Chapitre 1

Chapitre 1 : Média et actualité

1.1 Introduction

Dans toute activité de recherche ou de développement, la maitrise des concepts du domaine étudié est une étape primordiale avant toute démarche de conception ou de développement. Dans ce sens, ce premier chapitre sera consacré à la présentation de notre domaine d'études ainsi que tous les concepts qui y sont en relation. Il sera donc principalement question de présenter le domaine des média en général et de l'actualité (ou des news) en particulier.

1.2 Les médias

1.2.1 Définition

Depuis l'antiquité,les médias sont révélés comme l'un des outils les plus importants de transmission des informations et de l'actualité vers un publique de masse .

L'évolution de ce moyen est passé de plusieurs étapes.Au début le partage de l'information s'effectuait avec les peintures et des écritures.Mais maintenant, la création et l'invention des nouveaux modes et techniques de diffusion des informations a révolutionné le domaine du partage et la communication des messages avec les auditeurs, pour le but de les sensibiliser et affecter leurs perceptions.

On trouve cette définition du terme média dans le dictionnaire LAROUSSE : « Le terme média désigne tout moyen de distribution, de diffusion ou de communication, des œuvres, de documents, ou de messages écrits, visuels, sonores ou audiovisuels (comme la radio, la télévision, le cinéma, Internet, la presse, les télécommunications, etc ».[1]

1.2.2 Types des Médias

FIGURE 1.1 – Les types des médias [12]

Avant l'apparition des moyens technologiques les plus répandus de maintenant tel que l'internet et la télévision .La presse écrite était le seul moyen d'interaction et de partage de l'information avec l'audience.

Mais Aujourd'hui , après le développement extraordinaire de l'internet ,on peut maintenant avec un simple clic de surfer dans des sites WEB et de trouver touts types d'information et d'actualité en $ReelTime$ .Par exemple (les flashs news, les conférences de presse , les résultats et les statistiques des matchs de football en direct ex...) [1]

Les médias peuvent être classés en trois grandes familles :

1. **Les Médias écrites (imprimés)**

   1.1. **Journaux (NewsPapers)** – Imprimé qui contient un ensemble d'article et de publication ,présente l'actualité liées à touts les domaines ,publié sur une base quotidienne ou hebdomadaires ,c'est l'un des moyens les plus importants de PrintMedia .

   1.2. **Magazine** – Imprimé sur une base trimestrielle ou annuelle, elle contient des informations sur l'économie , la finance, la nourriture, la mode, etc [2]

2. **Les Médias de diffusion (ou de Broadcasting)**

   Appelé aussi un média de l'offre ,c'est un ensemble de moyens les plus répandus de divertissement et de communication des informations avec le grand publique dans ces dernières années. elle signifie le partage des contenus audio ou audiovisuels .

2.1. **Télévision** – Dans le passé, il y avait quelques chaînes qui partageaient des contenus générales , comme les chaines terrestres . Mais avec l'apparition des chaines satellites ,on a aperçu l'augmentation et diversification des types de contenu que chaque chaîne propose .On voit donc des chaines se spécialises dans l'actualités, les films, les sports, la nature etc.

C'est le média de diffusion le plus important en raison de sa portée auprès du publique dans le monde car au moins 1,67 milliards de foyers disposaient d'au moins un téléviseur en 2019 [5].

2.2. **Film** – Les films, les scénarios, les images animées ont une accessibilité mondiale. C'est le meilleur type de médias de masse pour envoyer des messages ciblés et influencer les grandes audiences. L'industrie du cinéma joue un rôle très important dans la diffusion de la conscience sociale.[2]

3. **Internet**
L'outil Internet a évolué pour devenir un média de la demande, c-à-d l'utilisateur qui choisit le type de contenu et d'actualités que souhaite voir et consulter .Par contre les médias de l'offre comme le Broadcast média ,ils offrent le type de contenu .Ce qui a permis l'augmentation de l'utilisation des moteurs de recherches[3]

3.1. **Réseaux sociaux ou sites Web** – incluant Facebook, Instagram, Twitter, YouTube, Tumblr, LinkedIn, Snapchat, Quora, Reddit, Pinterest, etc.

Ils sont largement utilisés par les gens du monde entier ,donc on peut trouver des actualités exclusifs dans un temps réel.

Mais ce moyen contient des inconvénients tel que le partage des fausses rumeurs, aussi des $FakeNews$ qui peuvent aidés à déstabiliser toute une société.[2]

3.2. **Les Forums** – Un endroit en ligne où se groupe un ensemble de gens pour le but de poster des publications ,commenter ,envoyer des messages ou discuter d'un sujet particulier. Les forums nous permettent de partager nos connaissances avec d'autres personnes ayant le même intérêt .[2]

1.2.3 Types de contenus médiatiques

1. **Actualités :** Ce sont les informations et news sur ce qui se passe dans le monde ,on trouve plus de détaille dans la section (1.3 Actualité) .[7]

2. **Événement courants :**Couverture d'un grand événement Comme (Tournoi Roland Garos) du sport de Tennis .

3. **Divertissements :**Des histoires sur la musique , l'industrie du cinéma ou encore le théâtre ,et si ce que cela vaut la peine de voir ou non .

4. **Rapports détaillés :** Par exemple (documentaires, un reportage d'investigation sur une catastrophe naturelle .)

5. **Opinions :** Article qui présente l'opinion d'une personne de façon subjective .Ce type permet de partager un avis d'un expert sur un cas ,par exemple "un politicien donne sa vision sur qui va gagner les élections ?" [4]

1.2.4 Les plus grands propriétaires de médias au monde

Position	Nom	Revenu	Description
1	Alphabet	$59.62 milliard	C'est la compagnie qui possède Google ,elle domine ce classement avec une large distance .La plupart de ses revenus vient des services de publicité comme AdSence
2	The Walt Disney Company	$22.45 milliard	C'est l'une les plus grandes compagnie de Cinéma et de création des dessin animes ,la plupart de ses revenus vient de la publicité
3	Comcast	$19.72 milliard	88 % de ses revenues vient de la diffusion des programmes de télévision
4	21st Century Fox	$18.67 milliard	C'est une grande entreprise de divertissement dans le monde ,elle est possédé par la compagnie Fox
5	Facebook	$11.49 milliard	La migration de la population vers l'internet et l'apparition des smart-phones a permis à Facebook d'être l'un des média les plus important dans dernières années

TABLE 1.1 – Classement des organisations selon une étude du site businessinside.fr en 2016[13]

1.3 L'actualité :

1.3.1 Définition

Une actualité ou une actu ,c'est un message ou une information d'un événement récent qui s'est passé dans la journée ou la semaine, délivré par les médias vers le grand publique .[6]

Dans ce moment actuelle ,il se passe un grand mélange d'événement dans tout les domaines de la vie ,ce qui causera un chaos d'information .Heureusement ,les journalistes essayent de structurer ce chaos de sorte que chaque jour ,le publique reçoit les nouvelles bien triées et soignées .Les nouvelles seront publiées dans les médias ,tel que la télévision et les journaux .[7]

On trouve cette définition du terme actualité dans le dictionnaire LAROUSSE : « Événements **actuels** intéressant un domaine d'activité ».

1.3.2 Critères d'actualité

Il existe une priorité d'affichage des informations dans les médias .Les nouvelles importantes qui feront le buzz seront affichées en premier et avec les détailles ,les nouvelles moins importantes sera affiché en dernier et avec moins de détaille .[7]

La distinction de l'importance des informations se fait par le jugement des journalistes ,en se basant sur le niveau d'intérêt de la société et l'importance relative de l'événement .[7]

1. **La nouveauté de l'actualité :**
   Le facteur de temps est très important pour déterminer l'importance d'une nouvelle ,chaque heure passe ,l'information ne sera pas nouvelle ,donc le publique ne le donnera pas le niveau d'intérêt qu'elle le mérite .[7]

2. **L'irrégularité (Inhabituelle) :**
   Ce sont des nouvelles qui ne se produisent pas souvent ou se passe rarement .Cependant ,un événement habituelle varie d'une société à une autre ,par exemple ,dans notre société : "Un chien mord l'homme" n'est pas une nouvelle ,mais "L'homme mord le chien" est une nouvelle .

   Mais dans des sociétés comme en Chine ,ce n'est pas une nouvelle car c'est habituelle de manger des chiens dans leur société ,donc cette information ne fera pas l'actualité .[7]

3. **Contenu intéressant :**
   Par exemple une information comme "Les scientifiques ont trouvé de l'eau dans mars ou dans une autre planète " est une information nouvelle et inhabituelle ,mais le problème elle ne trouvera pas le centre d'intérêt dans publique.

   Par contre une information comme "La chute de 70% des prix de pétrole" est une information qui fera le *buzz* car elle touchera directement la situation financière de la population d'un pays pétrolier .[7]

4. **L'exclusivité :**
   Chaque journal ou chaine TV ont des sources dans chaque grande entreprise dans le monde ,ces sources leur donnent des informations exclusifs .par exemple "Le président va démissionner la prochaine semaine ".

   Ce type de nouvelle va beaucoup attirer l'attention du publique de masse .Donc elle fera la une dans les journaux ou dans les chaines d'information .

