الجمهورية الجزائرية الديمقراطية الشعبية

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE YAHIA FARES DE MEDEA

FACULTE DES SCIENCES

Département Mathématique et Informatique



Projet de fin d'études

Présenté pour l'obtention du Diplôme de

Master en Informatique

Option : Ingénierie des Systèmes et Technologie du Web

THEME

La création d'une collection de test Arabe pour évaluer des algorithmes et des techniques des moteurs de recherche

Proposé et dirigé par :

Mr. Mazari A.C

Préparé et présenté par :

Melle. Ahmed Yousfi Wasna
Mr. Hayi Mohamed Yassine

Année Universitaire 2017-2018

Résumé

La langue arabe est une langue riche, mais cette langue n'a pas eu sa chance dans les recherches et le développement des outils informatiques, pour cela nous avons fait ce projet qui représente la création d'une collection de test arabe pour évaluer les techniques et les algorithmes des moteurs de recherche.

Notre travail est subdivisé en trois parties, la première partie représente la création du corpus bilingue (arabe et anglais), en s'appuyant sur la bibliothèque électronique mondial (Proquest).

La deuxième partie présente le calcul de la pertinence entre les documents et les requêtes par l'utilisation des trois moteurs de recherche open source plus une validation par un expert (être humain).

La dernière partie présente l'étude de la pertinence par la technique de la pertinence et par l'utilisation d'algorithme de Naïve Bayés, nous avons fait deux méthodes d'apprentissage dans cette troisième partie.

Les résultats finals de notre travail sont représentés par la création d'une collection qui offre aux utilisateurs de faire des évaluations sur les techniques et les algorithmes des systèmes de recherche d'information.

Aussi nous avons créé un site web qui offre aux utilisateurs de faire des recherche sur des projets de fin d'études, ou même pour avoir les informations sur la collections ou la télécharger et l'utilisé dans d'autre projets.

Mots Clés : Collection, recherche d'information, moteur de recherche, langue arabe, langue anglaise, bilingue, Naïve bayés, Python.

ملخص

اللغة العربية هي لغة ثرية، لكن هذه اللغة لم تتح لها فرصة في البحث والتطوير في أدوات الكمبيوتر، ولهذا قمنا بهذا المشروع الذي يمثل إنشاء مجموعة اختبار لتقييم التقنيات وخوارزميات محرك البحث.

ينقسم عملنا إلى ثلاثة أجزاء، الجزء الأول يمثل إنشاء مجموعة ثنائية اللغة (العربية والإنجليزية)، (Proquest) بالاعتماد على المكتبة الإلكترونية العالمية.

يقدم الجزء الثاني حساب العلاقة بين المستندات والطلبات عن طريق استخدام محركات البحث الثلاثة المفتوحة المصدر بالإضافة إلى التحقق من صحة العلاقة من طرف خبير (كائن بشري).

تتمثل النتائج النهائية لعملنا في إنشاء مجموعة تقدم للمستخدمين إجراء تقييمات على تقنيات وخوارزميات أنظمة استرجاع المعلومات لذلك أنشأنا موقعًا إلكترونيًا يتيح للمستخدمين إجراء أبحاث حول مشروعات نهاية الدراسة، أو حتى الحصول على معلومات حول المجموعة أو تنزيلها واستخدامها في مشاريع أخرى.

كلمات مفتاحية: اللغة العربية، اللغة الإنجليزية، محرك البحث، قاعدة بيانات، ثنائي اللغة، استرجاع المعلومات، بايثون، Naive Bayse.

Abstract

The Arabic language is a rich language, but this language has not had a chance in the research and development of computer tools, for this, we did this project, which represents the creation of an Arab test collection to evaluate the techniques and search engine algorithms.

Our work is divided into three parts, the first part represents the creation of the bilingual corpus (Arabic and English), relying on the global electronic library (Proquest).

The second part presents the computation of the relevance between the documents and the requests by the use of the three open source search engines plus a validation by an expert (human being).

The last part presents the study of the relevance by the technique of the relevance and by the use of algorithm of Naïve Bayes; we made two methods of learning in this third part.

The results of our work represented by the creation of a collection that offers users to make assessments on the techniques and algorithms of information retrieval systems.

Aussi nous avons créé un site web qui offre aux utilisateurs de faire des recherche sur des projets de fin d'études, ou même pour avoir les informations sur la collections ou la télécharger et l'utilisé dans d'autre projets.

Mots Clés : Collection, recherche d'information, moteur de recherche, langue arabe, langue anglaise, bilingue, Naïve bayés, Python.

Remerciements

Nous remercions d'abord le bon Dieu qui nous a aidés et qui nous a donné le courage et la patience pour réaliser ce modeste travail.

Notre profonde gratitude et sincère remerciement vont à notre promoteur Mr.Mazari A.C de nous avoir encadré, pour son suivi et son orientation.
Nous tenons également, a lui notre profonde reconnaissance pour le temps précieux qu'il nous a consacré, ses remarques pertinentes et précieuses, ses conseils et ses encouragements.

Nos remerciements vont aussi aux membres de jury pour avoir accepté de juger ce travail.

Nous tenons à remercier nos enseignants qui nous ont transmis le savoir et le bien.

Nous tenons à remercier également tous ceux qui nous ont aidé et contribué de près ou de loin à la réalisation de notre projet.

Dédicaces

Merci mon dieu de m'avoir donné la capacité d'écrire, de réfléchir et la force d'y croire,

Je dédie ce modeste travail à :

Notre promoteur monsieur Mazari A.C qui nous a donné tout ce qu'il peut, nous avons appris beaucoup de chose avec lui, nous le remercie pour ces efforts

Mes chers parents qui représentent mes deux yeux.

Mes deux adorables frères

Abdelmadjid et Mohamed Walid.

Toute ma famille paternelle et maternelle.

Mon formidable cousin Bachir.

Mon Binôme Ouàsnàà : elle a été sérieuse toute la période de la réalisation du projet, je te remercie pour tes efforts.

Tous mes enseignants en université Dr Yahia Farés à Médéa Et tous mes amis(es) et surtout promo Master ISTW 2017/2018

Hayi Mohamed Yassine

Je dédie ce travail à ceux qui me portent

dans leurs coeurs et m'accompagnent dans leurs pensées.

A ceux dont personne ne peut remplacer les sacrifices

qu'ils ont consentis pour mon éducation et

mon bien être

Mes très chers parents

A mes très cheres soeurs et leurs maries et leurs bébés et mon très chèr

frére Tamim, et toute ma famille sans exception

A mon partenaire yassine qui a fait un excellent travail

et aux membres de notre promotion

A tous mes ami(e)s sans exception et toute personne qui de près ou de

loin nous a aidé et a participé à réaliser notre projet

et qui nous ont encouragé tout au long de notre

formation

Wesnaà

Table de Matières

Introduction générale 01
Problématique et objectif 02
Organisation du mémoire..... 03

Chapitre -I- : Etat de l'art

1. Introduction ……..….04

2. Moteur de recherche ……05

2.1. Définition des moteurs de recherche ……...……....05

2.2. Composition des moteurs de recherche ……...…….....06

2.2.1. Base de données …….……..06

2.2.2. Spider ……06

2.2.3. Crawler L'exploration (crawling) ……06

2.2.4. L'indexeur ……06

2.2.5. L'interrogateur …….....07

2.3. Requête …….……07

2.4. Fonctionnalité d'un moteur de recherche …….……07

2.4.1. Recherche avancée ……..07

2.4.2. Suggestion ……08

2.4.3. Classement ……08

3. Système de recherche d'information (SRI) ……..……**09**

3.1. Définition de SRI …….……....09

3.2. Objectif de SRI …….09

3.3. Architecture générale des SRI ……..10

4. Recherche d'information ……...**13**

4.1. Définition de la recherche d'information ……13

4.2. Modelés de la recherche d'information ……13

4.2.1. Modèle booléen ……13

4.2.2. Modèle vectoriel ……..……13

4.2.3. Modèle probabiliste ……..……..……15

4.3. Concept de la recherche d'information ……..……..……15

4.4. Appariement requête-document ……..……..…… ..16

5. Collections ……...16

5.1. Définition d'une collection de documents ……16
5.2. Collections connues en Arabe ……..……..…… ….17
5.2.1. Arabic Concordancer ……...17
5.2.2. Blog de la langue arabe d'université du roi Abdulaziz ……....17
5.3. Composition d'une collection ……..18
6. Conclusion ……19

Chapitre -II- : Les travaux Connexes

1. Introduction ……20
2. TREC ……21
2.1. Définition du TREC ……..21
2.2. Historique du TREC ……..22
2.3. Objectifs de TREC ……22
2.4. Fonctionnement de TREC ……23
3. CLEF ……24
3.1. Définition du CLEF ……..24
3.2. Objectifs du CLEF ……26
3.3. Structure de CLEF ……....27
4. Travaux connexe ……27
5. Moteurs de recherche ……29
5.1. Avantages des moteurs de recherche ……...29
5.2. Inconvénients des moteurs de recherche ……..29
5.3. Meilleures moteurs de recherche au monde ……30
6. Mesures de performance ……...31
6.1. Rappel ……..32
6.2. Précision ……..32
6.3. F-Mesure ……32
6.4. Bruit et silence ……33
6.5. Relation entre rappel et précision ……33
7. Conclusion ……..34

Chapitre -III- : L'approche proposée

1. Introduction ……35

2. Description de l'approche ……..36

2.1. Plan global du travail ……....36

2.2. Plan détaillé du travail ……..37

3. Construction de la collection ……....38

3.1. Récupération des documents ……38

3.2. Construction de la base ……40

3.3. Extraction des requêtes ……44

4. Récupération de la pertinence à travers des moteurs de recherche«Open Source » ……..46

4.1. Validation par les trois moteurs de recherche ……..47

4.2. Regroupement des résultats par l'utilisation de l'algorithme de «Balanced Interleaving » .48

4.3. Validation de l'expert ……...49

5. Validation des résultats par l'utilisation de la technique d'AA ……50

5.1. Extraction des mots non vide ……...51

5.2. Calcul du nombre d'existence de chaque mot dans chaque document ……52

5.3. Choix de la partie de modèle et la partie de test ……..53

5.4. Calcul d'apprentissage par l'algorithme de naïve bayésienne ……54

6. Choix des deux langues (arabe et anglais) ……...56

6.1. Langue Arabe ……...56

6.2. Langue Anglaise ……..56

7. Différents types de la Collection ……...57

8. Conclusion ……..58

Chapitre -IV- : Implémentation et Evaluation

1. Introduction ……59

2. Différentes ressources utilisées ……60

2.1. Python ……..60

2.1.1. Définition de Python ……60

2.1.2. Bibliothèques de Python utilisés ……..61

2.2. NetBeans ……..62

2.3. Moteurs de recherche utilisée ……...63

2.3.1. Lucene ……..63

2.3.2. Solr ……...64

2.3.3. Hibernate ……..65

2.4. OCR « La reconnaissance optique de caractères » ……..66

3. Présentation du site web ……67

4. Résultats de l'expérimentation ……...70

4.1. Résultat de pertinence par les moteurs de recherche ……70

4.2. Résultats d'apprentissage automatique ……72

4.3. Comparaison entre les résultats de pertinence des moteurs de recherche et d'AA ……..75

5. Conclusion ……...79

Conclusion et perspectives 80
Annexe 1 .....……82
Annexe 2 86
Bibliographie 89

Tables des figures

FIGURE 1.1 : Suggestion des requêtes de recherche ……8

FIGURE 1.2 : Processus en U de recherche d'information ……11

FIGURE 1.3 : Différentes mesures de similarité ……14

FIGURE 1.4 : Pourcentage de distribution des textes par rapport à la source ……18

FIGURE 2.1 : Déférente procédures de TREC entre 1992 et 2006 ……22

FIGURE 2.2 : Graphe de nombres des participants de CLEF par rapport aux continents……25

FIGURE 2.3 : Groupes participants du CLEF entre 2000 ET 2008 ……25

FIGURE 2.4 : Pourcentage d'utilisation des meilleurs moteurs de recherche ……30

FIGURE 3.1 : Plan global de l'approche proposée ……36

FIGURE 3.2 : Plan détaillé qui explique les déférentes étapes de notre approche ……37

FIGURE 3.3 : Page de recherche de la bibliothèque électronique Proquest ……39

FIGURE 3.4 : Pourcentage des documents par apport au champ Under Topic……41

FIGURE 3.5 : Pourcentage des documents par apport au champ de Topic ……42

FIGURE 3.6 : Pourcentage des documents par apport au pays de l'université ……42

FIGURE 3.7 : Pourcentage des documents par apport au champ d'année ……43

FIGURE 3.8 : Code source python qui fait la traduction ……44

FIGURE 3.9 : Pourcentage des requêtes par rapport au nombre des mots de la requête ……46

FIGURE 3.10 : Algorithme de «Balanced interleaving» ……49

FIGURE 3.11 : Code source du programme qui retourne la liste des mots ……53

FIGURE 4.1 : Interface d'un programme Java sous NetBeans ……62

FIGURE 4.2 : Interface du moteur de recherche lucene ……63

FIGURE 4.3 : Interface du moteur de recherche Solr ……..65

FIGURE 4.4 : Page d'accueil du site free Online OCR ……66

FIGURE 4.5 : Fonctionnement de free Online OCR ……67

FIGURE 4.6 : Page de recherche du site web réalisé ……..68

FIGURE 4.7 : Page de résultat de la recherche en Arabe ……..68

FIGURE 4.8 : Page de résultat de la recherche en Anglais ……..69

FIGURE 4.9 : Page de statistique sur notre collection ……69

FIGURE 4.10 : Fonctionnement d'algorithme de «Balanced interleaving» ……..70

FIGURE 4.11 : Courbe de la précision de la requête Q001 par la deuxième méthode d'AA ……73

FIGURE 4.12 : Courbe des moyennes de précision des 600 documents ……74

FIGURE 4.13 : Histogramme sur les différentes catégories des requêtes ……..78

Tables des tableaux

TABLEAU 2.1 : Quelques informations sur les différentes collections du CLEF 26

TABLEAU 2.2 : Travaux connexe à notre PFE 27

TABLEAU 3.1 : Les différentes informations récupérées pour chaque document 40

TABLEAU 3.2 : Tableau des Topic et Under Topic 41

TABLEAU 3.3 : Tableau des requêtes 45

TABLEAU 3.4 : Nombre d'existence des requêtes par rapport au nombre des mots de la requête 45

TABLEAU 3.5 : Comparaison entre les trois moteurs de recherche utilisé 47

TABLEAU 3.6 : Explication des calculs des degrés de pertinences 48

TABLEAU 3.7 : Différence entre les deux méthodes d'apprentissage 51

TABLEAU 3.8 : Matrice de nombre d'existence de chaque mot par rapport au chaque document 52

TABLEAU 4.1 : Déférentes bibliothèques de Python utilisé 61

TABLEAU 4.2 : Explication du calcul de la pertinence 71

TABLEAU 4.3 : Matrice finale de pertinence 72

TABLEAU 4.4 : Matrice de pertinence du 1ere méthode 72

TABLEAU 4.5 : Comparaison entre le choix des parties tests dans les deux méthodes d'AA 75

TABLEAU 4.6 : Moyenne des mesures de performance 76

TABLEAU 4.7 : Comparaison entre quelque requête par apport aux mesures de performance 76

TABLEAU 4.8 : Tableau de comparaison entre les déférentes catégories des requêtes 77

Liste des acronyms

AA :	Aprentissage Automatique
BD :	Base Documentaire
BDD :	Base De Donnée
CLEF :	Cross-Language Evaluation Forums
NIST :	National Institute of Science and Technology
OCR :	Optical Character Recognition
PFE :	Projet de Fin d'Etude
RI :	Recherche d'Information
SRI :	Système de Recherche d'Information
T :	Topic
TREC :	Text REtrieval Conference
UT :	Under Topic
XML :	Extensible Marcup Language

Face à l'évolution rapide dans l'information et la communication, une masse importante d'information est produite tous les jours à travers des milliers de textes et d'articles de journaux.

La recherche d'information a facilité la vie de l'être humain, dès qu'une question se pose, nous tapons et nous trouvons des milliers et des milliers de réponses de différents types (textes, photos, graphes, vidéos … etc.) mais tous ces informations ne sont pas bien utilisé car derrières tous ces informations que se trouve sur le web nous avons un grand trésor.

Ce trésor d'informations doit être géré et manipulé par des moteurs de recherche, il existe beaucoup de moteurs de recherche de différents types et la différence qui caractérise les moteurs de recherche est la technique utilisée pour rechercher et récupérer de l'information.

Notre projet combine Entre l'évolution rapide dans l'information, la recherche d'information et les moteurs de recherche qui gère ce trésor d'information, notre travail consiste en deux langues l'arabe et l'anglais.

Dans ce mémoire de Master nous présentons notre approche de la création d'une collection de documents bilingues pour la recherche d'information qui offre la possibilité de tester et évaluer des différents algorithmes et modèles de RI, cette collection est similaire avec les collections proposées par les conférences organisées par TREC (Text REtrieval Conference) , TREC offre chaque année aux chercheurs du monde ces collections pour valider et évaluer leurs nouveaux algorithmes et nouveaux modèles en différentes langues et en différents formats.

Problématique

La recherche d'information en général et les moteur de recherche en particulier soufre de quelque point négative comme le nombre des documents retourné (trop d'informations retourné) et l'utilisateur ne peut pas lire tous les fichiers, au pire des cas il va consulter les 20 premiers documents, pour cela le classement des fichiers pertinent a un rôle fondamental et crucial pour juger la qualité d'un moteur de recherche.

Aussi la pertinence des documents par apport a une requête donnée, car si un moteur de recherche donne des documents non pertinents, tous simplement il va y'avoir une mauvaise réputation.

Objectif

Puisqu'il y a un manque dans les collections en arabe, nous avons décidé comme un premier objectif de crée une collection en arabe pour la rendre utilisable pour les chercheurs et les étudiants qui veulent faire leurs études sur la recherche d'information en arabe.

Comme un deuxième objectif nous voulons faire une comparaison entre l'étude de la pertinence entre l'utilisation des moteurs de recherche open source avec l'utilisation de l'apprentissage automatique par l'algorithme de naïve Bayes.

Notre dernier objectif est la création d'un outil qui offre aux utilisateurs la recherche des thèmes de projets de fin d'étude et les télécharger, la recherche se faite en deux langue arabe et anglais et par des requêtes prédéfinit dans les différentes domaines et les différentes universités arabes.

Ce présent mémoire est organisé en quatre chapitres comme suit ; Le premier chapitre est sur l'état de l'art, dans lequel nous expliquons les différentes composantes de la recherche d'information et aussi nous discutons sur les systèmes de la recherche d'information et les moteurs de recherche.
Dans le deuxième chapitre titré Travaux Connexes, nous présentons les différents travaux qui ont de relation avec notre travail, Aussi les meilleures collections connues au monde, et les meilleurs moteurs de recherche classée au monde.

Le troisième chapitre présente notre approche proposée pour la création de la collection, d'où nous allons expliquer en détail le plan de notre travail, depuis la récupération des documents vers la création de l'outil de la recherche des projets de fin d'études.

Au quatrième chapitre « Implémentation et Evaluation » nous présentons les différentes ressources utilisées dans le projet, et aussi l'outil qui offre la recherche sur notre collection, et finalement nous présenterons et nous expliquerons les différents résultats obtenus dans le travail.

CHAPITRE I :
Etat de l’art

1. Introduction

La recherche d'information (RI) est une branche de la technologie d'information, son but est de chercher les documents qui satisfont les besoins d'utilisateur dans une grande base documentaire. Les informations sont stockées au format électronique, il devient plus difficile de retrouver un ensemble d'informations au sein d'une base volumineuse.

Un système de RI permet à l'utilisateur de formuler sa demande et elle doit répondre le plus exactement possible à son besoins d'information.

La recherche d'information a connu un énorme développement, depuis les années soixante ou la recherche d'information représente seulement dans les simples recherches booléennes avec un géant ordinateur qui se trouve dans une grande salle.

Jusqu'à maintenant nous trouvons des systèmes d'informatique qui gère tous, les bibliothèques, les vidéos et des milliers d'informations, sans oublier les systèmes questions réponses qui donne des résultats correct a cent pour cent surtout ceux que nous trouvons dans les grandes usines ou les stations nucléaires.

Nous présentons dans ce chapitre les déférentes périodes du développement de la recherche d'information et des moteurs de recherche et la relation entre les deux, sans oublier de présenter le fonctionnement des moteurs de recherche par ces différentes étapes.

Pour la deuxième partie nous passons aux détails et aux composantes de la recherche d'information, cette dernière a besoin d'une collection pour qu'elle fonctionne, et la collections elle-même se compose d'un corpus et des requêtes et une table de pertinence.

Nous terminons par le besoin d'information et comment étudier la pertinence, sans oublier que ce chapitre contient les déférentes définitions des termes qui ont une relation avec notre projet de fin d'étude.

2. Moteur de recherche

2.1. Définition des moteurs de recherche

Un moteur de recherche est un outil qui fait la recherche sur le web, à l'aide des « robots », qui visite plusieurs sites automatiquement, chaque page visité va être indexé dans la BDD, pour qu'elle soit accessible par des mots-clés dans la partie d'interrogatoire.

Donc un moteur de recherche est un programme informatique qui contient un ou plusieurs algorithmes, ce moteur travaille à effectuer une recherche sur une base de données, les moteur de recherche les plus connus comme Google, Yahoo, AltaVista et Fast récupèrent des informations qui se trouve sur le web par des robots ou des spiders.

