.

Intelligence artificielle

- Approche statistique

Principe (wikipedia) : l'apprentissage automatique comporte généralement deux phases.

- La première consiste à **estimer un modèle à partir de données**, appelées observations, qui sont disponibles et en nombre fini, lors de la phase de conception du système. Cette phase dite « d'apprentissage » ou « d'entraînement » est généralement réalisée préalablement à l'utilisation pratique du modèle.
- La seconde phase correspond à **la mise en production** : le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à la tâche souhaitée.

  → En pratique, certains systèmes peuvent poursuivre leur apprentissage une fois en production, pour peu qu'ils aient un moyen d'obtenir un retour sur la qualité des résultats produits.

Intelligence artificielle

- Différents types d'apprentissage statistique :
  - apprentissage supervisé
  - apprentissage non supervisé
  - apprentissage semi-supervisé
  - apprentissage profond (deep learning)
  - apprentissage par renforcement
  - etc.

Intelligence artificielle

- Différents types d'apprentissage statistique :
  - apprentissage supervisé
  - apprentissage non supervisé
  - apprentissage semi-supervisé
  - apprentissage profond (deep learning)
  - apprentissage par renforcement
  - etc.

Plan

- Intelligence artificielle
  - Approche symbolique
  - Approche statistique
- Apprentissage supervisé
  - Définition(s)
  - Méthode et exemple
- Apprentissage non supervisé
  - Définition(s)
  - Méthode et exemple
- Présentation de l'Etude de cas

Apprentissage supervisé

- Prérequis : les ***classes*** sont prédéterminées et les ***exemples connus et étiquetés*** par un expert (ou *oracle*).
- Processus d'apprentissage : générer un **modèle** (dit de **prédiction**) à partir de l'observation des données étiquetées.

- Différents types d'apprentissage supervisé
  - **Prédiction de valeur** (ex : cours de la bourse) → **Régression linéaire**
  - **Classification** (binomiale ou à classes multiples) : prédire la catégorie d'une donnée (ex : affecter l'étiquette chat ou chien à une photo) → **Régression logistique**
  - **Détection des anomalies** : apprendre à quoi ressemble une activité normale et identifier tout ce qui est différent (ex: détection des fraudes)

→ Parfois il est préférable d'associer une donnée non pas à une classe unique, mais une probabilité d'appartenance à chacune des classes prédéterminées (on parle alors d'*apprentissage supervisé probabiliste*).

Apprentissage supervisé

- **Définition informelle** :
  - observations d'un phénomène
  - construction d'un modèle de ce phénomène
  - prévisions et analyse du phénomène grâce au modèle, le tout automatiquement (sans intervention humaine)

- **Définition formelle** :

  Soient Π la population, D l'ensemble des descriptions (caractéristiques), et $\{1,...,c\}$ l'ensemble des classes.

  $X : \Pi \rightarrow D$ est la fonction qui associe une description (caractéristiques) à tout élément de la population.

  $Y : \Pi \rightarrow \{1,...,c\}$ est la fonction de classement qui associe une classe à tout élément de la population.

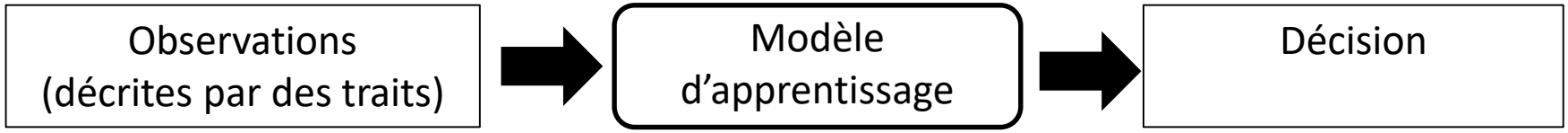
  une fonction $f : D \rightarrow \{1,...,c\}$ sera appelée *fonction de classement* ou *procédure de classification* ou ***modèle de prédiction***.

  Le but de l'apprentissage est alors de rechercher une procédure de classification f telle que

  $f \circ X = Y$ ou, de manière plus réaliste, telle que $f \circ X$ soit une bonne approximation de Y.

Apprentissage supervisé

Modèle d'apprentissage supervisé pour une classification multi-classes (**k** classes) dans un espace de dimension **d**.