5. **La proximité :**
   La place ou se passe l'événement est important pour les lecteurs ou les auditeurs ,par exemple "Le président des USA a fraudé les élections pour gagner la présidence" fera un bruit énorme dans le monde et l'information se propagera très rapidement . Mais si un petit pays de troisième tiers du monde ne fera pas la une dans les grands média du monde [7] .

Ceci étaient quelques critères principaux qui vont définir si l'information est de bonne qualité ou non ,et décider si cet article sera dans les premières page ou dans les dernières.

1.3.3 Les domaines d'actualité

Ceci sont les domaines les plus fréquents que le publique attende des nouvelles .

1. **Conflits :**
   Les guerres , les luttes militaires et les coups d'états feront toujours un sujet important des médias , par exemple l'invasion *Nato* en Libye l'année 2011 .[7]

2. **Politique :**
   Reporté les élections et les luttes pour acquérir le pouvoir politique est l'une des topics les plus importants couvertes par les médias . [7]

3. **Économie :**
   Comporte des informations sur les crises économique dans le monde aussi le prix des matières premiers dans les bourse mondiales .Ce domaine est très important ,il attire l'attention des auditeur parce-que ces crises toucheront directement leur mode du vie .Par exemple la crise économique mondiale de 2008 et la crise du pétrole en 2013 ont fait la une dans ces périodes .[7]

4. **Catastrophe et tragédie :**
   Comporte deux types : .

   - Catastrophe humaine :Comme l'accident d'usine de *Chernobyl* en 1986
   - Catastrophe naturelle :comme le Séisme de Haïti en 2010.[7]

5. **Crime :**
   Comporte tout type de crime comme le meurtre et le viol .Généralement les meurtres inhabituelles feront la Une dans les médias .

   Aussi la couverture des procès des grandes personnalités politiques et économiques fera la Une dans les médias .[7]

6. **Personnes célèbres :**

   Les célébrités sont des influenceurs ,ils attirent l'attention du publique lorsque par exemple ils visitent des lieux ou des pays des autres continent, faire les dons ,ou aussi leur implication dans les scandales . Alors reporter ces informations augmentera le nombre de vus de ces articles .[7]

7. **Santé :**
   Reporté l'état et les statistiques du propagation du coronaVirus dans le monde ,et aussi reporté les conférences de presse, les déclaration et décision des chefs d'état pour combattre ce virus pour un seul but "sensibiliser" les gens. [7]

8. **Sport :**
   Les médias spécialisées en sport essayent toujours de couvrir les sports populaires comme le football en Europe ou le basketball en USA . Pendent la période estivale des transferts de joueur ,les médias exploitent les sentiments des lecteurs envers les clubs qu'ils aiment ,en partagent beaucoup de rumeurs pour attirer leur attentions .Par exemple le titre "Ryad Mahrez rejoint Manchester City pour 60m euro" a fait la Une des presses en 2017 .[7]

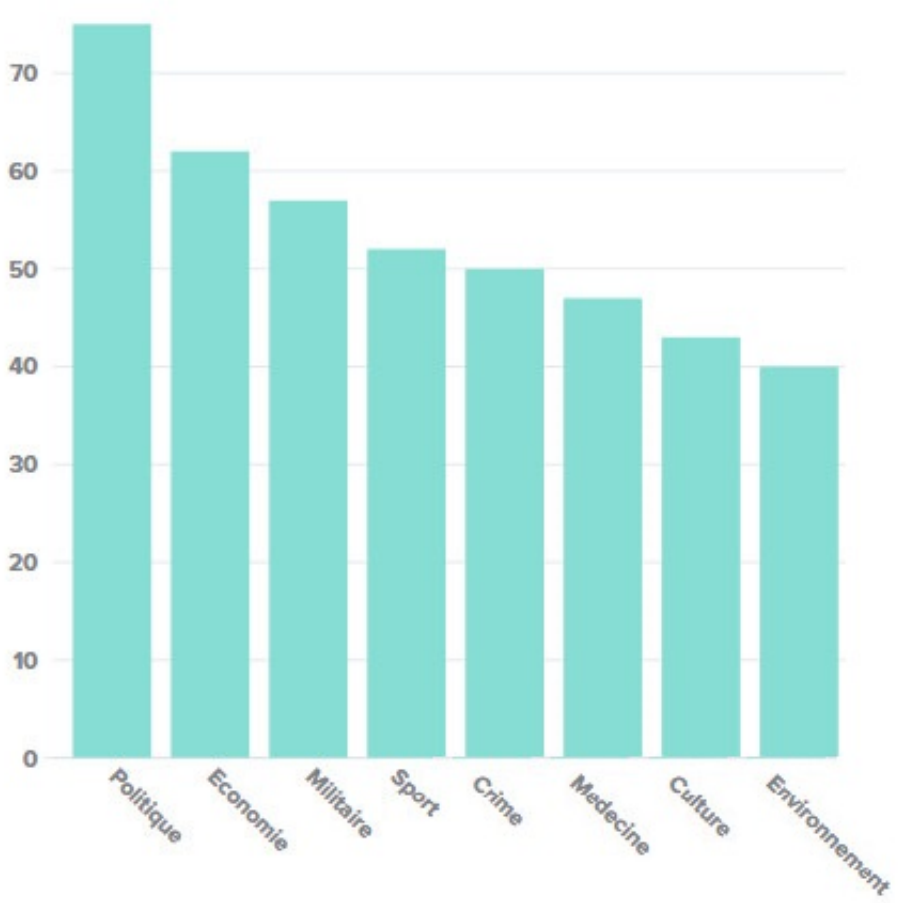

FIGURE 1.2 – Les topics le plus consulté par le publique de masse [14]

1.3.4 Diffusion des actualités

Récemment, il y a eu un déplacement de la consommation de l'actualité par le publique vers Internet Ce qui a baissé le nombre de vue des informations diffusées traditionnelles par l'audience .Le journalisme imprimé est donc menacé .

Une partie de la migration de l'audience a été vers les journaux en ligne associés à la presse . En outre, il existe des systèmes tels que les services de recherche de nouvelles, les moteurs de recherche et les aggregateurs de nouvelles ,qui permette d'accéder à un ensemble de sources d'information . Ce dernier ne constituent cependant pas une menace pour les organisations de médias existantes, tant que ces organisations s'adaptent a Internet pour suivre leur publique.

Les diffuseurs traditionnels tentent de se faire concurrence en fournissant des informations de dernière minute sur des développements importants, ceci mène que les nouvelles du jour ne sont plusieurs au moment des derniers tirages mais varient continuellement .[14]

1.3.5 L'intérêt publique sur les news

Pour analyser les intérêts publique sur les actualités . Plusieurs techniques ont été crées et déployées .Par exemple, plusieurs médias utilisent le calcule des nombre de clics dans les sites d'actualités pour calculer le nombre de vue pour chaque article. Aussi l'utilisation des techniques de l'intelligence artificielle pour analyser les avis et les commentaires des auditeurs pour détecter leur réaction sur un événement .

En plus de cette auto-surveillance , Il existe des questionnaires et des sondages pour le publique ,afin de détecter les sujets qui font la tendance .[14]

1.3.6 Classement des médias de l'actualité les plus consulté au monde

Pays et régions	Moy Durée de la visite	Pages/Visite	Position
yahoo.com	00 :07 :56	7.13	1
naver.com	00 :17 :54	12.08	2
qq.com	00 :05 :21	4.38	3
msn.com	00 :07 :54	4.72	4
globo.com	00 :06 :46	3.18	5

TABLE 1.2 – Classement des sites les plus populaires selon le trafic des visiteurs, de *similarweb.com*[12]

1.4 Conclusion

Au niveau de ce chapitre, nous avons donné une description bien détaillée sur notre domaine d'études, en présentant le domaine des médias, l'actualité et ce qui leur concerne.

On a vu que quotidiennement, plusieurs agences de médias publient des milliers d'articles contenant toutes sortes d'événement comme (politique, économique, santé, etc.). Ces articles seront consultés par des milliers d'auditeur dans le monde . On a constaté que plusieurs critères déterminent si ces articles sont de qualité, afin d'attirer l'attention du publique .

Le prochain chapitre sera consacrer pour présenter les concepts et techniques de base tel que l'annotation et classification d'un texte brute .

Chapitre 2

Annotation et classification

2.1 Introduction

Allons de La classification qui fait référence au développement, à l'analyse et à l'implémentation de méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible d'exécuter par des moyens algorithmiques classiques ,il sert à extraire et exploiter automatiquement l'information présente dans un jeu de données.

Avec ce principe, nous essayons dans ce chapitre de présenter les différents approches de classification, son objectif, citer quelques domaines d'application, ainsi quelques algorithmes de classification en détail.

2.2 Annotation

2.2.1 Définition :

Souvent , quand on veut analyser des donnés, on les trouve généralement sous forme numérique comme par exemple les donnés des ventes ,les mesures physiques (comme la masse ,le poids ...) et les catégories quantifiées .Les machines peuvent facilement analyser les données numériques , mais la questions qui se pose ,comment analysé les données de type "texte" ?

Pour cela ,afin de bien analyser un texte,il faut préparer les données de sorte que la machine trouve facilement les patterns et les inférence.

L'annotation C'est une technique consiste à ajouter des **méta-données** à un texte pour réaliser des taches ,tel que détection des frontières des phrases et détecter les entités nommés.C'est une étape très essentielle du **pré-traitement** des donnés dans l'apprentissage automatique .