La déférence entre les moteurs de recherche consiste dans ses propres techniques et critères qui déterminent la méthode de recherche et l'ordre des documents pertinents retournés, Alors nous ne trouvons jamais des moteur de recherche pareil et qui donne les mêmes résultats et dans le même classement.
Le mot Recherche Informatique (Information Retreival) fut utilisé pour la première fois par « Carlin N Mooers » en 1948 dans son mur de maitrise.

La RI généralement, a l'objectif d'extraire d'un document ou d'un ensemble de documents les informations pertinentes qui réponds aux besoins d'utilisateur.

L'accès à l'information demandé par rapport à l'utilisateur est effectué de manière souple à travers un système de recherche informatique [**Boucham, 2009**].

2.2. Composition des moteurs de recherche

Un moteur de recherche n'est pas seulement un petit programme qui récupère des informations, derrière ce petit programme, il existe plusieurs compositeur qui caractérise les différent type et catégorie des moteurs de recherche, donc les compositeurs des moteurs de recherche diffère d'un moteur de recherche à l'autre, En général les moteurs de recherche sont composés par :

2.2.1. Base de données

Un entrepôt ou se trouve les informations à chercher, chaque moteur de recherche fait ces recherches sur sa base de donnée, cette dernière est remplie par le spider que nous allons l'expliqué après, et il existe des moteurs de recherche qui cherchent directement sur le web.

2.2.2. Spider

Un robot ou un programme qui récupère des codes source des pages web, le spider peut récupérer les différentes ressources (pages web, images, vidéos, documents Word, PDF ou PostScript, etc.), Aussi par le spider les moteurs remplissent leurs bases.

2.2.3. Crawler L'exploration (crawling)

Un programme qui collecte les pages web récupérés, crawler provient du nom du tout premier moteur de recherche sur Internet (le Web Crawler).

2.2.4. L'indexeur

L'indexeur est un élément important, Un algorithme qui analyse, indexe et traite les pages, l'indexeur décompose les pages en ses diverses zones comme les titres, les entêtes, les liens... Etc.

2.2.5. L'interrogateur

L'algorithme qui fait la recherche sur la BDD à partir d'une requête donnée par le client et qui base sur Les critères de sélection pour récupère les meilleurs résultats dans le meilleur ordre.

2.3. Requête

La requête joue un rôle important sur les résultats retourné, car coté utilisateur, il doit écrire une requête qui vraiment a un sens avec son but de recherche, et sur le côté du moteur de recherche, la requête doit être bien analysé par la suppression des mots vide et par l'ajout des synonymes et le poids de chaque mot de la requête pour qu'ils nous facilitent la recherche et pour obtenir les meilleures résultats.

2.4. Fonctionnalité d'un moteur de recherche

Les moteurs de recherche doivent faire des calculs compliqués et utiliser des techniques et des méthodes pour que l'utilisateur soit satisfait, et les fonctionnalités les plus connue des moteurs de recherche sont :

2.4.1. Recherche avancée

C'est-à-dire le moteur de recherche donne la possibilité au utilisateur de chercher par un mot, une phrase, une partie d'un mot ou même par les relations logiques (et, ou, non) … etc.

2.4.2. Suggestion

Cette fonctionnalité facilitée la recherche pour l'utilisateur, par exemple quand il écrit « Univ de M » il reçoit quelque Suggestion des requêtes déjà faite par d'autre utilisateurs, la figure ci-dessous représente une simple explication sur la suggestion des requêtes de recherche.

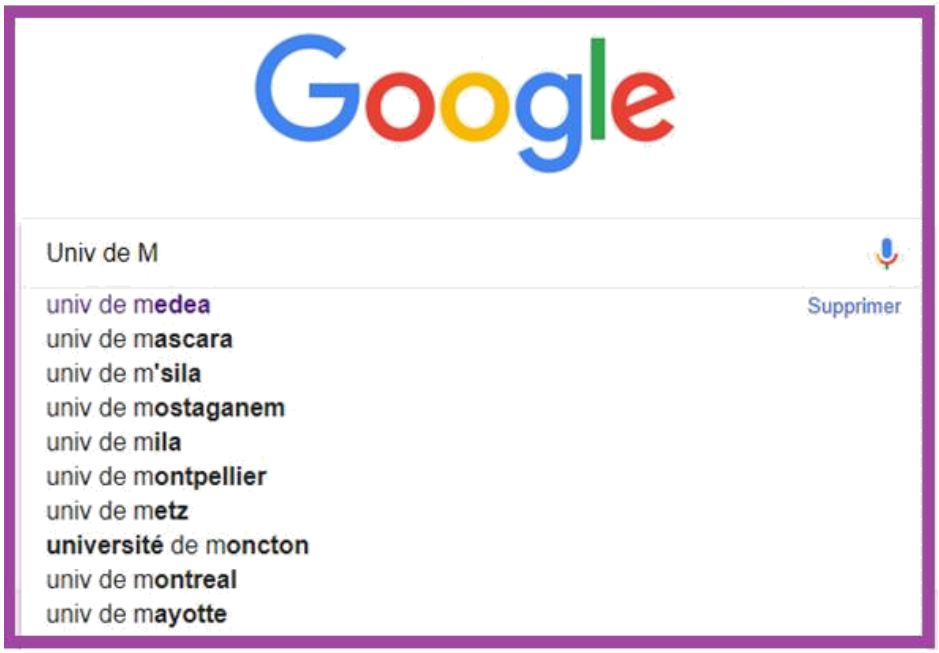


FIGURE 1.1 : Suggestion des requêtes de recherche

2.4.3. Classement

Chaque moteur de recherche utilise ces propres algorithmes et ces propres calculs seulement pour calculer précisément la pertinence des documents, mais aussi pour faire ensuite le classement des documents car le 1er document retourné n'a pas la même valeur que le 10ème document retourné par exemple. [Hacene et Taleb, 2015]

3. Système de recherche d'information (SRI)

3.1. Définition de SRI

Il existe plusieurs définitions d'un SRI, qui sont plus au moins proches :

Tomek Strzalkowski définit un SRI comme suit :
« La tâche typique d'un système de recherche d'information est de sélectionner des documents dans une base de données, en réponse à une requête de l'usager, et leur rangement par ordre de pertinence » [**Strzalkowski, 1993**].

Tandis qu'Alan Smeaton donne la définition suivante :
Le but d'un système de recherche d'information est de retrouver des documents en réponse à une requête des usagers, de manière à ce que les contenus des documents soient pertinents au besoin initial d'information de l'usager [**Smeaton, 1989**].

Un système de recherche d'information peut être vu comme un moyen ou un outils de prédiction de la pertinence des documents pour l'utilisateur, donc l'objectif de SRI est de chercher et trouver parmi une base de documents, les documents pertinents ou bien les documents qui répondant au besoin d'utilisateur qui se trouve dans la requête.

3.2. Objectif de SRI

Le système de recherche d'information ne consiste pas seulement à trouver des documents pertinents, il aussi d'autres buts qu'il doit les atteindre, ces buts ou bien les objectifs sont les points qui caractérisent les SRI et qui donne la déférence entre les milliers des SRI qui existe dans le monde.

- La bonne formulation de la requête soit par l'utilisateur ou bien par le système lui-même par l'ajout des synonymes et des mots proches de la requête.

- Le temps entre l'envoi d'une requête et la présentation de la réponse.
- Le bon style de la présentation des résultats qui influence l'utilisateur dans sa motivation à consulter les informations retrouvées.
- Le contenu du corpus et les informations qui contient car une bonne information attire l'utilisateur.
- La fiabilité et la bonne source d'information pour gagner la confiance de l'utilisateur.
- La capacité du système à identifier uniquement les informations pertinentes et à éliminer les autres.

3.3. Architecture générale des SRI

Le besoin d'un utilisateur exprimé par une requête, la notion de pertinence de cette requête peut être appréhendée à deux niveaux :

- *Niveau utilisateur* : la pertinence correspond à la satisfaction de l'utilisateur de l'ensemble des documents retournée par le SRI.
- *Niveau système* : le système mesure un degré de pertinence, une valeur de similitude entre un document et une requête.

Donc le SRI fait à rapprocher la pertinence système de la pertinence utilisateur, pour cela le SRI doit réaliser l'indexation des documents, la reformulation de la requête pour qu'elle soit plus détaillée mais cette partie n'est pas obligatoire, et à la fin la comparaison entre la requête et les documents par les différents modèles que nous allons l'expliquer prochainement.

Nous présentons schématiquement un SRI par la figure suivante qui représente le processus en U de recherche d'information.

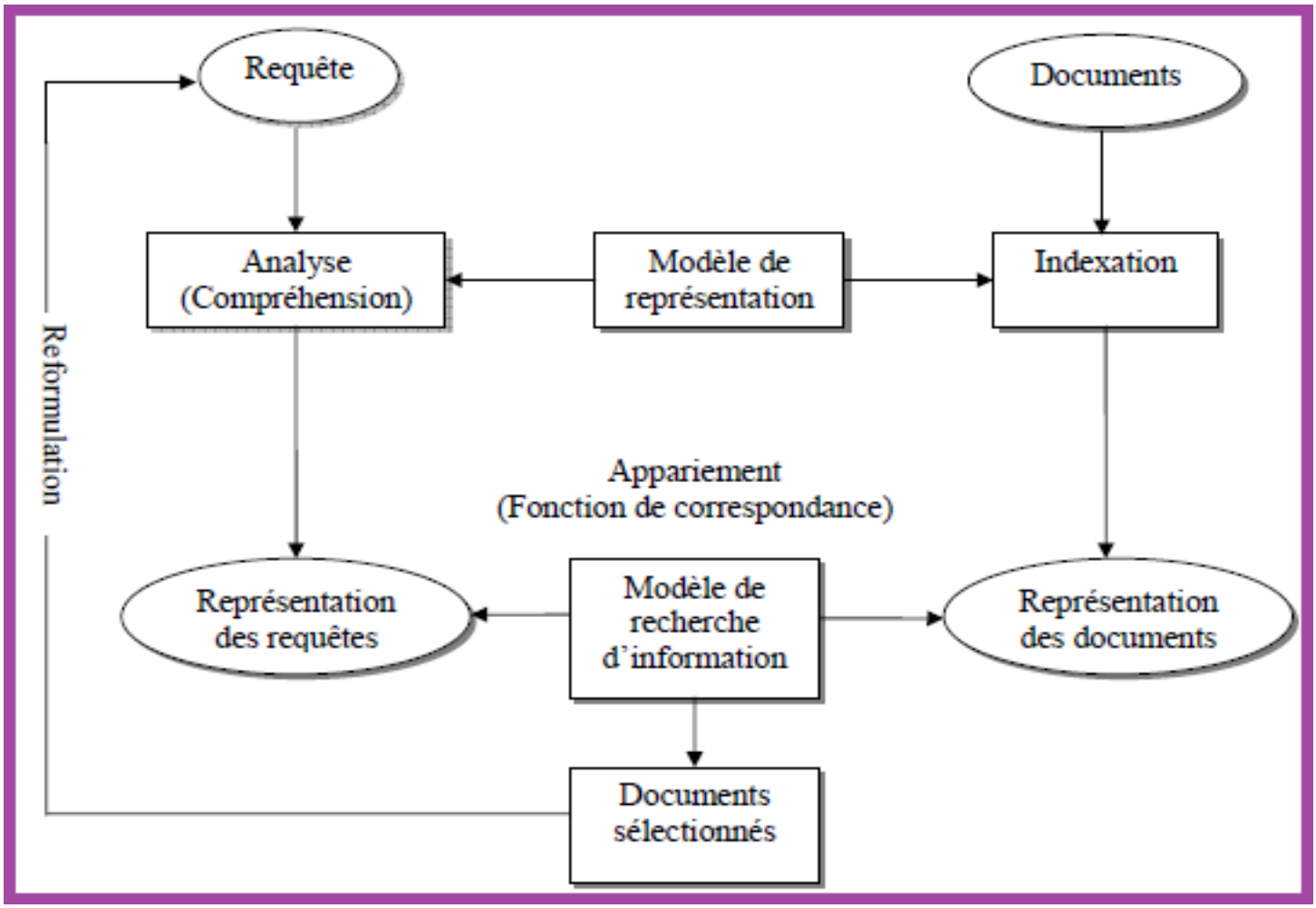


FIGURE 1.2 : Processus en U de recherche d'information

Du schéma précédent, nous pouvons dire que le système de recherche d'information se compose de deux processus, processus de représentation et processus de recherche d'information.
Modèle de représentation ou des fois nous le trouvons sous le nom indexeur, c'est un processus de représentation du contenu des textes d'où il permet d'extraire d'un document ou d'une requête.
Modèle de recherche d'information est un processus de comparaison entre les représentations des textes, issues du premier processus [**Boucham, 2009**]

D'après la figure précédente le processus de RI fonctionne comme suit :

- Un utilisateur fournit son besoin d'information sous la forme d'une requête.

- Le système indexe la requête et dans le même temps ou préalablement indexe la collection de documents.

- Grace à l'index, le système est en mesure de construire les représentations puis de mettre en correspondance la représentation de la requête avec les représentations des documents de la collection.

- Une liste des documents considérés par le SRI comme pertinents par rapport à la requête utilisateur

4. Recherche d'information

4.1. Définition de la recherche d'information

Il existe plusieurs définitions de la recherche information nous citons :

« 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 répondant aux besoins d'utilisateur » [**Kompaoré, 2008**]

4.2. Modelés de la recherche d'information

Un modèle de RI est le choix que le système de recherche d'information doit choisir pour obtenir les meilleures résultats, d'une autre façon c'est l'algorithme de calcul qui va nous donner les résultats, le SRI soit il se base sur un modèle ou plusieurs modèles et chaque modèle a plusieurs algorithmes, il existe trois catégories des modèle de RI :

4.2.1. Modèle booléen

« Le modèle booléen est le modèle le plus ancien et le plus simple de tous les modèles de la RI. Il est basé sur l'algèbre de Boole et la théorie des ensembles. Dans ce modèle chaque document et chaque requête sera représenté sous forme d'expression logique » **[Salton, 1971]**.

4.2.2. Modèle vectoriel

« Dans ce modèle la pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel. Un document de la collection est représenté sous forme d'un vecteur de termes pondérés » **[Salton, 1971]**.

Lorsque la collection contient T termes, un document Dj est défini par :

$$|\overrightarrow{D_j}| = (W_{1j}, \dots, W_{Tj})$$

Une requête est défini par :

$$|\vec{q}| = (W_{1q}, \dots, W_{Tq})$$

Avec :

W_{ij} : Le poids du terme ti dans le document Dj.

W_{iq} : Le poids du terme ti dans la requête Qq.

La fonction de correspondance mesure la similarité entre le vecteur requête et les vecteurs documents, il existe plusieurs mesures de similarité, la figure suivante donne quelque mesurés de similarité les plus connue **[Salton, 1983]**.

Inner product	$\|X \cap Y\|$	$\sum x_i * y_i$
Coef. de Dice	$\frac{2*\|X \cap Y\|}{\|X\|+\|Y\|}$	$\frac{2*\sum x_i * y_i}{\sum x_i^2 + \sum y_j^2}$
Mesure du cosinus	$\frac{\|X \cap Y\|}{\sqrt{\|X\|} * \sqrt{\|Y\|}}$	$\frac{\sum x_i * y_i}{\sqrt{\sum x_i^2 * \sum y_j^2}}$
Mesure du Jaccard	$\frac{\|X \cap Y\|}{\|X\|+\|Y\|-\|X \cap Y\|}$	$\frac{\sum x_i * y_i}{\sum x_i^2 + \sum y_j^2 - \sum x_i * y_i}$

FIGURE 1.3 : Différentes mesures de similarité

4.2.3. Modèle probabiliste

Ce modèle se base sur le calcul de la probabilité de pertinence d'un document par rapport à une requête, le principe principal consiste à trouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents.

Etant donne une requête utilisateur Q et un document D, la pertinence du document pour cette requête est mesurée par rapport à la probabilité que ce document soit pertinent pour cette requête, et la probabilité qu'il soit non pertinent [**Kadri, 2008**].

.

4.3. Concept de la recherche d'information

Chaque système de recherche d'information a un concept de représentation de la requête à chercher, et un mécanisme de la sélection des documents pertinent qui ce change d'un SRI a un autre [**Kompaoré, 2008**].

Et chaque Système de recherche d'information contient :

- Les documents
- La requête
- L'indexation (la représentation des documents)
- L'analyse (la représentation des requêtes)
- l'appariement requête-document

4.4. Appariement requête-document

Le noyau d'un système de recherche d'information représente L'appariement requête-document d'où cette appariement peut être exprimé dans une matrice (n * m) tel que n est le nombre des documents et m est le nombre des requêtes.

Chaque case de la matrice représente le degré de la pertinence (entre 0 et 1) ou bien elle peut contient seulement deux valeur (appartient ou appartient pas), le degré de pertinence peut être calculé par une mesure de pertinence comme la similarité RSV (requête, document) *RetrievalSattusValue.*

5. Collections

5.1. Définition d'une collection de documents

Il existe plusieurs définitions des collections de documents nous citons :

- ✓ La collection est un ensemble des documents de même source ou même type, elle contient des représentations simple pour que un simple utilisateur peut comprendre mais au même temps riche de l'information.

- ✓ Une collection (Aussi corpus ou fond documentaire) est un ensemble de document authentique (texte, page web, image, vidéo etc…) regroupés dans une optique précise selon un certain nombre de critères préalablement à tout travail linguistique et leur but est de déterminer la performance d'un système informatique.

5.2. Collections connues en Arabe

Les collections en Arabe n'existe pas trop par rapport au d'autres langues mais nous citons quelque collection connue dans le monde arabe :

5.2.1. Arabic Concordancer

Arabic Concordancer[1] est une collection de test en arabe qui consiste a récupéré ces textes depuis les projets de fin d'étude et les magazines, la collection a été créée en 2010 à l'université islamique internationale de Malaisie.

5.2.2. Blog de la langue arabe d'université du roi Abdulaziz

Blog de la langue arabe d'université du roi Abdulaziz[2] est une nouvelle collection de la part de l'université saoudienne du roi Abdulaziz de la science et la technologie, elle a été créée en 2018, elle contient 1 323 185 documents et plus de 9 millions mots.

La figure suivante représente pourcentage de distribution des textes par rapport à la source.

[1] http://efolio.iium.edu.my/arabicconcordancer/
[2] http://corpus.kacst.edu.sa

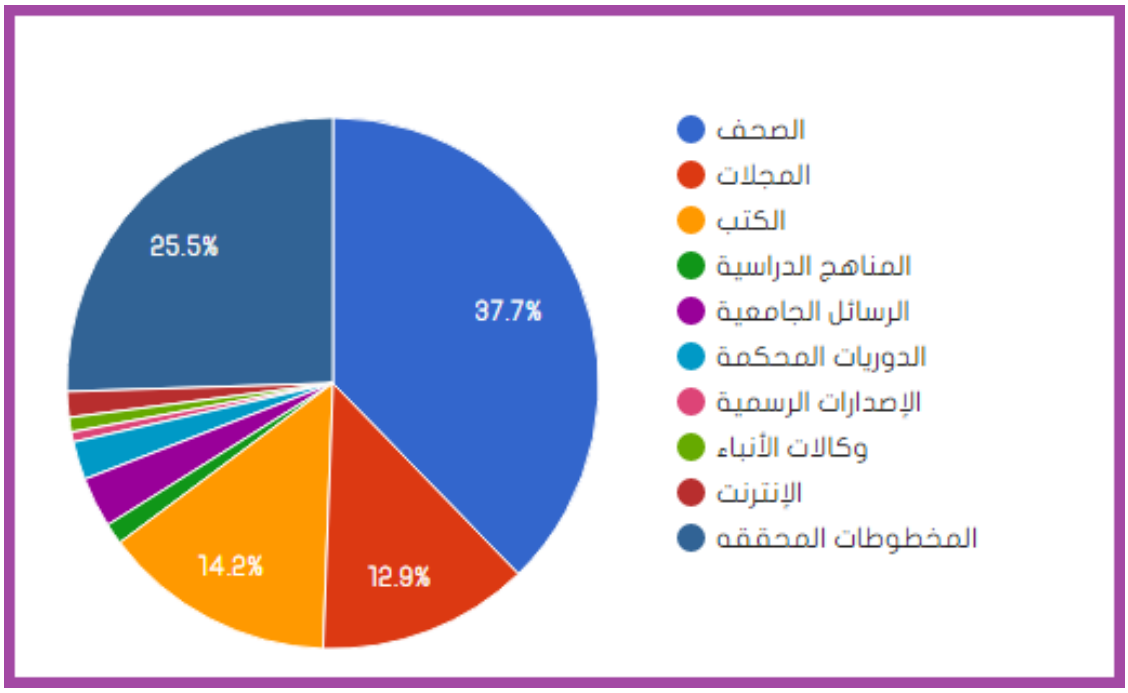


FIGURE 1.4 : Pourcentage de distribution des textes par rapport à la source

5.3. Composition d'une collection

Une collection de test comprend :

- corpus de documents : un ensemble des documents de même type et de même source de préférence.

- Liste de requêtes prédéfinies : Les requête sont nécessaire pour la collections, et de préférence la liste des requête doit avoir une relation avec les documents, par exemple si nous prenons des documents du domaine du sport, il faut avoir des requêtes sur le sport, leur but est de faire des tests et calculer les performances des systèmes et des algorithmes de la recherche d'information. Aussi la requête est l'expression du besoin en informations de l'utilisateur, c'est un ensemble de mots clés exprimés en langage naturel, booléen ou bien en langage graphique.

