REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre : 32 /RTIC/M2/2022

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : **Réseaux et Technologies de l'Information et de la Communication (RTIC)**

Une Application pour la gestion du flux Web

Par :

Khadraoui Aimen

Soutenu le 20/06/2023 devant le jury composé de :

Berghida Meryem	M C B	Président
Abdelli Belkacem	M C B	Rapporteur
Bendahmane Asma	M A A	Examinateur

Année universitaire 2022-2023

Remerciement

Je voudrais profiter de cet espace pour exprimer ma gratitude envers les personnes qui ont contribué à mon parcours académique et personnel.

Tout d'abord, je voudrais remercier mes professeurs et mes collègues pour leur soutien et leur encouragement tout au long de mon cursus universitaire. Leurs enseignements et leur expérience ont été inestimables pour mon développement personnel et académique.

Je tiens également à exprimer ma reconnaissance envers mon superviseur, le Dr. Abdelli, pour ses conseils éclairés et son soutien dans mes projets de recherche.

Je ne peux pas oublier de remercier ma famille et mes frères et sœurs pour leur amour, leur soutien et leur encouragement constants. Votre soutien inconditionnel a été une source de force pour moi.

Je voudrais également exprimer ma gratitude envers mon père décédé et lui adresser mes prières et mes pensées. Ta présence nous manque tous les jours.

Enfin, je voudrais exprimer ma gratitude envers mon ami Brahim Remaigui pour son soutien inébranlable et son amitié précieuse.

Merci infiniment à tous ceux qui ont contribué à mon parcours. Je n'aurais pas pu arriver où je suis aujourd'hui sans vous tous.

Cordialement,

aymen khadraoui

ملخص:

من الواضح أن تطور الويب وزيادة حجم المعلومات والمستندات المتاحة للمستخدمين ا في استرجاع المعلومات يشكل تحًدا كبير بطريقة دقيقة وفعالة. وبالتالي، يوجد حاجة ملحة إلى وجود تقنيات وأدوات تستطيع القيام بهذه المهمة بنجاح، وتوجيه المستخدمين إلى المعلومات الأكثر صلة وفائدة بالنسبة لهم .

وعليه، فإن إنشاء برنامج لرصد تدفق إدارة الويب واسترجاع المعلومات يعتبر الحل الأمثل لتلبية هذه الحاجة. ويتطلب ذلك استخدام التقنيات الحديثة والأدوات المتقدمة في مجال رصد واسترجاع المعلومات وإدارة تدفق الويب، وتصميم البرنامج بطريقة تتوافق مع احتياجات العميل المحددة.

الكلمات المفتاحية : إدارة تدفق الويب، تقنيات الرصد ، استرجاع المعلومات.

Abstract

The development of the Web has led to the creation of a considerable number of documents, which has complicated the management of this information flow. Tracking and retrieving relevant information in real-time has become challenging due to the abundance of available content. As a result, a project is currently underway to design a web flow management program that will enable the client to track and retrieve desired information in real-time. The program will also take into account the necessary funding to access the documents that meet the user's needs. This initiative aims to enhance the management of the Web flow and to facilitate access to pertinent information for users.

.

Keywords : *Web flow management, technological monitoring , Information retrieval.*

Résumé

Le développement du Web a engendré la création d'un nombre considérable de documents, ce qui a compliqué la gestion de ce flux d'informations. Le suivi et la récupération des informations pertinentes en temps réel sont devenus des tâches ardues en raison de la quantité importante de contenu disponible Ainsi, un projet est en cours de conception pour développer un programme de gestion de flux Web qui permettra de suivre et de récupérer les informations souhaitées par le client en temps réel. Ce programme tiendra également compte des financements nécessaires pour accéder aux documents correspondant aux besoins de l'utilisateur. Cette initiative vise à améliorer la gestion du flux Web et à faciliter l'accès aux informations pertinentes pour les utilisateurs

Keywords : *Gestion du flux web , Veille technologique , Recherche d'information.*

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

l'introduction générale

Depuis l'avènement d'Internet, la technologie de l'information a connu une évolution rapide et continue, qui a eu un impact significatif sur le monde entier. L'un des aspects les plus importants de cette évolution est le développement du web, qui a révolutionné la manière dont nous partageons et accédons à l'information. Cependant, cette évolution a également créé un défi important pour la gestion de la masse de données, car le volume croissant d'informations disponibles sur le web peut être difficile à gérer. La gestion du flux web est donc devenue une préoccupation majeure pour les entreprises et les organisations qui cherchent à tirer parti de la richesse d'informations disponibles sur le web.

Dans ce contexte, les techniques de veille et de récupération d'informations sont devenues essentielles pour surveiller et analyser efficacement les données du web, afin d'extraire des informations pertinentes et utiles pour prendre des décisions éclairées en matière de marketing, de développement de produits et de gestion de la concurrence. Toutefois, la quantité considérable d'informations disponibles sur le web, provenant de différentes sources, représente un véritable défi pour la gestion des flux web. Il est donc crucial de développer des outils et des techniques pour suivre, analyser et récupérer ces informations de manière efficace.

Dans cette optique, notre étude vise à présenter les concepts de gestion des flux web, d'indexation et de classement des documents, ainsi que les techniques de veille d'informations, Nous cherchons à développer un application de gestion des flux web qui permettra aux clients de suivre et de récupérer les informations selon leurs besoins et et leur domaine d'intérêt.Le application proposé a pour vocation de proposer un accès rapide et efficace à une multitude d'informations triées avec précision en fonction de la pertinence et de l'actualité des besoins du client.

Le application en question présente les résultats sous forme de liste claire, offrant une organisation optimale des données captées et récupérées. En effet, ce document récupéré est accompagné d'informations détaillées telles que la date et le lieu de publication, ainsi que le titre du sujet et un résumé du contenu, Cette approche permet de simplifier et d'accélérer le processus de récupération d'informations intégrées, constituant ainsi une solution efficace pour relever les défis de la gestion des flux web et améliorer l'expérience des utilisateurs.
L'objectif est de fournir un outil efficace pour aider les individus et les organisations à gérer la quantité croissante d'informations disponibles sur le web et à les exploiter de manière efficace.Nous allons donc présenter un bref résumé de l'organisation de mémoire :

- **le premier chapitre :**

  Nous avons défini les concepts fondamentaux de la recherche d'information, ainsi que la méthode d'analyse, de classification et d'indexation.

- **le deuxième chapitre :**

  nous avons abordé la question de la veille technologique, ses types et ses méthodes de mise en œuvre, et présenté certaines activités de veille technologique.

- **le troisième chapitre(Partie de conception) :**

  nous avons proposé la mise en place d'un système de gestion de flux web. Nous avons commencé par présenter une étude approfondie de ce système en établissant des organigrammes pour illustrer la structure et les différentes caractéristiques du système

- **le Quatrième chapitre(partie de implémentation et résultats) :**

  Nous avons ensuite présenté les résultats de notre étude, qui ont démontré l'efficacité et la pertinence de ce système de gestion de flux web pour améliorer la gestion de l'information et faciliter l'accès aux informations pertinentes pour les utilisateurs.

Chapitre 2

Gestion de Flux Web

2.1 Introduction

L'estimation de la capacité actuelle du Web et de la quantité d'informations qu'il contient n'est pas une tâche facile. Toutefois, une étude récente menée par des statisticiens danois et néerlandais a suggéré que le nombre de pages web pourrait varier entre 4,65 et 10 milliards. Cela souligne l'importance de la recherche d'informations sur le Web, en raison de la taille considérable de sa base de données et de la rapidité et de la facilité avec lesquelles les demandes des utilisateurs peuvent être traitées .
Dans ce chapitre, nous présenterons le développement du web et la définition de l'écriture dans le web, les concepts de base de la recherche d'information, ainsi que la méthode d'analyse, de classement et d'indexation.

2.2 L'avancement de Web

Le Web a beaucoup évolué depuis sa création en 1989 par Tim Berners-Lee. Voici quelques-unes des principales étapes de son développement [10] :

- Web 1.0 :

  La première phase du Web (les années 1990)principalement concernés par la navigation sur des pages statiques, c'est-à-dire de pages Web sans interactions ni mises à jour en temps réel. Les sites Web ont été construits à partir de langages comme HTML et CSS, et les navigateurs Web étaient principalement des outils de rendu[10].

- Web 2.0 :

  Dans les années 2000, le Web a évolué vers une version plus dynamique et interactive Grâce à l'avènement de technologies avancées telles que JavaScript connue sous le nom Web Collaboratif de qui permettait de mettre à jour des pages web en temps réel sans avoir à les recharger, ainsi que l'émergence de réseaux sociaux comme Facebook et Twitter, qui favorisaient la création et le partage du contenu par les utilisateurs[10].

- Web 3.0 :

  Le Web a évolué vers une troisième phase, appelée Web 3.0.De nouvelles technologies telles que le Web sémantique, l'ontologie, les données ouvertes et les API Il permet aux machines de comprendre et d'analyser plus efficacement le contenu Web. Grâce au Web 3.0, les utilisateurs devraient avoir un accès plus facile aux informations pertinentes et avoir des expériences plus individuelles en ligne[10].

- Web 4.0 :

  Certains professionnels du Web voient déjà l'idée d'une quatrième génération du Web appelée Web 4.0. L'idée principale est basée sur l'intelligence artificielle (IA) et sur la manière dont elle peut être utilisée pour améliorer l'expérience utilisateur en ligne. Le Web 4.0 peut également être marqué par l'arrivée de nouveaux dispositifs de communication tels que les interfaces cerveau-ordinateur, ainsi que par l'utilisation accrue de la réalité virtuelle et augmentée [10].

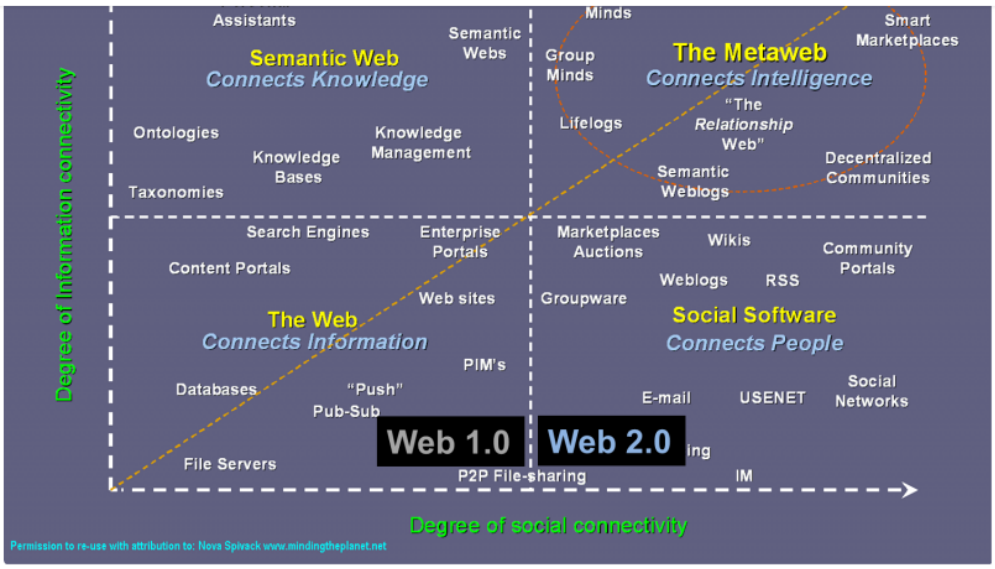

FIGURE 2.1 – L'avancement de Web[10].

2.3 La rédaction Web

est un type spécifique de rédaction qui vise à produire du contenu publiable sur des sites Web. Elle consiste à rédiger des textes qui s'adaptent aux limites et idiosyncrasies du web, notamment en termes de références naturelles (SEO)"optimisation pour les moteurs de recherche", de structure de l'information, de lisibilité et de facilité d'utilisation. La rédaction Web peut inclure la création d'articles de blog, de pages Web d'entreprise, de documents sur les produits, de newsletters, etc. Cela nécessite une bonne connaissance des règles de publication sur le web et une adaptation constante aux évolutions technologiques et aux attentes des internautes [9]. Le tableau 2.1 ci-dessous présente les dix règles à suivre pour écrire efficacement et de manière appropriée sur un sujet donné[41], en vue de la publication et de la diffusion du contenu. Ces règles visent à assurer la qualité et l'utilité du contenu, ainsi qu'à éviter tout risque de plagiat ou d'erreur. Le respect de ces règles permet d'augmenter les chances que votre contenu soit accepté et apprécié par votre public cible.

TABLE 2.1 – Les 10 règles d'écriture pour le web

Règle	Description
Choisir un sujet cohérent et pertinent	Comprendre les besoins des lecteurs et proposer des contenus pertinents pour votre activité et votre site
Écrire long	Rédiger un texte d'au moins 600 mots, de préférence plus de 1 000 mots, tout en gardant l'intérêt des lecteurs en tête
Construire un texte structuré et agréable à lire	Organiser le texte en paragraphes, utiliser des phrases courtes et non passives, et mettre en avant le plus intéressant dès le début
Ne pas copier/coller	Éviter de plagier d'autres contenus pour ne pas être sanctionné par les moteurs de recherche
Exploiter les Mots-clés (Requêtes cibles)	Choisir des mots-clés pertinents et les placer dans les titres, les textes alternatifs des images et le texte principal
Hiérarchiser les titres et intertitres	Utiliser des titres courts contenant le mot-clé et respecter les tailles minimum et maximum pour les différents niveaux de titres
Soigner la Meta description et le chapô	Rédiger une méta description concise et incitative pour inciter les lecteurs à cliquer sur le lien et un chapô concret et incitatif
Créer des liens internes et externes	Créer des liens vers d'autres contenus pour démontrer la cohérence du texte et l'expertise sur le sujet, et des liens vers d'autres sites pour montrer que vous vous inscrivez dans une communauté de sites de qualité
Utiliser des médias	Utiliser des médias tels que des images, des vidéos ou des infographies pour rendre le contenu plus attractif
Éviter les erreurs	Se relire pour éviter les erreurs et utiliser des outils de correction automatique pour aider à détecter les fautes

2.4 La Recherche d'Information

La recherche d'informations (IR) a été développée peu de temps après l'apparition des premiers ordinateurs personnels, ce qui en fait la première méthode d'accès aux documents électroniques via des ordinateurs [4]. Avec le lancement du colloque RIAO (Recherche d'Information Assistée par Ordinateur) en 1985 à Grenoble, le terme « RI » apparaît pour la première fois [4]

2.4.1 Définitions

Les définitions ont été nombreuses ces dernières années, voici trois exemples des plus utilisées :

- Définition 1 : la définition suggérée de la recherche d'information : L'objectif de cette activité est de localiser et de fournir à l'utilisateur les matériaux documentaires qui répondent le mieux à ses besoins d'information.[1]
- Définition 2 : La recherche d'informations, une ancienne branche de l'informatique, est la collecte, l'organisation, le stockage, la récupération et la sélection d'informations lorsqu'elles sont demandées.[2]
- Définition 3 : Un utilisateur qui a besoin d'informations spécifiques peut utiliser des stratégies de récupération d'informations pour utiliser et récupérer des informations qui correspondent à ses besoins. Ce domaine d'étude vise à accroître l'utilité, la pertinence et l'accessibilité de l'information.[3]

Toutes ces définitions sont basées sur le même contexte, et le but de la recherche d'information est de fournir des réponses précises et pertinentes aux questions posées par les utilisateurs en extrayant les informations nécessaires des documents disponibles.

2.4.2 Concepts de base de la recherche d'information

L'objectif premier de la recherche d'information est de mettre en place un processus permettant de retrouver des documents pertinents en réponse à la requête d'un utilisateur, à partir d'une large base de données de documents[4] Dans cette définition, on retrouve trois concepts principaux [5] : la documentation, la requête et la pertinence.

- la documentation Un document dans le contexte de la recherche d'informations peut prendre diverses formes, telles que du texte, un extrait de texte, une page Web, une image, une vidéo, etc.
- demande d'informations(Requête) Une demande d'informations représente les informations demandées exprimées par l'utilisateur. Il se présente généralement sous la forme d'une phrase contenant des mots-clés ou des modèles qui reflètent les besoins d'information spécifiques de l'utilisateur. En d'autres termes, une requête est l'expression d'une requête spécifique de l'utilisateur, qui permet de sélectionner les documents pertinents qui seront extraits de la base de données en réponse à cette requête.
- Pertinence de trouver uniquement les documents pertinents contenant les informations que l'utilisateur recherche. Le processus comprend différentes étapes, telles que la représentation, l'indexation, la recherche, la mise en correspondance et le classement des données pour répondre à la demande de l'utilisateur.

  Ce processus contient trois sous-processus :Indexation,Filtrage,Recherche[6]

  — -Indexation : Cela se réfère au processus d'indexation qui implique la présentation des documents sous forme de contenu résumé.

  — -Filtrage : Le filtrage consiste à supprimer tous les mots, espaces et balises courants afin de mieux sélectionner les mots clés pertinents pour rechercher des informations.

  — -Recherche : La recherche est le processus central de la RI, qui utilise diverses technologies pour récupérer les documents pertinents pour les utilisateurs Les trois éléments de base que le processus de recherche documentaire doit prendre en compte sont :

    - La représentation du contenu des documents
    - la représentation du besoin d'un utilisateur
    - la comparaison des deux représentations

  La figure 2.2 résume ces composants.

