République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Yahia Farès de Médéa
Faculté de Sciences

Département Mathématique et Informatique (MI)



Mémoire de fin d'études de Master

Filière : Informatique

Spécialité : Ingénierie des systèmes et technologie du web

Thème

Réalisation d'un moteur de recherche en arabe spécialisé en tourisme Algérien

---

Proposé et Dirigé par

Dr.HAYI Mohamed Yassine

Présenté par

ACHOUR Loubna

ASLI Noudjoud

Année Universitaire : 2023 – 2024

Résumé

L'arabe est l'une des langues les plus importantes au monde, avec près de 400 millions de locuteurs, la langue officielle de plus de 20 pays. Cependant, il n'a pas eu suffisamment de possibilités de recherche et de développement dans la technologie et les logiciels, car il a besoin d'algorithmes avancés et de recherche intensive du langage pour assurer une intégration efficace dans les applications technologiques modernes.

Sur cette base, Nous avons donc décidé de développer un moteur de recherche en langue arabe dédié au tourisme en Algérie, assurant une information complète et fiable.

Nous avons commencé par créer une base de données contenant des informations complètes sur les destinations touristiques populaires du pays.

Ensuite, une série de requêtes a été extraite, représentant des questions courantes que les touristes peuvent poser lors de la planification de leurs voyages.Dans la phase suivante, nous nous sommes concentrés sur le calcul de la pertinence entre les documents et les requêtes en utilisant trois moteurs de recherche open source, plus une validation par un expert (être humain).

Enfin, tous les résultats obtenus par les moteurs de recherche susmentionnés ont été compilés et liés, afin de garantir des résultats complets et fiables qui répondent aux besoins des utilisateurs et assurent une expérience de recherche efficace et fiable.

Mots-clés : Moteurs de recherche, Langue arabe, Recherche d'information, Tourisme, Apache Lucene, Apache Solr, Elasticsearch, Pertinence, Requête.

ملخص

تعد اللغة العربية واحدة من أهم اللغات في العالم، حيث يصل عدد الناطقين بها إلى ما يقارب ٤٠٠ مليون نسمة، وهي اللغة الرسمية لأكثر من ٢٠ دولة. ومع ذلك، لم تتح لها الفرصة الكافية للبحث والتطوير في مجال التكنولوجيا والبرمجيات، حيث انها بحاجة إلى خوارزميات متقدمة وأبحاث لغوية مكثفة لضمان التكامل الفعّال في التطبيقات التكنولوجية الحديثة.

لذا، استنادًا إلى هذا الاعتبار، قررنا التحرك نحو تطوير محرك بحث باللغة العربية مخصص للسياحة في الجزائر، يضمن توفير معلومات شاملة وموثوقة.

بدأنا بإنشاء قاعدة بيانات تحتوي على معلومات شاملة حول الوجهات السياحية الرائجة في البلاد، ومن ثم، تم استخراج مجموعة من الاستعلامات التي تمثل أسئلة شائعة يمكن أن يسألها السائحون أثناء تخطيط رحلاتهم.

في المرحلة التالية، ركزنا على تحليل العلاقة بين المستندات والاستعلامات باستخدام ثلاثة محركات بحث مفتوحة المصدر، مع التحقق من صحة العلاقة من خلال خبير (كائن بشري) . وأخيرًا، تم تجميع وربط جميع النتائج المتحصل عليها من طرف محركات البحث السابقة الذكر، لضمان الحصول على نتائج شاملة وموثوقة تلبي احتياجات المستخدمين وتضمن تجربة بحث فعالة وموثوقة .

الكلمات المفتاحية : محركات البحث، اللغة العربية، استرجاع المعلومات، السياحة، أباتشي لوسين، أباتشي سولر، إلاستيك سيرش، الصلة، الاستعلامات.

Abstract

Arabic is one of the most important languages in the world, with nearly 400 million speakers and the official language of more than 20 countries. However, it has not had sufficient opportunities for research and development in technology and software, as it requires advanced algorithms and intensive language research to ensure effective integration into modern technological applications.

Based on this, we decided to focus on developing an Arabic language search engine dedicated to tourism in Algeria, ensuring comprehensive and reliable information.

We started by creating a database containing complete information on popular tourist destinations in the country. Then, a series of queries were extracted, representing common questions tourists might ask when planning their trips.

In the next phase, we focused on calculating the relevance between documents and queries using three open-source search engines, plus validation by an expert (human).

Finally, all the results obtained by the aforementioned search engines were compiled and linked to ensure comprehensive and reliable results that meet user needs and provide an effective and reliable search experience.

Keywords : search engines, Arabic language, Information Retrieval, Tourism, Apache Lucene, Apache Solr, Elasticsearch, Relevance, Query.

Remerciements

En premier lieu, nous remercions Allah, dont l'aide, le courage et la patience nous ont permis de mener à bien ce travail.

Nos sincères remerciements au directeur de ce mémoire, **Dr. Hayi Mohamed Yacine**, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à enrichir notre réflexion.

Nous exprimons également notre gratitude envers les membres du jury pour avoir accepté de juger ce travail.

Nous tenons à adresser nos profonds remerciements au corps professoral et administratif de l'université de Medea pour tous les efforts déployés.

Nous joignons nos remerciements à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire.

Dédicace

D'abord et avant tout, je remercie ***Dieu*** *de m'avoir accordé son aide tout au long de mon parcours scolaire et de m'avoir donné la patience et la force pour être ce que je suis aujourd'hui.*

Je dédie ce mémoire à :

Mes chers parents, pour leur amour inconditionnel, leurs innombrables sacrifices, leur patience infinie et leur encouragement constant tout au long de mes études.Votre foi en moi m'a donné la force de persévérer même dans les moments les plus difficiles. Vous avez toujours été là pour me soutenir, et c'est en grande partie grâce à vous que ce mémoire a pu voir le jour.

Mon frère Abdou et ma sœur Marwa, pour leur amour inconditionnel, leur soutien constant et leur inspiration quotidienne.

Mes chères grands-mères et toute ma famille, pour leur amour, leur sagesse et leur soutien indéfectible.

À ma cousine Sarah, pour son soutien inébranlable et son inspiration constante.

À mon binôme Noudjoud, pour sa précieuse collaboration, son travail assidu et son soutien mutuel tout au long de ce projet.

Mes meilleures amies Amina et Nesrine, pour leur amitié précieuse et leur soutien inébranlable.

Achour Loubna

Dédicace

En premier lieu, je remercie ***Dieu*** *Tout-Puissant qui m'a fait capable de boire une dose de cette science, d'avoir ce plaisir que j'atteins ce niveau de ma carrière éducative, et m'avoir donné la volonté et la patience à finir ce travail.*

À mon cher père

Chaque mot semble bien trop modeste pour exprimer la profondeur de mon amour et de ma gratitude envers toi. Tu es la lumière qui illumine mon chemin, et la lampe qui ne s'éteint jamais. Tu es ma force et mon refuge après Dieu, et qui a fait l'effort des années pour gravir les échelles du succès.

À ma chère mère

Quoi que je fasse ou que je dise, je ne saurai point te remercier comme il se doit. Ton affection me couvre, ta bienveillance me guide et ta présence à mes côtés a toujours été ma source de force pour affronter les différents obstacles.

À ma chère sœur

Qui m'a encouragé en tout temps, et ses filles Rahef , Roua et Mariam , qui sont les source de joie et de bonheur.

À mes chers frères

Seddik et Ayoub ,qui m'ont soutenu dans les difficulté.

À mon binôme ***Loubna*** *pour son soutien moral, sa patience et sa compréhension tout au long de ce projet.*

À tous mes amies.

ASLI Noudjoud

Table des matières

Introduction générale **1**

Problématique et Objectifs **3**

Organisation du mémoire **4**

I Etat de l'art **5**

1 Introduction 6
2 Recherche d'information 6
2.1 Définitions 6
2.2 modèles de la recherche d'information 7
2.2.1 Le modèle booléen 7
2.2.2 Le modèle vectoriel 7
2.2.3 Le modèle probabiliste 8
2.3 Concepts de base de la recherche d'information 8
2.3.1 Collection de documents 8
2.3.2 Document 8
2.3.3 Requête 9
2.3.4 Besoin d'information 9
2.3.5 Pertinence 9
2.4 Traitement du langage naturel(NLP) 9
2.4.1 Définition 9
2.4.2 Applications du NLP 10
3 Systèmes de recherche d'information : 11
3.1 Définition 11
3.2 Processus de recherche d'information 11
3.3 Mesures d'évaluation des SRI 12
3.3.1 La précision (P) 12
3.3.2 Le rappel(R) 12
3.4 F-mesure 12
3.4.1 Le bruit(B) 12
3.4.2 Le silence(S) 12
4 Moteur de recherche : 13
4.1 Definition : 13

4.2 Composants d'un moteurs de recherche : . . . 13
4.2.1 Crawler (ou Spider) : . . . 13
4.2.2 Indexeur : . . . 13
4.2.3 L'interrogateur : . . . 13
4.3 Requête : . . . 14
4.4 Fonctionnalité d'un moteur de recherche : . . . 14
4.5 Type existants : . . . 15
4.5.1 moteurs de recherche en ligne « online » : . . . 15
4.5.2 moteurs de recherche hors ligne « offline » : . . . 15
5 langue arabe : . . . 16
5.1 L'historique de la langue arabe : . . . 17
5.2 La richesse de la langue arabe : . . . 17
5.3 Variétés de la langue arabe : . . . 18
5.3.1 L'arabe classique : . . . 18
5.3.2 Arabe standard moderne ou contemporain (ASM) : . . . 18
5.3.3 Arabe dialectal : . . . 18
6 Le tourisme . . . 19
6.1 Définition . . . 19
6.2 Rôles du tourisme . . . 19
6.2.1 rôle économique . . . 19
6.2.2 rôle social . . . 20
6.2.3 rôle culturel . . . 20
6.3 Le tourisme en Algérie . . . 20
7 Conclusion . . . 21

II Travaux Connexes **22**
1 Introduction . . . 23
2 Fondements et Évolutions en Recherche d'Information . . . 23
2.1 Trec . . . 23
2.1.1 Définition du Trec . . . 23
2.1.2 Historique de Trec . . . 24
2.1.3 Objectifs de Trec . . . 25
2.2 Clef . . . 25
2.2.1 Définition du Clef . . . 25
2.2.2 Objectifs de Clef . . . 27
2.3 Travaux connexe . . . 27
2.3.1 An intelligent use of stemmer and morphology analysis for Arabic information retrieval . . . 27
2.3.2 A new semantic similarity approach for improving the results of an Arabic search engine . . . 28
2.3.3 Formal Concept Analysis for Arabic Web Search Results Clustering . . . 29
2.3.4 REDUCE : a semi-supervised scalable approach for REsult-DUplication detection in Search Engines . . . 30
2.3.5 IBRI-CASONTO : Ontology-based semantic search engine . 31

2.3.6 Embedding arabic questions by feature-level fusion of wordrepresentations for questions classification : It is worth doing ? 33
3 Meilleures moteurs de recherche au monde . . . 35
3.1 google : . . . 36
3.2 Microsoft Bing : . . . 37
3.3 YANDEX : . . . 37
3.4 Yahoo ! : . . . 38
4 conclusion . . . 39

III L'approche proposée **40**
1 Introduction . . . 41
2 Description de l'approche . . . 41
2.1 Plan global du travail . . . 42
2.2 Plan détaillé du travail . . . 43
3 Construction du dataset et établissement de la liste des requêtes . . . 44
3.1 Collecte d'informations . . . 44
3.2 Construction de la base . . . 44
3.2.1 Base de données normale . . . 44
3.2.2 Différents types de la base de données . . . 46
3.2.3 Base de données améliorée . . . 47
3.3 Construction des requêtes . . . 48
3.3.1 Requêtes normales . . . 48
3.3.2 Requêtes améliorées : Optimisation des Recherches Touristiques 48
3.4 Extraction manuelle des documents pertinents : . . . 49
4 Récupération de la pertinence à travers des moteurs de recherche «Open Source » . . . 49
4.1 Validation par les trois moteurs de recherche . . . 50
4.2 Regroupement des Résultats des Trois Moteurs de Recherche . . . 51
5 Conclusion . . . 53

IV Implémentation Et Evaluation **54**
1 Introduction . . . 55
2 Différentes ressources utilisées . . . 55
2.1 Python . . . 55
2.2 Bibliothèques de Python utilisés . . . 55
2.3 Eclipse . . . 56
2.4 Wampserver . . . 56
3 Moteurs de recherche utilisée . . . 57
3.1 Lucene . . . 57
3.2 Elasticsearch . . . 58
3.3 Solr . . . 59
4 Résultats de la recherche . . . 60
4.1 Résultat de pertinence par les moteurs de recherche . . . 60
4.2 Comparaison des résultats des 04 parties . . . 61
4.3 Comparaison entre les résultats des moteurs de recherche . . . 62

4.4 Resultats de regroupements des trois moteurs 64
5 Présentation du site web 67
6 Conclusion 70

Conclusion générale **71**

Annexe1 **74**

Annexe2 **77**

Bibliographie **82**

Table des figures

1.1 Traitement du langage naturel. 10
1.2 Processus en U de la recherche d'information. 11
1.3 Fonctionnement d'un moteur de recherche. 14
1.4 Les Moteur de recherche online. 15
1.5 Moteur de recherche offline "Google Desktop". 16
1.6 Le monde arabe. 16
1.7 L'alphabet arabe. 17
1.8 Tourisme en Algérie 20

2.1 Nombre de participants par piste et nombre total de participants distincts dans chaque TREC. 24
2.2 Groupes participants du CLEF entre 2000 ET 2009. 26
2.3 Quelques informations sur les différentes collections du CLEF. 26
2.4 Les valeurs TF.IDF calculées obtenues par AMIR, LUCENE et FARASA. . . 28
2.5 Le processus de la méthodologie pour calculer la mesure de similarité sémantique. 29
2.6 Résultats de la requête (Éducation, التعليم) du cluster. 30
2.7 Architecture du Système de "REDUCE". 31
2.8 Structure de "IBRI-CASONTO". 32
2.9 Diagramme de flux d'un système de réponse aux questions 33
2.10 Part de marché des moteurs de recherche dans le monde en janvier 2024. . . 36
2.11 Moteur de recherche en ligne google. 36
2.12 Moteur de recherche en ligne Bing. 37
2.13 Moteur de recherche en ligne YANDEX. 38
2.14 Moteur de recherche en ligne Yahoo !. 38

3.1 Plan global de l'approche proposée. 42
3.2 Plan détaillé détaillant les diverses étapes de notre approche. 43
3.3 Pourcentage des documents par apport au champ Type. 45
3.4 Pourcentage des documents par apport au champ Zone. 46
3.5 Code d'amélioration de la BDD. 47
3.6 Code requêtes améliorées. 49
3.7 Code d'extraction du résultat final. 52

4.1 Interface d'un programme Java sous Eclipse. 56
4.2 Page d'accueil de WampServer. 57

4.3 Création d'une application avec Lucene. 58
4.4 Interface du moteur de recherche Elasticsearch. 59
4.5 Interface du moteur de recherche Solr. 60
4.6 Comparaison des 04 parties du Lucene. 62
4.7 Courbe de Comparaison du rappel des Trois Moteurs de Recherche. 63
4.8 Courbe de Comparaison de la précision des Trois Moteurs de Recherche. . . 63
4.9 Courbe de Comparaison de la f-mesure des Trois Moteurs de Recherche. . . . 64
4.10 Exemple de calcul du score. 65
4.11 Courbe de Rappel des 60 requêtes. 66
4.12 Courbe de la Précision des 60 requêtes. 67
4.13 Courbe de la F-mesure des 60 requêtes. 67
4.14 Page d'accueil du site web developpée. 68
4.15 Affichage des résultats de recherche des places pertinents. 68
4.16 Page d'informations touristiques de la place sélectionnée. 69

Liste des tableaux

1.1 Mesures de similarités utilisées dans le modèle vectoriel. 8

2.1 Travaux connexes à notre PFE. 35

3.1 Description des champs de la base de données. 45
3.2 Tableau des requêtes. 48
3.3 Comparaison entre les trois moteurs de recherche utilisé. 50
3.4 Quatre parties d'exécution de la recherche d'information. 50

4.1 Bibliothèques Python utilisées. 56
4.2 Les matrices de pertinence du Lucene. 61
4.3 Performances des moteurs de recherche selon les critères évalués. 62
4.4 Résultats finaux. 65
4.5 Matrice de pertinence finale. 66

Liste des abréviations et acronymes

ASM : Arabe Standard Moderne
BDD : Base De Données
CLEF : Cross-Language Evaluation Forum
CSV : Comma-Separated Value
DARPA : Defense Advanced Research Projects Agency
NIST : Institut National des Normes et de la Technologie
NLP : Natural Language Processing
OMT : Organisation Mondiale du Tourisme
PFE : Projet de Fin d'Etude
RI : Recherche D'information
RSV : Retrieval Status Value
SP : Nombre de Documents Pertinents Restitués par le Système
SRI : Système de Recherche d'Information
TF-IDF Term Frequency-Inverse Document Frequency.
TREC : Text REtrieval Conference
TLN : Traitement du Langage Naturel
XML : Extensible Markup Language

Introduction générale

Le tourisme joue un rôle crucial dans le paysage mondial contemporain, offrant des avantages économiques, culturels et sociaux tant aux voyageurs qu'aux destinations. En raison de l'importance croissante du tourisme, le développement de solutions innovantes pour faciliter la planification des voyages est essentiel.

L'Algérie se distingue par sa diversité naturelle et culturelle, en faisant une destination touristique unique. Le pays possède des plages enchanteuses sur la mer Méditerranée, les montagnes de l'Atlas idéales pour les aventures, et le Sahara qui offre des expériences exceptionnelles en tourisme désertique. De plus, l'Algérie regorge d'un riche patrimoine culturel qui se manifeste dans des villes historiques telles que Timgad et la Casbah d'Alger, ainsi que par des événements culturels et des festivals reflétant le folklore et les traditions ancestrales. Ces atouts font de l'Algérie un lieu idéal pour les touristes en quête d'expériences variées et de paysages magnifiques.

Malgré les importants atouts touristiques de l'Algérie, il existe un manque évident de promotion des zones touristiques, tant au niveau local qu'international. Ce manque conduit à une diminution du nombre de touristes et à une augmentation des défis auxquels le secteur touristique doit faire face. L'une des principales causes de ce manque est l'absence d'une stratégie de marketing touristique intégrée et l'utilisation sous-optimale des technologies modernes.

La langue joue un rôle important dans le développement du tourisme. En Algérie, la complexité de la langue arabe constitue un obstacle au développement des plateformes touristiques numériques capables d'attirer les touristes. Une grande partie du contenu touristique disponible sur Internet est rédigée dans des langues étrangères telles que le français et l'anglais, ce qui limite l'accès à l'information touristique pour les locuteurs de langue arabe. Ce déficit dans le développement des plateformes en arabe reflète la nécessité d'accroître les efforts pour renforcer le contenu numérique en langue arabe et améliorer l'expérience des touristes arabes.

De nos jours, le domaine de la recherche d'informations connaît une évolution rapide grâce aux progrès technologiques et aux innovations dans les techniques d'intelligence artificielle. Ce domaine contribue de manière significative à faciliter la vie des gens en fournissant un accès rapide à d'énormes quantités d'informations produites quotidiennement à travers des milliers de textes, d'articles et de nouvelles. Grâce aux moteurs de recherche, il est possible pour toute personne d'obtenir des réponses à ses questions en quelques secondes, qu'il s'agisse de textes, d'images, de graphiques ou de vidéos.

Problématique

La prolifération des informations touristiques en ligne a entraîné des défis dans l'accès à du contenu pertinent et fiable, en particulier pour les utilisateurs arabophones intéressés par le tourisme en Algérie. La nature complexe de la langue arabe rend cette tâche encore plus difficile, car il est complexe de développer des systèmes de recherche efficaces et précis dans ce contexte. Les moteurs de recherche actuels manquent souvent de la sensibilité linguistique et culturelle nécessaire pour répondre aux besoins de cette démographie, entraînant des expériences utilisateur suboptimales .

De plus, il est important de noter qu'il n'existe pas un base de données centralisé sur le tourisme en Algérie, ce qui complique encore davantage la création de solutions technologiques adaptées. Cette absence de données structurées limite la capacité à développer des outils de recherche et des applications mobiles efficaces qui pourraient bénéficier aux touristes potentiels et aux acteurs du secteur du tourisme en Algérie.

Ce projet vise à de remédier à ces lacunes et résoudre ces défis en identifiant des sources de données fiables et en structurant l'information de manière accessible et pertinente pour les utilisateurs arabophones désireux de découvrir les attraits touristiques de l'Algérie.

Objectifs

En raison du manque évident de ressources touristiques en arabe, nous avons décidé de créer une plateforme dédiée aux informations touristiques sur l'Algérie en langue arabe. et pour cela notre premier objectif est de créer une base de données contenant les destinations touristiques les plus célèbres des 58 wilayas afin de rendre ces informations plus accessibles aux utilisateurs arabophones.

Notre deuxième objectif est d'améliorer la qualité des résultats de recherche fournis par notre plateforme touristique. Pour ce faire, nous voulons comparer la pertinence par l'utilisation des moteurs de recherche open source entre les différentes versions de la base de données et les requêtes.

Enfin, nous visons à créer une plateforme intuitive et conviviale qui répond aux besoins spécifiques des utilisateurs arabophones intéressés par le tourisme en Algérie.

Organisation du mémoire

Ce mémoire est structuré en quatre chapitres comme suit :

Le premier chapitre, consacré à l'état de l'art, explore les différentes composantes de la recherche d'information ainsi que les systèmes et moteurs de recherche existants.

Dans le deuxième chapitre, intitulé Travaux Connexes, nous examinons les diverses études en lien avec notre recherche, les principales collections de données reconnues mondialement pour leur rôle essentiel dans l'évaluation et l'amélioration des systèmes de recherche d'information, ainsi que les moteurs de recherche les mieux classés au monde.

Le troisième chapitre détaille notre approche proposée pour la création d'un moteur de recherche, où nous expliquons en profondeur notre méthodologie, depuis la collecte des informations jusqu'à la conception de notre moteur de recherche touristique.

Enfin, le quatrième chapitre, « Implémentation et Évaluation », décrit les différentes ressources utilisées dans le projet et présente ainsi que discute les résultats obtenus au cours de notre travail, et finalement, nous présenterons notre site web qui est la matérialisation concrète des résultats de notre recherche.

Chapitre I

Etat de l'art

1 Introduction

Un autre aspect très important pour la description des intérêts de l'utilisateur est l'aspect La recherche d'information (RI) est une branche de l'informatique et des sciences de l'information qui permet à l'utilisateur un accès facile à l'information pertinente répondant à son besoin d'information à partir de vastes ensembles de données, telles que des documents textuels, des bases de données, des pages Web. . .etc Au moyen d'un Système de Recherche d'Information (SRI) qui est un ensemble de programmes informatiques conçus dans le but de trouver les informations les plus pertinents qui répondent aux besoins spécifiques des utilisateurs, exprimés sous forme de requête. Le domaine de la recherche d'informations reste en évolution constante depuis ses débuts dans les années soixante, où il se limitait à des recherches booléennes simples effectuées sur de grands ordinateurs dans des salles spacieuses. Au fil du temps, il a prospéré pour s'adapter aux changements technologiques, sociaux et économiques.
Dans ce chapitre,nous avons abordé en détail le domaine de la recherche d'information, en examinant toutes les nuances importantes qui y sont liées ,De plus, nous avons également abordé le sujet des systèmes de recherche d'information (SRI). Ensuite, nous avons passé aux moteurs de recherche, expliquant leur fonctionnement par les différentes étapes qu'ils suivent. nous avons discuté aussi de la langue arabe, qui constitue l'une des caractéristiques distinctives de notre moteur de recherche, en présentant divers sujets essentiels liés à cette langue. Enfin, nous avons examiné le tourisme, un élément fondamental de notre sujet, sans oublier que ce chapitre contient les déférentes définitions des termes qui ont une relation avec notre projet de fin d'étude.

2 Recherche d'information

2.1 Définitions

Plusieurs définitions de la recherche d'information ont vu le jour dans ces dernières années, nous citons :

- Définition 1 : La recherche d'information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations [Hernandez, 2005].

- Définition 2 : La recherche d'information est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information partinente [Sicard, 2010].

- Définition 3 : La recherche d'information est un ensemble de techniques offrant la possibilité de retrouver à partir d'une grande masse de documents ceux qui sont susceptibles correspondre aux besoins d'un utilisateur utilisant souvent une requête dans un langage naturel [Zargayouna, 2005].

D'apres ces definitions, on conclut que le RI a un objet pour d'extraire d'un document ou d'un ensemble de documents les informations pertinentes pour répondre aux besoins des utilisateurs.

