UNIVERSITÉ SAAD DAHLEB BLIDA 1
Faculté des Sciences
Département d'informatique

Mémoire de Master
En Informatique

Option : Ingénierie des logiciels

Indexation et recherche des documents textuels

Réalisé par :
Abderrahmane Wassim Mehdaoui
Morsli Idriss

Supervisé par :
Dr. BACHA Siham

Devant le jury composé de :
Président :
Examinateur :

Juin 2023

Résumer

De nos jours, la disparition du patrimoine algérien et le manque de moyens d'archivage numérique ont conduit à une perte irréversible d'informations précieuses. La préservation du patrimoine culturel et historique de l'Algérie est devenue une préoccupation majeure, mais les ressources et les infrastructures nécessaires pour archiver numériquement ces informations font cruellement défaut.

La recherche d'informations dans ce contexte représente un défi majeur, car les utilisateurs sont confrontés à la difficulté de retrouver et d'accéder aux connaissances et aux ressources disponibles. L'énorme quantité d'informations perdues ou non archivées rend cette tâche encore plus complexe. Les systèmes de recherche d'informations doivent donc faire face à ce problème de disparition du patrimoine algérien et de manque de moyen d'archivage numérique.

Une solution possible à cette problématique serait de mettre en place des initiatives visant à numériser et à archiver de manière systématique le patrimoine culturel et historique de l'Algérie. Cela permettrait de préserver ces informations pour les générations futures et de faciliter leur accès grâce à des systèmes de recherche d'informations performants. L'indexation jouerait un rôle clé dans ce processus, en extrayant les éléments descripteurs des documents numérisés pour permettre une recherche pertinente et efficace.

Il est donc essentiel de mobiliser les ressources et les compétences nécessaires pour résoudre ce problème de disparition du patrimoine algérien et de manque de moyen d'archivage numérique. Cela permettrait de préserver l'histoire et la culture de l'Algérie, tout en facilitant l'accès à ces informations pour les chercheurs, les étudiants et le grand public.

Notre projet propose un outil sophistiqué qui permet a l'utilisateur de publié et navigué du contenus textuel,et ce dans les deux langues (Arabe et Anglais). Ceci est accompli en utilisant des techniques d'intelligence artificielle, de traitement automatique du langage et d'apprentissage automatique.

*Mots clés :*Indexation, recherche d'information, documents textuels, embedding.

الملخص

في يومنا هذا، أدت اختفاء التراث الجزائري ونقص وسائل التخزين الرقمي إلى فقدان لا رجعة فيه للمعلومات الثمينة. أصبح الحفاظ على التراث الثقافي والتاريخي للجزائر أمرًا هامًا، ولكن الموارد والبنية التحتية اللازمة لأرشفة هذه المعلومات بوسعها تفتقر بشدة.

تشكل بحث المعلومات في هذا السياق تحديًا كبيرًا، حيث يواجه المستخدمون صعوبة في العثور على المعرفة والموارد المتاحة والوصول إليها. يجعل الكم الهائل من المعلومات المفقودة أو غير المؤرشفة هذه المهمة أكثر تعقيدًا. لذا يجب على أنظمة بحث المعلومات التعامل مع مشكلة اختفاء التراث الجزائري ونقص وسائل التخزين الرقمي.

إحدى الحلول الممكنة لهذه المشكلة هي إطلاق مبادرات تهدف إلى تحويل وأرشفة تراث الجزائر الثقافي والتاريخي بطريقة منهجية. سيساعد ذلك على الحفاظ على هذه المعلومات للأجيال القادمة وتيسير الوصول إليها من خلال أنظمة بحث معلوماتية فعالة. ستلعب فهرسة دورًا رئيسيًا في هذه العملية من خلال استخراج عناصر وصفية من المستندات المرقمة لتمكين البحث ذي الصلة والفعال.

لذا من الضروري تعبئة الموارد والمهارات اللازمة لحل مشكلة اختفاء التراث الجزائري ونقص وسائل التخزين الرقمي. سيساعد ذلك على الحفاظ على تاريخ وثقافة الجزائر، مع تسهيل الوصول إلى هذه المعلومات للباحثين والطلاب والجمهور العام.

يقترح مشروعنا أداة متطورة تسمح للمستخدمين بنشر وتصفح المحتوى النصي باللغتين العربية والإنجليزية. يتم تحقيق ذلك باستخدام تقنيات الذكاء الاصطناعي ومعالجة اللغة الطبيعية والتعلم الآلي.

الكلمات المفتاحية : فهرسة، استرجاع معلومات، وثائق نصية، التضمين

Abstract

Nowadays, the disappearance of Algerian heritage and the lack of digital archiving means have led to an irreversible loss of valuable information. Preserving Algeria's cultural and historical heritage has become a major concern, but the resources and infrastructure needed to digitally archive this information are sorely lacking.

Information retrieval in this context poses a significant challenge as users face difficulty in finding and accessing available knowledge and resources. The vast amount of lost or unarchived information further complicates this task. Information retrieval systems must therefore address this issue of Algerian heritage disappearance and lack of digital archiving means.

One possible solution to this problem would be to establish initiatives aimed at systematically digitizing and archiving Algeria's cultural and historical heritage. This would help preserve this information for future generations and facilitate access through efficient information retrieval systems. Indexing would play a key role in this process by extracting descriptive elements from digitized documents to enable relevant and effective searching.

It is therefore crucial to mobilize the necessary resources and expertise to address this issue of Algerian heritage disappearance and lack of digital archiving means. This would preserve Algeria's history and culture while making these valuable resources more accessible to researchers, students, and the general public.

Our project proposes a sophisticated tool that allows users to publish and navigate textual content in both Arabic and English languages. This is achieved through the use of artificial intelligence, natural language processing, and machine learning techniques.

Keywords : Indexing, information retrieval, textual documents, embedding.

REMERCIEMENTS

Avant tout, je remercie ALLAH le tout-puissant de m'avoir donné le courage, la volonté et la patience de mener à terme ce présent travail dans les meilleures conditions.

Je tiens à exprimer toute ma reconnaissance à mon encadreur principal, Madame Bacha Siham. Je la remercie de m'avoir encadré, orienté, aidé et conseillé, et pour la liberté de travail qu'elle m'a laissée tout au long de ce semestre. Je la remercie aussi pour sa disponibilité, sa patience et surtout ses judicieux conseils, qui ont grandement contribué à alimenter ma réflexion.

J'adresse mes sincères remerciements à tous les professeurs de l'USDB , intervenants, et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé mes réflexions et ont accepté de me rencontrer et répondre à mes questions durant mon cursus et mes recherches et qui doivent voir dans ce travail la fierté d'un savoir bien acquis. Je remercie mes très chers parents, qui ont toujours été là pour moi. Ma mère, qui a oeuvré pour ma réussite, de par son amour, son soutien, tous les sacrifices consentis et ses précieux conseils, pour toute son assistance et sa présence dans ma vie : reçois à travers ce travail aussi modeste soit-il, l'expression de mes sentiments et de mon éternelle gratitude. Mon père, qui peut être fier et trouver ici le résultat de longues années de sacrifices et de privations pour m'aider à avancer dans la vie. Puisse Dieu faire en sorte que ce travail porte son fruit ; merci pour les valeurs nobles, l'éducation et le soutien permanent venu de toi. « Vous avez tout sacrifié pour vos enfants n'épargnant ni santé ni efforts. Vous i m'avez donné un magnifique modèle de labeur et de persévérance. Je suis redevable d'une éducation dont je suis fier ».

J'adresse tous mes voeux de réussite à Morsli Idriss (mon binôme), ainsi que tous les autres étudiants en Master 2 IL spécialement et les autres Masters en général. Enfin, je remercie tous mes Amis que j'aime tant, et tous mes autres ami(e)s pour leur sincère amitié et confiance, et à qui je dois ma reconnaissance et mon attachement. À tous ces intervenants, je présente mes remerciements, mon respect et ma gratitude

Wassim

REMERCIEMENTS

Tout d'abord, je souhaite adresser mes remerciements les plus sincères à Allah. Sa présence bienveillante et Sa bénédiction ont été des sources inépuisables de force et de guidance tout au long de ce parcours académique exigeant. Je suis profondément reconnaissant envers Allah pour m'avoir accompagné à chaque étape de ce chemin et m'avoir permis d'atteindre mes objectifs.

Je tiens aussi à exprimer mes sincères remerciements à mes parents pour leur amour, leur soutien inconditionnel et leurs sacrifices constants tout au long de ce parcours académique.

Je remercie mon défunt père , pour son amour, ses conseils et son dévouement indéfectible. Je te suis reconnaissant pour tout ce que tu as fait .Ton influence positive restera à jamais gravée dans mon cœur et je suis honoré d'être ton fils "alah yatrahmak". Je souhaite aussi exprimer mes plus sincères remerciements à ma maman, à mon frère , mes deux sœurs et mon oncle pour leur amour, leur soutien inconditionnel et leur présence précieuse dans ma vie. Je souhaite également exprimer ma gratitude à ma chère encadrante, Madame Bacha Siham . Votre expertise, votre patience et votre dévouement m'ont permis de mener à bien ce mémoire. Vos conseils éclairés et votre encadrement précieux ont été d'une aide inestimable pour moi. Je vous suis infiniment reconnaissant pour votre soutien constant. Je remercie très sincèrement les membres du jury d'avoir bien voulu accepté d'évaluer mon travail

Enfin, je tiens à remercier mes amis et ma famille qui m'ont soutenu tout au long de cette aventure. Vos encouragements, vos mots d'encouragement et votre présence bienveillante ont été des éléments indispensables dans ma réussite. Votre soutien moral et votre croyance en moi ont été des moteurs puissants qui m'ont aidé à surmonter les défis et à donner le meilleur de moi-même. À vous tous qui avez contribué, de près ou de loin, à l'aboutissement de ce mémoire, je vous adresse mes plus sincères remerciements. Votre soutien, vos encouragements et votre confiance en moi ont été des cadeaux inestimables. Je suis profondément reconnaissant pour votre apport à ma réussite académique.

Idriss

DÉDICACES

Ce modeste travail est dédié, À mes chers parents, à qui aucun hommage ne pourrait être à la hauteur de leurs sacrifices. Que Dieu leur procure bonheur et longue vie. À ma soeurs Ahlem que j'aime énormément. À tous mes amis en particulier Amine, Chems Eddine, Fady, Nadhir, Khaled, Zyneb, Manel et Adel, et à tous les Apôtres pour leur présence et leur soutien. À tous les membres du groupe Pav6, ma deuxième famille. Et à toute personne qui, de près ou loin, m'a apporté soutien, conseil ou réconfort.

Mille mercis.

Wassim

DÉDICACES

À mes chers parents, à qui aucun hommage ne pourrait être à la hauteur de leurs sacrifices. Que Dieu leur procure bonheur et longue vie. À mon frère Riadh et mes sœurs Madina et Amel et à mes nièces et neveux et mon oncle Mohammed que j'aime énormément. À tous mes amis en particulier O Fella, M cherif, C Mohamed, H Hakim et à tous les Apôtres pour leur présence et leur soutien. À Madame D Hayat. À tous les membres d'IT Community ,ma deuxième famille. Et à toute personne qui, de près ou loin, m'a apporté soutien, conseil ou réconfort.

Mille mercis.

Idriss

Table des matières

Liste des tableaux

Table des figures

Introduction générale

Dans les méandres du temps, l'Algérie se dresse comme un témoin vivant d'une histoire riche et complexe. De l'influence berbère aux conquêtes romaines, en passant par les périodes arabes, ottomanes et françaises, ce pays incarne une véritable mosaïque culturelle. Son patrimoine, aussi diversifié que précieux, constitue l'essence même de son identité nationale. Comme le souligne si justement le proverbe, "Un peuple sans passé est comme un arbre sans racines".

Cependant, cette richesse culturelle est aujourd'hui menacée. Les conflits, les catastrophes naturelles et les changements socio-économiques fragilisent la préservation de ce patrimoine unique. De plus, le manque de moyens modernes et durables d'archivage ajoute une autre dimension critique à cette situation. Les archives traditionnelles, telles que les bibliothèques, jouent certes un rôle essentiel dans la préservation et l'accès aux documents textuels patrimoniaux. Cependant, elles présentent des limites évidentes, notamment en termes d'accessibilité au grand public et de disponibilité des exemplaires à travers le pays. Ainsi, de nombreux passionnés de culture sont contraints de parcourir de longues distances pour consulter les précieux documents qui les intéressent.

Face à ces défis majeurs, une solution innovante s'impose : la création d'une plateforme d'archivage numérique, véritable gardienne du patrimoine algérien. Cette plateforme ambitionne de regrouper l'ensemble des documents textuels liés à la culture et au patrimoine de l'Algérie, offrant ainsi un refuge virtuel à ces trésors culturels. Bien plus qu'un simple outil d'accès, cette plateforme constitue une passerelle vers un univers captivant, permettant aux utilisateurs de plonger dans une expérience immersive et sans frontières.

Parmi les éléments clés de cette plateforme, un processus d'indexation avancé jouera un rôle crucial en visant à améliorer la recherche d'informations et l'extraction précise d'informations des documents. Ce processus permettra d'organiser efficacement les documents, de les catégoriser et de les étiqueter avec précision, facilitant ainsi leur recherche et leur consultation ultérieure. L'objectif principal de l'indexation intelligente sera d'améliorer la capacité des systèmes de recherche à comprendre le contenu des documents, offrant ainsi des résultats plus pertinents et précis. En combinant des techniques d'intelligence artificielle et de traitement automatique du langage, ce processus d'indexation assurera une expérience fluide et confortable aux utilisateurs, en rendant la recherche d'informations aussi intuitive que possible. De plus, en extrayant des informations spécifiques des documents, telles que des dates, des noms de personnes ou des lieux, il permettra aux utilisateurs de trouver rapidement des informations précises, sans avoir à parcourir l'intégralité des documents.

Notre travail consiste donc à proposer une solution sophistiquée et pérenne à ce défi majeur. Notre plateforme d'archivage national offrira aux utilisateurs la possibilité de consulter facilement tous les documents qu'elle contient, à tout moment et en tout lieu. Elle encouragera également le partage des travaux des chercheurs, historiens et anthropologues, tout en respectant leurs droits et reconnaissant leurs mérites. Pour atteindre ces objectifs, notre plateforme tirera parti des avancées de l'intelligence artificielle, du traitement automatique du langage, de l'apprentissage automatique et de la recherche d'information. Une composante essentielle de notre plateforme sera un outil d'indexation intelligent, qui organisera les documents de manière intelligente, extraira des informations pertinentes et facilitera la recherche ciblée.

Dans ce mémoire, nous on entamerons une présentation approfondie dans enjeu qui flotte autour du patrimoine et des problèmes qui le menace puis nous basculerons sur une étude théorique de la recherche d'information et d'indexation avant de passer a l'examinassions littéraire des divers approches et algorithmes utilisé dans le domaine de la recherche d'information. Nous exposerons également notre contribution à ce domaine. La première partie de notre travail permettra au lecteur de recueillir une compréhension complète des techniques existantes pour la recherche d'information. Nous aborderons ensuite l'approche que nous avons pour améliorer la qualité des résultats de recherche. Nous détaillerons le fonctionnement de cette approche et expliquerons en quoi se différencie-t-elle des autres méthodes. Nous présenterons également les résultats de nos expérimentations et les comparerons au résultats obtenus par les autres approches. Dans la troisième partie, nous présenterons le produit final ainsi que sa réalisation.

Enfin, nous conclurons notre travail avec une conclusion générale. Nous aborderons les limites de notre approche et les pistes d'amélioration possibles. Nous présenterons également les prochaines étapes de notre travail.

Chapitre 1

État de l'art

1.1 Contexte et problématique

1.1.1 Le patrimoine immatériel

On entend par "patrimoine immatériel"[37] les pratiques, représentations, expressions, connaissances et savoir-faire, ainsi que les instruments, objets, artefacts et espaces culturels qui leur sont associés, que les communautés, les groupes et, le cas échéant, les individus reconnaissent comme faisant partie de leur patrimoine culturel. Ce patrimoine culturel immatériel, transmis de génération en génération, est recréé en permanence par les communautés et groupes en fonction de leur milieu, de leur interaction avec la nature et de leur histoire, et cela leur procure un sentiment d'identité et de continuité, contribuant ainsi à promouvoir le respect de la diversité culturelle et la créativité humaine. Aux fins de la présente Convention, seul sera pris en considération l'héritage conforme aux instruments internationaux existants relatifs aux droits de l'homme, ainsi qu'à l'exigence du respect mutuel entre communautés, groupes et individus, et d'un développement durable.

Celui-ci se manifeste par ailleurs dans les domaines suivants :
- les traditions et expressions orales, y compris la langue comme vecteur du patrimoine culturel immatériel.
- les arts du spectacle.
- les pratiques sociales, rituels et événements festifs.
- les connaissances et pratiques concernant la nature et l'univers.
- les savoir-faire liés à l'artisanat traditionnel.