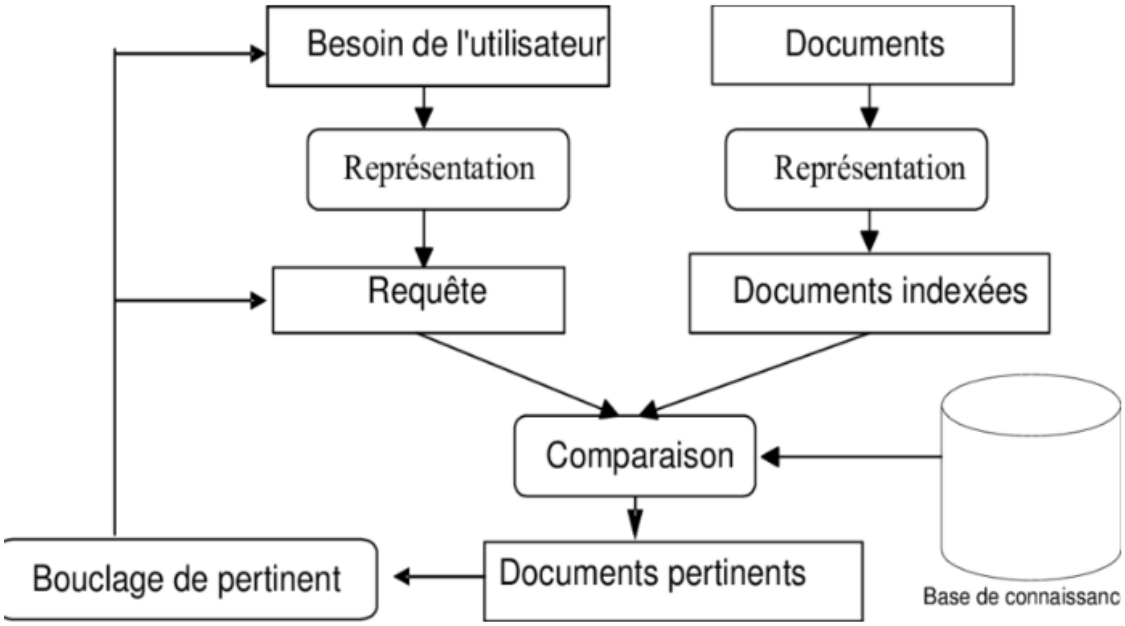

FIGURE 2.2 – Concepts de base de la recherche d'information [4].

- L'évaluation L'évaluation est un processus très critique dans le domaine de la recherche d'information (RI) car elle mesure la qualité des résultats obtenus,Nous mentionnerons les deux types de mesures les plus importants

  — La précision : Mesure le nombre de documents associés récupérés par rapport au nombre total de documents récupérés pour une requête donnée. Cette métrique est utilisée pour évaluer la qualité des résultats de recherche.La formule mathématique pour calculer la précision est la suivante [4] :

$$Précision = \frac{\text{Nombre de documents pertinents récupérés}}{\text{Nombre total de documents récupérés}}$$

  — Rappel : Dans le contexte de la recherche d'informations (IR), le rappel mesure la proportion de documents connexes récupérés par le système de recherche parmi tous les documents pertinents de l'ensemble de documents. En d'autres termes, le rappel fait référence à la capacité du système de recherche à trouver tous les documents pertinents pour une requête donnée.

$$Rappel = \frac{\text{Nombre de documents pertinents retournés}}{\text{Nombre total de documents pertinents}}$$

2.5 Les modèles de RI

Un modèle IR est un ensemble de paramètres qui définissent le fonctionnement d'un système de recherche d'informations (RI) pour classer et récupérer les documents pertinents pour une requête donnée. Il est souvent représenté par un quadrilatère [D, R, F, Sim (ri, dj)], où [4] :

- D représente la L'ensemble des documents à indexer
- RIl s'agit d'un ensemble de représentations logiques des besoins d'information d'un utilisateur
- F C'est un processus qui visualise les formulaires de document et de demande(requêtes) et les relations entre eux.
- Sim(ri, dj) C'est une fonction d'analogie qui établit une correspondance numérique entre la requête Ri et le document Dj. Cette similitude permet de classer un document par rapport à tous les documents du groupe, selon son importance par rapport à la requête Ri.

2.5.1 Modèle booléen

Le modèle booléen repose sur la théorie des ensembles. Dans ce modèle, les documents et les requêtes sont représentés par des ensembles de mots clés. Chaque document est représenté par une conjonction logique des termes non pondérés qui constituent l'index du document[4,6]. Ainsi, un document d peut être représenté comme une expression logique : $d = t_1 \wedge t_2 \wedge t_3 \wedge \ldots \wedge t_n$

Une requête est une expression booléenne dont les termes sont reliés par des opérateurs logiques (OR, AND, NOT) qui permettent d'effectuer des opérations d'union, d'intersection et de différence entre les ensembles de résultats associés à chaque terme. Par exemple, une requête q peut être représentée comme suit : $q = (t_1 \wedge t_2) \vee (t_3 \wedge t_4)$

La fonction de correspondance est basée sur l'hypothèse que la présence ou l'absence des termes de la requête dans le document est suffisante pour déterminer sa pertinence. Elle vérifie si l'index de chaque document dj implique l'expression logique de la requête q en utilisant les opérateurs logiques AND, OR et NOT. Le résultat de cette fonction est binaire et est décrit comme suit : RSV(q,d) = 1,0 où 1 indique que le document est pertinent pour la requête et 0 indique le contraire. Les opérations logiques utilisées pour évaluer la pertinence d'un document pour une requête sont représentées par les symboles pour l'opération AND, pour l'opération OR et ¬ pour l'opération NOT [8].

Le tableau 2.2 résume les avantages et les inconvénients du modèle booléen [4].

Avantages
— Correspondance exacte : un document correspond ou ne correspond pas à la requête .
— Le modèle booléen donne aux utilisateurs un sentiment de contrôle sur le système de recherche
— Le modèle est robuste .
— Les résultats sont prévisibles et relativement faciles à expliquer.
Inconvénients
— La notion de classement des documents n'existe pas dans un système de recherche d'informations booléen. Les documents récupérés sont soit classés 0 ou 1
— Difficile pour les utilisateurs non formés de manipuler correctement le modèle ;
— Tous les termes ont le même poids
— Le modèle est très strict et la correspondance exacte peut récupérer trop peu ou trop de documents
— Les requêtes complexes sont difficiles à écrire ;
— Manque de connaissance sur la façon d'utiliser ses possibilités de recherche ;
— Le modèle considère chaque document et requête comme simplement un ensemble de mots.

TABLE 2.2 – Avantages et inconvénients du modèle booléen

2.5.2 Le modèle vectoriel

Le modèle vectoriel représente les documents et les requêtes sous forme de vecteurs dans un espace vectoriel à n dimensions[7], où chaque dimension correspond à un terme du vocabulaire d'indexation. L'indice du document dj est représenté par le vecteur = (w1, j, w2, j, w3, j, ..., wn, j), où wk, j est le poids du terme tk dans le document dj , et chaque composante wk, j est un

nombre réel compris entre 0 et 1.

De même, la requête q est représentée par un vecteur V = (w1, q, w2, q, w3, q, ..., wn, q), où wk, q est le poids du terme tk dans la requête q. Le poids d'un terme dans une requête est souvent déterminé par une méthode appelée pondération, qui peut être basée sur l'importance du terme dans l'ensemble de documents ou sur d'autres critères.La correspondance est mesurée par la similarité entre le vecteur de requête et les vecteurs documents [6]. L'échelle classique utilisée dans le modèle vectoriel est le cosinus L'angle formé par les deux vecteurs [8]

$$RSV(q,d) = \frac{\sum\limits_{t \in d \cap q} w_{t,d} \times w_{t,q}}{\sqrt{\sum\limits_{t \in d} w_{t,d}^2} \times \sqrt{\sum\limits_{t \in q} w_{t,q}^2}} \tag{2.1}$$

La formule RSV (q, d) calcule le score de similarité entre une requête q et un document d dans le modèle vectoriel de RI. Voici comment chaque terme de la formule contribue à ce score :

—

$$\sum_{t \in d \cap q} w_{t,d} \times w_{t,q}$$

: cette somme calcule le produit scalaire entre les vecteurs représentant la requête et le document, c'est-à-dire la similarité cosinus entre ces deux vecteurs. Pour chaque terme t qui est présent à la fois dans la requête et le document, on multiplie le poids $w_{t,d}$ du terme dans le document par le poids $w_{t,q}$ du terme dans la requête, et on ajoute ces produits à la somme totale.

—

$$\sum_{t \in d} w_{t,d}^2$$

: cette racine carrée calcule la norme euclidienne du vecteur représentant le document d. Pour chaque terme t qui est présent dans le document, on calcule le carré du poids $w_{t,d}$ du terme dans le document, et on ajoute ces carrés à la somme totale. Ensuite, on prend la racine carrée de cette somme pour obtenir la norme euclidienne.

—

$$\sum_{t \in q} w_{t,q}^2$$

: cette racine carrée calcule la norme euclidienne du vecteur représentant la requête q. Pour chaque terme t qui est présent dans la requête, on calcule le carré du poids $w_{t,q}$ du terme dans la requête, et on ajoute ces carrés à la somme totale. Ensuite, on prend la racine carrée de cette somme pour obtenir la norme euclidienne.

En somme, la formule RSV (q, d) mesure la similarité cosinus entre les vecteurs représentant la requête et le document, tout en prenant en compte la longueur de ces vecteurs (normes euclidiennes). Plus la somme des produits scalaires est élevée par rapport aux normes euclidiennes, plus le score de similarité RSV (q, d) sera grand, ce qui indique une pertinence plus forte du document d par rapport à la requête q .

Le tableau 2.3 résume les avantages et les inconvénients du modèle vectoriel [4].

Avantages
— Les utilisateurs utilisent principalement des requêtes en texte libre, c'est-à-dire qu'ils saisissent un ou plusieurs mots plutôt que d'utiliser un langage précis avec des opérateurs pour construire des expressions de requête (modèles centrés sur l'utilisateur). — Le modèle vectoriel attribue des poids non binaires aux termes d'index dans les requêtes et les documents. — Le modèle vectoriel peut être le mieux caractérisé par sa tentative de classer les documents par similitude entre la requête et chaque document. — Cette formulation empêche le système de recherche de favoriser les documents courts par rapport à leurs homologues plus longs. — Des mesures de similarité très simples ou des schémas de pondération de termes peuvent être utilisés ; — Le défi consiste principalement à trouver un bon schéma de pondération ; — Le modèle fonctionne assez bien en pratique malgré des faiblesses évidentes.
Inconvénients
— Le principal inconvénient du modèle vectoriel est qu'il ne définit pas de valeurs appropriées pour les composantes vectorielles. — Il y a une hypothèse d'indépendance des termes ; — Le modèle vectoriel prend souvent beaucoup de temps pour calculer un espace multidimensionnel dans lequel existe une énorme quantité de termes différents. De plus, le modèle vectoriel ignore les relations sémantiques entre les termes et ne préserve aucun ordre séquentiel dans un texte.

TABLE 2.3 – Avantages et inconvénients du modèle vectoriel

2.5.3 modèle de recherche probabiliste

Le modèle de recherche probabiliste utilise la théorie des probabilités pour calculer la probabilité qu'un document soit pertinent pour une requête donnée, sur la base de deux conditions de probabilité. Une équation Bayes est utilisée pour calculer cette probabilité, en supposant l'indépendance des variables "documents liés" et "documents non pertinents"[4,6]. Le processus de recherche consiste à évaluer cette possibilité pour chaque document, et à classer les documents par ordre décroissant depertinence[8].

Soit D i (t1, t2, t3,...., tN) où

$$T_i = \begin{cases} 1 & \text{si } t_i \text{ indexe le document } D_j \\ 0 & \text{sinon} \end{cases} \tag{2.2}$$

a) P(pert/Di) = (P(Di/pert) x P(pert)) / P(Di)

b) P(nonpert/Di) = (P(Di/nonpert) x P(nonpert)) / P(Di)

où

- P(Di/pert) : Probabilité d'obtenir un document Di à partir des documents pertinents
- P(pert) :probabilité de parenté. Possibilité d'en choisir un au hasard document connexe
- P(pert/Di) : est la probabilité de pertinence du document Di sachant sa description
- P(Di) : Probabilité de choisir le i document :

  P(Di)=p(Di /pert) * p(pert) + p(Di /Nonpert) * p(Nonpert)

Le tableau 2.4 résume les avantages et les inconvénients du modèle probabilistes [4].

Avantages
— Les méthodes probabilistes ont montré de bonnes performances dans une variété de tâches, y compris la récupération ad hoc, la récupération d'informations en langue croisée, la récupération d'informations distribuée, la prédiction de difficulté de requête, la récupération de passage, etc.
— Très compétitifs et largement utilisés aujourd'hui en raison de leur solide fondement théorique dans la réflexion sur l'incertitude.
Inconvénients
— Ils ne modélisent pas de manière exhaustive le processus de récupération d'informations.

TABLE 2.4 – Avantages et inconvénients du modèle probabilistes

2.6 Analyse du Web

L'analyse de texte Web implique l'utilisation de systèmes informatiques pour comprendre le texte écrit par les humains et extraire des informations commerciales utiles. Avec un logiciel d'analyse de texte, il est possible de classer, trier et extraire des informations de manière autonome à partir de plusieurs sources textuelles, telles que les e-mails, les documents, les contenus des réseaux sociaux et les avis sur les produits. Cela permet d'identifier des tendances, des relations, des sentiments et d'autres informations précieuses de manière efficace et précise, comme le ferait un humain.

Dans le domaine de l'analyse textuelle, on distingue deux approches : l'approche traditionnelle, axée sur la morphologie et la syntaxe du langage, et l'approche plus récente, centrée sur la sémantique et la pragmatique linguistiques. La sémantique lexicale étudie le sens des mots individuels, tandis que la sémantique propositionnelle s'intéresse au sens global des phrases et des énoncés. Ces deux approches sont complémentaires et offrent des perspectives différentes pour l'analyse des textes et des discours. En somme, il est important de distinguer ces approches pour mieux comprendre comment l'analyse textuelle peut être réalisée de manière efficace et précise

[18].

2.6.1 L'analyse lexicale

L'analyse lexicale est une étape clé dans le traitement automatique du langage naturel qui implique la segmentation d'un texte en unités lexicales, appelées "tokens", telles que des mots, des nombres et des symboles de ponctuation. Cette étape permet également de déterminer la catégorie grammaticale de chaque token, par exemple, si un mot est un verbe, un nom, un adjectif. Trois étapes importantes de l'analyse lexicale d'un document texte :Suppression de mots vides Le, stemming, Lemmatisation [18] .
Voici un exemple d'analyse lexicale pour l'interrogation "Quand l'Algérie a-t-elle remporté la Coupe d'Afrique?" :

— La première étape consiste à segmenter le texte en tokens : "Quand", "l'", "Algérie", "a", "-", "t-", "elle", "remporté", "la", "Coupe", "d'", "Afrique", "?"

— Ensuite, on peut supprimer les mots vides (ou stopwords) qui ne sont pas utiles pour l'analyse sémantique : "Algérie", "remporté", "Coupe", "Afrique"

— La troisième étape est de déterminer la racine de chaque mot (stemming) ou sa forme de base (lemmatisation) pour faciliter la recherche : "Algérie", "remport", "Coup", "Afrique"

2.6.1.1 Suppression de mots vides

Certains mots comme les prépositions, les conjonctions ou les articles n'ont pas d'impact sur la signification des mots qu'ils accompagnent, car ils ne contiennent aucune information sémantique. Par conséquent, ces mots peuvent être supprimés des documents pour réduire la taille du lexique. Ce processus de prétraitement est couramment utilisé pour améliorer l'efficacité et la précision des modèles d'analyse de texte en éliminant le bruit et en se concentrant sur les informations les plus importantes[15].
Voici quelques exemples en français et anglais :

— exemples en français : "Les ,De, Et, Un, Le,des ,la,en"

— exemples en anglais :"That, This ,But, Or , The, Of, A, In "

2.6.1.2 Le stemming

Le stemming est un processus de traitement de texte utilisé pour réduire les mots à leur forme de base, ou racine, en supprimant les affixes à la fin des mots tels que les préfixes et les suffixes. Le but est de réduire les variantes d'un mot à leur forme de base commune, ce qui peut améliorer l'efficacité des algorithmes de recherche et d'analyse de texte[19].Des exemples sont illustrés dans la figure 2.3.

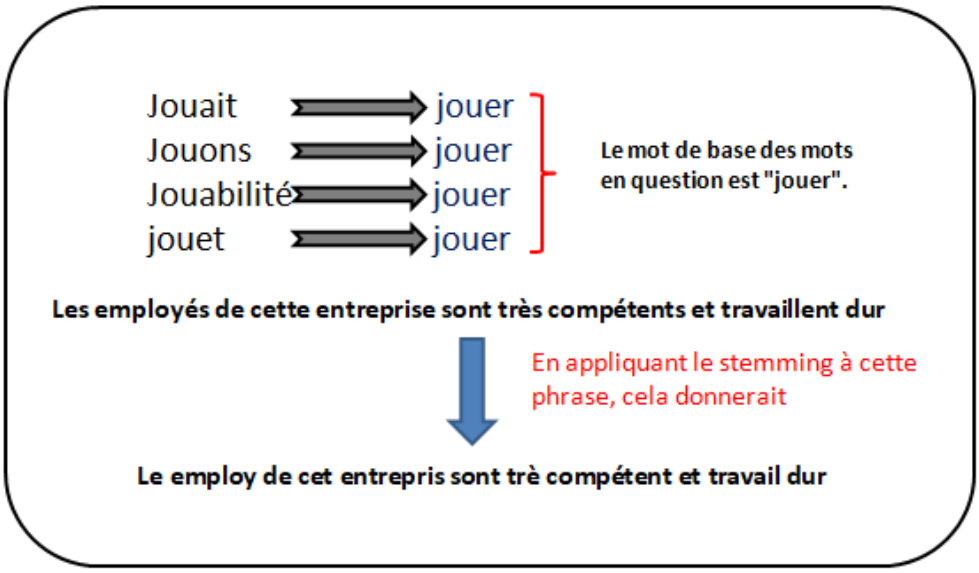

FIGURE 2.3 – Exemple Le stemming [20] .

2.6.1.3 Lemmatisation

La lemmatisation est un processus d'analyse de texte qui consiste à identifier la forme de base d'un mot en utilisant des connaissances linguistiques et morphologiques plus avancées que le stemming. Contrairement au stemming qui se contente de supprimer les affixes, la lemmatisation prend en compte le contexte et la syntaxe des mots pour déterminer leur forme de base[19]. Voici deux phrases qui illustrent l'utilité de la lemmatisation :

— "Les chats ont attrapé des souris hier soir." En utilisant la lemmatisation, on peut identifier que le lemme de "attrapé" est "attraper" et que le lemme de "souris" est "souris". Ainsi, la phrase pourrait être transformée en "Les chats ont attraper des souris hier soir", ce qui est incorrect grammaticalement mais qui montre comment la lemmatisation peut aider à

identifier les formes de base des mots.