2.2 modèles de la recherche d'information

Un modèle de recherche d'information (RI) offre une interprétation théorique de la notion de pertinence. Plusieurs modèles de RI ont été proposés dans la littérature, chacun s'appuyant sur des cadres théoriques différents tels que la théorie des ensembles, l'algèbre ou encore les probabilités. Globalement, on distingue trois principales catégories de modèles :

2.2.1 Le modèle booléen

C'est le premier modèle de RI [Salton et al., 1966], Il est basé sur la théorie des ensembles et l'algèbre de Boole. Dans ce modèle, un document **d** est représenté par un ensemble de mots-clés (termes), La requête **q** de l'utilisateur est egalement représentée par une expression logique, composée de termes reliés par des opérateurs logiques : AND($\wedge$), OR ($\vee$) et NOT ($\neg$).
La fonction de la récupération du modèle booléen considère qu'un document soit pertinent ou non pertinent,le résultat de cette fonction est donc binaire décrite comme suit :

$$RSV(d,q) = \begin{cases} 1 & \text{si } d \text{ appartient à l'ensemble décrit par } q \\ 0 & \text{sinon} \end{cases} \tag{1.1}$$

2.2.2 Le modèle vectoriel

Initialement proposé par Salton et implémenté dans le système SMART [Salton et al., 1966], C'est un modèle algébrique, il représente les documents ainsi que les requêtes par des vecteurs de termes dans un espace multidimensionnel (N dimensions, avec N : le nombre des termes d'indexation de la collection).
L'index d'un document dj est le vecteur :

$$d_j = (W_{1j}, W_{2j}, \ldots, W_{nj}) \tag{1.2}$$

Une requête est également représentée par un vecteur :

$$q = (W_{1q}, W_{2q}, \ldots, W_{nq}) \tag{1.3}$$

Avec :

$W_{ij} \in [0,1]$: le poids du terme t_i dans le document d_j.

W_{iq} : le poids du terme t_i dans la requête q.

Il y a plusieurs façons de calculer la similarité entre deux vecteurs, les plus répandues sont (Table 1.1) :

Mesure	Formule
Produit scalaire	$\text{RSV}(q, d_i) = \sum_{j=1}^{t} W_{qj} * W_{ij}$
Mesure de cosinus	$\text{RSV}(q, d_i) = \frac{\sum_{j=1}^{t} W_{qj} * W_{ij}}{\sqrt{\sum_{j=1}^{t} {W_{qj}}^2} * \sqrt{\sum_{j=1}^{t} {W_{ij}}^2}}$
Mesure de Dice	$\text{RSV}(q, d_i) = \frac{2 * \sum_{j=1}^{t} W_{qj} * W_{ij}}{\sum_{j=1}^{t} {W_{qj}}^2 + \sum_{j=1}^{t} {W_{ij}}^2}$
Mesure de Jaccard	$\text{RSV}(q, d_i) = \frac{2 * \sum_{j=1}^{t} W_{qj} * W_{ij}}{\sum_{j=1}^{t} {W_{qj}}^2 + \sum_{j=1}^{t} {W_{ij}}^2 - \sum_{j=1}^{t} W_{qj} * W_{ij}}$

TABLE 1.1 – Mesures de similarités utilisées dans le modèle vectoriel.

2.2.3 Le modèle probabiliste

Proposé par Robertson et Sparck Jones [Robertson and Jones, 1976], Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Etant donné une requête utilisateur Q et un document D, il s'agit de calculer la probabilité de pertinence du document pour cette requête.
Le score d'appariement entre le document D et la requête Q, noté RSV (Q, D) est donné par :

$$RSV(Q, D) = \frac{p(R/D)}{p((R/D))} \tag{1.4}$$

2.3 Concepts de base de la recherche d'information

Dans le domaine de la recherche d'informations, il existe de nombreux concepts de base, notamment :

2.3.1 Collection de documents

La collection de documents (ou corpus) constitue l'ensemble des informations (des documents) exploitables et accessibles.

2.3.2 Document

Un document peut être un texte, un morceau de texte, une image, une bande vidéo, etc. On appelle document toute unité d'information qui peut constituer une réponse à un besoin en information/requête d'un utilisateur [Menegon et al., 2009].

2.3.3 Requête

La requête est l'expression du besoin en information de l'utilisateur. Elle est un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique.

2.3.4 Besoin d'information

Le besoin d 'information est une expression mentale de ce que l'utilisateur souhaite a rechercher. Ingwersen [Ingwersen, 1994] a défini trois types :

- Besoin vérificatif : l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà.

- Besoin thématique connu : l'utilisateur cherche clarifier, au revoir ou à trouver de nouvelles informations dans un sujet et domaine connu.

- Besoin thématique inconnu : : pour ce type de besoin l'utilisateur cherche de nouveau concept ou de nouvelles relations hors des sujets ou domaine qui lui sont familiers.

2.3.5 Pertinence

La pertinence est une notion fondamentale dans le domaine de la RI. La pertinence peut être définie comme la correspondance entre un document et une requête, ou encore une mesure d'informativité du document à la requête [Boughanem and Savoy, 2008]. nous citons les quatre les plus importantes : Pertinence algorithmique, Pertinence thématique, Pertinence cognitive et Pertinence situationnelle [Boughanem and Savoy, 2008] [Tamine and Calabretto, 2008] .

2.4 Traitement du langage naturel(NLP)

2.4.1 Définition

Le Traitement du Langage Naturel (TLN), appelé le Natural Language Processing en anglais, est un domaine de l'intelligence artificielle. Le NLP a pour mission de permettre aux machines de comprendre, d'analyser et de générer le langage humain de manière naturelle Comme indiqué dans la figure 1.1. Contrairement à un langage de programmation formel, le langage naturel est complexe, contextuel et parfois ambigu. Le NLP implique, entre autres, l'usage de techniques d'apprentissage et de traitement automatique des langues. Elles servent à extraire des informations significatives à partir de données textuelles et linguistiques. Les ordinateurs sont ainsi en mesure de saisir le sens des mots et les relations entre eux. Ils appréhendent le contexte global d'une conversation ou d'un texte [Baclic et al., 2020].
Dans le même ordre d'idées, on peut conclure que le Traitement du Langage Naturel (NLP) est multidisciplinaire car il emprunte à d'autres disciplines, principalement la linguistique, la psycholinguistique, les sciences cognitives, l'informatique et les statistiques.

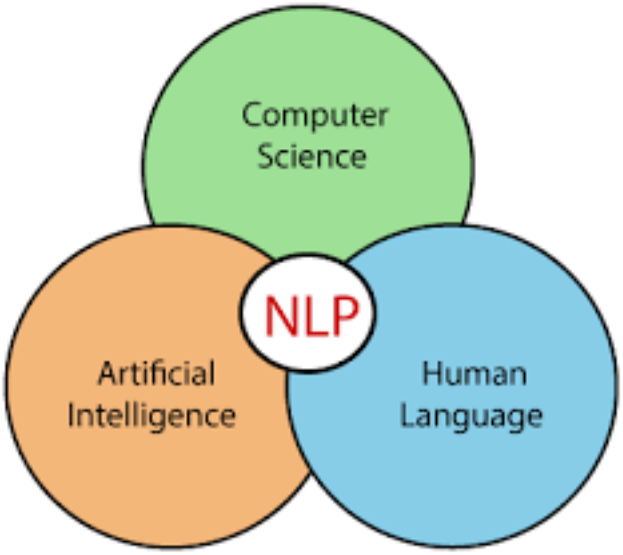


FIGURE 1.1 – Traitement du langage naturel.

2.4.2 Applications du NLP

Parmi les utilisations du Traitement du Langage Naturel (NLP), on peut énumérer les suivantes :

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

- Analyse du ressenti sur les médias sociaux : Le traitement du langage naturel est devenu un outil commercial essentiel pour découvrir des informations cachées sur les médias sociaux. L'analyse du ressenti permet d'analyser le langage utilisé dans les messages des médias sociaux, les réponses, les critiques, etc., afin d'extraire les attitudes et les émotions en réponse aux produits, aux promotions et aux événements. Ces informations peuvent ensuite être utilisées par les entreprises dans la conception des produits, les campagnes publicitaires, etc.

- Classification de texte : La classification de texte permet d'attribuer des catégories prédéfinies à un document et de l'organiser pour vous aider à trouver les informations dont vous avez besoin ou simplifier certaines activités. Par exemple, une application de catégorisation de texte est le filtrage du spam dans les e-mails.

- Résumé automatique : Les méthodes NLP sont également utilisées pour produire des résumés courts, précis et fluides d'un document texte plus long.

- Chatbots : Les méthodes NLP sont au cœur du fonctionnement des Chatbots actuels. Bien que ces systèmes ne soient pas totalement parfaits, ils peuvent aujourd'hui facilement gérer des tâches standards telles renseigner des clients sur des produits ou services, répondre à leurs questions, etc. Ils sont utilisés par plusieurs canaux, dont l'Internet, les applications et les plateformes de messagerie. L'ouverture de la plateforme Facebook Messenger aux Chatbots en 2016 a contribué à leur développement .

3 Systèmes de recherche d'information :

3.1 Définition

Un Système de Recherche d'Informations (SRI) est un système informatique qui facilite l'accès à un ensemble de documents, pour permettre de retrouver ceux dont le contenu correspond le mieux à un besoins d'Informations d'un utilisateur [Diallo, 2023].

3.2 Processus de recherche d'information

Les différentes étapes du processus de RI, sont représentées schématiquement par le processus en U (voir Figure. 1.2) . La figure illustre particulièrement : les notions de documents et de requêtes qui sont des conteneurs d'informations, les opérations d'analyse, d'indexation et d'appariement qui permettent globalement de traiter la requête dans le but de sélectionner des documents à présenter à l'utilisateur.

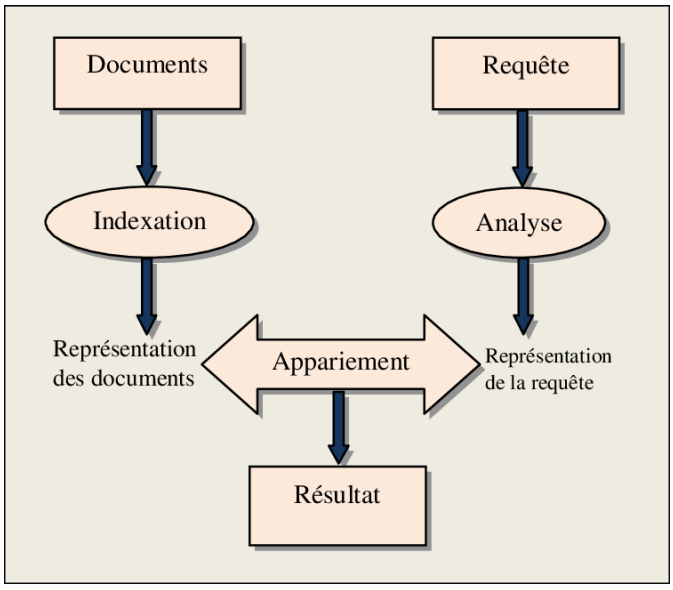


FIGURE 1.2 – Processus en U de la recherche d'information.

3.3 Mesures d'évaluation des SRI

Le principal objectif d'un système de recherche d'Information est de restituer à l'utilisateur tous les documents pertinents et de rejeter tous les documents non pertinents. Cet objectif est évalué à l'aide de différentes mesures d'évaluation [Sanderson et al., 2010]. Nous présentons ci dessous les plus utilisées : La précision (P) , Le rappel (R) , Le Bruit (B) , Le silence(S) , La F-mesure.

3.3.1 La précision (P)

La précision est le rapport du nombre de documents pertinents restitués par le système (SP) sur le nombre total de documents restitués (R), exprimée ainsi :

$$\text{Précision} = \frac{\text{SP}}{R} \tag{1.5}$$

3.3.2 Le rappel(R)

Le rappel est calculé par le rapport du nombre de documents pertinents restitués(SP) sur le nombre total de documents pertinents(P), exprimé ainsi :

$$\text{Rqppel} = \frac{\text{SP}}{P} \tag{1.6}$$

3.4 F-mesure

Est une mesure qui estime une moyenne harmonique pondérée de la précision et du rappel et est décrite par la formule suivante [Amer, 2020] :

$$F\text{-}mesure = \frac{2 \times \text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{1.7}$$

Des mesures complémentaires au rappel et précision ont été définies, il s'agit de bruit et de silence.

3.4.1 Le bruit(B)

La mesure d'évaluation bruit est une notion complémentaire à la précision, elle est définie par : B = 1 - P
où P est la précision du SRI.

3.4.2 Le silence(S)

La mesure d'évaluation «silence» est une notion complémentaire au rappel, elle est définie par : S = 1 - R
où R est le rappel du SRI.

4 Moteur de recherche :

4.1 Definition :

Les moteurs de recherche sont des applications Web conçues pour faciliter l'accès à différentes sources d'information en les indexant et en permettant aux utilisateurs de trouver des documents via une interface de recherche. Ils effectuent des recherches sur le Web en utilisant des mots-clés et des termes de recherche, également connus sous le nom de requêtes. En réponse à ces requêtes, le moteur de recherche présente automatiquement les informations pertinentes, les proposant dans un ordre défini en fonction de ses propres critères. [Le Ny, 2013]

4.2 Composants d'un moteurs de recherche :

Considérant la complexité des opérations impliquées dans le fonctionnement d'un moteur de recherche afin de fournir des résultats adaptés à la demande de l'utilisateur, il est évident qu'il ne s'agit pas simplement d'un outil basique. En effet, un moteur de recherche est composé de divers éléments essentiels qui interagissent de manière synergique pour permettre son fonctionnement optimal. Généralement, les moteurs de recherche sont constitués de trois éléments fondamentaux :

4.2.1 Crawler (ou Spider) :

Un crawler, également appelé robot d'indexation ou spider, est un logiciel essentiel au fonctionnement des moteurs de recherche. Sa principale mission consiste à explorer activement le Web en suivant les liens hypertexte pour découvrir de nouvelles pages et extraire des informations pertinentes de ces pages. Le crawler parcourt systématiquement les sites, collectant des données telles que le contenu textuel, les liens vers d'autres pages, les titres de pages et les mots-clés. Ces données sont ensuite indexées dans la base de données du moteur de recherche, permettant ainsi une recherche efficace. En somme, le crawler joue un rôle crucial dans l'acquisition et la mise à jour des informations contenues dans l'index du moteur de recherche. [Boudry, 2002]

4.2.2 Indexeur :

Un indexeur est une partie importante d'un moteur de recherche. Son travail consiste à organiser les informations trouvées sur les pages web pour que le moteur puisse les rechercher rapidement. Il découpe les pages en différentes parties, comme les titres et les liens, pour les ranger correctement. Il s'assure également que les informations sont mises à jour régulièrement. En résumé, l'indexeur aide les utilisateurs à trouver ce qu'ils cherchent sur internet. [Boudry, 2002]

4.2.3 L'interrogateur :

L'interrogateur, souvent désigné comme le moteur de recherche, est un composant crucial d'un moteur de recherche. Son rôle principal est de traiter les requêtes des utilisateurs. Lors-

qu'un utilisateur effectue une recherche, l'interrogateur examine cette requête et parcourt l'index du moteur de recherche pour trouver les pages web les plus appropriées. Il utilise un algorithme spécifique pour effectuer cette recherche dans la base de données. Ensuite, l'interrogateur classe ces pages en fonction de leur pertinence et les présente comme résultats de recherche à l'utilisateur. En résumé, l'interrogateur aide les utilisateurs à trouver ce qu'ils cherchent en analysant leurs demandes et en utilisant un algorithme de recherche pour interroger la base de données du moteur de recherche.

4.3 Requête :

Une requête est l'expression utilisée par un utilisateur pour obtenir des informations précises. Cette expression peut prendre la forme de mots-clés, de phrases ou de questions. De plus, dans les moteurs de recherche modernes, une requête peut prendre la forme d'une image. L'utilisateur utilise cette requête pour indiquer ce qu'il recherche, tandis que le système de recherche utilise cette requête pour localiser et présenter les résultats pertinents. Il est donc essentiel que la requête soit compréhensible et a un sens avec le but de recherche. [Bouzeghoub and Kostadinov, 2005]

4.4 Fonctionnalité d'un moteur de recherche :

Bien qu'il soit connu que le fonctionnement des moteurs de recherche varie d'un moteur à l'autre, Le fonctionnement du moteur de recherche peut être simplifié en trois étapes principales, qui peuvent être exprimées dans le schéma suivant :

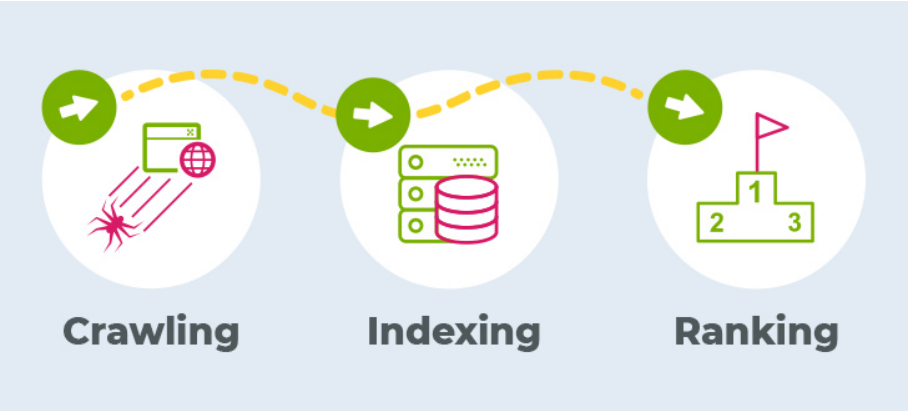


FIGURE 1.3 – Fonctionnement d'un moteur de recherche.

Il existe trois étapes de base partagées par tous les moteurs de recherche, à savoir : Le crawling : découverte de nouvelles pages web par les robots d'indexation L'indexation : trie et stockage des pages indexées par le moteur Le classement : définition de l'ordre des pages selon leur pertinence vis-à-vis des mots-clés recherchés. [Farchy and Méadel, 2013]

Le crawling implique l'exploration systématique des sites web par des robots appelés « spiders ou crawlers », qui collectent des informations à partir des pages en ligne. Ces données sont ensuite transmises à l'index, qui agit comme le cerveau du moteur de recherche. L'index

évalue les informations et les classe en fonction de leur pertinence [Vitali, 2006]. Ensuite, une étape de classement intervient où le moteur de recherche utilise un algorithme complexe et confidentiel pour analyser les données indexées. Cet algorithme prend en compte une série de critères de pertinence tels que la qualité du contenu, comprenant l'originalité, la longueur et la richesse sémantique, la réputation du site, évaluée à travers des aspects tels que son autorité et le nombre de liens entrants, la concordance des mots-clés avec la requête de l'utilisateur, l'ergonomie et la convivialité du site, ainsi que la localisation géographique de l'internaute. Ce processus permet au moteur de recherche de présenter les résultats de recherche de manière à fournir les réponses les plus adaptées aux requêtes des utilisateurs. [Farchy and Méadel, 2013] [Foenix-riou, 2022]

4.5 Type existants :

4.5.1 moteurs de recherche en ligne « online » :

Un moteur de recherche en ligne, également appelé moteur de recherche online, est un outil permettant aux utilisateurs d'effectuer des recherches sur Internet en entrant des mots-clés, des phrases ou d'autres requêtes afin de trouver des informations pertinentes. par exemple Google, Yahoo et Bing.

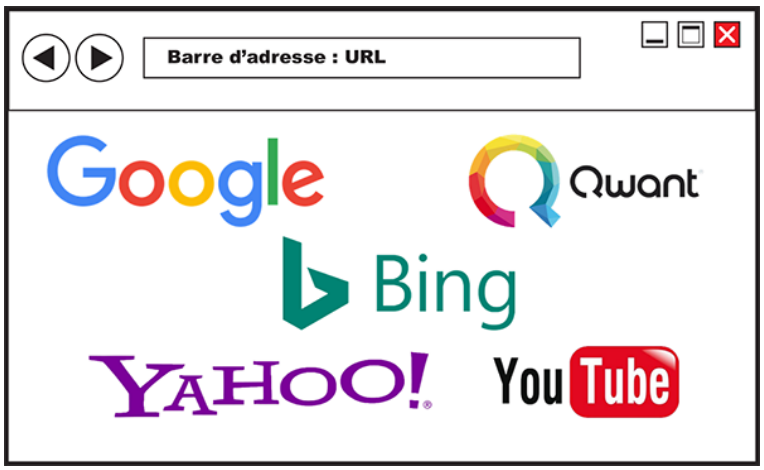


FIGURE 1.4 – Les Moteur de recherche online.

4.5.2 moteurs de recherche hors ligne « offline » :

Un moteur de recherche offline est un outil qui peut être utilisé sans connexion internet, permettant ainsi de rechercher des informations stockées localement sur l'appareil de l'utilisateur. Un exemple notable est l'application Google Desktop, Google Desktop [1] est un moteur de recherche offline de bureau qui facilite la recherche de documents sur votre ordinateur. Cet outil n'étant pas limité aux noms de fichiers, il est en mesure d'effectuer des recherches dans le contenu de divers types de documents, Word, PDF, PowerPoint, Excel, Images, Vidéos, musiques ainsi que vos e-mails et leurs pièces jointes, Google Desktop.

1. https://www.01net.com/telecharger/internet/moteur_rech/google-desktop.html



FIGURE 1.5 – Moteur de recherche offline "Google Desktop".

5 langue arabe :

La langue arabe occupe une place significative à l'échelle mondiale. Parlée par plus de 500 millions de personnes en tant que langue maternelle, elle est également la langue officielle dans 24 pays et reconnue comme l'une des six langues officielles par les Nations Unies. Ces pays englobent les 22 États membres de la Ligue arabe, ainsi que l'Érythrée et le Tchad. S'étendant sur plusieurs continents, la moitié de ces nations arabophones se trouvent en Afrique, incluant l'Algérie, l'Égypte, la Libye et la Tunisie, entre autres, tandis que l'autre moitié réside en Asie, avec des pays tels que l'Arabie saoudite, l'Irak et la Jordanie, entre autres. Cette diversité linguistique reflète l'influence étendue et l'importance de l'arabe à travers les frontières et les cultures [Bernier, 2021].

De plus, l'arabe n'est pas simplement une langue, mais un riche ensemble de dialectes et de variantes, avec une variante particulière se démarquant comme la norme écrite formelle utilisée dans les médias, l'éducation et l'expression culturelle à travers le monde arabe [Habash, 2010].



FIGURE 1.6 – Le monde arabe.

5.1 L'historique de la langue arabe :

Il est difficile d'aborder l'étude d'une langue sans référer à l'histoire qu'elle véhicule. La langue arabe est Née dans la péninsule arabique et ses marges syro-iraquiennes, la langue arabe appartient à la grande famille des langues sémitiques apparues plusieurs siècles avant l'ère chrétienne.
Sa trajectoire aura été marquée par une interaction dynamique avec d'autres aires linguistiques et culturelles. Ce fait est sensible dès l'émergence d'une langue « proto-arabique », exprimée à travers des alphabets d'abord empruntés à d'autres langues (Akkadien). L'arabe « ancien » est attesté ensuite à travers des inscriptions en caractères nabatéens, dont la plus célèbre est celle d'Al Namâra datée de 328, découverte en Syrie méridionale par deux archéologues français. Jusqu'au VIe siècle, un dialecte littéraire supra-tribal s'affirme comme la langue de la poésie. L'apparition d'une nouvelle religion au VIIe siècle lui donne sa dimension sacrée à travers ce que Charles Pellat appelle « le premier monument littéraire des Arabes », le Coran. C'est le corpus conjoint du texte coranique et de la poésie antéislamique qui constituera les fondements de l'arabe dit « classique » [Rady et al., 2022].

5.2 La richesse de la langue arabe :

La langue arabe est un monde à part entière, Si on procède par le nombre de mots, c'est la langue qui occupe la première place et de loin, sa richesse créa un vocabulaire fort étendu et une rare souplesse de formes, comme étant une « langue savante,compliquée et très riche, elle s'écrit au moyen de 28 lettres (voir figure 6). Cet alphabet est un abjad : terme décrivant un système d'écriture ne notant que les consonnes de la langue. »
Ces consonnes peuvent devenir également des longues voyelles comme (ا), (و) et (ي).

خ	ح	ج	ث	ت	ب	أ
khah	hah	jeem	theh	teh	beh	alef
ص	ش	س	ز	ر	ذ	د
sad	sheen	seen	zain	reh	thal	dal
ق	ف	غ	ع	ظ	ط	ض
qaf	feh	ghain	ain	zah	tah	dad
ي	و	ه	ن	م	ل	ك
yeh	waw	heh	noon	meem	lam	kaf

FIGURE 1.7 – L'alphabet arabe.

L'écriture arabe comporte également des voyelles qui ne sont pas essentielles à l'écriture, il existe de plus une série d'autres diacritiques de syllabation dont les plus courants sont

l'indication de l'absence de voyelle آ (السكون) et la gémination des consonnes آ.(الشدة)

« Les grammairiens arabes prétendent que toutes les racines de leur langue, au nombre de 6000, ont été primitivement des verbes. Les mots composés de ces racines se complètent, soit au moyen de lettres, soit par le redoublement des radicales, ou encore par le changement des voyelles. C'est ainsi qu'une même racine peut donner des verbes, des substantifs, des adjectifs, des adverbes, enfin des dérivés de toute sorte.
L'arabe dispose également d'une richesse de vocabulaire et de figures rhétoriques inimaginables. Il y aurait ainsi 80 termes différents pour représenter le miel, 200 pour le serpent, 500 pour le lion, 1000 pour le chameau, autant pour l'épée, et jusqu'à 4000 pour rendre l'idée de malheur. De la même manière. Cette richesse sémantique témoigne du raffinement de la civilisation et de la littérature arabo-islamique. Il existe une multitude de nuances d'idées, chacune traduite par un terme spécifique. » [Chouikha, 2010]
Malheureusement, Malgré la richesse de la langue arabe, mais elle n'arrive pas à suivre l'évolution et le développement de certaines langues comme le français ou l'anglais, ce qui pousse les arabophone à reprendre des mots à ces langues étrangères pour désigner certaines choses au lieu de créer des mots propres à la langue arabe.

5.3 Variétés de la langue arabe :

Les linguistes arabes ont structuré la langue arabe en deux variétés principales, la première variété dite « l'arabe classique » ou « l'arabe littéraire » et la deuxième dite « l'arabe dialectal ». D'autres linguistes ont ajouté une variété intermédiaire écrite et parlée, et désignée par le nom « arabe standard contemporain ».