En outre, il convient de souligner que la préservation du patrimoine culturel immatériel ne se limite pas à la simple reconnaissance et transmission de ses éléments. Elle implique également la promotion de sa vitalité et de sa viabilité à travers des mesures appropriées. Cela peut inclure des actions visant à soutenir les porteurs et les praticiens de ce patrimoine, à favoriser leur participation active, à sensibiliser le public à son importance et à encourager une coopération pour sa sauvegarde.

D'une part, ils attestent des événements passés ainsi que des idées et croyances antérieures, telles que les manuscrits de la bibliothèque de Tlemcen, datant de l'époque médiévale, qui

représentent une précieuse preuve de la vie intellectuelle et culturelle de cette période. Ils sont aussi témoins de l'évolution de la société à travers le temps, en révélant diverses pratiques culturelles, traditions, événements historiques, personnalités importantes ou encore modes de vie passés. Cela offre une fenêtre vers la compréhension du passé et aide à façonner l'avenir. Ces derniers représentent par ailleurs une source de connaissances pour les chercheurs, les étudiants, les enseignants, les historiens et le grand public.

Ensuite, ces derniers peuvent également avoir une valeur symbolique et émotionnelle pour les individus et les communautés qui les considèrent comme faisant partie de leur identité culturelle. Une mention spéciale doit être faite pour les chants traditionnels, qui sont également des documents textuels et un témoignage important de la culture orale du pays. Ces chants, transmis de génération en génération, reflètent très bien les traditions, les valeurs et l'identité culturelle des communautés. Ils constituent un exemple de la valeur symbolique et émotionnelle que peuvent avoir ces documents patrimoniaux et historiques.

D'autre part, il est vrai que les sites patrimoniaux et les manifestations culturelles attirent un grand nombre de touristes et de visiteurs, ce qui peut avoir un impact économique positif pour la région. Cela peut stimuler la recherche et le développement de projets culturels et contribuer à renforcer l'attractivité de la région en offrant des expériences uniques et en créant des opportunités d'emplois locaux.

1.1.2 Mesures existantes

Ce n'est qu'après avoir saisi les points citer précédemment que l'on comprend pourquoi nous retrouvons des pays comme le Japon qui va jusqu'à crée des agences spécialisé chargé de promouvoir et de protéger le patrimoine culturel et artistique du pays, y compris les œuvres littéraires, musicales, théâtrales et cinématographiques. on a la Chine aussi avec leurs administration nationale du patrimoine culturel qui elle aussi veille a la protection et la préservation du patrimoine culturel et intellectuel du pays. sans oublié l'organisation des Nations Unies pour l'éducation, la science et la culture (UNESCO) qui a mit au point "La Convention pour la sauvegarde du patrimoine culturel immatériel"[37] en 2003 visent à protéger les traditions culturelles et les expressions orales et symboliques de diverses communautés du monde entier. La convention reconnaît que les formes d'expression culturelle sont en constante évolution et que les communautés sont les gardiennes de leur propre patrimoine immatériel, celle-ci a été ratifiée par plus de 180 États membres de l'UNESCO dont l'Algérie, ce qui en fait l'un des instruments juridiques les plus largement ratifiés de l'organisation. La Convention a contribué à accroître la sensibilisation et l'appréciation du patrimoine culturel immatériel dans le monde entier, ainsi qu'à promouvoir la coopération internationale.

Elle invite donc les États membres à identifier et à protéger leur patrimoine culturel immatériel, à promouvoir la coopération internationale et à respecter les droits des communautés détentrices du patrimoine culturel immatériel. Les États membres sont encouragés à établir des inventaires nationaux de leur patrimoine culturel immatériel et à élaborer des politiques et des mesures de sauvegarde pour préserver ce patrimoine.

Depuis son adoption, la convention a été ratifiée par plus de 180 États tous membres de l'UNESCO, ce qui en fait l'un des instruments juridiques les plus largement ratifiés de l'UNESCO. La Convention a contribué à accroître la sensibilisation et l'appréciation du patrimoine culturel immatériel dans le monde entier, ainsi qu'à promouvoir la coopération internationale pour sa sauvegarde et sa transmission aux générations futures.

D'ailleurs elle va même jusqu'à définir "la sauvegarde" du patrimoine comme étant les mesures visant à assurer la viabilité du patrimoine culturel immatériel, y compris l'identification, la documentation, la recherche, la préservation, la protection, la promotion, la mise en valeur, la transmission, essentiellement par l'éducation formelle et non formelle, ainsi que la revitalisation des différents aspects de ce patrimoine.

Autant que signataire de cette convention l'Algérie a prit quelques mesures pour mettre en place des dispositifs tel que la désignation du " mois du Patrimoine" qui se tiens du 18 Avril au 18 mai chaque année et qui se résume en une série d'activités dans plusieurs musées et sites culturels du pays et cette année les activités se concentrent sur la profondeur africaine du patrimoine culturel algérien, exemple le musée national de Cherchel a mit la lumière sur la Maurétanie Césarienne, tandis que le musée national Cirta de Constantine organisa des journées portes ouvertes et des ateliers pour enfants. pendant que le musée des arts et traditions populaires du Palais du Bey expose principalement des peintures et tint des conférences sur le patrimoine culturel[7]. Aujourd'hui, on constate même la naissance de business souvent sous le titre de start-up qui auront comme service la présentation et la préservation du patrimoine Algériens tel que El Fnardjia (الفنارجية) une start-up Algérienne qui a vu naissance en fin 2021 et qui, aujourd'hui, organise des sorties guidées a plusieurs endroits historiques du nord comme du sud Algériens tels que la casbah d'Alger ou les ruines romaines à Tipaza et qui a pour but d'informer et d'éduquer leurs clients sur l'importance et la diversité du patrimoine culturel et historique de notre pays.

1.1.3 Archives inaccessibles

Malgré toutes les mesures prises par l'État, nous manquons encore aujourd'hui de moyens adéquats pour la préservation et le partage de notre patrimoine, en particulier lorsqu'il s'agit de document textuel. Un exemple d'un document textuel patrimonial et historique en Algérie revêtant une importance intellectuelle est "Kitab al-Ibar" d'Ibn Khaldoun. Cet ouvrage est considéré comme l'un des chefs-d'œuvre de la littérature arabe et est une référence majeure pour les études historiques et sociologiques. Ibn Khaldoun y développe une théorie sur l'histoire et la société, basée sur l'observation et l'analyse des événements passés. Ses idées ont eu une influence sur l'histoire et la culture de l'ensemble de la région du Maghreb et sont encore étudiées aujourd'hui. Un autre exemple est "Kitab al-Istibsar" d'Ibn Rushd qui traite le domaine de la jurisprudence islamique, également connu sous le nom de fiqh. Cet ouvrage aborde les différentes écoles de pensée en matière de fiqh et leurs différences, ainsi que les méthodes utilisées pour interpréter le Coran et la Sunna. Il s'agit d'un ouvrage important dans l'histoire de la pensée islamique et a eu une influence significative sur le développement de la jurisprudence musulmane. Sans oublier les divers ouvrages de Malek Bennabi, également connu sous le nom d'Ibn Badis tel que "La

Méditerranée dans la pensée d'Ibn Khaldoun", où il examine les idées de l'historien et philosophe du XIVe siècle Ibn Khaldoun sur la Méditerranée en tant que point de convergence des civilisations et des cultures. Il convient aussi de mentionner qu'il est l'auteur de l'ouvrage intitulé "L'Islam face aux défis de la civilisation", dans lequel il aborde de manière approfondie les nombreux défis auxquels l'Islam est confronté dans le monde moderne. Dans cet ouvrage, il évoque la nécessité pour les musulmans de redécouvrir leur véritable essence et d'approfondir leur compréhension de l'Islam, afin de faire face aux défis actuels de manière éclairée et responsable. À travers une analyse rigoureuse de l'histoire et de la philosophie islamique, l'auteur invite les musulmans du monde entier à s'engager dans une réflexion critique sur les enjeux contemporains et à travailler ensemble pour construire un avenir meilleur pour l'Islam et l'humanité dans son ensemble.

On peut également retrouver d'autres ouvrages plus récents tel que

دولة الجزائر البحرية دولة جزائرية أم مجرد ولاية عثمانية؟

de l'auteur/chercheur Mourad Kahlouche qui traite de la position de l'Algérie avant 1830 et de sa position par rapport a l'empire ottoman, ce dernier s'est spécialisé dans l'histoire algérienne avant la colonisation Française. Il a, par ailleurs, fait plusieurs apparitions dans divers magazines et revues culturelles et historiques. Malheureusement, Mourad Kahlouche nous a quitté il y a de cela un an laissant derrière lui le fruit de plusieurs années de travail sauf que les seuls supports numériques de son travail se trouve sur ses comptes réseaux sociaux personnels (e.g : figure 1.1) qui pourrait être supprimés à la décision de l'un de ses proches et par conséquent entraînerait la perte de ces derniers.

FIGURE 1.1 – Dernière publication de Mourad Kahlouche dans le magazine "المؤرخ الحر" [1]

Les documents pris comme exemple précédemment font partie d'un point de départ essentiel pour la recherche dans divers domaines et malgré la présence de quelques centres

1. twitter.com/ka_mourad

de recherche tel que le CRASC, créé en 1992 a Oran à part leurs bulletins et rapports le grand public ou les chercheurs non affiliés n'ont pas vraiment accès a leurs archives ainsi qu'avec le temps, il est devenu très difficile de se procurer d'un exemplaire de ses documents historiques dû à la dégradation du papier ou au diverses conditions naturelles (incendie 1.2, tremblement de terre...ect).

FIGURE 1.2 – Témoignage montrant la perte de documents historiques dû à un incendie [2]

1.1.4 Problème de protection des droits d'auteurs

Et cela représente un énorme danger vis a vis de notre patrimoine du a l'absence de moyens de digitalisions et de diffusion de documents ou des archives. Nous retrouvons aussi certaine personne ayant en leurs possession de ancien livre/écrit disparu des librairie aujourd'hui, mais qui n'ont pas les moyens de les éditer voir publier et donc de les diffuser alors certes il existe aujourd'hui des plateformes de publication et consultation de document

2. www.facebook.com/faizariache ?mibextid=ZbWKwL

textuelle gratuit tell qu'Archive.org mais on l'a vu a plusieurs reprise que l'information est d'abord disponible gratuitement au début et deviens payante avec le temps une fois que la plateforme utilisé se vulgarise on peut prendre l'exemple de Medium.com qui est une plateforme de blogging et d'écriture en ligne qui permet aux utilisateurs de publier des articles sur une variété de sujets. La plateforme était gratuite à ses débuts, mais depuis 2019, Medium a introduit un modèle d'abonnement payant qui donne accès à des fonctionnalités supplémentaires et à des articles exclusifs. ou d'autre qui ont fini par fermer tell que Google Reader qui est un agrégateur de flux RSS populaire qui permettait aux utilisateurs de suivre les dernières mises à jour de leurs sites web préférés, ce derniers a été mit hors service par Google en 2013. Il est également important de prendre en considération le fait que lorsque du contenu est publié sur ces plateformes, cela peut donner à l'entreprise une liberté complète sur l'utilisation et la reproduction de ce contenu. En d'autres termes, l'entreprise peut avoir le droit de réutiliser le contenu publié comme bon lui semble, sans avoir à obtenir une autorisation supplémentaire ou à rémunérer l'auteur du contenu 1.3. Cette pratique est souvent stipulée dans les conditions d'utilisation des plateformes en question, et il est essentiel de bien les lire avant de publier du contenu.

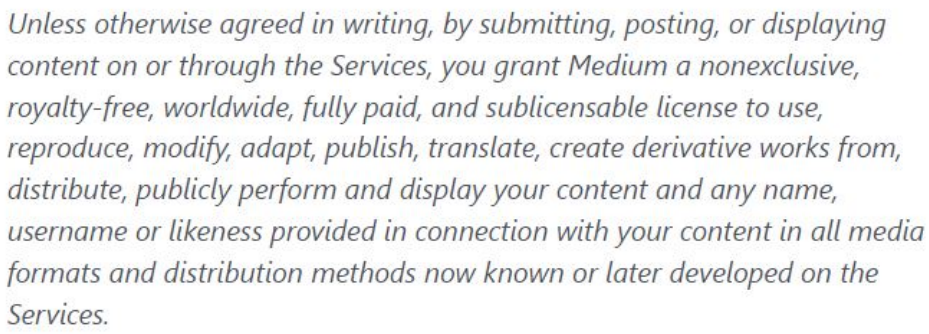

Unless otherwise agreed in writing, by submitting, posting, or displaying content on or through the Services, you grant Medium a nonexclusive, royalty-free, worldwide, fully paid, and sublicensable license to use, reproduce, modify, adapt, publish, translate, create derivative works from, distribute, publicly perform and display your content and any name, username or likeness provided in connection with your content in all media formats and distribution methods now known or later developed on the Services.

FIGURE 1.3 – Extrait de la charte d'utilisation de la plateforme medium.com [3]

1.1.5 Contribution

Il est donc essentiel de prendre en compte ces conséquences potentielles pour de l'adoption d'une solution qui implique une perte de l'identité intellectuelle de notre nation, car cela porte risque a la perte ou la capitalisation de notre information par des tiers, ce qui va à l'encontre de notre objectif initial d'où la vocation de ce travail. Notre objectif était de développer une plateforme totalement ouverte et accessible, destinée à l'archivage et la diffusion de touts documents touchant de près ou de loin au patrimoine immatériel de notre pays.

Par conséquent, dans le cadre de ce mémoire, nous allons principalement nous concentrer sur la mise en place d'un moteur de recherche efficace, rapide et adaptable pour notre

3. www.medium.com

plateforme. celui-ci devra être en mesure d'explorer et d'indexer de vastes quantités de données, tout en garantissant une pertinence maximale des résultats. Afin de concevoir ce moteur de recherche, nous allons explorer les différentes approches et techniques utilisées autrefois pour la recherche d'information, ainsi que les outils les plus couramment employés dans ce domaine.

En résumé, la prochaine section de ce chapitre abordera en détail les défis associés à la recherche d'information dans le contexte de notre plateforme de conservation et de diffusion du patrimoine immatériel. Nous allons également examiner les différentes approches, ainsi que les divers plateforme de partage de textes affin de les comparer et choisir celle qui répondra le mieux au besoin de notre projet.

1.2 Principe de l'indexation et de la recherche d'information

Le domaine de la recherche d'information [33] peut être considéré, dans une certaine mesure, comme le domaine appliqué le plus réussi de la NLP. La vitesse et l'échelle de l'adoption du Web dans le monde entier n'a été rendue possible que par le bias des moteurs de recherche efficaces et disponibles gratuitement. Ces outils sont utilisés par environ 85% des internautes lorsqu'ils recherchent des informations spécifiques [52].

Mais qu'est-ce que la RI précisément ? nous pourrions la définir comme tell : "La RI traite tout se qui est représentation, stockage, organisation et accès a l'information. L'information en question peut être des références à de vrais documents, des documents eux-mêmes, voire des paragraphes uniques, ainsi que des pages Web, documents oraux, des images, photos, musique, vidéo, etc." [8]. Les systèmes de recherche modernes n'interagissent pas directement avec les documents (ou les requêtes). Ils utilisent différentes techniques et stratégies(La figure 1.4) pour représenter les principaux aspects sémantiques des documents et des requêtes. On nomme cela le processus d'indexation.

L'indexation de documents au sein d'une base de données est de nature complexe.La conception de l'index doit être réalisée de manière à faciliter la recherche de documents, tout en tenant compte des critères de recherche spécifiques à chaque domaine d'application. Chaque utilisateur possède ses propres critères de pertinence, en fonction de ses besoins particuliers. Par conséquent, il n'existe pas de solution unique à ce problème. Deux approches principales sont distinguées pour définir la requête. La première approche est la recherche par caractéristiques, dans laquelle l'utilisateur "décompose" son objectif et décrit, à l'aide de primitives, ce qu'il recherche. La seconde approche est la recherche par l'exemple, dans laquelle l'utilisateur fournit un document en requête et le système recherche les documents les plus similaires au sein de la base de données.Dans ce travail, nous nous focaliserons principalement sur les documents écrits en anglais et en arabe, mais les idées et les concepts que nous introduisons peuvent également être appliqués, avec certaines adaptations, à d'autres supports (musique, image, photo, vidéo) et à d'autres langues (par exemple, allemand, espagnol, russe, chinois, japonais, etc...).