- Jugements de pertinence : dans une collection nous pouvons définir la pertinence des documents existe par apport à les requête proposer dans le corpus par une matrice où chaque case de cette matrice contient la relation entre une requête et un document (pertinent/ non pertinent) ou bien un degré de pertinence qui peut être donné entre zéro et un.

6. Conclusion

Dans ce chapitre nous avons expliqué les moteurs de recherche en, nous avons donné quelque définition sur les moteurs de recherche, et aussi nous avons présenté leurs différentes compositions, ensuite nous avons décrit le fonctionnement des moteurs de recherche.

Aussi, nous avons expliqué le système de la recherche d'information par la définition du SRI et son objectif et aussi par la discussion sur l'architecture des systèmes des recherches d'information.

Ensuite nous avons passé à la recherche d'information par la présentation des différents modèles de la recherche d'information : booléen, vectoriel et probabiliste, sans oublier que nous avons présenté le concept de la recherche d'information.

Finalement, nous avons donné quelque collection en arabe connue au monde et aussi les différents compositeurs d'une collection.

CHAPITRE II :
Travaux Connexes

1. Introduction

Dans ce chapitre, nous sommes intéressés à prendre des travaux qui ont une relation avec notre projet, le meilleur exemple qui est proche de notre PFE est la vaste collection de tests TREC qui a les même but de notre travail, mais TREC est une collections mondiale qui offre leur bases au chercheurs qui travaillent sur la recherche d'information.
Ensuite nous expliquons la collection CLEF qui a les mêmes caractéristiques que la collection TREC, nous parlons sur la définition et l'objectif et aussi ça structure.

Apres l'explication des exemples des travaux qui sont proche de notre travail, nous prenons quelque exemple sur les moteurs de recherche puisque notre travail est une combinaison entre la collection et le moteur de recherche. Les moteurs de recherche sont nécessaires dans la vie de l'être humain, surtout depuis quelque année ou les moteurs de recherche sont devenus l'oxygène des utilisateurs, car chacun de nous utilise les moteurs de recherche dans les différents domaines de la vie.

Nous voulons faire des définitions et donnons des historiques sur quelque moteurs de recherche, nous avons choisie de parler sur Google, Yahoo, Baidu et Bing.

Finalement nous parlons sur les mesures d'évaluation, comment nous les calculons, et nous discutons en détail sur les trois mesures les plus connue

- ✓ Rappel
- ✓ Précision
- ✓ Mesure harmonique (F-mesure)

Aussi nous donnons les déférentes formules des mesures d'évaluation.

2. Collection de TREC

2.1. Définition du TREC

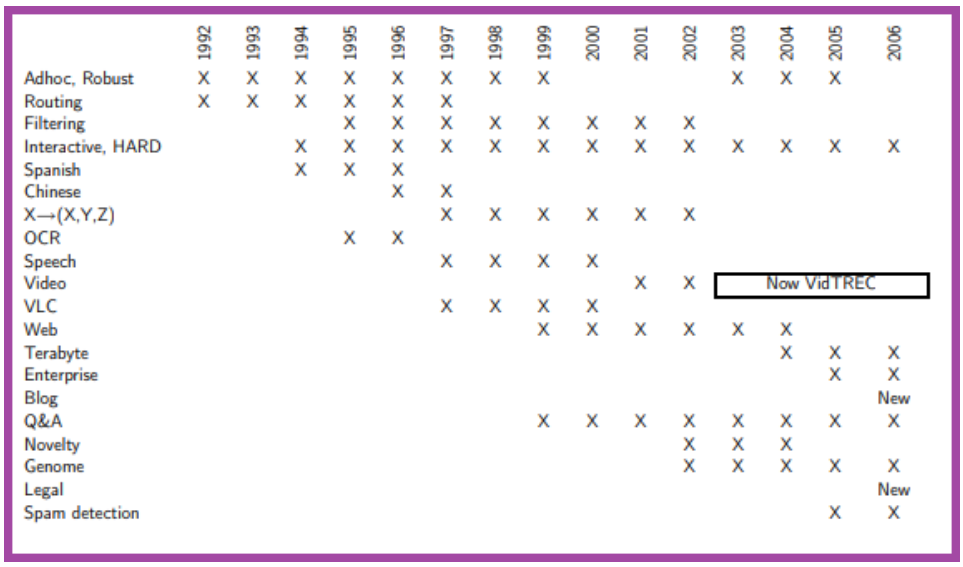
TREC[1] (Text Retrieval Conference) est une conférence d'évaluation qui encourage les chercheurs qui sont intéressé du domaine de la recherche d'information, TREC offre une vaste collection de teste et des procédures de notation uniformes [**Baccini, 2010**].

TREC a été créée la première fois en 1992 par NIST (National Institute of Science and Technology), Les campagnes de TREC sont devenues la référence en ce qui concerne l'évaluation des systèmes de recherche d'information, cette série d'évaluations annuelles base en général sur les chercheurs pour construire les bases de données.

Aujourd'hui le TREC est considéré comme le développement le plus important dans la recherche d'informations expérimentales, et demeure le plus cité et le plus utilisé par la communauté de la recherche d'information [**Bouramoul, 2011**].

TREC à plusieurs versions avec déférente procédure et la figure suivante explique le but du TREC entre 1992 et 2006[**Beigbeder, 2006**].

[1] https://trec.nist.gov/



	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Adhoc, Robust	X	X	X	X	X	X	X	X				X	X	X	
Routing	X	X	X	X	X	X									
Filtering				X	X	X	X	X	X	X	X				
Interactive, HARD			X	X	X	X	X	X	X	X	X	X	X	X	X
Spanish			X	X	X										
Chinese					X	X									
X→(X,Y,Z)						X	X	X	X	X	X				
OCR				X	X										
Speech						X	X	X	X						
Video										X	X	Now VidTREC			
VLC						X	X	X	X						
Web								X	X	X	X	X	X		
Terabyte													X	X	X
Enterprise														X	X
Blog															New
Q&A								X	X	X	X	X	X	X	X
Novelty											X	X	X		
Genome											X	X	X	X	X
Legal															New
Spam detection														X	X

FIGURE 2.1 : Déférente procédures de TREC entre 1992 et 2006

2.2. Historique du TREC

La conférence Text Retrieval (TREC), coparrainée par l'Institut national des normes et de la technologie (NIST) et le Département de la défense des États-Unis, a été lancée en 1992 comme nous avons dit précédemment, dans le cadre du programme TIPSTER Text.

Son but était de soutenir la recherche au sein de la communauté de recherche d'information en fournissant l'infrastructure nécessaire pour l'évaluation à grande échelle des méthodologies de récupération de texte. [**Beigbeder, 2006**].

2.3. Objectifs de TREC

La série d'ateliers TREC a les objectifs suivants :

- Encourager la recherche en recherche d'information basée sur de grandes collections de tests.

- 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 des problèmes du monde réel.

- Accroître la disponibilité de techniques d'évaluation appropriées à l'usage de l'industrie et des universités, y compris le développement de nouvelles techniques d'évaluation plus applicables aux systèmes actuels. [**Bouramoul, 2011**].

2.4. Fonctionnement de TREC

TREC est dirigé par un comité de programme composé de représentants du gouvernement, de l'industrie et des universités.

1. Le NIST fournit pour chaque TREC un ensemble de test de documents et de questions.
2. Les participants utilisent leurs propre systèmes de récupérations sur les données et retourné au NIST une liste des docs les mieux classés.
3. Le NIST regroupe les résultats individuels, évalue l'exactitude des documents récupérés et évalue les résultats
4. Le cycle TREC se termine par un atelier qui permet aux participants de partager leurs expériences.

3. Collection de CLEF

3.1. Définition du CLEF

Les campagnes CLEF[2] (Cross-Language Evaluation Forum) ou aussi (Conférence et laboratoires du Forum d'évaluation, anciennement Forum d'évaluation inter linguistique) qui se rattachent plus particulièrement aux systèmes multilingues, le projet européen d'évaluation des systèmes de recherche d'information est lancée en 2000 par un but de promouvoir la recherche dans le domaine des systèmes multilingues, CLEF travaille à construire des campagnes d'évaluations annuelles [**Bouramoul, 2011**].

Les deux figures suivantes représentent quelque statistique sur CLEF, première figure est un graphe qui représente le nombre des participants de CLEF par rapport aux continents, nous remarquons que les européens représente le meilleur pourcentage des participants, pour les participants africains nous remarquons que presque n'existant pas pendant toute la période entre 2000 et 2008.

La deuxième figure représente le nombre des participant par groupe car CLEF devise leur participant par des groupes, Nous remarquons que AdHoc et QA@TREC (Groupe de trec du Question Answer) sont les deux groupe les plus dominante en TREC [**Ferro et Peters, 2008**].

[2] http://www.clef-initiative.eu

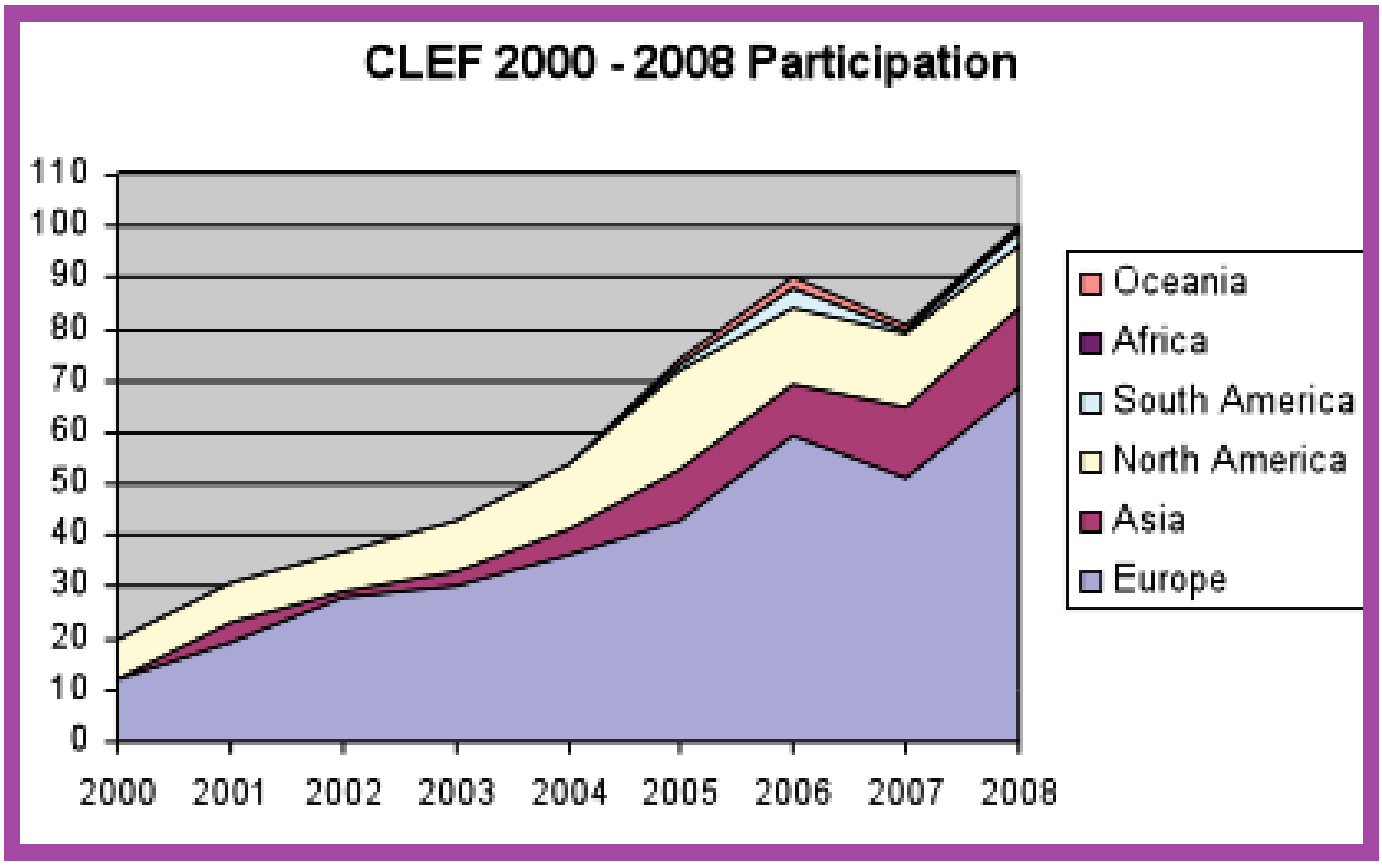


FIGURE 2.2 : Graphe de nombres des participants de CLEF par rapport aux continents

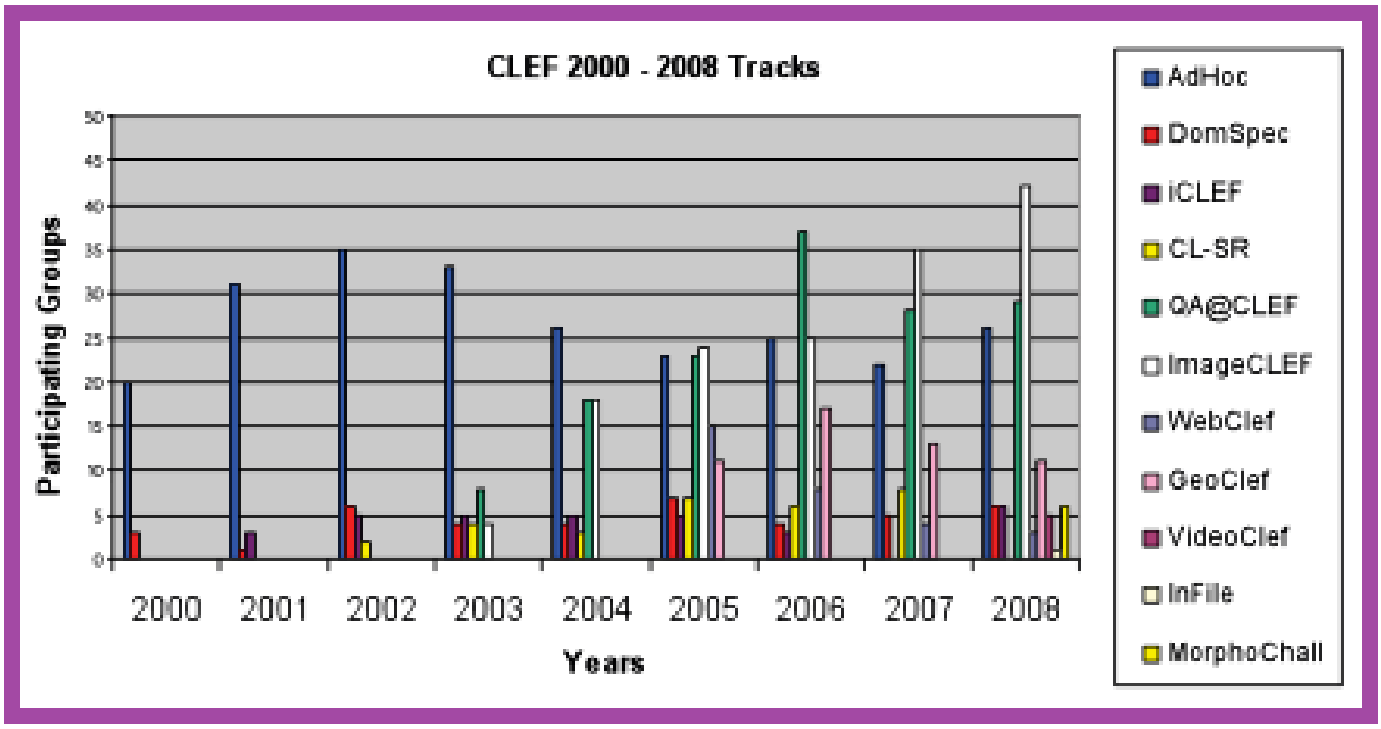


FIGURE 2.3 : Groupes participants du CLEF entre 2000 ET 2008

Le tableau suivant représente quelque informations comme l'année d'addition au groupe de TREC ou la taille ou nombre des documents des différentes collections du TREC, nous remarquons que TREC base sur plusieurs langues comme l'Anglais, le Français, l'Allemand, la langue Espagnole, italienne et aussi la langue russe.

Collection	Added in	Size (MB)	No. of Docs	Median Size of Docs. (Bytes)	Median Size of Docs. (Tokens)[7]	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: AGZ 94	2001	86	50527	1454	187	129
Italian: AGZ 95	2003	85	48980	1474	192	132
Russian: Izvestia 95[8]	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

TABLEAU 2.1 : Quelques informations sur les différentes collections du CLEF

3.2. Objectifs du CLEF

CLEF favorise la recherche et le développement en fournissant une infrastructure pour :

1. Tests, réglages et évaluations de systèmes multilingues et multimodaux.
2. Etude de l'utilisation de données non structurées, semi-structurées, hautement structurées et sémantiquement enrichies dans l'accès à l'information.

3. Création de collections de tests réutilisables pour l'analyse comparative.
4. L'exploration de nouvelles méthodes d'évaluation et de moyens novateurs d'utiliser des données expérimentales.
5. Discussion des résultats, comparaison des approches, échange d'idées et transfert de connaissances.

3.3. Structure de CLEF

L'Initiative CLEF est structurée en deux parties principales :

1- Une série de laboratoires d'évaluation, c'est-à-dire des laboratoires pour effectuer l'évaluation des systèmes d'accès à l'information et des ateliers pour discuter et piloter des activités d'évaluation novatrices.
2- Une conférence à comité de lecture sur un large éventail de questions, y compris :
   - Enquête poursuivant les activités des laboratoires d'évaluation.
   - Expériences utilisant des données multilingues et multimodales ; en particulier, mais pas uniquement, les données résultant des activités CLEF.
   - La recherche sur les méthodes d'évaluation et les défis.

4. Travaux connexe

Plusieurs projets sont faite pour crée des collections de test soit pour la langue arabe ou d’autre langue, aussi il existe plusieurs projet qui travaille à évaluer les algorithmes et les techniques des moteur de recherche, nous présentons dans le tableau suivants quelque travaux connexe avec notre PFE.

Titre	les auteurs	Année	la problématique abordée	Université
Une approche de recherche sémantique dans les documents semi-structurés	Rami Harrathi Sylvie Calabretto	2010	ce document parle sur la mesure de similarité entre la requette et une partie du texte a cherché	INSA de Lyon, France
Recherche d'Information Translinguistique sur les Documents en Arabe	Youssef Kadri	2008	Les corpus paralleles, et Aussi il a travaillé sur des documents en Arabe	Universite de Montreal, Canada
IRC CHANNEL DATA ANALYSIS USING APACHE SOLR	Nikhil Reddy Boreddy	2015	Ce document ce base sur les SRI, comme Lucene et surtout Apache Solr	Purdue University West Lafayette, Indiana
Stemming Methodologies Over Individual Query Words for an Arabic IRS	Hani Abu-Salem Mahmoud Al-Omari Martha W. Evens	1999	étalement sur des mots de requête individuels pour un système de recherche d'information en arabe	University, P.O. Box (7), Mu'tah, Karak, Jordan.
Arabic information retrieval perspectives	Ahmed abdelali Jim Cowie Hamdy s. soliman	2008	Perspectives de recherche d'information en arabe	New Mexico State University, Las Cruces, USA
Examining the Effect of Improved Context Sensitive Morphology on Arabic Information Retrieval	Kareem Darwish Hany Hassan Ossama Emam	2011	l'effet de l'amélioration de l'analyse morphologique sur l'information arabe unilingue Récupération (IR).	German University in Cairo New Cairo, Cairo, Egypt

TABLEAU 2.2 : Travaux connexe à notre PFE

5. Moteurs de recherche

5.1. Avantages des moteurs de recherche

- La possibilité de faire des recherches dans une grande masse d'informations.
- Le fait d'obtenir rapidement des informations précises sur des sujets divers et variés.
- Le fait de pouvoir trouver des documents spécifiques.
- Le fait de repérer des sites récents ou ayant été écartés des annuaires.
- Le fait de pouvoir réaliser des recherches complexes en utilisant la logique booléenne.

5.2. Inconvénients des moteurs de recherche

- Absence de contrôle des informations (présence d'URL périmées dans les résultats).
- Des interrogations qui semblent complexes car les interfaces changent d'un moteur de recherche à l'autre.
- Des résultats parfois surprenants car l'indexation est automatique.
- Des résultats qui sont classés selon un ordre qui donne la priorité à la popularité des informations.

5.3. Meilleures moteurs de recherche au monde

Selon Net Applications[3], sur l'ensemble de l'année 2015, Google a représenté près de 62.3% des requêtes sur lesmoteurs de recherche dans le monde, Apres Google nous trouvons au classement le moteur de recherche américain Baidu qui domine sur 19,8 des requêtes dans le monde.

Nous Trouvons aussi au classement des Top 5 des moteurs des recherches les plus utilisé au monde le moteur Bing et Aussi Yahoo, nous présentons quelque information sur les moteurs de recherche les plus connue au monde :

- ✓ Google[4] : Le géant du monde dans la recherche d'information, une entreprise américaine de services technologiques crée en 1998, en Californie, par Larry Page et Sergueï Brin.

- ✓ Baidu[5] : est une entreprise internet chinoise, crée à Pékin en 2004, le site le plus consulter au Chine en juin 2013, et le 4ᵉ site le plus consulté sur Internet en 2016.