L'annotation d'un texte est un lien très important dans le développement des techniques de traitement de langage naturelle . [16]

2.2.2 Création du document annotée :

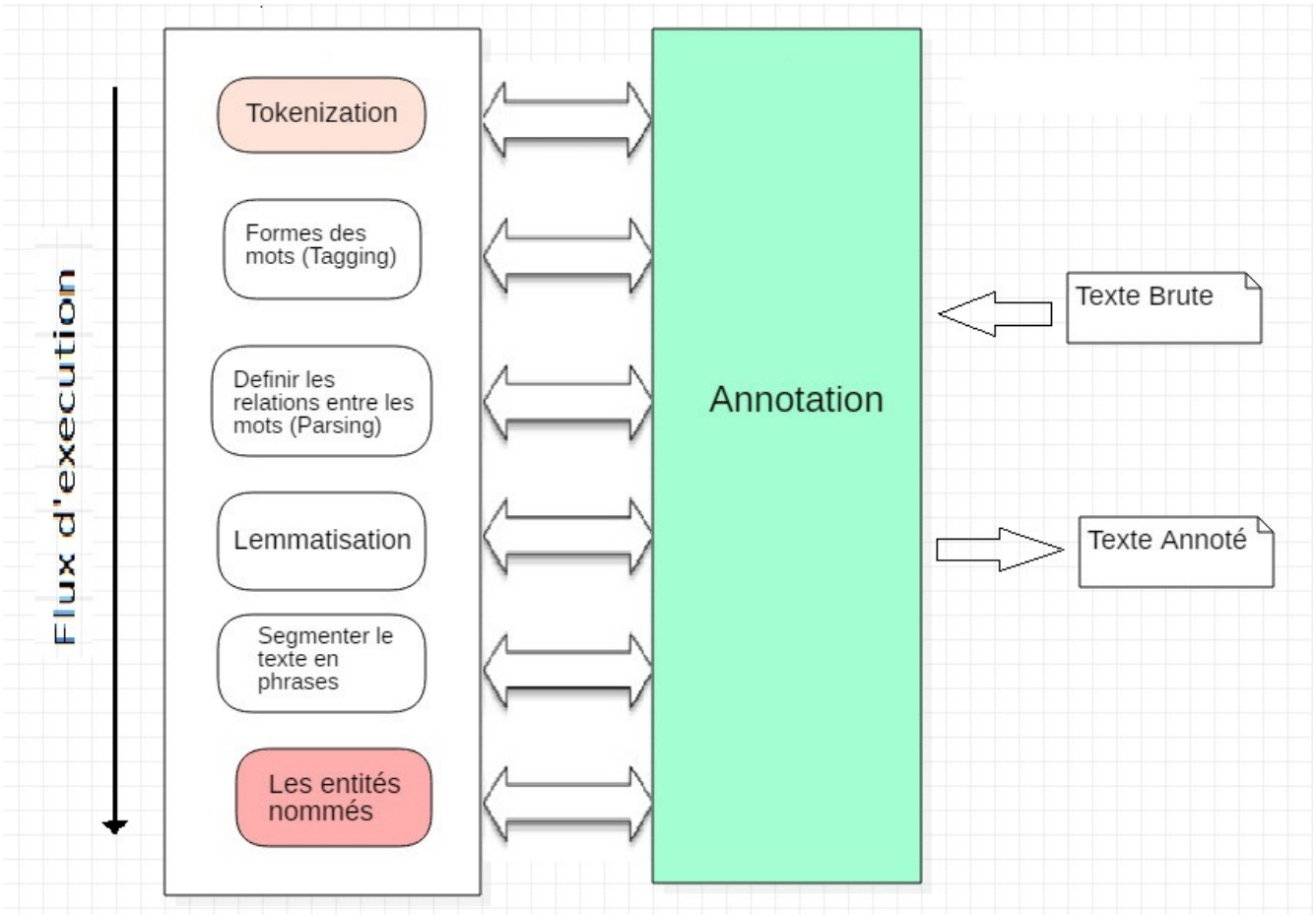

FIGURE 2.1 – Les étapes d'annotation d'un texte brute [15]

1. **Tokenization :**

   Tokenization est une tache qui permet de deviser un texte en plusieurs composant qui s'appelle *token* . Les token sont des blocs de base qui constitue le document afin de comprendre le sens du texte .

(a) **Comment faire de la Tokenization :**

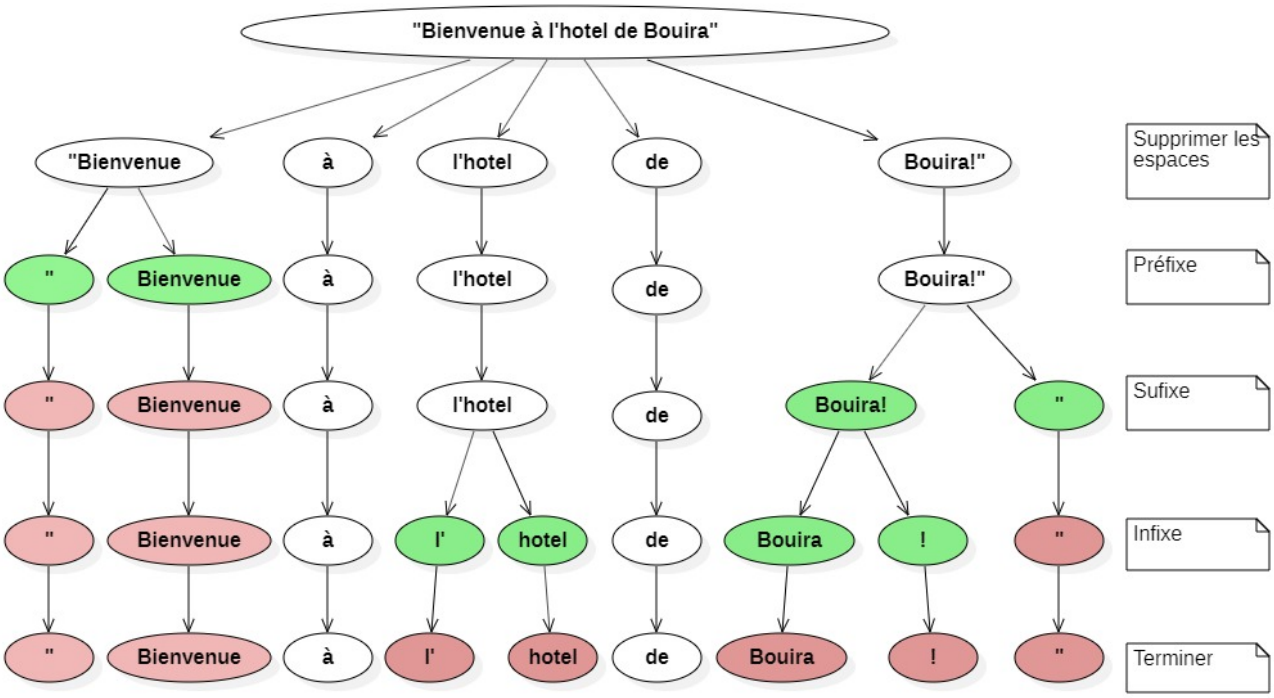

FIGURE 2.2 – Les étapes de le Tokenization [18]

(b) **Les étapes de le Tokenization :**

- **Préfixe** : Ce sont les caractères qui viennent en premier. par exemple (*parenthseouvrante*"
- **Suffixe** : Ce sont les caractères qui viennent dernier ,par exemple (*parenthsefermente*")"*le*
- **Infixe** : Les règles d'exception comme les apostrophes ,*pointd'exclamation*"!", *d'interrogat* ...

2. **Lemmatisation :** La lemmatisation est une forme compliquée de la racinisation des mots ,puisqu'elle met en évidence l 'analyse grammatical ,elle consiste à analyser les termes de manière à identifier sa forme canonique (lemme) afin de réduire les différentes formes (pluriel, féminin, conjugaison, etc.) .

   Par exemple le Lemme du mot "sont" est le verbe "être"

3. **Définir les formes grammaticaux des mots : (Tagging)** Ce processus permet de correspondre chaque token à sa forme grammatical ,exemple :nom ,verbe ,auxiliaire .

   L'annotation utilise un modèle de prédiction pour déterminer quelle est la forme du mot dans la phrase .Par exemple les mots qui suivent "le,les..." sont des noms ,mais cette prédiction peut être trompeuse comme dans cet exemple : "je le trouve sympa" .

   Aussi il y a des mots qui ne se ressemblent pas .Mais ils désignent le même sens ,par contre il existe des mêmes mots dans des positions différentes dans la même phrase qui n'ont pas le même sens .

3.1. **Exemple la forme grammatical des mots d'une phrase :**

L'	Algerie	est	le	plus	grand	pays	du	continent	africain.
DET	NOUN	AUX	DET	ADV	ADJ	NOUN	DET	NOUN	ADJ

FIGURE 2.3 – Les formes grammaticales des mots [18]

4. **Dépendance grammatical des mots dans une phrase (Parsing) :** Ce processus permet de trouver les relations entre les mots ,par exemple :sujet ,verbe ,complément ,Groupe nominale ...