Les méthodes d'indexation comprennent en général les éléments suivants :

1. **La signature** ou *index* : ce processus consiste a utiliser un ensemble de caractéristiques d'indexation pour refléter le contenu sémantique des documents (ou des requêtes). Ces unités d'indexation sont le plus souvent des mots pour les textes, des notes de musique pour la musique, des valeurs de couleur pour les images, etc. Si nous nous limitons aux supports écrits, nous pouvons considérer les numéros de classe de thésaurus ou des formulations composées comme "traitement du langage naturel", ainsi que des mots simples comme "langage" ou "naturel", comme unités d'indexation. l'établissement de la signature d'un document textuelle doit également répondre a quelque questions, il faut décider si tous les détails d'un document doivent être pris en compte ou seulement ses caractéristiques essentielles, en tenant compte de l'importance du document par rapport à certains objectifs. De plus, il est nécessaire de déterminer le degré de précision ou de spécificité des termes d'indexation choisis.
2. Une **Métrique de similitude** (ou de distance) : une fonction qui permet de comparer les signatures et d'associer les documents similaires, le choix de cette fonction dépend lourdement de nos ambition, parmi les mesures de similarité les plus commune on retrouve :
   — **Mesure par appariement de mots**[11] cette mesure est basé sur l'alignement d'un texte par rapport à un autre, où chaque mot d'un texte est associé à un seul mot, au maximum, d'un autre texte. Ainsi, une mesure contrastive a été développée, dans une approche d'appariement entre les mots de chaque texte. Dans cette approche, chaque mot X_i de poids non nul du vecteur X est apparié à son plus proche voisin d'indice vois(i) de poids non nul du vecteur Y :

$$vois(i) = argmax_{\mathrm{iY}_j} \neq 0 m_i j \tag{1.1}$$

   — **Mesure entre représentations moyennes de textes**[11] :l'idée principale derrière cette approche est de calculer une représentation moyenne pour chaque texte en agrégeant les embeddings (représentations vectorielles) des mots qui le composent. Si v_i,est la k-ème composante du mot i dans l'espace des embeddings de mots, la k-ème composante du vecteur représentatif X' du texte X est obtenue par :

$$X'_k = \frac{1}{\sum_i (X_i)} \sum_i (x_i v_{i,k}) \tag{1.2}$$

   — **Mesure cosinus**[40] cette métrique est la plus commune dans le domaine de la recherche d'information car elle calcule l'angle entre les deux vecteur et de prendre en compte l'intégralité des deux documents comme suit :

$$\text{Similarité\_cosinus}(X, Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} \tag{1.3}$$

   La similarité cosinus varie entre -1 et 1, où une valeur de 1 indique une similarité maximale (les vecteurs sont parfaitement alignés), une valeur de -1 indique une similarité minimale (les vecteurs sont opposés) et une valeur de 0 indique une absence de similarité linéaire (les vecteurs sont orthogonaux).

3. Des **algorithmes de recherche** qui, bas´es sur les deux outils précédents, permettent de retrouver rapidement parcourir l'intégralité de la base donnée et retourner les documents rechercher.
4. **Interface utilisateur** qui rend transparente la procédure de recherche et facilite la communication avec l'utilisateur et l'introduction de la requête.

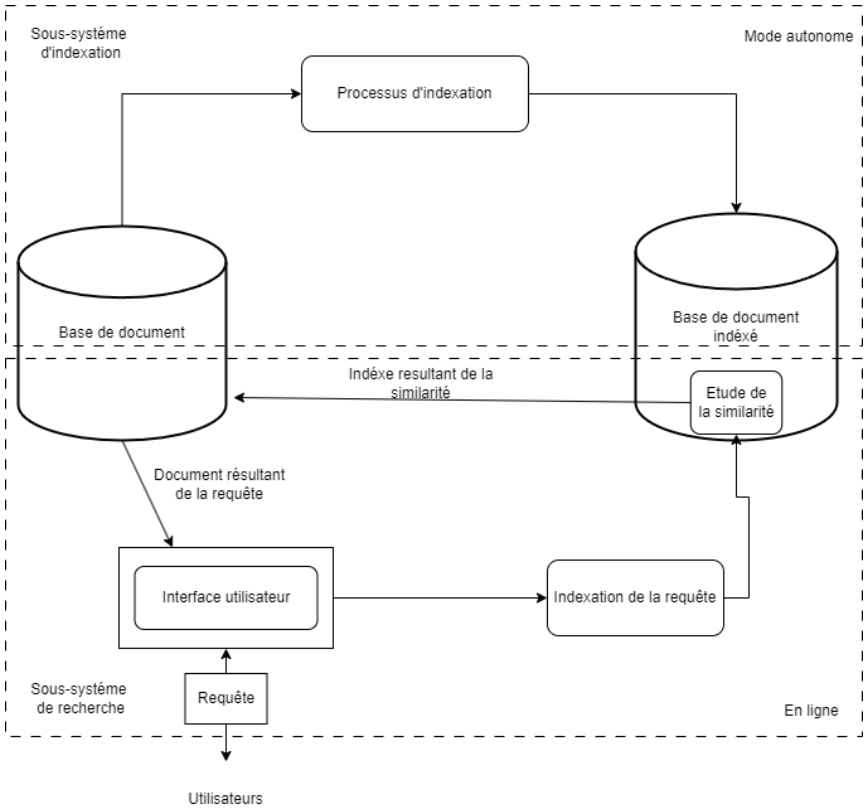


FIGURE 1.4 – Schéma générale de système de recherche d'information[39]

1.3 Travaux existants

1.3.1 Approches de la recherche d'information

Un système de recherche d'information (SRI)[21] est basé sur différents modèles théoriques de recherche d'informations qui déterminent comment les tâches d'indexation et d'appariement sont effectuées. mettre en œuvre. Toute IRM se compose de trois modèles, le modèle d'index de document, le modèle d'index de requête et le modèle de calcul. Pertinence des documents pour les requêtes.

Modèle IR classique

Les modèles classiques [1] de recherche d'informations, tels que les modèles IR vectoriels, booléens et probabilistes, sont largement utilisés et relativement faciles à mettre en œuvre. Ces modèles reposent sur des concepts mathématiques fondamentaux et sont conçus pour récupérer des informations en fonction de documents contenant des ensembles de requêtes spécifiées. Contrairement aux systèmes de recherche d'informations plus avancés, les modèles classiques n'impliquent pas de classements ou de scores. Au lieu de cela, ils se concentrent sur la correspondance entre les documents et les requêtes.

Modèle booléen

Le modèle booléen [29], en particulier, a été le premier modèle de recherche d'informations développé et est souvent critiqué pour sa simplicité. Ce modèle repose sur l'utilisation des opérateurs logiques (AND, OR, NOT) de George Boole[22] pour combiner les termes de requête et former de nouveaux ensembles de documents correspondants. Par exemple, une requête contenant les termes "économique" et "finance" avec l'opérateur AND retournera les documents qui contiennent à la fois les termes économiques et financiers.

Cependant, bien que le modèle booléen soit simple et facile à comprendre, il présente certaines limitations. Il ne tient pas compte des nuances sémantiques ou de la similarité entre les termes, ce qui peut conduire à une récupération d'informations moins précise. C'est pourquoi d'autres modèles plus avancés, tels que les modèles vectoriels et probabilistes, ont été développés pour améliorer la précision de la recherche d'informations.

Modèle vectoriel

Le modèle vectoriel[21] est un modèle largement accepté et étudié dans le domaine de la recherche d'informations. Introduit par Salton en 1970 [41] son succès peut être attribué à sa simplicité de conception et de mise en œuvre. L'utilisation de mesures de similarité par le modèle permet de trier les documents pertinents et de placer les plus similaires en haut de la liste. Cependant, le modèle vectoriel a ses limites. Par exemple, il ne peut pas tenir compte des interdépendances entre les termes d'indexation car chaque terme est

considéré comme indépendant. De plus, des critiques ont été faites concernant le manque de fondement théorique solide du modèle dans la représentation des documents et des requêtes, ainsi que dans sa fonction de correspondance.

Modèle probabiliste

Dans le domaine de la recherche d'informations, la modélisation probabiliste consiste à utiliser un modèle pour classer les documents en fonction de leur probabilité d'être pertinents pour les besoins d'information d'un utilisateur. L'objectif est de déterminer la probabilité que chaque document soit pertinent pour une requête donnée. Le premier modèle probabiliste de recherche d'informations a été introduit au début des années 1960 par Maron et Kuhn[47], et depuis lors, de nombreux modèles ont été développés, dont beaucoup sont basés sur le principe de classement des probabilités (PRP) décrit par Robertson[17]. Le PRP consiste à séparer les documents en deux classes - pertinentes (R) et non pertinentes (NR) - en fonction des besoins d'un utilisateur, et la tâche du modèle probabiliste est d'estimer la probabilité qu'un document appartienne à la classe pertinente.

Les Modèles	**Les Avantages**	**Les inconvénients**
Modèle booléen	-La simplicité du modéle. -Dans le cas de corpus explorés par des spécialistes, ayant notamment une très bonne connaissance du vocabulaire, cela le rend très efficace.	-Adapté à des recherches sur des corpus généralistes. - La formulation des requêtes devient vite laborieuse. - La pondération binaire des termes du vocabulaire limite la pertinence des résultats et ne permet pas de les ordonner.
Modèle vectorielle	- Relativement simple à appréhender et est facile à implémenter. -Son efficacité dépendant pour une grande part de la qualité de la représentation.	-Suppose que les termes représentatifs sont indépendants. -L'ordre des mots n'est pas pris en compte.
Modèle probabiliste	- Permet d'obtenir une vision plus réaliste et nuancée d'un problème. - Il permet également de quantifier l'incertitude associée à ces prédictions, ce qui est essentiel pour prendre des décisions éclairées.	- Complexes à construire et à ajuster en fonction des données disponibles. - Ils peuvent être sensibles à la qualité des données d'entrée.

TABLE 1.1 – Tableau comparatif des divers modéle de recherche d'information

Modèle graphique

Les modèles graphiques [20] fournissent une manière normalisée de représenter les relations entre les données et les ressources de connaissances, ce qui facilite leur organisation et leur raisonnement. Le type spécifique de raisonnement dépend du domaine d'étude et d'application. Dans la recherche d'informations, les modèles graphiques représentent les connexions entre les documents et les termes pour créer un système simple permettant de répondre rapidement et efficacement aux demandes des utilisateurs. Ce processus est également connu sous le nom d'indexation dans le domaine de la recherche d'informations.

Les réseaux bayésiens [20] sont un type de modèle graphique utilisé pour représenter les connaissances sur un domaine spécifique. La structure du graphique est utilisée pour montrer les relations entre les variables aléatoires, chaque nœud représentant une variable et chaque arête représentant la dépendance probabiliste entre elles. Essentiellement, un réseau bayésien est un moyen de modéliser la distribution de probabilité des variables dans le modèle.

Le problème principale de ce type d'approche est que lorsque la dimensionnalité des données augmente, l'estimation des paramètres du modèle peut devenir plus difficile, et l'inférence peut devenir coûteuse en termes de temps de calcul. Les modèles graphiques peuvent également être sensibles au phénomène de la "malédiction de la dimensionnalité", où la quantité de données nécessaire pour estimer les paramètres du modèle augmente exponentiellement avec le nombre de variables.

Modèle basé sur algorithme génétique (GA)

Les algorithmes génétique[36] sont une méthode adaptative utilisé pour résoudre les problèmes de recherche et d'optimisation initialement inspiré par la théorie de l'évolution de charles Darwin[13] et introduit par John Henry Holland[24] se sont des algorithmes qui se base sur le principe de la sélection naturelle et de l'évolution en attribuant une fitness fonction a chaque "individuelle".En faisant évoluer itérative ment une population de solutions et en utilisant des concepts tels que le croisement (crossover) et la mutation, les algorithmes génétiques peuvent explorer efficacement l'espace des solutions, converger vers de meilleures solutions et s'adapter aux conditions changeantes du problème. Cependant, les performances d'un algorithme génétique dépendent fortement de la conception de la fonction de fitness, des stratégies de sélection et des opérateurs génétiques, qui doivent être adaptés au problème spécifique en cours. La force des GA viens du fait que la technique est robuste et peut résoudre une large plage de problème, n'est pas garantie de trouver la solution optimal globale mais assure de trouver une solution grandement acceptable. Mais, ces derniers rencontre aussi de nombreuses limitations telles que la convergence prématurée, où l'algorithme se fixe sur une solution sous-optimale et échoue à exploiter l'intégralité des possibilités. des espace de recherche Cela peut se produire si la population converge trop rapidement ou reste piégée dans un optimum local, empêchant la découverte de meilleures solutions. Ils peuvent également avoir des difficultés à gérer les contraintes, en particulier dans les problèmes présentant des contraintes complexes et diverses. S'assurer que les solutions générées satisfont toutes les contraintes du problème peut être difficile,

et le processus de recherche peut nécessiter des mécanismes supplémentaires ou des modifications pour gérer efficacement les contraintes. Enfin les algorithmes génétiques peuvent être coûteux en termes de calcul, notamment lorsqu'ils traitent de grandes populations, de fonctions de fitness complexes et d'espaces de recherche de grande dimension. L'évaluation de la fitness de chaque individu de la population peut être chrono-phage, rendant les GA moins efficaces pour les problèmes ayant un coût de calcul élevé.

Modèle basé sur les colonies

les algorithmes basé sur les colonies inspirées de la nature[19] sont des algorithmes qui s'inspire du comportement sociale des animaux lors de la recherche de nourriture, souvent utilisé pour résoudre des problèmes d'optimisation ou de recherche l'idée principale de ces algorithmes est de modéliser le comportement collectif des animaux et d'exploiter leur capacité à trouver des solutions efficaces dans des environnements complexes. Les exemples les plus connus d'algorithmes de recherche basés sur les colonies animales incluent l'algorithme des colonies de fourmis (Ant Colony Optimization - ACO)[14] et l'algorithme de l'essaim de particules (Particle Swarm Optimization - PSO)[27].

Dans le contexte de l'ACO, les solutions potentielles à un problème sont représentées sous forme de chemins, et les fourmis artificielles suivent ces chemins tout en déposant des phéromones. Les phéromones sont ensuite évaporées avec le temps, ce qui permet aux chemins les plus courts d'accumuler plus de phéromones, ce qui attire davantage de fourmis vers ces chemins. Comme les algorithmes génétiques, au fil du temps, cette méthode converge vers une solution optimale. Tant qu'a l'algorithme de l'essaim de particules est inspiré du comportement de vol des oiseaux ou de la recherche de nourriture des abeilles. Les solutions potentielles sont représentées par des "particules" dans un espace de recherche multidimensionnel. Ces particules se déplacent dans l'espace de recherche en fonction de leur propre expérience et de l'expérience des meilleures particules voisines. Les meilleures solutions sont conservées et combinées pour guider les particules vers des régions prometteuses de l'espace de recherche. Cela permet de trouver progressivement des solutions de meilleure qualité. Bien que ces algorithmes ont prouver être efficace ces derniers ont aussi plusieurs inconvénients dont une lente convergence ou une grande sensibilité a topologie de l'espace de recherche, si l'espace de présente des caractéristiques complexes ou des optima locaux, les algorithmes peuvent avoir du mal à trouver des solutions de haute qualité. Ainsi qu'une grande dépendance au connaissance du problème. Par exemple, dans l'algorithme des colonies de fourmis, la modélisation des problèmes sous forme de graphes et la sélection appropriée des paramètres dépendent souvent de la structure du problème à résoudre.

Modèle base sur apprentissage en profondeur

L'apprentissage en profondeur est puissant car il peut aider à apprendre automatiquement des représentations de différentes données dans différentes tâches[31]. En réalité cette approche fait partie des plus répandue aujourd'hui. Les représentations apprises sont toutes sous la forme de vecteurs réels, également appelées représentations distribuées. De

cette manière, la correspondance en recherche d'information peut être réalisée à travers les représentations vectorielles, ce qui permet d'améliorer considérablement les performances de certaines tâches en recherche d'information et de mener à bien des tâches qui étaient auparavant considérées comme impossibles. Par exemple, il est possible d'apprendre directement des représentations à partir d'images et de leurs textes associés, d'exploiter ces représentations pour faire correspondre des questions et des images, et d'atteindre une grande précision dans la recherche d'images (e.g [26, 50]). Il a également été prouver possible d'améliorer la réponse traditionnelle aux questions à partir de document a l'aide de l'apprentissage en profondeur[44]. L'avantage avec cette approche est que non seulement elle fournis des résultat remarquable sur le terrain mais aussi qu'elle ne nécessite quasiment aucune connaissance linguistique dans la réalisation du système (e.g : [45]).

Néanmoins, il existe également de nombreux défis. La plus grande question est de savoir comment combiner le calcul neuronal (ou l'apprentissage profond) avec le traitement symbolique traditionnel, car les deux semblent nécessaires pour la recherche d'information, et c'est la qu'une toute nouvelle technique a vu le jour.

EBR(embedding based retrieval)[53] une technique de RI qui consiste d'améliorer l'efficacité et la qualité des résultats de recherche en combinant a la fois le calcul neuronal et règles explicites et représentations du langage utilisant de dense vecteurs capturant le sens sémantique ou contextuelle des mots ou document, avant d'en tirer conclusion soit en se servant comme la plus part des cas d'une métrique de distance (e.g : [40, 35, 15] , des algorithmes de clustering (e.g : [10, 28, 16] ou bien en passant par les mécaniques d'inférence.

Les EBRs offrent des avantages significatifs en termes de capacité à traiter des données complexes, de compréhension sémantique, d'apprentissage automatique des caractéristiques, d'amélioration continue et d'adaptabilité. Ces avantages peuvent améliorer la précision, la pertinence et l'efficacité de notre système de recherche d'information, ce qui en fait une option attrayante à considérer. Dans ce travail, nous allons exploiter les EBRs pour proposer une approche d'indexation des documents textuels. Cette approche sera détaillée dans le cahpitre 2.