— "Il est difficile de différencier les mots « lire » et « livre »." La lemmatisation peut aider à distinguer les différentes formes d'un mot dans le contexte. Dans cette phrase, le lemme de "lire" est "lire" et le lemme de "livre" est "livre", même s'ils ont la même racine. Cela permet de mieux comprendre la signification des mots dans le contexte de la phrase.

2.6.2 Analyse sémantique

Sélection des attributs : Les critères les plus utilisés sont [15] :

2.6.2.1 fréquence documentaire

La méthode de "fréquence documentaire" supprime les mots apparaissant dans un nombre limité de documents, considérés comme peu utiles pour prédire la catégorie d'un texte ou améliorer les performances du classificateur. Elle permet de réduire rapidement le nombre d'attributs, mais il faut faire attention à ne pas éliminer des termes potentiellement informatifs ayant une fréquence faible ou moyenne, qui peuvent être utiles pour comprendre le sens global du texte.

2.6.2.2 gain d'information

La méthode du "gain d'information" mesure le pouvoir discriminatoire d'un mot en calculant le nombre de bits d'information que la présence ou l'absence de ce mot apporte à la prédiction de la catégorie d'un texte. Cette méthode est souvent utilisée dans la construction d'arbres de décision pour choisir l'attribut qui permet de diviser l'ensemble des instances en deux groupes homogènes. Elle permet ainsi de sélectionner les mots les plus pertinents pour la classification des textes.

2.6.2.3 d'information mutuelle

La mesure d'information mutuelle (ou "mutual information") est basée sur le nombre de fois qu'un mot apparaît dans une catégorie donnée pour évaluer sa pertinence dans la prédiction de

la classe d'un document. Plus le nombre d'apparitions d'un mot dans une catégorie est élevé, plus son score d'information mutuelle est élevé.

2.6.2.4 force du terme

La méthode de la "force du terme" (term strength) est différente des autres techniques présentées. Elle vise à estimer l'importance d'un terme en se basant sur sa tendance à apparaître dans des documents similaires. Tout d'abord, des paires de documents sont formées en utilisant la similarité cosinus supérieure à un seuil prédéfini. Ensuite, la force d'un terme est calculée en utilisant la probabilité conditionnelle qu'il apparaisse dans le deuxième document d'une paire, sachant qu'il apparaît dans le premier. Cette méthode prend en compte la distribution des termes dans les documents similaires plutôt que leur relation avec une catégorie spécifique, ce qui peut être utile pour des tâches telles que la recommandation de contenu.

2.7 La Classification des pages web

Dans le monde actuel, avec des milliards de documents textuel sur le web, il n'est plus possible de se fier aux méthodes conventionnelles de recherche. Le manque d'organisation des documents disponibles nécessite un effort supplémentaire de la part de l'utilisateur pour récupérer les résultats de recherche pertinents. classement de documents textuel peut jouer un rôle important dans l'organisation efficace des documents selon les besoins de l'utilisateur. Cette technique permet de catégoriser(class) un nouveau document dans l'un des groupes prédéfinis en fonction de ses caractéristiques[12].

2.7.1 Définition de classification

La classification des pages web est le processus de catégorisation des pages web en fonction de leur contenu, de leur sujet ou de leur mais.La classification des pages Web est utilisée pour de nombreuses raisons, notamment pour faciliter la recherche d'informations sur un sujet particulier, pour la publicité ciblée, pour la sécurité Internet et pour l'analyse des tendances sur le Web [11].

2.7.2 types de classification

La classification documents textuel Web peut être divisée en plusieurs sous-problèmes plus spécifiques en fonction du type de classification requis. Voici quelques exemples courants de classification des pages Web [13] :

2.7.2.1 Classification par sujet

Classification par sujet cela implique de classer les pages Web en fonction de leur sujet ou de leur thème. Par exemple, un site d'actualités peut être classé dans des catégories telles que sport, affaires, politique et divertissement

2.7.2.2 Classification fonctionnelle

Classification fonctionnelle cela implique de classer les pages Web en fonction de leur objectif ou de leur fonction. Par exemple, une page Web peut être classée comme une page de produit, une page d'accueil, une page de contact ou un article de blog.

2.7.2.3 Classification de sentiment

Classification de sentiment cela implique de classer les pages Web en fonction du sentiment exprimé dans le contenu. Par exemple, un site d'avis peut classer les avis comme positifs, négatifs ou neutres.

2.7.2.4 Autres types de classification

Autres types de classification il peut y avoir d'autres types de classification, en fonction des besoins spécifiques de l'application. Par exemple, un moteur de recherche peut classer les pages Web en fonction de leur pertinence par rapport à une requête particulière, ou une application de sécurité peut classer les pages Web en fonction de leur niveau de menace.

2.7.2.5 La classification plate et La classification hiérarchique

La classification plate La classification plate, également appelée classification non hiérarchique, consiste à créer un ensemble de classes ou de catégories dans lequel chaque élément

ou objet ne peut appartenir qu'à une seule classe. Cela signifie que chaque classe est considérée comme indépendante et sans rapport avec les autres classes[13].

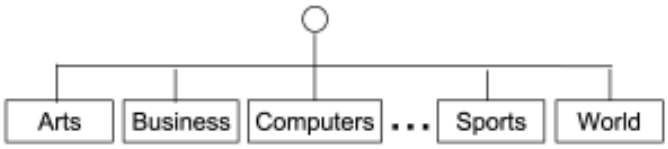

FIGURE 2.4 – La classification plate[13].

la classification hiérarchique la classification hiérarchique consiste à créer une structure hiérarchique de classes ou de catégories dans laquelle chaque classe est subsumée par une classe plus générale. Dans cette approche, chaque objet peut appartenir à plusieurs classes, en fonction de ses attributs ou caractéristiques. La classification hiérarchique est couramment utilisée dans les situations où il y a un grand nombre de classes ou lorsque les objets à classifier partagent de nombreuses caractéristiques entre eux[13].

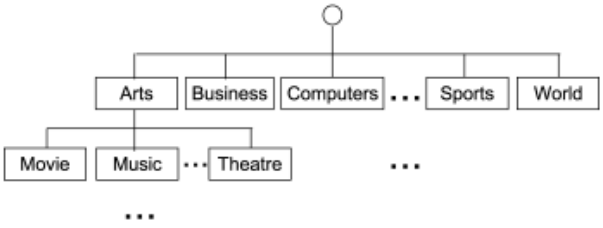

FIGURE 2.5 – La classification hiérarchique[13].

2.7.2.6 Classification binaire et multiclasse

La classification peut être divisée en classification binaire et classification multiclasse, en fonction du nombre de classes impliquées [13] : Classification binaire Dans la classification binaire, les instances sont catégorisées exactement dans l'une des deux classes

Classification multiclasse la classification multiclasse, il y a plus de deux classes impliquées. Par exemple, une classification multiclasse à quatre classes peut inclure des catégories telles que les arts, les affaires, l'informatique et les sports. La classification multiclasse peut être soit une classification à étiquettes simples, où une seule étiquette de classe peut être attribuée à une

instance, soit une classification à étiquettes multiples, où une instance peut appartenir à une, deux ou toutes les classes

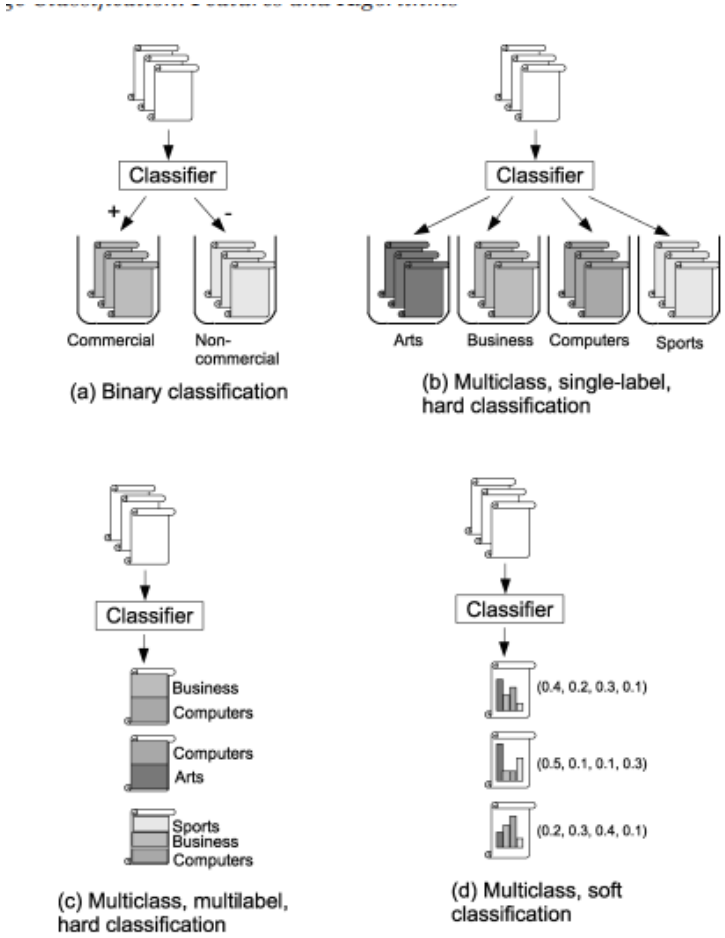

FIGURE 2.6 – Classification binaire et multiclasse [13].

2.7.3 Classification Automatique de textes

La classification de texte consiste principalement à identifier le contenu d'un document en quelques mots, dans le but de faciliter la recherche ciblée dans un ensemble de documents textuels. Cette étiquetage permet de filtrer efficacement les résultats de recherche en fonction des critères précis définis par l'utilisateur[15,17]. La création d'un algorithme de classification se compose généralement de quatre étapes [14,17] :

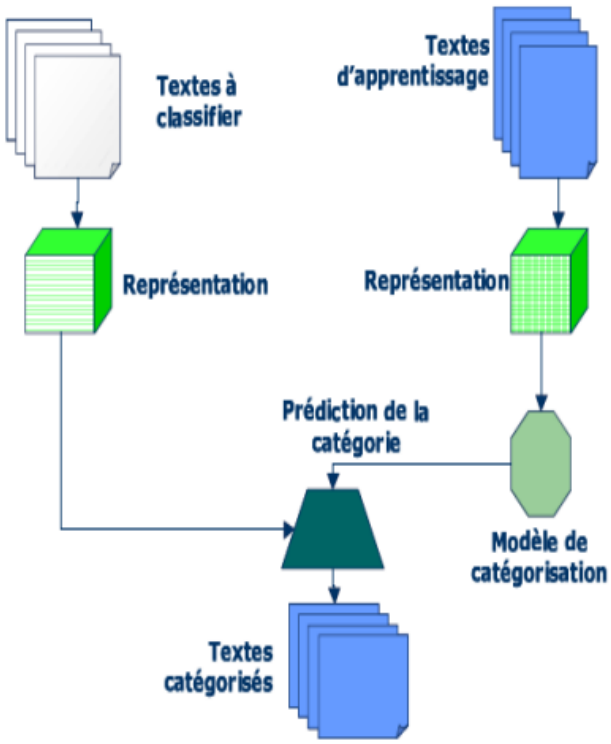

FIGURE 2.7 – étapes de Classification [15].

2.7.3.1 Définition des Classes

La première étape dans la conception d'un algorithme de classification consiste à définir les catégories ou étiquettes que l'on souhaite attribuer aux documents en fonction de notre besoin et contexte[14].

2.7.3.2 Vectorisation

Les algorithmes d'apprentissage automatique ne comprennent pas le sens d'un texte directement car ils ne fonctionnent pas comme le cerveau humain. Pour ces algorithmes, un texte n'est qu'une suite de caractères binaires. Pour que l'algorithme puisse comprendre le texte, il est nécessaire de le transformer en un vecteur numérique. Ce processus est appelé "vectorisation" et permet de représenter le texte sous forme de données numériques que l'algorithme peut utiliser pour apprendre à classifier les documents en fonction des catégories définies [14].

- Bag of Word / Bag of ngrams (Méthode par Fréquence)

Il s'agit d'un réseau de neurones qui utilise une méthode simple pour transformer un texte en un vecteur[16]. Cette méthode se base sur l'apparition ou le nombre d'apparitions de mots dans un vocabulaire prédéfini. Par exemple, si le vocabulaire choisi est composé de 'man', 'woman', 'boy', 'girl', 'prince', 'princess', 'queen', 'king', 'monarch', alors le mot "prince" serait représenté par le vecteur [0,0,0,0,1,0,0,0,0].

Pour transformer une phrase, on compte le nombre d'apparitions des mots clés définis dans le vocabulaire. Par exemple, la phrase "This man is not the prince, it is the king" serait représentée par le vecteur [1,0,0,0,1,0,0,1,0], et la phrase "Boy, boy, boy" serait représentée par le vecteur [0,0,3,0,0,0,0,0,0].

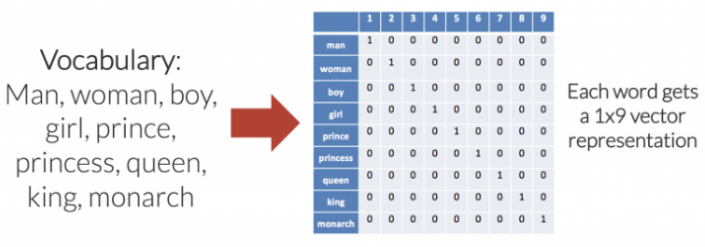

FIGURE 2.8 – Bag of Word [16].

2.7.3.3 Mise en place du modèle de catégorisation

Une fois les caractéristiques extraites, plusieurs algorithmes d'apprentissage peuvent être utilisés pour la classification automatique de textes. Pour citer les plus connus[15] :

- k-NN (k-Nearest Neighbour)

  Le k-NN est une méthode de catégorisation automatique qui représente chaque texte dans un espace vectoriel. L'algorithme ne repose pas sur des prototypes de catégorie, mais compare chaque nouveau texte avec l'ensemble des textes du jeu d'apprentissage pour déterminer la catégorie la plus proche en termes de similarité. La catégorie choisie sera celle qui contient en moyenne le plus de textes voisins[15].

Algorithm 1 Classification par k-PPV

Data: un échantillon de l textes classés en C = c1, c2, ..., cn classes

Input: k : le nombre de voisins à considérer pour la classification

Input: t : le texte à classer

Output: c : la classe attribuée au texte t

for *chaque texte l dans l* **do**

transformer le texte l en vecteur l = (x1, x2, ..., xm) calculer la distance entre t et l selon une métrique de distance

trier les textes de l par ordre croissant de leur distance à t sélectionner les k premiers textes de l compter le nombre d'occurrences de chaque classe parmi les k textes sélectionnés attribuer la classe la plus fréquente à t

- SVM (Support Vector Machine)

  SVM (Support Vector Machine) est une méthode de classification binaire utilisée en apprentissage supervisé. Elle a été introduite par Vapnik en 1995 et est considérée comme une alternative récente pour la classification. Cette méthode est basée sur des fondements théoriques solides et repose sur l'existence d'un classificateur linéaire dans un espace approprié. Pour résoudre des problèmes de classification à deux classes, SVM utilise un ensemble de données d'apprentissage pour apprendre les paramètres du modèle. Elle est également basée sur l'utilisation de fonctions noyau (kernel) qui permettent une séparation optimale des données[15].

  1. consiste à transformer les données en utilisant une transformation non linéaire pour les plonger dans un espace de grande dimension
  2. es classifieurs linéaires sont utilisés pour séparer les classes dans l'espace transformé en maximisant la marge, qui est la distance entre les classes.
  3. Les vecteurs supports, un nombre limité de points, sont utilisés pour déterminer les hyperplans qui définissent la frontière entre les classes[15].

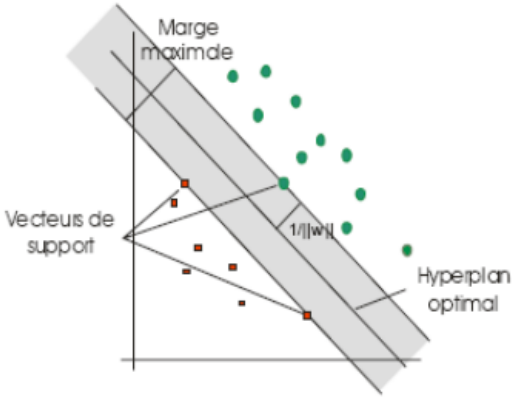

FIGURE 2.9 – SVM (Support Vector Machine) [15].

2.7.3.4 Prédiction

La phase de prédiction implique l'utilisation du modèle d'apprentissage créé lors de la phase 3 pour classifier les textes ou les documents que l'on souhaite traiter[14].

2.8 similarité entre les sites web

En général, une mesure de similarité est une fonction qui permet de mesurer la relation entre deux objets lorsqu'ils sont comparés entre eux. Elle permet de quantifier le degré de similitude ou de ressemblance entre ces deux objets[21]. La figure 2.10 illustre un processus systématique de calcul du rapport de similarité.

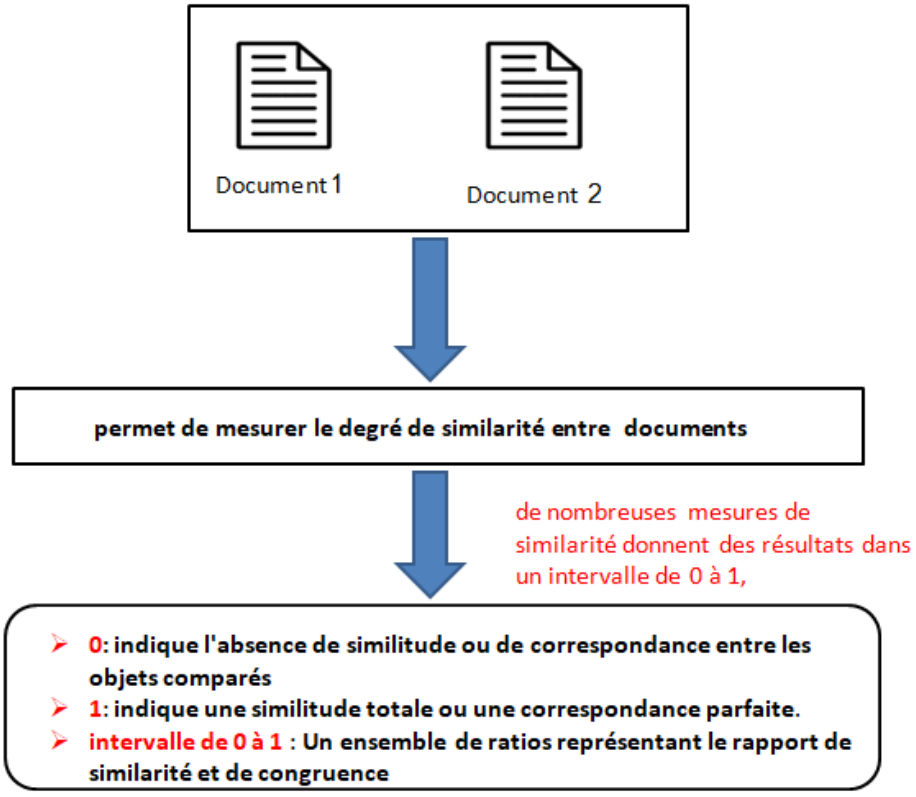

FIGURE 2.10 – processus de calcul du rapport de similarité.