4.1. **Exemple de représentation des dépendances des mots**

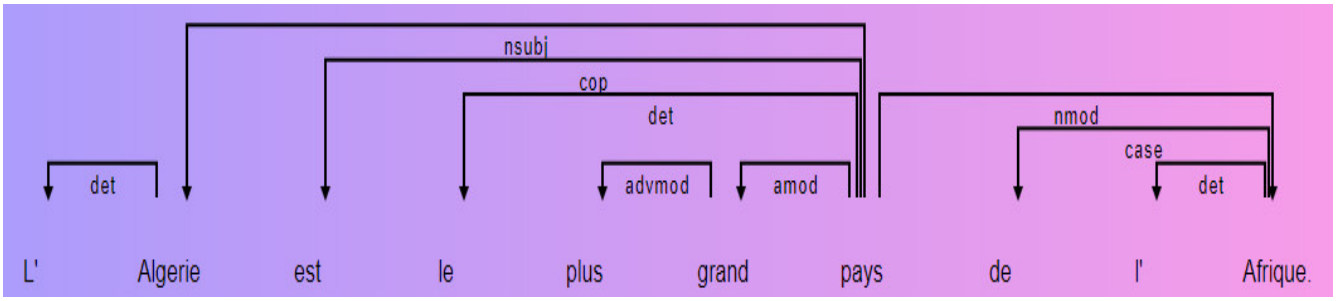

FIGURE 2.4 – Les dépendances des mots d'une phrase [18]

Cette opération aide à détecter les frontières des phrase .

5. **Segmenter le texte en phrases :** Cette étape sert a deviser le texte en plusieurs phrase selon des condition prédéfinis .Par exemple segmenter le texte lorsque la phrase se termine par un point ou une virgule .

6. **Les entités nommées** L'extraction de l'information qui désigne l'opération de conversion du texte en données structurée nous mène à mettre en évidence les entités nommées .Ce sont tous les éléments du langage qui font référence à une entité unique et concrète, appartenant à un domaine spécifique (ie.humain, économique, géographique, etc.)tel que les noms propre ,des expressions de temps te de quantité.[8]

6.1. **Extraction des entités nommées**

Cette opération à pour objectif de repérer et catégoriser toute en contenue dans un texte .il existe plusieurs approches :

6.1.1. **Symbolique** : Approche à base règle , cette approche est très utilisé par les processus de détection d'entité nommés .Les règles d'extraction sont écrites par des experts du domaine linguistiques .

6.1.2. **Statistique** :C'est une approche appelé aussi approche par apprentissage , elle utilise des processus automatique pour l'extraction d'information. Son principe est de mettre en point des modèles d'analyse à partir d'une grande masse de donnés .

6.1.3. **Hybride(Symbolique+Statique)** : Elles permettent de comprendre que dans une phrase comme « Orange n'est pas cotée en bourse », « Orange » réfère à une entreprise, alors que dans « Notre voyage à Orange s'est bien terminé », « Orange » réfère à la ville et que dans « J'ai fait de la confiture à l'orange », « Orange » réfère au fruit et non pas à une entité nommée comme dans les deux précédents .Cette forme aide à traiter les homographes (Les mots qui ont la même forme écrite ,mais qui ont une signification différente) .[17]

6.2. **Exemple d'une phrase contient des entités nommées**

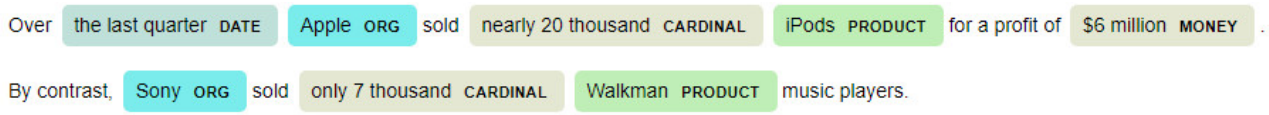

FIGURE 2.5 – Exemple des entités nommés qui content des nom d'organisation ,des dates ,produits ,devises ... [18]

2.3 Classification

2.3.1 Objectif de la classification

La classification est une discipline relié de près ou de loin a plusieurs domaines donc Il ne suffit pas de collecter des montagnes d'informations, de les stocker dans des bases de données, mais il faut les exploiter c.-à-d. en tirer des connaissances.

1. **Les étapes d'une classification :**
   (a) Choix des données .
   (b) Calcul des similarités entre les n individus à partir des données initiales .
   (c) Choix d'un algorithme de classification et exécution .
   (d) L'interprétation des résultats :
   - Évaluation de la qualité de la classification .
   - Description des classes obtenue .[10]

2.3.2 Type de classification

1. **Apprentissage supervisée (classification ou discrimination) :**

1.1. **Définition :** Le «classement» est une méthode supervisée qui consiste à définir une fonction qui attribue une ou plusieurs classes à chaque donnée. Dans cette approche les classes sont connues à priori. Selon [Govaert, 2003] . La conception supervisée d'un classificateur à C classe (ensemble fini de classe c_i) est le fait de classifier N objets (x_i) de même nature (des phonèmes, caractères manuscrits,. . .) sachant que ces N objets sont supposés avoir

été préalablement « étiquetés » par un « superviseur » en C ensembles qui forme un ensemble d'apprentissage. c'est-à-dire , on cherche à prédire si un objet (élément) « x_i» décrit par un ensemble de descripteurs «d», appartient ou non à une classe « c_j » parmi N Classes, pour faire ça on a une méthode décrit par la formule suivante :

$$(x_1, c_2)(x_2, c_4)(x_3, c_2)...(x_i, c_j)/x_i \in R^d, c_j \in C$$

La performance de la classification dépend notamment de l'efficacité de la description. De plus, si l'on veut obtenir un système d'apprentissage, la procédure de classification doit permettre de classer efficacement tout nouvel exemple (pouvoir prédictif).

1.2. **Les algorithmes de classification supervisée :**

i. **Les K plus proches voisins (K-PPV) :** La méthode des plus proches voisins (noté parfois k-PPV ou k-NN pour-Nearest-Neighbor) consiste à déterminer pour chaque nouvel individu que l'on veut classer, la liste des plus proches voisins parmi les individus déjà classés.
En fait, le k-NN est un type spécial d'algorithme qui n'utilise pas de modèle statistique. Il est "non paramétrique" et il se base uniquement sur les données d'entraînement. Ce type d'algorithme est appelé $memory - based$. Si l'on a une nouvelle entrée dont on veut prédire la classe, on va simplement regarder les k voisins les plus proches de ce point et regarder quelle classe constitue la majorité de ces points, afin d'en déduire la classe du nouveau point.

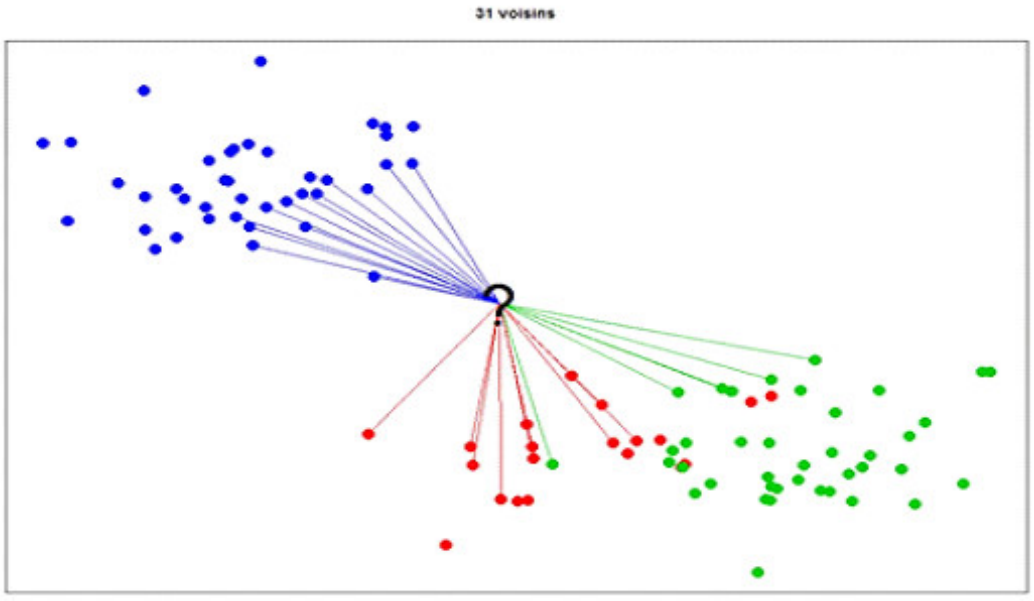

FIGURE 2.6 – Illustration d'une classification avec K meilleur voisin [14]

ii. **Classificateur Naïf Bayes :**Cette méthode fondé sur le Théorème de Bayes.
Leur particularité est de prédire la valeur des paramètres du modèle en termes de probabilité. Cette classification permet d'apprendre un modèle de classification à partir des données, l'ensemble d'apprentissage «A» est connue et chaque objet est étiqueté par sa classe ,l'objectif est de chercher à classer un nouveau objet « X_{new} » non encore étiqueté. Le classificateur bayésien va choisir la classe « y» qui a la plus grande probabilité, on parle de règle MAP (maximum a posteriori) .

$$Y_{MAP} = \underset{y}{\operatorname{argmax}}\, P(y|X_{new}) = \underset{y}{\operatorname{argmax}}\, P(X_{new}|y)P(y)$$