1.3.2 Platformes open source dédiées aux données textuelles

Le but de ce travail est principalement de crée un moteur de recherche a base d'indexe pour ce faire il nous faut une plateforme ou déployer notre modèle une fois l'entraînement effectuer pour ça nous avons exploré un ensemble de plateforme open source de gestion et partage de contenu textuelle durant cette section nous discuterons les différents points de chacune d'entre elle.Nous sommes motivés par plusieurs causes qui ont conduit à la création de cette initiative.

Tout d'abord, l'absence de plateformes similaires en Algérie constitue une lacune importante. Il est essentiel de combler ce manque en offrant un espace centralisé où les textes algériens pourront être stockés, préservés et facilement accessibles à tous. En fournissant une telle plateforme, nous permettons aux écrivains, chercheurs, étudiants et passionnés

de la culture algérienne d'explorer et d'enrichir leur connaissance de manière pratique et efficace.

Ensuite, nous cherchons à préserver la richesse culturelle de l'Algérie. Les textes jouent un rôle fondamental dans la transmission de l'histoire, des traditions, des valeurs et des idées d'une nation. En créant une plateforme d'archivage, nous contribuons à préserver la mémoire collective du pays et à garantir que les œuvres littéraires, les documents historiques et autres textes importants ne soient pas perdus ou oubliés.

Enfin, en développant un moteur de recherche intelligent, nous améliorons l'accessibilité et la convivialité de la plateforme. Les utilisateurs pourront effectuer des recherches précises, découvrir de nouveaux textes pertinents et interagir de manière dynamique avec les contenus archivés. Cela facilitera la diffusion et le partage des connaissances, tout en encourageant l'innovation et la créativité parmi les utilisateurs.

Nom	Repo	Technology	point positive	point negative
BookStack	https ://github.com/BookStackApp/BookStack	Laravel	-Bon support de la comunauté -facile a modifier	
Open Encyclopedia	https ://github.com/open-encyclopedia-system/oes-demo	PHP	-core isolé -idée relative a la notre	-documentation indisponnible
MediaWiki	https ://github.com/wikimedia/mediawiki	PHP		pas de support de la communauté
Realms	https ://github.com/scragg0x/realms-wiki	Python	supporte les systéme de recherche customusé - supporte diver gestionnair de base de donnée	-plus mis a jour depuis 5ans -beaucoup de framework
Outline	https ://github.com/outline/outline	Typescript	-documentation detailé -architecteur orienté composants	ne posséde que Markdown comme éditeur de texte
Gollum	https ://github.com/gollum/gollum	Ruby	-met a disposition divers éditeur de texte	git-powered (pas beginer friendly pour l'utilisateur)
WikiJs	https ://github.com/Requarks/wiki	NodeJs	-core isolé -supporte divers gestionnaire de base de données -mets a disposition divers éditeur de texte - inclura un pdf scrapper dans le future	-difficile a metre en place

TABLE 1.2 – Divers plateforme de content management Textuelle

Comme on peut le constater dans le tableau 1.2 chacune des plateforme a ces bon et mauvais points. Cependant après plusieurs testes notre choix se portera sur la plateforme Wiki.JS de part ça rapidité et fluidité d'exécution, sa structure de base de donnée, la flexibilité qu'elle offre en terme de choix de SGBD nous offre ainsi la liberté d'échanger entre le SQL et le NoSQL selon nos désires, sa documentation détaillé, son architecture basé sur les micro services nous sera d'une grande utilité lors du déploiement de notre modèle et facilité d'utilisation (UI friendly).

1.4 Conclusion

Vu l'importance des moteur de recherche aujourd'hui, leur place est plus que jamais primordiale à l'élaboration de plateforme d'archivage capable de traiter des informations en temps réel et de les fournir de manière personnalisée aux utilisateurs.

Dans ce chapitre nous avons d'abord présenter le contexte de notre plateforme puis les raisons qui nous ont poussé a faire ce travail, ensuite nous avons également présenter de manière générale les système de recherche et d'indexation de texte avant de traiter des divers approche et travaux élaboré dans le domaine de la recherche d'information puis nous avons conclus par l'étude des plateforme de publication de document textuelle disponible en open-source pour la base de notre plateforme.

Quant au chapitre suivant abordera l'approche que nous avons choisi de suivre ainsi que notre solution globale et ses différente fonctionnalité.

Chapitre 2

Solution proposée

2.1 Introduction

La réalisation d'un logiciel ou d'un système informatique doit être obligatoirement précédée d'une étape d'analyse et de conception. Cette étape a pour objectif de définir et de formaliser les étapes nécessaires du développement de l'application afin de rendre cette dernière plus fidèle aux besoins.

Dans ce chapitre, nous allons présenter le schéma global de l'approche proposée, les différentes étapes de pré-traitement et d'indexation. Par la suite, nous présenterons le modèle d'embedding utilisé, puis parlerons des étapes d'intégration de notre modèle. Pour finir, nous présenterons les schémas conceptuels de notre plateforme globale El Mahroussa Tech.

2.2 Schéma globale de l'approche

Comme présenté dans la figure 2.1, notre approche consiste d'abord à traiter le document dès son arrivée en créant son index grâce à l'embedding. Cet index est ensuite sauvegardé et sera sollicité lorsque la requête arrive pour qu'elle soit traitée. Le calcul de similarité est effectué entre la requête et les documents indexés, et les K meilleurs documents sont renvoyés en tant que résultats.

1. Embedding Based Retrieval

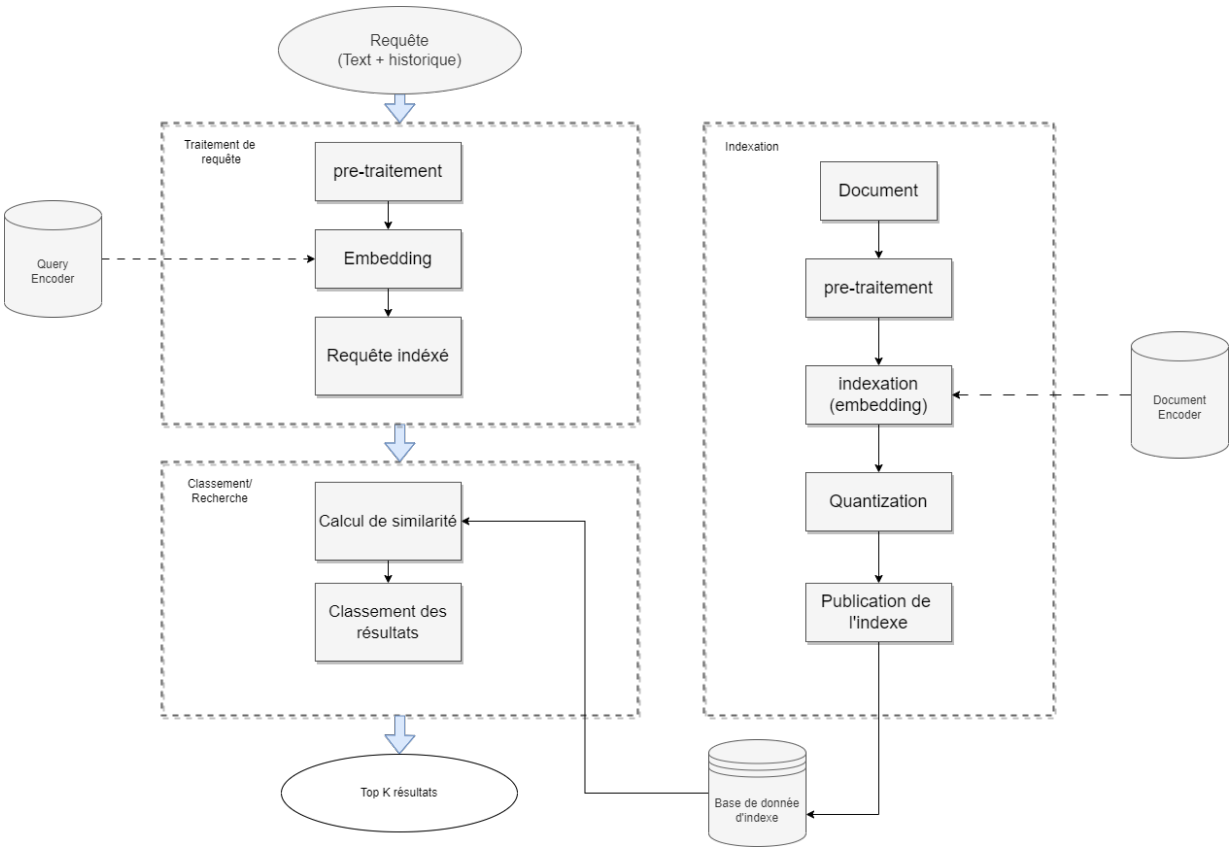


FIGURE 2.1 – Schéma globale du système de recherche EBR [1]

2.3 Approche proposée

Il est possible d'exprimer la tache de recherche d'information comme étant un problème d'optimisation de rappel. En d'autre terme, en partant d'une requête Q et d'un ensemble des documents ciblés T :

$$T = t_1, t_2, t_3, ...t_N$$

On obtient les top K documents retournés par le modèle,

$$d_1, d_2, d_3, ...d_k$$

Nous cherchons a maximiser notre rappel par rapport au top K résultat :

$$rappel@K = \frac{\sum_{i=1}^{K} d_i \in T}{N}$$

Afin de palier ce problème, nous illustrerons cette tache comme un problème de classement (ranking) basé sur le calcul de distance entre le document d et la requête Q. Ces derniers seront encodés à l'aide d'un modèle de réseaux de neurones, notamment appelé modèle d'embedding, dans des vecteurs denses sur qui nous appliquerons une similarité cosinus comme métrique de distance. Pour finir, nous nous servirons de la fonction de perte Triplet [43] afin d'approximer notre rappel.

2.4 Processus de pré-traitement des données

Dans le but d'alimenter notre modèle en données pour l'entraînement, plusieurs opérations de pré-traitements ont été effectuées sur chaque document afin de rendre le format des données conforme au format d'entrée de l'algorithme d'apprentissage. Pour ce faire, nous commençons par analyser le format de sauvegarde employé par wiki.Js. on remarque que la platforme utilise des balises HTML afin de préserver le format du texte (paragraphe, saut de ligne...ect), on prendra comme exemple le paragraphe suivant :

La lueur dorée du crépuscule caressait délicatement les sommets enneigés de la montagne majestueuse. Ci-après les étapes de pré-traitement suivies :

1. **Segmentation et suppression des mots vides** : Ces opérations permettent de récupérer tous les mots ou expressions significatifs d'un article donné. Nous entamons par diviser le texte en segments, puis nous éliminons les mots ou symboles qui ne sont pas pertinents, tels que la ponctuation, les pronoms et les déterminants. Après segmentation nous obtenons :
   "La", "lueur", "dorée", "du", "crépuscule", "caressait", "délicatement", "les", "sommet", "enneigés", "de", "la", "montagne", "majestueuse". Après suppression des mots vides, nous nous retrouvons avec : "lueur", "dorée", "crépuscule", "caressait", "délicatement", "sommets", "enneigés", "montagne", "majestueuse".

2. **Racinisation (Stemming)** : Cette étape consiste à extraire la racine de chaque mot d'un document et de le représenter ainsi. Cela permet de regrouper plusieurs variante d'un même terme dans un seul mot et, par conséquent, réduire le nombre de mots nécessaires à la représentation d'un document. Un exemple de résultats après l'étape de racinisation :
"lueur", "doré", "crépuscul", "caress", "délicat", "somm", "enneig", "montagn", "majestueus".

2.5 Modèle d'embedding et calcule de similarité

2.5.1 Indexation : vectorization des données

Une fois le dataset pre-traité, la prochaine étape consiste à vectoriser et formater le jeu de données afin de l'utiliser pour entraîner le modèle. Cependant, ce processus ne se limite pas simplement à la conversion des données en vecteurs. Une étape clé dans ce processus est d'appliquer une vectorisation des données, ce qui permet de convertir un document textuel à un vecteur numérique ainsi que les requêtes. Nous allons utilisé de l'embedding pour généré cette représentation vectorielle des documents textuels. Ces vecteurs serviront ensuite de base pour entraîner notre modèle.

Nous allons choisir la fonction de perte Triplet pour l'optimisation de notre représentation embeddind. Pour cela, notre modèle doit comprendre trois éléments : un encodeur de document appelé *Document Encoder* $E_D = g(D)$ qui générera l'embedding des documents, un autre encodeur pour la requête appelé *Query Encoder* $E_Q = f(Q)$ qui se chargera de générer l'embedding des requêtes, et enfin une mesure de similarité $S(E_Q, E_D)$ qui s'occupera de calculer le score de la requête Q et du document D comme illustré dans la figure 2.2. Ces encodeurs, sont des réseaux de neurones qui transforment une entrée en un vecteur dense de faible dimension.

Pour ce qui est de la fonction de similarité, nous avons opté pour la similarité cosinus pour les trois raisons suivantes :

Premièrement, *l'indépendance linguistique* : La similarité cosinus est indépendante de la langue, ce qui signifie qu'elle peut être appliquée à des documents représentés sous forme de vecteurs, quel que soit le langage utilisé. Cela nous convient particulièrement car nous cherchons à déployer notre modèle sur deux langues très différentes l'une de l'autre.

Deuxièmement, *la mesure intuitive* : La similarité cosinus fournit une mesure simple et intuitive de la similarité entre deux vecteurs. Elle calcule le cosinus de l'angle entre les vecteurs, ce qui représente la similarité en terme d'orientation plutôt que de magnitude. Cela convient aux tâches de recherche d'information où l'accent est souvent mis sur la capture de la similarité sémantique ou de la relation entre les documents plutôt que sur leurs valeurs absolues.

Enfin, *la popularité* : La similarité cosinus est la mesure de similarité la plus sollicitée lorsqu'il s'agit de recherche d'information, indépendamment de l'algorithme ou de l'approche utilisée.

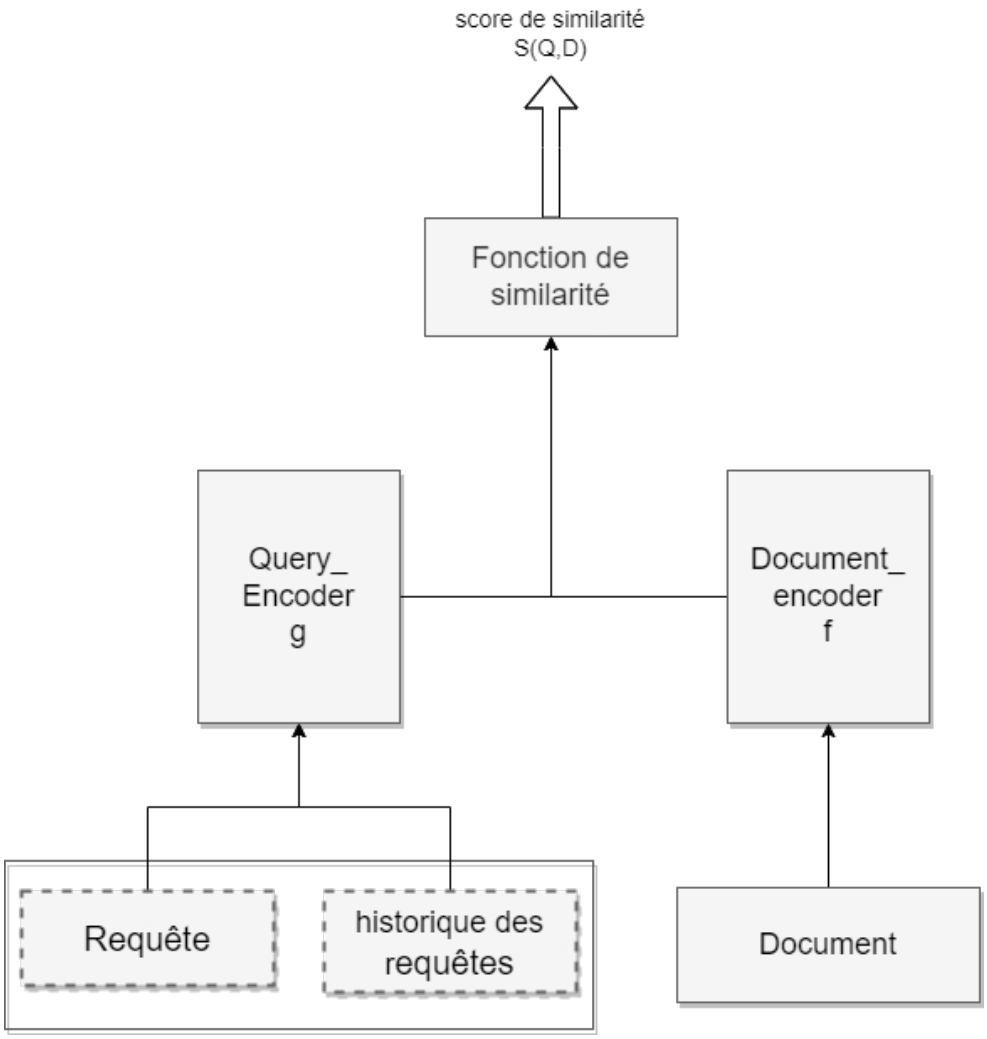


FIGURE 2.2 – Schéma du modèle embedding

Algorithm 1 Algorithme de calcule de similarité entre une requête et des documents

Input : Q : requête, $ListeDocument$: documents
Output : $taux\_de\_similarité$: reel
segmentation(requête)
suppression_mots_vides(requête)
racinisation(requête)
pour chaque document dans ListDocument faire
(document)
suppression_mots_vides(document)
racinisation(document)
fin
$anchor$ = Vectorization(requête)
$documents$ = Vectorization(ListeDocument)
$tauxdesimilarité$ = Tri(cosinus_similarité($anchor$,$documents$))
retourner $mat\_similarité$[: 5]

2.6 Déploiement du modèle

Une fois l'entraînement terminé, il nous faudra évidemment un moyen de déployer notre modèle, ainsi qu'un pont de communication pour permettre à notre plateforme d'échanger des données avec le modèle.