2.8.1 La similarité cosinus

La similarité cosinus est une mesure de similarité très couramment utilisée dans l'analyse de texte et le traitement du langage naturel. Elle permet de mesurer le degré de similarité entre deux vecteurs en considérant l'angle formé entre ces vecteurs dans un espace vectoriel. La formule mathématique de la similarité cosinus est la suivante [21,22] :

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{2.3}$$

où A et B sont deux vecteurs de dimensions n, et est l'angle formé entre ces vecteurs. Le produit scalaire A . B représente la somme des produits de chaque composante des deux vecteurs. Les normes ||A|| et ||B|| représentent la racine carrée de la somme des carrés de chaque composante des vecteurs A et B, respectivement. La valeur de la similarité cosinus varie entre -1 (vecteurs opposés) et 1 (vecteurs identiques), avec une valeur de 0 indiquant une indépendance linéaire

entre les vecteurs .figure 2.11 illustre un processus de calcul du rapport de La similarité cosinus.

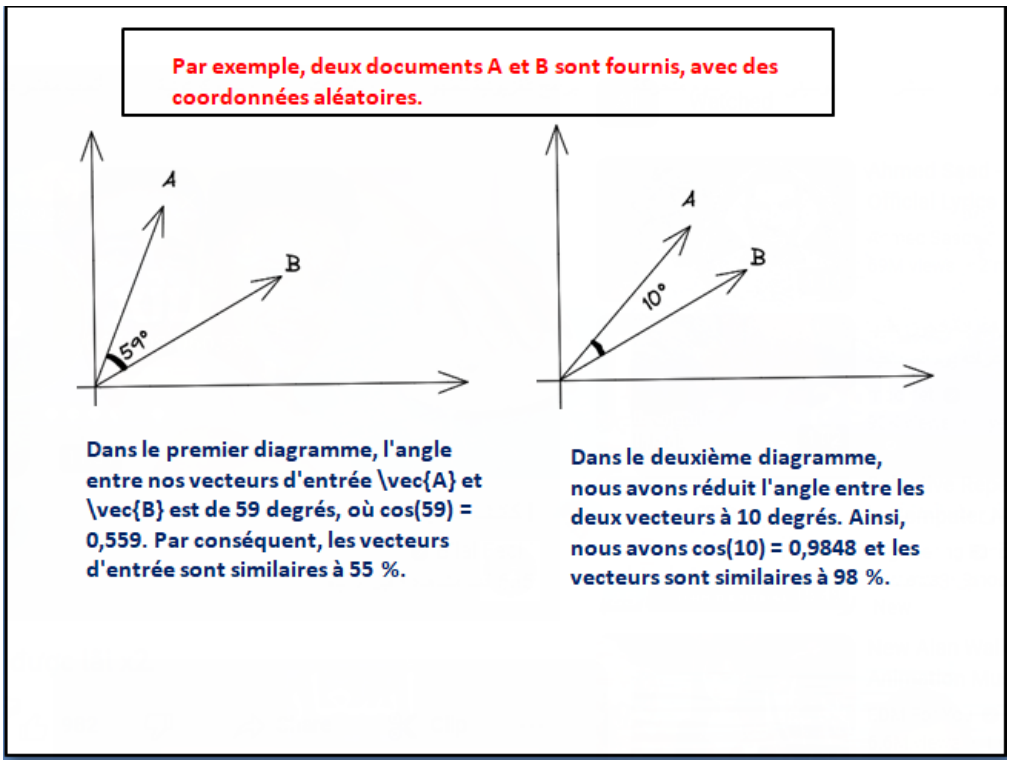

FIGURE 2.11 – La similarité cosinus [40].

2.8.2 L'indice de Dice

L'indice de Dice peut également être utilisé pour mesurer la similarité entre deux textes. Pour ce faire, chaque texte est d'abord divisé en ensembles de mots, puis l'indice de Dice est calculé sur la base de ces ensembles de mots. La formule pour calculer l'indice de Dice pour deux textes A et B est la suivante [21,22] :

$$Dice(A,B) = \frac{2\,|A \cap B|}{|A| + |B|} \tag{2.4}$$

où A et B sont les ensembles de mots extraits des textes A et B, et left| A right| représente le nombre de mots dans l'ensemble A. L'indice de Dice renvoie une valeur comprise entre 0 et 1, où une valeur de 1 indique une similitude parfaite entre les textes A et B, tandis qu'une valeur de 0 indique une absence de similitude entre les deux textes.figure 2.12 illustre un processus de calcul du rapport de La similarité cosinus.

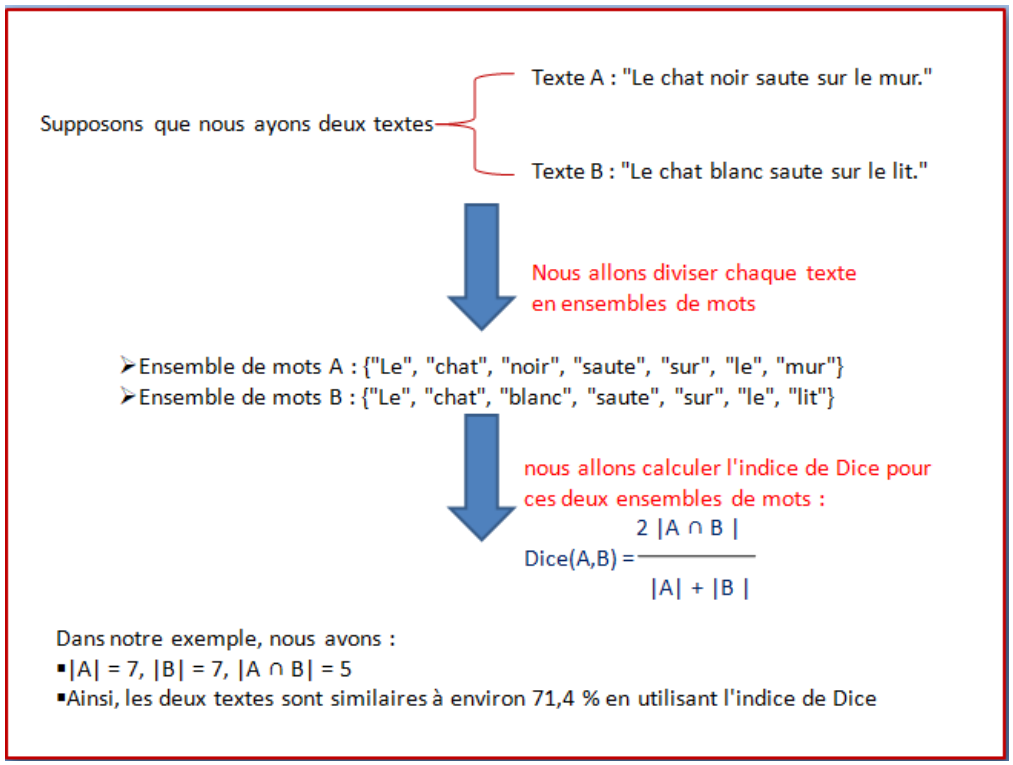

FIGURE 2.12 – L'indice de Dice.

2.9 Conclusion

La gestion de flux web fait référence à la manière dont les informations sont organisées, traitées et diffusées sur le web. Il s'agit d'un processus complexe qui implique la collecte, la transformation et la distribution de données numériques en temps réel. Les flux web sont utilisés dans une variété de contextes, tels que la diffusion de nouvelles en ligne, la surveillance des médias sociaux et la collecte de données sur les tendances du marché.

La veille sur la gestion de flux web est une activité cruciale pour les entreprises qui souhaitent suivre les tendances et les développements en temps réel sur le web et sur la vielle. Elle permet aux entreprises de collecter, d'analyser et de diffuser des informations pertinentes, de prendre des décisions éclairées et de maintenir leur position concurrentielle

Chapitre 3

La veille technologie

3.1 Introduction

La veille est un processus qui vous permet de surveiller et de recueillir des informations en ligne sur un sujet précis de fournir une veille stratégique pour les entreprises les organisations et les particuliers Les clients seraient en mesure de rester informé(e) des dernières actualités, des événements et les réglementations propres à un domaine particulier De nombreux outils, notamment les moteurs de recherche, les agrégateurs d'informations, les alertes d'actualités, les réseaux sociaux et les blogs, peuvent être utilisés pour surveiller Internet. Selon les demandes de l'utilisateur, des processus manuels ou automatisés peuvent être utilisés. Dans ce chapitre nous parlerons de Définitions, fonctions et moyensde la veille.

3.2 1 Qu'est-ce que la veille ?

3.2.1 Définitions de la veille

peut être défini comme l'activité d'observation active et systématique d'un environnemental organisé ou d'une personne, dans le but de collecteur, d'analyseur et de diffuseur d'informations pertinentes pour prendre des décisions éclairées ou anticiper les évolutions futures. Cette activité peut être formalisée et réglementée, et elle peut concerner divers domaines tels que la concurrence, les évolutions technologiques, les tendances du marché, la réglementation, etc. La veille consiste à surveiller régulièrement un sujet donné pour collecter et analyser des informations pertinentes provenant de différentes sources. Les nouvelles technologies de l'information,

notamment Internet, ont grandement facilité cette activit [23].

3.2.1.1 La veille documentaire

La veille documentaire implique la surveillance active de sources de documents pertinents pour une entreprise ou une organisation, telles que les bases de données, les bibliothèques et les sites web spécialisés. Son but est de trouver des documents utiles et pertinents pour l'entreprise et de les rassembler pour une utilisation à la demande[24].

3.2.1.2 la veille informationnelle

La veille informationnelle est un processus automatisé permettant de surveiller les publications récentes dans les domaines d'intérêt de l'utilisateur, tels que la science, la technologie, l'économie, le droit et les affaires. Elle permet de gagner du temps et d'avoir des informations actualisées. Pour éviter d'être submergé d'informations peu pertinentes, il est important d'être précis dans les critères de sélection de la veille, appelé "profil de recherche"[25].

3.2.2 Types de veille

Il n'y a pas une veille mais plusieurs veilles segmentées. De ce fait, la veille stratégique en regroupe différents types tels que[26] :

FIGURE 3.1 – les différents types de veille[26].

3.2.2.1 la veille concurrentielle

La veille concurrentielle consiste à surveiller les concurrents actuels ou potentiels d'une entreprise, en collectant des informations sur leurs produits, services, évolutions, rachats, etc. Cela permet à l'entreprise de mieux comprendre leur positionnement sur le marché et d'identifier les opportunités et les menaces.

3.2.2.2 la veille commerciale

La veille commerciale consiste à surveiller les besoins des clients, les évolutions du marché, les tendances et les stratégies des concurrents pour adapter les produits ou services de l'entreprise en conséquence et rester compétitif.

3.2.2.3 la veille technologique

La veille technologique consiste à identifier les nouveautés et innovations dans son domaine d'activité, tant en termes de produits, de services, de connaissances, de technologies informatiques, d'outils numériques, etc.

3.2.2.4 la veille juridique et réglementaire

Le processus d'examen de la législation (lois et décrets) permet d'anticiper les impacts sur l'activité de l'entreprise. Cette pratique est cruciale pour respecter les réglementations et éviter les risques juridiques

3.2.2.5 la veille ereputation

Le suivi de l'eréputation consiste à surveiller en permanence les informations publiées en ligne concernant une marque, telles que les avis, les témoignages et les commentaires sur les réseaux sociaux. Cette pratique permet de comprendre la perception de la marque par les clients et de réagir rapidement en cas de problème.

3.3 La veille digitale

La veille digitale est une pratique consistant à effectuer une surveillance permanente et méthodique du web, en vue de collecter et d'analyser des informations pertinentes sur un sujet spécifique. Elle permet de se tenir informé des dernières actualités, des tendances émergentes, des opinions des utilisateurs et des réactions du marché en temps réel. Cette pratique est particulièrement utile pour les entreprises qui cherchent à anticiper les changements du marché et à prendre des décisions stratégiques éclairées Voici les différentes sources complémentaires qui peuvent être suivies dans le cadre d'une veille numérique[27] :

- Résultats de recherche sur les principaux moteurs de recherche tels que Google.
- Nouvelles publications sur les réseaux sociaux majeurs tels que Facebook, Twitter et Reddit.
- Grands agrégateurs d'informations, tels que Google Actualités.
- Toutes sortes de sources disponibles sur Internet, telles que des blogs spécialisés ou des communications d'entreprises concurrentes.

3.4 Le processus de veille :

Le processus de veille se compose habituellement de quatre étapes principales[28,29] :

1. La définition des besoins d'information
2. La recherche et la collecte d'informations
3. Le traitement de l'information
4. La diffusion de l'information.

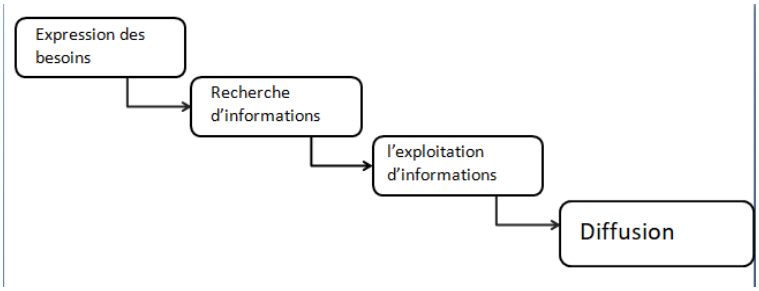

FIGURE 3.2 – Le processus de veille[28].

3.4.1 La définition des besoins d'information

La première étape du cycle de la veill econsiste à identifier les besoins en informations est couramment appelée "ciblage",est cruciale pour la mise en place d'un système de veille efficace en informatique.Son objectif est de délimiter le champ et la direction de la surveillance, ainsi que de classer les domaines ou axes de veille selon leur importance pour le système d'information , La mise à jour régulière des besoins en informations est essentielle pour assurer la pertinence et l'efficacité du système de veille en informatique[29].

3.4.1.1 L'étendue de la veille

L'étendue de la veille correspond au champ d'application de la surveillance, c'est-à-dire l'ensemble des sources et des domaines à surveiller. L'orientation de la veille, quant à elle, fait référence à l'objectif de la surveillance et à la manière dont l'information collectée sera utilisée pour répondre aux besoins de l'entreprise. En d'autres termes, l'étendue et l'orientation de la veille déterminent la portée et l'efficacité du système de veille en matière de collecte et d'utilisation des informations.

3.4.1.2 La hiérarchisation des axes de veille

La hiérarchisation des thèmes de suivi consiste à catégoriser les domaines ou sujets de suivi selon leur importance pour le système d'information. Cette étape permet de hiérarchiser les zones de suivi en fonction de leurs enjeux stratégiques et opérationnels, et d'orienter les efforts de collecte et de traitement des informations. Il permet également de cibler les sources d'information les plus pertinentes et les plus utiles.

3.4.1.3 La mise à jour de l'identification des besoins

La mise à jour de l'identification des besoins en information est une étape essentielle dans le processus de veille d'un système d'information. Aide à maintenir la pertinence et l'efficacité du suivi en s'adaptant aux changements et aux besoins d'information.

3.4.2 La recherche et la collecte informations

Il est à ce stade du processus crucial de se poser la question de la nature, de la diversité, de la quantité et de la qualité des sources d'information et des informations collectées[30,31].

3.4.2.1 La nature de l'information

Il est important de rappeler qu'avant de se pencher sur les différentes sources d'information, il existe traditionnellement trois niveaux d'accès distincts :

1. L'information dite "blanche" constitue 80 de l'ensemble de l'information disponible. Elle est facilement accessible de manière licite et ne nécessite aucune mesure de sécurité particulière, pouvant être recherchée avec des outils grand public.
2. Quant à l'information dite "grise", elle représente 15 de l'ensemble de l'information disponible, et bien qu'elle soit accessible de manière licite, sa connaissance et son accès peuvent être difficiles. L'accès à cette information nécessite donc des techniques avancées de recherche et de traitement de l'information.
3. l'information dite "noire" ne représente que 5 de l'ensemble de l'information disponible et se caractérise par une diffusion limitée et un accès explicitement protégé. Son acquisition ne peut être réalisée que par le biais d'actions d'espionnage.

3.4.2.2 Les sources d'information

Il est possible d'identifier deux catégories de sources : les sources formelles et les sources informelles.

— Les sources formelles Les sources d'information formelles peuvent être classées en sources externes ou internes. Les sources externes, qui sont accessibles publiquement, comprennent la presse, les livres, les médias, les rapports d'activité des entreprises, les bases de données et les études. Les sources internes, quant à elles, regroupent les informations accessibles via l'intranet de l'entreprise, telles que les monographies de concurrents, les études diverses et les lettres d'informations.