Si on a besoin juste de déterminer la classe la plus probable pour l'instance Xnew, on fait le calcul suivant pour un exemple de test :

$$y = \underset{y}{\operatorname{argmax}}\, P(y)$$

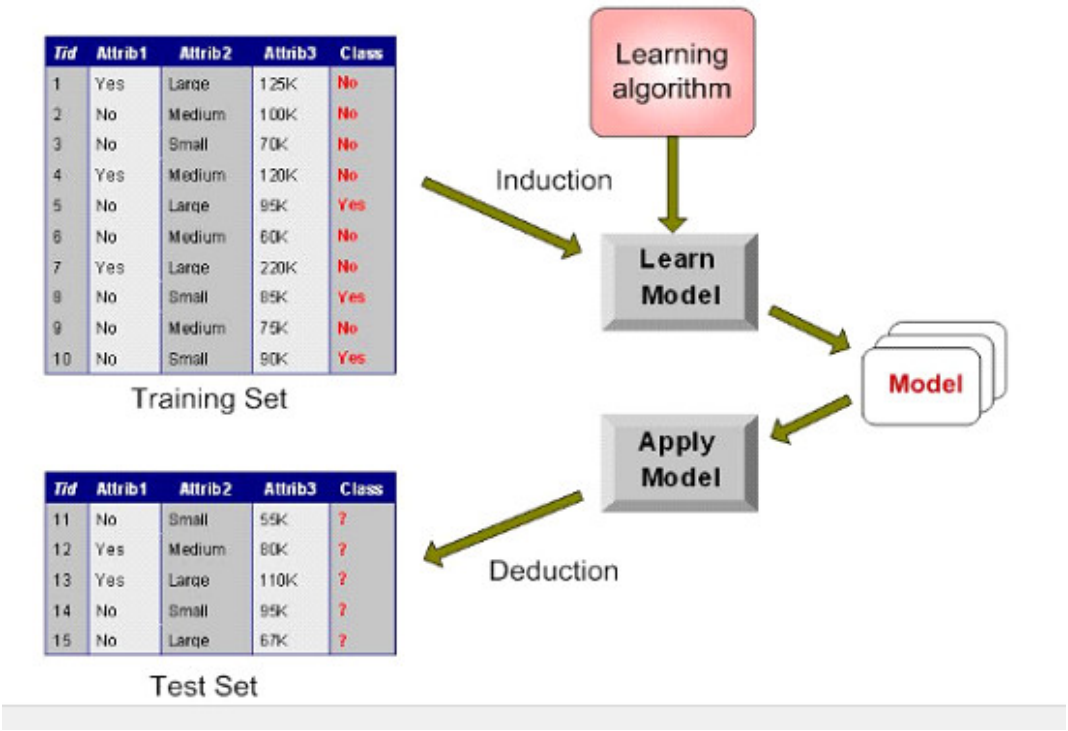

FIGURE 2.7 – Illustration d'une classification avec Naïf Bayes [14]

iii. **Machine à Vecteur de Support (MVS) :** Le MVS est un classificateur dit linéaire, ça veut dire que, dans le cas parfait, les données doivent être linéairement séparables. Ainsi nos données sont représentées comme étant un espace vectoriel. La problématique maintenant est de trouver le meilleur séparateur (ligne, plan, hyperplan) qui partage nos données en deux catégories. L'espace entre ces deux catégories est appelé marge, qui est définie par les points (Vecteurs de support) les plus proches du séparateur. Le but étant essentiellement de maximiser cette marge, plus elle est grande meilleurs est le résultat. Considérons le problème de séparer l'ensemble D composé de N paires de données/classes d'apprentissage appartenant à deux classes (A_1,A_2) :

$$D = (x_1, z_1)(x_2, z_2)...(x_i, z_j)/x_i \in R^p, z_i \in -1, +1$$

Chaque exemple d'entrée x_i est caractérisé par un ensemble de p variables descriptives. Ainsi, le vecteur x est donné comme suit :

$$X_i = (x_{i,1}, x_{i,2}, x_{i,3}...x_{i,p})$$

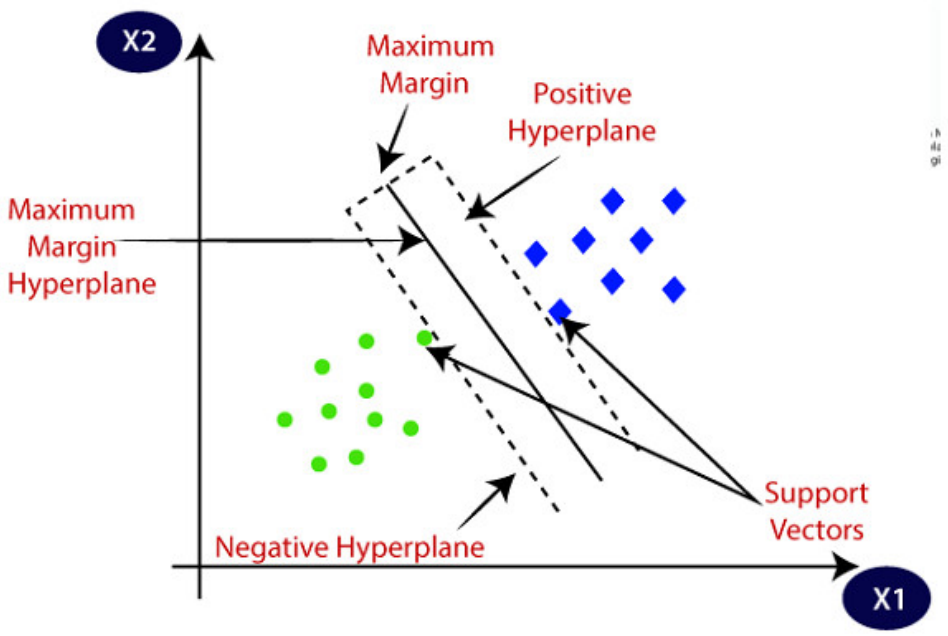

FIGURE 2.8 – L'hyperplan séparateur pour des données bidimensionnelles [14]

2. Apprentissage non supervisée(clustering)

2.1. Définition :

Le **clustering** est une méthode d'apprentissage non supervisé permettant de trouver des patterns dans les données, sert principalement à segmenter ou classifier une base de données ou extraire des connaissances pour tenter de relever des sous-ensembles de données difficiles à identifier à l'œil nu (cluster). Ca signifie un regroupement ou partitionnement des données en fonction de leurs similarités(regroupement des données qui se ressemblent). On dispose d'éléments non classés sans aucune prédiction spécifique dans le but de trouver de modèles communs. Un «Cluster» est donc une collection d'objets qui sont similaires entre eux et qui sont dissemblables par rapport aux objets appartenant à d'autres groupes.comme l'illustration suivante indique :

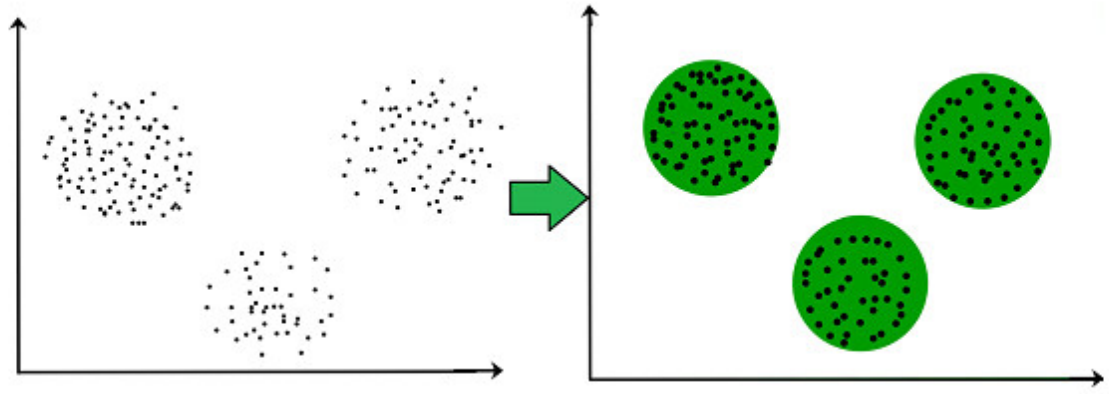

FIGURE 2.9 – Illustration de composition des clusters [15]

2.2. Les étapes du processus de clustering

D'une manière générale, le processus du clustering se divise en trois étapes principales :

A. Le pré-traitement de données .
B. Le choix de l'algorithme de clustering .
C. La validation et l'interprétation des résultats .