Pour cela, nous utiliserons une API (Application Programming Interface), qui est un ensemble de règles, de protocoles et de définitions permettant à différentes applications de communiquer entre elles. Elle définit les méthodes et les formats de données que les applications doivent utiliser pour interagir les unes avec les autres. La figure 2.3 reflète la manière dont notre API assure la communication entre l'interface client et le modèle intelligent.

Cette API peut prendre plusieurs formes, mais elle est généralement constituée d'un ensemble de points de terminaison appelés "endpoints", qui correspondent à des URLs. Chaque endpoint est associé à une action ou une opération particulière effectuée par l'application.

L'API que nous allons construire sera basée sur la norme de facto pour les API Web[34], également appelée REST (Representational State Transfer). Elle repose sur l'utilisation des méthodes HTTP telles que GET, POST, PUT et DELETE pour effectuer des opérations sur les ressources. Cela est souvent réalisé via un ASGI HTTP (Asynchronous Server Gateway Interface), qui fournit une interface normalisée entre les serveurs Web compatibles async, les frameworks et les applications Python. C'est grâce à cette interface que la communication entre le modèle et les clients s'établira.

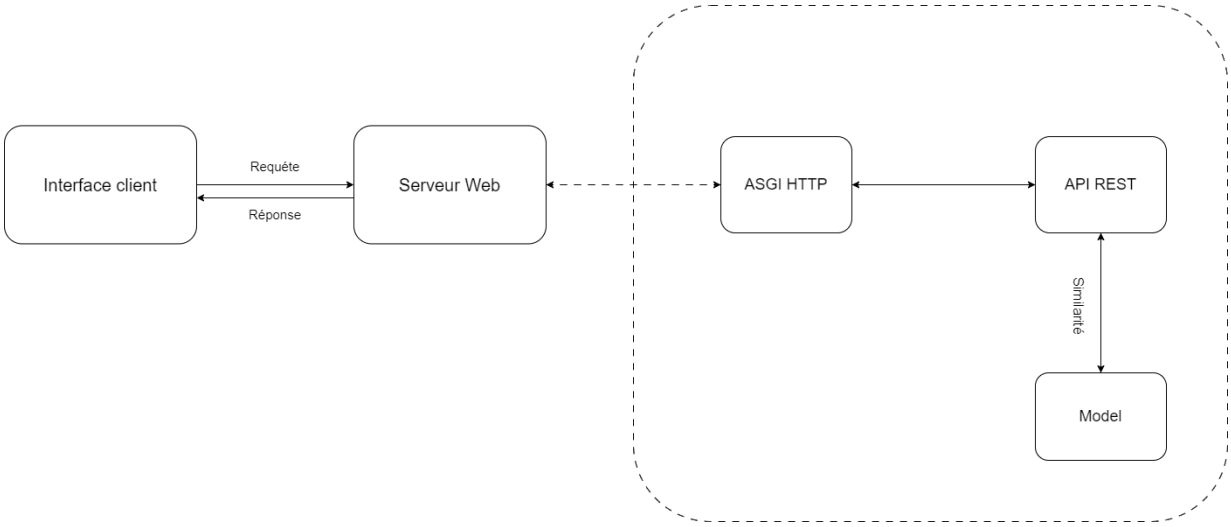


FIGURE 2.3 – Schéma de communication entre client et modèle via API

2.7 Platforme El Mahroussa Tech

Bien que ce travail se focalise principalement sur l'archivage de documents textuels, notre plateforme générale aura également un volet dédié à l'archivage de vidéos. Le tout sera regroupé au sein d'une plateforme principale appelée El Mahroussa Tech. Afin de réaliser cela, nous devrons effectuer une conception de la structure de la base de données ainsi que de l'architecture système.

2.7.1 Conception de base donnée

Afin de s'adapter à nos besoins très stricts en termes de structure, nous avons opté pour une base de données SQL. Cela nous offrira un certain degré d'intégrité et de sécurité pour nos données, au prix de la flexibilité que nous aurait offerte le NoSQL.

Le SGBD PostgreSQL employé est un système de gestion de base de données relationnelles (SGBDR) open source. Il offre une grande fiabilité, une extensibilité et une conformité aux normes SQL. PostgreSQL a été initialement développé à l'Université de Californie à Berkeley dans les années 1980, et il est aujourd'hui maintenu et amélioré par une communauté de développeurs à travers le monde[46].

WikiJS possède déjà sa propre structure de base de données supportée par PostgreSQL. Cependant, nous apporterons quelques modifications à celle-ci :

1. Dans la table des documents (nommée "pages"), nous ajouterons un attribut appelé "index". Cet attribut sera utilisé pour stocker l'index du document après l'application de notre modèle d'embedding. Cette opération sera effectuée lors de la publication d'un nouveau document. Cela nous permettra de ne pas avoir à exécuter l'opération sur l'ensemble de la table à chaque recherche.

Après avoir effectué cette modification, chaque document sera représenté de la manière suivante :
- id : l'id du document
- path : contient l'Url vers la page du document
- title : le titre
- isPrivate : status du document True si il est privé False pour public (False par defaut)
- content : contient le text corp du document
- createdAt : date de création
- updatedAt : derniére date de modification
- authorId : l'id du l'auteur
- index : l'index du document apres embedding

2. La table userHistory : celle ci sera crée et contiendra les 5 dernière requêtes effectuer par l'utilisateur. la table sera représenté comme suit :
   - userID : l'id de l'utilisateur en question
   - queries : regroupe les 5 requêtes plus récente de l'utilisateur

Pour la page principale, la base de données sera constituer d'une seul table "User". Celle-ci servira à alimenter la table "user" des deux autres plateformes de données. La représentation de cette table sera la suivante :
- Nom
- Prénom
- Email
- isVerified
- mot de passe

2.7.2 Architecture système

Dans cette partie, nous allons présenter l'architecture globale de notre système. Cette dernière est déployée sous la forme d'une architecture Micro-services (AMS)[18]. Cette architecture propose une solution basée sur le découpage d'une application en petits services autonomes appelés micro-services. Chaque micro-service expose une API REST que les autres micro-services peuvent utiliser et consommer. Cette approche permet une plus grande modularité, une évolutivité simplifiée et une meilleure gestion des fonctionnalités de l'application.

Présentation de l'architecture

Les micro-services désignent à la fois une architecture et une approche de développement logiciel. Cette approche consiste à décomposer les applications en éléments simples et indépendants les uns des autres. Contrairement à une approche monolithique classique, où tous les composants sont regroupés en une seule entité, les micro-services fonctionnent en synergie pour accomplir les mêmes tâches, tout en étant séparés. Chaque composant ou processus représente un micro-service, qui est granulaire et léger. Cette approche permet d'utiliser des processus similaires dans plusieurs applications, favorisant ainsi la réutilisation du

code et la flexibilité. Les micro-services sont un élément clé pour optimiser le développement des applications en vue de l'adoption d'un modèle cloud-native, où les applications sont conçues pour être déployées et fonctionner de manière efficace dans un environnement cloud. La figure 2.4 illustre cette architecture.

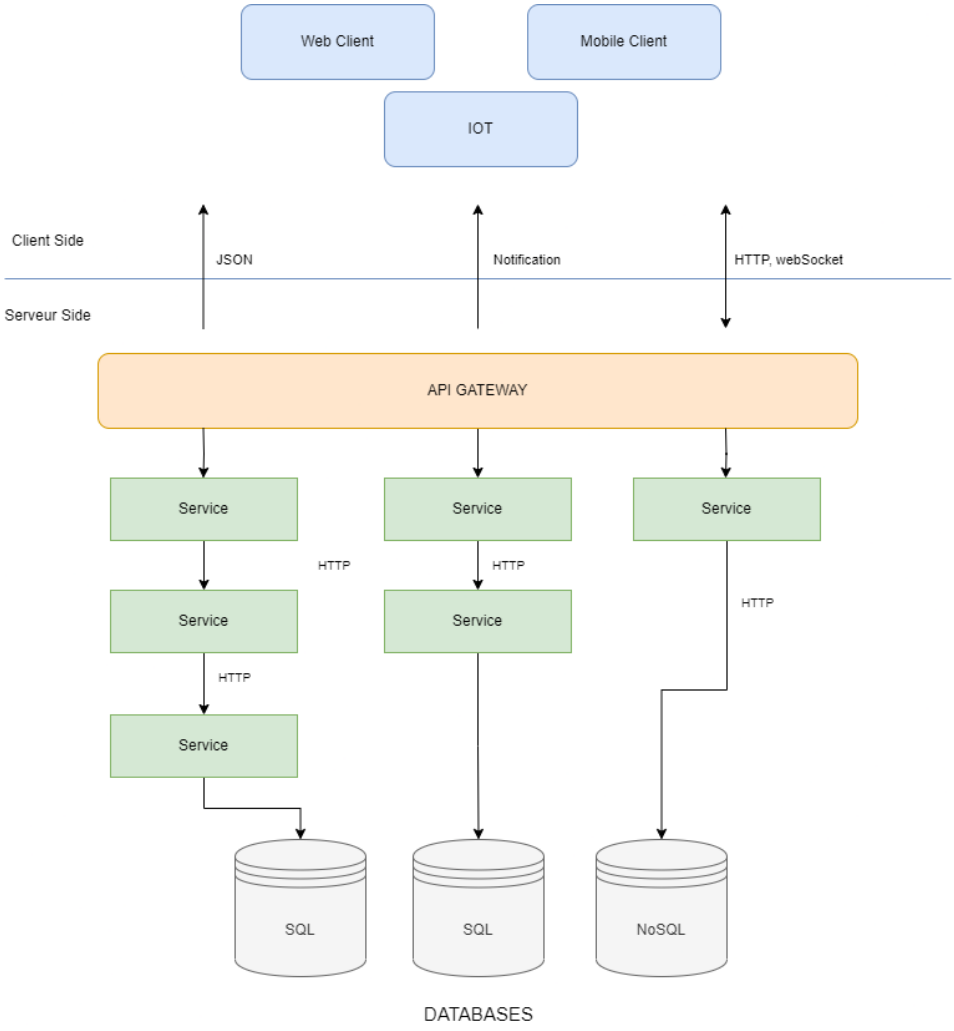


FIGURE 2.4 – Schéma de l'architecture micro-service [2]

2. www.pinterest.com/pin/435301120209947864/

Avantage de l'architecture micro-service

Parmi les avantages de l'utilisation du modèle architectural des micro-services pour les applications Web :

- Scalabilité : L'architecture des micro-services facilite le passage à l'échelle des applications Web en fonction des besoins. Cela est possible car chaque service peut être mis à l'échelle de manière indépendante.
- Flexibilité : L'architecture des micro-services facilite l'ajout de nouvelles fonctionnalités aux applications Web. Cela est possible car chaque service peut être développé et déployé de manière indépendante.
- Résilience : L'architecture des micro-services rend les applications Web plus résilientes aux pannes. Cela est possible car chaque service peut être redémarré ou remplacé de manière indépendante.
- Maintenabilité : L'architecture des micro-services rend les applications Web plus faciles à entretenir et à mettre à jour. Cela est possible car chaque service est autonome et peut être développé et déployé de manière indépendante.

Dans l'ensemble, le modèle architectural des microservices est un bon choix pour le développement d'applications Web complexes. Il offre la scalabilité, la flexibilité, la résilience et la facilité de maintenance.

2.8 Conclusion

Dans ce chapitre, nous avons présenté la conception de la plateforme "El Mahroussa Tech". Nous avons commencé par l'approche de notre système de recherche et d'indexation d'information, ainsi que son schéma global. Nous avons également introduit le principe des API et du déploiement des modèles intelligents. Enfin, nous avons abordé les schémas conceptuels de "El Mahroussa Tech".

Dans le chapitre suivant, nous aborderons l'implémentation et les résultats de notre modèle, ainsi que les étapes de réalisation de notre plateforme "El Mahroussa Tech" et de son hébergement.

Chapitre 3

IMPLÉMENTATION ET RÉSULTATS

3.1 Introduction

Après avoir finalisé l'étape de conception nous allons dédiez ce chapitre a la réalisation et l'évaluation. Les diverses problématiques ont fait l'objet d'une analyse approfondie, ce qui a engendré la mise en œuvre des modules de développement visant à atteindre un produit final fonctionnel pour les utilisateurs.

Nous allons d'abord présenter l'environnement de travail ainsi que les outils et langages employés dans la réalisation du travail. Nous présenterons également en détails les différente étapes de réalisation du modèle et de son évaluation ainsi que de son déploiement au sein de la plateforme. Enfin, nous clôturerons par la suite les divers étapes d'hébergement de notre plateforme finale.

3.2 Environnement de développement

Dans cette section nous allons introduire les machines, outils et langages utiliser tout le long du travail :

3.2.1 Caractéristiques des machines

Pendant tout le projet, nous avons utilisé deux différentes machines physique pour la partie développement ainsi qu'un serveur cloud pour effectuer pour le traitements gourmands en terme de ressource et temps d'exécution.

1. Poste de travail 1

Système d'exploitation	GNU/Linux Elementory OS 18.0 64bits
RAM	16GB
CPU	Intel core I7-7700K 4.5GHZ

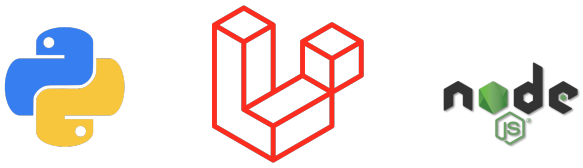
2. Poste de travail 2

Système d'exploitation	Windows 11 64bits
RAM	16GB
CPU	Intel core I7-12700k 5.0GHZ

3. Serveur (Machine virtuelle Cloud)

Système d'exploitation	GNU/Linux Ubuntu Data Science 64bits
RAM	32GB
CPU	Intel Xeon CPU E5-2673 v4 @ 2.295GHz

3.2.2 Langages de programmation et logiciels



(a) (b) (c)

FIGURE 3.1 – Logos des langages de programmation utilisés

1. **Python** : Python est un langage de programmation de haut niveau. Il supporte la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort et d'une gestion automatique de la mémoire. Plusieurs bibliothèques sont fournies afin de faciliter les développements[2].
2. **Laravel** : Laravel est un framework d'application web doté d'une syntaxe expressive et élégante. Un framework web fournit une structure et un point de départ pour la création de votre application, vous permettant de vous concentrer sur la création de quelque chose d'exceptionnel tandis que nous nous occupons des détails techniques[3].
3. **Nodejs** : NodeJS est un outil libre codé en Javascript et orientée pour des applications en réseau. Cet outil JavaScript est devenu célèbre dans l'univers du développement web depuis quelques années. D'ailleurs, il est très apprécié des géants du web comme Netflix, PayPal, LinkedIn, Uber, la NASA, etc. [4].