$D = \{ Xi=(x_i^1, x_i^2, ..., x_i^d)\}_{i=1,...,n}$

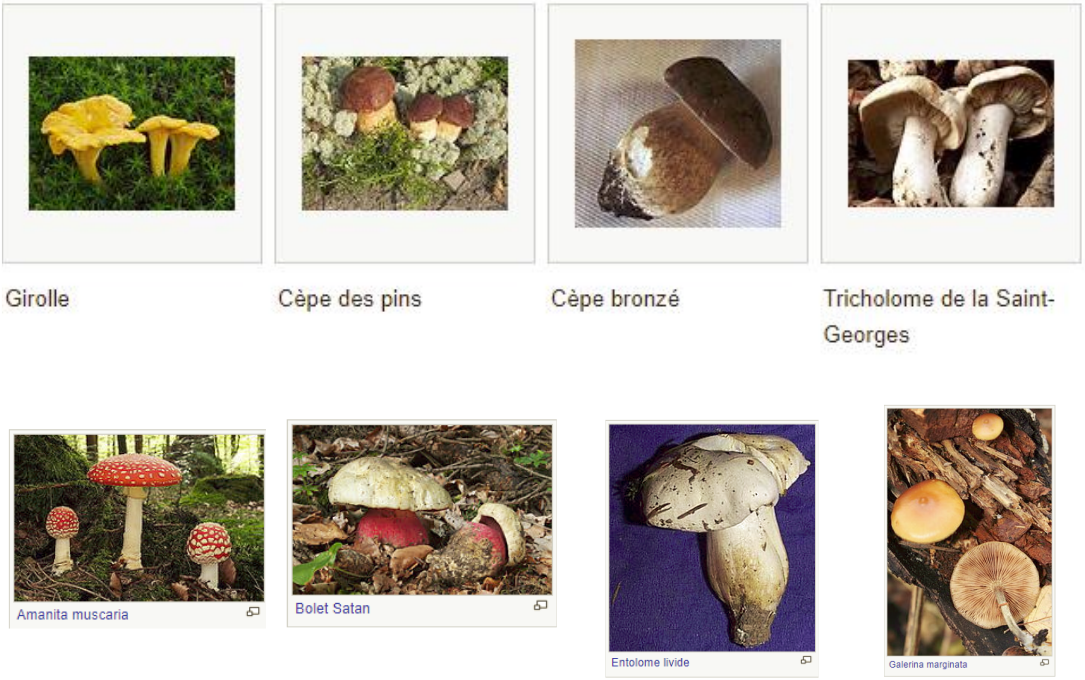
fonction $f(X_i)=y_i$

$\{ y_i, y_i \in \{Y_1, Y_2, ..., Y_k\})$

Apprendre, c'est trouver une fonction *f* telle que $f \circ X \approx Y$

Apprentissage supervisé : un exemple

- Exemple emprunté à Marc Boullé (http://www.vincentlemaire-labs.fr/cours/2.1-ApprentissageSupervise.pdf) : distinguer les champignons comestibles des champignons non comestibles, à partir de leurs caractéristiques.



Apprentissage supervisé : un exemple

→ Etiquetage des données par un expert (ici un pharmacien)

Apprentissage supervisé

- Les **caractéristiques/traits** (**features** en anglais) peuvent être des valeurs numériques, alphanumériques, des images… Quant à la variable prédite Y, elle peut être discrète ou continue.

Apprentissage supervisé : un exemple

→ Définition d'un ensemble de caractéristiques (traits)

- Couleur du pied
- Texture du pied
- Hauteur du pied (de min à max)
- Couleur du chapeau ⁉
- Largeur du chapeau (de min à max)
- Forme du chapeau
- Texture du chapeau
- Couleur des lamelles
- Espacement des lamelles
- etc.

Apprentissage supervisé

- Les caractéristiques/traits d'un exemple sont regroupées dans un vecteur pour former le **vecteur de traits**. Chaque exemple d'apprentissage a son vecteur de trait.