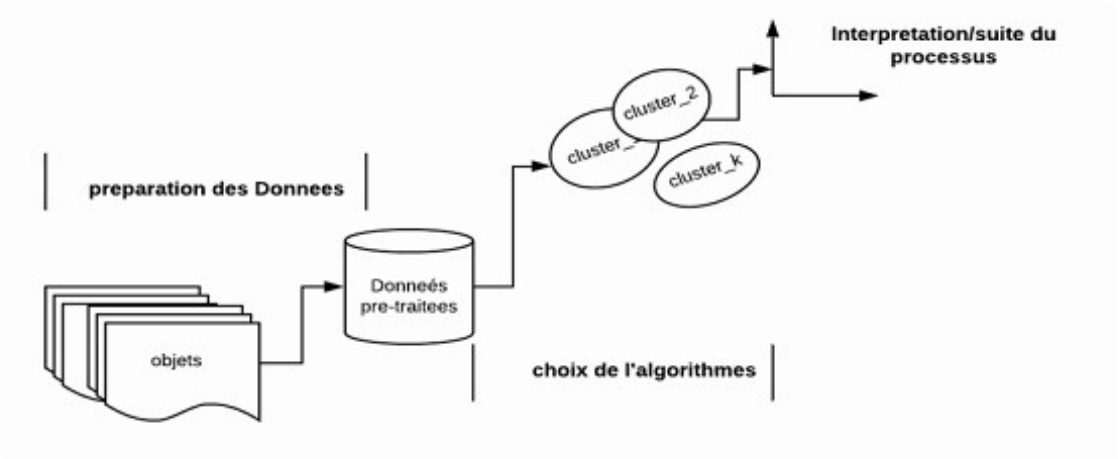

FIGURE 2.10 – Illustration du processus de clustering [15]

2.3. **Les domaine d'applications de clustering**

Les algorithmes de clustering sont les plus souvent utilisés pour une analyse exploratoire des données . Il s'agit par exemple d'identifier :

- Des clients qui ont des comportements similaires (segmentation de marché).
- Des utilisateurs qui ont des usages similaires d'un outil .
- Des communautés dans des réseaux sociaux .
- Des motifs récurrents dans des transactions financières.

2.4. **Typologie de clustering :**

Il existe plusieurs types ou bien stratégies de construction des clusters :

A. **Clustering hiérarchique**

Dans ce cas en vas faire une décomposition en arborescence des groupes. La sortie principale du clustering hiérarchique est un dendrogramme , qui montre la relation hiérarchique entre les clusters [voire la figure 2.10],il se divise en deux types :

1. **Clustering agglomératif :**
   Dans le cas du clustering agglomératif (ou bottom-up), on commence par considérer que chaque point est un cluster à lui tout seul. Ensuite, on trouve les deux clusters les plus proches, et on les agglomère en un seul cluster. On répète cette étape jusqu'à ce que tous les points appartiennent à un seul cluster, constitué de l'agglomération de tous les clusters initiaux.
2. **Clustering divisif :** c'est une approche inverse de la précédente, le clustering divisif (ou top-down), consiste à initialiser avec un unique cluster contenant tous les points, puis à itérativement séparer chaque cluster en plusieurs, jusqu'à ce que chaque point appartiennent à son propre cluster.

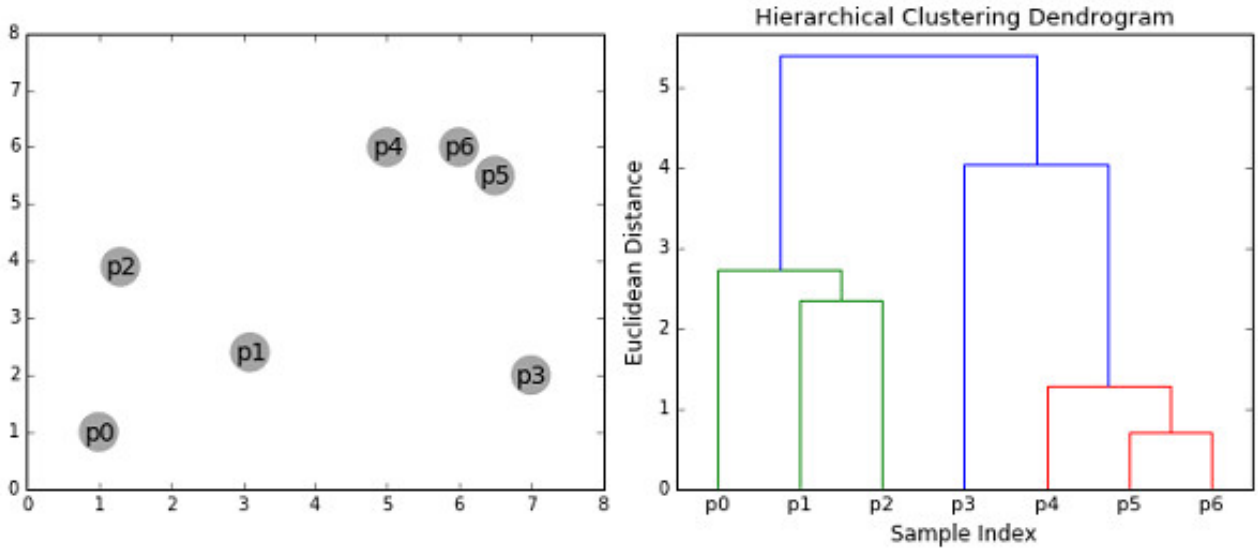

FIGURE 2.11 – Dendrogramme d'un clustering hiérarchique

B. **Clustering non-hiérarchique**
le clustering non hiérarchique vise à trouver un regroupement d'objets qui maximise ou minimise certains critères d'évaluation. Beaucoup de ces algorithmes assigneront de manière itérative des objets à différents groupes tout en recherchant une valeur optimale du critère,il sert à décomposer l'ensemble d'objets en K

2.5. **Les algorithmes de clustering les plus connus :**

A. **K-Means :**

1. **Définition :**

K-Means est un algorithme d'apprentissage non-supervisé ,il est de loin l'algorithme le plus populaire et le plus simple des algorithme de clustering .
L'algorithme est utilisées pour trouver des groupes et ensembles qui n'ont pas été étiquetés ,trouver les patterns afin de choisir des meilleurs décisions .[23]
Cette méthode suit une procédure très simple pour classer une ensemble de données à travers un certain nombre(**K**) de cluster .[22]

K-means est définie par la fonction objective qui minimise les distances entre les données dans un cluster , elle sera appliquer dans touts les clusters .[23]

2. **La fonction objective de K-means :**

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \Big(\sum_{x_j \in S_i} \big\| x_j - \mu_i \big\| \Big)$$

Avec

- x_j sont les données qui sont dans le dataset .
- S_i est un cluster .

- μ_i est le centroide du cluster S_i .[23]

3. **Les avantages est les inconvénients de K-means :**
   - **Les avantages :**
     - Algorithme rapide pour le clustering des données.
     - Facilité de l'implémentation .
     - K-means produit des clusters plus serrés que le clustering hiérarchique [23]
   - **Les inconvénients :**
     - Le résultat peut ne pas être globalement optimal, car on sélectionne plusieurs valeurs de K dans le début .
     - Cet algorithme est lent dans le traitement des large dataset ,car il accède à touts les données de cette dataset . [23]
4. **Fonctionnement de l'algorithme K-means :**
   - L'algorithme divise le data et en N dimension (dans cet exemple deux dimension).
   - Choisir la valeur **K** ,le nombre de cluster qui sera généré .
   - Initialiser k points dans le dataset comme étant le centroide initial du cluster .
   - Pour chaque donné dans le dataset faire
     - Calculer la distance entre les données du dataset et le centroide du cluster .
     - Affecter les données vers le cluster du centroide le plus proche .
   - Déplacer les centroides vers la moyenne($mean$) de l'emplacement des données du cluster .
   - Répéter les étapes 4 et 5 jusqu'au nombre maximale d'itération ,ou que les centroides arrêtent de déplacer . [23]

5. **Organigramme de K-means :**

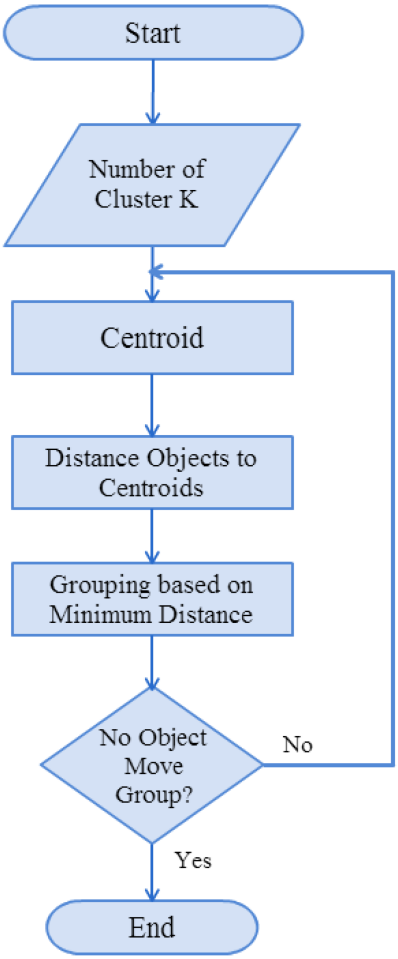

FIGURE 2.12 – Fonctionnement de l'algorithme K-means

6. **Le pseudo-algorithme de K-means :**

Input : D Data-set
m //Le nombre des données dans le dataset
$(x_1, x_2, x_3...x_m)$ //Les données du dataset
K //Initialiser le nombre des clusters .
Initialiser de façon aléatoires les centroides des clusters $(\mu_1, \mu_2...\mu_k \in R^n)$
répéter
  pour *i = 1 jusqu'au m* **faire**
    //assigner chaque donné à un cluster
    c_i = l'index de centeroide de cluster le plus proche de la données x_i
  fin
  pour *j = 1 jusqu'au k* **faire**
    μ_j= La moyenne(mean) de points (données) qui sont dans le cluster
  fin
jusqu'à μ_j *arrête de changer*;