— Internet est devenu une source primordiale pour toute entreprise souhaitant effectuer de la veille. Cependant, il est essentiel de prendre en compte quatre caractéris-

tiques des sources disponibles sur le web :

- La diversité des sources : elon Google, il existe plus de 200 millions de sites web et 1000 milliards de pages sur Internet, ce qui montre une grande diversité de sources disponibles en ligne.
- La qualité de l'information en ligne peut être incertaine, car tout le monde peut publier sur Internet sans qu'il y ait de validation préalable.
- La stabilité de l'information : le contenu des sites évolue, des sources peuvent disparaître.
  - Le Web visible L'ensemble des pages facilement accessibles par les moteurs de recherche généralistes est considérablement plus restreint que l'ensemble de l'Internet dans son ensemble. En effet, il est estimé à environ 10 milliards de pages, avec un taux de croissance d'environ 1,5 million de documents ajoutés quotidiennement.
  - Le Web invisible Le web profond, englobe la partie de l'Internet qui n'est pas directement accessible via les moteurs de recherche classiques. Cette partie comprend des bases de données, des bibliothèques en ligne (gratuites ou payantes), des sites accessibles uniquement via un mot de passe, des portails sectoriels, des sites institutionnels, des publications en ligne et des réseaux sociaux.

- Les sources informelles Les sources informelles désignent les informations ou les données qui ne sont pas produites ou validées par des sources officielles ou reconnues comme fiables. Ces sources peuvent inclure des blogs personnels, des forums en ligne, des réseaux sociaux et d'autres sources similaires. Bien qu'elles puissent fournir des informations utiles, il est important de prendre en compte leur fiabilité et leur crédibilité avant de les utiliser pour la prise de décision ou la recherche.

3.4.3 Le traitement de informations

Il ne suffit pas de simplement rechercher et collecter de l'information. Cette dernière doit être traitée, c'est-à-dire analysée, synthétisée et mise en forme afin d'être exploitée efficace-

ment[29].

3.4.3.1 L'analyse des informations

L'analyse des informations consiste à examiner les sources collectées pour extraire les éléments pertinents en réponse à une problématique donnée, en utilisant des techniques et des outils tels que les tableaux, les graphiques, les statistiques et les modèles de données. Elle peut inclure l'examen de la qualité et de la fiabilité des sources, la comparaison de différentes perspectives et points de vue, la recherche de corrélations et de tendances, et l'identification de relations de cause à effet[29].

3.4.3.2 La synthèse des informations

La synthèse consiste à transformer une grande quantité d'informations brutes ou interprétées en un ensemble cohérent et concis. Elle résume la problématique et met en évidence les éléments clés du sujet traité, tout en intégrant le point de vue personnel de l'auteur[29].

3.4.4 La diffusion de l'information

Après la collecte et le traitement de l'information, il est important de la diffuser aux utilisateurs concernés. Les responsables de la veille doivent alors répondre à quatre questions clés : à qui l'information doit-elle être diffusée ? à quel moment ? de quelle manière ? et à travers quels canaux de communication ? Il est également crucial de prendre en compte les éventuels obstacles à la circulation et à la diffusion de l'information, et de mettre en place des stratégies pour les surmonter[29].

3.5 Système informatique de veille

Un système informatique de veille est un système automatisé qui surveille et analyse des sources d'informations sélectionnées en temps réel ou périodiquement, afin de fournir des informations utiles pour la prise de décision stratégique et opérationnelle dans un domaine d'intérêt donné. Les informations collectées sont présentées sous forme de rapports, d'alertes ou de

tableaux de bord pour aider les utilisateurs à prendre des décisions éclairées en temps opportun.Il existe de nombreux exemples de systèmes informatiques de veille, dont voici quelques-uns [39] :

3.5.1 Twitter :

Twitter est une plateforme de médias sociaux qui peut être utilisée comme source d'informations pour la veille. Les utilisateurs peuvent surveiller des hashtags, des comptes et des mots-clés spécifiques pour suivre les tendances, les opinions et les conversations en temps réel.

3.5.2 Facebook :

Facebook est également une plateforme de médias sociaux qui peut être utilisée pour la veille. Les utilisateurs peuvent surveiller des pages, des groupes et des comptes pour suivre les tendances et les conversations.

3.5.3 Feedly :

Feedly est un agrégateur de contenu qui permet aux utilisateurs de rassembler des informations provenant de différentes sources (blogs, sites d'actualités, réseaux sociaux, etc.) en un seul endroit pour une consultation facile et rapide. Il peut être utilisé pour surveiller les tendances et les sujets pertinents.

3.5.4 Pinterest :

Pinterest est une plateforme de partage de photos et d'images qui peut également être utilisée pour la veille. Les utilisateurs peuvent créer des tableaux thématiques pour rassembler des images et des liens pertinents pour surveiller les tendances et les sujets d'intérêt.

3.5.5 Pocket :

Pocket est une application de gestion de signets qui permet aux utilisateurs de sauvegarder des articles, des vidéos et des pages Web pour une consultation ultérieure. Il peut être utilisé

pour surveiller les tendances et les sujets pertinents.

3.5.6 Netvibes :

Netvibes est une plateforme de veille en temps réel qui permet aux utilisateurs de surveiller les tendances, les médias sociaux, les actualités et les données de l'entreprise en temps réel, en utilisant des tableaux de bord personnalisables.

3.5.7 Flipboard :

Flipboard est un agrégateur de contenu qui permet aux utilisateurs de créer des magazines personnalisés en rassemblant des articles, des images et des vidéos provenant de différentes sources. Il peut être utilisé pour surveiller les tendances et les sujets pertinents.

3.5.8 Google Alerts :

Google Alerts est un outil de surveillance d'Internet qui permet de suivre des mots-clés spécifiques et de recevoir des alertes par e-mail lorsque de nouveaux contenus sont publiés en ligne.

3.5.9 Scoop.it :

Scoop.it est une plateforme de curation de contenu qui permet aux utilisateurs de créer des magazines en ligne personnalisés en rassemblant des articles, des vidéos, des images et des infographies à partir de différentes sources. Les utilisateurs peuvent également surveiller des sujets spécifiques pour suivre les tendances et les conversations.

3.6 Conclusion

Au cours de ce chapitre, nous avons présenté une définition complète du processus de veille, en expliquant ses différentes typologies ainsi que ses étapes clés. De plus, nous avons abordé

certaines des caractéristiques des systèmes de veille de l'information les plus couramment utilisés. Dans le prochain chapitre, nous aurons l'occasion de mettre en pratique ces connaissances en créant un système de veille de l'information, dédié à la gestion du flux web.

Chapitre 4

Conception du système de gestion de flux web

4.1 l'introduction

La phase de conception conceptuelle est considérée comme l'étape la plus cruciale d'un projet informatique, car elle vise à déterminer les choix d'informations et de traitements à prendre en compte dans le système d'information. Dans cette section, nous examinons la structure générale du système et procédons à une brève définition de ses différentes unités. Nous discuterons également des diagrammes UML utilisés pour modéliser le système.

4.2 Architecture du système proposé

Nous avons proposé un système de gestion de flux web et de récupération d'informations pertinentes, car il dépend de la comparaison entre la requête représentée par des mots-clés et l'indexation, et il se compose de plusieurs étapes, la phase de demande d'information est cruciale pour assurer la pertinence des résultats retournés à l'utilisateur. Cette étape implique une analyse requête en deux étapes distinctes, l'analyse lexicale qui consiste à identifier les mots-clés de la requête, et l'analyse sémantique qui permet de comprendre le sens de la demande en utilisant des techniques de traitement du langage naturel. Par la suite, le système recherche les données liées à la demande d'information en utilisant une base de données Internet. Les documents sont stockés dans cette base de données grâce à un processus d'indexation qui per-

met d'attribuer des mots-clés à chaque document. Cette étape facilite l'accès à l'information recherchée en réduisant le temps de recherche et en augmentant la précision des résultats retournés. Ainsi, le processus d'indexation joue un rôle crucial dans la performance et l'efficacité du système de recherche d'information.Tel qu'illustré dans la figure 4.1 suivante, les composants peuvent être représentés schématiquement..

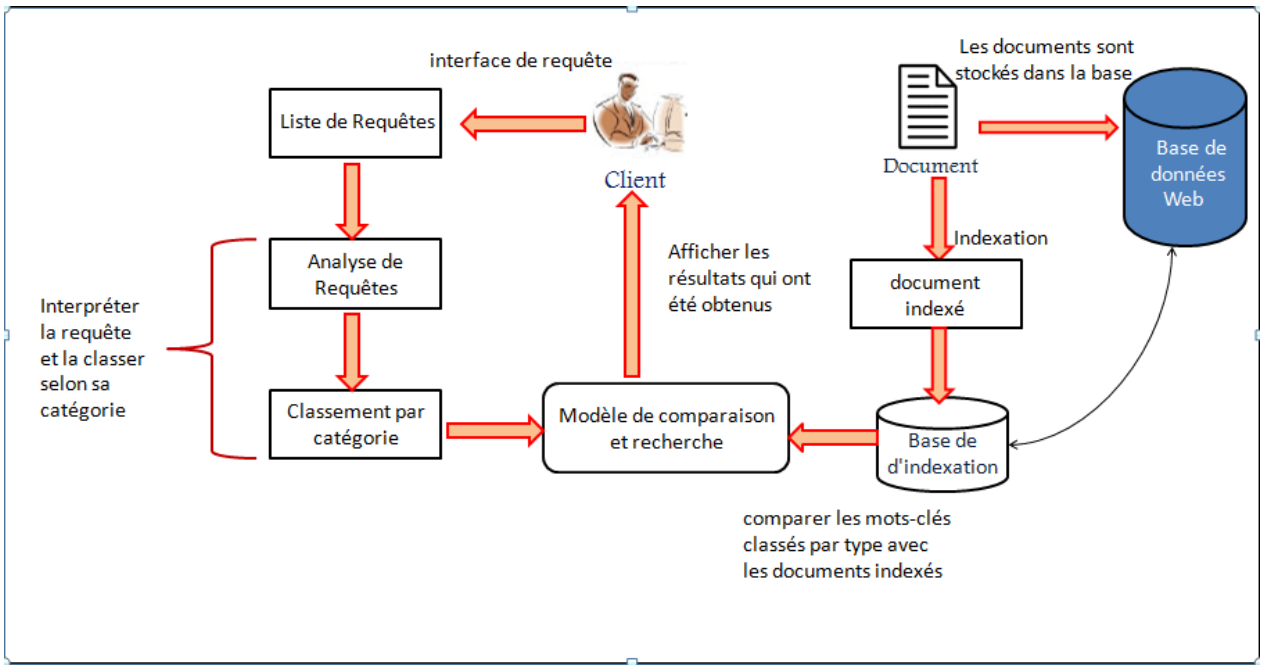

FIGURE 4.1 – Architecture du système proposé .

Le système se compose de :

— Requête (Demande d'information) : Une demande formulée par un utilisateur pour obtenir des informations spécifiques. Elle peut prendre la forme d'une question, d'une recherche ou d'une demande de clarification.

— Analyse requête : Le processus d'examen et de compréhension de la demande d'information pour déterminer son objectif et ses éléments clés. Cela peut impliquer l'identification des mots-clés, la suppression des mots inutiles et l'analyse syntaxique.

— Classement par catégorie : Le regroupement des requêtes ou des informations en catégories ou en thèmes spécifiques. Cela permet d'organiser les données et de faciliter la recherche et l'accès ultérieurs.

— L'indexation de documents : Le processus de création d'un index structuré et organisé des documents, des pages Web ou d'autres sources d'informations. Cela permet de faciliter la recherche et l'accès ultérieurs en optimisant la vitesse et la précision des recherches.

— Modèle de comparaison et Recherche : Un cadre ou une méthode utilisée pour comparer et trouver des informations pertinentes par rapport à une requête donnée.

— Résultats de recherche : Les informations ou les documents pertinents qui sont retournés en réponse à une requête de recherche donnée. Ces résultats sont généralement classés selon leur pertinence ou leur adéquation par rapport à la requête initiale.

4.2.1 Requête (Demande d'information)

La requête est l'expression du besoin de l'utilisateur en matière d'information. Elle représente l'interface entre le système d'information et l'utilisateur[32], le terme "requête" est également utilisé pour décrire les termes de recherche saisis dans un moteur de recherche. Par exemple, il est possible de trouver des articles traitant des requêtes les plus fréquentes sur Internet .

4.2.2 Analyse lexicale

Nous allons également aborder quelques problèmes linguistiques liés à la recherche d'informations, tels que la radicalisation et la lemmatisation, la tokenisation, les mots vides et la normalisation.

4.2.2.1 La tokenisation

La tokenisation est la division d'un requête en unités de sens appelées tokens. Cela permet de convertir le requête en une forme que le système peut traiter dans le cadre du traitement du langage naturel. Les tokens peuvent être des mots, des phrases ou des symboles de ponctuation [33]. Voici un exemple4.2 de tokenisation

FIGURE 4.2 – exemple de tokenisation [34].

4.2.2.2 Les mots vides

Les mots vides, également appelés Les stop words, sont des mots couramment utilisés dans une langue tels que "le, la, les" en français, et "the, of" en anglais,", , " en Arab . Ces mots sont souvent ignorés lors de la recherche d'informations car ils sont très fréquents et n'ont pas beaucoup de valeur sémantique en eux-mêmes . L'exclusion de ces mots permet de gagner du temps et et de rendre les résultats de recherche plus pertinents [35]. Tel qu'illustré dans la figure4.3,

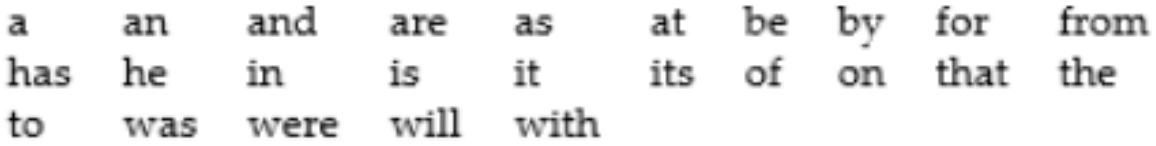

FIGURE 4.3 – exemple de Les mots vides[34].

4.2.3 Analyse sémantique

L'analyse sémantique est une technique courante en traitement automatique du langage naturel utilisée pour comprendre la signification et le sens des textes. Cette méthode consiste à indexer des documents avec des mots clés pertinents, puis à analyser leur contenu en comparant les mots clés utilisés avec ceux qui ont été indexés. De cette manière, elle permet de déterminer la pertinence des documents en réponse à une requête donnée et de les classer en conséquence. Les moteurs de recherche utilisent largement cette technique pour fournir des résultats pertinents aux requêtes des utilisateurs [36]. L'analyse sémantique est une technique qui permet d'analyser le sens d'un texte en se concentrant sur les intentions, les ressentis et les

émotions qui dictent le sens d'un message. Contrairement à l'analyse syntaxique qui s'intéresse à la structure grammaticale d'une phrase, l'analyse sémantique se focalise sur le sens des mots utilisés Il est important de noter les différences entre l'analyse syntaxique et l'analyse sémantique[37] :

— L'analyse syntaxique se concentre sur la structure grammaticale d'une phrase et la relation entre les mots qui la composent. Elle s'intéresse donc à la "forme" du texte.

— 'analyse sémantique s'intéresse au sens des mots et des expressions utilisés, ainsi qu'aux intentions, émotions et ressentis qu'ils véhiculent. Elle se concentre donc sur le "fond" du texte.

Il convient de souligner que ces deux types d'analyse sont complémentaires et permettent de mieux comprendre et interpréter un texte.

4.2.4 L'indexation de documents

Un système de recherche d'informations (IRS) utilise l'indexation pour convertir les requêtes et les documents en une représentation intermédiaire pour faciliter la recherche sémantique. L'indexation consiste à extraire les termes représentatifs du contenu sémantique d'un document ou d'une requête et à les disposer dans une structure appelée dictionnaire. Le résultat de l'indexation est un descripteur qui peut être une liste de termes pondérés pour leur degré de représentativité du contenu sémantique. Les groupes de mots, également appelés thésaurus, peuvent être utilisés pour ajouter de la richesse sémantique à l'indexation. La qualité de l'indexation est essentielle pour obtenir des résultats de recherche de qualité[38]. Il existe trois méthodes principales d'indexation des documents : manuelle, automatique et semi-automatique :

- L'indexation manuelle : consiste à analyser chaque document par un spécialiste du domaine ou un documentaliste. Cette méthode peut être très précise mais elle est également très coûteuse en termes de temps et de ressources humaines.
- L'indexation automatique : utilise un processus entièrement automatisé pour analyser chaque document. Bien que cette méthode soit plus rapide et moins coûteuse que l'indexation manuelle, elle peut parfois être moins précise

- l'indexation semi-automatique : combine les avantages des deux méthodes précédentes. Le processus d'analyse automatique est utilisé pour générer une première liste de termes clés, mais le spécialiste du domaine ou le documentaliste intervient ensuite pour établir des relations sémantiques entre les mots clés et choisir les termes significatifs.

La fonction de pondération permet d'attribuer à chaque terme d'indexation une valeur qui mesure son importance dans le document où il apparaît. Le pouvoir de discrimination des termes pour décrire le contenu des documents n'est pas identique pour tous les termes. Pour trouver les termes du document qui représentent le mieux son contenu sémantique, on a défini la fonction de pondération d'un terme dans un document, connue sous le nom de Tf, Idf, qui est utilisée dans différentes versions par la majorité des systèmes de recherche d'information (IRS).

- TF (fréquence des termes) : cette mesure est proportionnelle à la fréquence du terme dans le document. L'idée sous-jacente est que plus un terme est utilisé dans un document, plus il est important dans la description de ce document [22] .
- Idf (inverse de la fréquence des documents) : mesure l'importance d'un terme dans toute la collection. L'idée sous-jacente est que les termes qui apparaissent dans peu de documents de la collection sont plus représentatifs du contenu de ces documents que ceux qui apparaissent dans tous les documents de la collectio[22].

4.2.5 Modèle de comparaison et Recherche

Les systèmes de recherche d'informations (IRS) intègrent un processus de recherche/décision qui vise à sélectionner les informations pertinentes pour l'utilisateur. Pour cela, une mesure de similarité (correspondance) entre la requête de l'utilisateur et les descripteurs des documents dans la collection est calculée. Seuls les documents dont la similarité dépasse un seuil prédéfini sont sélectionnés par l'IRS. La fonction de correspondance est un élément clé d'un SRI car la qualité des résultats dépend de la capacité du système à calculer une pertinence des documents aussi proche que possible du jugement de pertinence de l'utilisateur.