5.3.1 L'arabe classique :

C'est la forme linguistique ancienne de l'arabe dont les règles de la grammaire étaient fixées entre le 8^{e} et le 10^{e} siècle siècle. Il est appris dans les établissements d'enseignement à travers la littérature arabe et les cours de théologie [Mars, 2016].

5.3.2 Arabe standard moderne ou contemporain (ASM) :

C'est une variante moins formelle que l'arabe classique. L'arabe standard est utilisé dans la vie officielle, universitaire et administrative. De plus, c'est par le biais de l'arabe standard, que deux locuteurs arabophones d'origines dialectales différentes sont susceptibles de se comprendre [Mars, 2016].

5.3.3 Arabe dialectal :

Il est souvent utilisé dans l'expression de la vie quotidienne locale. Elle est considérée comme la langue vernaculaire de l'ensemble des arabophones. Par définition, les dialectes arabes sont les langues maternelles des populations des pays arabes, et ces formes linguistiques sont parfois très différentes d'une région à l'autre. [Mars, 2016]
On peut regrouper ces dialectes en quatre grands groupes [Boulaknadel, 2008] :
• Les dialectes arabes, parlés dans la Péninsule Arabique : dialectes du Golfe, dialecte du

najd, yéménite .
• Les dialectes maghrébins : algérien, marocain, tunisien, hassaniya de Mauritanie .
• Les dialectes proche-orientaux : égyptien, soudanais, syro-libano-palestinien, irakien (nord et sud) .
• La langue maltaise : est également considérée comme un dialecte arabe.

6 Le tourisme

6.1 Définition

Le mot « Tourisme » vient de la transcription Anglaise d'un vocable français « THE TOUR », qui a été utilisé pour la première fois en 1841 désignant la personne qui faisait le grand tour, cette expression désigne le voyage sur le continent, c'est à ce moment-là que le tourisme naisse. Le tourisme a connu une multitude de définitions relatives et variables selon le temps et le lieu, donc difficile à définir d'une manière précise car il existe une diversité de définitions dont ont privilège la suivante :

L'encyclopédie Universalise de 9éme édition qui indique que « le tourisme est l'expression d'une mobilité humaine et sociale fondée sur un excédent budgétaire susceptible d'être consacré au temps libre passé à l'extérieur de la résidence principale, il implique au moins un découché ». [Hélène, 2009]

Selon l'Organisation Mondiale du Tourisme (OMT) « L'art de satisfaire les aspirations les plus divers qui incitent l'homme à se déplacer hors deson univers quotidien et de son cache habituel, à la recherche de dépaysement et d'évasion» [2]

Définition selon le conseil économique et social [Unies, 2003] :
Ensemble des activités liées :

- Au déplacement des personnes sur une certaine distance dans le cadre d'une activité de loisir.
- Sur le plan statistique la définition inclut les déplacements de plus de 24h.

6.2 Rôles du tourisme

6.2.1 rôle économique

• Rentrée de devises et circulation des monnaies.
• Création d'emplois et donc la lutte contre le chômage.
• Développement économique du pays.
• Apport positif dans l'aménagement du territoire.

2. https://www.unwto.org/fr

6.2.2 rôle social

• Echange de cultures et de savoir du au contact des personnes des différents recoins du monde.
• Diminution des pensées racistes et ouverture sur le monde extérieur.

6.2.3 rôle culturel

• Faire connaître aux touristes les traditions, l'histoire, La culture d'un peuple.

• Développement de l'artisanat et de l'art.

• La mise en valeur des potentialités du pays en matière de patrimoine historique et architectural. [HADRI, 2017]

6.3 Le tourisme en Algérie

L'Algérie possède diverses ressources touristiques dont la grande partie n'est pas exploitée. C'est un pays de contrastes, du littoral au désert passant par les hauts plateaux et les steppes. Il est situé au nord de l'Afrique, s'étendant sur une superficie de 2.381.741 m^2. En effet le nord Algérien s'ouvre sur la mer Méditerranée, sur une longueur de 1200 km de côte, son climat extrêmement favorable, permet de voyager en toute saison, soit en bord de mer en saison estivale, soit à la montagne (les hauts plateaux). La saison hivernale permet de mieux découvrir la splendeur du Sahara et les sommets enneigés des montagnes de Kabylie et des Aurès. [Fahima, 2023]



FIGURE 1.8 – Tourisme en Algérie .

7 Conclusion

Dans ce chapitre, nous avons abordé plusieurs sujets. Tout d'abord, nous avons examiné la recherche d'information (RI) en fournissant quelques définitions et en présentant les différents modèles : le modèle booléen, le modèle vectoriel et le modèle probabiliste, ainsi que les principaux concepts qui les accompagnent. Ensuite, nous avons expliqué le fonctionnement des systèmes de recherche d'information (SRI) et le processus de la RI, en soulignant les mesures d'évaluation utilisées pour évaluer leur performance. De plus, nous avons exploré les moteurs de recherche, en détaillant leurs composants et leur fonctionnement, ainsi que les différents types existants. Ensuite, nous avons abordé la langue arabe, son histoire, sa richesse et sa diversité. Enfin, nous avons discuté du tourisme et de son rôle économique, social et culturel, en présentant également le tourisme en Algérie. Donc, une synthèse exhaustive des thèmes principaux traités dans ce chapitre a été fournie, ouvrant la voie à une exploration et une analyse plus approfondies dans les chapitres à venir.

Chapitre II

Travaux Connexes

1 Introduction

Dans ce chapitre, nous mettons en lumière les travaux connexes étroitement liés à notre projet de fin d'études, en accordant une attention particulière à deux initiatives majeures dans le domaine de la recherche d'information : le Text Retrieval Conference (TREC) et le Forum d'évaluation et de laboratoires (CLEF).

Depuis ses débuts en 1992, TREC s'engage à évaluer les systèmes de recherche d'information à grande échelle en fournissant des ensembles de données de test et des tâches d'évaluation spécifiques. De même, CLEF organise des campagnes annuelles d'évaluation pour les systèmes de recherche.

En mettant l'accent sur TREC et CLEF dans ce chapitre, nous mettons en évidence leur définition, leur histoire, ainsi que leurs objectifs, reconnaissant ainsi leur importance pour notre projet de fin d'études. L'évaluation des systèmes de recherche dans CLEF se concentre sur le même domaine que TREC, reflétant ainsi leur objectif commun d'améliorer les performances des systèmes de recherche et de fournir des opportunités de développement.

Après avoir expliqué TREC et CLEF, nous avons procédé à une sélection minutieuse de six travaux connexes à notre projet de fin d'études, couvrant divers domaines allant de la recherche en langue arabe à l'étude des moteurs de recherche. Notre objectif dans ce chapitre est de présenter succinctement les détails des travaux choisis. Tout en abordant quelques définitions et concepts fondamentaux des moteurs de recherche renommés tels que Google, Bing,YANDEX et Yahoo.

2 Fondements et Évolutions en Recherche d'Information

Les initiatives telles que TREC et CLEF sont essentielles pour évaluer et améliorer les technologies de recherche d'information. Elles fournissent des plateformes standardisées pour tester et comparer différents systèmes, stimulant ainsi l'innovation et l'excellence dans ce domaine.

2.1 Trec

2.1.1 Définition du Trec

La Conférence sur la Récupération de Texte (TREC) est une série d'ateliers qui encouragent la recherche en matière de récupération d'informations à partir d'applications réalistes en fournissant de vastes collections de tests, des procédures de notation uniformes, et un forum pour les organisations intéressées par la comparaison des résultats. Le premier atelier a eu lieu en novembre 1992, et des ateliers (suivant des évaluations) ont été organisés chaque année depuis lors. En plus de la tâche principale, six sous-tâches supplémentaires, appelées "tracks", permettent aux participants de se concentrer sur des sous-problèmes communs particuliers dans la récupération, tels que la récupération à travers les langues, la récupération

de la parole, et la récupération à des niveaux élevés de précision. TREC est co-parrainé par l'Institut National des Normes et de la Technologie (NIST) et le Bureau de la Technologie de l'Information de l'Agence des Projets de Recherche Avancée de Défense (DARPA). [Harman, 1998]

La figure (2.1) répertorie tous les pistes (track) qui ont été organisés lors de chaque TREC et donne le nombre de groupes participants à chaque track chaque année. La dernière ligne du tableau fournit le nombre total de participants distincts par TREC. Les tâches proposées au sein des parcours offerts pour un TREC donné ont divergé au fur et à mesure de l'évolution du TREC.

	TREC											
Track	92	93	94	95	96	97	98	99	00	01	02	03
Ad hoc	18	24	26	23	28	31	42	41	—	—	—	—
Routing	16	25	25	15	16	21	—	—	—	—	—	—
Interactive	—	—	3	11	2	9	8	7	6	6	6	—
Spanish	—	—	4	10	7	—	—	—	—	—	—	—
Confusion	—	—	—	4	5	—	—	—	—	—	—	—
Database merging	—	—	—	3	3	—	—	—	—	—	—	—
filtering	—	—	—	4	7	10	12	14	15	19	21	—
Chinese	—	—	—	—	9	12	—	—	—	—	—	—
NLP	—	—	—	—	4	2	—	—	—	—	—	—
Speech	—	—	—	—	—	13	10	10	3	—	—	—
Cross-language	—	—	—	—	—	13	9	13	16	10	9	—
High precision	—	—	—	—	—	5	4	—	—	—	—	—
Very large corpus	—	—	—	—	—	—	7	6	—	—	—	—
Query	—	—	—	—	—	—	2	5	6	—	—	—
Question answering	—	—	—	—	—	—	—	20	28	36	34	33
Web	—	—	—	—	—	—	—	17	23	30	23	27
Video	—	—	—	—	—	—	—	—	—	12	19	—
Novelty	—	—	—	—	—	—	—	—	—	—	13	14
Genome	—	—	—	—	—	—	—	—	—	—	—	29
HARD	—	—	—	—	—	—	—	—	—	—	—	14
Robust	—	—	—	—	—	—	—	—	—	—	—	16
Total participants	25	31	33	36	38	51	56	66	69	87	93	93

FIGURE 2.1 – Nombre de participants par piste et nombre total de participants distincts dans chaque TREC.

2.1.2 Historique de Trec

En 1990, le NIST a été chargé de construire une grande collection de tests pour évaluer la technologie de recherche de texte développée dans le cadre du projet TIPSTER de l'Agence des projets de recherche avancée de défense (DARPA). Cette collection devait comporter environ un million de documents texte complets : environ cent fois plus grande que les collections de tests non propriétaires existantes. L'année suivante, le NIST a proposé que cette grande collection de tests soit mise à la disposition de l'ensemble de la communauté de recherche par la création de TREC. On croyait que les deux éléments manquants dans l'expérimentation en récupération de l'information (RI) pourraient être abordés en fournissant

une grande collection de tests, des procédures de notation uniformes et un forum pour les organisations intéressées par la comparaison des résultats. [Voorhees and Harman, 2005]

2.1.3 Objectifs de Trec

— Encourager la recherche en récupération de texte basée sur de grandes collections de tests. Pour accroître la communication entre l'industrie, le milieu universitaire et le gouvernement en créant un forum ouvert pour l'échange d'idées de recherche.

— Accélérer le transfert de technologie des laboratoires de recherche vers les produits commerciaux en démontrant des améliorations substantielles dans les méthodologies de récupération sur des problèmes du monde réel.

— Augmenter la disponibilité de techniques d'évaluation appropriées à utiliser par l'industrie et le milieu universitaire, y compris le développement de nouvelles techniques d'évaluation plus applicables aux systèmes actuels.

— Pour servir de vitrine aux systèmes de récupération de données à la pointe de la technologie pour la DARPA et ses clients. [Harman, 1996] [Voorhees and Harman, 2005]

2.2 Clef

2.2.1 Définition du Clef

CLEF (Cross-Language Evaluation Forum) est un projet qui évalue les systèmes de recherche d'informations, en mettant l'accent sur les langues européennes et en abordant les défis de l'extraction de texte interlingue. CLEF vise à surmonter les barrières linguistiques dans la recherche d'informations en évaluant les méthodes permettant d'accéder à l'information dans des documents rédigés dans différentes langues ou lorsque la langue de l'utilisateur diffère de la langue du document. Il sert de plateforme pour faire progresser la recherche en matière de recherche d'informations multilingue et promouvoir le développement de technologies qui facilitent l'accès efficace à l'information dans divers contextes linguistiques. [Cancedda and Renders, 2010]

La figure (2.2) représente le nombre de participants par groupe au CLEF, où les participants sont regroupés selon différents critères. On remarque que les groupes AdHoc et QA@TREC (groupe TREC Question Answer) se distinguent comme les plus dominants. [Tsikrika et al., 2013]

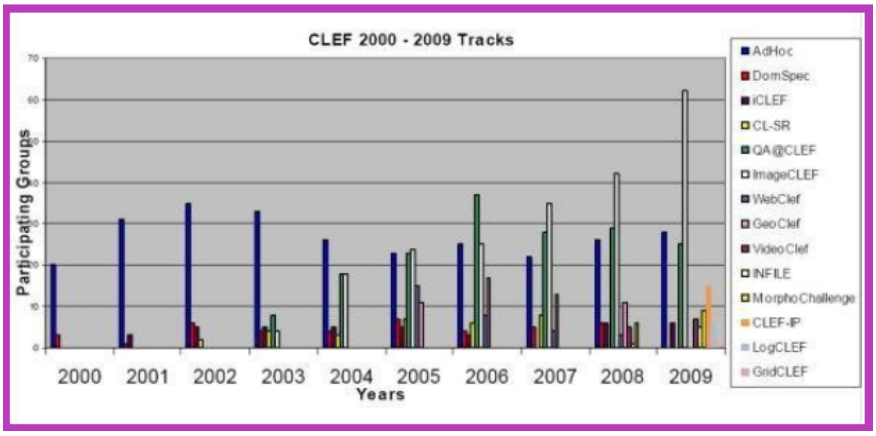


FIGURE 2.2 – Groupes participants du CLEF entre 2000 ET 2009.

Le tableau suivant représente les sources et les dimensions principales de la collection de documents du CLEF. Chaque entrée indique le nom du journal, l'année d'ajout à la collection, la taille en mégaoctets (MB), le nombre de documents, le nombre de tokens (mots), et le nombre de caractéristiques extraites. Les collections couvrent une variété de langues telles que le néerlandais, l'anglais, le finnois, le français, l'allemand, l'italien, le russe, l'espagnol, et le suédois, chacune avec ses spécificités et ses caractéristiques uniques. Cette diversité linguistique et les caractéristiques extraites illustrent la richesse et la complexité des données disponibles pour l'évaluation des systèmes de recherche d'information dans le cadre du CLEF.

Collection	Added in	Size (MB)	No. of docs	Median size of docs. (Bytes)	Median size of docs. (Tokens)[a]	Median size of docs. (Features)
Dutch: Algemeen Dagblad 94/95	2001	241	106483	1282	166	112
Dutch: NRC Handelsblad 94/95	2001	299	84121	2153	354	203
English: LA Times 94	2000	425	113005	2204	421	246
English: Glasgow Herald 95	2003	154	56472	2219	343	202
Finnish: Aamulehti late 94/95	2002	137	55344	1712	217	150
French: Le Monde 94	2000	158	44013	1994	361	213
French: ATS 94	2001	86	43178	1683	227	137
French: ATS 95	2003	88	42615	1715	234	140
German: Frankfurter Rundschau 94	2000	320	139715	1598	225	161
German: Der Spiegel 94/95	2000	63	13979	1324	213	160
German: SDA 94	2001	144	71677	1672	186	131
German: SDA 95	2003	144	69438	1693	188	132
Italian: La Stampa 94	2000	193	58051	1915	435	268
Italian: AGZ94	2001	86	50527	1454	187	129
Italian: AGZ 95	2003	85	48980	1474	192	132
Russian: Izvestia 95[b]	2003	68	16761			
Spanish: EFE 94	2001	511	215738	2172	290	171
Spanish: EFE 95	2003	577	238307	2221	299	175
Swedish: TT 94/95	2002	352	142819	2171	183	121

FIGURE 2.3 – Quelques informations sur les différentes collections du CLEF.

2.2.2 Objectifs de Clef

— Fournir une infrastructure pour le test et l'évaluation des systèmes de recherche d'informations fonctionnant sur les langues européennes, à la fois dans des contextes monolingues et multilingues.
— Construire des ensembles de tests de données réutilisables pouvant être utilisés par les développeurs de systèmes à des fins de référence.
— Créer une communauté de recherche et développement dans le domaine de la recherche d'informations multilingue (CLIR).

2.3 Travaux connexe

De nombreuses études ont été réalisées ces dernières années sur la mesure de la similarité entre documents afin de faciliter la recherche d'informations dans des systèmes d'information complexes tels que les moteurs de recherche. Ces recherches ont conduit à plusieurs projets visant à créer des moteurs de recherche pour différentes langues, y compris l'arabe, ainsi qu'à des initiatives spécifiques axées sur l'amélioration du traitement automatique du langage naturel pour la langue arabe. Dans la section suivante, nous présenterons quelques travaux similaires à notre projet de recherche.

2.3.1 An intelligent use of stemmer and morphology analysis for Arabic information retrieval

Ali Alnaied, Mosa Elbendak, et Abdullah Bulbul ont développé une approche novatrice appelée Arabic Morphology Information Retrieval (AMIR) pour extraire les racines des mots arabes en utilisant des règles linguistiques spécifiques. Leur objectif était d'améliorer les techniques de racination et d'analyse morphologique pour optimiser le traitement du langage naturel en arabe.

Ils se sont concentrés sur l'amélioration de la performance des systèmes de recherche d'informations arabes en résolvant les problèmes liés aux stemmers infixes. La méthode AMIR utilise des règles sur les relations entre les lettres arabes pour extraire les racines des mots. Comparée à des méthodes existantes comme LUCENE et FARASA, AMIR a démontré une précision supérieure lors des tests sur l'ensemble de données EveTAR, atteignant une précision moyenne de 0,34 % contre 0,27 % pour LUCENE, 0,28 % pour FARASA et 0,21 % pour la méthode non-stem.

Les travaux futurs pourraient inclure l'extension de l'ensemble de données, l'amélioration des performances du système et l'intégration de fonctionnalités avancées, ainsi que l'exploration des mots informels pour enrichir encore davantage les systèmes de recherche d'informations en arabe.

La figure suivante représente les valeurs TF. IDF calculées obtenues par AMIR, LUCENE et FARASA :

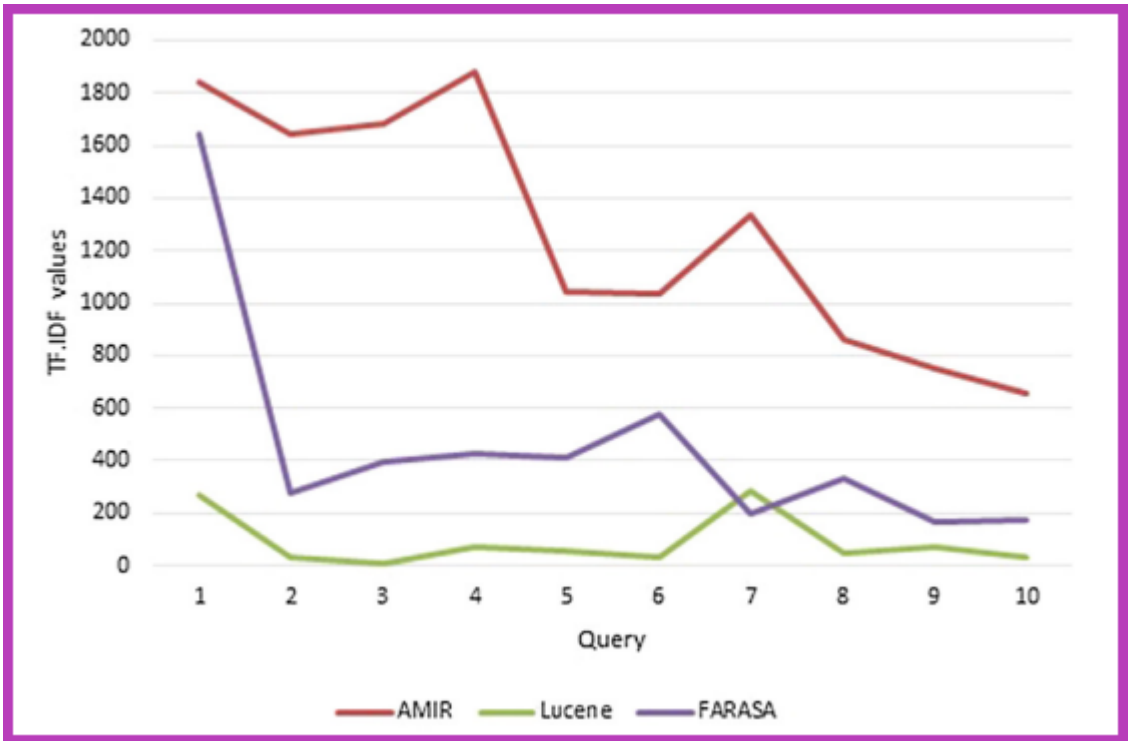


FIGURE 2.4 – Les valeurs TF.IDF calculées obtenues par AMIR, LUCENE et FARASA.

2.3.2 A new semantic similarity approach for improving the results of an Arabic search engine

Amine EL HADI, Youness MADANI, Rachid EL AYACHI et Mohamed ERRITALI ont cherché à améliorer la détection de similarité sémantique dans les moteurs de recherche arabes. Leur méthode impliquait l'utilisation du cadre MapReduce et de WordNet pour mesurer la similarité sémantique. Les résultats ont montré une technique efficace avec des performances élevées dans l'identification des documents pertinents. En intégrant la racinisation arabe, la traduction et un algorithme MapReduce pour le calcul de la similarité, leur approche a surpassé les méthodes existantes telles que Wu and Palmer et Leacock and Chodorow.

Les mesures de performance de l'étude ont révélé des améliorations significatives par rapport aux approches précédentes. Les taux de précision pour Wu and Palmer et Leacock and Chodorow étaient de 40 %, tandis que leur approche a atteint une précision de 60 %. En termes de rappel, Wu and Palmer ont obtenu 33,34 %, Leacock and Chodorow 50 %, et leur approche 50 %. De plus, les scores F1 pour Wu and Palmer et Leacock and Chodorow étaient respectivement de 36,75 % et 44,45 %, tandis que leur approche a excellé avec un score de 54,54 %.

À l'avenir, l'intégration de cette approche innovante dans les moteurs de recherche multilingues promet d'améliorer les performances globales et la précision de la détection de similarité sémantique. Les résultats de l'étude soulignent le potentiel de tirer parti de technologies avancées telles que MapReduce et WordNet pour optimiser les capacités des moteurs de recherche dans des contextes linguistiques diversifiés.

La figure suivante représente Le processus de la méthodologie pour calculer la mesure de similarité séman-tique :

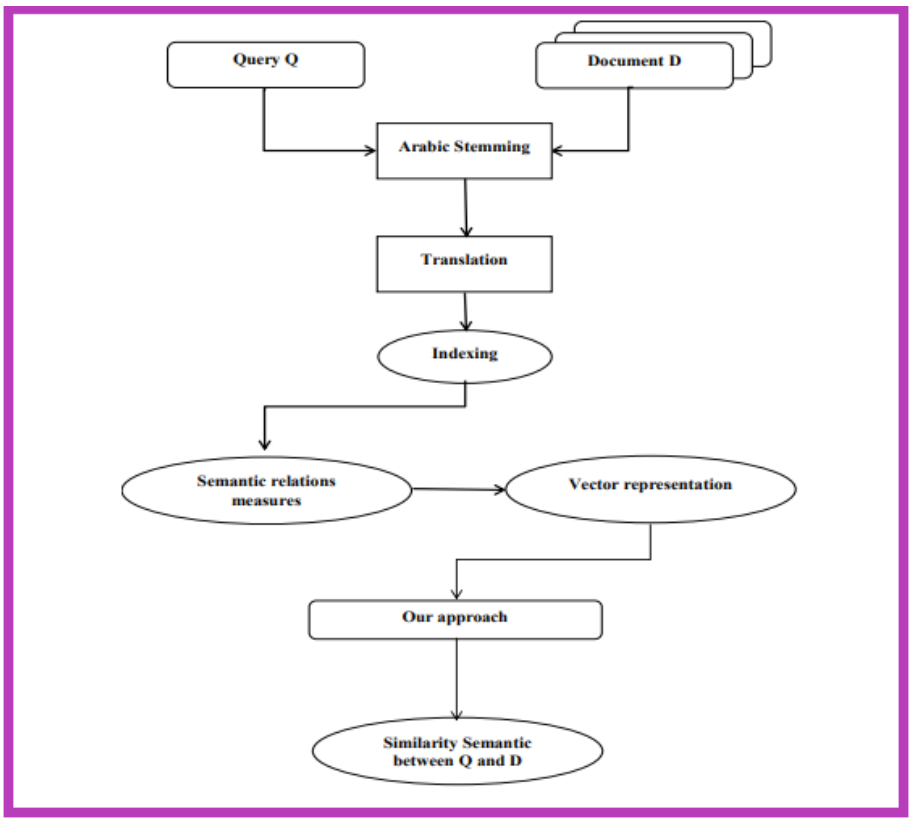


FIGURE 2.5 – Le processus de la méthodologie pour calculer la mesure de similarité sémantique.

2.3.3 Formal Concept Analysis for Arabic Web Search Results Clustering

Issam Sahmoudi et Abdelmonaime Lachkar ont présenté un système innovant de regroupement des résultats de recherche sur le Web en arabe, en utilisant l'Analyse de Concepts Formels (ACF). Ce système vise à améliorer la navigation des utilisateurs arabophones en catégorisant les extraits Web des moteurs de recherche en clusters étiquetés distincts, adaptés aux caractéristiques uniques de la langue arabe. Contrairement aux solutions conventionnelles conçues pour les langues européennes, cette méthode exploite l'ACF pour offrir une solution sur mesure aux utilisateurs arabes d'Internet.