- ✓ Bing[6] : un moteur de recherche élaboré par la société Microsoft lancé en juin 2009 à Washington, il est connu aussi par le nom MSN Search, Live Search et Windows Live Search.

- ✓ Yahoo[7] : Ou Yahoo search, lancé en 1995 par le langage de programmation PHP, le plus ancien et classé le 4eme au monde par un pourcentage de 7,7% des utilisateurs aumonde.

[3] http://www.netapplications.com/

[4] https://www.google.com/

[5] https://www.baidu.com/

[6] https://www.bing.com/

[7] https://search.yahoo.com/

La figure suivante représente le pourcentage d'utilisation des moteurs de recherche au monde d'après Net Applications :



FIGURE 2.4 : Pourcentage d'utilisation des meilleurs moteurs de recherche

6. Mesures de performance

Dans le domaine de la recherché d'information, il existe quelques mesures de performance d'un système de recherche d'information, le rappel et la précision sont deux mesures pour évaluer les performances des systèmes de recherche d'information, ces deux mesures base sur la connaissance des documents pertinents de la collection d'une part.

D'autre part la partition de l'ensemble des documents restitués par le système de recherche d'information en deux catégories : documents pertinents et documents non pertinents, nous citons quelques mesures connues et utilisées en recherche d'information :

6.1. Rappel

C'est le pourcentage de documents pertinents restitués par le système relativement à l'ensemble des documents pertinents contenus dans la base, le rappel mesure la capacité du système à restituer l'ensemble des documents pertinents, cette mesure a une relation avec la mesure du silence que nous allons expliquer prochainement, le rappel est calculée selon la formule suivante :

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

6.2. Précision

Précision mesure la capacité du système à ne sélectionné que des documents pertinents d'où cette mesure a une relation avec la mesure du bruit que nous va l'expliquer prochainement, la Précision égal la proportion de documents pertinents relativement à l'ensemble des documents restitués par le système, la précision est calculée selon la formule suivante : Précision

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

6.3. F-Mesure

F-Mesure ou bien mesure harmonique est une mesure qui combine le rappel et la précision, en effet rappel et précision ayant tendance à varier en sens inverse.

Nous n'avons pas un bon système de recherche d'information qui se base seulement sur le rappel ou seulement sur la précision, il faut baser sur les deux car le bon SRI doit sélectionner tous les documents pertinent et au même temps doit sélectionner que des documents pertinent, f-mesure est calculée selon la formule suivante :

$$\text{F-mesure} = 2 * \frac{(\text{Précision} * \text{Rappel})}{(\text{Précision} + \text{Rappel})}$$

6.4. Bruit et silence

Bruit est l'ensemble de documents non pertinents sélectionné par le système de recherche d'information, lors d'une recherche d'information.

Silence est l'ensemble des documents pertinents non sélectionné par un SRI lors d'une recherche, ces deux mesures sont aussi importantes pour évaluer les résultats obtenus par un SRI.

6.5. Relation entre rappel et précision

Un système de recherche d'information idéal est un système qui sélectionné tous les documents pertinent (rappelle = 1), et tous les documents qu'il retrouve sont pertinents (précision=1), en pratique ce seuil idéal n'est jamais atteint puisque ces deux mesure évoluent en sens inverse, intuitivement, si nous augmentons le rappel en retrouvent plus de documents pertinents, et si nous diminuons la précision en retrouvant aussi plus de documents non pertinents.

7. Conclusion

Dans ce chapitre nous avons décrit quelque travail connexe et proche de notre projet, le meilleur projet qui peut donner un exemple sur notre PFE est la base de TREC d'où chaque année NIST (National Institute of Science and Technology) travail à créer une nouvelle collection de TREC, un géant projet qui donne des bénéfices record.

Par la même méthode et les mêmes buts CLEF la version européenne de TREC donne aussi des résultats fantastiques, et donne des collections bien étudier dans les différentes langues du continent européen et mondial.

Ensuite nous avons cité quelque travaux connexe en langue arabe, il existe plusieurs chercheures qui ont construit des collections valide en arabe, nous avons prenez des exemples qui ont les mêmes caractères et les mêmes buts de notre travail.

Les moteurs de recherche ont aussi des relations avec notre PFE car nous voulons crée un moteur de recherche, nous avons donné des exemples sur les meilleures moteurs classé dans le monde.

Finalement nous avons donné les différentes formules des mesures qui évaluer les SRI, rappel, précision et f-mesure sont les mesures les plus connue au monde dans le domaine de la recherche d'information et chaque moteur de recherche travail à améliorer son système de recherche d'information par l'amélioration des valeurs des mesures auparavant.

CHAPITRE III :
L’approche proposée

1. Introduction

La pertinence d'un document par rapport à une requête doit être sûre pour pouvoir obtenir les meilleurs résultats pour l'utilisateur.

Ce chapitre présente la structure générale de notre approche que nous avons proposé pour créer une collection de RI.

Elle consiste à construire une collection de test arabe comme TREC[1] et CLEF[2], par la récupération des documents et la construction de la base et des requêtes, et après avoir eu la collection, nous allons récupérer les documents pertinents de chaque requête et donner un degré de pertinence.

Dans l'état de l'art actuel en RI, les résultats sont faites à partir des calculs d'existence des mots clés de la requête, par contre pour notre travail, dans un premier temps nous avons fait des combinaisons entre trois moteurs de recherche « Open Source » pour avoir seulement les documents qui appartient vraiment aux requêtes données.

Après nous avons validé la pertinence par la validation d'un être humain, et par l'utilisation d'un algorithme qui s'appelle «Balanced Interleaving» que nous allons l'expliquée prochainement.

Finalement, nous avons bénéficié de l'existence de la matrice finale de la pertinence pour faire des techniques d'apprentissage afin d'évaluer et comparer les résultats des moteurs de recherche avec les résultats trouvés par la technique d'apprentissage.

[1] TREC Text REtrieval Conference
[2] CLEF Cross-Language Evaluation Forums

2. Description de l'approche

Notre travail est devisé en trois étapes principales :

I. Construction de la base documentaire et la liste des requêtes.

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

III. Validation des résultats par l'utilisation de la techniqued'apprentissage automatique.

2.1. Plan global du travail

Dans la figure suivante, nous expliquons en général les étapes principales de l'approche proposée pour notre projet :

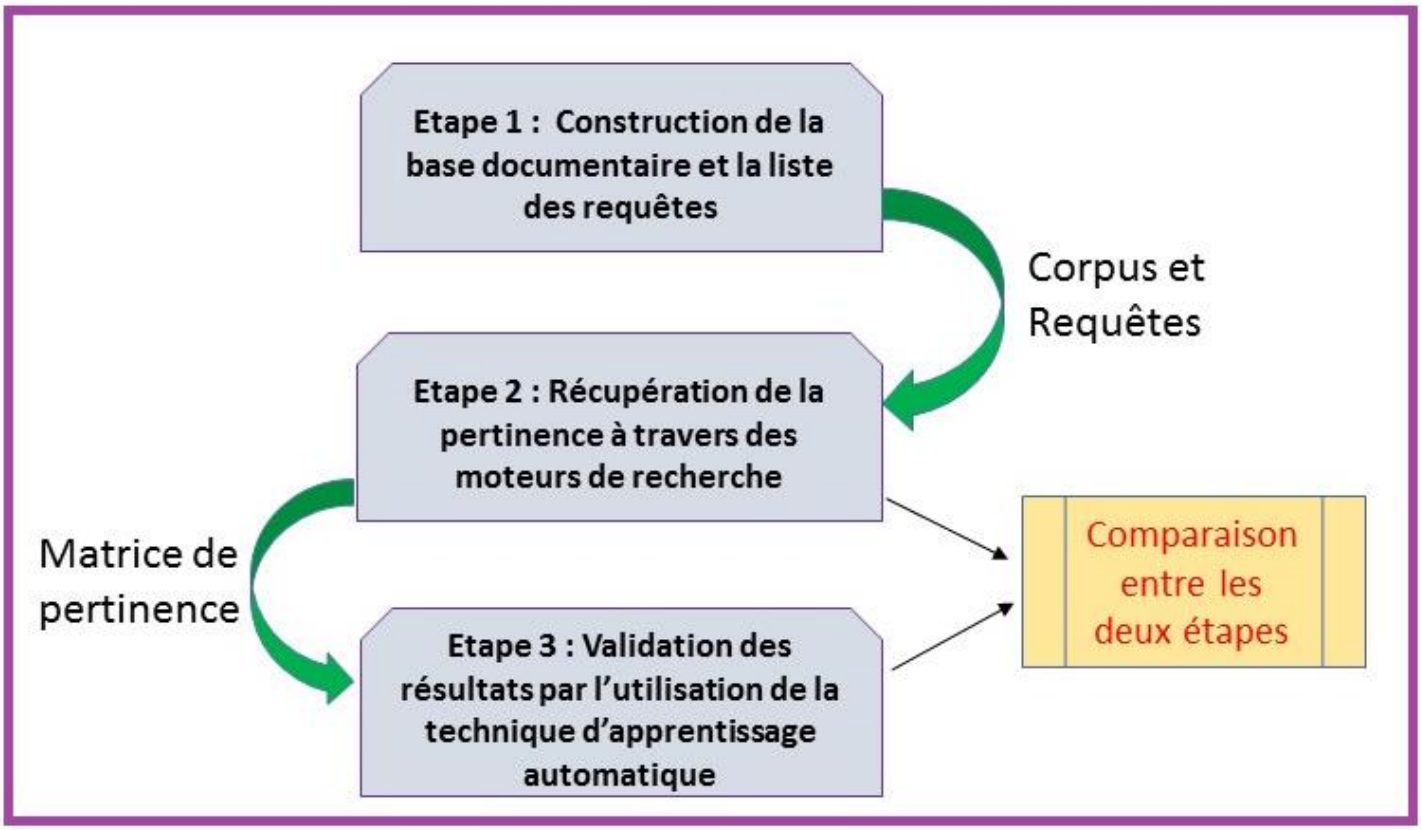


FIGURE 3.1 : Plan global de l'approche proposée

2.2. Plan détaillé du travail

Dans la figure suivante, nous expliquons en détails les étapes principales de l'approche proposée pour notre projet :

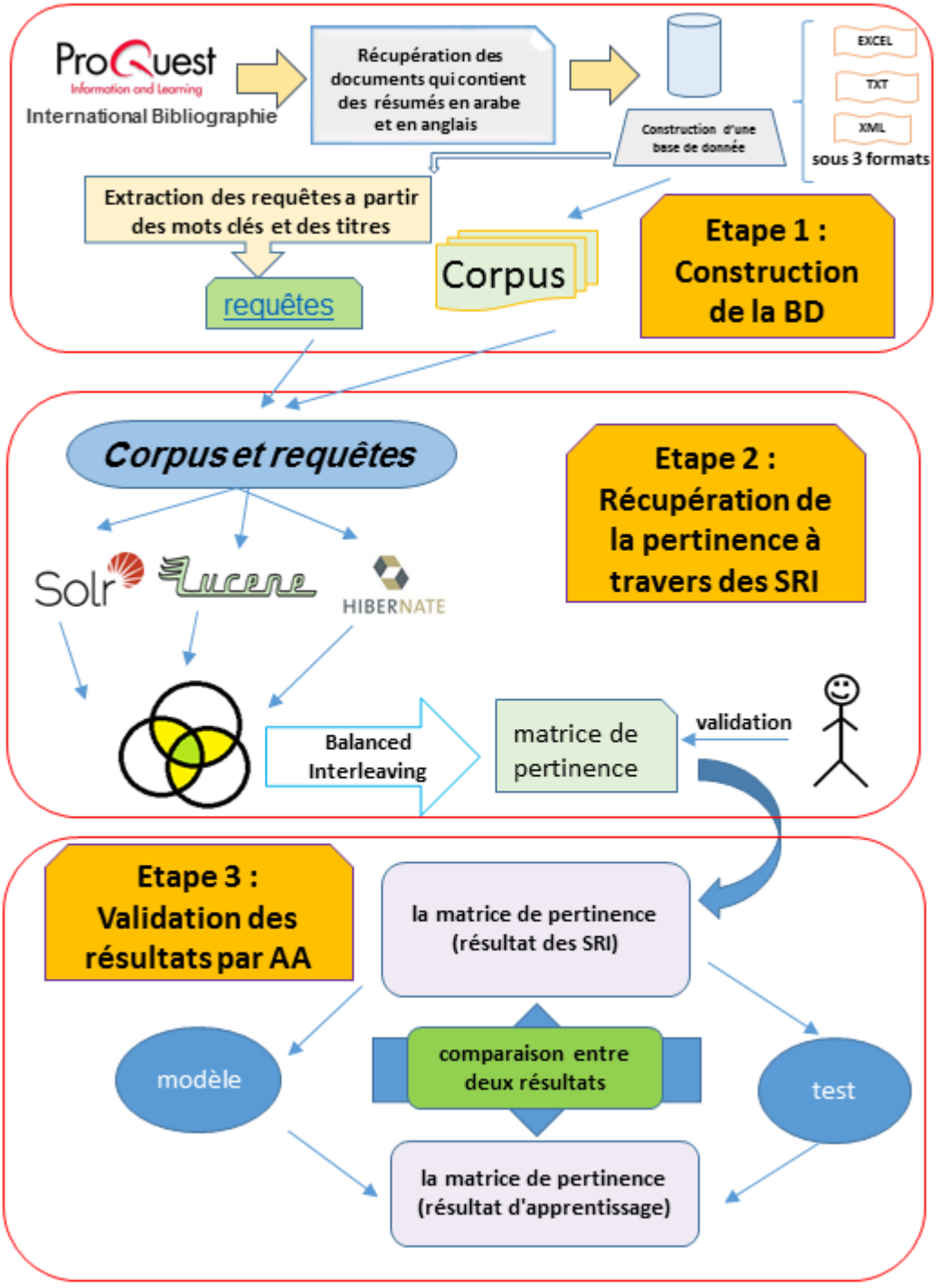


FIGURE 3.2 : Plan détaillé qui explique les déférentes étapes de notre approche

Nous avons mentionné précédemment que notre projet contient trois étapes principales. Maintenant nous allons les détaillées par la suite :

3. Construction de la collection

Pour améliorer les techniques utilisées dans les moteurs de recherche, il nous faut une base de données fiable, pour cela nous avons construit notre base de données, en utilisant les projets de fin d'étude de doctorat et de master. Et exactement leurs résumés, car le résumé d'un projet de fin d'étude contient des mots scientifiques, et utilise un lexique d'écriture direct, claire et compréhensif.

La construction de la base de données est la première étape dont il faut chercher des documents et les analysés manuellement pour sélectionner les meilleures par rapport aux critères de recherche, pour cela nous avons devisé cette partie en trois étapes :

3.1. Récupération des documents

Nous avons construit une collection qui contient 600 articles scientifiques sous format PDF au début dont plus de 90% des documents sont téléchargés de la bibliothèque Proquest[3], cette bibliothèque américaine offre des articles scientifiques pour les différents domaines et les différentes universités du monde, alors avons choisi cette bibliothèque pour deux raisons :

Premièrement notre université Dr.Yahia Farés à Médéa a donné un offre pour leurs étudiants afin d'utiliser la bibliothèque gratuitement pour une période de 20 jours.

[3] Proquest est une bibliothèque électronique qui offre aux chercheurs les déférents projets de fin d'étude au monde https://www.proquest.com/

Deuxièmement car cette Big Data facilite la recherche aux documents des pays arabe et surtout les pays du Moyen-Orient, et ça nous aide à trouver des résumés en arabe et en Anglais pour le même document.

Notre objectif dans cette première phase est de trouver des documents qui contiennent des résumés en Anglais et en Arabe au même temps pour l'avantage d'avoir une collection bilingue.

La figure suivante représente l'interface de recherche de la bibliothèque utilisée pour télécharger des documents bilingue afin de construire notre base documentaire.

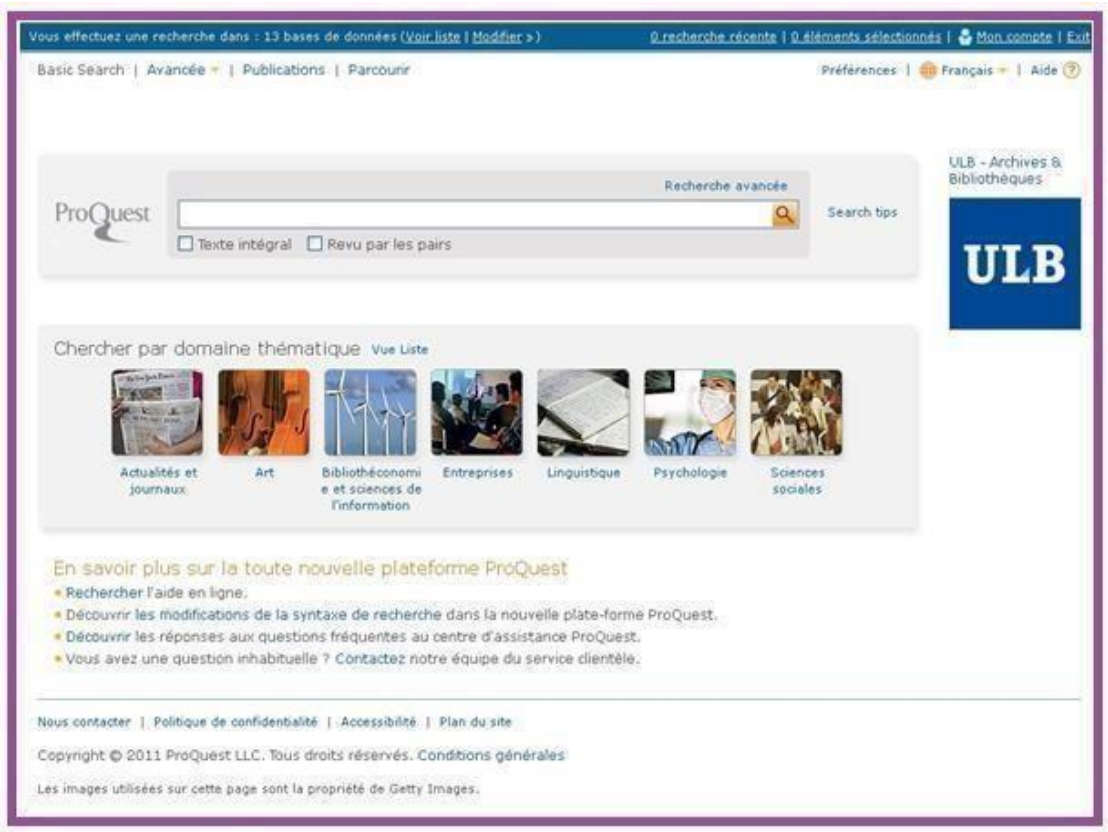


FIGURE 3.3 : Page de recherche de la bibliothèque électronique Proquest

3.2. Construction de la base

Apres le téléchargement des 600 documents bilingue (en arabe et en anglais) nous avons créé une base de données sous Excel, et depuis chaque document nous avons récupéré 11 informations, et le tableau suivant montre les types des informations récupérées.

champs	Type	Min	Max	Moyenne
Id	chaine de caractères	D001	D600	/
Titre Ar	chaine de caractères	4 mots	12 mots	6,8
Titre	chaine de caractères	4 mots	12 mots	6,6
Résumé Ar	chaine de caractères	/	/	196
Résumé	chaine de caractères	/	/	203
Année	entier	1982	2017	2008
Université	chaine de caractères	/	/	/
Mots clés Ar	chaine de caractères	3	9	5,3
Mots clés	chaine de caractères	3	9	5,3
Topic	chaine de caractères	T1	T6	/
Under Topic	chaine de caractères	UT1	UT17	/

TABLEAU 3.1 : Les différentes informations récupérées pour chaque document

Les «Topic» et «Under Topic» sont récupérés directement à partir des documents, et nous les avons classifiées comme suit :

Topic Id	Topic Ar	Topic	Under Topic Id	Under Topic Ar	Under Topic
T1	العلوم الطبيعية	Natural Sciences	UT01	كيمياء طبيعية	Natural Chemistry
			UT02	الهندسة البيئية	Environmental Engineering
			UT03	جيولوجيا	Geology
T2	تكنولوجيا	Technology	UT04	هندسة البترول	Petrol Engineering
			UT05	هندسة ميكانكية	Mechanical Engineering
			UT06	هندسة كهربائية	Electrical Engineering
			UT07	الاعلام الالي	Computer Science
			UT08	هندسة الطيران	Aviation Engineering
			UT09	هندسة الاتصالات	Communications Engineering
T3	العلوم الدقيقة	Science Exact	UT10	الرياضيات	Mathematics
			UT11	الفيزياء	Physics
			UT12	الكيمياء	Chemistry
T4	هندسة المدينة	City Engineering	UT13	هندسة المواصلات	Transportation Engineering
			UT14	الهندسة المدنية	Civil Engineering
			UT15	هندسة الإدارة و البناء	Management and Construction Engineering
T5	الاقتصاد	Economy	UT16	الاقتصاد	Economy
T6	العلوم الإسلامية	Islamic Sciences	UT17	العلوم الإسلامية	Islamic Sciences

TABLEAU 3.2 : Tableau des Topic et Under Topic

La figure suivante représente le pourcentage des documents par rapport au :