4.2.6 Résultats de recherche

subsectionRésultats de recherche Les résultats consistent en des documents textuels d'actualité qui présentent le titre de l'événement, le nom du site d'information et la date de publication de l'article.

4.3 Diagramme de cas d'utilisation du système

Le diagramme de cas d'utilisation est précis et permet de spécifier les principaux objectifs fonctionnels de l'application, ainsi que la relation entre ces objectifs et les utilisateurs. Dans une large mesure, la conception entière est basée sur ce diagramme4.4. Nous allons présenter deux acteurs, l'utilisateur et l'administrateur, interagissant avec le système de récupération d'informations. Par exemple, les fonctionnalités incluent l'inscription, la recherche d'informations et la gestion du système.

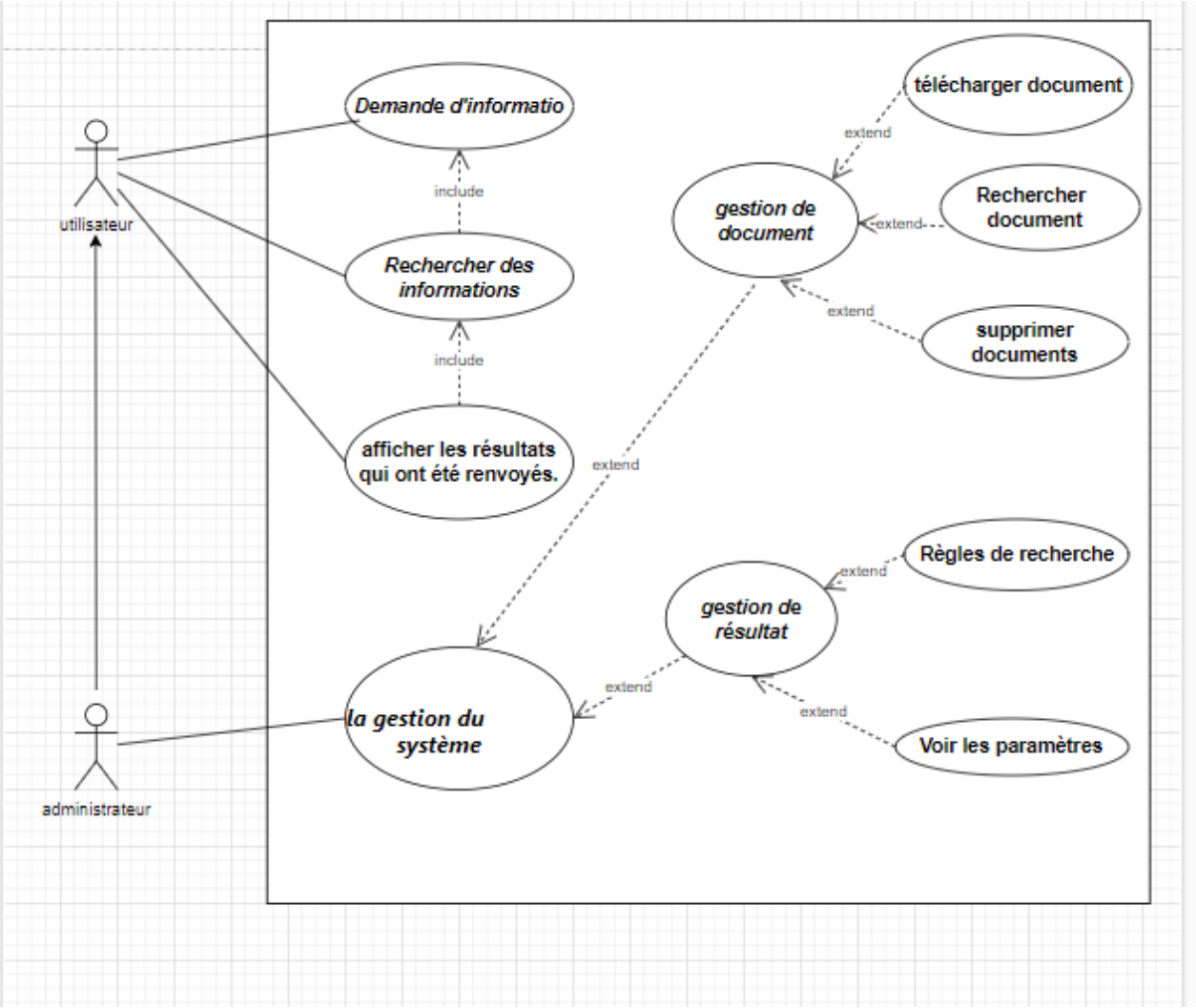

FIGURE 4.4 – Diagramme de cas d'utilisation (Activity Diagram :System Inforamtion retreival) .

4.4 Diagramme d'activité du système

Diagramme d'activité 4.5. pour la recherche d'une tâche d'entrée d'informations qui comprend :

— Vérification du document saisi.

— Traitement de la liste de mots.

— Calcul des paramètres.

— Calcul de la similarité

— affichage du résultat.

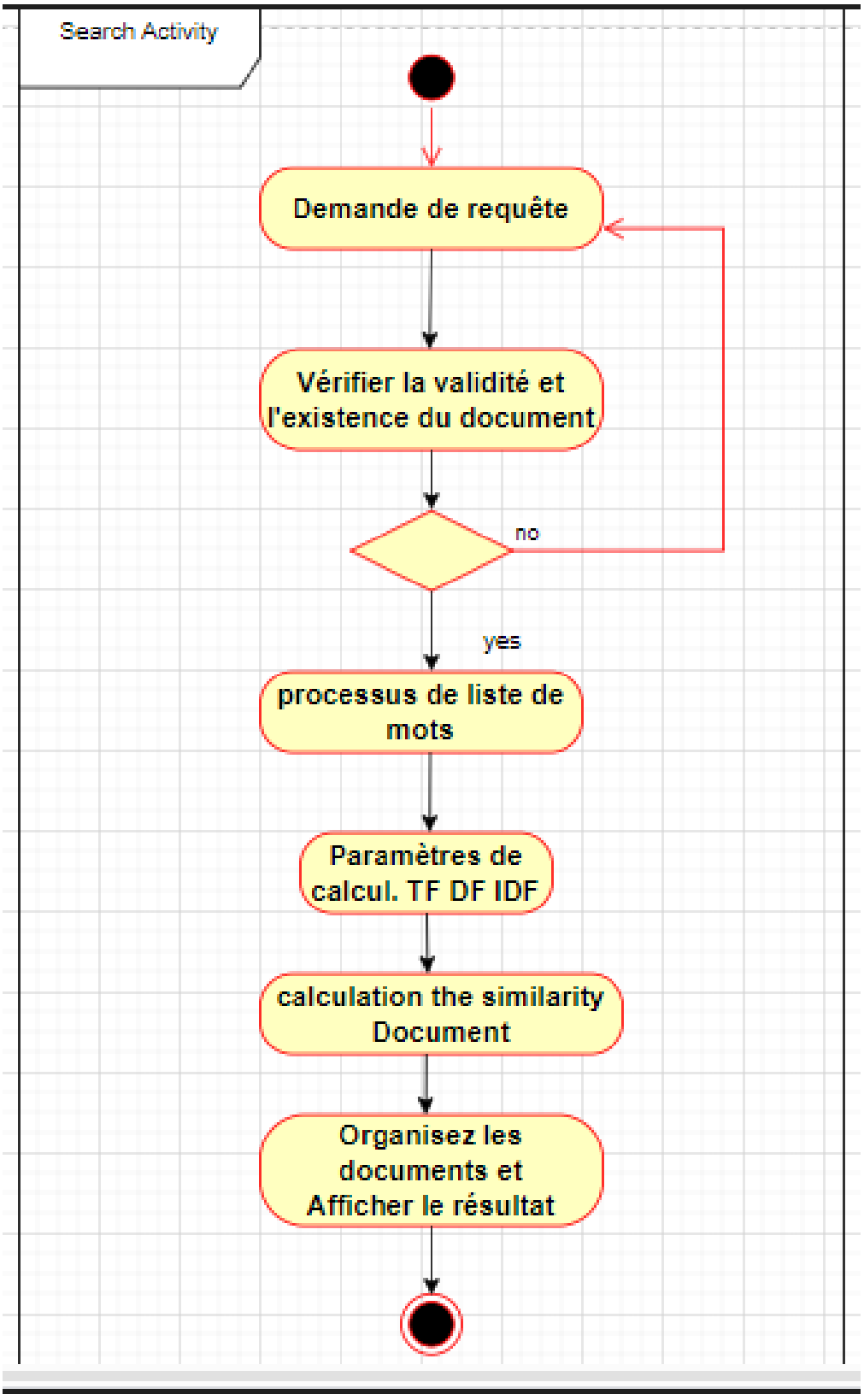

FIGURE 4.5 – Diagramme d'activité (Activity Diagram :System Inforamtion retreival) .

4.5 diagramme de séquence du système

Le processus de recherche d'informations de requête suit un diagramme 4.6de séquence qui débute par la vérification de la nature des termes saisis par l'utilisateur. Par la suite, une liste de mots est préparée à partir de la requête, et le poids Tf-Idf est calculé pour chaque mot de la liste. Enfin, la similarité est mesurée et les résultats sont affichés. Ce processus est couramment utilisé dans le domaine de la recherche d'informations et permet de fournir des résultats pertinents à l'utilisateur.

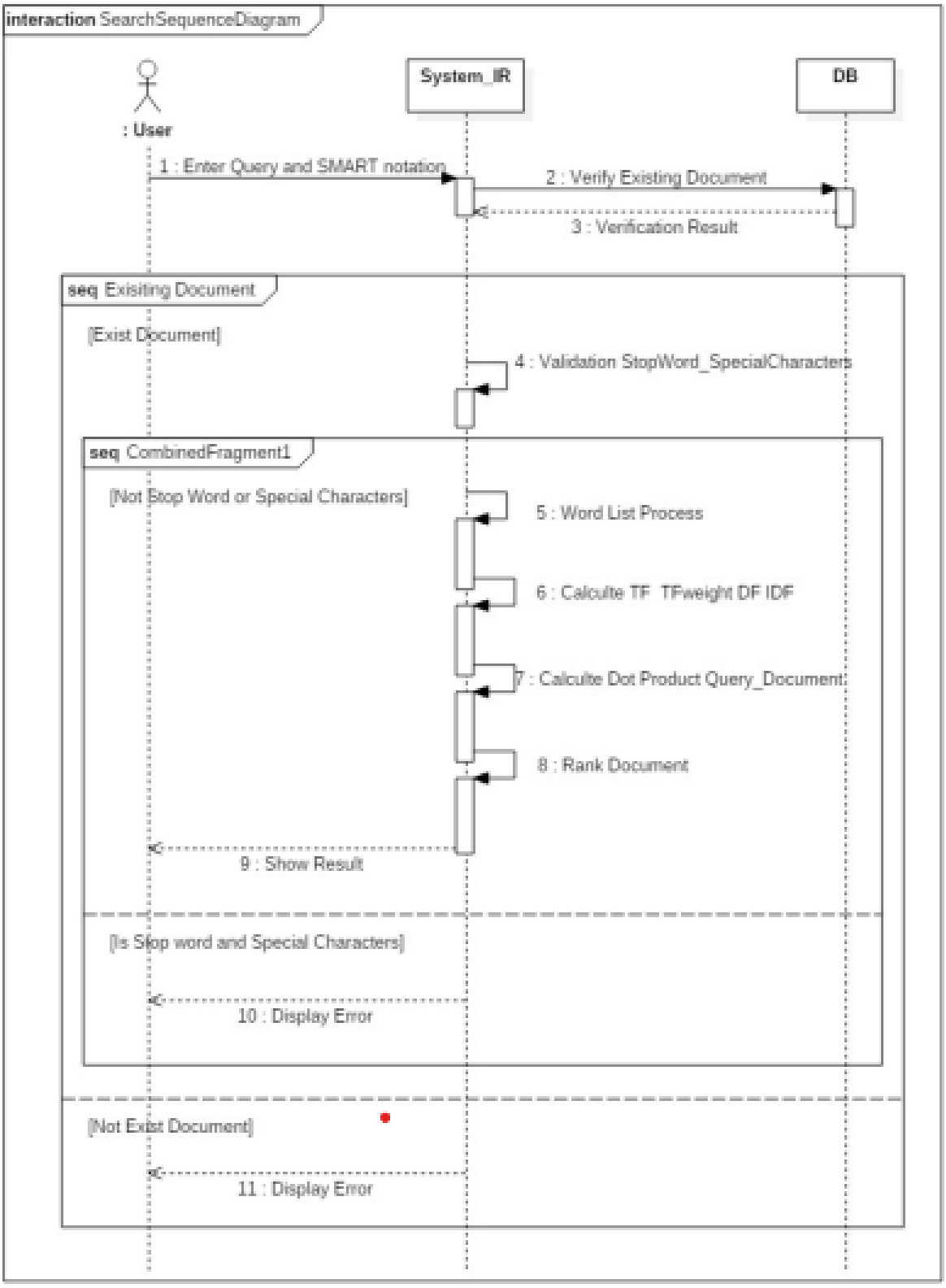

FIGURE 4.6 – diagramme de séquence .

4.6 Conclusion

Dans ce chapitre, nous avons révélé l'architecture de notre système proposé ainsi que défini les différents composants et fonctions du schéma proposé. Nous avons inclus des diagrammes UML pour illustrer ces éléments. Dans le chapitre suivant, nous mettrons en œuvre le programme et présenterons les résultats

Chapitre 5

Implémentation et résultats du système de gestion de flux web

5.1 l'introduction

Dans ce chapitre, nous présentons la mise en œuvre d'une application de gestion de flux web qui surveille et récupère les actualités pertinentes en fonction des requêtes des utilisateurs. Nous mettons en évidence les outils de développement utilisés, le framework de développement choisi, ainsi que la source de données utilisée. Enfin, nous présentons les résultats obtenus grâce à cette application.
Ce chapitre offre une description détaillée du processus de développement de notre application. Nous commençons par présenter les outils de développement que nous avons utilisés, tels que Visual Studio Code pour l'environnement de développement. Ensuite, nous expliquons le framework de développement que nous avons sélectionné, tel que Express.js, qui facilite la création d'un serveur web pour gérer les requêtes des clients. Nous décrivons également la source de données utilisée, qui peut être l'API NewsAPI pour obtenir des actualités en temps réel.

5.2 Architecture et composants de l'application

Concevoir et développer un site Web interactif avec intégration front-end et back-end à l'aide de JavaScript et du framework Express.js. Dans cette version du site Web, nous uti-

lisons le langage de modélisation JSON (JavaScript Object Notation) pour structurer et stocker les données nécessaires. Le fichier JSON joue un rôle maître dans l'organisation des informations et la communication entre le front-end et le back-end d'une application

FIGURE 5.1 – Architecture de l'application

5.2.1 Development environment :

— Operating system : Microsoft Windows10

— Tools of development : Visual Studio Code

— framework de développement : Express.js qui est un framework de développement d'applications web pour Node.js.

— sources des données : l'API News

— Application Server : serveur Express qui utilise l'API News pour fournir des données à l'application.

5.2.2 Visual Studio Code :

Visual Studio Code est un éditeur de code source léger mais puissant qui s'exécute sur votre ordinateur de bureau et est disponible pour Windows, macOS et Linux. Il offre une prise en charge intégrée de JavaScript, TypeScript et Node.js, et dispose d'un écosystème

riche en extensions pour d'autres langages et environnements d'exécution (comme C++, C, Java, Python, PHP, Go, .NET)[42].

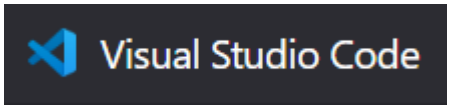

FIGURE 5.2 – logo de Visual Studio Code

5.2.3 Node.js

Node.js est une plateforme permettant de créer rapidement et efficacement des applications serveur évolutives en utilisant JavaScript. Node.js est l'environnement d'exécution, tandis que npm est le gestionnaire de packages pour les modules Node.js[43].

FIGURE 5.3 – logo de Node.js

5.2.4 Express.js :

Express est un cadre d'application web minimaliste et flexible pour Node.js. Il offre un ensemble robuste de fonctionnalités pour les applications web et mobiles. Avec ses nombreuses méthodes utilitaires HTTP et middleware, il est facile de créer une API solide. Express fournit également des performances optimales en offrant une couche mince de fonctionnalités essentielles, tout en préservant les fonctionnalités appréciées de Node.js. De plus, de nombreux frameworks populaires reposent sur Express, ce qui en fait un choix courant pour le développement d'applications web[44].

5.2.5 l'API News :

L'API News est une interface HTTP REST permettant de rechercher et de récupérer des articles en direct provenant de diverses sources sur le web[45].

Pour obtenir une clé (key) pour l'API News, vous devez suivre les étapes suivantes

— Rendez-vous sur le site web de l'API News.

— Remplissez le formulaire d'inscription en fournissant les informations requises, telles que votre nom, votre adresse e-mail et éventuellement des détails supplémentaires.

— Une fois votre inscription terminée, vous devriez recevoir un e-mail de confirmation contenant votre clé d'API.

— Utilisez cette clé dans vos requêtes pour accéder aux fonctionnalités de l'API News.

La première figure 5.4 représente la page d'accueil de l'API et La deuxième figure 5.5 représente le processus de récupération de la clé l'API News :

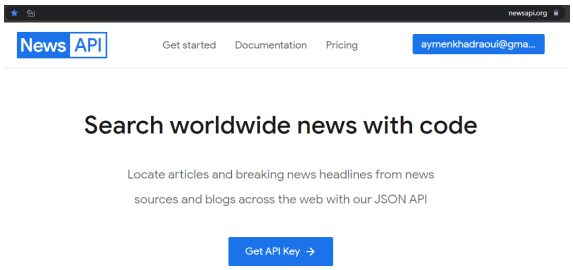

FIGURE 5.4 – La page d'accueil de l'API News

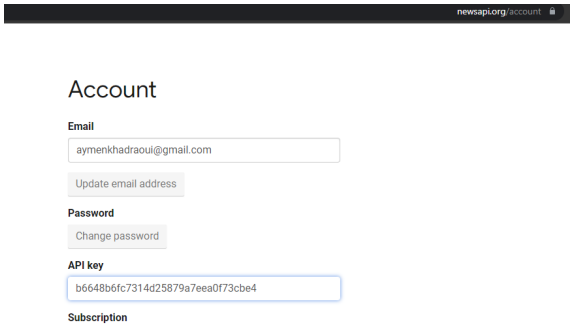

FIGURE 5.5 – la clé l'API News

5.3 Langages de programmation

5.3.1 HTML (Hyper Text Markup Language)

"HTML, acronyme de Hyper Text Markup Language, est le langage de balisage standard pour créer des pages Web. Il décrit la structure d'une page Web en utilisant une série d'éléments. Ces éléments HTML indiquent au navigateur comment afficher le contenu et

permettent d'étiqueter des éléments tels que "ceci est un titre", "ceci est un paragraphe", "ceci est un lien", etc."[43]

5.3.2 CSS(stands for Cascading Style)

C'est un langage de style utilisé pour décrire l'apparence d'un document écrit en HTML ou XML. CSS est responsable de la présentation visuelle d'une page Web, y compris des aspects tels que la mise en page, les couleurs, les polices et l'espacement[43].