L'efficacité du système a été démontrée par une enquête empirique utilisant des documents Web arabes provenant de la hiérarchie du projet Open Directory. Les résultats montrent une qualité de regroupement supérieure par rapport à des méthodes traditionnelles telles que le regroupement par arbre de suffixes et Lingo. L'évaluation utilise des mesures comme l'Information Mutuelle Normalisée (NMI) et l'Entropie Complémentaire Normalisée (NCE) pour juger de la qualité du regroupement. Le système suit un flux de travail structuré incluant

la collecte de données, le prétraitement de texte, la construction du treillis conceptuel, la sélection de clusters et la génération d'étiquettes basées sur des mots-clés.

Pour améliorer les performances du système, les auteurs suggèrent d'intégrer des ressources de connaissances externes telles que l'Arabic Word-Net et l'Arabic Wikipedia. Cela pourrait fournir un contexte supplémentaire et une meilleure compréhension sémantique, affinant ainsi le processus de regroupement et enrichissant l'expérience utilisateur.

La figure suivante représente Résultats de la requête (Éducation, التعليم) du cluster :

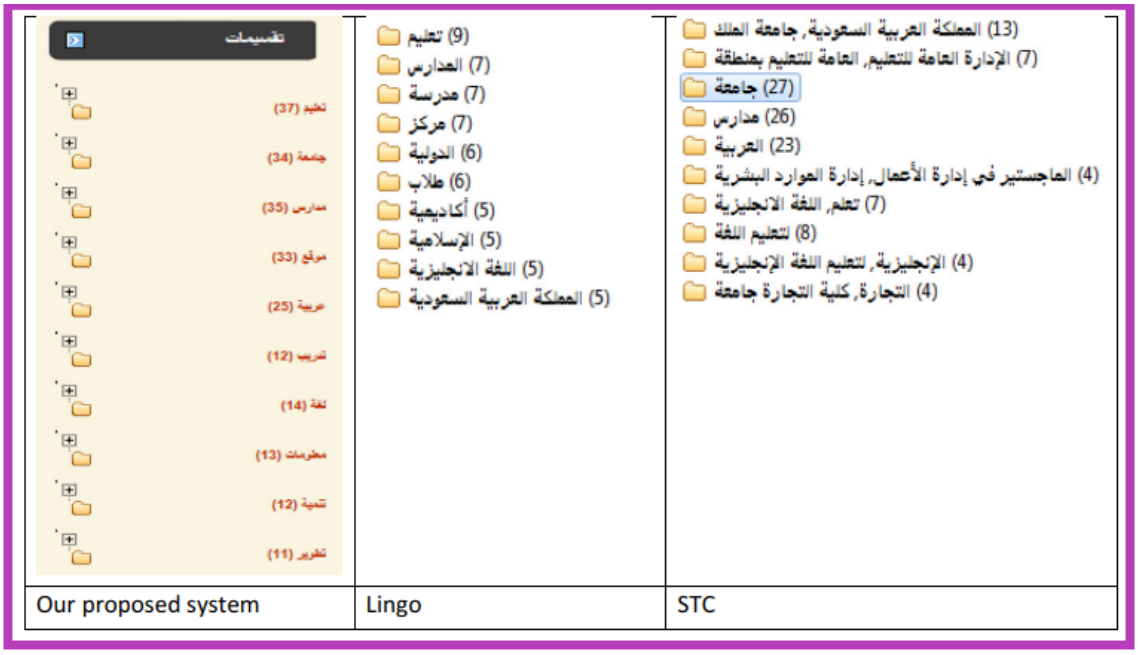


FIGURE 2.6 – Résultats de la requête (Éducation, التعليم) du cluster.

2.3.4 REDUCE : a semi-supervised scalable approach for REsultDUplication detection in Search Engines

REDUCE–une approche semi-supervisée et évolutive pour la détection de la duplication des résultats dans les moteurs de recherche" aborde le problème de la duplication des résultats dans les moteurs de recherche. Les auteurs mettent en avant les défis associés à cette duplication, tels que son impact sur la pertinence des résultats pour les utilisateurs et la nécessité de trouver des solutions efficaces pour améliorer l'expérience de recherche en ligne.

Pour résoudre cette problématique, les auteurs proposent une approche semi-supervisée et évolutive nommée REDUCE. Cette approche vise à repérer les résultats dupliqués en collectant des données à partir de diverses sources en ligne, en extrayant le contenu des pages Web et en comparant les similitudes entre les résultats. En utilisant des techniques de prétraitement et de classification, l'approche REDUCE permet d'identifier et de supprimer les résultats dupliqués, offrant ainsi aux utilisateurs des résultats de recherche plus pertinents et uniques.

En mettant en pratique cette approche, les auteurs ont observé une amélioration significative de la qualité des résultats de recherche, avec une réduction notable de la redondance des informations. L'article souligne également l'importance de l'évolutivité des modèles d'apprentissage automatique dans les systèmes de recherche en ligne, en insistant sur la nécessité de traiter efficacement de grandes quantités de données pour garantir des performances optimales.

En conclusion, l'approche REDUCE représente une solution prometteuse pour détecter et réduire la duplication des résultats dans les moteurs de recherche, ouvrant la voie à des recherches futures visant à améliorer davantage l'efficacité et la précision des algorithmes de détection de duplication.
La figure suivante représente l'architecture du système de "REDUCE" :

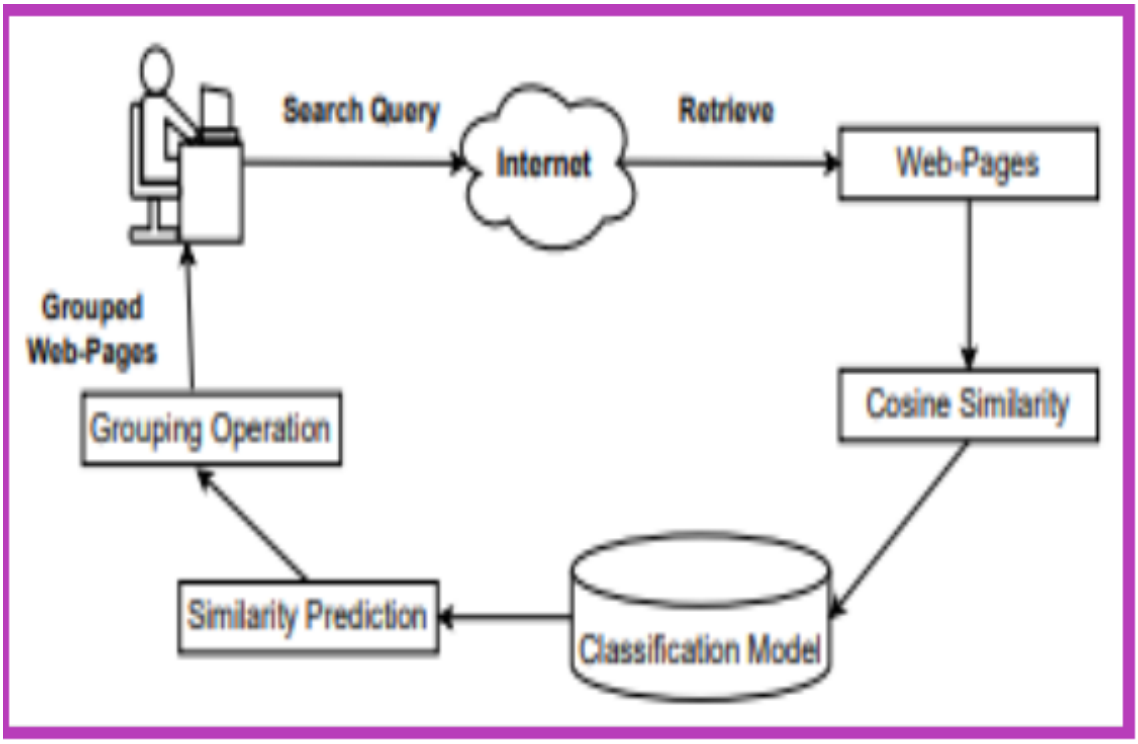


FIGURE 2.7 – Architecture du Système de "REDUCE".

2.3.5 IBRI-CASONTO : Ontology-based semantic search engine

L'article "IBRI-CASONTO : Moteur de recherche sémantique basé sur l'ontologie" présente le développement d'un moteur de recherche visant à améliorer l'extraction d'informations à partir de référentiels de données. Ce moteur prend en charge les langues arabe et anglaise et utilise des techniques de recherche basées sur les mots-clés et la sémantique. Il exploite des technologies telles que le Framework de Description des Ressources (RDF) et les graphes ontologiques pour améliorer les capacités de recherche.

Une des caractéristiques clés d'IBRI-CASONTO est l'utilisation du raisonneur HerMiT dans l'environnement Protégé. Le raisonneur aide à inférer des sous-classes, des classes équivalentes, des classes disjointes, des propriétés d'objets, des propriétés de données et d'autres informations liées à l'ontologie. En exploitant l'ontologie et les technologies sémantiques, IBRI-CASONTO vise à fournir des résultats de recherche plus précis et pertinents par rapport aux moteurs de recherche traditionnels basés sur les mots-clés.

Le document aborde les défis posés par la complexité de la langue arabe, qui présente des aspects morphologiques, grammaticaux et sémantiques complexes. Malgré ces défis, le développement de systèmes de recherche sémantique en arabe est crucial pour améliorer l'extraction d'informations dans le contenu en arabe.

IBRI-CASONTO est spécifiquement adapté aux collèges de sciences appliquées à Oman, en se concentrant sur les départements académiques, le personnel et les informations sur les étudiants. Le moteur vise à offrir une expérience de recherche plus efficace et précise aux utilisateurs dans ce domaine. Le document inclut également des métriques de performance du moteur IBRI-CASONTO, démontrant sa capacité à récupérer des informations pertinentes avec précision (Précision : 52,49

En conclusion, l'article souligne l'importance des moteurs de recherche sémantiques basés sur l'ontologie comme IBRI-CASONTO pour surmonter les limitations des technologies de recherche traditionnelles basées sur les mots-clés. La conception du moteur, basée sur l'ontologie RDF et les graphes ontologiques, offre une approche prometteuse pour améliorer l'extraction d'informations dans des domaines spécifiques tels que les collèges de sciences appliquées.
La figure suivante représente Structure de "IBRI-CASONTO" :

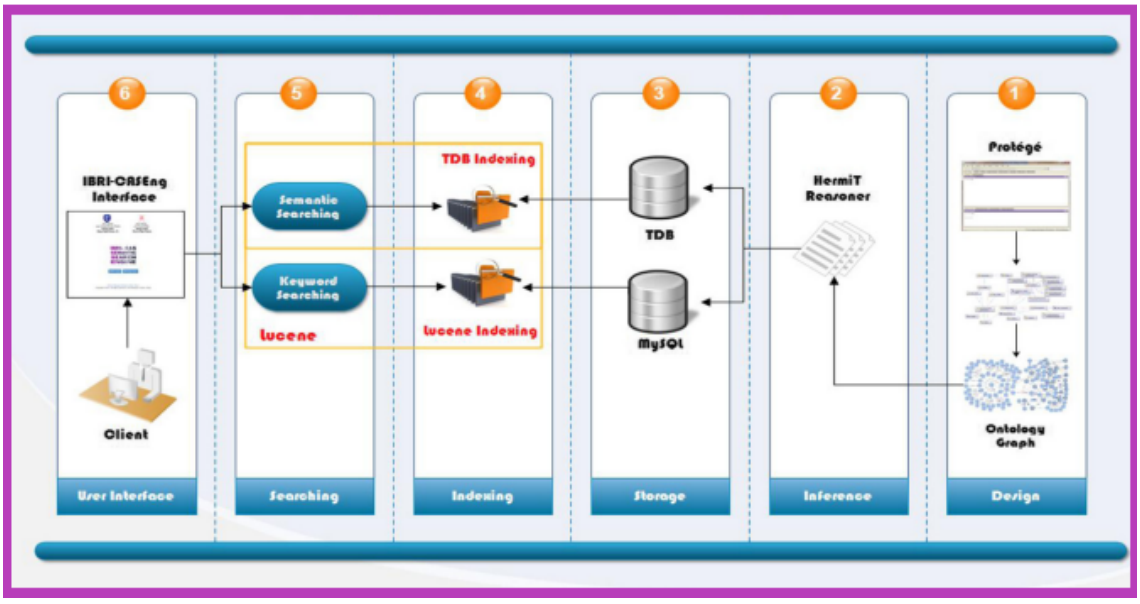


FIGURE 2.8 – Structure de "IBRI-CASONTO".

2.3.6 Embedding arabic questions by feature-level fusion of wordrepresentations for questions classification : It is worth doing ?

L'article "Embedding Arabic questions by feature-level fusion of word representations for questions classification" se concentre sur l'amélioration de la classification des questions en arabe en utilisant des méthodes de fusion au niveau des caractéristiques des représentations de mots. Les auteurs explorent l'utilisation de différentes représentations de mots, notamment BERT, ELMo et les embeddings enrichis par des informations sur les sous-mots, tels que W2V. Trois contributions principales sont mises en avant dans l'étude :

1. Gestion des mots hors vocabulaire : La méthode proposée gère les mots hors vocabulaire, ce qui est crucial pour la classification précise des questions en arabe.
2. Utilisation de la représentation BERT : Les auteurs appliquent la représentation BERT pour extraire les caractéristiques les plus précieuses des mots et construire une meilleure représentation des questions.
3. Étude de l'impact de la fusion de différentes représentations de mots : L'article analyse l'effet de la fusion de divers embeddings de mots sur la classification des questions en arabe.

Pour évaluer les modèles proposés, une validation croisée stratifiée à 5 plis est réalisée sur un ensemble de données contenant 3173 questions étiquetées selon les taxonomies arabes et de Li & Roth. Les résultats expérimentaux montrent que tous les modèles surpassent les travaux précédents liés à la tâche de classification des questions en arabe. En utilisant la taxonomie arabe, le modèle basé sur AraBERT atteint une précision maximale de 94,20%, tandis que pour la taxonomie de Li & Roth, le modèle basé sur la concaténation d'AraBERT et de W2V obtient la précision globale la plus élevée de 93,51%.

En conclusion, l'article met en lumière l'importance des représentations de mots contextualisées et de la fusion de ces représentations au niveau des caractéristiques pour améliorer la classification des questions en arabe, offrant ainsi des perspectives prometteuses pour les systèmes de question-réponse et le traitement automatique du langage naturel en arabe.
La figure suivante représente le diagramme de flux d'un système de réponse aux questions :

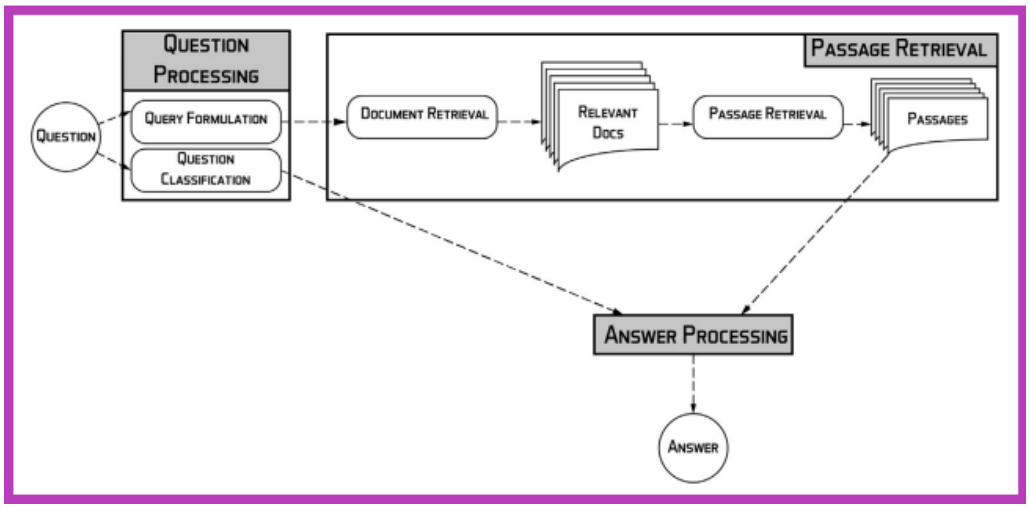


FIGURE 2.9 – Diagramme de flux d'un système de réponse aux questions .

Nous présentons dans le tableau suivants quelque travaux connexe avec notre PFE :

Titre	Réferences	Université	La problématique abordée	Résultat
An intelligent use of stemmer and morphology analysis for Arabic information retrieval.	[Alnaied et al., 2020]	Northumbria University, UK	L'extraction des racines dans le contexte de la recherche d'information en arabe.	une précision moyenne de 0.34%, tandis que LUCENE 0.27%, FARASA 0.28% et le système sans stemmer 0.21% .
REDUCE : a semi-supervised scalable approach for REsultDUplication detection in Search Engines	[Hazimeh et al., 2021]	Lebanese University, Beirut, Lebanon	La duplication des résultats dans les moteurs de recherche.	- Accuracy 0.967 - Précision 0.5413 - Rappel 0.8
IBRI-CASONTO : Ontology-based semantic search engine	[Sayed and Al Muqrishi, 2017]	Minia University, Egypt	Développer un moteur de recherche sémantique performant en arabe.	- Précision 52.49% - Rappel 75% - L'exactitude 45%.
Embedding arabic questions by feature-level fusion of word representations for questions classification : It is worth doing ?	[Hamza et al., 2022]	IbnTofail University, Kenitra, Morocco	La classification critique des questions en arabe.	- F1 score 94.19%.
A new semantic similarity approach for improving the results of an Arabic search engine	[Amine et al., 2019]	Sultan Moulay Slimane University ,Marocco	Améliorer la pertinence des résultats en arabe avec une similarité sémantique basée sur MapReduce et WordNet.	-Précision 60% - Rappel 50% - F1-score 54.54%.

Formal Concept Analysis for Arabic Web Search Results Clustering	[Sahmoudi and Lachkar, 2016]	USMBA, Fez, Morocco	Regroupement optimisé des résultats de recherche web en arabe via l'Analyse de Concept Formel (FCA).	- Précision 68% - Rappel 0.41 % - F-Mesure 0.51 %
Semantic Annotation of Arabic Web Documents using Deep Learning	[Albukhitan et al., 2018]	King Fahd University,Saudi Arabia.	Utilisation efficace des embeddings de mots pour l'annotation arabe de documents Web, en mettant l'accent sur la santé et la nutrition.	- Précision de 86.6% - Rappel de 82.0% - F-Mesure de 84.2 %
Semantic Annotation of Arabic Web Resources using Semantic Web Services	[Albukhitan et al., 2016]	King Fahd University,Saudi Arabia.	Développer des services Web sémantiques pour l'annotation de documents arabes.	- Précision 84.8% - Rappel 68.9%

TABLE 2.1 – Travaux connexes à notre PFE.

3 Meilleures moteurs de recherche au monde

Selon les données de StatCounter, Google reste le moteur de recherche numéro 1 dans le monde en janvier 2024, avec 91% de part de marché. Cela le place loin devant Microsoft Bing qui détient 3% du marché. Voici les statistiques des parts de marché de Google Search dans le monde :

- Google à 91,47%
- Bing à 3,43%
- YANDEX à 1,78%
- Yahoo ! à 1,1%

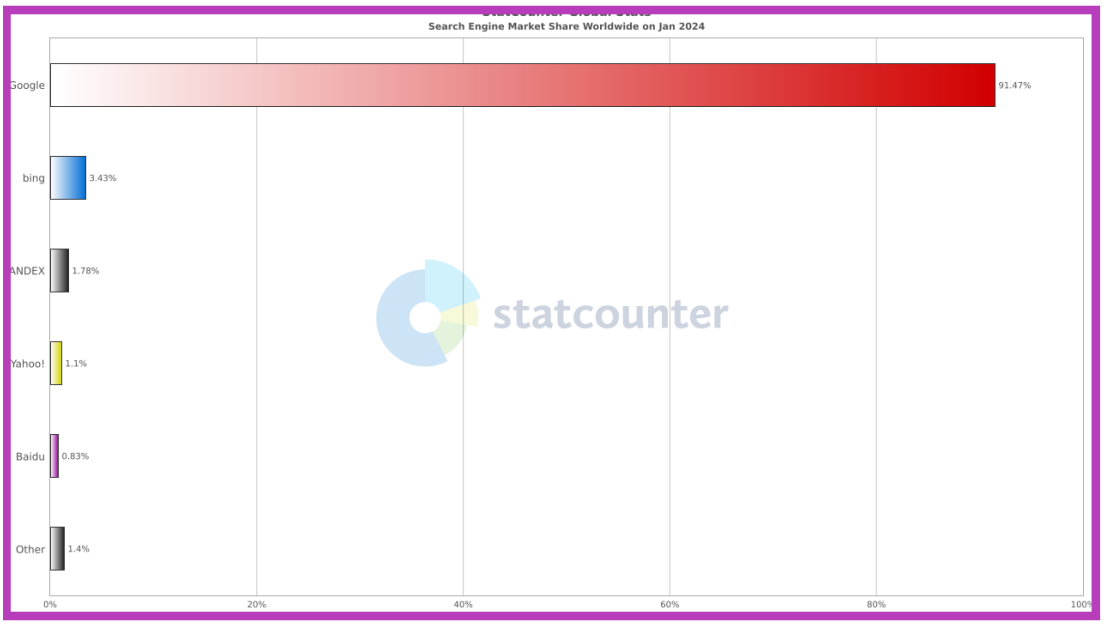


FIGURE 2.10 – Part de marché des moteurs de recherche dans le monde en janvier 2024.

Voici quelques informations sur les moteurs de recherche les plus réputés à l'échelle mondiale :

3.1 google :

Créé en 1995 à l'Université Stanford par Larry Page et Sergey Brin, Google est devenu le leader incontesté de la recherche en ligne, contrôlant environ 92 % du marché mondial. À ses débuts, il portait le nom de "Backrub", mais il a depuis évolué pour proposer une multitude de services gratuits et payants, allant de la recherche sur le web à l'hébergement de sites, en passant par les boîtes de messagerie et la messagerie instantanée.

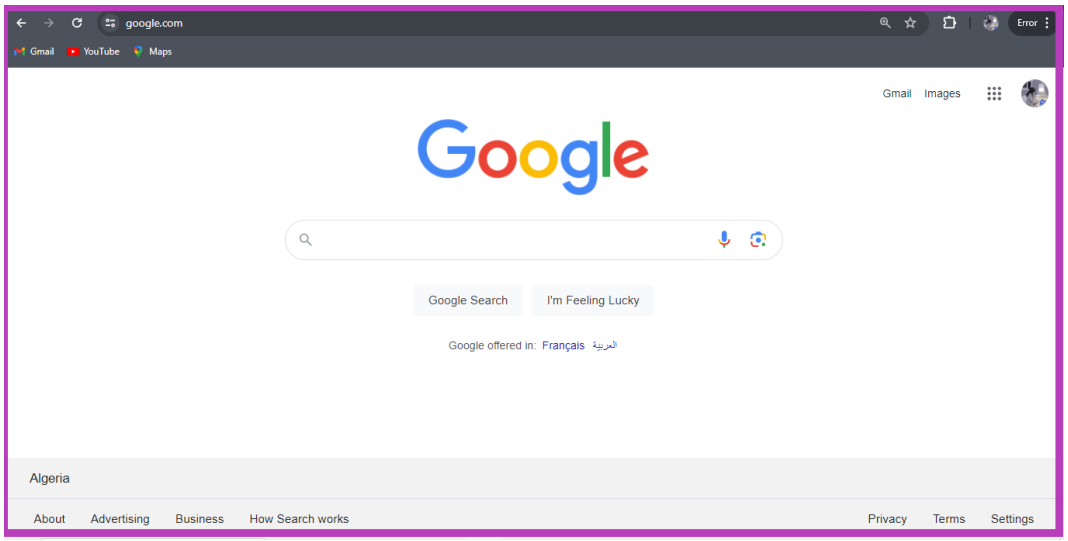


FIGURE 2.11 – Moteur de recherche en ligne google.

3.2 Microsoft Bing :

Créé par Microsoft et introduit en 2009 pour concurrencer Google, Bing est devenu une présence significative dans le domaine des moteurs de recherche. En partenariat avec Yahoo ! depuis cette période, Bing est aujourd'hui le moteur de recherche par défaut du navigateur Microsoft Edge. Bien que sa part de marché soit inférieure à celle de Google, Bing se démarque grâce à ses fonctionnalités exclusives.