3.2.3 Librairies et bibliothèques

Dans ce qui suit, nous présentons les langages et bibliothèques utilisés pour mise en oeuvre de notre plateforme :

1. **Pytorch** : PyTorch est une bibliothèque d'IA, développée par Meta (ex-Facebook), écrite en Python pour se lancer dans le deep learning (ou apprentissage profond) et le développement de réseaux de neurones artificiels. À partir de plusieurs variables, elle peut servir à réaliser des calculs de gradients ou à utiliser des tableaux multidimensionnels obtenus grâce à des tenseurs[38].
2. **re** : la bibliothèque python standard "re" (pour "regular expression", expressions régulières en français) fournit des fonctionnalités pour travailler avec des motifs de correspondance de texte. Elle permet de rechercher des motifs spécifiques dans des chaînes de caractères, d'effectuer des opérations de recherche et de remplacement avancées, ainsi que de diviser une chaîne de caractères en fonction d'un motif donné[48].
3. **Farasa** : L'équivalent arabe de "perspicacité", Farasa est une Toolbox de traitement du langage naturel arabe développé au sein de l'institut Qatar Computing Research Institute. Elle est composée de plusieurs modules : segmentation, étiquetage, etc. Farasa surpasse ou égalise les deux fameuses Toolbox pour l'arabe Stanford NLP et MADAMIRA[5].
4. **NLTK** : est une bibliothèque utilisée pour créer des programmes Python qui travaillent avec des données de langage humain pour les appliquer dans le traitement automatique du langage naturel (TALN) statistique. Elle contient des bibliothèques de traitement de texte pour la tokenisation, l'analyse syntaxique, la classification, la racinisation, l'étiquetage et le raisonnement sémantique. Elle comprend également des démonstrations graphiques et des ensembles de données d'exemple, ainsi qu'un livre de recettes et un livre expliquant les principes sous-jacents des tâches de traitement du langage prises en charge par NLTK[9].
5. **Scikit-learn** : Scikit-learn est une bibliothèque essentielle pour le langage de programmation Python, généralement utilisée dans des projets d'apprentissage automatique. Scikit-learn se concentre sur les outils d'apprentissage automatique, y compris les algorithmes mathématiques, statistiques et à usage général, qui constituent la base de nombreuses technologies d'apprentissage automatique. En tant qu'outil gratuit, Scikit-learn est extrêmement important dans de nombreux types de développement d'algorithmes pour l'apprentissage automatique et les technologies connexes [25].
6. **Numpy** :NumPy est le package fondamental pour le calcul scientifique en Python. C'est une bibliothèque Python qui fournit un objet de tableau multidimensionnel, divers objets dérivés (tels que des tableaux masqués et des matrices) et une gamme de fonctions pour des opérations rapides sur les tableaux, comprenant des opérations mathématiques, logiques, de manipulation de forme, de tri, de sélection, d'E/S, des transformées de Fourier discrètes, de l'algèbre linéaire de base, des opérations statistiques de base, des simulations aléatoires et bien plus encore.

7. **FastAPI** :FastAPI est un framework Python moderne et rapide pour la création d'API web. Il est basé sur les annotations de type Python pour une meilleure validation des données et une documentation automatique.
8. **Psycopg2** : est une bibliothèque Python largement utilisée pour interagir avec des bases de données PostgreSQL. Elle fournit une interface efficace et simple à utiliser pour établir une connexion à une base de données PostgreSQL, exécuter des requêtes SQL, récupérer les résultats et gérer les transactions[49].
9. **Uvicorn** : Uvicorn est un serveur Web asynchrone basé sur Python qui permet de déployer rapidement des applications web basées sur des frameworks tels que FastAPI et Starlette. Il est conçu pour offrir des performances élevées grâce à son support natif de l'asynchronisme, ce qui le rend adapté aux applications nécessitant une manipulation efficace des requêtes HTTP simultanées. Uvicorn est facile à configurer et à utiliser, et il est souvent utilisé comme serveur de développement ou pour le déploiement d'applications web en production[30].

3.3 Corpus et base de données

En vue d'implémenter l'approche proposer, nous avons tenter d'utiliser des corpus destiner a la recherche d'information textuel. Pour ce faire nous avons tenté de solliciter *Text REtrieval Conference* (TREC) [51] en raison de ça grande réputation, de ça fiabilité et de sa disponibilité en plusieurs langue en question de recherche d'information mais sans réponse de la part de l'entreprise nous étions dans l'obligation d'explorer d'autre solution alternative décrites ci-dessous.

3.3.1 ir_datasets

ir_datasets[32] est une bibliothèque Python open-source conçue spécifiquement pour faciliter l'accès à une large gamme de base de donnée textuelles utilisés dans le domaine de la recherche d'informations. Parmis la panoplie de base de donnée qu'offre *ir_datasets* on retrouve les corpus suivants

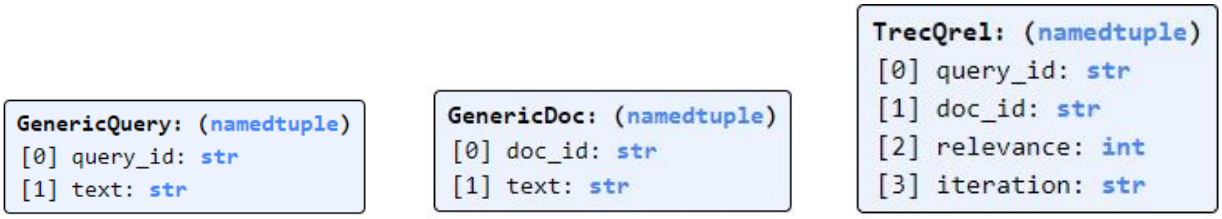
— ANTIQUE
— AQUAINT
— BEIR (benchmark suite)
— C4
— ClueWeb09
— MSMARCO
— mr-tydi
— NFCorpus

Après une longue sélection seul les deux base de donnée suivante finissent par nous intéresser :

Antique

Antique[23] est une collection de questions réelles et de jugements de pertinence qui ont été développés pour améliorer la recherche de passages de réponse dans les systèmes modernes de recherche d'informations. Les questions ont été prélevées à partir d'un large éventail de catégories sur Yahoo ! Answers[6].

1. **structure de la base de donnée**
   la base d'*Antique* se divise en 3 sous ensembles :
   — antique/train : ensemble d'entraînement officiel du jeu de donnée Antique.
   — antique/test : ensemble de test officiel du jeu de donnée Antique.
   — antique/test/non-offensive : antique/test sans un ensemble de requêtes jugées par les auteurs d'*ANTIQUE* comme "offensantes" (bruyante)
   chaqu'un de ses sous ensembles se compose de 3 tables la figure 3.2 ci-dessous montre la structure ainsi que le type de chqu'un des attribus des tables des collections de données :



(a) requête (b) document (c) qrel

FIGURE 3.2 – structure de la base *Antique*[1]

   Dans le cadre d'entraîner et d'évaluer notre modèle nous n'aurons recours qu'aux ensembles Antique/train et antique/test/non-offensive.

2. **Statistique du jeu de donnée** La pertinence des documents par rapport au requête et est distribuer comme suit :
   — label 4 : Cela semble raisonnable et convaincant. Sa qualité est au rendez-vous avec ou mieux que la "réponse possiblement correcte". Noter que il n'a pas à fournir la même réponse que le "Peut-être Bonne réponse".
   — label 3 : Cela peut être une réponse à la question, cependant, ce n'est pas suffisamment convaincant. Il devrait y avoir une réponse avec beaucoup meilleure qualité pour la question.
   — label 2 :Elle ne répond pas à la question ou si elle y répond, elle fournit une réponse déraisonnable, cependant, elle n'est pas hors contexte. Donc, vous ne pouvez pas l'accepter comme réponse à la question.
   — label 1 : Elle est complètement hors contexte ou ne fait aucun sens.
   Le tableau 3.1 et graphe 3.3 illustre la répartition des réponses de la base de donnée *Antique* :

1. ir-datasets.com/antique

# entraînement (test) question	2,426(200)
# entraînement (test) réponse	27,422(6589)
# mot/question	10.51
# mot/réponse	47.75

TABLE 3.1 – statistique d'Antique[23]

Quant au réponses, elles sont répartie comme suit :

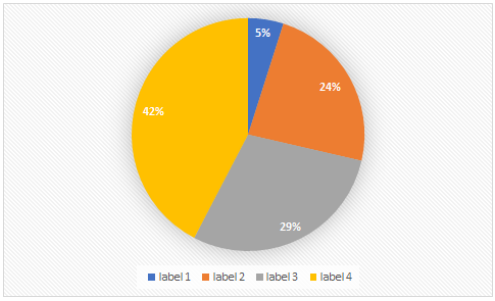


FIGURE 3.3 – Distribution de la pertinence des documents sur les questions d'Antique

3. **Avantages et inconvénients** L'utilisation d'une base de donnée de question non factuelle tell qu'*Antique* dans la réalisation d'un système de recherche d'information permettra au modèle de se familiarisé avec la future incertitude et incomplétude des requête des utilisateurs plus tard, de proposer une variété de question et de mieux évaluer le potentiel de notre approche. L'inconvénient est que la base de donnée est assez peut balancer qu'elle est basé sur l'opinion de plusieurs personnes et donc nous nous retrouvons avec plusieurs doublant que nous devrons traiter plus tard se qui va impacter la taille de donnée.Ajouté à cela la masse importante de données que nous ne pouvons modifier au risque de biaiser les résultats et la difficulté de traitements des fichiers dans un poste de travail (1 ou 2) ce qui nécessite l'utilisation du serveur.

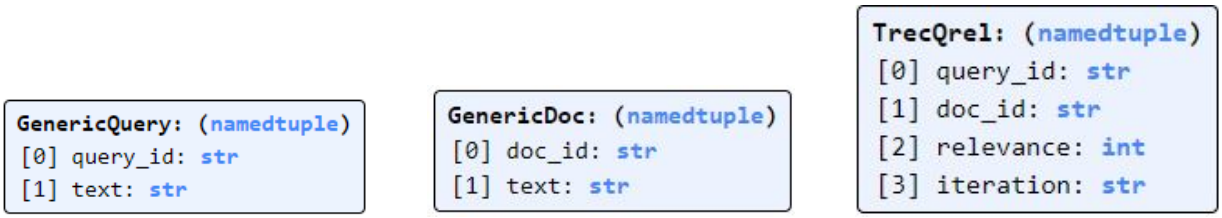
Mr-tydi

Mr-tydi[54, 12] est une suite de base de données multilingue construite à partir du TyDi QA Benchmark[54, 12]. Conçues pour évaluer le classement avec des représentations denses apprises. la base est conçu de onze langues typologiquement diverses, mais nous utiliserons uniquement celle en langue arabe *Mr-tydi/ar*.

1. **Structure de la base de donnée**

   la base de donnée *Mr-tydi/ar* se divise en 3 sous ensembles :
   - **mr-tydi/ar/dev** : ensemble de développement officiel du jeu de donnée *Mr-tydi/ar*.
   - **mr-tydi/ar/test** : ensemble de test officiel du jeu de donnée *Mr-tydi/ar*.
   - **mr-tydi/ar/train** : ensemble d'entraînement officiel du jeu de donnée *Mr-tydi/ar*.

Chacun de ses sous ensembles se compose de 3 tables la figure 3.4 ci-dessous montre la structure ainsi que le type de chqu'un des attribus des tables des collections de données :



(a) requête (b) document (c) qrel

FIGURE 3.4 – Structure de la base Mr-tydi [2]

Dans le cadre d'entraîner et d'évaluer notre modèle nous n'aurons recours qu'aux ensembles Antique/train et antique/test/non-offensive.

2. **Statistique du jeu de donnée** la base de donnée est composé de plus de deux millions de document et de 12,377 questions les paires de question réponse sont jugé comme suit :
   — 1 : Passage identifié comme pertinent.
   — 0 : passage identifié comme non pertinent.

Le tableau 3.2 et graphe 3.5 illustre la répartition des réponses de la base de donnée Mr-Tydi/ar : Quant au réponses, elles sont répartie comme suit :

# entraînement (test) question	12,377(1,081)
# entraînement (test) réponse	2,089,837(1,257)
# mot/question	12.32
# mot/réponse	51.67

TABLE 3.2 – statistique de Mr-Tydi/ar[23]

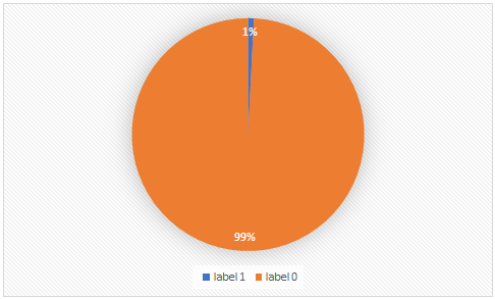


FIGURE 3.5 – Distribution de la pertinence des documents sur les questions de Mr-Tydi/ar

2. ir-datasets.com/mr-tydi.html.mr-tydi/ar

3.4 Métriques d'évaluation

Le but principal de notre projet est de développer un système de recherche de bout en bout, qui sera évalué à travers des métrique d'évaluation. L'élaboration d'une métriques hors ligne implique de définir des critères d'évaluation spécifiques pour mesurer la pertinence, la précision et l'efficacité de notre système de recherche. Cela nous permettra de quantifier les résultats obtenus par rapport à des réponses de référence, des annotations humaines ou d'autres indicateurs de qualité pertinents. Les métriques hors ligne peuvent inclure des mesures telles que la précision, le rappel, le F1-score, la similarité cosinus, ou encore des mesures spécifiques à notre domaine d'application.

Une fois que nous aurons développé ces métriques hors ligne et établi un ensemble de références de qualité, nous serons en mesure d'évaluer notre modèle en comparant ses résultats aux résultats attendus. Cette évaluation nous permettra de comprendre les forces et les faiblesses du système et d'identifier les domaines où des améliorations sont nécessaires.

En procédant ainsi, nous pourrons peaufiner notre modèle et affiner ses performances avant de le déployer dans un environnement en ligne. Cette approche préliminaire hors ligne nous permettra de gagner du temps et de l'efficacité en isolant les complications spécifiques à l'installation en ligne et en nous concentrant sur l'évaluation précise de la qualité du modèle lui-même.

Pour évaluer notre modèle nous aurons recours au métrique d'évaluation suivante

Fonction de perte triplet

La fonction de perte triplet[42] est une fonction qui mesure à quel point les prédictions d'un modèle diffèrent des valeurs réelles attendues. Pour ce faire pour chaque triplé donnée (q^i, d^i_+, d^i_-) où q^i représente la requête de l'utilisateur d^i_+ le document positif a la requête et d^i_- le document négatif a la requête, respectivement la perte triplet est définit comme telle :

$$L = \sum_{i=1}^{N} max(0, D(q^i, d^i_+)D(q^i, d^i_-) + m)$$

Avec $D(u, v)$ étant la distance entre les vecteurs u et v, et m la marge appliqué entre les paires positive et négative, ainsi que N étant le nombre totale de triplé sélectionnée depuis notre ensemble de données. L'intuition de cette fonction sera de séparé la paire positive de la paire négative par une marge de distance. il est aussi important de souligné que le réglage de la valeur de la marge a son impact sur les performances du modèle, la valeur optimale de la marge varie considérablement selon les différentes tâches d'entraînement, pour affiner cette dernière il a suffit d'appliquer un Grid search nous nous sommes retrouver avec une valeurs de marge de 1.

Matrice de confusion

La matrice de confusion présente les résultats des prédictions en comparaison avec les vraies étiquettes des données. Ses éléments indiquent le nombre de prédictions correctes et incorrectes, permettant ainsi d'analyser les erreurs de classification, les taux de vrais positifs, de faux positifs, de vrais négatifs et de faux négatifs, la figure 3.6 illustre la répartition de cette matrice.

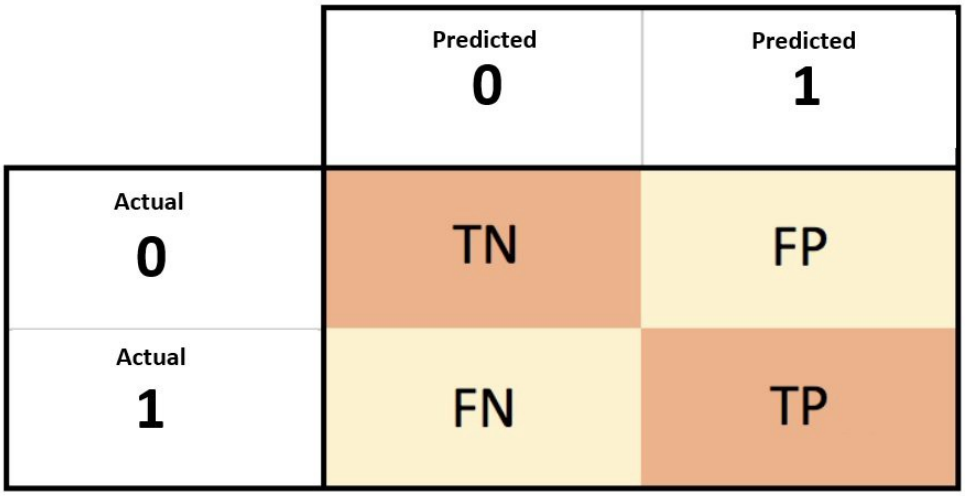


FIGURE 3.6 – Matrice de confusion

Grâce a cette matrice il est possible d'extraire plusieurs métrique d'évaluation du modèle :

— Précision

$$Precision = TP/(TP + FP)$$

— Rappelle

$$RecallScore = TP/(FN + TP)$$

— F1-Score

$$F1Score = 2 * PrecisionScore * RecallScore/ \\ (PrecisionScore + RecallScore)$$

3.5 Pré-traitement et structure des corpus

Avant de passer a l'entraînement de notre système il est nécessaire d'analyser le format des deux corpus choisi et d'identifier toute restructuration et pré-traitement nécessaire affin de permettre l'exploitation de ces derniers.