Apprentissage supervisé : un exemple

→ Définition des **vecteurs de traits (trait = caractéristique)**

X_1=(jaune-or; ; 1; 5; jaune-or; 3; 10; convexe-entonnoir; plis ramifiés; ; ; … ; COMESTIBLE)

X_2=(brun clair; ferme; ; ; brun; 5; 20; hémisphère convexe; épais charnu; ; … ; COMESTIBLE)

X_3=(chamois roux; robuste; 1; 3; brun foncé; 10; 20; ; charnu; ; … ; COMESTIBLE)

X_4=(blanc; ; 1; 8; blanc chamois grisâtre; 5; 13; convexe; charnu; blanc; serrées … ; COMESTIBLE)

X_5=(blanc; pelucheux; 1; 30; rouge orangé; 8; 20; globuleux convexe; verrues blanches; blanc jaunâtre; serrées … ; NON-COMESTIBLE)

X_6=(rouge orangé; ; ; blanc grisâtre; 10; 30; hémisphérique convexe; marge lisse; ; … ; NON-COMESTIBLE)

X_7=(blanc grisâtre; ferme; ; blanc grisâtre; 8; 20; convexe déprimé; soyeux; blanc jaunâtre; libres … ; NON-COMESTIBLE)

X_8=(crème ocre; ; ; jaune rouge brun; 2; 5; convexe; lisse brillant; crème; … ; NON-COMESTIBLE)

…

etc.

Sur quelles règles décider si un champignon est comestible ou non → **complexité**

Définir $\boldsymbol{f(X_i) = y_i}$ avec $\boldsymbol{y_i \in}$ **{COMESTIBLE, NON-COMESTIBLE}**

Apprentissage supervisé

- **Plusieurs méthodes** qui dépendent du type des données (texte, image, son, etc.), du choix de f, etc.
  - SVM (Support Vector Machine)
  - Arbres de décision
  - Réseaux bayésiens
  - Régression logistique
  - Réseaux de neurones
  - etc.
- Qui dépendent :
  - de la taille, de la qualité, de la nature des données
  - de l'exploitation du résultat
  - de l'accessibilité à des données annotées
  - du temps dont on dispose
  - du concours d'un ou de plusieurs experts (annotation, validation)
  - etc.

Apprentissage supervisé : un exemple

Choix du modèle :

Y est une **variable discrète** (COMESTIBLE ou NON-COMESTIBLE)

Classification binaire

Modèle de régression logistique

Apprentissage supervisé

- **5 étapes principales pour la mise en œuvre :**
  - étape 1 : récupération des données + étiquetage (annotation) des données
  - étape 2 : construction des vecteurs de traits
    - définition des traits
    - Programmation des fonctions de traits
  - étape 3 : construction des ensembles
    - d'**apprentissage** ou **training set** (2/3 des vecteurs de traits - équilibré). Cet ensemble sert à construire le modèle
    - de **test** ou **test set** (1/3 des vecteurs de traits). Cet ensemble, disjoint de l'ensemble d'apprentissage, sert à évaluer le modèle
  - étape 4 : construction du **modèle de prédiction** (définir la fonction de prédiction) à partir du training set
    - choix d'un algorithme, en fonction de la nature des données, du type de classification (binaire, multi-classes), du type de traits, etc.
    - paramétrage de l'algorithme
  - étape 5 : application du modèle sur le test set et **évaluation** du modèle

Apprentissage supervisé : un exemple

- Retour à notre exemple des champignons : classification binaire (COMESTIBLE, NON-COMESTIBLE)

  → étapes 1 et 2 réalisées

  → étape 3 : si n vecteurs de traits (donc n champignons étiquetés) :

  n_{com} = nombre de champignons comestibles

  $n_{non\text{-}com}$ = nombre de champignons non comestibles

  $n = \min(n_{com}, n_{nom\_com})$

  training set =(2*n)/3 exemples COM + (2*n)/3 exemples NON-COM

  test set = n/3 exemples COM + n/3 exemples NON-COM

  → étape 4 : choix d'un algorithme qui va générer le modèle de prédiction

  régression logistique binaire

Apprentissage supervisé

→ étape 5 : application à l'ensemble de test et évaluation

- Dupliquer l'ensemble de test et remplacer dans l'un, les classes des exemples du test set par « ? » (catégorie que l'algorithme va chercher à définir)

  ex : X_1=(jaune-or; ; 1; 5; jaune-or; 3; 10; convexe-entonnoir; plis ramifiés; ; ; … ; **?**)

- Appliquer le modèle obtenu à l'étape précédente sur cet ensemble de test modifié → classification automatique de ces exemples

  ex : X_1=(jaune-or; ; 1; 5; jaune-or; 3; 10; convexe-entonnoir; plis ramifiés; ; ; … ;
  [COMESTIBLE;0.87887] [NON-COMESTIBLE;0.12113])

- Evaluation en comparant cette classification automatique avec les exemples de l'ensemble de test de départ

  ex : X_1=(jaune-or; ; 1; 5; jaune-or; 3; 10; convexe-entonnoir; plis ramifiés; ; ; … ;
  [COMESTIBLE;0.87887] [NON-COMESTIBLE;0.12113])

  vs.