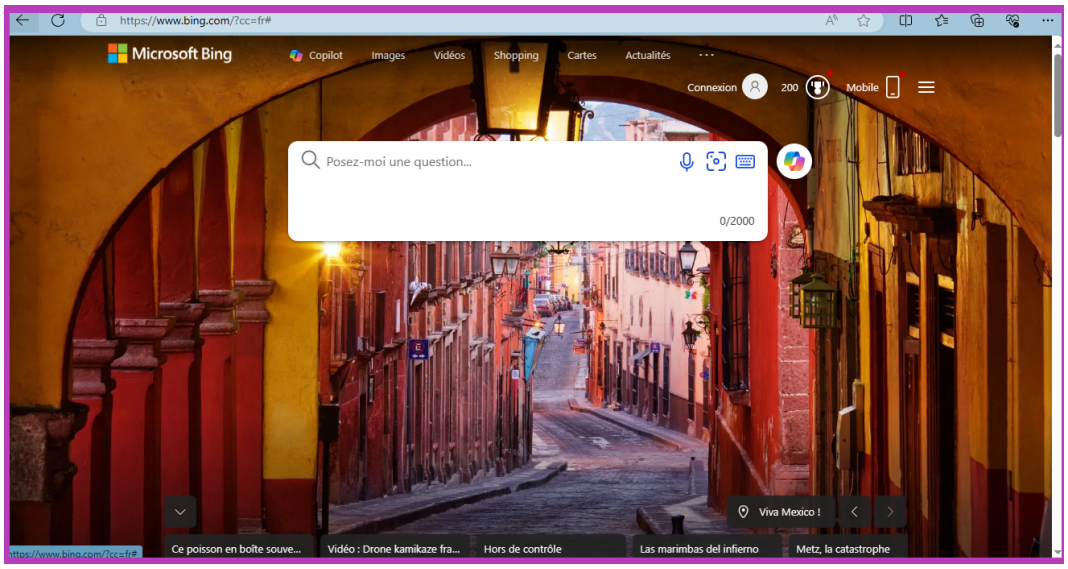


FIGURE 2.12 – Moteur de recherche en ligne Bing.

3.3 YANDEX :

YANDEX est une entreprise technologique russe qui développe des produits et services intelligents alimentés par l'apprentissage automatique. Depuis 1997, Yandex propose des services de recherche et d'information de classe mondiale, pertinents localement. Son moteur de recherche, basé à Moscou, est devenu le plus important et le plus utilisé en Russie et dans les pays de la CEI. Il offre une variété de services similaires à ceux de Google, tels que la recherche, la messagerie électronique et bien plus encore.

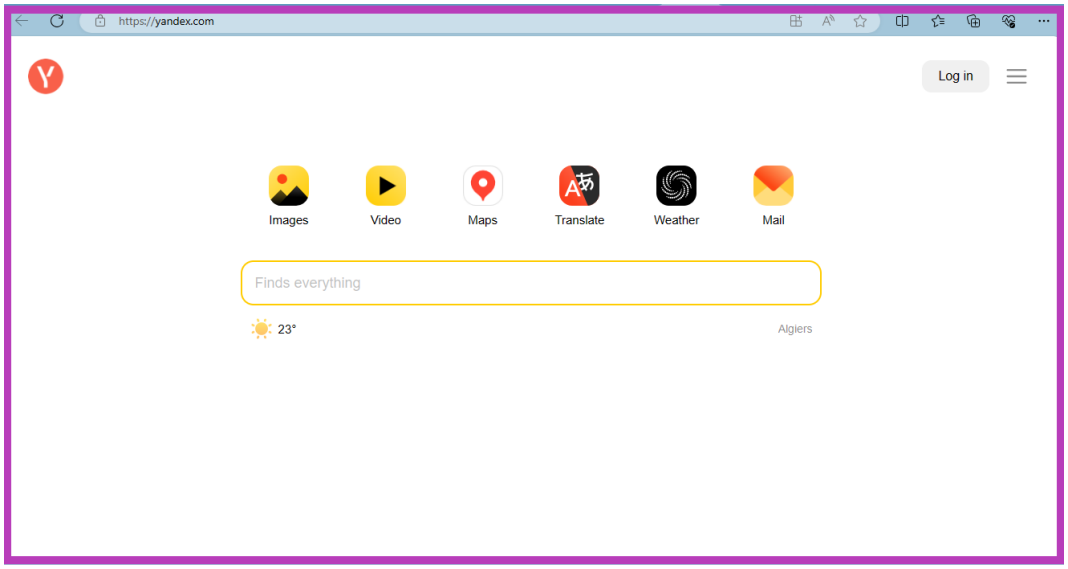


FIGURE 2.13 – Moteur de recherche en ligne YANDEX.

3.4 Yahoo ! :

Yahoo Search, plus connu sous le nom de Yahoo !, est l'un des plus anciens moteurs de recherche. Il est aujourd'hui le quatrième moteur de recherche le plus utilisé dans le monde. Il base son fonctionnement sur un aspect communautaire en donnant aux internautes la possibilité de sélectionner et d'annoter les pages et sites web qui leur paraissent intéressants pour les partager avec d'autres. Le moteur de recherche de Yahoo ! est totalement gratuit. Vous pouvez l'utiliser via n'importe quel navigateur web ou via application mobile.

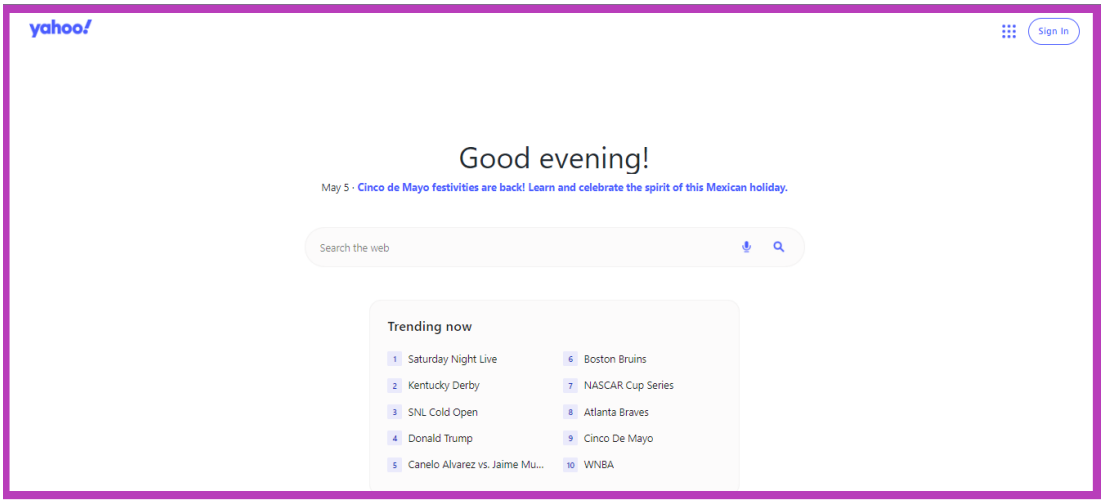


FIGURE 2.14 – Moteur de recherche en ligne Yahoo !.

4 conclusion

Dans ce chapitre, nous avons passé en revue plusieurs travaux en lien avec notre projet de fin d'études. Nous nous sommes concentrés sur deux initiatives majeures dans le domaine de la recherche d'information : le Text Retrieval Conference (TREC) et le Forum de laboratoire et d'évaluation (CLEF). À travers ces initiatives, nous avons exploré les définitions, objectifs et évolutions clés du domaine de la recherche d'information.

Nous avons également sélectionné six travaux pertinents pour notre projet, couvrant divers aspects tels que la recherche en langue arabe et l'analyse de moteurs de recherche. Nous avons présenté les concepts fondamentaux et les définitions de certains moteurs de recherche populaires comme Google, Bing, Yandex et Yahoo.

Donc,on peut dire que ce chapitre fourni une analyse précise et complète des travaux antérieurs qui constituent une base de connaissances importante pour notre projet. En passant en revue et en évaluant ces travaux.

Chapitre III

L'approche proposée

1 Introduction

Au cœur de toute expérience il est crucial de garantir la pertinence d'un document par rapport à une requête afin d'obtenir les meilleurs résultats pour l'utilisateur.

Dans le cadre de notre projet de fin d'études, notre ambition est de créer un moteur de recherche touristique en arabe spécialement dédié à l'Algérie.

Cependant, au lieu d'opter pour une approche linéaire, nous avons délibérément choisi une méthode stratégique et itérative que nous détaillerons dans ce chapitre.

Tout d'abord, Nous avons concentré surla récupération de la pertinence via l'utilisation de moteurs de recherche open source.

Ensuite, nous procédons à l'exécution des requêtes en les soumettant à ces moteurs de recherche sélectionnés. Par la suite, nous recueillons les résultats retournés par ces moteurs de recherche pour chaque requête.

Enfin ,Afin d'assurer la qualité et la fiabilité de nos résultats, nous avons extrait une matrice finale qui regroupe les résultats des trois moteurs de recherche. Cette matrice nous a permis de collecter des informations pertinentes et de les intégrer dans notre site web dédié.

2 Description de l'approche

Notre travail est devisé en trois étapes principaux :

I. Construction du dataset et établissement de la liste des requêtes

- Création du dataset : Rassemblement des données pertinentes dans un fichier Excel, comprenant toutes les informations nécessaires pour l'analyse ultérieure.
- Établissement de la liste des requêtes : Identification et formulation des questions ou des requêtes pertinentes qui seront utilisées pour évaluer la pertinence des résultats.

II. Récupération de la pertinence à travers des moteurs de recherche "Open Source"

- Choix des moteurs de recherche "Open Source" : Sélection des moteurs de recherche disponibles en open source qui seront utilisés pour récupérer les résultats pertinents.
- Exécution des requêtes : Soumission des requêtes établies à l'étape précédente aux moteurs de recherche sélectionnés.
- Extraction des résultats : Collecte des résultats retournés par les moteurs de recherche pour chaque requête.

III. Regroupement des Résultats des Trois Moteurs de Recherche

- Création de la Matrice Finale : À l'aide d'un code Python, une matrice finale est créée en consolidant les résultats des trois moteurs de recherche.

- Intégration dans le Site Web : La matrice finale a été utilisée dans notre site web dédié, permettant aux utilisateurs d'accéder facilement aux informations pertinentes recueillies à partir des moteurs de recherche.

2.1 Plan global du travail

La figure ci-dessous présente de manière générale les étapes principales de l'approche proposée pour notre projet.

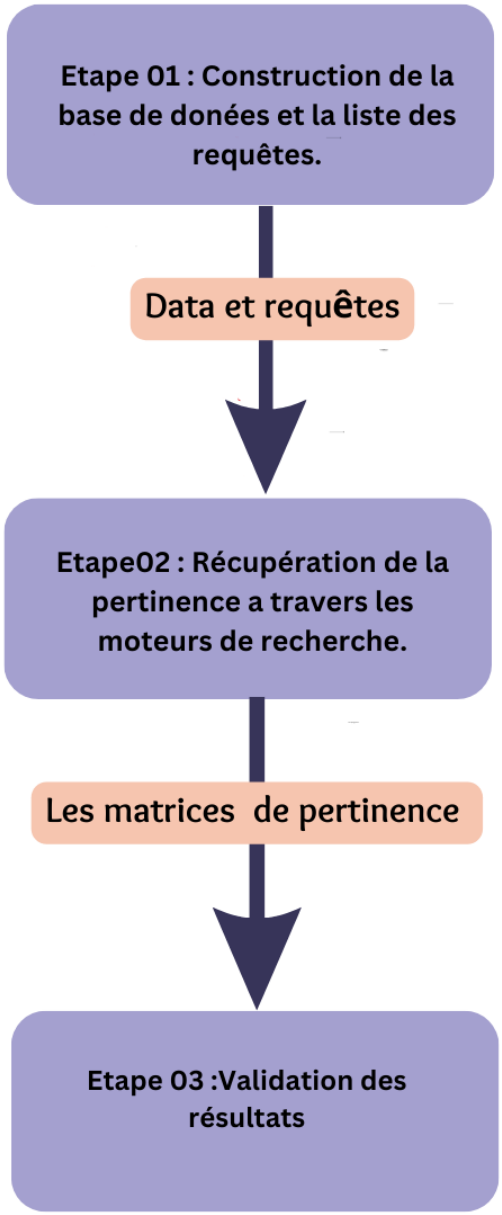


FIGURE 3.1 – Plan global de l'approche proposée.

2.2 Plan détaillé du travail

Les étapes principales de l'approche proposée pour notre projet sont expliquées en détail dans la figure suivante :

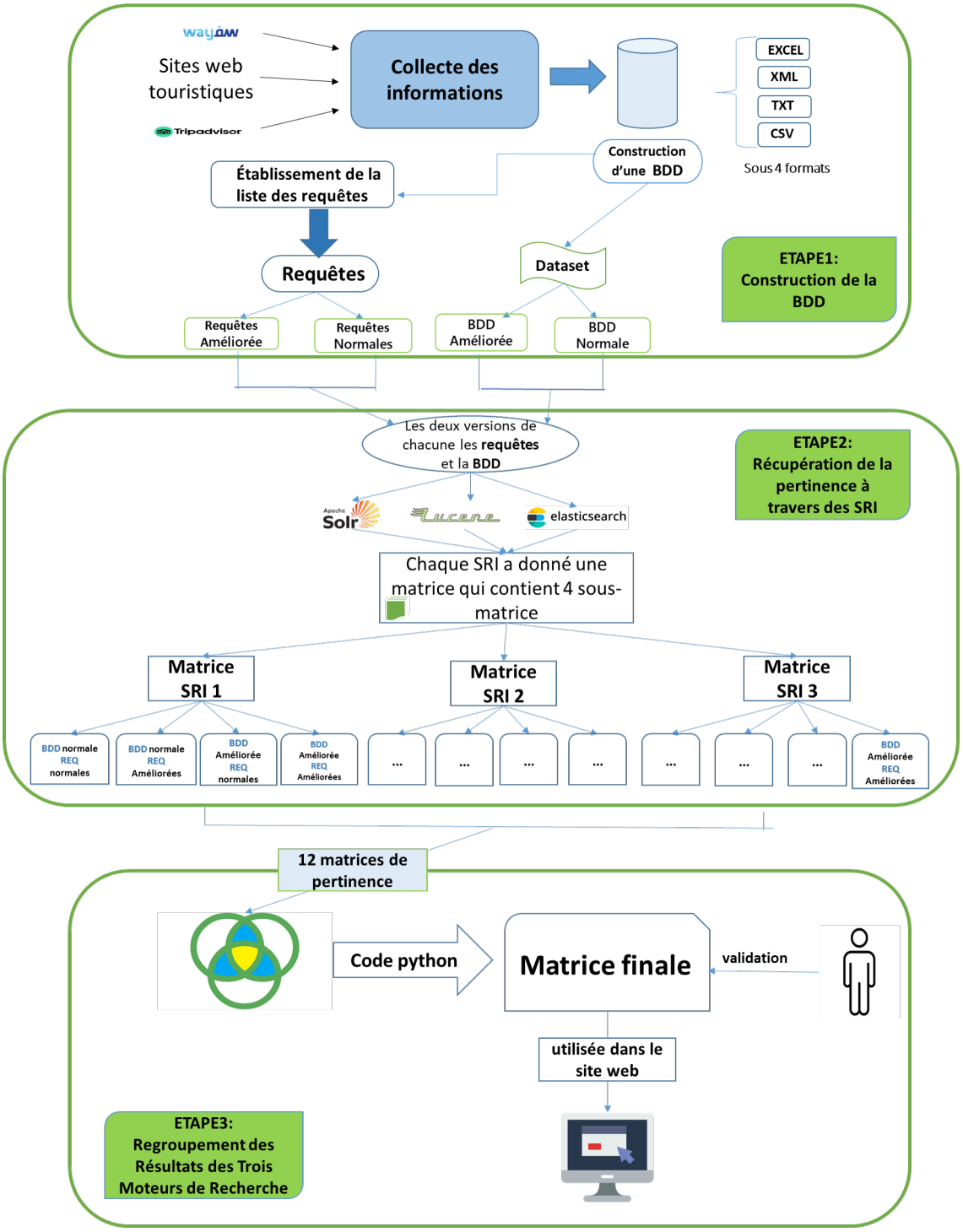


FIGURE 3.2 – Plan détaillé détaillant les diverses étapes de notre approche.

Comme nous l'avons déjà souligné, notre projet se compose de trois étapes principales. Maintenant, nous allons les détailler plus en détail.

3 Construction du dataset et établissement de la liste des requêtes

3.1 Collecte d'informations

Le processus de collecte d'informations a été le plus difficile, car aucune base de données précédente ne correspondait exactement à nos besoins. La meilleure solution consistait à accéder à Internet et à rechercher manuellement les informations nécessaires. Nous avons consulté divers sites touristiques algériens tels que le Journal national du tourisme et du voyage, ainsi que d'autres sites incluant l'Algérie comme destination touristique, notamment le populaire site « Safarway ». Différentes sources en ligne ont été utilisées pour recueillir les informations après validation.

Les informations recueillies étaient les suivantes :
- Le nom de chaque lieu
- L'évaluation des visiteurs
- La classification (historique, naturel, plage, etc.)
- La wilaya
- Le matricule de la wilaya
- Sa localisation sur Maps pour faciliter l'accès
- Le contact du lieu
- La direction : (Est, Ouest, Centre, Désert)

Une base documentaire a également été créée pour chaque lieu pour lequel des renseignements ont été recueillis, de sorte que 355 documents représentant un type de données seront ajoutés à notre base de données, ce qui constitue la description.

3.2 Construction de la base

3.2.1 Base de données normale

Après avoir rassemblé les informations nécessaires, nous avons élaboré une base de données sous Excel. L'information était répartie en 10 champs. Ils sont représentés dans le tableau suivant :

Champ	Type
Id	Entier
Nom du lieu	Chaîne de caractères
Ville	Chaîne de caractères
Description	Chaîne de caractères
Type	Chaîne de caractères
Note	Réel
Emplacement	Réel
Image	PNG/JPG
Contact	Chaîne de caractères/entier
Matricule	Entier
La zone	Chaîne de caractères

TABLE 3.1 – Description des champs de la base de données.

Les figure suivante représentent le pourcentage des documents par rapport au champ "TYPE" et le champs "Zone (direction de la wilaya)" :

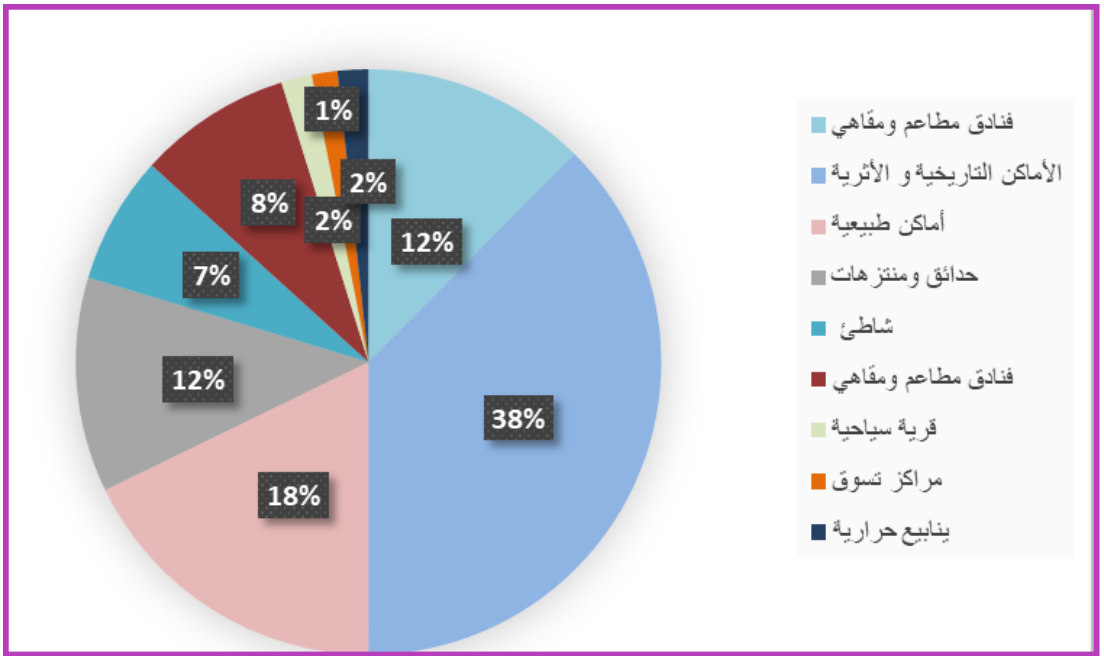


FIGURE 3.3 – Pourcentage des documents par apport au champ Type.

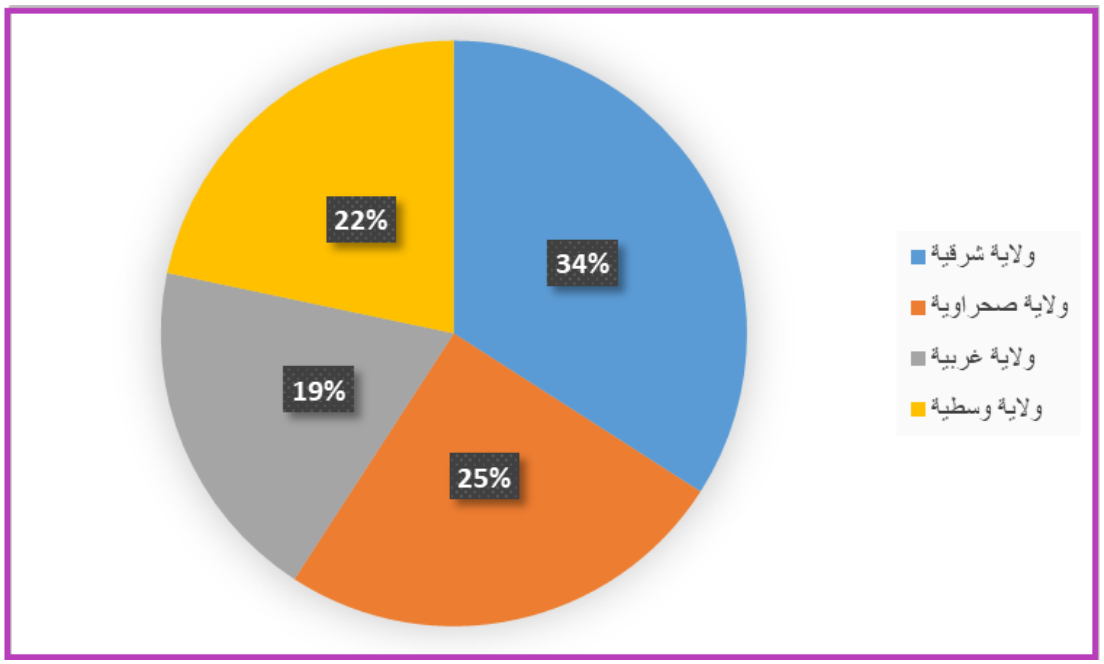


FIGURE 3.4 – Pourcentage des documents par apport au champ Zone.

3.2.2 Différents types de la base de données

La diversité et la complexité des besoins actuels nous ont conduit à organiser notre base de données en différents types. Il existe une grande variété d'outils et de méthodes dans chaque domaine, y compris la programmation. Nous pouvons mieux saisir cette richesse en utilisant cinq types distincts pour notre base de données et offrir des analyses plus précises et adaptées aux besoins variés des utilisateurs.

— **Excel :** Le format est sélectionné en raison de sa facilité d'utilisation et de la possibilité de structurer les données en tableaux qui peuvent être utilisés et manipulés avec des langages comme Python.

— **XML :** Ce format est standard dans de nombreux systèmes modernes en raison de sa solidité et de sa capacité à hiérarchiser des données complexes.

— **TXT :** Ce format texte simple est idéal pour les besoins d'indexation des moteurs de recherche. Il permet une recherche rapide et efficace, notamment avec des outils tels que Lucene.

— **CSV :** Ce format est très utilisé dans les applications web et les transferts de données car il est simple à utiliser et compatible avec de nombreux logiciels d'analyse de données.

— **SQL :** Choisi pour gérer les bases de données relationnelles selon la norme mondiale et pour gérer des requêtes complexes et garantir l'intégrité des données.

3.2.3 Base de données améliorée

Puisque les moteurs de recherche se concentrent généralement sur les documents .txt, nous avons amélioré les descriptions pour qu'elles soient meilleures et contiennent davantage de mots clés permettant aux moteurs de recherche de générer des résultats plus pertinents que leurs prédécesseurs non améliorés.

L'image suivante représente le code utilisé dans le processus d'amélioration :

```
import os
import nltk
from nltk.corpus import wordnet
import csv

# Download the Arabic WordNet resource
nltk.download('omw-1.4')

def find_synonyms(word):
    synonyms = []
    try:
        for synset in wordnet.synsets(word, lang='arb'):
            for lemma in synset.lemmas(lang='arb'):
                synonyms.append(lemma.name())
    except Exception as e:
        print("An error occurred:", e)
    return synonyms

def find_synonyms_in_directory(input_dir, output_csv):
    with open(output_csv, 'w', newline='', encoding='utf-8-sig') as outfile:
        csv_writer = csv.writer(outfile)
        for filename in os.listdir(input_dir):
            if filename.endswith('.txt'):
                file_path = os.path.join(input_dir, filename)
                with open(file_path, 'r', encoding='utf-8') as infile:
                    text = infile.read()
                    words = text.split()
                    word_synonyms_map = {}
                    for word in words:
                        word = word.strip()  # Remove leading/trailing spaces
                        synonyms = find_synonyms(word)
                        if synonyms:
                            word_synonyms_map[word] = ', '.join(synonyms)
                        else:
                            word_synonyms_map[word] = word  # Use the original word if no synonyms
found
                    csv_writer.writerow([filename] + [word_synonyms_map[word] for word in words])


find_synonyms_in_directory('D:/description', 'descriptionameliorer.csv')
```

FIGURE 3.5 – Code d'amélioration de la BDD.

3.3 Construction des requêtes

3.3.1 Requêtes normales

De notre base de données nous avons extrait 60 requêtes sous forme des questions, Les requêtes ont été faites à partir de la question : "Qu'est-ce que quelqu'un pourrait se demander qui veut voyager à une destination particulière ?" Les réponses étaient nombreuses à titre d'exemple : "Quelles sont les meilleures plages que je peux visiter à l'est ?".
Le tableau ci-dessous présente un échantillon de la base des requêtes (Voir l'annexe pour le tableau complet)

Id	Query
Q01	الأماكن و الأنشطة المناسبة للعائلات في ولاية تلمسان؟
Q02	أماكن تستحق الزيارة في ولاية تيميمون
Q03	ماهي المعالم السياحية البارزة في ولاية جيجل ؟
.	.
.	.
.	.
Q60	اين يمكنني تناول وجبة تقليدية في ولاية بسكرة ؟

TABLE 3.2 – Tableau des requêtes.