Algorithme 1 : K-means [23]

7. **Exemple du Fonctionnement de l'algorithme :**

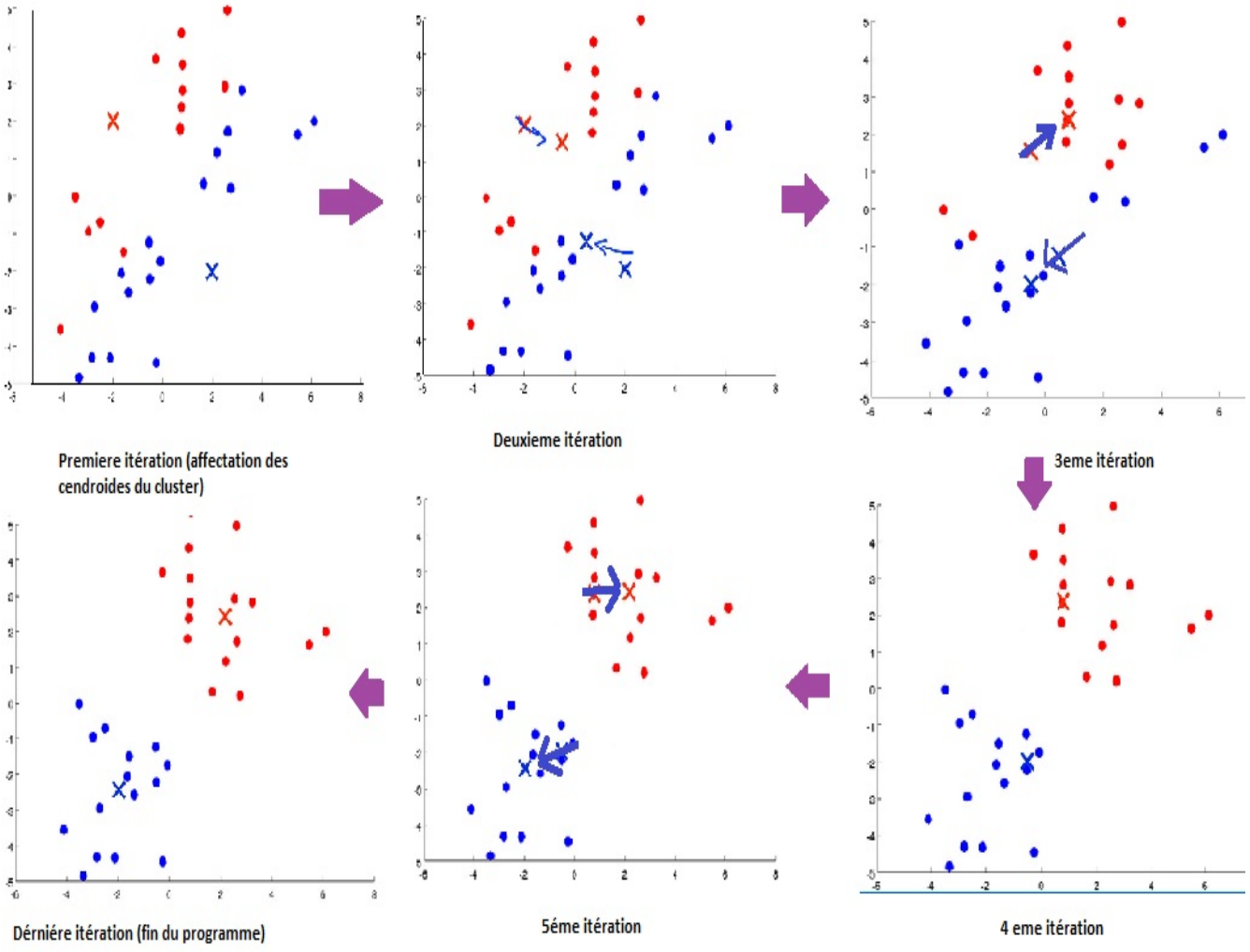

FIGURE 2.13 – Fonctionnement de l'algorithme K-means

8. **Comment trouver le meilleur K :**
Pour utiliser l'algorithme de k-means ,les utilisateur sont censés de trouvé le meilleur K (nombre de groupes) afin d'acquérir des bons résultats .
Cependent ,il existe une méthode pour trouver le meilleur K ,c'est la **méthode du coude** (elbow method) .
L'idée de cette technique ,est d'exécuter l'algorithme K-means sur une plage de valeur de K et calculer le SSE (somme carré de la distance entre une donnée du cluster et son centroide). Après ,on trace le graphe (qui correspond à un bras) ,la valeur qui se situe dans le coude du bras c'est la valeur du meilleur K de l'algorithme .(pour plus de détaiile voire la figure 2.12)

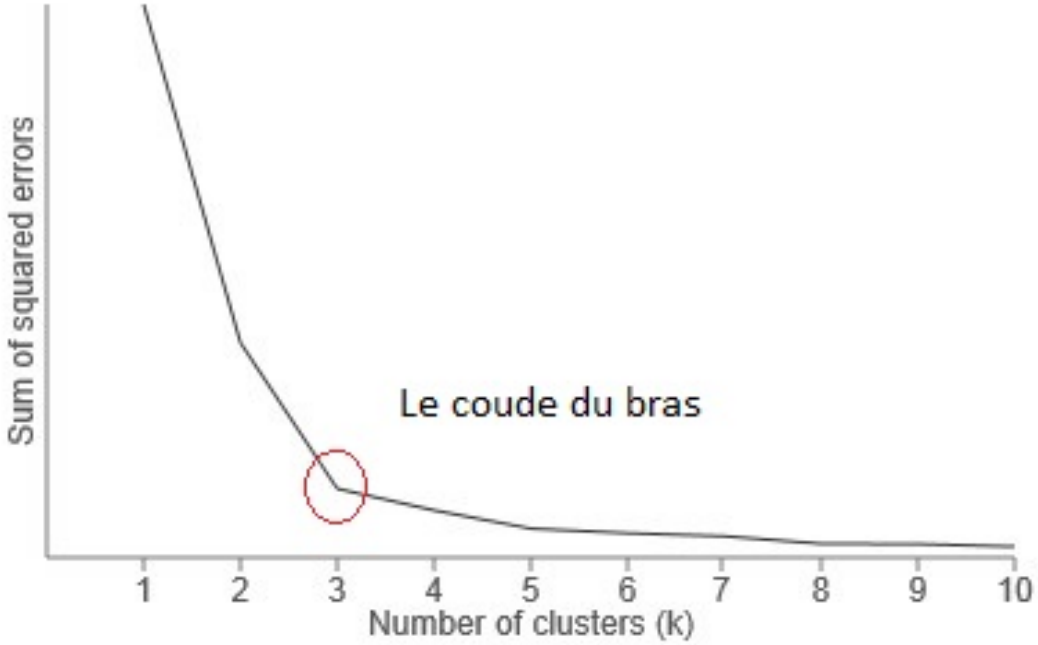

FIGURE 2.14 – Fonctionnement de la méthode *Elbow* [22]

B. **L'algorithme DBSCAN :**

1. **Définition :**

   DBSCAN (aussi appelé algorithme de clustering basé sur la densité) est une méthode utilisé par l'apprentissage automatique ,le but de cet algorithme est de séparer les clusters qui ont une petite et une grande densité . cette méthode est utilisée pour construire les modèles et les algorithmes de machine learning . [19]

2. **Fonctionnement de l'algorithme DBSCAN**

   - L'algorithme divise le data et en N dimension (dans cet exemple deux dimension).
   - Sélection de deux paramètres
     - Le nombre de donnée minimum dans un cluster (**minPts**) .
     - La distance minimale entre deux données dans un cluster appelé **epsilon** .
   - Choisir une donnée aléatoire et voir si elle est une donnée noyau d'un cluster ou non selon deux critères
     - La distance minimale entre la donnée noyau et un voisin est inférieur à epsilon .
     - Le nombre de donnée voisin supérieur ou égal du nombre de donnée minimum (**minPts**)
   - **Important :** Les données qui sont à l'intérieur de l'epsilon des données voisines ,et qui contienne le nombre minimale de nœuds dans leur epsilon ,sont aussi intégré dans le cluster (voir la figure 2.13 pour plus de détaille).
   - Les données qui n'appartiennent à aucun cluster sont appelées valeur aberrante (*Noise*) . [19]