5.3.3 JavaScript

JavaScript désigne un langage de développement informatique, et plus précisément un langage de script orienté objet. On le retrouve principalement dans les pages Internet. Il permet, entre autres, d'introduire sur une page web ou HTML des petites animations ou des effets. Il existe de nombreux frameworks JavaScript orientés vers les interfaces web (ou "orientés client"). Les trois plus connus sont Jquery, AngularJS (qui a été initialement développé par Google) et React (qui, lui, est né chez Facebook). Il existe néanmoins quelques infrastructures JavaScript open source orientées serveur, même si ce langage n'avait pas été conçu dans cette optique au départ. La plus célèbre d'entre elles n'est autre que NodeJS [43].

5.4 Structure et fonctions du programme

La figure 5.6 illustre la conception et l'architecture d'une application.chaque fichier de cette application est une représentation qui inclut des fonctions et des méthodes.

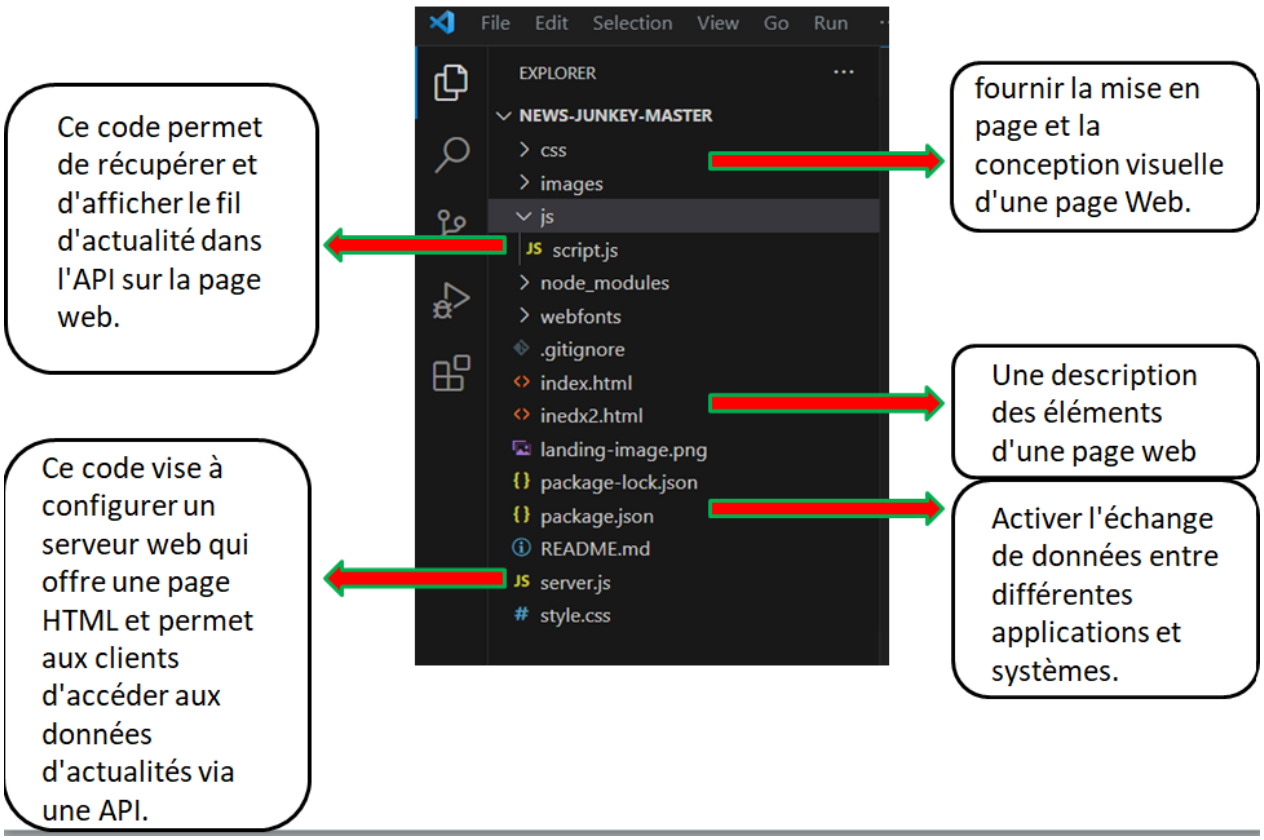

FIGURE 5.6 – l'architecture d'une application

5.4.1 Script.js

Le code fourni a pour objectif d'interagir avec une API de nouvelles afin de récupérer des articles en fonction d'une requête spécifique. Il utilise JavaScript pour effectuer des opérations telles que la récupération des paramètres de la requête et du numéro de page à partir de l'URL, l'appel à l'API de nouvelles, la manipulation des données de réponse et la mise à jour du contenu HTML de la page.

La fonction "fetchNews" est une fonction asynchrone qui prend deux paramètres : "query" et "pageNo". Elle utilise l'API fetch pour effectuer une requête à une API de nouvelles en utilisant les valeurs de ces paramètres.

La fonction utilise le mot-clé "await" pour attendre la réponse de la requête fetch. Une fois que la réponse est obtenue, elle est assignée à la variable "a".

En résumé, la fonction "fetchNews" effectue une requête à une API de nouvelles en utilisant les paramètres "query" et "pageNo", et attend la réponse de cette requête.

La figure 5.7représente déclaration des variables

```
let articlesPerPage;
let totalPages;
prev = document.getElementById("pbutton")
console.log("Hey I am javascript")
let query = window.location.search.split("?")[1].split("&")[0].split("=")[1];
let page = parseInt(window.location.search.split("?")[1].split("&")[1].split("=")[1]);
console.log(query, page)
```

FIGURE 5.7 – déclaration des variables

Le code ci-dessous 5.8 est utilisé pour récupérer des articles d'actualité à partir de l'API en fonction d'une requête spécifique.

```
const fetchNews = async (query, pageNo) =>
{
    let a = await fetch(`/api?q=${query}&apiKey=d99e1b37ae004e808b0ee06e896192eb&pageSize=8&page=${pageNo}`)
    let r = await a.json()
    console.log(r)
    queryText.innerHTML = query.replace("+", " ")
    queryResults.innerHTML = r.totalResults
    totalPages = Math.ceil(r.totalResults / articlesPerPage)
        next.href = `/?q=${query}&pageno=${page + 1}`
        pre.href = `/?q=${query}&pageno=${page - 1}`
        console.log(page)
        if(page>1){
            prev.style.display="block"
        }else{
            prev.style.display="none"
        }
    let str = ""
    for (let item of r.articles)
    {
        let date = new Date(item.publishedAt).toLocaleDateString()
        str = str + `
        <div class="col-sm-3">
              <div class="card text-bg-dark m-2" style="width: 18rem;">
                <img src="${item.urlToImage}" class="card-img-top" alt="...">
                <div class="card-body text-bg-dark">
                    <h5 class="card-title">${item.title}</h5>
                    <span class="fw-bold">Published </span> : ${date}
                    <p class="card-text">${item.description}</p>
                    <a target="_blank" href="${item.url}" class="btn btn-success">Read More...</a>
                </div>
            </div>
        </div>
        `
    }
    content.innerHTML = str;
```

FIGURE 5.8 – Le code récupérer des articles d'actualité

5.4.2 package json

package-lock.json ou package.json qui spécifie les dépendances et les versions des packages utilisés dans un projet Node.js. Ces packages sont des bibliothèques ou des modules externes qui fournissent des fonctionnalités spécifiques au projet.

— "api" : une dépendance non spécifiée dans le contenu fourni. Il peut s'agir d'un package personnalisé spécifique au projet.

— "axios" : une bibliothèque populaire pour effectuer des requêtes HTTP.

— "express" : un framework web utilisé pour créer des applications web en Node.js.

— "news" : une dépendance non spécifiée dans le contenu fourni. Il peut s'agir d'un package personnalisé spécifique au projet.

5.4.3 server .js

Ce code d'utilisation d'Express pour créer un serveur web utilisant le port 3002. Il utilise également les modules path, axios et express pour différentes fonctionnalités :

— `express` : Le module principal d'Express qui permet de créer une instance de l'application Express.

— `path` : Le module permettant de manipuler les chemins de fichiers et de répertoires.

— `axios` : Le module qui permet d'effectuer des requêtes HTTP vers des API externes.

— `app` : L'instance de l'application Express qui sera utilisée pour définir les routes et les middlewares.

— `port` : Le numéro du port sur lequel le serveur écoutera les requêtes entrantes.

Le code définit deux routes principales :

— La route racine : Elle renvoie le fichier "index.html" en utilisant la méthode res.sendFile() avec le chemin absolu spécifié par path.join.

— La route "/api" : Elle gère les requêtes GET vers "/api" en effectuant une requête vers l'API de NewsAPI en utilisant Axios. Les résultats sont ensuite renvoyés en tant que réponse JSON avec la méthode res.json().

```
const express = require('express')
const path = require('path')
const axios = require('axios')
const app = express()
const port = 3002

app.get('/', (req, res) => {
  res.sendFile('index.html',{root: path.join(__dirname)})
})

app.get('/api', async(req, res) => {
  console.log(req._parsedUrl.query)
  let url = "https://newsapi.org/v2/everything?"+req._parsedUrl.query
  let r = await axios(url)
  let a = r.data
  res.json(a)
})

app.use(express.static(path.join(__dirname, "js")));

app.listen(port, () => {
  console.log(`Example app listening on port ${port}`)
})
```

FIGURE 5.9 – serveur web

Pour lancer le serveur web, vous pouvez utiliser le terminal et exécuter la commande suivante : `node server.js` ,Tel qu'illustre dans la figure 5.10 .

FIGURE 5.10 – Exécuter le serveur web

5.4.4 index.HTML et CSS

Une page Web a été créée en utilisant html et css. Le site Web contient une barre des tâches et une barre de recherche en haut, en plus d'une définition de page, ainsi qu'une section dédiée à l'affichage des actualités au milieu. En bas, il comprend des informations supplémentaires telles que le nom du programmeur et les liens Facebook et Instagram. Comme indiqué dans la image5.11 et la deuxième image 5.12

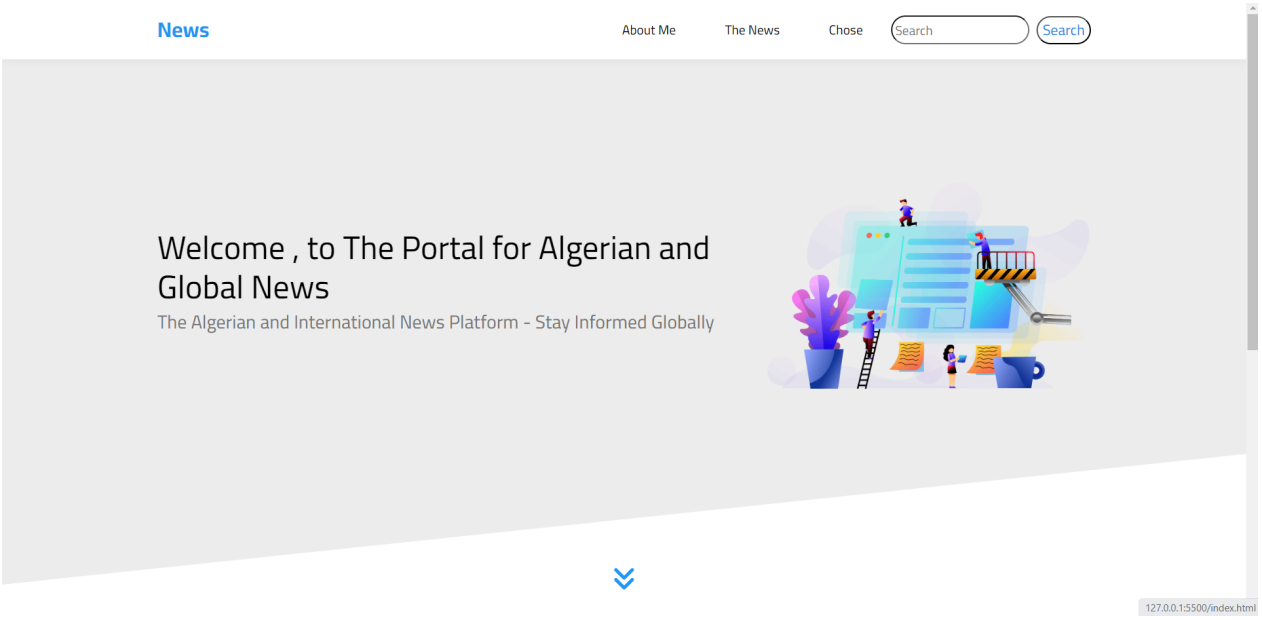

FIGURE 5.11 – Contenu des pages Web

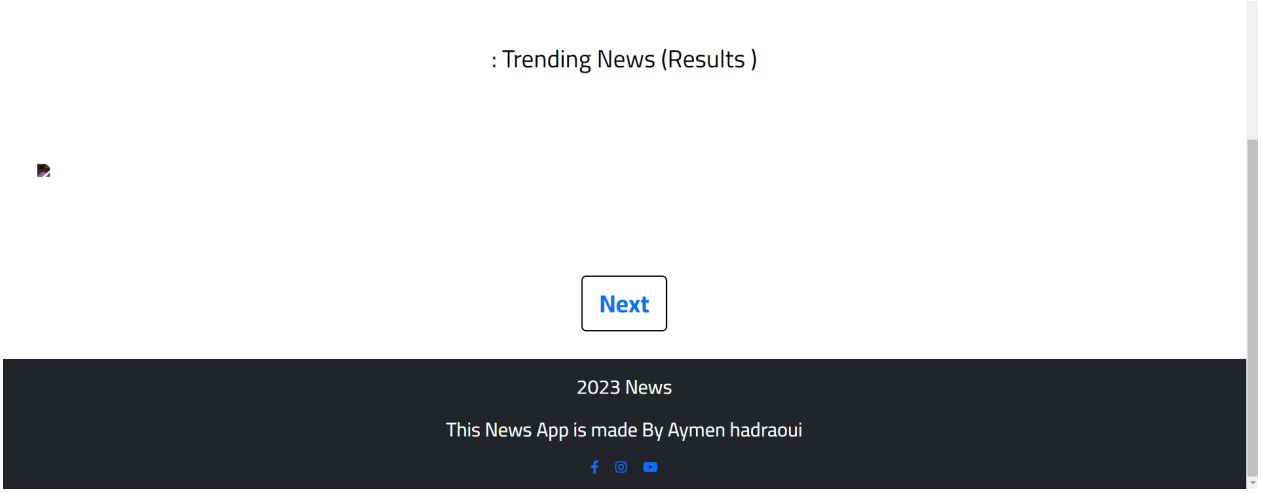

FIGURE 5.12 – Contenu des pages Web

5.5 Les résultats obtenus

Nous avons sélectionné les actualités liées à "Bein-sport" à partir de la barre des tâches. Comme indiqué sur la figure 5.14 et la figure5.13 Les résultats ont été capturés le 27 mai 2023, ce qui explique pourquoi certains résultats ont été récupérés en temps réel.