1. Topic.
2. Under topic.
3. Année de création du document.
4. Pays de l'université.

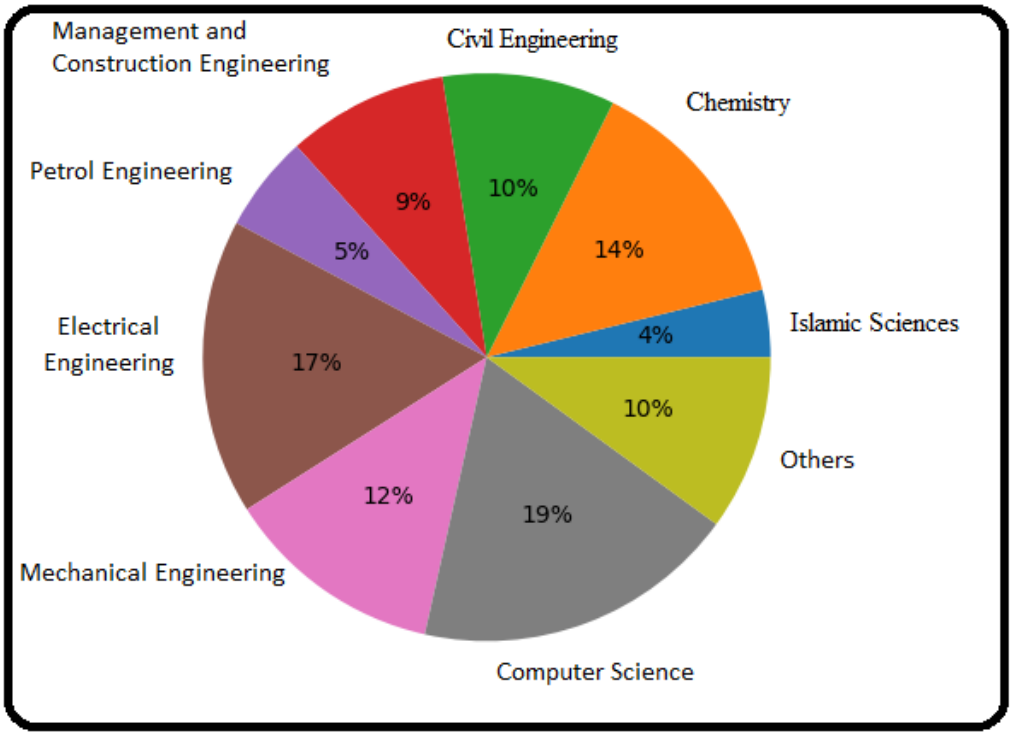


FIGURE 3.4 : Pourcentage des documents par apport au champ Under Topic

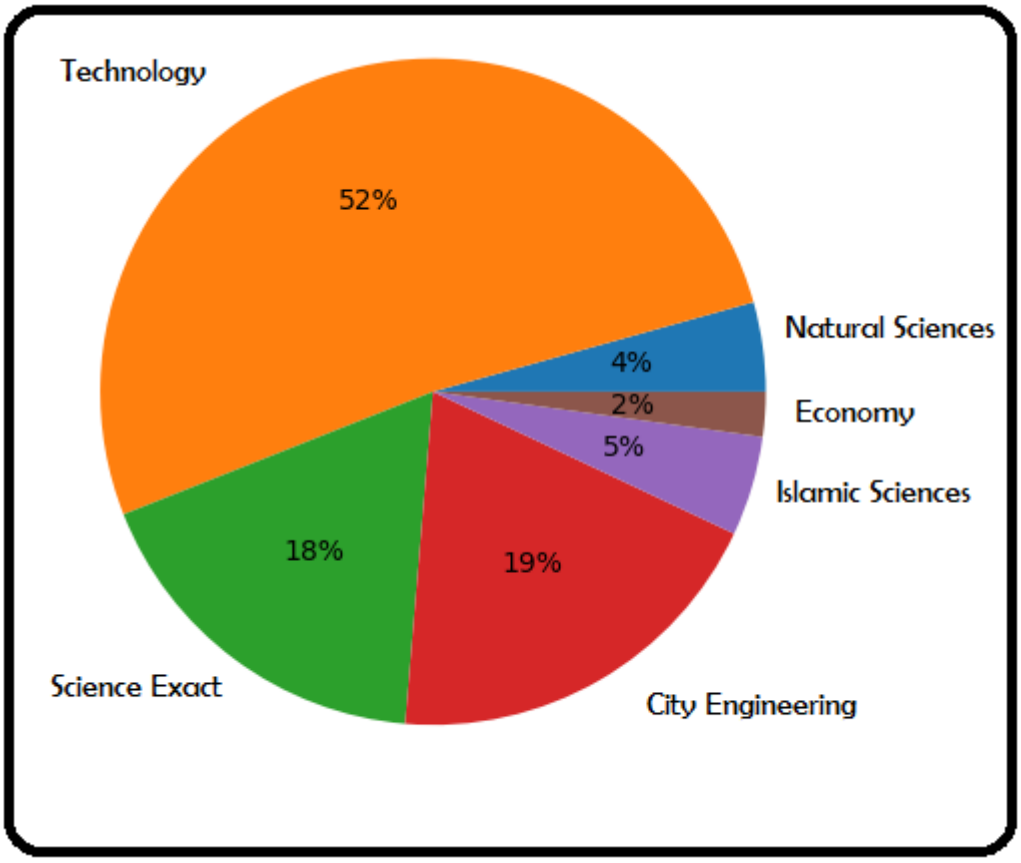


FIGURE 3.5 : Pourcentage des documents par apport au champ de Topic

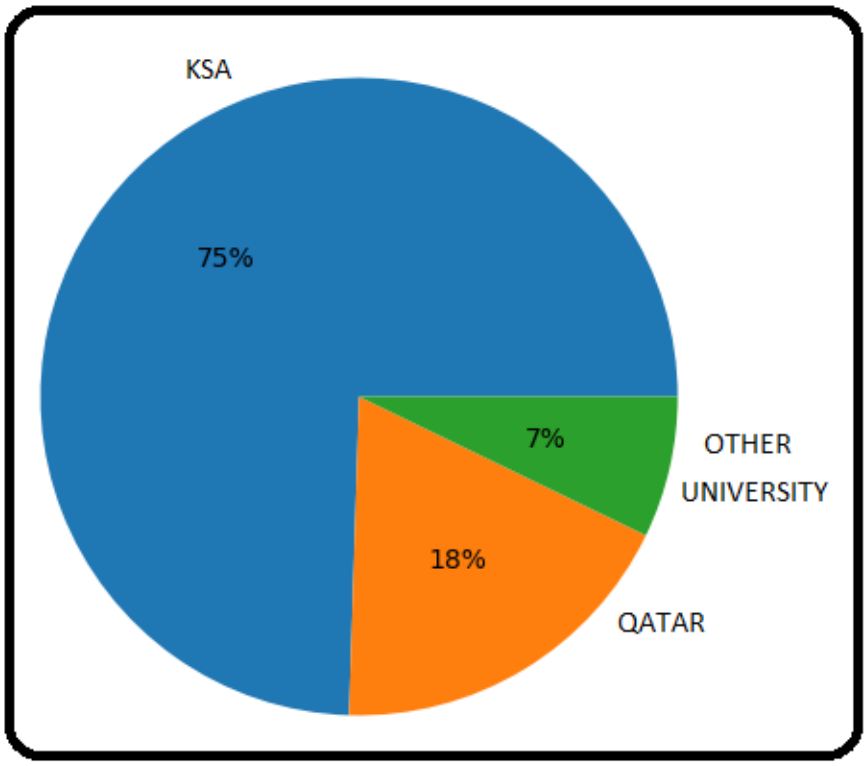


FIGURE 3.6 : Pourcentage des documents par apport au pays de l'université

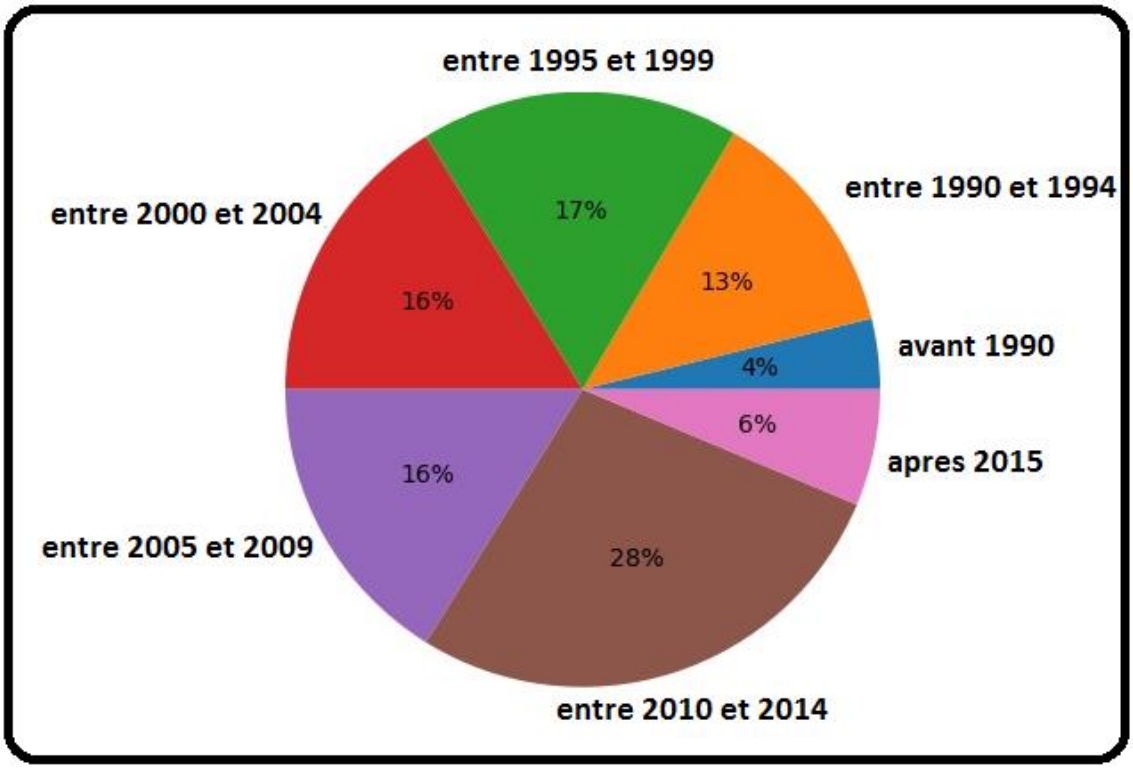


FIGURE 3.7 : Pourcentage des documents par apport au champ d'année

Concernant les informations qui ne sont pas disponibles sur certains documents, par exemple des fois nous ne trouvons pas les mots clés en Anglais ou en arabe, alors nous avons fait des traductions en utilisant un programme Python et en plus nous avons fait une validation humaine pour avoir des résultats corrects et aussi pour avoir des tableaux complète.

La figure suivante représente le code source qui fait la traduction par l'utilisation de la bibliothèque Translate.

```
#!/usr/bin/env python
#-*- coding: utf-8 -*-
import xlrd
import xlwt
from translate import Translator
a= xlrd.open_workbook("la base.xlsx")
sh = a.sheet_by_name('1')
trad = Translator(from_lang="ar",to_lang="en")
output = xlwt.Workbook()
p1 = output.add_sheet('résultat traduction')


for i in range (0,161):
                result = trad.translate(sh.cell(i,0).value)
                p1.write( i,0 , result)

output.save('resultat.xls')
```

FIGURE 3.8 : Code source python qui fait la traduction

3.3. Extraction des requêtes

Pour proposer des requêtes nous avons fait une analyse sur les titres et les mots clés en Arabe en utilisant un programme Python que nous allons créer pour calculer le nombre d'occurrence des « Un-gramme, Bi-gramme, trigramme et 4-gramme ».

Nous avons évité les résultats qui ont une grande occurrence (V_max = 100) car ce sont des mots vides.

Et aussi nous avons ignoré les résultats qui existent moins de 15 fois (V_min = 15) car une requête qui n'existe pas trop dans la collection qui va minimiser les résultats du projet.

Et bien sur ces deux valeurs V_max et V_min peuvent être modifié dans les futurs travaux.

Les résultats donnés par le programme python sont traités manuellement pour sélectionner seulement les requêtes qui ont un sens, pour cela nous avons gardé 161 requêtes.

Le tableau suivant montre un échantillon de la base des requêtes : (Tableau complet voir l'annexe)

Id Query	Query Ar	Query
Q001	الأمن المعلوماتي	Information security
Q002	اتصالات الشبكة الذكية	Smart Network Connections
Q003	أداء أنظمة الاتصالات	Performance of communication Systems
.	.	.
.	.	.
.	.	.
Q160	نظام الشبكات اللاسلكية	Wireless Networking System
Q161	نمو الحبوب	Grain growth

TABLEAU 3.3 : Tableau des requêtes

Et comme nous avons expliqué précédemment, les requêtes sont devisées sur 4 parties par rapport au nombre de mots dans chaque requête.

nombre de mots	**nombre de requêtes**
un-gramme	21
bi-gramme	96
trigramme	34
4-gramme	10

TABLEAU 3.4 : Nombre d'existence des requêtes par rapport au nombre des mots de la requête

Et la figure suivant montre la distribution des requêtes selon le nombre de mots :

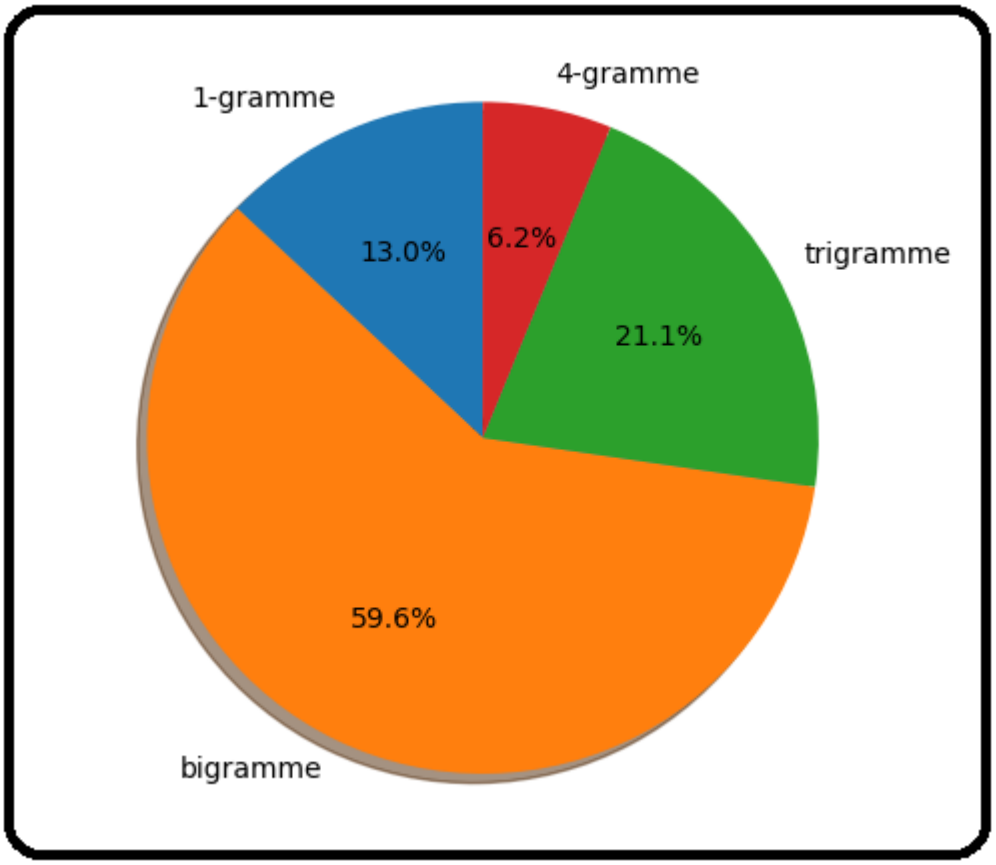


FIGURE 3.9 : Pourcentage d'existence des requêtes par rapport au nombre des mots de la requête

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

Le noyau de notre travail est représenté dans l'étude de la pertinence des documents construits précédemment par rapport aux requêtes que nous les avons extraits, l'étude de la pertinence peut être contrôlé par rapport à quelques critères, par exemple le K (signifie le nombre des documents retournés par rapport à une requête donnée). Généralement l'utilisateur n'arrive pas à lire tous les résultats, au maximum il va consulter les 20 premières pages,

Pour cela nous avons pris (K=50) par défaut, mais cette valeur peut être modifiée soit dans ce projet ou dans les futurs travaux.

Nous allons deviser cette étape en trois parties :

4.1. Validation par les trois moteurs de recherche

Les moteurs de recherche Open Source offrent à leurs utilisateurs de faire des tests (des fois limités) mais ces offres aident les chercheurs et les étudiants a évalué et juger leurs travail par la comparaison entre les résultats des moteurs de recherche mondial et les simples algorithmes des chercheures.

Les trois moteurs de recherche choisis sont :

- Lucene
- Hibernate
- Apache Solr

Le tableau suivant explique quelques informations sur les trois moteurs de recherche utilisés dans le projet :

le moteur	Programmé en	Fondation	licence	Dernière version
Apache Solr	Java	Apache	libre	15/01/2018
Hibernate	Java	Red Hat	publique générale limitée	19/10/2017
Lucene	Java,Perl, C++, PHP, C#, Python	Apache	libre	27/04/2017

TABLEAU 3.5 : Comparaison entre les trois moteurs de recherche utilisé

Chaque moteur de recherche nous a donné une matrice de pertinence, nous avons fait un degré de pertinence, si nous prenons un k=50 par exemple (les 50 premiers documents pertinent), les premiers 20% des documents pertinents ont un degré de (1) et les derniers 20% des documents pertinents ont un degré de (0.2).

La pertinence doit exister au minimum dans en trois matrice parmi les six matrices existent, ça veut dire si la pertinence existe seulement dans deux matrice sur les six matrice, nous prenons que le document qui n'appartient pas au requête.

Le tableau suivant explique comment calculer un degré de pertinence entre un document et une requête.

degré de pertinence	K_4 supérieur a :	K_4 inferieur a :
1	0%	20%
0,8	20%	40%
0,6	40%	60%
0,4	60%	80%
0,2	80%	100%

TABLEAU 3.6 : Explication des calculs des degrés de pertinences

4.2. Regroupement des résultats par l'utilisation de l'algorithme de «Balanced Interleaving »

Apres la recherche des trois systèmes de recherche d'information, nous avons eu six matrices de pertinence.

Le but de l'algorithme de «Balanced Interleaving» est de regroupé les différente résultats qui se trouve dans les six matrices précédente dans une seule matrice dans un ordre choisie par l'algorithme de «Balanced Interleaving».

Nous avons utilisé un programme Python qui va donner le classement des documents pertinents pour chaque requête à partir des trois matrices de pertinence de chaque moteur de recherche utilisé [Chapelle, 2012].

[4] : (K) pourcentage de documents retournés

```
//Algorithm: Balanced Interleaving
Input Rankings A= (a1; a2 ;...) and B= (b1; b2 ;...)
I :=(); //output ranking list, initialized to empty list
k_a:=1; k_b:=1; //counters for the lists A and B, respectively
AFirst:= randomly select 0 or 1 //decide which ranking gets priority
while (k_a<=|A| and k_b <=|B|) do // if not at end of A or B
      if (k_a< k_b or (k_a== k_b and AFirst ==1))then
        if A[k_a] <= I then I.append(A[k_a]) //append a result from A
             k_a := k_a+1
      else
        if B[k_b] <= I then I.append(B[k_b]) //append a result from B
             k_b:=k_b +1
     end if
end while
Output : interleaved ranking I =(i1;i2;…)
```

FIGURE 3.10 : Algorithme de «Balanced interleaving» [kharitonov, 2013]

4.3. Validation de l'expert

C'est une partie supplémentaire pour que notre base doive être plus performante.

Les programmes d'informatiques donnent des résultats perfectifs mais ils ne restent pas valides toujours, pour cela l'être humain a une validité plus forte que le programme, alors l'être humain (un expert) a aussi la possibilité de valider la pertinence des documents pertinents retournés par les trois moteurs de recherches, donc il peut annuler la validité de pertinence si 'il trouve que la requête n'a aucune relation avec le document.

5. Validation des résultats par l'utilisation de la technique d'apprentissage automatique

L'apprentissage est une méthode dans l'intelligence artificiel, cette méthode facilite à trouver une information perdue ou inexistante par la comparaison des informations qui existent déjà, l'apprentissage est fait par des calculs et des comparaisons compliqués.

Cette méthode a connu un développement incroyable dans les dernières années et elle contient plusieurs algorithmes, nous mentionnons la classification par l'algorithme de naïve bayésienne (naïve bayes en anglais) que nous avons utilisé dans notre approche.

Naïve bayésienne est un type de classification qui consiste à faire des calculs compliqués de probabilité basé sur le théorème de Bayes, il appartient à la famille des classifications linéaires.

Pour la troisième partie de notre projet, nous avons fait un apprentissage sur la matrice de pertinence qui représente le résultat du deuxième partie de notre PFE pour comparer les résultats obtenue par les moteur de recherche (deuxième étape du PFE) et les résultats d'apprentissage (troisièmes partie du PFE).

Nous avons réalisé deux tests d'apprentissage automatique sur la matrice de pertinence, le premier test n'a pas donner des résultats satisfaisante, donc nous avons ajouté un deuxième test pour améliorer les résultats d'apprentissage et le tableau suivant va expliquer la différence entre ces deux tests sans parler sur les résultats qui vont être expliqués dans le prochain chapitre.