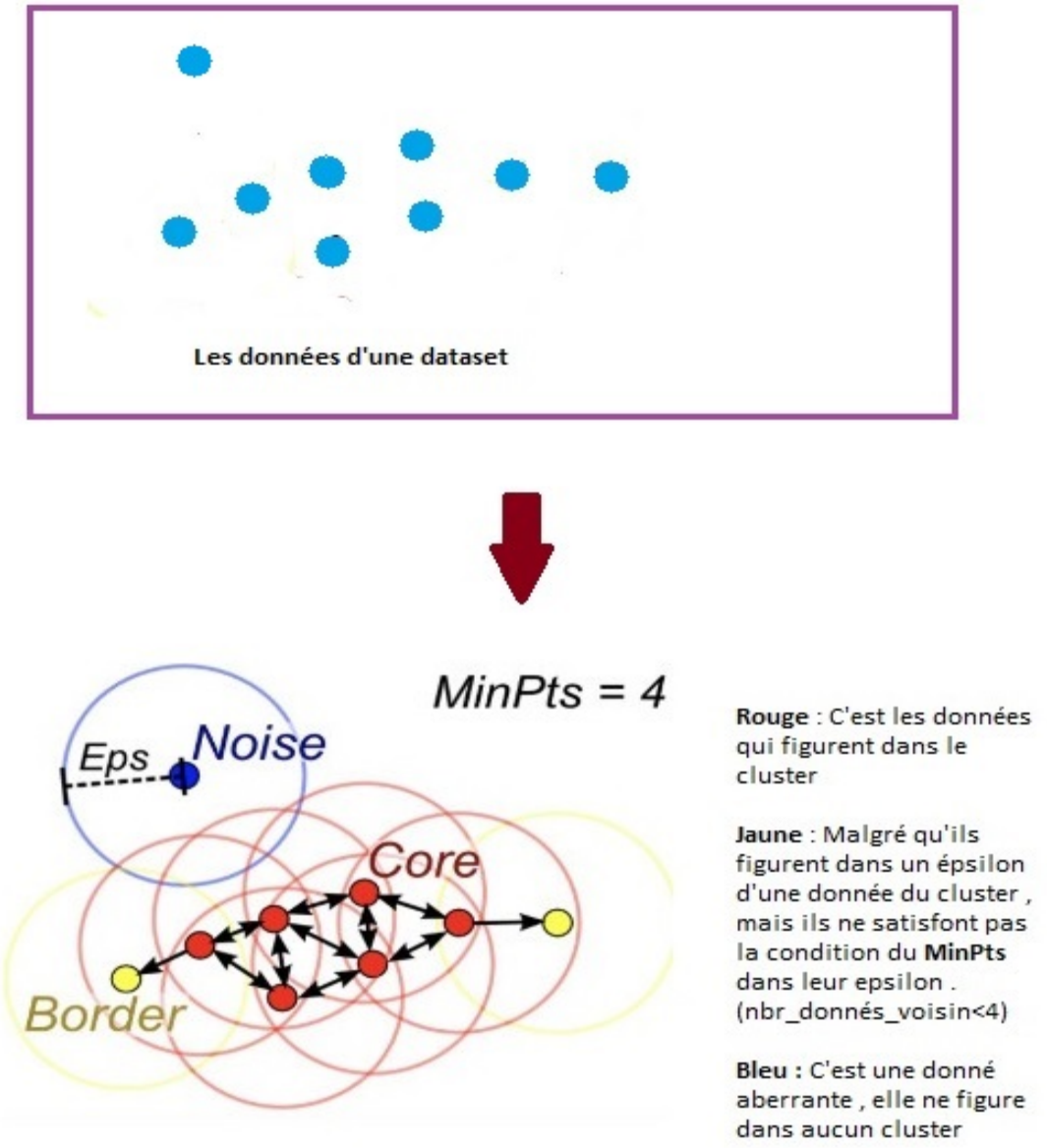

FIGURE 2.15 – Fonctionnement de l'algorithme DBSCAN [19]

3. **Le pseudo-algorithme de DBSCAN :**

```
Input : D            Data-set
        eps          //La distance maximale entre l'objet principale
        minPts       //nombre minimale de points dans un cluster.
        nombre_cluster = 0
pour chaque objet p dans D faire
    si p n'est pas traité alors
        C = trouver_les_voisins(p,eps)
        si |C| < minPts alors
            Marquer p comme valeur aberrante
        sinon
            pour chaque m dans C faire
                K = trouver_les_voisins(m,eps)
                si |K| > = minPts alors
                    ajouter la donnée m dans C
                fin
            fin
            Signaler C comme un cluster
            nombre_cluster = nombre_cluster + 1
        fin
    fin
fin
```

Algorithme 2 : DBSCAN [20]

4. **Les avantages est les inconvénients de DBSCAN :**
   - **Les avantages :**
     - Il est excellent dans la séparation des clusters qui ont une grande densité ,et les clusters qui ont une petite densité .
     - Excellent pour la gestion des données aberrantes (outliers) dans une dataset . [19]
   - **Les inconvénients :**
     - L'algorithme souffre lors de la séparation des clusters qui ont la même densité . [19]
5. **Différence entre K-Means et DBSCAN :**

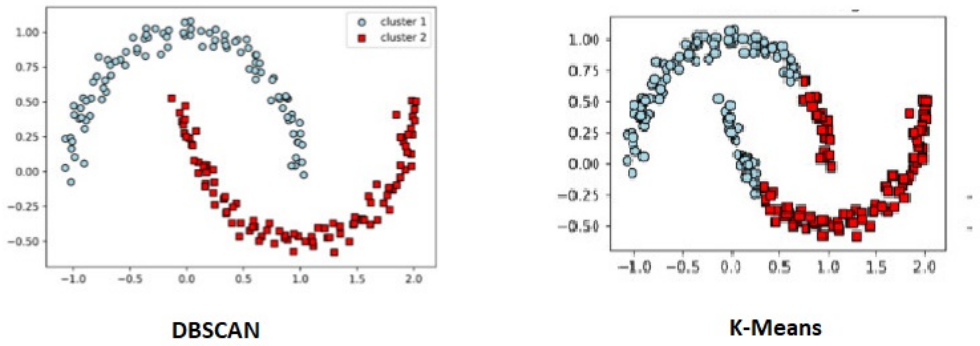

FIGURE 2.16 – Les étapes d'annotation d'un texte brute [21]

2.4 Conclusion

Dans ce chapitre nous avons étudié la conception des méthodes d'apprentissage automatique (discrimination et clustering) de manière générale. Nous avons vue comment une machine peut apprendre dans la première approche (la discrimination ou la classification) à partir d'un échantillon d'exemples classé (étiquetés) de classer un nouvel exemple non étiqueté .La classification peut être basée sur plusieurs stratégies, probabiliste (Classifieur naïf de Bayes) ou bien approximative (plus proches voisins, MVS). On va passer à la deuxième approche (le clustering) si la première ne peut pas traiter un certain problème c'est-à-dire, d'apprendre à partir d'une base sans aucune connaissance préalable. Mais dans certains moments, on se trouve dans un cas ou on est obligé de faire un pré-traitement des données avant de classifier (classification des données de type texte), il faut ajouter des méta-données afin de passer d'un texte brute au texte annoté pour que la machine comprend et apprend de façon efficace, comme nous avons montré dans la première partie.

Bibliographie

[1] https ://www.marketing91.com/types-of-media/
Consulté le : 10/02/2020-19 :30

[2] https ://whatagraph.com/blog/articles/different-types-of-media
Consulté le : 12/02/2020-22 :12

[3] https ://www.ecrirepourleweb.com/outil-internet-media-difference/
Consulté le : 15/02/2020-22 :39

[4] http ://www.fao.org/elearning/Course/FCOM/fr/pdf/trainerresources/learnernotes0861.pdf
Consulté le : 17/02/2020-20 :49

[5] https ://www.statista.com/statistics/268695/number-of-tv-households-worldwide/

[6] https ://fr.wikipedia.org/wiki/Actualité Consulté le : 02/04/2020-19 :44

[7]

[8] http ://infolingu.univ-mlv.fr/Bibliographie/Elsa/Expose_LG2007
Consulté le : 02/04/2020-18 :37

[9] http ://www.ieee.ma/uaesb/pdf/algo-classification.pdf
Consulté le : 02/04/2020-19 :43

[10] http ://dspace.univ-tlemcen.dz/bitstream/112/1045/4/Memoire.pdf
Consulté le : 02/04/2020-20 :56

[11] https ://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203345-clustering-definition/ Consulté le : 04/04/2020-19 :03

[12] https ://www.similarweb.com/fr/top-websites/category/news-and-media
Consulté le : 05/04/2020-17 :56

[13] https ://www.businessinsider.fr/us/the-30-biggest-media-owners-in-the-world-2016-5
Consulté le : 06/04/2020-15 :12

[14] https ://d-scholarship.pitt.edu/25116/2/paper327.html
Consulté le : 02/04/2020-15 :12

[15] https ://stanfordnlp.github.io/CoreNLP/pipelines.html
Consulté le : 15/04/2020-17 :12

[16] https ://www.oreilly.com/library/view/natural-language-annotation/9781449332693/ch01.html
Consulté le : 17/04/2020-14 :40

[17] http ://e-biblio.univ-mosta.dz/bitstream/handle/123456789/6143/MINF170.pdf
Consulté le : 17/04/2020-15 :10

[18] https ://spacy.io/usage/linguistic-features-dependency-parse
Consulté le : 18/04/2020-20 :56

[19] https ://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818
Consulté le : 19/04/2020-13 :20

[20] https ://www.sciencedirect.com/science/article/pii/S0169023X06000218
Consulté le : 19/04/2020-21 :15

[21] Sebastian Raschka , Vahid Mirjalili. *Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*
Consulté le : 19/04/2020-19 :31 ,page 393

[22] https ://medium.com/@dilekamadushan/introduction-to-k-means-clustering-7c0ebc997e00
Consulté le : 22/04/2020-15 :31

[23] https ://medium.com/datadriveninvestor/k-means-clustering-b89d349e98e6
Consulté le : 22/04/2020-15 :45