BeIN Sports: Trending News (526Results)

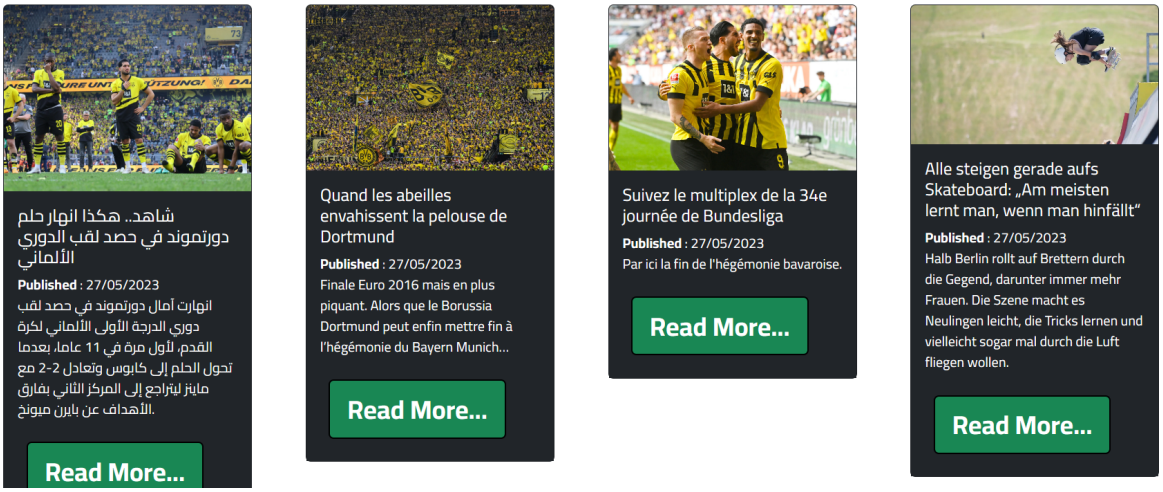

FIGURE 5.13 – Les résultats obtenus

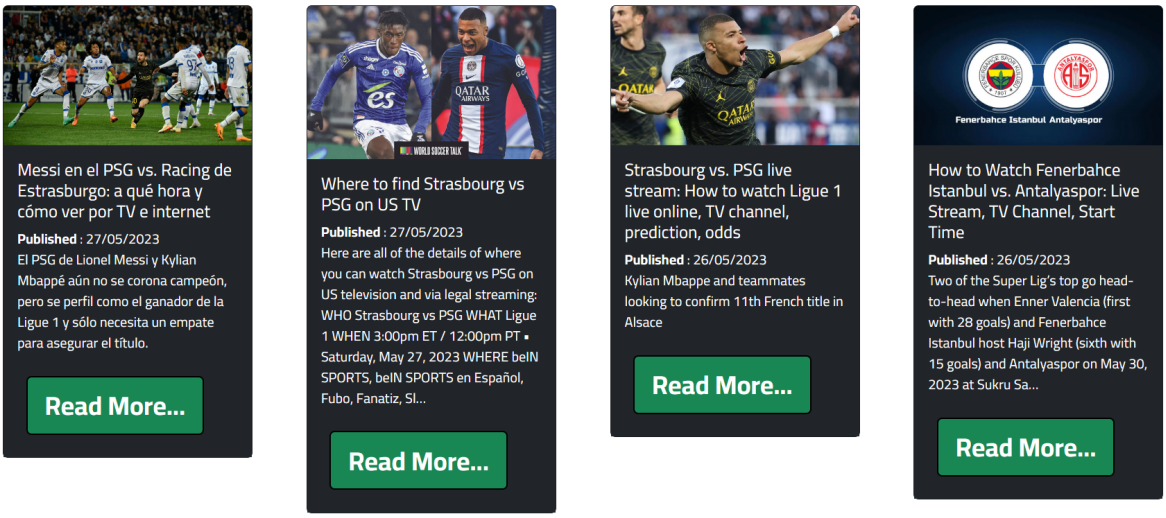

FIGURE 5.14 – Les résultats obtenus

Il est également possible de rechercher des actualités sur l'Algérie en arabe à l'aide du moteur de recherche dédié sur le site d'actualités. Comme indiqué sur la figure 5.15 et la figure5.16

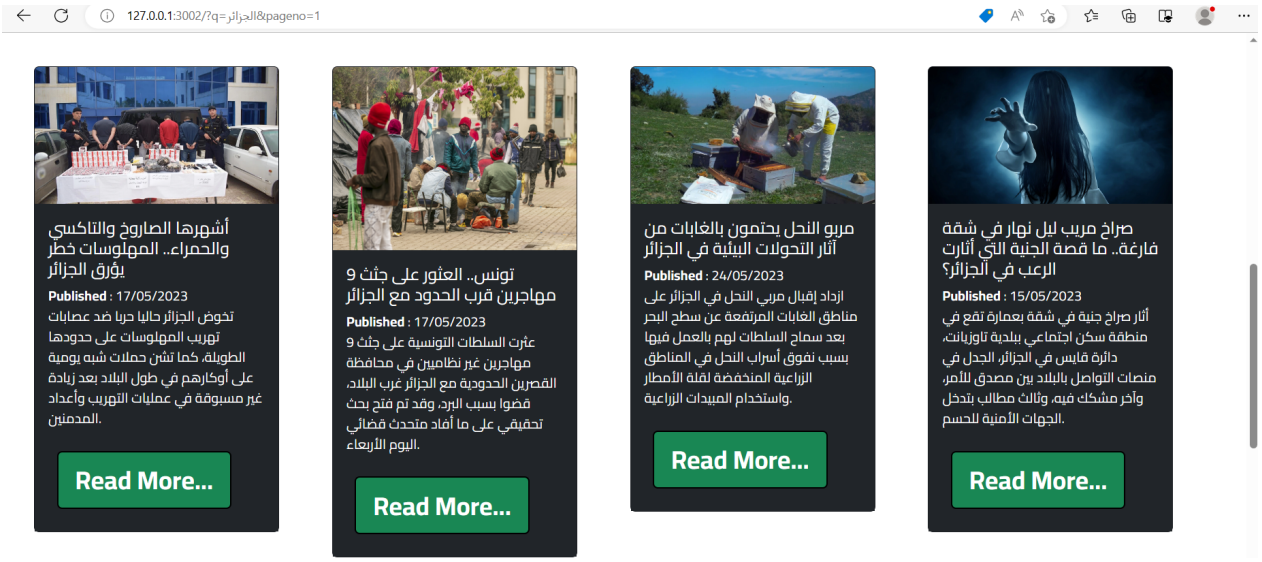

FIGURE 5.15 – Les résultats obtenus

FIGURE 5.16 – Les résultats obtenus

Il est également possible de rechercher des actualités sur Riyad Mahrez en à l'aide du moteur de recherche dédié sur le site d'actualités. Comme indiqué sur la figure 5.17 et la figure5.18

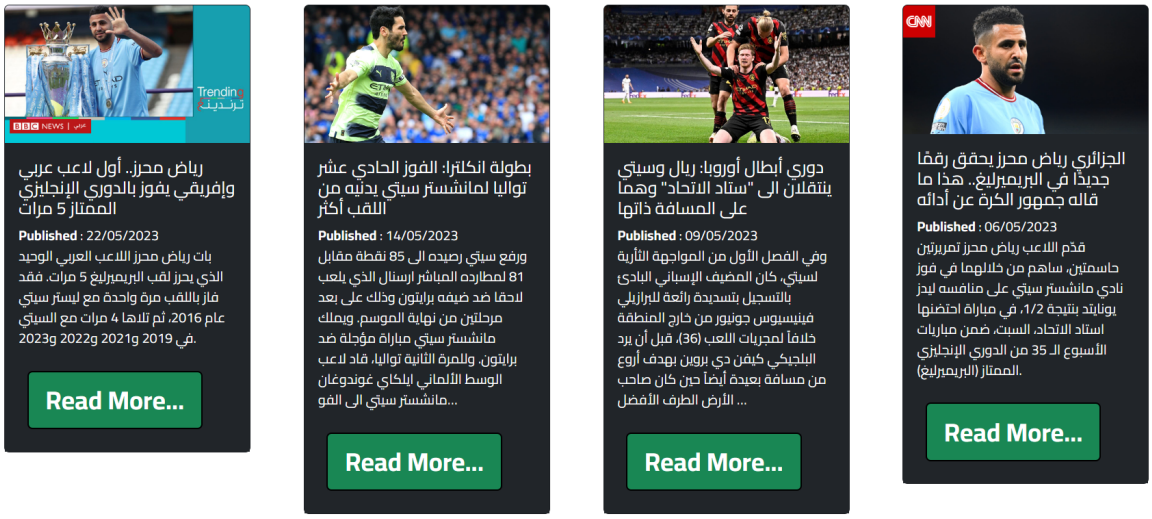

FIGURE 5.17 – Les résultats obtenus

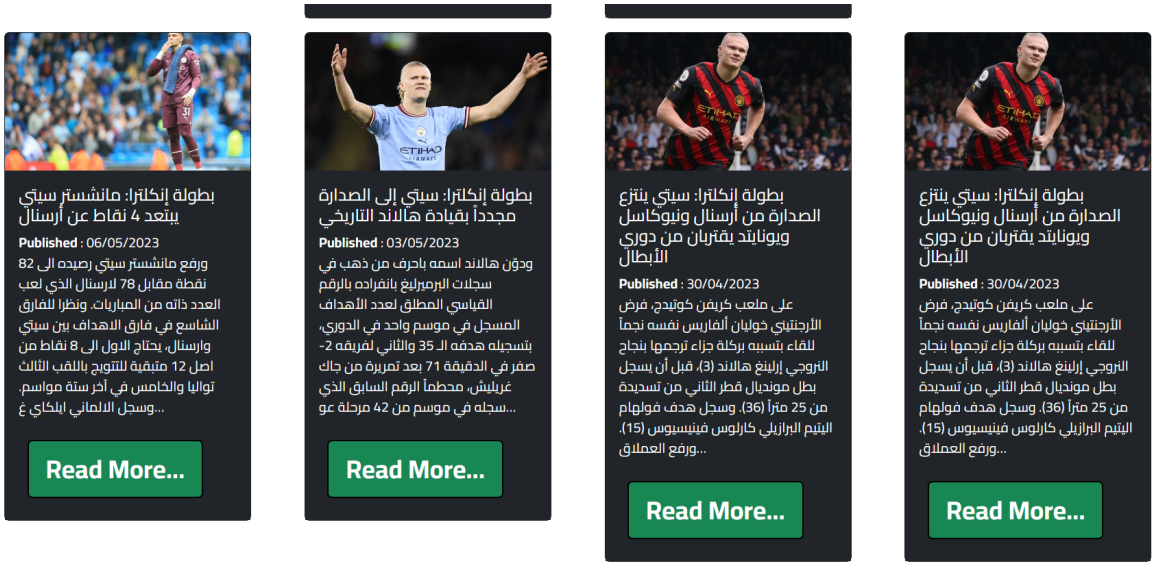

FIGURE 5.18 – Les résultats obtenus

5.6 Conclusion

Les différentes composantes de l'application ont été définies, y compris les outils de programmation et de structuration du programme. Les fonctionnalités diverses ont été discutées et les résultats ont été présentés.

Chapitre 6

Conclusion générale

Dans le cadre de notre étude approfondie de la gestion du streaming Web, nous avons examiné tous les aspects du domaine. Nous avons étudié les thèmes de la surveillance, de la recherche et de la récupération des données, ainsi que les défis liés à l'accès à une grande base de données de données.

Pour mettre notre étude en pratique, nous avons développé un site Web côté client qui donne des informations et des sujets sur leurs intérêts, tels que le sport, la politique et les dernières nouvelles, à partir d'une variété de sources d'information mondiales et locales.

La fonction principale de notre site Web est de fournir aux utilisateurs un accès facile et pratique à des informations pertinentes et fiables.

Maintenant, nous souhaitons élargir notre portée en créant une application mobile qui offrira à chaque client un compte unique. Cette application mobile permettra aux utilisateurs d'accéder facilement aux données collectées, ainsi qu'à d'autres fonctionnalités complémentaires. L'objectif principal de cette application est de fournir une expérience utilisateur pratique et personnalisée à nos clients, en leur permettant d'accéder rapidement aux informations pertinentes sur le sport et la politique.En résumé, notre objectif est de développer une application mobile conviviale, offrant à chaque client un compte unique et permettant une gestion simplifiée des flux web. Nous visons à fournir une expérience utilisateur optimale en facilitant l'accès

Bibliographie

[1] Ioannis Kouris, Christos Makris, Evangelos Theodoridis, and Athanasios Tsakalidis. Indexing and compressing text. In Encyclopedia of Information Science and Technology, Third Edition, pages 1800–1808. IGI Global, 2015

[2] Richard S Segall and Shen Lu. Linkage discovery with glossaries. In Encyclopedia of Business Analytics and Optimization, pages 1411–1421. IGI Global, 2014

[3] María-Dolores Olvera-Lobo and Juncal Gutiérrez-Artacho. Searching health information in question-answering systems. In Handbook of Research on ICTs for Human-Centered Healthcare and Social Care Services, pages 474–490. IGI Global, 2013. .

[4] Mezzi, M. "Système de recherche sensible au contexte : Contribution a un Modelé sémantiquement enrichi pour la recherche d'information textuelle basée sur les folksonomies", thèse de doctorat en informatique, Université de Blida 1. (2018)

[5] Bouramoul, A. " Recherche d'information Contextuelle et Sémantique sur le Web", thèse de doctorat en informatique, Université MENTOURI de Constantine. (2011)

[6] Anwar A. Alhenshiri, "Web Information Retrieval and Search Engines Techniques", 2010, Al- Satil journal, PP : 55-92.

[7] A.bouramoul, «Recherche d'information contextuelle et semantique sur le web », thèse doctorat, université de MENTOURI de Constantine, 2011

[8] Herzallah Abdelkarim «RECHERCHE D'INFORMATION »,support de cours, Université Akli Mohand Oulhadj - Bouira,2017

[9] definitions-marketing, http ://www.definitions-marketing.com/definition/ecriture-web/ ,consulté (03/03/2023).

[10] C marketing, https ://c-marketing.eu/du-web-1-0-au-web-4-0/, consulté (13/03/2023)

[11] Marina Santini , Some Issues in Automatic Genre Classification of Web Pages,scholar google,University of Brighton, Lewes Rd, Brighton, UK,2006

[12] Mitali Desai and Mayuri A. Mehta," A HYBRID CLASSIFICATION ALGORITHM TO CLASSIFY ENGINEERING STUDENTS' PROBLEMS AND PERKS " Department of Computer Engineering, Sarvajanik College of Engineering and Technology, Surat, India ,

International Journal of Data Mining Knowledge Management Process (IJDKP) Vol.6, No.2, March 2016

[13] Xiaoguang Qi and Brian D. Davison. Web Page Classification : Features and Algorithms. ACM Computing Surveys, 41(2), February 2009. An earlier draft was published as Technical Report LU-CSE-07-010

[14] Headmind Partners, https ://www.headmind.com/fr/text-mining-classification-automatique-de-textes/,consulté (03/04/2023).

[15] Aouine Mohammed , CATEGORISATION AUTOMATIQUE DE TEXTE ARABE,diplôme de Magister en Informatique UNIVERSITE DE GUELMA,2009

[16] shanelynn,https ://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/ ,consulté (04/04/2023).

[17] Sami Laroum, Nicolas Béchet, Hatem Hamza, Mathieu Roche. Classification automatique de documents bruités à faible contenu textuel,Livre 2009

[18] CIIA 2009 : Mohammed El Amine Abderrahime 2éme conférence internationale sur l'informatique et ses applications .vers une interface pour l'enrichissement des requêtes en arabe dans un systéme de recherche d'information. Université Abou Bekr Belkaid Tlemcen Algérie.2009

[19] Lahiru Liyanapathirana. Nlp chronicles : Introduction to natural language processing with nltk,avril 2023. URL https ://heartbeat.fritz.ai/ nlp-chronicles-intro-to-nlp-with-nltk-b2c369fbb9a7

[20] Hafsa Jabeen. Stemming and lemmatization in python. https ://www.datacamp.com/community/tutorials/stemminglemmatization-python, avril 2023.

[21] Elsa Negre. Comparaison de textes : quelques approches.... ,Laboratoire d'Analyses et Modélisation de Systèmes pour l'Aide à la Décision UMR 7243,2013. ffhal-00874280

[22] Matthieu Constant, cour Similarité entre les mots ,Traitement Automatique des Langues Master Informatique Université Paris-Est Marne-la-Vallée,2003

[23] Fabien Picarougne, Gilles Venturini, et Christiane Guinot. Un algorithme génétique

parallèle pour la veille stratégique sur internet. In VSST 2004, Toulouse, France, 25-29 octobre 2004. à paraître

[24] Bolens, S. (2017). Mise en place d'un système de veille documentaire : Comment capitaliser des ressources numériques spécifiques et optimiser leur accès au Centre de documentation et bibliothèque du CNP. Certificat en gestion de documentation et de bibliothèque, Yverdon-les-Bains

[25] Gaska, S. (2023, 15 février). Veille informationnelle : Guide complet. Leptidigital. Récupéré le 9 avril 2023,
à partir de https ://www.leptidigital.fr/inbound-marketing/veille-informationnelle-guide-complet-36088/

[26] GaskaFacilacliker,Veille stratégique : Démarrer une veille efficace, à partir de https ://facilacliker.fr/veille-strategique-demarrer-une-veille-efficace/,Récupéré le 9 avril 2023,

[27] Yumens. Définition de la veille. Récupéré le 9 avril 2023,
à partir de https ://www.yumens.fr/expertise/veille/definition/

[28] Benhadji, Y., Laouedj, Z. (2020). La veille stratégique : levier de compétitivité de l'entreprise [Strategic scanning : a leverage of company competitiveness]. Les Cahiers du MECAS, 16(2), 72.

[29] Miaux, J.-F. (2010). Mise en œuvre d'une activité de veille : Le cas de Réseau Ferré de France (Mémoire de Titre professionnel "Chef de projet en ingénierie documentaire" INTD, niveau I). Conservatoire National des Arts et Métiers, Institut National des Techniques de la Documentation.

[30] COHEN Corine. Veille et intelligence stratégiques. Paris, Hermès Lavoisier, 2004. 286p. ISBN : 2-7462-0851-2

[31] Rouach, D. (2010). La veille technologique et l'intelligence économique. Collection Que sais-je ?, Economie. Paris : Presses universitaires de France.

[32] A.bouramoul, «Recherche d'information contextuelle et semantique sur le web », thèse doctorat, université de MENTOURI de Constantine, 2011.

[33] Praveen Dubey. An introduction to bag of words and how to code it in python for nlp,Récupérer le Avril 2023.
URL https ://www.freecodecamp.org/news/ an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9

[34] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Natural Language Engineering, 16(1) :100–103, 2010.

[35] Margaret Rouse. stop word, September 2005.
URL https ://searchmicroservices.techtarget.com/definition/stop-word.Récupérer le 23Avril 2023

[36] Jurafsky, D., Martin, J. H. (2019). Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson.

[37] Semanticall,"L'analyse sémantique et l'analyse lexicale".
URL https ://semanticall.com/lanalyse-semantique/,Récupéré le 1 avril 2023

[38] N.D.Y. Kompaoré. « Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes : vers un processus de RI adaptatif ». Thèse de doctorat en informatique, Université Paul Sabatier de Toulouse, 2008.

[39] Mohamed bacha,"Outils pour organiser et gérer sa veille sur le web",2016, URL https ://www.academia.edu, Récupéré le 7 avril 2023

[40] Baeldung. "What Is Cosine Similarity?" 14 May 2023.
https ://www.baeldung.com/cs/cosine-similarity

[41] Russir. "10 règles d'écriture SEO pour être visible par les moteurs sur internet." 10 May 2023. https ://xn–russir-en-b4a.fr/10-regles-decriture-seo-pour-etre-visible-par-les-moteurs-sur-internet/

[42] Visual Studio Code , https ://code.visualstudio.com/,Date d'accès : 27 mai 2023

[43] w3schools,https ://www.w3schools.com/nodejs/nodejs ,Date d'accès : 29 mai 2023

[44] javatpoint,https ://www.javatpoint.com/expressjs-tutorial,Date d'accès : 30 mai 2023

[45] NewsApi,https ://newsapi.org/docs,Date d'accès : 24 mai 2023