	Test 1	Test 2
Modèle	200	40
classe (yes)	entre 10 et 15%	50%
classe (no)	entre 85 et 90%	50%
Test	400	560
classe (yes)	environ 10%	entre 2 et 5 %
classe (no)	environ 90%	entre 95 et 98%

TABLEAU 3.7 : Différence entre les deux méthodes d'apprentissage

Les étapes d'apprentissage automatique :

- ✓ Extraction des mots non vides
- ✓ Calculer le nombre d'existence de chaque mot dans chaque document
- ✓ Choisir la partie de modèle et partie de test
- ✓ Calculer l'apprentissage par la technique de naïve bayes

5.1. Extraction des mots non vide

À partir des résumés en arabe de 600 documents et par un programme python nous avons récupéré tous les mots, d'où nous avons trouvé 27 514 mots.

Tandis que la taille des résumés en arabe contient plus de 80000 mots, ensuit nous avons utilisé une liste des mots d'arabe vides **[**El Hachani, 1998**]** pour ignorer les mots vides qui n'ont pas un sens ou qui appartient à une autre langue (anglais ou français) ou bien des chiffres et des caractères spéciaux, enfin nous avons gardé seulement 20 278 mots.

5.2. Calcul du nombre d'existence de chaque mot dans chaque document

Par un programme sous python nous avons construit une matrice de taille 600*20278 qui contient le nombre d'existence de chaque mots dans un document, nous avons utilisé cette matrice pour le calcul d'apprentissage de la pertinence par la méthode de naïve bayes, et exactement par la comparaison d'existence des mots non vides dans les différents documents, et le tableau suivant explique cette matrice :

id mot	le mot	D001	D002	.	.	.	D600
Mot 1	ابتكار	0	1	.	.	.	0
Mot 2	ابجديات	2	0	.	.	.	0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
Mot 20 278	يمكن	0	3	.	.	.	0

TABLEAU 3.8 : Matrice de nombre d'existence de chaque mot par rapport au chaque document

La figure suivante représente le code source d'n programme sous python qui retourne la liste total des mots :

```
#!/usr/bin/env python
# coding: utf-8
import xlrd
import xlwt
liste_mots=[]
a= xlrd.open_workbook("3.xlsx")
sh = a.sheet_by_name('bdd')
book = Workbook()
feuil1 = book.add_sheet('final')
for i in range (0,600):
    var=str(sh.cell(i,0).value)
    liste=liste+var
l=liste.split( )
print("nombre de mots = ",len(l))
for i in range (0,len(l)):
    rep=0
    k=len(liste_mots)
    for j in range (0,len(liste_mots)):
        if l[i]== liste_mots[j]:rep=1
    if rep==0 :
        liste_mots.append(l[i])
print(k)
for i in range (0,k):
    feuil1.write(i,0,liste_mots[i])
```

FIGURE 3.11 : Code source du programme qui retourne la liste des mots

5.3. Choix de la partie de modèle et la partie de test

La technique de Naïve bayes contient deux partie : modèle et test, dans la 1ere expérience nous avons fait un modèle de 33% du nombre total des documents ce qui représente 200 documents et 66% pour la partie test qui représente 400 documents, mais comme nous avons dit précédemment les résultats obtenus sont pas vraiment satisfaisant.

Dans la deuxième expérience nous avons choisi le même nombre entre les documents pertinents et les documents non pertinents, ce point n'est pas pris en considération dans la 1ere expérience car il faut faire le choix au hasard (les 200 premiers documents).

Puisque nous avons choisi le même nombre de documents pour les deux catégories (appartient et appartient pas) il est obligatoire de choisir un petit nombre de documents, alors nous avons choisi 40 documents d'où 20 sont pertinents et 20 autres documents qui ne sont pas pertinents pour avoir une égalité entre les deux classes.

5.4. Calcul d'apprentissage par l'algorithme de naïve bayésienne

La phase programmation est la dernière phase de la partie d'apprentissage, cette dernière est la même pour les deux expériences car le même programme est utilisé pour calculé l'apprentissage par la technique de naïve bayésienne.

Cette méthode est spécialisé aux sacs de mots ou qui s'appelle la place dans quelque source, Sacs de mots veut dire une grande liste des mots d'où nous avons appliqué l'algorithme d'apprentissage sur cette liste.

La seule déférence est que nous avons changé l'échantillon du modèle et du test, ce programme prend comme entré la matrice final de pertinence (600*161) et la matrice de nombre d'existence de chaque mot dans chaque document (20278 *60).

Et voilà en détail les différentes étapes d'apprentissage :

- ✓ Le calcul de la probabilité des deux classes (appartient et appartient pas) d'échantillon de modèle.

$$\hat{P}(c) = \frac{N_c}{N}$$

- ✓ Le calcul de la probabilité de chaque mot du document de test par rapport à la classe (appartient) après la classe (appartient pas), dont cette probabilité égal au nombre d'existence de ce mot dans un échantillon du modèle de la classe appartient (appartient pas) plus un sur le nombre des mots de la classe appartient (appartient pas) plus le nombre total de notre échantillon de modèle.

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

- ✓ Enfin nous calculons la somme des probabilités des mots du document de test par rapport aux deux classes et chaque probabilité est multipliée par la puissance de nombre d'existence de ce mot dans le document de test, après cela nous comparons les deux probabilités (appartient et appartient pas) et la classe la plus grande va être la nouvelle classe du document testé [**Lemaire, 2014**], [**Leray, 2002**].

- ✓ Nous calculons l'apprentissage pour tout l'échantillon du modèle avec tous les requêtes pour comparer les résultats obtenus par l'apprentissage avec les résultats obtenus par les SRI, la comparaison est représentée par trois normes :

  - Rappel
  - Précision
  - F-mesure

6. Choix des deux langues (arabe et anglais)

Notre base a été créée le mois de mars 2018, c'est une base bilingue et les deux langues utilisées sont choisies pour les raisons suivantes :

6.1. Langue Arabe

Notre langue mère, c'est la langue du Coran sacré, et aussi car elle est classée la 4ème langue la plus utilisée dans le monde d'après Merker Al Bourhan[5].

La langue arabe est utilisée par plus de 150 millions de locuteurs, sans oublier la richesse de la langue arabe qui contient 12 millions et 300 milles mots, et la difficulté de l'utilisée et de l'apprendre.

Notre langue n'a pas eu son vrai droit dans le développement de la recherche d'information .si nous la comparons avec les autres langues comme l'Anglais, Chinoise ou l'Espagnol, nous trouvons des centaines de recherches universitaires sur ces langues.

6.2. Langue Anglaise

La langue Arabe utilisable seulement dans les pays du nord d'Afrique et du moyen d'orient, et pour donner plus de valeur à notre travail, il faut utiliser une langue mondiale.

Nous avons choisi la langue Anglaise qui représente la langue des études scientifique et qui est la seule langue dominante sur le monde en général et l'informatique ou la recherche d'information en particulier, pour cela nous avons fait une collection en Arabe et aussi en Anglais qui est la langue maternelle pour plus de 400 millions de personnes.

5 : https://merkez-al-bourhan.com/rang-occupe-langue-arabe-monde

7. Différents types de la Collection

Nous sommes à l'ère de l'évolution et aujourd'hui dans chaque domaine nous trouvons beaucoup de types d'extension, par exemple dans le coté de programmation nous trouvons des centaines de langages de programmation, pour ce point-là nous avons exprimé notre base de donnée par cinq types différents :

- ✓ **EXCEL** : sous forme de tableau car il est compréhensible, facile à chercher, et rapide a utilisé par le langage de programmation python qui est utilisé dans les déférentes étapes du projet.

- ✓ **XML** : connu mondialement et toutes les bases construites ressèment sont sous format d'XML.

- ✓ **TXT** : pour faire la recherche d'information par les moteurs de recherche il nous faut cette format, comme lucene il index les fichiers (TXT).

- ✓ **CSV** : nouvelle format, elle est utilisé dans les plusieurs domaine, par exemple dans les Template prédéfinis des sites web.

- ✓ **SQL :** la forme la plus connue au monde, facile a manipulé et elle est utilisable par tout.

8. Conclusion

Dans ce chapitre nous avons expliqué en détail les déférentes étapes de notre projet, depuis le choix de la bibliographie qui nous offre des documents que nous cherchons jusqu'à la comparaison entre les deux résultats obtenus par les trois moteurs de recherche et la validation de l'être humain avec les résultats d'apprentissage.

Nous avons parlé de la base de données et expliqué comment la créer, Aussi nous avons donnée quelque statistique sur notre base.

Ensuite nous avons expliqué comment nous avons étudié la pertinence par l'utilisation des moteurs de recherche et l'algorithme de «Balanced interleaving» et la validation d'un expert.

Finalement, notre approche consiste à faire des tests d'apprentissage pour faire une comparaison entre les deux méthodes (moteurs de recherche et apprentissage).

CHAPITRE IV : Implémentation ET Evaluation

1. Introduction

Dans ce chapitre nous présentons l'implémentation de notre système proposé, nous commençons tout d'abord par la présentation de l'environnement de développement, en détaillant les différents outils utilisés, les langages de programmation ainsi que les moteurs de recherche utilisés dans le projet.

Puis nous expliquons le déroulement de l'outil, qui se présente sous forme d'un site, ce dernier explique notre travail ainsi que les services offert à l'utilisateur comme la recherche sur notre base documentaire depuis les déférentes requêtes récupérées par les titres et les mots clés des documents en deux langues ou d'obtenir des informations sur la collection.

Ensuite nous présentons les résultats obtenus dans les différentes étapes de notre projet par des tableaux et des graphes pour faciliter la compréhension des résultats obtenue, et la comparaison entre les résultats obtenue par les deux méthodes (moteurs de recherche et apprentissage automatique), cette comparaison est représentée par les critères de mesure des performances (Rappel, précision et f-mesure).

Enfin nous interprétons et nous commentons les résultats obtenus pour analyser notre travail et voir si nous avons atteint nos objectifs.

2. Différentes ressources utilisées

Chaque projet d'informatique dépend sur plusieurs ressources et outils car chaque ressource va nous aider dans un point donnée, pour cela nous présentons les différents langages de programmation ou outils utilisé dans les trois étapes de notre projet :

2.1. Python

Le choix du langage Python[1] utilisé dans le projet n'est pas une coïncidence car le Python est connu mondialement et notre choix est bien expliqué dans les prochaines lignes.

2.1.1. Définition de Python

Python est un puissant langage de programmation objet, il a été créé en 1991 comme une première version et il a connu un développement énorme, il détermine implicitement les types, Il a été développé par Guido van Rossum [Doty, 2008].

Dans notre projet nous utilisons la version 3.6.4[2] qui est sortie le 19 Décembre 2017 et ce magnifique puissant langage est classé le 3eme meilleure langage de programmation au monde d'après Le classement Tiobe[3] qui se base sur les données de consultation issues des moteurs de recherche.

Nous avons opté pour le langage de programmation Python en raison de ces caractéristiques, Python est sous licence open source, il nous offre une utilisation gratuite et libre et il est indépendant de toute plateforme.

---

[1] https://www.python.org/.

[2] https://www.python.org/downloads/release/python-364/: Lien d'installation du python version 3.6.4.

[3] TIOBE: The Software Quality Company https://www.tiobe.com/tiobe-index/.

Ce langage offre la manipulation de la langue arabe[4] et aussi il fonctionne sur tous les principaux systèmes d'exploitation.

2.1.2. Bibliothèques de Python utilisés

Nous avons ajouté des bibliothèques de Python qui nous aide à lire et écrire dans les fichiers Excel, faire des traductions, ajouter des fonctions mathématiques, ou même de faire des interfaces et des graphes.

Le tableau suivant cite les déférentes bibliothèques de Python utilisé dans notre projet :

La bibliothèque	Son but
Xlrd	lire un fichier Excel
Xlwt	écrire dans un fichier Excel
Translate	faire la traduction des informations qui n'existe pas dans les documents PDF
matplotlib.pyplot	pour faire les déférents diagrammes qui explique les résultats de notre travail
Math	pour utiliser des fonctions de calcule prédéfini
Tkinter	faire des interfaces

TABLEAU 4.1 : Différentes bibliothèques de Python utilisé

L'installation des bibliothèques de Python est simple, il faut accéder au site officiel du python qui est mentionné précédemment et tapez le nom de la bibliothèque et la télécharger puis l'installer comme un simple programme.

L'installation ce faite aussi par l'utilisation de la boite de commande CMD directement par l'accès au dossier ou se trouve Python et taper la commande suivante «python –m pip install nom_de_bibliothèque».

[4] (#-*- coding: utf-8 -*-) la commande sous python pour l'encodage UTF-8.

2.2. NetBeans

NetBeans Est un environnement de développement intégrer open source qui nous offre la manipulation des différents langages de programmation et qui facilite le travail, Dans notre projet nous utilisons NetBeans pour manipuler le moteur de recherche Lucene qui se base sur le langage Java, la version utiliser est Netbeans IDE 8.2[5].

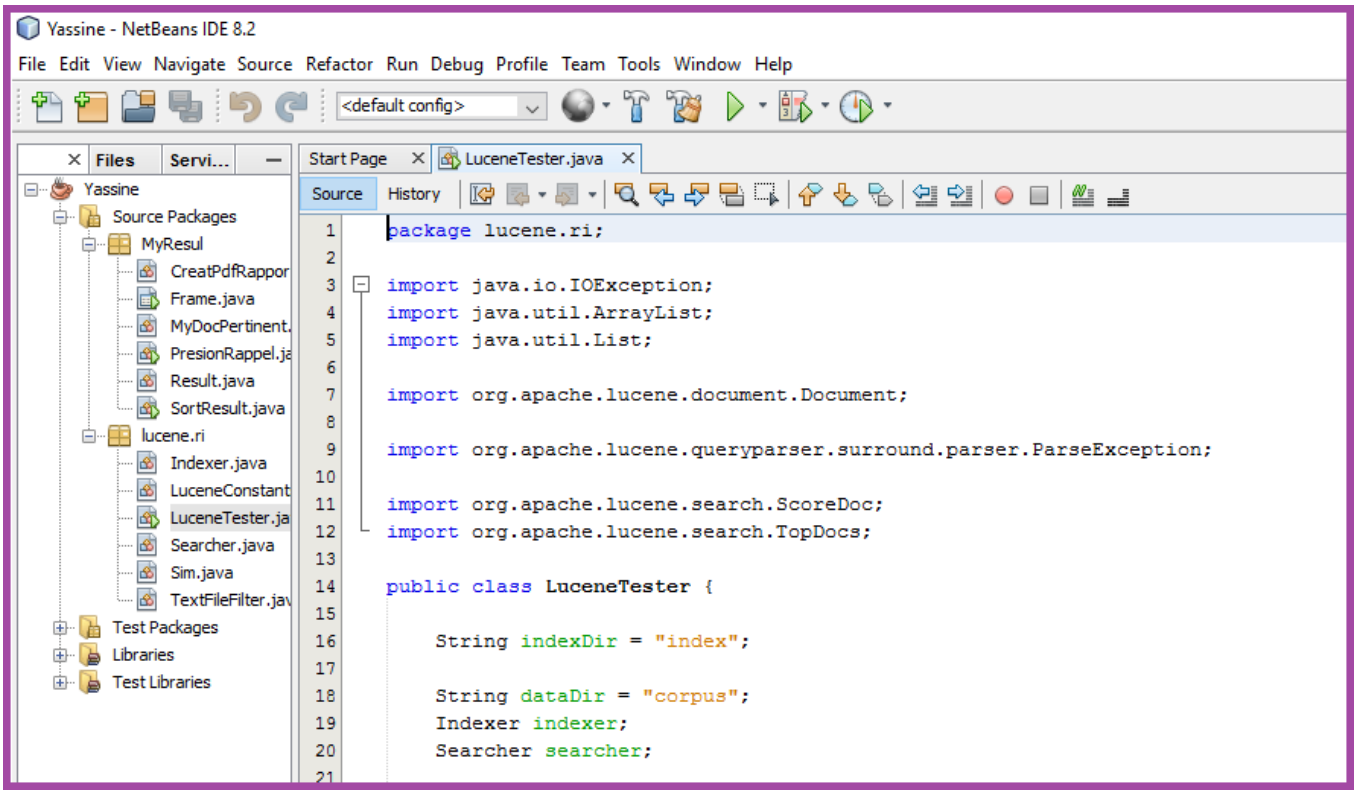


FIGURE 4.1 : Interface d'un programme Java sous NetBeans

[5] https://netbeans.org/community/releases/82/install.html

2.3. Moteurs de recherche utilisée

2.3.1. Lucene

Lucene est un moteurs de recherche d'où il utilise des techniques et des algorithmes pour donner la liste des documents pertinent par rapport à une requête donnée sans oublier que le corpus est donné par l'utilisateur de Lucene, aussi ce dernier est programmer par Java, la version utiliser est la dernière (6.5.1)[6], elle est créé le 27 avril 2017 par la fondation apache [**Balipa, 2015**].

Nous avons programmé en java une interface pour faciliter la recherche via lucene et choisir le type des résultats voulu ainsi que les résultats sont donné dans un fichier Excel qu'on peut l'import après vers Python qui va récupère les résultats des trois moteurs de recherche et le résultat de l'être humain pour donner la matrice de pertinence finale.

Et voilà notre interface crée pour faire l'exécution de lucene.

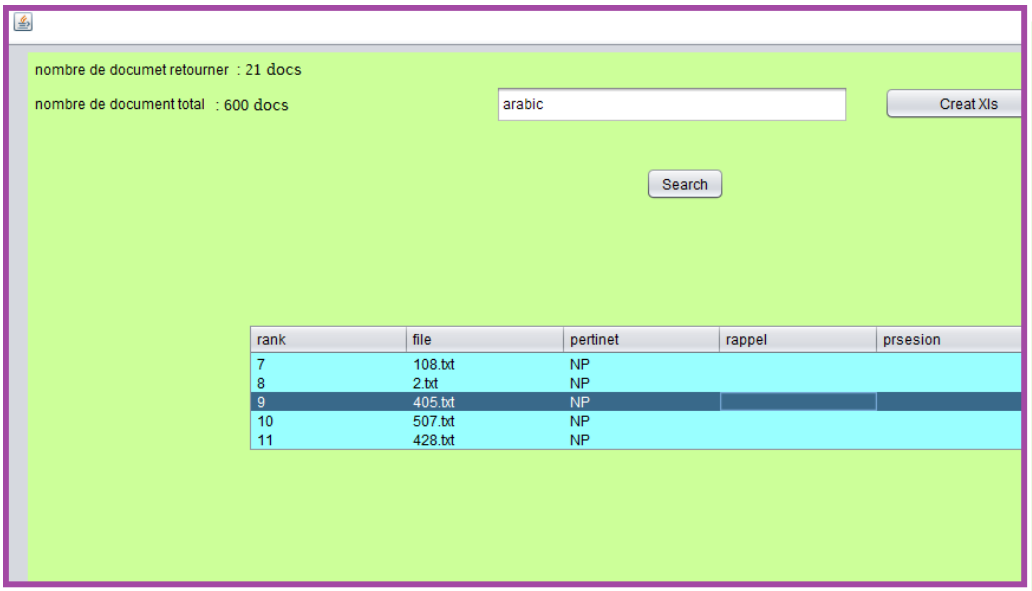


FIGURE 4.2 : Interface du moteur de recherche lucene

[6] https://lucene.apache.org/core/6_5_1/index.html

2.3.2. Solr

Solr est un moteur de recherche open source programmé par java sous la fondation d'apache avec une licence libre, Solr nous offre des recherches sur des documents de format XML, simple à utiliser [Rajani, 2013], et voilà les différentes étapes pour utiliser ce moteur de recherche open source :

- Télécharger le fichier (solr-7.3.0.zip)[7] via internet
- Par la boite de commande CMD, nous choisissons le dossier où se trouve le fichier et nous exécutons la commande (solr start)
- Création d'une base par la commande (solr create -c nom_bdd)
- Déplacer tous nos fichiers XML dans le dossier (solr-7.3.0\example\exampledocs)
- Copier tous les fichiers XML dans notre base documentaire par la commande (java -Dc=nom_bdd -jar post.jar *.xml)
- Accéder à l'adresse local (localhost:8983/solr/) et faire des recherches.

Et la figure suivante représente l'interface du moteur de recherche Solr

[7] http://www-us.apache.org/dist/lucene/solr/7.3.0/



FIGURE 4.3 : Interface du moteur de recherche Solr

2.3.3. Hibernate

Hibernate[8] est un moteur de recherche open source crée par Gavin King qui est un membre d'une équipe de développement de JBOSS.

Hibernate Search fournit la fonctionnalité de recherche de texte qui est particulièrement adaptée aux recherches

Hibernate fait la recherche sur les documents indexés manuellement par l'utilisateur sous format XML et d'autre format aussi, il est programmé en Java sous une licence publique générale mais limité [**Bernard, 2017**].

[8] http://in.relation.to/2018/05/11/hibernate-orm-5114-final-release/

2.4. OCR « La reconnaissance optique de caractères »

OCR[9] *: optical character recognitionn en Anglais ou* La reconnaissance optique de caractères en français.

Quelque informations sont indisponibles dans les documents que nous avons obtenue par la bibliothèque Proquest, et des fois nous ne pouvons pas copier ces informations car nous avons téléchargé sous format d'une photo (.JPG), pour cela nous avons trouvé une solution par les convertir d'une photo vers un texte par l'utilisation d'un OCR qui se trouve gratuitement sur internet sur le site cité au pied de la page.