3.3.2 Requêtes améliorées : Optimisation des Recherches Touristiques

Nous avons amélioré l'efficacité de nos requêtes en utilisant des techniques de traitement du langage naturel (NLP) et des ressources lexicales adaptées à l'arabe. À l'aide de codes Python, nous avons généré des synsets pour chaque mot-clé de nos requêtes, permettant ainsi d'élargir le champ lexical et d'améliorer la pertinence des résultats de recherche. Ensuite, nous avons manuellement filtré les mots générés pour chaque requête, en supprimant ceux qui ne semblaient pas pertinents. Cela a permis de créer des requêtes plus riches et mieux adaptées aux moteurs de recherche, optimisant ainsi la pertinence des résultats.

```python
# code queries amelioree
import nltk
from nltk.corpus import wordnet
import csv

# Download the Arabic WordNet resource
nltk.download('omw-1.4')

def find_synsets(word):
    synsets = []
    try:
        for synset in wordnet.synsets(word, lang='arb'):
            synsets.append(synset.name())
    except Exception as e:
        print("An error occurred:", e)
    return synsets

def find_synsets_in_csv(input_csv, output_csv):
    with open(input_csv, 'r', newline='', encoding='utf-8') as infile, \
         open(output_csv, 'w', newline='', encoding='utf-8-sig') as outfile:
        csv_reader = csv.reader(infile)
        csv_writer = csv.writer(outfile)
        for row in csv_reader:
            synsets_columns = []
            for column in row:
                synsets = [find_synsets(word) for word in column.split()] if column else []
                synsets_columns.append(synsets)
            csv_writer.writerow(synsets_columns)

find_synsets_in_csv('C:/Users/loubna achr/processedq12.csv', 'qameliorer12.csv')
```

FIGURE 3.6 – Code requêtes améliorées.

3.4 Extraction manuelle des documents pertinents :

Nous avons procédé à l'extraction manuelle des documents pertinents pour chaque requête formulée. Cette extraction a été systématiquement organisée et documentée dans un fichier Excel. Chaque requête est associée à une liste spécifique de documents pertinents, soigneusement vérifiée et validée pour assurer la fiabilité des données. Ces données serviront par la suite à l'évaluation des performances de différents moteurs de recherche, en calculant les métriques de précision, de rappel et la mesure F. Cette approche nous permet d'établir une comparaison rigoureuse et objective des moteurs de recherche en termes de pertinence des résultats fournis.

4 Récupération de la pertinence à travers des moteurs de recherche «Open Source »

Retrouver la pertinence a travers les moteurs de recherche est un processus complexe visant a fournir des resultats appropries pour les requetes des utilisateurs. Pour cela, on a étudié la pertinence avec les données pré-construites pour les requêtes que nous les avons extraites. Et parce que l'homme ne peut pas lire tout ce qu'il reçoit, nous avons utilisé seulement 10 premiers résultats,et nous pouvons modifier.

4.1 Validation par les trois moteurs de recherche

Les moteurs de recherche Open Source offrent aux utilisateurs la possibilité d'explorer et d'expérimenter avec les algorithmes sous-jacents qui guident la recherche sur Internet. Bien que ces tests puissent être limités en fonction des ressources disponibles, ils fournissent une fenêtre précieuse pour les chercheurs et les étudiants afin d'évaluer et de juger leur travail.En comparant les résultats obtenus par ces moteurs avec ceux des géants du web, pour affiner les stratégies de recherche et pour s'assurer que les informations collectées sont fiables et pertinentes.
Les trois moteurs de recherche choisis sont :

- Lucene
- Elasticsearch
- Solr

Le tableau suivant explique quelques informations sur les trois moteurs de recherche qui nous avons utiliser :

Le moteur	**Lucene**	**Elasticsearch**	**Solr**
Anée de création	2000	2010	2006
Programmation	Java, Ruby, le Perl, le PHP, le C++	Java	Java
fondation	Apache	Apache	Apache
Base	Bibliothèque de recherche plein texte.	Basé sur Lucene	Distribué basé sur Lucene

TABLE 3.3 – Comparaison entre les trois moteurs de recherche utilisé.

Nous avons effectué une évaluation de trois moteurs de recherche : Lucene, Solr et Elasticsearch, et testé quatre parties(Table 3.4) distinctes de données et de requêtes : (1) Bdd normales et requêtes normales, (2) Bdd améliorées et requêtes normales, (3) Bdd normales et requêtes améliorées, et (4) Bdd améliorées et requêtes améliorées. Pour chaque partie, nous avons exécuté 60 requêtes distinctes, en prenant en compte uniquement les 10 premiers résultats retournés par chaque moteur de recherche. Cela nous a permis de créer des matrices de pertinence où chaque cellule indique la pertinence d'un résultat pour une requête spécifique. Et donc, pour chaque moteur de recherche et chaque partie, nous avons une matrice de pertinence. Au total, nous avons donc 12 matrices de pertinence (3 moteurs $\times$ 4 parties).

Partie	**BDD**	**Requetes**
01	Normale	Normales
02	Améliorée	Normales
03	Normale	Améliorées
04	Améliorée	Améliorées

TABLE 3.4 – Quatre parties d'exécution de la recherche d'information.

4.2 Regroupement des Résultats des Trois Moteurs de Recherche

Le processus de regroupement des résultats de trois moteurs de recherche représente un effort significatif visant à améliorer l'expérience des utilisateurs sur notre moteur de recherche. Après avoir collecté les 12 matrices de pertinence générées, nous avons procédé à la création d'une matrice de pertinence finale .

Pour extraire ce matrice , nous avons utilisé le code suivant (Figure 3.7) :

Ce code Python représente un travail minutieux et stratégique visant à consolider les résultats de recherche provenant de trois moteurs de recherche distincts. Tout d'abord, il importe les bibliothèques nécessaires pour lire et écrire des données à partir de fichiers Excel.

Ensuite, il parcourt les données des trois moteurs de recherche, stockées dans des feuilles de calcul différentes, et procède à l'analyse de chaque requête. Pour chaque requête, il évalue la pertinence des résultats Les 10 meilleurs résultats sont sélectionnés pour chaque requête, puis triés par pertinence décroissante.

Enfin, ces résultats consolidés sont écrits dans une nouvelle feuille de calcul Excel, fournissant ainsi une matrice finale contenant les informations les plus pertinentes pour chaque requête. Ce processus complexe vise à améliorer l'expérience de recherche des utilisateurs, en leur fournissant un accès rapide et efficace aux informations les plus pertinentes .

```
import xlwt
import networkx as nx
#import matplotlib.pyplot as plt
import xlrd
import random

workbook0 = xlwt.Workbook()
sheet0 = workbook0.add_sheet('1')

myBook = xlrd.open_workbook('code 1 .xls')
a = myBook.sheet_by_index (0)
b = myBook.sheet_by_index (1)
c = myBook.sheet_by_index (2)


for i1 in range(1,61):#requettes #61
    liste=[]
    for i2 in range(1,5):#etapes#5
        if i2==1 : lig=1
        if i2==2 : lig=12
        if i2==3 : lig=23
        if i2==4 : lig=34
        #i=0
        for i3 in range(0,10):#10 premiers resultats
            for i4 in range(1,4):#3 page exel
                if i4==1 : page=a
                if i4==2 : page=b
                if i4==3 : page=c

                trouver=False
                if i4==1:
                        ca=page.cell_value(rowx=lig+i3, colx=i1)
                        case=""
                        for ii in range(8,len(ca)-4):
                            case=case+ca[ii]
                else:
                        case=str(page.cell_value(rowx=lig+i3, colx=i1))
                i=0
                while (i < len(liste)and(trouver==False)):
                    if i4==1:
                        ca=page.cell_value(rowx=lig+i3, colx=i1)
                        if ca=="" : case=0
                        else:
                          case=""
                          for ii in range(8,len(ca)-4):
                            case=case+ca[ii]
                          case=int(case)
                    else:
                        if page.cell_value(rowx=lig+i3, colx=i1)=="" : case=0
                        else:case=int(page.cell_value(rowx=lig+i3, colx=i1))
                    #print(type(case),type(liste[i][0]))
                    #if (case==48):
                        #print(liste[i][0],case)
                        #print(liste)

                    if (liste[i][0]==case):
                     #   print('yes')
                        liste[i][1]=liste[i][1]+ ((10-i3)/10)
                        trouver = True
                    i=i+1

                #print(case)
                if trouver == False :
                    ccc=[]
                    ccc.append(int(case))
                    ccc.append(((10-i3)/10))
                    liste.append(ccc)
                #print(liste)

    for iii in range(0,len(liste)) :
        for iiii in range(0,len(liste)) :
            if (liste[iii][1] > liste[iiii][1]):
                vide=liste[iii]
                liste[iii]=liste[iiii]
                liste[iiii]=vide
    for i in range(0,10) :
      sheet0.write(i, i1,liste[i][0] )
      sheet0.write(i, i1+60,liste[i][1] )

workbook0.save('resultats_finalfinal.xls')
```

FIGURE 3.7 – Code d'extraction du résultat final.

5 Conclusion

Dans ce chapitre, nous avons exploré en détail notre approche stratégique pour la réalisation de notre projet.Notre approche se décompose en trois étapes principales : la construction du dataset et l'établissement des requêtes, la récupération de la pertinence via des moteurs de recherche open source, et enfin le regroupement des résultats obtenus.

Nous avons collecté manuellement des informations pertinentes à partir de diverses sources en ligne, ce qui nous a permis de créer une base de données structurée dans Excel.

Ensuite, nous avons sélectionné des moteurs de recherche open source tels que Lucene, Solr et Elasticsearch pour exécuter les requêtes et recueillir des résultats pertinents.

Le processus de fusion des résultats de ces moteurs, via du code Python, a permis de créer une matrice finale de pertinence que nous avons intégrée à notre moteur de recherche dédié. Cette méthodologie a prouvé son efficacité en améliorant la pertinence des résultats de recherche, offrant ainsi une expérience utilisateur optimale.

Chapitre IV

Implémentation Et Evaluation

1 Introduction

Ce chapitre est consacré à l'implémentation de notre système Proposé. Nous débutons en présentant l'environnement de développement, en détaillant les outils utilisés, les langages de programmation, ainsi que les moteurs de recherche utilisés dans le projet.

Ensuite, nous analysons et expliquons les résultats obtenus dans les différentes étapes de notre projet à l'aide des tableaux et des graphiques pour une meilleure compréhension, Nous comparons également les performances des trois moteurs de recherche selon les critères standard (rappel, précision et F-mesure).

Enfin, nous décrivons le fonctionnement de notre outil, présenté sous forme de site web. Ce dernier explique notre travail ainsi que les services offerts à l'utilisateur, tels que la recherche dans notre base de données à partir des différentes requêtes prédéfinies qui prennent la forme de questions que les touristes peuvent poser.

2 Différentes ressources utilisées

Dans la réalisation de tout projet informatique, la diversité des ressources et des outils utilisés est cruciale, chaque élément contribuant à un aspect spécifique du développement. Ainsi, nous exposons ici les langages de programmation et les outils employés à travers les différentes phases de notre projet : de sa conception à son déploiement. Chacun de ces éléments est sélectionné avec soin pour répondre aux exigences propres à chaque étape du processus de développement, garantissant ainsi une progression efficace et cohérente tout au long du projet.

2.1 Python

Python est un langage de programmation open source créé par le programmeur Guido van Rossum en 1991. Il s'agit d'un langage de programmation interprété, qui ne nécessite donc pas d'être compilé pour fonctionner. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser. [Benoslan and Benhamadou, 2020]

2.2 Bibliothèques de Python utilisés

Dans ce projet, nous avons utilisé déférentes bibliothèques Python pour faciliter le processus de développement,ces bibliothèques sont illustrées dans le tableau suivant :

Bibliothèque	Description
Pandas	La manipulation et l'analyse de données
Wordnet	L'analyse linguistique et le traitement automatique des langues
Matplotlib.pyplot	Créer des visualisations statiques
Xlwt	Créer et écrire des fichiers Excel
Xlrd	Lire des fichiers Excel

TABLE 4.1 – Bibliothèques Python utilisées.

2.3 Eclipse

Eclipse est une plate-forme d'outils universels un environnement de développement intégré (IDE) ouvert et extensible pour tout. L'une des initiatives les plus excitantes nées du monde du développement d'applications depuis longtemps et bénéficie du soutien considérable des plus grandes entreprises et organisations dans le secteur de la technologie. Eclipse est de plus en plus répandu dans les domaines commercial et universitaire.
Nous avons choisi d'utiliser l'environnement de développement intégré (IDE) Eclipse,Pour utiliser Lucene. Eclipse offre une interface conviviale et de nombreuses fonctionnalités qui facilitent le développement et le débogage des applications Java. En utilisant Eclipse, nous avons pu intégrer facilement la bibliothèque Lucene.

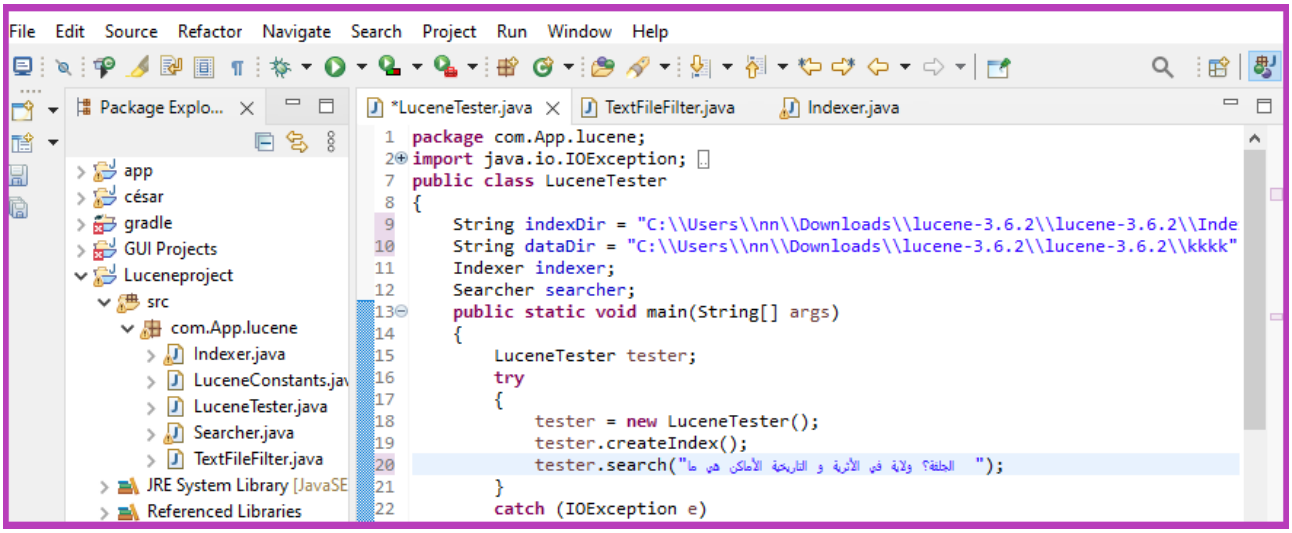


FIGURE 4.1 – Interface d'un programme Java sous Eclipse.

2.4 Wampserver

WampServer (anciennement WAMP5) est une plateforme de développement Web de type WAMP, permettant de faire fonctionner localement (sans se connecter à un serveur externe) des scripts PHP. WampServer n'est pas en soi un logiciel, mais un environnement comprenant

deux serveurs (Apache et MySQL), un interpréteur de script (PHP), ainsi qu'une administration pour les deux bases SQL PhpMyAdmin et SQLiteManager. Il dispose d'une interface d'administration permettant de gérer et d'administrer ses serveurs au travers d'un tray-icon (icône près de l'horloge de Windows). La grande nouveauté de WampServer 2 réside dans la possibilité d'installer et d'utiliser n'importe quelle version de PHP, Apache ou MySQL en un clic. Ainsi, chaque développeur peut reproduire fidèlement son serveur de production sur sa machine locale. [Ndiaye, 2023]

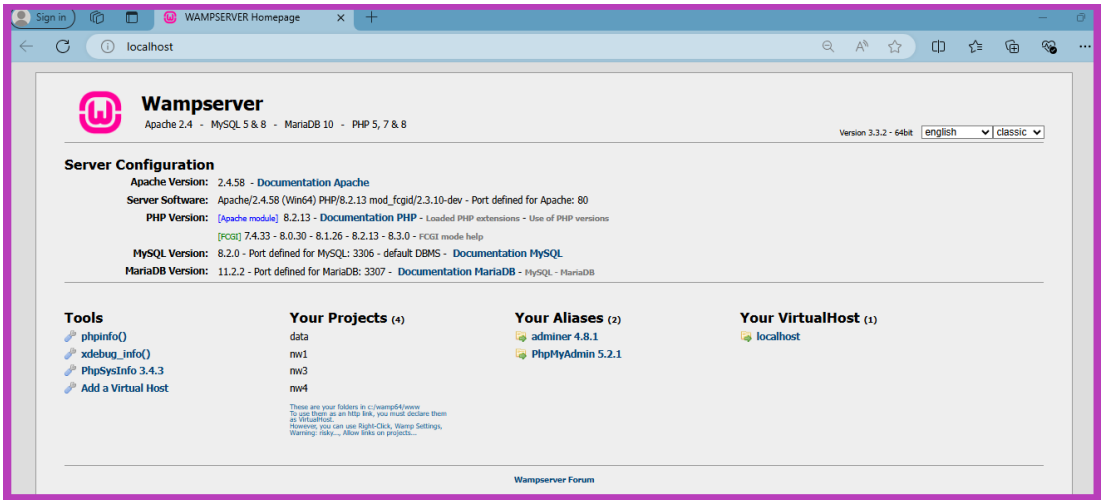


FIGURE 4.2 – Page d'accueil de WampServer.

3 Moteurs de recherche utilisée

3.1 Lucene

Lucene est une bibliothèque de moteur de recherche open source développée par la Fondation Apache Software, qui fournit des interfaces de programmation pour l'indexation et la recherche de documents texte. Utilisée principalement pour mettre en œuvre des moteurs de recherche textuels, Lucene permet d'indexer efficacement des documents et de rechercher du texte en utilisant des algorithmes de recherche précis. Bien que basée sur Java, Lucene peut être utilisée avec d'autres langages de programmation comme Perl, Python et .Net. Elle prend en charge diverses requêtes telles que PhraseQuery, WildcardQuery, RangeQuery, FuzzyQuery et BooleanQuery. Lucene est employée par des sites importants tels que Wikipédia et LinkedIn pour rechercher sur le web et dans des bases de données, suivant les principes de la "Apache Way". [Zhao et al., 2012]

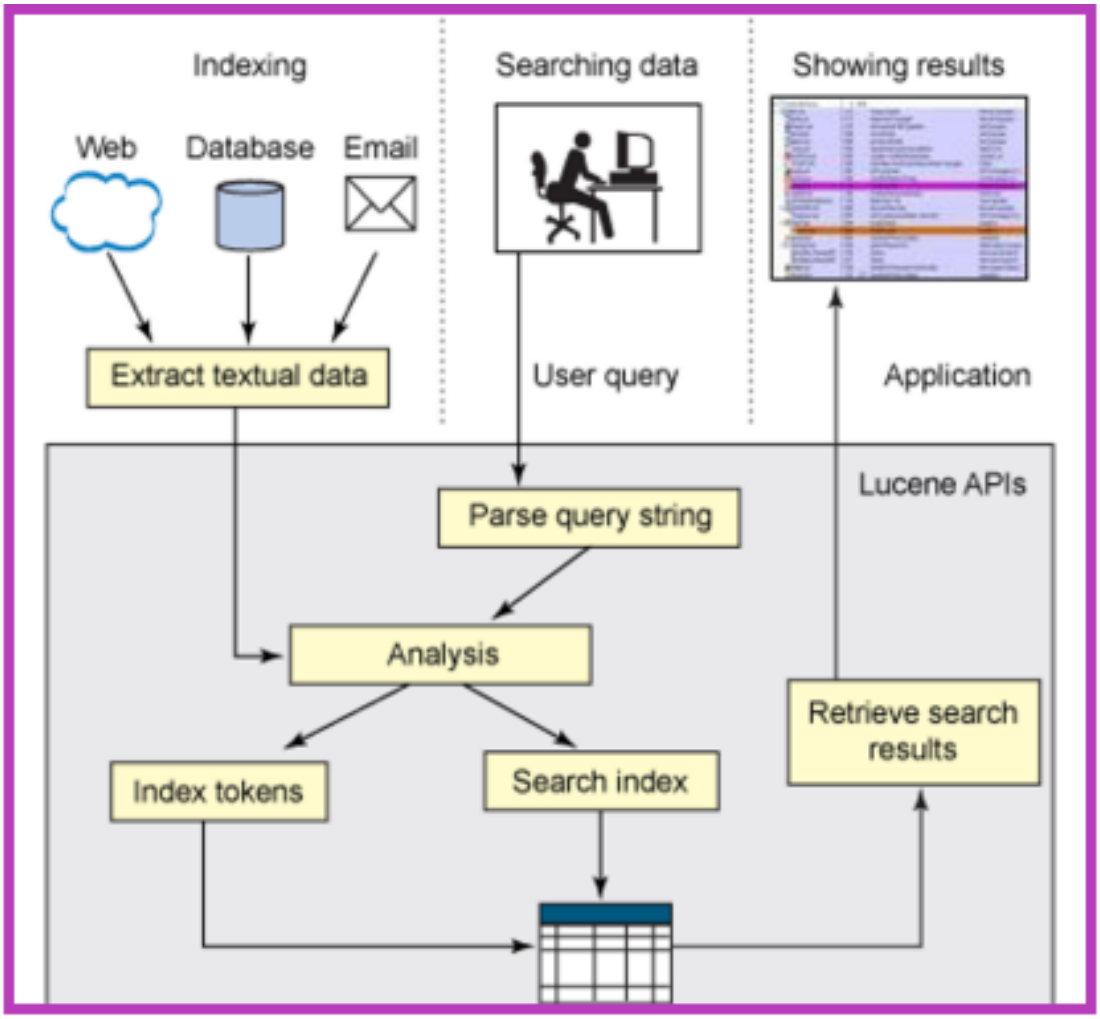


FIGURE 4.3 – Création d'une application avec Lucene.

3.2 Elasticsearch

Elasticsearch est un moteur de recherche et un outil d'analyse de données qui a été développé à partir de la sous-structure Apache Lucene, et qui est léger, facile à installer, codé en open source et scalable. Ce moteur de recherche, qui offre des services via une API Restful, est très rapide et pratique grâce à ses outils visuels tiers et ses options de sécurité. Elasticsearch est moins complexe et moins détaillé que le moteur de recherche Solr, également produit à partir d'Apache Lucene, mais il est tout aussi performant. Il est robuste et flexible. L'un des avantages les plus importants est qu'il est en temps réel et distribué. Aujourd'hui, Elasticsearch est utilisé dans la recherche de contenu, l'analyse de données et les requêtes dans des projets tels que Mozilla et Foursquare. [Kılıc and Aksakalli, 2016]

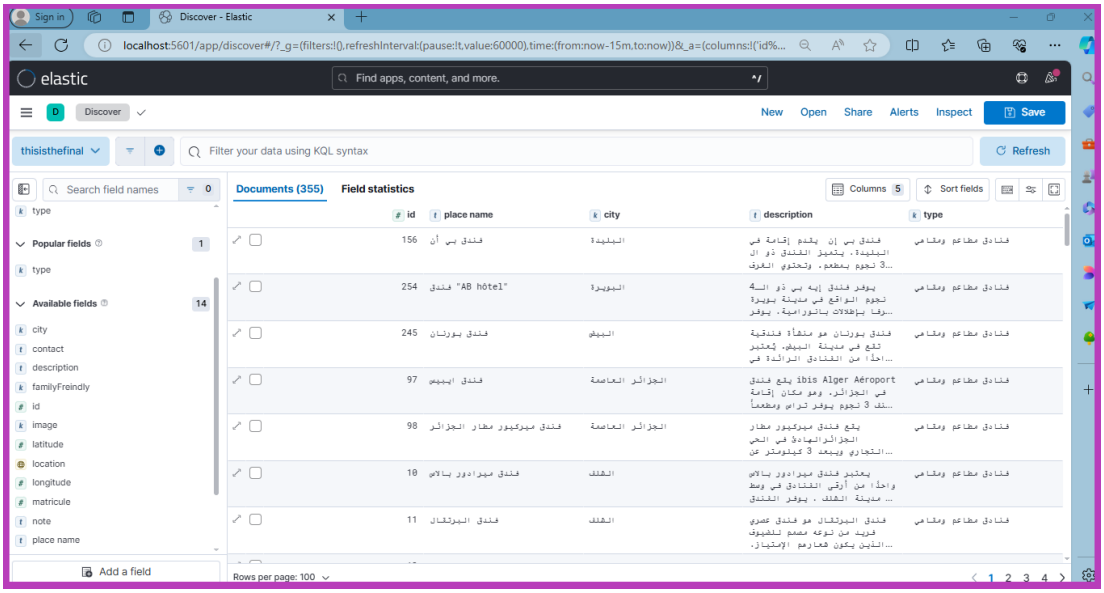


FIGURE 4.4 – Interface du moteur de recherche Elasticsearch.