1. **Suppression de doublons** : nous remarquons l'apparition de doublons dans le premiers corpus la même réponse a la même question avec deux valeur de pertinence différente cela est du au fait qu'*Antique* soit basé sur des points de vu de plusieurs personnes et donc il arrive que deux juriste est des avis diffèrent sur es-que tell document répond a telle question, le tableau 3.3 illustre un exemple de document on doublons : Pour remédier a ça nous avons décidé de garder uniquement la cellule

348777	How about a bumper sticker thats says : ?	348777_2	I would buy one. It is what they think/say.	4
348777	How about a bumper sticker thats says : ?	348777_1	I would buy one. It is what they think/say.	2

TABLE 3.3 – Exemple de doublons existant dans le corpus d'Antique

ayant le score de pertinence le plus élever parmi tout les doublons.

2. **Suppression de caractère spéciaux** :Pour ce faire nous solliciterons la bibliothéque python re pour les deux corpus anglais et arabe, tel qu'indiqué dans le tableau 3.4 :

How about a bumper sticker thats says : ?	How about a bumper sticker thats says
;أخاف عطفك غلب حبك.لا يا /حبيبي ابغىا هـواك ؛	أخاف عطفك غلب حبك لا يا حبيبي ابغى هـواك

TABLE 3.4 – Exemple de suppression de charactere spéciale

3. **Segmentation (Tokenization) et suppression des mots vides** : Pour l'Anglais nous avons utilisé le Tokenizer natif de NLTK, et la FARASA Toolbox pour la langue Arabe, comme montré dans le tableau 3.5 :

How about a bumper sticker thats says	How bumper sticker says
أخاف عطفك غلب حبك لا يا حبيبي ابغى هـواك	أخاف عطفك غلب حبك يا حبيبي ابغى هـواك

TABLE 3.5 – résultat après segmentation et suppression de mot vide

4. **Racinisation (Stemming)** :La librairie Snowball Stemmer de NLTK a été utilisée pour l'Anglais ; quant à l'Arabe c'est toujours la Toolbox FARASA, le tableau 3.6 illustre un exemple de cela :

How bumper sticker says	how bumper sticker say
أخاف عطفك غلب حبك يا حبيبي ابغى هـواك	خاف عطف قلب حب يا حبيب ابغى هـوك

TABLE 3.6 – Résultat après l'application du Stemming sur les deux langues

3.5.1 Annotation des documents

Après avoir suivi toutes ces étapes, nous sommes presque prêts à utiliser notre ensemble de données et à entraîner notre modèle. Toutefois, il est important d'annoter et organiser les corpus de benchmark pour pouvoir entraîner et évaluer notre modèle. Chaque requête est associée a un ensemble de documents. Ces derniers représentent le résultat de recherche en utilisant Cette requête. L'ensemble de documents assignés à une requête donnée est réparti en deux classes des documents positifs et des documents négatifs. Un document positif représentent un document qui a été sélectionné par un utilisateur lors d'une recherche avec une requête donnée. Un document positif a un contenu qui correspond et répond à une requête d'un utilisateur. Cependant, un document négatif est un document qui a été proposé à utilisateur dans les résultats d'une recherche lancée avec une requête donnée, mais ce document négatif a été écarté par l'utilisateur car son contenu ne répond pas à ce qu'il cherche. Par conséquent, ce document a été classé négatif. Dans le cas du corpus *Antique*, nous considérons tout document ayant un score de 3 ou plus comme étant positif et le reste comme négatif, quant a *Mr-tydi* les documents positifs et négatifs sont déjà tous étiquetés à 1 et 0, respectivement. Cependant, il est important de noter que nous ne pouvons pas inclure tous les documents négatifs disponibles dans notre modèle. Cela est dû à la nécessité de maintenir un équilibre entre le nombre de documents positifs et négatifs pour chaque requête. Pour éviter un déséquilibre significatif entre les exemples positifs et négatifs et un éventuel problème de sur-apprentissage, nous avons choisi de prendre un nombre de documents négatifs égal au nombre de documents positifs pour chaque requête. Cette approche permet de garantir un rapport équilibré entre les exemples positifs et négatifs, ce qui est crucial pour effectuer un apprentissage correct et assurer les performances du modèle. Après l'étape de l'annotation, nous avons chargé toutes nos donnés dans un Dataloader et utilisé la bibliothèque python pytorch. Nous nous sommes retrouvés avec un total de 2,426 requêtes et 4,852 documents pour l'anglais et 12,364 requêtes 24,728 documents au total pour la langue arabe. La distribition des données annotées dans les deux corpus de tests en langue arabe et anglais est détaillée dans le tableau 3.7.

Total	**Corpus anglais**	**Corpus arabe**
Requête	2,426	12,364
Documents	4,852	24,728
Documents positifs	2,426	12,364
Documents négatifs	2,426	12,364

TABLE 3.7 – Distribution des données annotées dans les corpus de tests.

3.6 Représentation des données textuelles

Après avoir traiter et formater nos données il est maintenant nécessaire de représenter nos donnée textuelle numériquement et d'appliquer notre signature ou aussi appelé indexation, pour ce faire la classe *Countvectorizer* de la bibliothèque python *scikit-learn* a

était utilisé dans l'extraction des caractéristiques pour l'Anglais et l'Arabe. Ensuite, nous allons appliqué a chaque document et requête notre signature grâce a nos deux réseaux de neurones encodeur (document_encoder et query_encoder) que nous utiliserons comme embedding le but est de transformer un tenseur d'entrée de grande dimension en un tenseur de sortie de dimension inférieur comme illustré dans le tableau 3.8 :

How about a bumper sticker thats says : ?	[0.2332, 0.3126, 0.3665, 0.8011, 0.0000, 0.3408, 0.0000, 0.3374]
أخاف عطفك غلب حبك يا حبيبي ابغى هـواك. '	[0.0000, 0.0000, 0.3486, 0.0000, 0.0947, 0.0000, 0.0000, 0.1773]

Table 3.8 – Exemple d'indexation

3.7 Résultats

Comme citer dans la section 2.6 notre approche est basé sur le calcul de similarité entre la requête et les documents. Le tableau 3.9 présente un exemple de score de similarité, extrait des données en langues anglaise, entre différents documents et la même requête.

Document	Requête	Similarité
1+1 is 2 in maths. 1+1 is 1 in love	what is 1 plus 1 ?	0.6907
they don't know any better	what is 1 plus 1 ?	0.4938
unless you want prison time you dont.	what is 1 plus 1 ?	0.4214
if you get caught its federal.	what is 1 plus 1 ?	0.2375
they don't know any better	what is 1 plus 1 ?	0.1968

Table 3.9 – Exemple de calcul de similarité

Les tableaux 3.10 et 3.11 présentent les matrices de confusions obtenus par les modèles entraînés sur les données en langue anglaise et en langue arabe, respectivement. Nous remarquons que le taux des faux positifs et des faux négatifs sont beaucoup plus élevés dans le modèle entraîné sur les données en langue arabe que celui entraîné sur les données anglaises. La complexité de langue arabe nécessite encore un travail de recherche plus approfondi afin d'améliorer les résultats. Par ailleurs, le nombre d'exemples utilisés pour le test en langue arabe était beaucoup plus important que celui en langue anglaise : un total de 2,426 requêtes et 4,852 documents pour l'anglais et 12,364 requêtes 24,728 documents au total pour la langue arabe. Par conséquent, le taux d'erreur sur les données en langue arabe est beaucoup plus élevé que celui sur les données en langue anglaise.

Les tableaux 3.12 et 3.13 représentent les résultats obtenus sur les données en langue arabe en utilisant plusieurs métriques d'évaluation.

	Positive (Actuel)	**Négative (Actuel)**
Positive (Prédit)	$TP \approx 1351$	$FP \approx 167$
Négative (Prédit)	$FN \approx 1077$	$TN \approx 2256$

TABLE 3.10 – Matrice de confusion du modèle entraîné sur les données en langue anglaise.

	Positive (Actuel)	**Négative (Actuel)**
Positive (Prédit)	$TP \approx 6,578$	$FP \approx 2,193$
Négative (Prédit)	$FN \approx 5,786$	$TN \approx 10,171$

TABLE 3.11 – Matrice de confusion du modèle entraîné sur les données en langue arabe.

Métrique	**Valeur**
Fonction de perte triplet	0.125
Precision	0.893
Recall	0.556
F1-score	0.690

TABLE 3.12 – Résultat obtenus sur les données en langue anglaise

Métrique	**Valeur**
Fonction de perte triplet	0.230
Precision	0.751
Recall	0.532
F1-score	0.551

TABLE 3.13 – Résultat obtenus sur les données en langue arabe

Les résultats d'évaluation de performance obtenus sur nos corpus de test sont satisfaisants en comparaison aux performances des travaux existants testés sur les différents benchmarks sélectionnés comme illustré sur les tableaux 3.7 et 3.8. Ces performances moyennes montrent que la recherche et l'indexation d'information restent des tâchent complexes à effectuer malgré les longues années de recherche. Une fois notre plateforme hébergée, nous collecterons des données réelles sur lequel notre modèle pourra être raffiné.

3.8 Plateforme Kateb et intégration du module de recherche intelligent

Après avoir finalisé la création de notre API, nous procédons à son intégration complète dans notre plateforme. Lors de cette intégration, nous prenons en compte les fonctionnalités de recherche déjà présentes dans la barre de recherche de wiki.js. Il est important de noter

Method	MAP	MRR	P@1	P@3	P@10	nDCG@1	nDCG@3	nDCG@10
BM25	0.1977	0.4885	0.3333	0.2929	0.2485	0.4411	0.4237	0.4334
DRMM-TKS [6]	0.2315	0.5774	0.4337	0.3827	0.3005	0.4949	0.4626	0.4531
aNMM [17]	0.2563	0.6250	0.4847	0.4388	0.3306	0.5289	0.5127	0.4904
BERT [4]	**0.3771**	**0.7968**	**0.7092**	**0.6071**	**0.4791**	**0.7126**	**0.6570**	**0.6423**

FIGURE 3.7 – Le benchmark résulte d'une grande variété de modèles d'extraction sur l'ensemble de données ANTIQUE[23]

	Ar
BM25 (default)	0.368
BM25 (tuned)	0.367
mDPR	0.260
hybrid	0.491†

FIGURE 3.8 – Le benchmark résulte du rappelle@100 sur d'une grande variété de modèles d'extraction sur l'ensemble de données MR-Tydi/ar[54]

que la plateforme wiki.js propose déjà cinq types d'algorithmes de recherche dans cette barre de recherche.

Pour commencer, nous créons notre propre fichier de recherche dédié à notre API. Ce fichier nous permet de personnaliser les résultats de recherche en utilisant notre propre modèle et algorithme. Dans ce fichier, nous ajoutons également un script.js qui agit comme un intermédiaire entre la barre de recherche de l'utilisateur et notre API. Ainsi, lorsque l'utilisateur saisit un texte de recherche dans la barre, le script.js envoie cette requête à notre API.

L'API traite alors la requête et retourne les identifiants des documents pertinents à afficher en réponse à la recherche de l'utilisateur. Ces identifiants sont ensuite utilisés pour récupérer les documents correspondants dans notre base de données. Ensuite, nous pouvons présenter à l'utilisateur les résultats pertinents de sa requête.

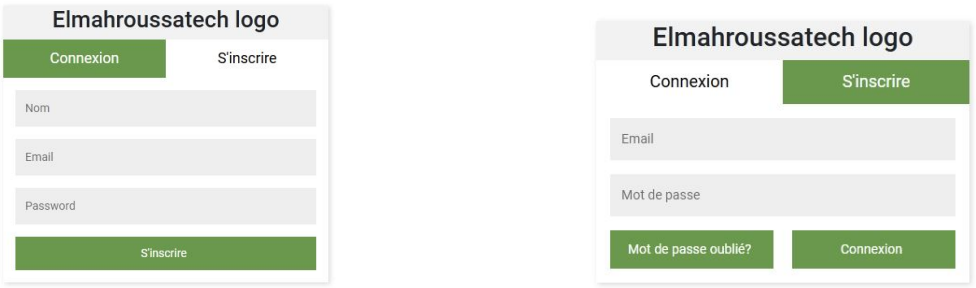
En fin de compte, grâce à cette intégration, nous sommes en mesure de fournir à l'utilisateur une réponse précise et pertinente à sa requête de recherche. Notre API joue un rôle clé dans ce processus, en personnalisant les résultats et en améliorant l'expérience globale de recherche au sein de notre plateforme wiki.js.

3.9 Plateforme globale El Mahrousha Tech

3.9.1 Présentation de la plateforme

Dans le cadre de la mise en place de notre plateforme principale, une procédure d'authentification sera nécessaire à l'arrivée de l'utilisateur l'interface présenter dans la figure

3.9 lui sera afficher afin de lui permettre d'accéder à la sélection des plateformes disponibles. L'utilisateur devra fournir des informations minimales, à savoir un nom, une adresse e-mail et un mot de passe, pour pouvoir s'inscrire. Lors de la première inscription, une confirmation par courrier électronique sera requise pour activer le compte. Les données liées à l'authentification seront sauvegardées et utilisées uniquement pendant la durée de la session.



(a) inscription (b) connexion

FIGURE 3.9 – Interface d'inscription et de connexion de "El Mahroussa Tech"

Une fois que l'utilisateur a été authentifié avec succès, l'interface de sélection de plateforme illustré dans la figure 3.10 lui est présentée. À ce stade, il lui est demandé de choisir la catégorie d'archives qu'il souhaite consulter. Il convient de rappeler que, dans le cadre de ce projet, seule la composante textuelle des archives est prise en charge.

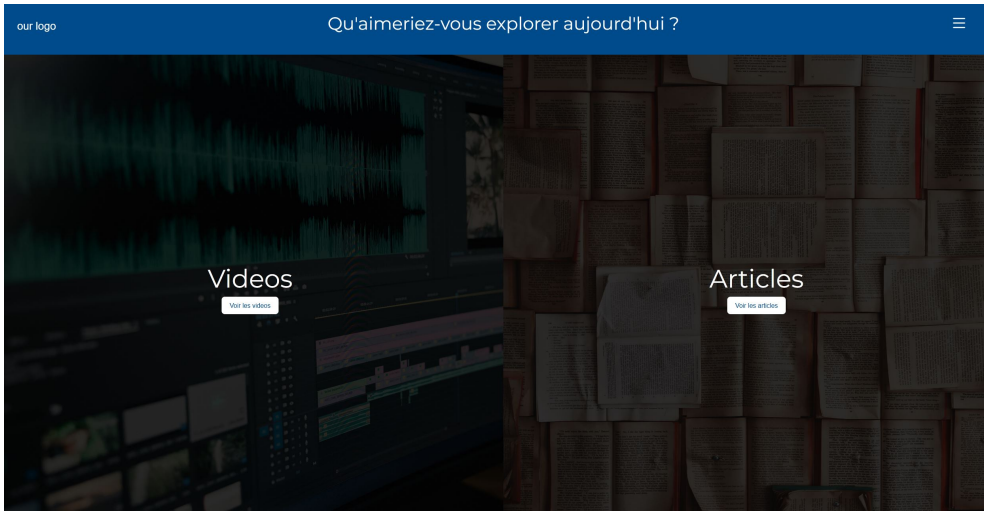
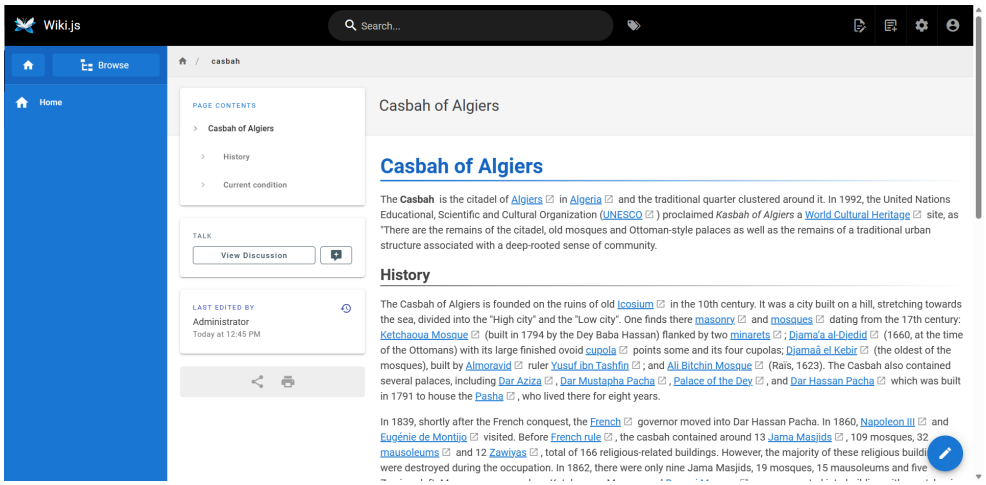
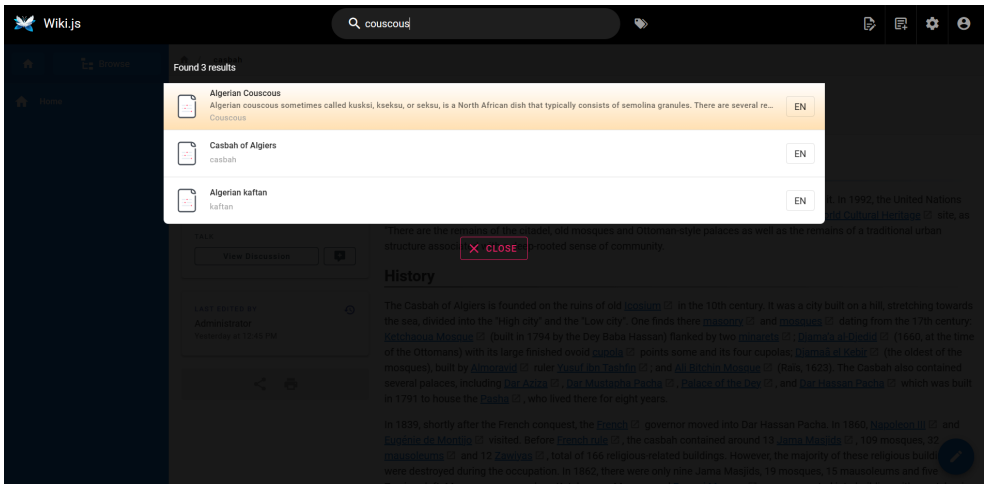