Et la figure suivante présente la page d'accueil du site choisie d'où nous choisissons la photo que nous voulons la convertir et aussi nous choisissons la langue (Arabe, Anglais ou les deux).

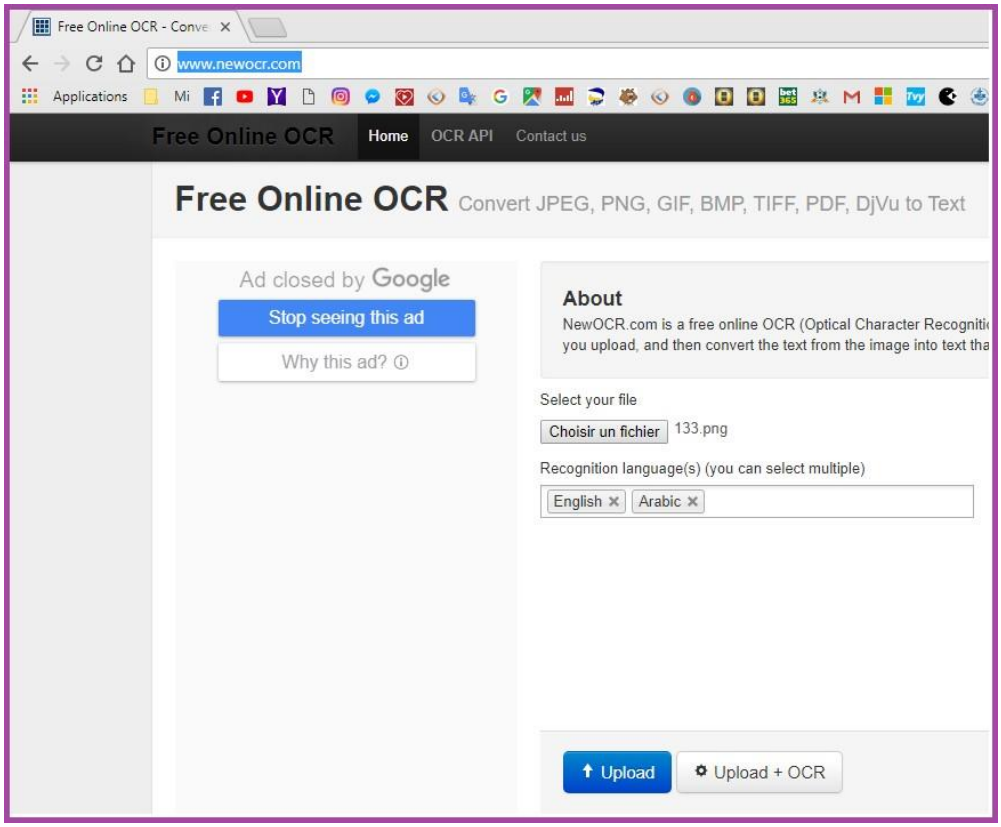


FIGURE 4.4 : Page d'accueil du site free Online OCR

[9] http://www.newocr.com/

Et la figure suivante explique un exemple de transfert d'un document en arabe sous image vers un texte.

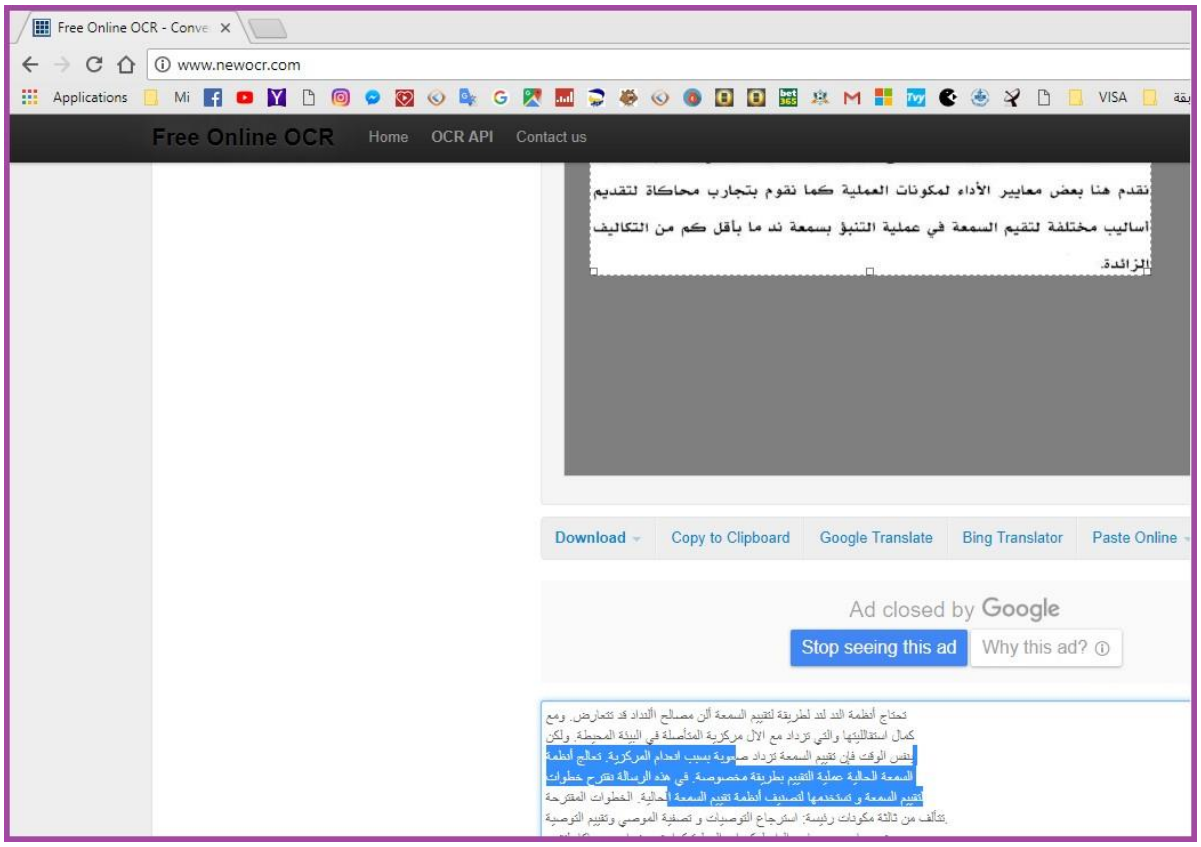


FIGURE 4.5 : Fonctionnement de free Online OCR

3. Présentation du site web

Pour l'implémentation de notre approche nous avons développé un site web en PHP à l'aide de l'environnement HTML, le simple site nous offrons une page d'accueil qui donne aux utilisateurs la possibilité de choisir la langue de la recherche, soit Arabe ou bien Anglais, Ensuite l'utilisateurs doit choisir une requête permis les 161 Requêtes que nous avons extrait précédemment, la figure suivante représente la page d'accueil du site web développer.

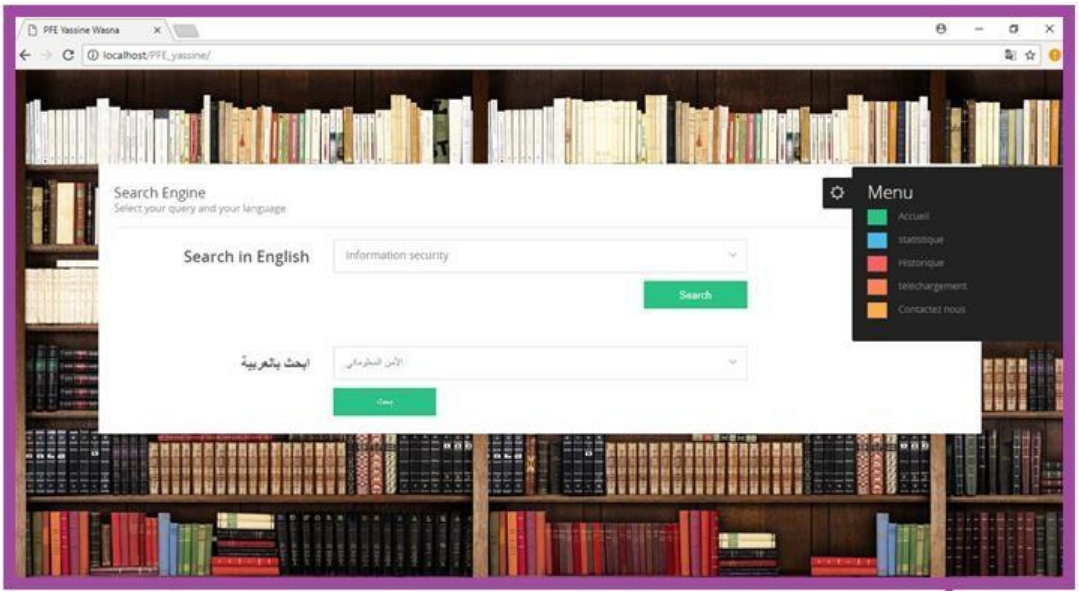


FIGURE 4.6 : Page de recherche du site web réalisé

Apres le choix de la requête, le site web affiche les résultats par 5 champs :

- Titre du document
- Université
- Année de la création du document
- Domaine du document
- Un bouton qui permette de télécharger le document sous format PDF

Les deux figures suivantes représentent la page de résultat pour la recherche en arabe et en anglais respectivement :

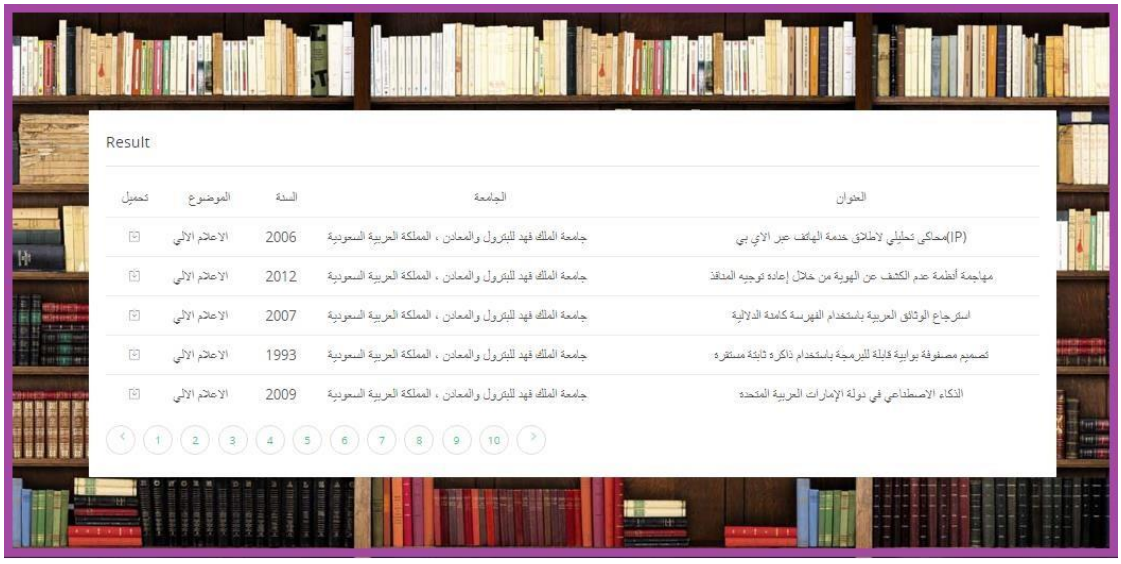


FIGURE 4.7 : Page de résultat de la recherche en Arabe

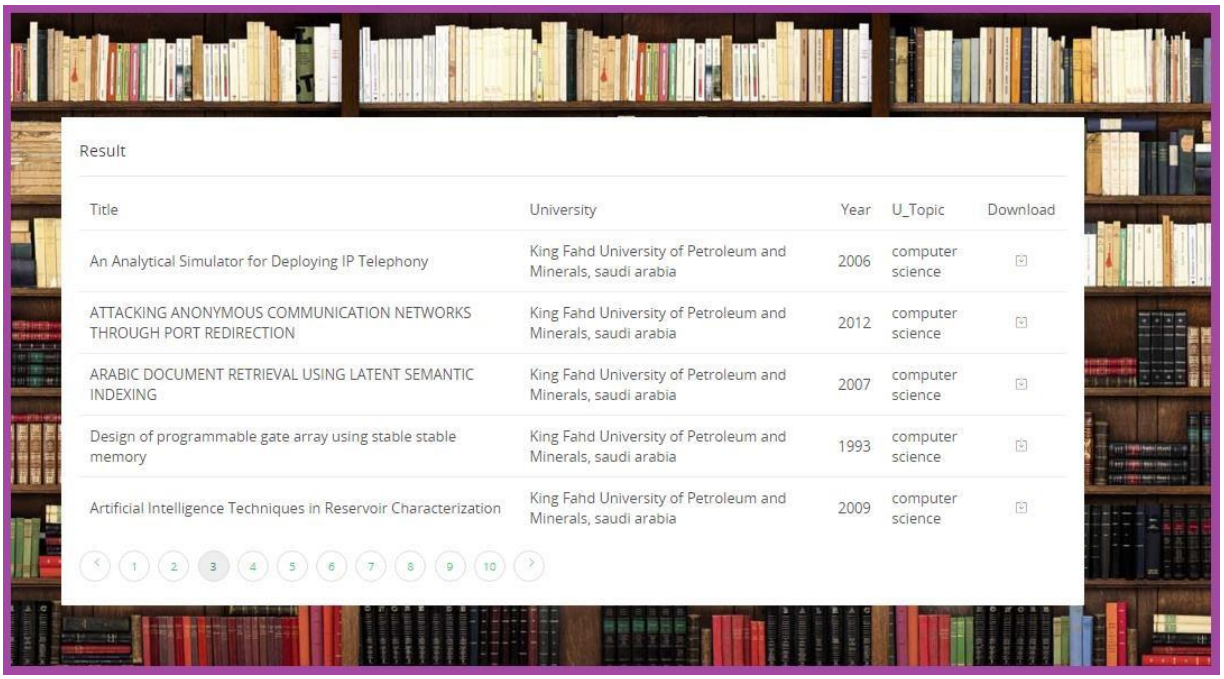


FIGURE 4.8 : Page de résultat de la recherche en Anglais

Aussi notre site web offre à l'utilisateur de télécharger la collection sous plusieurs formats, oui voir quelque statistique sur notre travail, la figure suivante donne un exemple sur quelques statistiques sur notre collection



FIGURE 4.9 : Page de statistique sur notre collection

4. Résultats de l'expérimentation

4.1. Résultat de pertinence par les moteurs de recherche

Apres l'utilisation des trois moteurs de recherches par les deux langues arabe et anglaise, nous avons eu six matrices de pertinences qui sont regroupés dans une seule matrice par l'algorithme de «Balanced interleaving».

Comme nous avons expliqué dans le chapitre précédent, Balanced interleaving Est un programme qui s'applique sur des matrices pour avoir une nouvelle matrice qui représente tous les matrices précédentes et qui donne un nouveau classement des documents [kharitonov, 2013].

La figure suivante explique le but de cet algorithme.

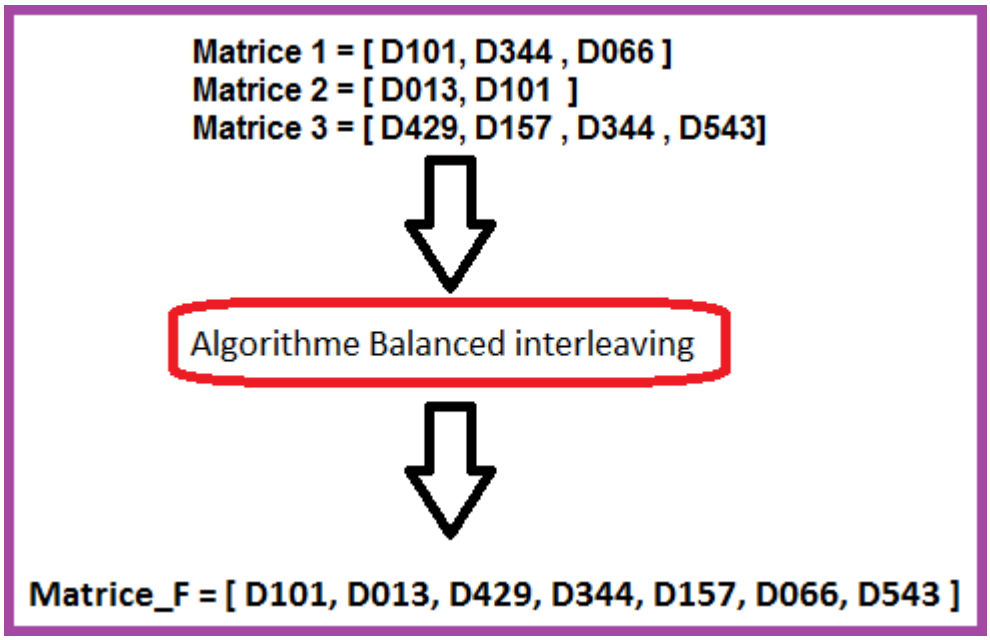


FIGURE 4.10 : Fonctionnement d'algorithme de «Balanced interleaving»

Aussi nous avons fait une validation manuelle par la lecture des résumés et les requêtes, et la décision est faite par l'être humain si cette requête appartient à un document ou non, donc la validation de l'être humain consiste seulement de valider le résultat du deuxième partie (les moteurs de recherche) car l'être humain ne peut pas valider tous le corpus, les résultats de l'être humain sont représenté dans la matrice final de la pertinence.

Le degré de pertinence final égal la moyenne des degrés de pertinence des six matrices, le tableau suivant explique comment nous avons calculé le degré dans la matrice de pertinence final.

Query	Document	Lucene Arabe	Lucene Anglais	Solr Arabe	Solr Anglais	Hibernate Arabe	Hibernate Anglais	validation	Resultat final
Q001	D090	0.4	0.2	0.4	0	0.4	0.2	oui	0.266
Q001	D103	0.2	0.2	0	0.6	0	0.4	non	0
Q001	D135	0.2	0.6	0.4	0.8	0.2	0.6	oui	0.466
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
Q161	D380	0.4	1	0.4	0.6	0.6	0.8	oui	0.633

TABLEAU 4.2 : Explication du calcul de la pertinence

Apres la validation par l'algorithme de «Balanced interleaving» et le calcul du degré de pertinence et aussi la validation de l'être humain, Nous avons eu la matrice finale de pertinence qui représente le résultat de la deuxième partie de notre approche.

le tableau suivant représente la matrice finale de pertinence :

Query	Document	Topic	Under Topic	Degree of relevance
Q001	D090	T6	UT17	0.266
Q001	D135	T3	UT11	0.466
Q001	D366	T1	UT03	0.066
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Q161	D380	T2	UT05	0.633

TABLEAU 4.3 : Matrice finale de pertinence

4.2. Résultats d'apprentissage automatique

L'apprentissage est la dernière étape de notre approche, et comme nous avons dit précédemment que dans cette étape nous avons fait deux méthodes, les résultats de la première étape sont ignorés car cette méthode n'a pas donné des bons résultats, pour la deuxième étape les résultats sont satisfaisants.

Query		Q001	Q001	Q002	Q002	.	.	.	Q161	Q161
Méthode		SRI	Apprentissage	SRI	Apprentissage	.	.	.	SRI	Apprentissage
partie modèle	D001	0.415	/	0	/	.	.	.	0.516	/
	D002	0	/	0	/	.	.	.	0	/
	D003	0	/	0.815	/	.	.	.	0.448	/
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	D200	0.388	/	0	/	.	.	.	0	/
partie Test	D201	0.826	Oui	0	non	.	.	.	0.219	oui
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	D600	0	Non	0.215	non	.	.	.	0	non

TABLEAU 4.4 : Matrice de pertinence du 1ere méthode

Nous avons fait un modèle d'apprentissage de Naïve Bayes par 200 document (les 200 premiers documents) ce qui représente un tiers de nombre total des documents et le reste (400 documents) qui représente deux tiers des documents pour la partie test.

Nous avons ignoré ce test car il n'a pas donné des bons résultats, la majorité des résultats obtenue sont (n'appartient pas) car permis les 200 documents de la partie modèle, nous avons trouvé au maximum 50 documents qui appartient et les autre n'appartient pas.

Cette différence a donné une avance pour la classe (n'appartient pas) ce qui ne donne pas une égalité entre les deux classes.