3.3 Solr

Solr est une plateforme de recherche d'entreprise open source de la Fondation Apache. Développée par Yonik Seeley chez CNET Networks en 2004, Solr est principalement écrite en langage Java et fonctionne également de manière multiplateforme. Solr fonctionne comme un serveur de recherche en texte intégral autonome et offre différentes fonctionnalités telles que la recherche en texte intégral, la mise en surbrillance des résultats, l'indexation en temps réel, le clustering dynamique, l'intégration de bases de données, des fonctionnalités NoSQL et, enfin, la gestion de documents riches. Solr est l'un des plus grands outils pour les systèmes de récupération d'information et est utilisé dans de nombreuses applications réelles comme les médias, le commerce électronique, les portails d'emploi et de carrières, la recherche d'entreprise et la recherche sur les réseaux sociaux. [Gondaliya and Joshi, 2017] la figure suivante représente l'interface du moteur de recherche Solr :

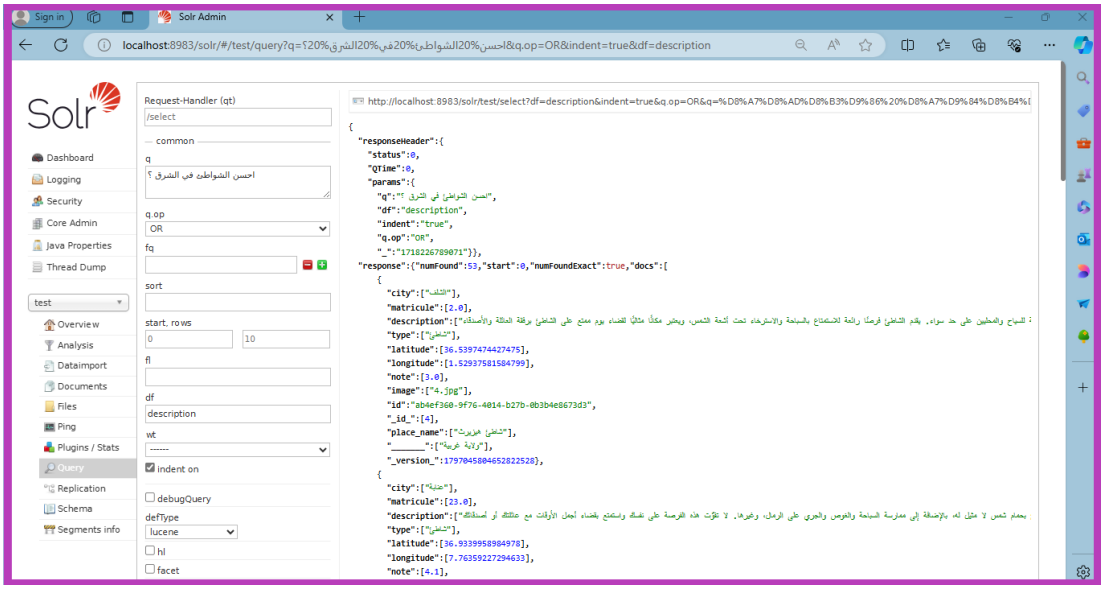


FIGURE 4.5 – Interface du moteur de recherche Solr.

4 Résultats de la recherche

4.1 Résultat de pertinence par les moteurs de recherche

Après l'utilisation des trois moteurs de recherches,nous avons extrait 12 matrices"moteurs × 4 parties"(comme on a expliqué en détail dans le chapitre précédent),Nous avons donné dans le tableau suivant un exemple d'extraction des matrices de pertinence du moteur1 (LUCENE) :

Parties	Q1	Q2	Q3	...	...	...	...	...	Q58	Q59	Q60
01	19	355	74	...	...	...	...	...	258	85	219
	20	314	83	...	...	...	...	...	103	148	258
	..	..	..	...	...	...	...	...	..	..	..
	..	..	..	...	...	...	...	...	..	..	..
	..	..	..	...	...	...	...	...	..	..	..
	14	336	76	...	...	...	...	...	351	256	218
	15	288	216	...	...	...	...	...	129	335	129
02	19	314	74	...	...	...	...	...	258	134	212
	21	274	80	...	...	...	...	...	103	187	216
	..	..	..	...	...	...	...	...	..	..	..
	..	..	..	...	...	...	...	...	..	..	..
	..	..	..	...	...	...	...	...	..	..	..
	275	260	231	...	...	...	...	...	272	206	232
	20	288	314	...	...	...	...	...	136	274	215
03	19	274	76	...	...	...	...	...	258	114	219
	21	314	74	...	...	...	...	...	232	82	232
	..	..	..	...	...	...	...	...	..	..	..
	..	..	..	...	...	...	...	...	..	..	..
	..	..	..	...	...	...	...	...	..	..	..
	17	336	78	...	...	...	...	...	12	258	79
	48	265	300	...	...	...	...	...	48	46	167
04	9	352	74	...	...	...	...	...	232	9	219
	19	353	78	...	...	...	...	...	258	191	232
	..	..	..	...	...	...	...	...	..	..	..
	..	..	..	...	...	...	...	...	..	..	..
	..	..	..	...	...	...	...	...	..	..	..
	21	355	300	...	...	...	...	...	167	4	82
	13	5	77	...	...	...	...	...	82	157	217

TABLE 4.2 – Les matrices de pertinence du Lucene.

4.2 Comparaison des résultats des 04 parties

Après avoir utilisé les 3 moteurs de recherche pour 4 étapes et 60 requêtes, nous avons extrait 12 matrices. Ensuite, nous avons calculé la précision, le rappel et le f-mesure. Les résultats ont montré que la configuration (4), avec bdd et des requêtes améliorées, a produit les meilleures performances en termes de précision, rappel et F-mesure.par contre, la configuration (1), avec des bdd et des requêtes normales, a donné les pires résultats. Les configurations (2) et (3), avec des améliorations soit dans la bdd soit dans les requêtes, ont offert de meilleures performances que la configuration (1), mais n'ont pas atteint les niveaux de performance de la configuration (4). Cela indique que l'amélioration des deux éléments - données et requêtes - est cruciale pour optimiser les résultats des moteurs de recherche.

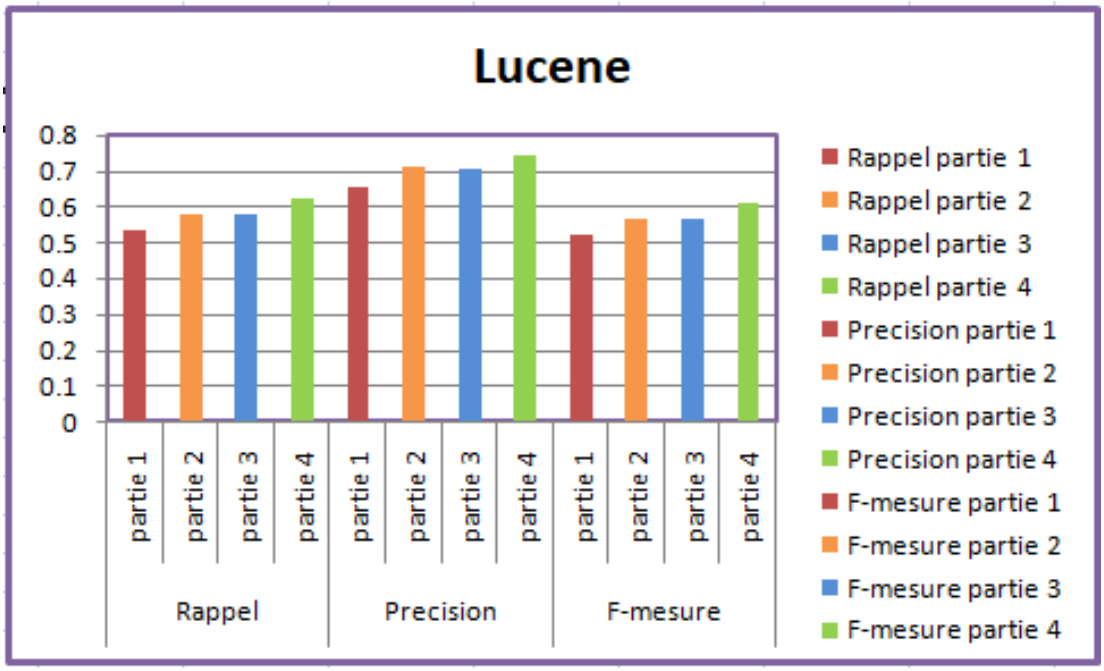


FIGURE 4.6 – Comparaison des 04 parties du Lucene.

4.3 Comparaison entre les résultats des moteurs de recherche

Après le choix des moteurs de recherche open source, qui sont Lucene, Elasticsearch et Solr, reconnus pour leurs performances et largement utilisés dans divers domaines, les résultats retournés par chaque moteur ont été collectés, puis analysés pour évaluer la précision, le rappel et la F-mesure. Ces évaluations ont été basées sur la liste manuellement extraite des documents pertinents pour chaque requête (expliqué au chapitre 3).

Lucene s'est avéré supérieur à Elasticsearch et Solr, avec une précision moyenne (avg-precision) de 0.704, suivie par Elasticsearch avec 0.681, et Solr avec 0.652.

Le tableau suivant présente les résultats de chaque moteur en termes de précision moyenne, rappel moyen et F-mesure moyenne :

Moteur	Avg-Precision	Avg-Rappel	Avg-F-Mesure
Lucene	0.704	0.579	0.566
Elasticsearch	0.681	0.557	0.5415
Solr	0.652	0.530	0.521

TABLE 4.3 – Performances des moteurs de recherche selon les critères évalués.

Pour illustrer ces comparaisons, les figures suivantes présentent des courbes de chaque mesure pour les trois moteurs.

Le moteur 1 correspond à Lucene, le moteur 2 à Elasticsearch et le moteur 3 à Solr :

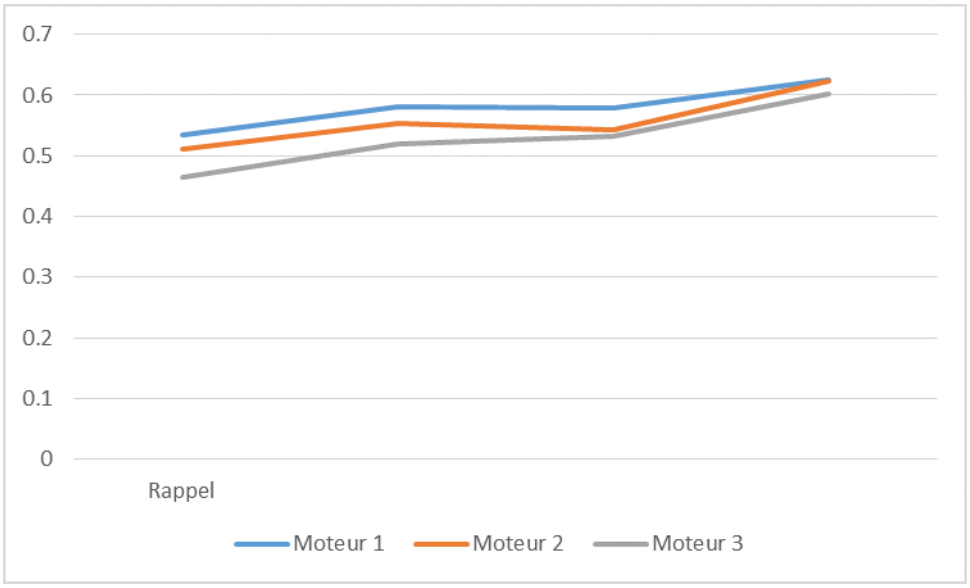


FIGURE 4.7 – Courbe de Comparaison du rappel des Trois Moteurs de Recherche.

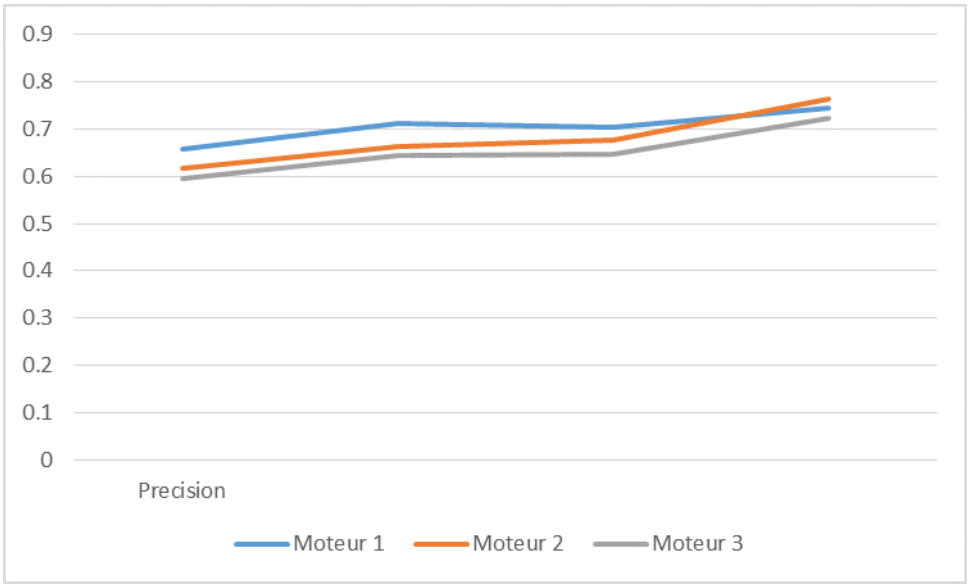


FIGURE 4.8 – Courbe de Comparaison de la précision des Trois Moteurs de Recherche.

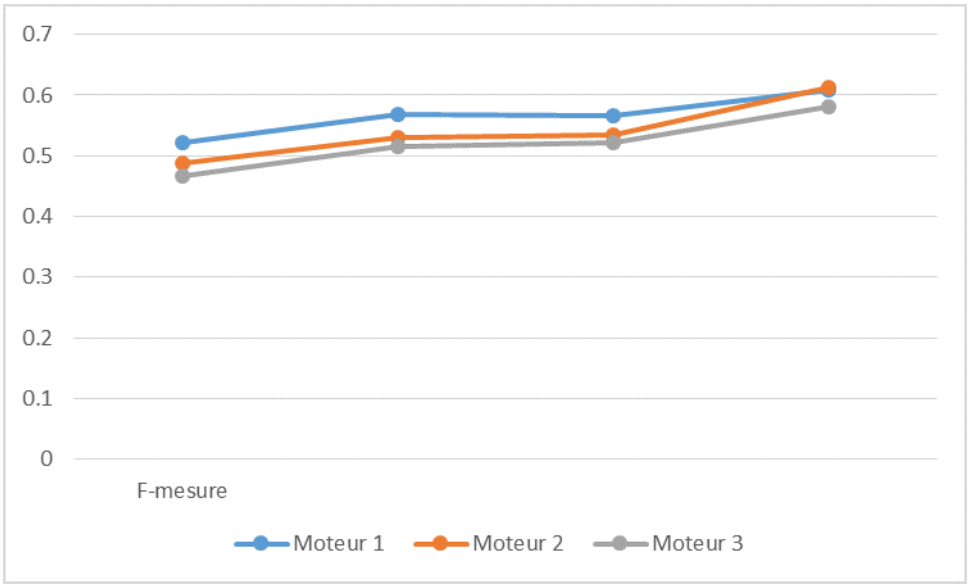


FIGURE 4.9 – Courbe de Comparaison de la f-mesure des Trois Moteurs de Recherche.

4.4 Resultats de regroupements des trois moteurs

Pour regrouper les résultats de trois moteurs de recherche (représentés par 12 matrices de pertinence) dans une seule matrice , on a utilisé un code Python suivant une méthodologie efficace. Chaque requête génère 10 résultats pertinents par partie, et avec 4 parties par requête, cela totalise 40 résultats pertinents par requête et pour chaque moteur de recherche. Pour chaque résultat de pertinent, un score est attribué en fonction de sa pertinence dans différentes parties de la requête. Ces scores sont normalisés pour refléter une échelle standard de pertinence de 0 à 1, où 1 représente une pertinence maximale et 0 une pertinence minimale. Une fois que tous les scores individuels ont été calculés pour chaque résultat pertinent dans chaque partie et pour chaque moteur de recherche, les résultats sont agrégés en additionnant les scores pertinents pour chaque requête. Ensuite, les résultats sont classés par ordre décroissant de pertinence. Les 10 résultats les plus pertinents sont sélectionnés dans cet ordre pour chaque requête.

Nous avons fourni dans la figure suivante un petit exemple simplifié illustrant comment les scores sont calculés :

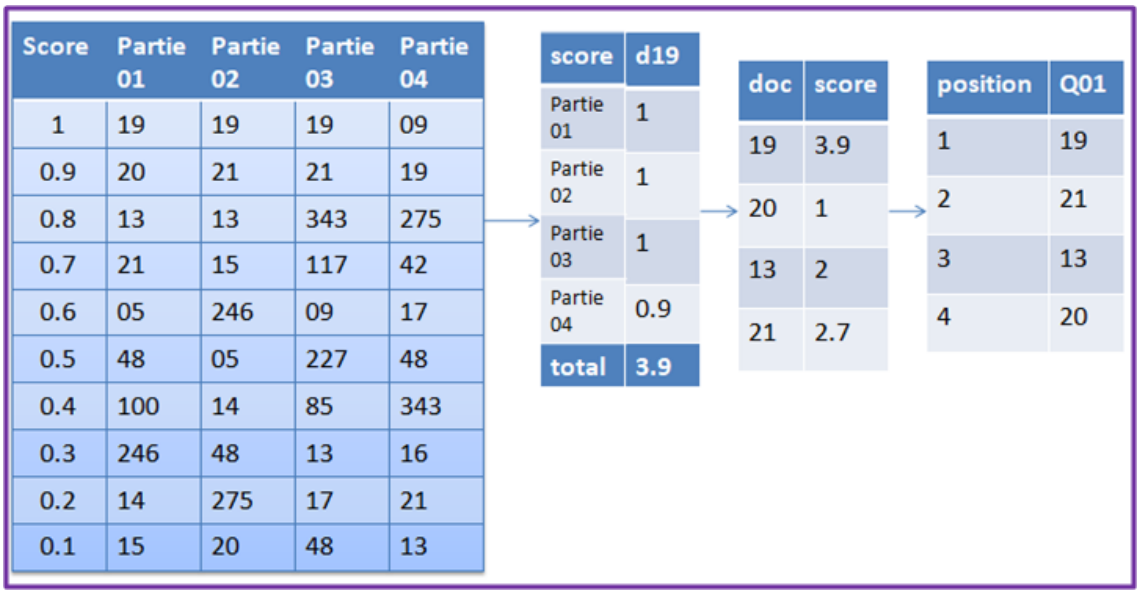


FIGURE 4.10 – Exemple de calcul du score.

Après avoir calculé le score de chaque document pertinent, nous avons regroupé les 10 meilleurs résultats (les documents pertinents avec le score le plus élevé) pour chaque requête, comme indiqué dans le tableau suivant :

	Q 01		Q 02		...	Q 59		Q 60	
Position	**D.Pert**	**Score**	**D.Pert**	**Score**	...	**D.Pert**	**Score**	**D.Pert**	**Score**
1	19	8.7	354	8	...	9	5.5	219	7.6
2	100	7.5	352	7.4	...	114	5.4	136	7.6
3	9	5.4	314	6.6	...	4	5.4	258	5.7
4	13	5	355	6.5	...	184	3.5	232	5.4
5	48	4	61	6.2	...	134	2.7	216	5.4
6	21	3.8	274	5.1	...	19	2.6	244	4.4
7	20	3.4	353	4.5	...	11	2.6	213	3.6
8	343	2.7	40	2.5	...	58	2.5	212	2.8
9	17	2.7	48	2	...	191	2.4	351	2.7
10	246	2.1	5	1.8	...	39	2.1	82	2.6

TABLE 4.4 – Résultats finaux.

La matrice de pertinence finale résultante est présentée dans le tableau ci-dessous (Table 4.4), où les documents pertinents pour chaque requête (Q1 à Q60) sont récapitulées.

Requête	**Q1**	**Q2**	**Q3**	...	...	...	...	...	**Q58**	**Q59**	**Q60**
Document pertinent	19	354	74	...	...	...	...	...	258	9	219
	100	352	76	...	...	...	...	...	232	114	136
	9	314	79	...	...	...	...	...	219	4	258
	13	355	75	...	...	...	...	...	48	148	232
	48	61	80	...	...	...	...	...	344	134	216
	21	274	300	...	...	...	...	...	66	19	244
	20	353	78	...	...	...	...	...	244	11	213
	343	40	77	...	...	...	...	...	136	58	212
	17	48	83	...	...	...	...	...	247	191	351
	246	5	171	...	...	...	...	...	99	39	82

TABLE 4.5 – Matrice de pertinence finale.

Après avoir regroupé les résultats des trois moteurs dans une seule matrice(Matrice de pertinence finale) à l'aide du code Python, tel qu'expliqué précédemment, nous avons évalué les performances en termes de rappel, de précision et de F-mesure.
Les courbes de rappel et de précision pour les 60 requêtes sont illustrées respectivement dans les figures 4.10 et 4.11, tandis que la courbe de la F-mesure est représentée dans la figure 4.12.

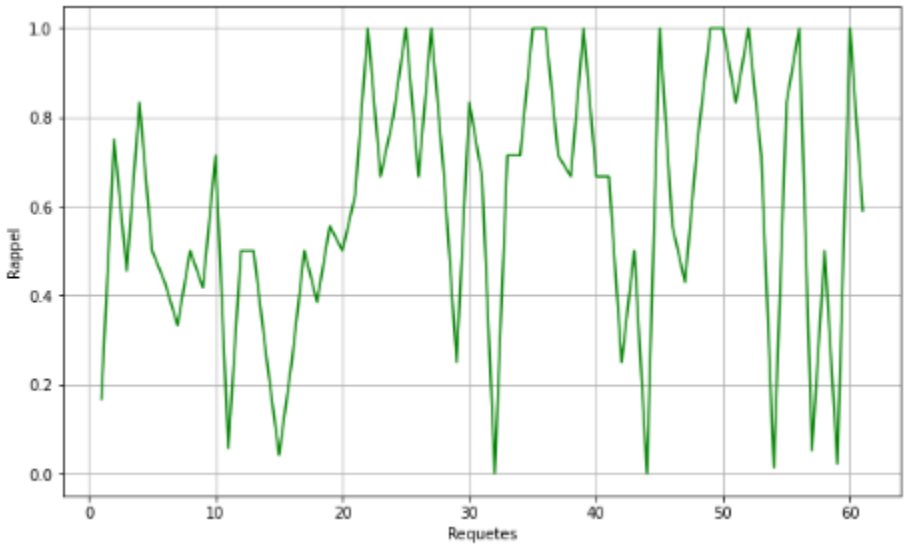


FIGURE 4.11 – Courbe de Rappel des 60 requêtes.

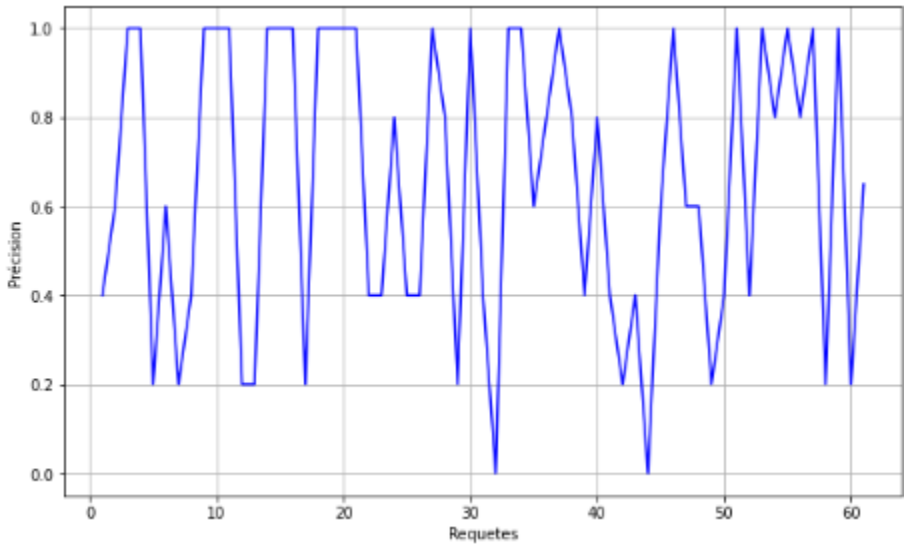


FIGURE 4.12 – Courbe de la Précision des 60 requêtes.

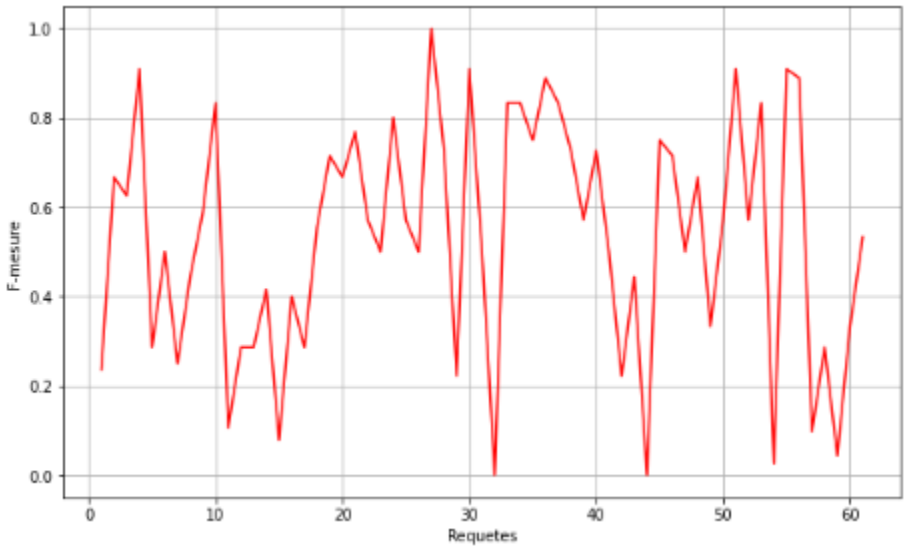


FIGURE 4.13 – Courbe de la F-mesure des 60 requêtes.