FIGURE 3.10 – Interface de choix de plateforme

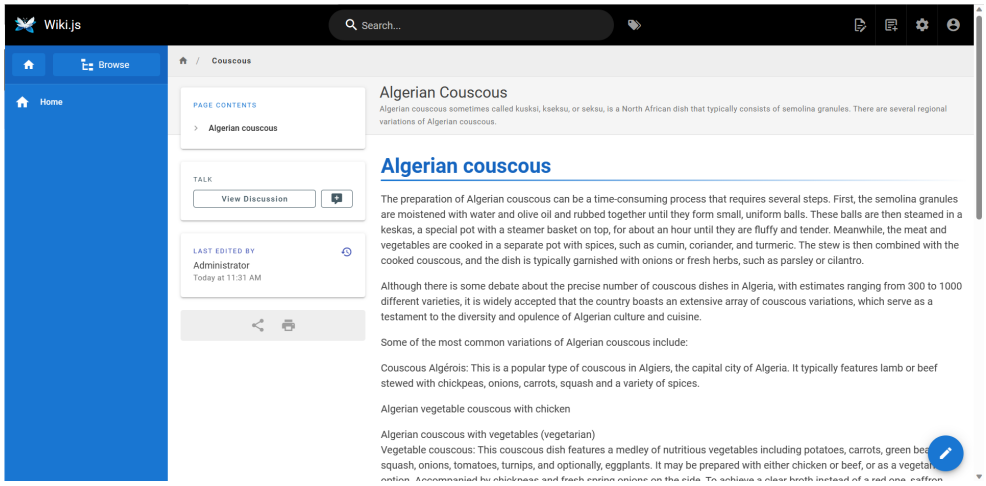
Après avoir effectué son choix, l'utilisateur sera redirigé vers la plateforme choisi (celle axée sur les document textuelle). Les données d'authentification de l'utilisateur seront également transférées via un partage de session et il se retrouvera sur la page principale de la plateforme, où il pourra effectuer des recherches selon ses besoins conformément à l'exemple visuel fourni dans la figure 1.



(a) page Principale



(b) Barre de recherche



(c) Document recherché

FIGURE 3.11 – Interface d'archive de document textuelle

3.9.2 Hébergement

Afin de garantir l'accessibilité de notre site à tout utilisateurs, nous avons pris la décision de l'héberger sur un serveur VPS avec un sous-domaine dédié à notre plateforme de texte. Pour le serveur web, nous avons choisi de configurer un serveur NGINX en raison de sa flexibilité et de sa capacité à fournir une protection efficace contre les attaques de type DDOS. Le serveur NGINX joue également un rôle essentiel en facilitant la communication entre les différents ports de chacune de nos sous-plateformes.

3.10 Conclusion

Dans ce chapitre, nous avons d'abord présenté l'environnement de travail ainsi que les divers outils et langage utilisé lors de la réalisation du projet, l'implémentassions de notre module d'indexation et de recherche les résultat obtenus et les évaluations en utilisant les métrique adéquates. Nous avons par la suite traité de son intégration a notre sous-plateforme "Kateb" puis nous avons présenté la plateforme globale El Mahroussa Tech ainsi que quelque détails sur son hébergement.

Conclusion Générale

De nos jours, l'accès à l'information est devenu une nécessité majeure dans tous les domaines de la société. Les sites de blogs et les plateformes de partage de médias ont pleinement conscience de cette demande et en ont fait leur principal modèle économique. Dans un monde où l'histoire d'un peuple définit son identité, il est regrettable que notre pays ne dispose d'aucun moyen archivé permettant d'accéder ses informations. Ces derniers représente un point de départ important pour les chercheurs et historiens qui se spécialise sur la culture de notre pays.

La principal motivation de se travail a été de proposer un outil d'archivage de document textuelle patrimonial qui permet au lecteur de s'instruire de faire le recherche et se documenter sur les origines de notre nation mais aussi de permettre au chercheurs et historien de diffuser le fruit de leurs travail sans pour autant perdre leurs droit ou mérite vis a vis de celui-ci.Nous réalisons à présent qu'entreprendre ce projet s'étend bien au-delà du cadre d'un projet de fin d'études et s'inscrit dans une optique d'innovation, dont l'objectif est d'apporter une réelle contribution et apporter une solution a un problème sociale existant dans la société actuelle.

Dans le présent mémoire nous avons au premiers lieu étudié l'importance du patrimoine culturelle et présenter les dangers qui rode autour de ces derniers. Nous avons ensuite présenté en détail le domaine de la recherche d'information et de l'indexation, Et ce afin d'enlever toute ambiguïté sur leurs usage pour bien cerner les possibilités d'exploitation de ses techniques.

Puis nous avons effectué une profonde étude de l'état de l'art existant sur les divers méthode de recherche d'information employer avant d'identifier les plateformes open-source les plus fiables en matière de partage de contenu textuelle.

Une longue période du projet a été dédiée à la recherche, la récolte et le pré-traitement des données. Une fois le dataset de chaque langue de l'application créés,une étude exploratoire a été réalisée sur ces ensembles afin d'identifier les caractéristiques homogènes des données disponibles et de mettre en évidence les indicateurs distinctifs (documents positif et négatif) des exemples dans les corpus.

Néanmoins, un des principaux obstacle de ce projet a été le manque flagrant de corpus et d'outils pour le traitement d'un modèle de recherche d'information surtout pour la langue arabe.

Nous avons également pris le soin de fournir des détails exhaustifs sur la solution que nous proposons. Nous avons minutieusement détaillé les différentes démarches que nous

avons suivies et nous avons clairement défini l'approche que nous avons adopté, en justifiant chaque choix entrepris.

Les résultats du calcule de similarité on était assez satisfaisant comparé au divers benchmark employé pour évaluer les corpus de donnée que nous avons employé. Nous avons également de très bon résultats sur le terrain après le déploiement du module sur la plateforme dans les deux langues.

La dernière marche de notre projet a étais la réalisation de la plateforme "Elmahroussatech". "Elmahroussatech" est une plateforme en ligne qui intègre notre modèle de recherche et d'indexation et qui permet a l'utilisateur d'exploiter pleinement toutes les fonctionnalités présenter précédemment.L'expérience utilisateur et le design des interfaces ont été conçus de telle sorte à compléter les fonctionnalités par l'intuitivité et le plaisir de l'utilisation.

La dernière partie du mémoire présente les caractéristiques et les interfaces de l'application, ainsi que les différentes étapes de développement, en fournissant des justifications pour le choix des outils, des bibliothèques et des langages de programmation utilisés.

Nous pouvons aujourd'hui affirmer que notre objectif a été atteint, puisque l'application que nous avons réalisé satisfait largement les principales ambitions définies.

Ce projet nous as étais bénéfique sur multiple plan. Il nous as permit de mettre en pratique plusieurs connaissance majeur acquise tout le long de notre cursus de Master en ingénierie du logicielle, de développé une toute nouvelle méthodologie de recherche et de différenciation des travaux, de découvrir de toute nouvelle notion dans le milieu de l'intelligence artificiel notamment dans le domaine TAL et de la recherche d'information. Grâce à ce mémoire, nous avons pu développé nos compétences en rédaction et devenir plus rigoureux en matière de plagiat, de citation des sources d'information et de référencement. La collaboration en binôme, sous la supervision de notre promoteur, a grandement renforcé notre capacité à travailler en équipe.

Nonobstant, il est nécessaire de souligner que la solution proposé est loin d'être parfaite et possède de grande fenêtre d'amélioration. certains point demande encore a être revus et retravailler, L'une des perspective qui s'avérait être primordiale serais d'ajouter la langue française comme langue supporter par notre modèle.

quand ta "Elmahroussatech" elle pourrait être enrichie également par d'autre fonctionnalités tell que l'ajout d'un système de traduction automatique de document affin de permettre a tout le monde d'accéder a tout information peut importe sa langue de confort. Mais aussi par le fait d'ajouter des images affin de permettre au auteurs de mieux illustré leurs recherche ou information. Et bien sûr d'un système de recommandation automatique qui sera recommander de nouveaux documents aux utilisateurs selon leurs préférence a leurs arrivé.

References

[1] *Information retrieval system explained : Types, comparison amp ; components*, Nov 2022.

[2] *Wikibooks.org*. Website, Septembre 2022. Mai 2023.

[3] *laravel.com*. Website, Fevrier 2023. Mai 2023.

[4] *pappleweb.org*. Website, 2023. Mai 2023.

[5] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, *Farasa : A fast and furious segmenter for arabic*, in 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Association for Computational Linguistics, 2016, pp. 11–16.

[6] L. A. Adamic, J. Z. 0008, E. Bakshy, and M. S. Ackerman, *Knowledge sharing and yahoo answers : everyone knows something.*, in WWW, J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, eds., ACM, 2008, pp. 665–674.

[7] S. B, *Mois du patrimoine en algérie : programme d'activités varié dans les musées*, 2023. Accessed : 27/04/2023.

[8] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval addison*, 1999.

[9] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python : analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.", 2009.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*, Journal of machine Learning research, 3 (2003), pp. 993–1022.

[11] D. Charlet and G. Damnati, *Simbow : une mesure de similarité sémantique entre textes (simbow : a semantic similarity metric between texts)*, in Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 - Articles courts, Orléans, France, 6 2017, ATALA, pp. 126–133.

[12] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki, *TyDi QA : A benchmark for information-seeking question answering in typologically diverse languages*, Transactions of the Association for Computational Linguistics, (2020).

[13] C. Darwin and L. Kebler, *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life*, J. Murray, London, 1859. PDF.

[14] M. Dorigo, M. Birattari, and T. Stutzle, *Ant colony optimization*, IEEE computational intelligence magazine, 1 (2006), pp. 28–39.

[15] K. L. Elmore and M. B. Richman, *Euclidean distance as a similarity metric for principal component analysis*, Monthly weather review, 129 (2001), pp. 540–549.

[16] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., *A density-based algorithm for discovering clusters in large spatial databases with noise.*, in kdd, vol. 96, 1996, pp. 226–231.

[17] N. Faessel, *Indexation et interrogation de pages Web décomposées en blocs visuels*, PhD thesis, Aix-Marseille 3, 2011.

[18] C. N. Fateh, *Architecture logicielle, chapitre 4 : Styles d'architecture*, 2022. Support de cours reçu par mail.

[19] I. Fister Jr, X.-S. Yang, I. Fister, J. Brest, and D. Fister, *A brief review of nature-inspired algorithms for optimization*, arXiv preprint arXiv :1307.4186, (2013).

[20] F. Fkih, *Modèles d'Indexation et Algorithmes de Recherche d'Information à partir de Documents non Structurés*, PhD thesis, 11 2016.

[21] K. Garrouch, *Modèles de Recherche d'information basés sur les Réseaux Bayésiens et les Réseaux Possibilistes*, PhD thesis, 02 2017.

[22] I. Grattan-Guinness and G. Bornet, *George Boole : Selected Manuscripts on Logic and its Philosophy*, vol. 20, Springer Science & Business Media, 1997.

[23] H. Hashemi, M. Aliannejadi, H. Zamani, and W. B. Croft, *Antique : A non-factoid question answering benchmark*, 2019.

[24] J. H. Holland, *Adaptation in Natural and Artificial Systems : An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT press, 1992.

[25] Inria, *Scikit-learn : Donner de l'intelligence à nos systèmes.* Inria Centre Saclay - Île-de-France, May 2018. Consulté le 06-06-2023.

[26] A. Karpathy and L. Fei-Fei, *Deep visual-semantic alignments for generating image descriptions*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.

[27] J. Kennedy and R. Eberhart, *Particle swarm optimization*, in Proceedings of ICNN'95-international conference on neural networks, vol. 4, IEEE, 1995, pp. 1942–1948.

[28] T. Kohonen, *Self-organizing maps*, vol. 30, Springer Science & Business Media, 2012.

[29] A. H. Lashkari, F. Mahdavi, and V. Ghomi, *A boolean model in information retrieval for search engines*, in 2009 International Conference on Information Management and Engineering, IEEE, 2009, pp. 385–389.

[30] M. Lathkar, *Getting started with fastapi*, in High-Performance Web Apps with FastAPI : The Asynchronous Web Framework Based on Modern Python, Springer, 2023, pp. 29–64.

[31] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, nature, 521 (2015), pp. 436–444.

[32] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N. Goharian, *Simplified data wrangling with $ir_datasets$, in$SIGIR$*, 2021.

[33] C. D. Manning, P. Raghavan, and H. Schütze, *Xml retrieval*, Introduction to Information Retrieval, (2008).

[34] K. Memarian, J. Matthiesen, J. Lingard, K. Nienhuis, D. Chisnall, R. N. Watson, and P. Sewell, *Into the depths of c : elaborating the de facto standards*, ACM SIGPLAN Notices, 51 (2016), pp. 1–15.

[35] J. M. Merigó and M. Casanovas, *A new minkowski distance based on induced aggregation operators*, International Journal of Computational Intelligence Systems, 4 (2011), pp. 123–133.

[36] S. Mirjalili and S. Mirjalili, *Genetic algorithm*, Evolutionary Algorithms and Neural Networks : Theory and Applications, (2019), pp. 43–55.

[37] l. s. e. l. c. Organisation des Nations Unies pour l'éducation, *Textes fondamentaux de la convention de 2003 pour la sauvegarde du patrimoine culturel immatériel*. Accessed : 24/04/2023.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *Pytorch : An imperative style, high-performance deep learning library*, in Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[39] G. Quellec, *Indexation et fusion multimodale pour la recherche d'information par le contenu. Application aux bases de données d'images médicales.*, PhD thesis, Télécom Bretagne, 2008.

[40] F. Rahutomo, T. Kitasuka, and M. Aritsugi, *Semantic cosine similarity*, in The 7th international student conference on advanced science and technology ICAST, vol. 4, 2012, p. 1.

[41] G. Salton, A. Wong, and C.-S. Yang, *A vector space model for automatic indexing*, Communications of the ACM, 18 (1975), pp. 613–620.

[42] J. Savoy and E. Gaussier, *Information retrieval*, (2010).

[43] F. Schroff, D. Kalenichenko, and J. Philbin, *Facenet : A unified embedding for face recognition and clustering*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[44] Y. Sharma and S. Gupta, *Deep learning approaches for question answering system*, Procedia computer science, 132 (2018), pp. 785–794.

[45] R. Socher, E. Huang, J. Pennin, C. D. Manning, and A. Ng, *Dynamic pooling and unfolding recursive autoencoders for paraphrase detection*, Advances in neural information processing systems, 24 (2011).

[46] M. Stonebraker and G. Kemnitz, *The postgres next generation database management system*, Communications of the ACM, 34 (1991), pp. 78–92.

[47] P. Thompson, *Looking back : On relevance, probabilistic indexing and information retrieval*, Information processing & management, 44 (2008), pp. 963–970.

[48] G. Van Rossum, *The Python Library Reference, release 3.8.2*, Python Software Foundation, 2020.

[49] D. Varrazzo, *Psycopg 2.9.6 documentation.*

[50] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, *Show and tell : A neural image caption generator*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[51] E. M. Voorhees, D. K. Harman, et al., *TREC : Experiment and evaluation in information retrieval*, vol. 63, Citeseer, 2005.

[52] D. Wolfram, A. Spink, B. J. Jansen, T. Saracevic, et al., *Vox populi : The public searching of the web*, JASIST, 52 (2001), pp. 1073–1074.

[53] C. Zeng, L. Luo, Q. Ning, Y. Han, Y. Jiang, D. Tang, Z. Wang, K. Chen, and C. Guo, {*FAERY*} *: An* {*FPGA-accelerated*} *embedding-based retrieval system*, in 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22), 2022, pp. 841–856.

[54] X. Zhang, X. Ma, P. Shi, and J. Lin, *Mr. TyDi : A multi-lingual benchmark for dense retrieval*, arXiv :2108.08787, (2021).