  X_1=(jaune-or; ; 1; 5; jaune-or; 3; 10; convexe-entonnoir; plis ramifiés; ; ; … ;
  COMESTIBLE)

Apprentissage supervisé

→ étape 5 : application à l'ensemble de test et évaluation

- Termes de l'évaluation :
  - VP (**vrai positif**) = exemple COMESTIBLE classé comme COMESTIBLE
  - VN (**vrai négatif**) = exemple NON-COMESTIBLE classé comme NON-COMESTIBLE
  - FP (**faux positif**) = exemple NON-COMESTIBLE classé comme COMESTIBLE
  - FN (**faux négatif**) = exemple COMESTIBLE classé comme NON-COMESTIBLE
- Différents indicateurs :
  - Précision (bruit) P = #VP/(# VP + #FP)
  - Rappel (silence) R =#VP/(#VP + #FN)
  - F-mesure = (2*P*R)/(P + R)
  - Exactitude = (#VP + #VN) / (#VP + #VN + #FP + #FN)
- Outil = matrice de confusion servant à mesurer la qualité d'un système de classification.

exemple : classification de mails

		Classe estimée	
		normal	**pourriel**
Classe réelle	**normal**	95	5
	pourriel	3	97

Apprentissage supervisé

→ étape 5 : application à l'ensemble de test et évaluation

- Si bons indicateurs alors Modèle ok
- Sinon processus itératif : ajustement des paramètres de l'algorithme comme
  - le nombre d'itérations,
  - l'ajout de traits
  - La suppression de traits (sélection de traits),
  - etc.

  et retour à l'étape 2…

→ au final, production de différents modèles

Plan

- Intelligence artificielle
  - Approche symbolique
  - Approche statistique (ou automatique)
- Apprentissage supervisé
  - Définition(s)
  - Méthode et exemple
- Apprentissage non supervisé
  - Définition(s)
  - Méthode et exemple
- Présentation de l'Etude de cas

Apprentissage non supervisé

- Principe : diviser un groupe hétérogène de données, en sous-groupes de manière que les données considérées comme les plus similaires soient associées au sein d'un groupe homogène et qu'au contraire les données considérées comme différentes se retrouvent dans d'autres groupes distincts
- Mise en œuvre :
  - Récupérations des données **non annotées**
  - Construction des vecteurs de traits (représentation vectorielle numérique = embedding), sans la catégorisation
  - Algorithme de classification basée sur des similarités ou des dissimilarités (distance)
  - Un expert doit ensuite identifier la sémantique des classes…

Apprentissage non supervisé

- **Plusieurs méthodes** dont **Clustering** (partitionnement des données) :
  - K-Means
  - Clustering hiérarchique
  - Clustering par densité

- Qui dépendent :
  - de la taille, de la qualité, de la nature des données
  - de l'exploitation du résultat
  - etc.

Apprentissage non supervisé

ex :

X_1=(jaune-or; ; 1; 5; jaune-or; 3; 10; convexe-entonnoir; plis ramifiés; ; ; …)

X_2=(brun clair; ferme; ; ; brun; 5; 20; hémisphère convexe; épais charnu; ; …)

X_3=(chamois roux; robuste; 1; 3; brun foncé; 10; 20; ; charnu; ; …)

X_4=(blanc; ; 1; 8; blanc chamois grisâtre; 5; 13; convexe; charnu; blanc; serrées …)

X_5=(blanc; pelucheux; 1; 30; rouge orangé; 8; 20; globuleux convexe; verrues blanches; blanc jaunâtre; serrées …)

X_6=(rouge orangé; ; ; blanc grisâtre; 10; 30; hémisphérique convexe; marge lisse; ; …)

X_7=(blanc grisâtre; ferme; ; blanc grisâtre; 8; 20; convexe déprimé; soyeux; blanc jaunâtre; libres …)

X_8=(crème ocre; ; ; jaune rouge brun; 2; 5; convexe; lisse brillant; crème; …)

…

etc.