5 Présentation du site web

Afin de mettre en pratique notre approche, nous avons créé un site web en PHP en utilisant l'environnement HTML.

Notre site offre une interface simple avec une page d'accueil qui permet aux utilisateurs de sélectionner une requête parmi les 60 que nous avons préalablement extraites. La page d'accueil du site web développé est représentée dans la figure suivante .

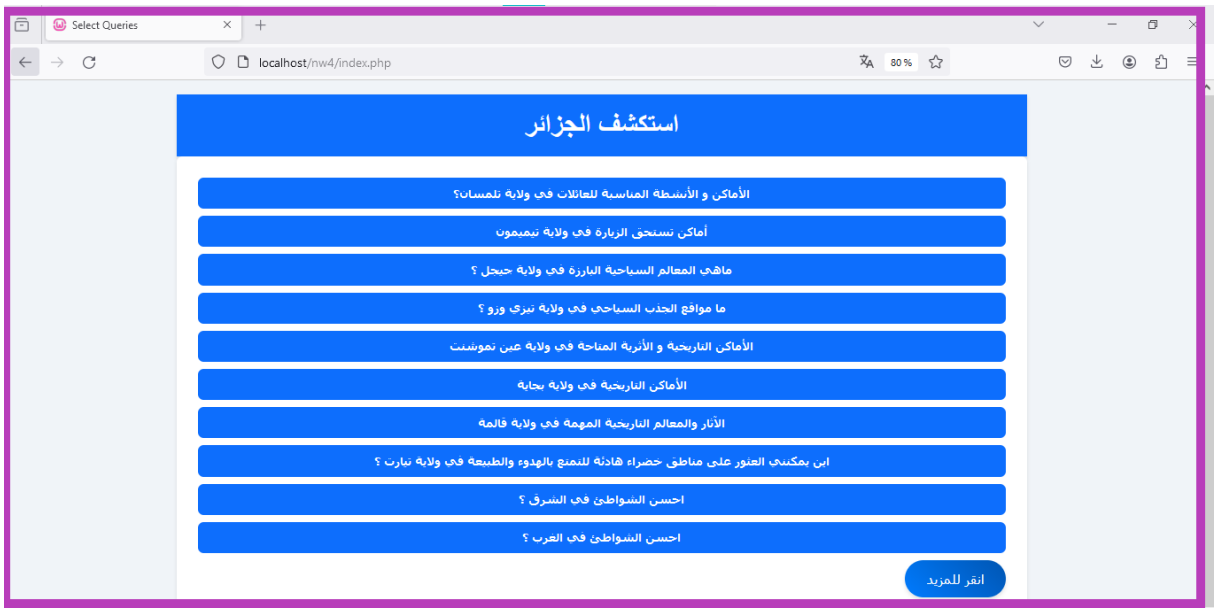


FIGURE 4.14 – Page d'accueil du site web developpée.

Une fois que la requête a été choisie, le site web affiche les résultats de la recherche sous forme d'une liste de places pertinentes correspondant à celle-ci, accompagnées d'une image et du nom de chaque place. La figure 4.14 présente la page de résultat d'une requête sélectionnée.

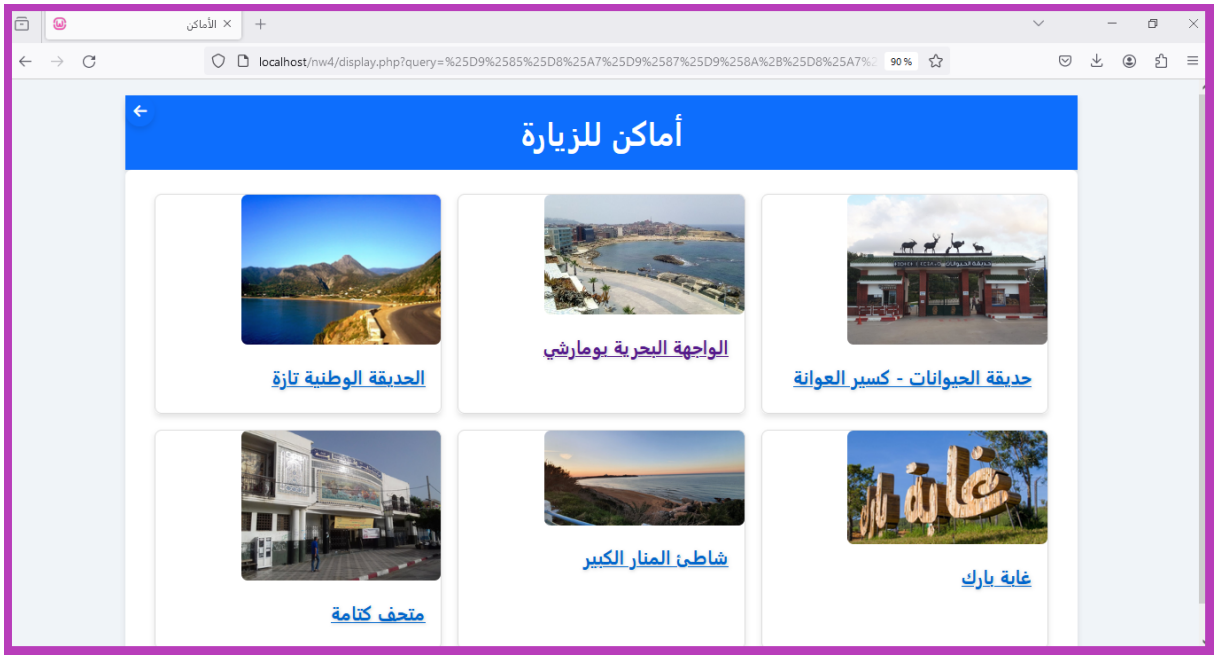


FIGURE 4.15 – Affichage des résultats de recherche des places pertinents.

Après que l'utilisateur a cliqué sur le nom ou l'image d'une place, une nouvelle page s'affiche

avec toutes les informations touristiques sur la place sélectionnée :

- Une image
- Description de la place
- Nom de la wilaya
- Matricule de la wilaya
- Direction de la wilaya
- Type de place
- Contact
- Indication si le lieu est familial ou non (danger pour les enfants et les femmes)
- Localisation sur Maps

La figure suivante présente la page des informations détaillées :



FIGURE 4.16 – Page d'informations touristiques de la place sélectionnée.

6 Conclusion

Dans ce chapitre, nous avons détaillé le processus d'implémentation et d'évaluation de notre projet. Nous avons d'abord présenté les différentes ressources utilisées, en particulier les langages de programmation et les outils tels que Python et ses bibliothèques, ainsi que l'IDE Eclipse. Ensuite, nous avons examiné les moteurs de recherche employés (Lucene, Elasticsearch, et Solr),en évaluant leurs performances à travers diverses configurations de bases de données et de requêtes.

Nous avons discuté sur les résultats obtenus dans ce travail, par la comparaison entre les étapes du projet ou par des tableaux et des courbes qui donnent une vision totale sur le projet Nous avons discuté des résultats obtenus dans notre étude, à travers des tableaux et des courbes permettant une vision globale du travail.Les résultats obtenus montrent une variation des performances de chaque moteur de recherche et leur adéquation aux différents besoins. Nous avons également démontré comment le regroupement des résultats de plusieurs moteurs de recherche peut améliorer l'expérience utilisateur en fournissant des informations plus précises et pertinentes.

Enfin, nous avons présenté le site web développé pour afficher les résultats de recherche, permettant aux utilisateurs d'accéder facilement aux informations touristiques recherchées.

Conclusion générale

Conclusion Générale et Perspectives

Dans ce mémoire, nous avons développé une plateforme touristique en arabe dédiée à promouvoir le tourisme en Algérie, basée sur des études menées à l'aide des moteurs de recherche open source.

La première étape réalisée est la création de la base de données. Cette base contient 355 lieux touristiques pour 58 wilayas, avec leurs informations détaillées.

Après la création de cette base de données, nous avons extrait la liste des requêtes spécifiques. Ensuite, à l'aide d'un code Python, nous avons généré une version améliorée de notre base de données ainsi que de la liste des requêtes.

Nous avons ensuite procédé à l'analyse approfondie et aux tests de chaque version de la base de données avec chaque version de la liste des requêtes, en utilisant les moteurs de recherche open source.

Les résultats ont montré que les versions améliorées de la base de données et des requêtes ont donné des résultats beaucoup plus pertinents par rapport aux versions non améliorées.

Nous avons également réalisé des comparaisons de performances entre trois moteurs de recherche open source : Lucene, Elasticsearch et Solr.

Les résultats obtenus sont les suivants : Lucene s'est avéré supérieur à Elasticsearch et Solr, avec une précision moyenne (avg-precision) de 0,704. Lucene a également obtenu un rappel de 0.579 et une F-mesure de 0.566. Elasticsearch suit avec une précision moyenne de 0,681, un rappel de 0.557 et une F-mesure de 0.5415.Solr arrive en troisième position avec une précision moyenne de 0,652, un rappel de 0.530 et une F-mesure de 0.521.Ainsi,nous avons atteint nos objectifs en répondant à la problématique posée dans ce travail.

Cette étude est menée dans le but de faciliter l'accès à des informations précises et fiables sur les destinations touristiques populaires du pays, spécifiquement adaptées aux besoins des visiteurs arabophones en Algérie.

ce travail ouvre une voie vers diverses prespectives qui se situent sur les plans suivants :

• Améliorer l'expérience utilisateur en ajoutant une barre de recherche pour faciliter la recherche des informations nécessaires.

• Élargir le public potentiel en rendant la plateforme disponible en plusieurs langues, en commençant par le français.

• Rendre la plateforme disponible en ligne pour faciliter l'accès et l'utilisation par les utilisateurs à tout moment et en tout lieu.

• Mener une étude pour comparer les algorithmes des moteurs de recherche open source afin d'inclure les caractéristiques distinctives de chaque moteur dans le nôtre, dans le but de fournir de meilleurs résultats.

• Élargir la base de données pour inclure davantage d'informations touristiques diversifiées et complètes.

ANNEXE 1

Une requête est utilisée pour interroger ou manipuler les données présentes dans des tables. Elle est constituée de termes spécifiques et peut être améliorée par les systèmes de recherche d'information en ajoutant des synonymes, pour élargir les résultats.

Dans le cadre de notre travail, nous avons effectué une extraction des requêtes provenant des informations collectées, nous avons détaillé cette étape dans le troisième chapitre.

Voici le tableau global des 60 requêtes :

Query Id	**Query**
Q01	الأماكن و الأنشطة المناسبة للعائلات في ولاية تلمسان؟
Q02	أماكن تستحق الزيارة في ولاية تيميمون
Q03	ماهي المعالم السياحية البارزة في ولاية جيجل ؟
Q04	ما مواقع الجذب السياحي في ولاية تيزي وزو ؟
Q05	الأماكن التاريخية و الأثرية المتاحة في ولاية عين تموشنت
Q06	الأماكن التاريخية في ولاية بجاية
Q07	الآثار والمعالم التاريخية المهمة في ولاية قالمة
Q08	اين يمكنني العثور على مناطق خضراء هادئة للتمتع بالهدوء والطبيعة في ولاية تيارت ؟
Q09	احسن الشواطئ في الشرق ؟
Q10	احسن الشواطئ في الغرب ؟
Q11	ما هي أفضل الوجهات السياحية في الصحراء الجزائرية ؟
Q12	اين يمكنني العثور على أفضل المطاعم في ولاية عنابة ؟
Q13	اجود المطاعم في ولاية جيجل ؟
Q14	ما هي أفضل الوجهات السياحية في ولاية المدية ؟
Q15	أين تقع أجمل الأماكن السياحية في الشرق ؟
Q16	ما هي أفضل الأماكن السياحية في ولاية الجزائر العاصمة ؟
Q17	يقال ان تيبازة مدينة اثرية بامتياز لكن لم استطع تحديد وجهة اثرية و تاريخية
Q18	الاماكن المثيرة للاهتمام في ولاية وهران؟
Q19	انا سائح اجنبي و ابحث عن اهم المعالم السياحية بولاية تمنراست ؟
Q20	أبرز الأماكن السياحية في ولاية قسنطينة ؟
Q21	ابرز الوجهات السياحية في ولاية البويرة ؟
Q22	الأماكن التاريخية في ولاية عين الدفلى ؟

Q23	أماكن للاستمتاع بالطبيعة في ولاية بومرداس ؟
Q24	انا متواجد في ولاية سيدي بلعباس و لم اعرف بعد اين اقضي افضل الاوقات
Q25	ابحث عن أفضل الفنادق في ولاية سطيف؟
Q26	ما هي الشواطئ الجميلة التي يمكن زيارتها في ولاية الشلف ؟
Q27	أجمل الأماكن المقترحة لزيارتها في ولاية الطارف ؟
Q28	ما هي المناطق التاريخية الاثرية المشهورة في ولاية باتنة؟
Q29	ما هي الأماكن التاريخية التي يمكن زيارتها في ولاية سطيف؟
Q30	ما هي الوجهات السياحية في ولاية تبسة؟
Q31	ما هي الأماكن التاريخية و الأثرية في ولاية الجلفة؟
Q32	الأماكن التاريخية و الأثرية في ولاية غرداية؟
Q33	ما الأماكن التي تتميز بها ولاية البليدة؟
Q34	ماهي المعالم السياحية البارزة في ولاية سكيكدة؟
Q35	أفضل الأماكن للسياحة في ولاية ام البواقي؟
Q36	ما هي الوجهات السياحية الشهيرة في ولاية ميلة ؟
Q37	ما الأماكن التي يجب زيارتها عند التواجد بولاية مستغانم ؟
Q38	الينابيع الحرارية التي يمكن زيارتها للاسترخاء في الجزائر؟
Q39	أفضل شاطئ في ولاية وهران؟
Q40	المواقع الأثرية الموجودة في ولاية تلمسان؟
Q41	ماهي اجمل الحدائق والمنتزهات في الجزائر العاصمة ؟
Q42	ما هي الأماكن الطبيعة الخلابة التي يجب زيارتها في ولاية سعيدة؟
Q43	ماهي الأماكن الأثرية التي يجب زيارتها في ولاية معسكر؟
Q44	أين يمكنني الإقامة في ولاية خنشلة؟
Q45	ما هي الأماكن التي يمكن زيارتها في ولاية أدرار؟
Q46	الوجهات السياحية الشهيرة في ولاية بسكرة
Q47	الاماكن الطبيعية الخلابة الموجودة في ولاية تمنراست؟
Q48	أفضل مكان سياحي ديني في ولاية بشار؟
Q49	ماهو افضل مركز تسوق في ولاية سطيف
Q50	محمية طبيعية وطنية في تيزي وزو

Q51	ما هي أفضل الأماكن السياحية في ولاية برج بوعريريج ؟
Q52	ماهي اهم القصور والمساجد التي يجب زيارتها في ولاية قسنطينة ؟
Q53	ماهي الينابيع الحرارية المعروفة بفوائدها الصحية؟
Q54	اين يمكنني ان استمتع بوقتي في الجزائر
Q55	ابحث عن قرية سياحية لتمضية عدة أيام بعيدا على صخب المدينة
Q56	اريد اقتراح واحات صحراوية بالجزائر
Q57	اريد قضاء أوقات مسلية بالشرق الجزائري
Q58	اين يمكنني تناول وجبة تقليدية بالغرب الجزائري
Q59	لدي أطفال صغار و اريد تمضية يوم لا ينسى معهم
Q60	اين يمكنني تناول وجبة تقليدية في ولاية بسكرة

ANNEXE 2

Document de la BDD sous format de XML :

Comme nous l'avons mentionné précédemment, nous avons construit notre base de données (BDD) en 5 formats. Nous souhaitons donner un exemple du format XML (Extensible Markup Language) pour un document de la base. À l'aide de ce langage de balisage, nous avons inclus 10 informations pour chaque document : id, nom du lieu, wilaya, matricule de la wilaya, description du lieu, type (catégorie) du lieu, latitude, longitude, sa note dans le Maps, contact (numéro de téléphone, site web, page Facebook, etc.), et enfin la zone (direction) de la wilaya.

<**doc**>

<**id**> 3 </**id**>

<**place name**> فندق جنان مالك٣ </**place name**>

<**city**> أدرار </**city**>

<**matricule**> 01 </**matricule**>

<**description**>

فندق جنان مالك في أدرار هو واحد من الوجهات السياحية المميزة في المنطقة، حيث يوفر تجربة إقامة مريحة وفاخرة للنزلاء. يتميز الفندق ببنيته الحديثة والفاخرة، وتصميمة الجميل الذي يجمع بين الأصالة والراحة العصرية

</**description**>

<**type**> فنادق مطاعم ومقاهي </**type**>

<**latitude**> 29.2663622318795 </**latitude**>

<**longitude**> 0.224185669307999 </**longitude**>

<**note**> 3.9 </**note**>

<**contact**> 0560 00 61 88 </**contact**>

<**zone**> ولاية صحراوية </**zone**>

</**doc**>

Bibliographie

[Albukhitan et al., 2016] Albukhitan, S., Alnazer, A., and Helmy, T. (2016). Semantic annotation of arabic web resources using semantic web services. *Procedia Computer Science*, 83 :504–511.

[Albukhitan et al., 2018] Albukhitan, S., Alnazer, A., and Helmy, T. (2018). Semantic annotation of arabic web documents using deep learning. *Procedia computer science*, 130 :589–596.

[Alnaied et al., 2020] Alnaied, A., Elbendak, M., and Bulbul, A. (2020). An intelligent use of stemmer and morphology analysis for arabic information retrieval. *Egyptian Informatics Journal*, 21(4) :209–217.

[Amer, 2020] Amer, N. O. (2020). *Recherche sociale et personnalisée d'Information.* PhD thesis, Université Grenoble Alpes [2020-....].

[Amine et al., 2019] Amine, E., Madani, Y., El Ayachi, R., and ErrerERRITALI, M. (2019). A new semantic similarity approach for improving the results of an arabic search engine. *Procedia Computer Science*, 151 :1170–1175.

[Baclic et al., 2020] Baclic, O., Tunis, M., Young, K., Doan, C., Swerdfeger, H., and Schonfeld, J. (2020). Le traitement du langage naturel (tln), une sous-zone d'intelligence artificielle.

[Benoslan and Benhamadou, 2020] Benoslan, M. O. and Benhamadou, S. (2020). *Conception et réalisation d'un système de télésurveillance en temps réel à base de Raspberry.* PhD thesis.

[Bernier, 2021] Bernier, A. (2021)). Notre guide de la langue arabe et ses dialectes.

[Boudry, 2002] Boudry, C. (2002). Typologie et mode de fonctionnement des outils de recherche d'information sur internet en biologie/médecine. *médecine/sciences*, 18(5) :616–622.

[Boughanem and Savoy, 2008] Boughanem, M. and Savoy, J. (2008). *Recherche d'information : état des lieux et perspectives.* Hermes Science Publications.

[Boulaknadel, 2008] Boulaknadel, S. (2008). *Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de spécialité : Apport des connaissances morphologiques et syntaxiques pour l'indexation.* PhD thesis, Université de Nantes.

[Bouzeghoub and Kostadinov, 2005] Bouzeghoub, M. and Kostadinov, D. (2005). Personnalisation de l'information : aperçu de l'état de l'art et définition d'un modèle flexible de profils. *CORIA*, 5 :201–218.

[Cancedda and Renders, 2010] Cancedda, N. and Renders, J.-M. (2010). Cross-lingual text mining.

[Chouikha, 2010] Chouikha (2010). *La langue arabe, son histoire, son originalité, et son influence.*

[Diallo, 2023] Diallo, A. K. (2023). Détection automatique des contenus offensifs en bambara sur les réseaux sociaux.

[Fahima, 2023] Fahima, B. (2023). Pour un tourisme durable : la stratégie de développement touristique dans les littoraux algériens for sustainable tourism : the tourism development strategy in the algerian coasts. *Revue Recherches et études en Développement*, 10(01) :606–619.

[Farchy and Méadel, 2013] Farchy, J. and Méadel, C. (2013). Moteurs de recherche et référencement : chassez le naturel... *Hermès*, (2) :147–154.

[Foenix-riou, 2022] Foenix-riou, B. (2022). Connaître le fonctionnement des moteurs de recherche web.

[Gondaliya and Joshi, 2017] Gondaliya, T. P. and Joshi, H. D. (2017). Journey of information retrieval to information retrieval tools-ir&irt a review.

[Habash, 2010] Habash, N. Y. (2010). *Introduction to Arabic natural language processing.* Morgan & Claypool Publishers.

[HADRI, 2017] HADRI, K. (2017). le tourisme durable.

[Hamza et al., 2022] Hamza, A., En-Nahnahi, N., El Mahdaouy, A., and Ouatik, S. E. A. (2022). Embedding arabic questions by feature-level fusion of word representations for questions classification : It is worth doing ? *Journal of King Saud University-Computer and Information Sciences*, 34(9) :6583–6594.

[Harman, 1996] Harman, D. (1996). The text retrieval conferences (trecs). In *TIPSTER TEXT PROGRAM PHASE II : Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996*, pages 373–410.

[Harman, 1998] Harman, D. (1998). The text retrieval conferences (trecs) and the cross-language track. In *LREC*, pages 517–522.

[Hazimeh et al., 2021] Hazimeh, H., Chreim, Z., Noureldine, A., Harb, H., Mugellini, E., Abou Khaled, O., and Hannoun, F. (2021). Reduce : a semi-supervised scalable approach for result duplication detection in search engines. *Procedia Computer Science*, 192 :893–902.

[Hélène, 2009] Hélène, M. (2009). Comment intégrer les populations locales au développement d'une structure touristique ?

[Hernandez, 2005] Hernandez, N. J. (2005). *Ontologies de domaine pour la modélisation du contexte en recherche d'information.* PhD thesis, Université Paul Sabatier-Toulouse III.

[Ingwersen, 1994] Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities elements of a cognitive theory for information retrieval interaction : Elements of a cognitive theory for information retrieval interaction. In *SIGIR'94 : Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 101–110. Springer.

[Kılıc and Aksakalli, 2016] Kılıc, U. and Aksakalli, K. (2016). Comparison of solr and elasticsearch among popular full text search engines and their security analysis. In *Future Internet of Things and Cloud Workshops, 2015 6th International Conference on*, pages 163–168.

[Le Ny, 2013] Le Ny, Y. (2013). Un moteur de recherche multilingue (choix, paramétrage et industrialisation). Master's thesis.

[Mars, 2016] Mars, A. (2016). *Le TAL au service desenseignants des langues : mise en oeuvre d'une plate-forme pour l'enseignement du français et de l'arabe, langues étrangères.* Theses, Université Grenoble Alpes.

[Menegon et al., 2009] Menegon, D., Mizzaro, S., Nazzi, E., Vassena, L., et al. (2009). Benchmark evaluation of context-aware web search. In *CIRSE 2009 : ECIR 2009 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation*, pages 30–33.

[Ndiaye, 2023] Ndiaye, A. (2023). Etude conception et réalisation d'une plateforme pour la gestion, le suivi et l'archivage des mémoires : le cas de l'ufr ses de l'uasz.

[Rady et al., 2022] Rady, G. S., Mohamed, S. S., Mohamed, M. F., and Hussain, K. F. (2022). High dimensional autonomous computing on arabic language classification. *Computers and Electrical Engineering*, 100 :108020.

[Robertson and Jones, 1976] Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3) :129–146.

[Sahmoudi and Lachkar, 2016] Sahmoudi, I. and Lachkar, A. (2016). Formal concept analysis for arabic web search.

[Salton et al., 1966] Salton, G. et al. (1966). The smart system–retrieval results and future plans. *Information Storage and Retrieval*, pages 1–9.

[Sanderson et al., 2010] Sanderson, M. et al. (2010). Test collection based evaluation of infor-

mation retrieval systems. *Foundations and Trends® in Information Retrieval*, 4(4) :247–375.

[Sayed and Al Muqrishi, 2017] Sayed, A. and Al Muqrishi, A. (2017). Ibri-casonto : Ontology-based semantic search engine. *Egyptian Informatics Journal*, 18(3) :181–192.

[Sicard, 2010] Sicard, J. (2010). *Contributions à la modélisation, évaluation et conception de systèmes de recherche d'information en pair-à-pair*. PhD thesis, Evry, Institut national des télécommunications.

[Tamine and Calabretto, 2008] Tamine, L. and Calabretto, S. (2008). Recherche d'information contextuelle et web. *Recherche d'information : état des lieux et perspectives*, pages 201–224.

[Tsikrika et al., 2013] Tsikrika, T., Larsen, B., Müller, H., Endrullis, S., and Rahm, E. (2013). The scholarly impact of clef (2000–2009). In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization : 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings 4*, pages 1–12. Springer.

[Unies, 2003] Unies, N. (2003). Conseil économique et social. *Commission économique pour l'Afrique*.

[Vitali, 2006] Vitali, X. (2006). Problématique de la recherche textuelle et intranet qualité.

[Voorhees and Harman, 2005] Voorhees, E. M. and Harman, D. K. (2005). The text retrieval conference. *TREC : Experiment and evaluation in information retrieval*, pages 3–19.

[Zargayouna, 2005] Zargayouna, H. (2005). *Indexation sémantique de documents XML*. PhD thesis, Paris 11.

[Zhao et al., 2012] Zhao, L., Liu, X., and Callan, J. (2012). Wikiquery-an interactive collaboration interface for creating, storing and sharing effective cnf queries. In *OSIR@ SIGIR*, pages 1–8.