Apprentissage non supervisé

- Recodage des données (attribuer un code aux couleurs, vectorisation des données (embeddings), etc.)
- Préciser le nombre de classes voulues
- Distances calculées entre vecteurs
  - Distance euclidienne
  - Norme
  - Distance de Levenshtein
  - etc.
- Chaque classe regroupe les vecteurs les plus proches
- Il reste à caractériser la sémantique de la classe

Plan

- Intelligence artificielle
  - Approche symbolique
  - Approche statistique (ou automatique)
- Apprentissage supervisé
  - Définition(s)
  - Méthode et exemple
- Apprentissage non supervisé
  - Définition(s)
  - Méthode et exemple
- Présentation de l’Etude de cas

Présentation de l'étude de cas

- Objectif : détecter automatiquement si un commentaire de série ou de film est positif ou négatif

- Apprentissage sur un corpus de commentaires de séries télévisées issus du site ALLOCINE
  - Apprentissage supervisé en utilisant les évaluations (en nombre d'étoiles) des spectateurs
  - Apprentissage non supervisé

- Corpus : commentaires en langue française sur des séries télévisées
  (~1000 séries), issus de Allo Ciné
  (http://www.allocine.fr/series/ficheserie-17907/critiques/presse/#pressreview40022939)

Présentation de l’étude de cas

- Exemple de commentaires ALLOCINE :

Ti Nou

LeClub300

Suivre son activité
173 abonnés
Lire ses 2 673 critiques

Critique de la série

5,0 Publiée le 12 janvier 2020

Au fil des années, "Game of thrones" se sera imposé comme un monument grâce à un art extrêmement habile du rebondissement, bousculant un genre très codé. Une série donnant plus de place à son intrigue et ses personnages qu’à son univers fantaisiste, elle parvient pourtant au fil des saisons à assurer un spectacle grandiose pour la télévision. Elle adapte habilement l’œuvre de George R. R. Martin en sachant s’en détourner pour fluidifier la narration.

5 0

J.Dredd59

Suivre son activité
34 abonnés
Lire ses 675 critiques

Critique de la série

1,5 Publiée le 23 avril 2019

Ouais j'adore me faire détester, mais ce n'est pas pour ça que je mets une note si basse à cette série. Par contre, c'est en partie pour son succès immérité, reposant en majeure partie sur la décadence de notre société, que je la descends. Sérieux, même Emilia Clarke l'avoue : "la plupart des scènes de sexe dans les films ou à la télévision sont gratuites et généralement là juste pour attirer du public", et à quoi avons nous droit à chaque saison ? A part ça ? Des guerres, des nipple count, des complots, un hiver qui ne vient pas, des trahisons, des zombies, encore du sexe, des morts vu de près... pire que le JT, mais c'est révélateur. La seule chose appréciable est qu'aucun épisode ne se ressemble et qu'on ne sait pas à quoi s'attendre, pour aucun personnage. Pas mal mais trop peu pour faire une réussite que je dégomme car trop de gogos l'encensent (et l'auront oublié 6 ans après la fin de diffusion).

7 5

Présentation de l'étude de cas

- **Étape 1** : **Constitution du corpus**
    - Aspiration de pages web à partir du site ALLOCINE
    - Extraction des commentaires et des évaluations
- **Étape 2** : **Génération des vecteurs de trait**
    - Définition des traits
    - Programmation des fonctions de trait
    - Génération des vecteurs de trait
- **Étape 3** : **Génération de modèles prédictifs**
    - Construction des ensembles d'apprentissage et d'évaluation
    - Classification binaire apprentissage supervisé et évaluation du modèle
    - Classification binaire apprentissage non supervisé et évaluation du modèle

Présentation « Etude de cas »

Apprentissage supervisé

lundi am (MK)	Présentation de l'apprentissage statistique
	Présentation de l'étude de cas
lundi am + pm (IU)	Aspiration de pages web ALLOCINE
	parcours d'une arborescence
lundi pm + mardi am (PDS)	Parser XML
	Extraction des commentaires et des évaluations (format XML) à partir des pages aspirées
mardi pm (MK)	Parser XML en python
	Définition des traits
mercredi am + pm (MK)	Programmation des fonctions de traits Constitution des ensembles d'apprentissage et de test
jeudi am + pm (SVP)	Apprentissage supervisé : Modèle de régression logistique
	Evaluation du modèle
vendredi am (SVP)	Apprentissage non supervisé : Modèle K-MEANS
	Evaluation du modèle
vendredi pm (SVP + MK)	Discussion
	Evaluations de fin de module