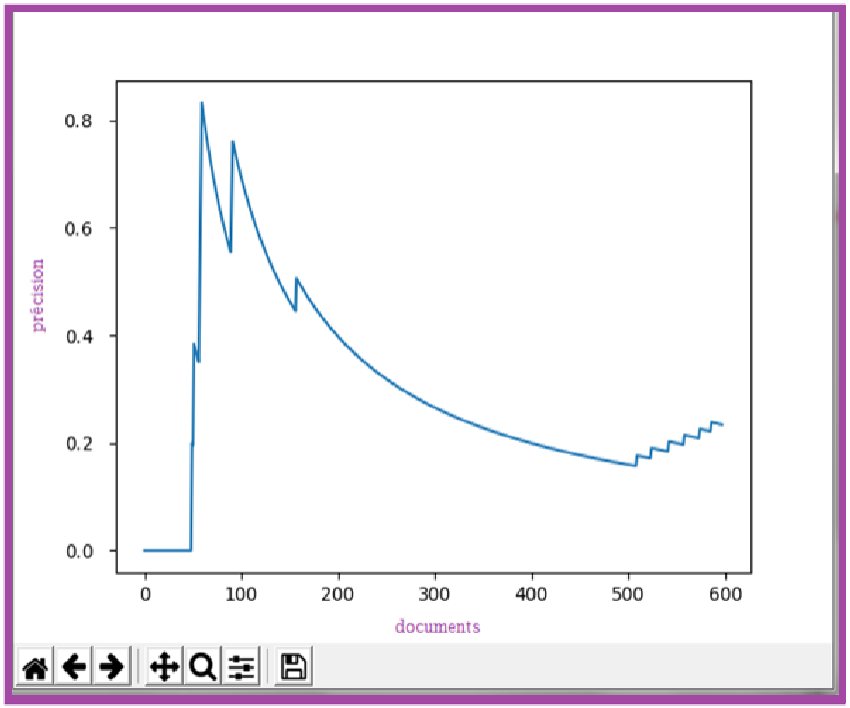


FIGURE 4.11 : Courbe de la précision de la requête Q001 par la deuxième méthode d'apprentissage automatique

La figure précédente représente une courbe de la précision des 600 documents, nous remarquons que la précision égal à zéro pour les 40 premières documents car ces documents sont utilisé pour la partie du modèle, ensuite nous remarquons que la précision est enlevé au début des documents retourné après la précision diminue jusqu'à 0.2, depuis cette courbe nous jugeons que la deuxième méthode d'apprentissage automatique a donné des bon résultats pour la première requête.

La première méthode n'a pas donné des résultats voulus donc nous avons décidé de faire un apprentissage automatique par une condition d'égalité entre les deux classes, pour cela nous avons choisi pour la partie modèle un tiers des documents pertinents et le même nombre des documents non pertinent, et nous avons laissé les documents restants pour la partie test.

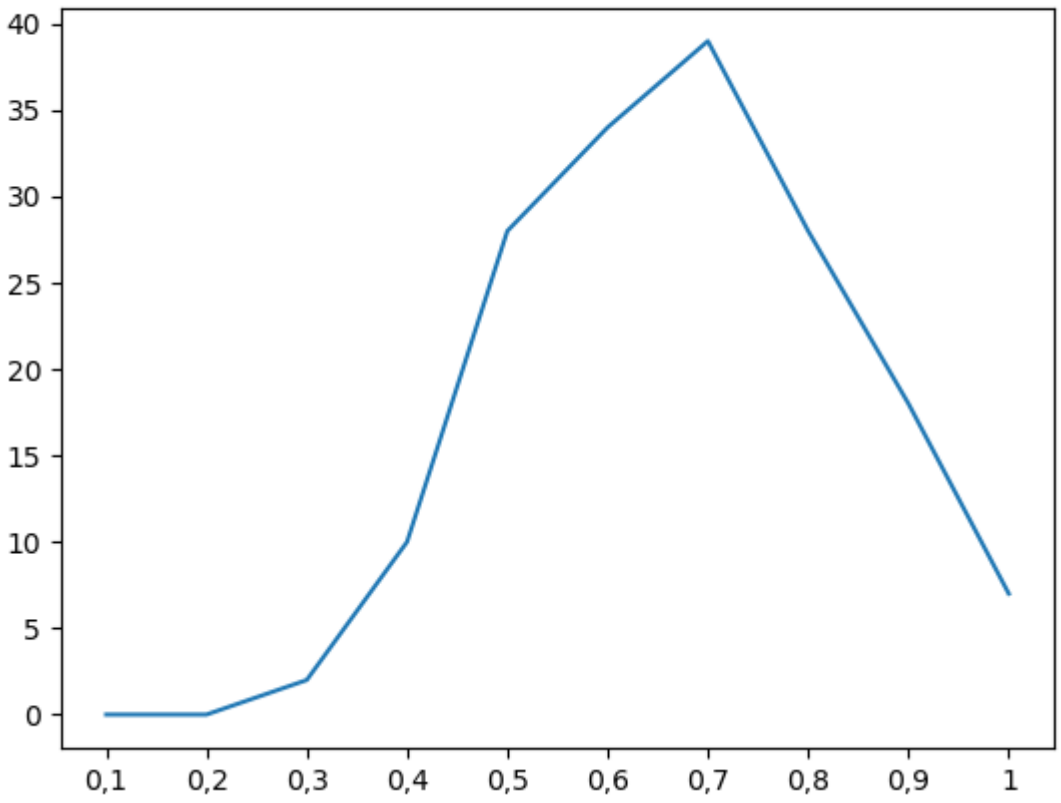


FIGURE 4.12 : Courbe de la moyenne de F-mesure des 161 requêtes

Le tableau suivant explique le choix des deux parties modèles et la partie du test dans les deux méthodes d'apprentissage automatique.

La méthode		méthode1			méthode 2		
Query	Doc Pert	Nb Doc Test	Test (oui)	Test (non)	Nb Doc Test	Test (oui)	Test (non)
الأمن المعلوماتي	44	200	12	182	30	15	15
اتصالات الشبكة الذكية	42	200	15	185	28	14	14
أداء أنظمة الاتصالات	34	200	11	189	22	11	11
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
نمو الحبوب	17	200	5	195	12	6	6

TABLEAU 4.5 : Comparaison entre le choix des parties testS dans les deux méthodes d'apprentissage automatique

4.3. Comparaison entre les résultats de pertinence des moteurs de recherche et d'apprentissage automatique

Pour évaluer les résultats des deux dernières étapes du projet (pertinence par moteur de recherche et par apprentissage automatique) nous avons fait une comparaison entre les deux étapes par le calcul des documents pertinents (résultat des moteurs de recherche) et le nombre des documents pertinent après la validation de l'être humain.

Ensuite le nombre des documents à chercher (partie test d'apprentissage) et le nombre des documents retourné par l'apprentissage et finalement le nombre total des documents pertinent retourné qui représente le noyau de notre travail.

Le tableau suivant donne chaque mesure avec sa moyenne :

Mesure	Moyenne
doc pertinents (SRI)	**33.383**
doc après validation (SRI)	**29.956**
doc du partie test (Apprentissage)	**19.890**
doc sélectionnés (Apprentissage)	**17.885**
doc pertinent sélectionné (Apprentissage)	**11.886**
Précision	**0.663**
Rappel	**0.597**
F-mesure	**0.627**

TABLEAU 4.6 : Moyenne des mesures de performance

Nous remarquons que les résultats de la deuxième méthode d'apprentissage automatique sont satisfaisants par rapport aux résultats des moteurs de recherche, car la précision et le rappel sont supérieure de 0.5, aussi pour la F-mesure qui représente la relation entre le rappelle et la précision, ça moyenne du tous l'échantillon égal à 0.627.

Avec tous les résultats obtenus nous pouvons dire que l'apprentissage automatique à donner des bons résultats, mais aussi ça reste des résultats limités qui ne sont pas vraiment proche des résultats des moteurs de recherche.

Le tableau suivant contient les 5 mesures que nous avons expliquées précédemment pour quelque requête de notre projet.

Technique	Technique 1 par SRI		Technique 2 par Apprentissage Automatique		
Query	doc pertinents	doc après validation	doc du partie test	doc sélectionnés	doc pertinent sélectionné
الأمن المعلوماتي	50	44	29	25	18
اتصالات الشبكة الذكية	50	42	28	26	19
أداء أنظمة الاتصالات	41	34	23	29	13
استثمار أموال الوقف	50	38	25	36	17
استعمال الانزيمات	37	29	19	26	9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
نظام الشبكات اللاسلكية	36	31	20	26	14
نمو الحبوب	22	17	11	18	7

TABLEAU 4.7 : Comparaison entre quelque requête par apport aux mesures de performance

Le tableau précèdent ne peut pas expliquer ou donner un point de vue sur les résultats obtenue pour cela nous avons devisé les résultats obtenue par rapport au nombre des mots des requêtes.

Le tableau suivant donne une vue sur la différence des mesures de performance entre les déférentes catégories des requêtes par rapport au nombre des mots de requêtes.

Query / Mesure	un-gramme	Bi-gramme	Trigramme	4-gramme	moyenne
doc pertinents	29.15	34.23	32.54	37.02	33.383
doc après validation	26.53	30.17	29.88	35.36	29.956
doc du partie test	17.09	20.33	19.27	23.66	19.890
doc sélectionnés	16.02	18.33	15.94	24.15	17.885
doc pertinent sélectionné	10.44	12.06	10.19	19.03	11.886
Rappel	0.610	0.593	0.528	0.804	0.595
Précision	0.651	0.651	0.657	0.639	0.652
k-mesure	0.630	0.621	0.586	0.712	0.620

TABLEAU 4.8 : Tableau de comparaison entre les déférentes catégories des requêtes

Nous remarquons que les mesures de performance sont plus satisfaisantes dans les requêtes en 4-gramme par rapport aux autres requêtes.

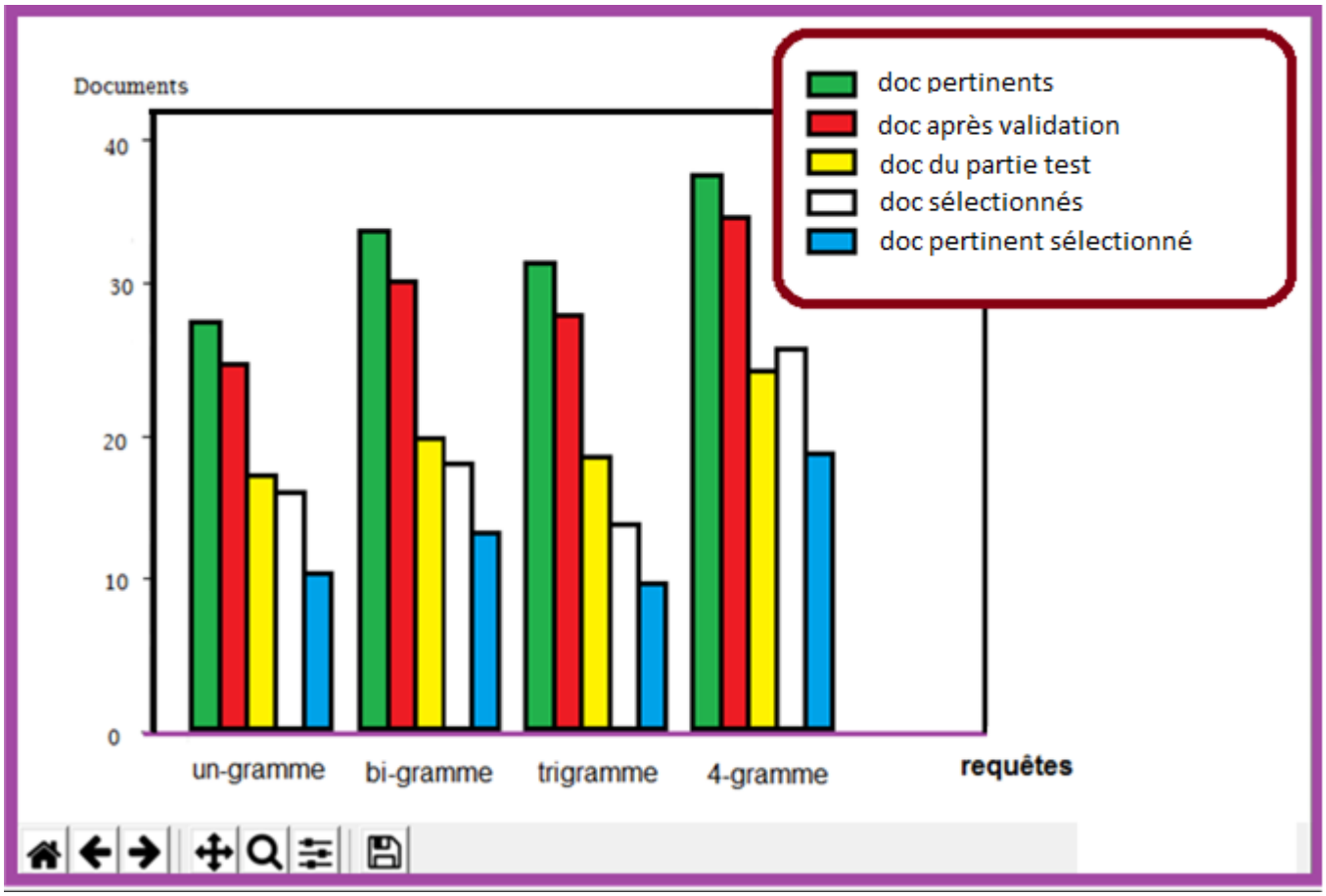


FIGURE 4.13 : Histogramme sur les différentes catégories des requêtes

La figure précédente montre que les requête en 4-gramme nous donnons des résultats plus satisfiable par apport au d'autre requêtes, car les requêtes en 4-gramme sont plus détaillé car elles contiennent 4 mots, par contre une requête qui contient seulement un mot n'est pas vraiment détaillé, c'est un mot général et la recherche sur un seul mot ne donne pas vraiment les résultats que nous cherchons.

La validation de l'être humain représente 89.73% par rapport aux résultats retournés par les moteurs de recherche, pour cela cette partie est importante dans la recherche d'information, la validation de l'être humain élimine les documents retourné par les moteurs mais qui sont par pertinent par rapport au requête.

59.75% est le pourcentage de la réussite de la deuxième méthode d'apprentissage automatique par rapport aux résultats retournée par les moteurs de recherche et valider par l'être humain, ce pourcentage est satisfait est ne pouvons utiliser l'apprentissage automatique à la place des systèmes de recherche d'information, mais cette méthode reste limité et ne donne pas des résultats épicés.

5. Conclusion

Dans ce chapitre nous avons parlé sur l'implémentation et l'évaluation de notre projet, nous avons conclu que l'apprentissage automatique par l'algorithme de naïve bayes a donné des résultats acceptables et satisfait.

Nous avons présenté dans ce chapitre les différentes ressources utilisées, comme Pyhton et ces différentes bibliothèques, Aussi l'environnement de développement NetBeans qui aide à manipuler les codes en Java, ensuite nous avons expliqué le fonctionnement des trois moteurs de recherche utilisé : Lucene, Solr et Hibernate, et finalement nous avons présenté l'OCR qui nous a aidés à convertir les photos en textes.

Après les ressources, nous avons présenté l'outil qui explique les résultats de notre travail, il est représenté dans un site web qui offre à ces utilisateurs de faire des recherches sur des projets de fin d'étude, la recherche se fait depuis des requêtes connues dans les différents domaines, et la recherche se fait en bilingue (Arabe et Anglais), le site web aussi donne la possibilité de télécharger les documents en format PDF.

Finalement 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.

Conclusion et perspectives

Dans notre projet nous avons obtenu les trois objectifs que nous a parlés au début du projet, nous avons créé une collection bilingue (arabe et anglaise) qui contient 600 documents et 161 requêtes, la collection permet de faire des évaluations sur les algorithmes et les techniques des moteurs de recherche, la collection contient au total plus de 80 000 mots en arabe.

La collection est téléchargeable sous 5 formats : Xml, Txt, Excel, Csv et Sql, elle permet aux étudiants universitaires ou les chercheures de faire des tests sur leurs algorithmes de recherche et faire des comparaisons avec nos résultats.
La collection aussi a été utilisée pour construire un site web qui offre la recherche sur des projets de fin d'étude dans les différents domaines et en bilingue arabe et anglais, Le site web a aussi l'avantage de télécharger la collection ou de consulter les différentes statistiques sur le travail et la collection.

Le troisième objectif représente la comparaison entre les résultats de la recherche par des moteurs de recherche open source et par l'apprentissage automatique, l'apprentissage a sélectionné presque de 60% des résultats sélectionnés par les moteurs de recherche et ce pourcentage est positif qui dépasse les 50%.

Les résultats obtenus par l'apprentissage automatique sont satisfiable car nous avons marqué une moyenne de précision qui égale à 0.663 et la moyenne des rappels qui égale à 0.597 et même la valeur moyenne de f-mesure qui a dépassé 60%, les valeurs de ces trois mesures exprime la validité de notre travail.

Nous avons eu quelque difficulté à la récupération des documents qui continents des résumés en arabe et en anglais, mais nous avons surmonte sur ce problème par l'utilisation de la bibliothèque électronique Proquest, aussi nous avons eu des difficultés d'encodage, par exemple si nous copions des résumés en arabe, nous obtenons des caractères incompréhensible ou des point d'interrogation, ce problèmes est résolu par l'utilisation des OCR ou par fois d'écrire les résumés en arabe manuellement.

Python il nous à aider à éliminer les mots vides et à construire les différentes formats de la collection, aussi python facilite les calculs des différentes formules utilisé au projet.

Les résultats obtenue sont satisfaisable, mais nous pouvons avoir des résultats plus convaincant par l'ajout des documents qui construit la collections et par l'utilisation des ordinateurs puissants pour utiliser les différents algorithmes d'apprentissage automatique ou d'utiliser plus de trois moteurs de recherche pour valider la collection.

Nous avons appris dans ce travail beaucoup de points positives, nous avons développé nos capacités au domaine de la recherche d'information et nous avons aussi connue plusieurs nouveaux algorithmes et méthodes de la RI ou en apprentissage automatique, nous avons rencontré par plusieurs travaux de recherche sous format d'article ou des thèses de doctorat dans le domaine.

Comme des perspectives nous voulons améliorer notre travail par l'ajoute des autres moteurs de recherche open source, et aussi de faire l'apprentissage avec d'autres techniques, et le plus important est d'ajouter des documents pour avoir une collection plus vaste.

ANNEXE 1

Requetés :

Une requête permet de poser une question sur les données stockées dans les tables ou d'accomplir une action sur ces données, elle est composé d'un ensemble de mots, elle peut être développé par les systèmes de recherche d'information par l'ajout des synonymes.

Dans notre travail nous avons fait une extraction des requêtes depuis les titres et les mots clés des documents et nous avons expliqué cette étape sur le troisième chapitre.

Nous présentons maintenant le tableau global des 161 requêtes :

Id Query	Query Ar	Query
Q001	الأمن المعلوماتي	Information security
Q002	اتصالات الشبكة الذكية	Smart Network Connections
Q003	أداء أنظمة الاتصالات	Performance of communication systems
Q004	استثمار أموال الوقف	Investment of Waqf funds
Q005	استعمال الانزيمات	The use of enzymes
Q006	أعمال الصيانة	Maintenance work
Q007	اكسيد الألمنيوم	Aluminum oxide
Q008	الأثر الإسلامي	Islamic Impact
Q009	الأجسام المضادة للفيروسات	Antivirals
Q010	الاستدامة	Sustainability
Q011	الاصلاح الأسري	Family Reform
Q012	الإطارات الصناعية	Industrial Tires
Q013	الاكسجين	Oxygen
Q014	الألياف الزجاجية	Fiber glass
Q015	الأمراض	Diseases
Q016	الانتشار الحراري	Thermal diffusion
Q017	الإنجليزية	English
Q018	الأنظمة	Systems
Q019	البحث عبر الإنترنت	Search online
Q020	البرمجيات	Software
Q021	البلاطات الخرسانية	Concrete slabs
Q022	البنسلين	Penicillin
Q023	البنية التحتية	Infrastructure
Q024	البيئة	The environment
Q025	التأمل البصري	Visual meditation
Q026	التأمين	Insurance
Q027	التحليل الكيميائي	Chemical analysis
Q028	التداخلات الكهرومغناطيسية المتولدة	Electromagnetic interference generated

Q029	التدفق النقدي	Cash flow
Q030	التربية الإسلامية	Islamic education
Q031	الترجمة الآلية	Automatic translation
Q032	الترجمة العربية	Arabic translation
Q033	التعرف الآلي	Automatic identification
Q034	التكسرات الطبيعية	Natural crackers
Q035	التوافقيات	Harmonics
Q036	التيار المستمر	DC power supply
Q037	الجمل العربية	Arabic sentences
Q038	الجودة و الضمان	Quality and Warranty
Q039	الحاسبات فائقة السرعة	High speed computers
Q040	الحركة المرورية	traffic
Q041	الحقل الصوتي	Audio field
Q042	الحلول الكلاسيكسة	Classic solutions
Q043	الحمض النووي	DNA
Q044	الخصائص المجهرية	Microscopic properties
Q045	الخصائص الميكانيكية	Mechanical properties
Q046	الخصائص الميكانيكية والحرارية	Mechanical and thermal properties
Q047	الخط العربي المطبوع	Printed Arabic Calligraphy
Q048	الخلية و المحفزات	Cell and Catalysts
Q049	الخوارزميات	Algorithms
Q050	الخواص الجيوتقنية	Geotechnical properties
Q051	الخواص الميكانيكية	Mechanical properties
Q052	الخورزميات المتطورة	Advanced algorithms
Q053	الدولة الإسلامية	Islamic country
Q054	الذكاء الاصطناعي	Artificial intelligence
Q055	الراحة البصرية	Visual comfort
Q056	السحنات الحيوية المجهرية	Micro - biological loads
Q057	السلوك التقاربي	Approximate behavior
Q058	الشبكات الاجتماعية	social networks
Q059	الشبكات العصبية الاصطناعية	Artificial Neural Networks
Q060	الشبكات العصبية المتكيفة	Adaptive Neural Networks
Q061	الشريعة	Sharia
Q062	الشيخوخة	Aging
Q063	الضغط العالي	High pressure
Q064	الطاقة الكهربائية	electricity
Q065	العلماء المسلمين	Muslim scholars
Q066	العمليات الحسابية	mathematical calculations
Q067	الفرضيات المعارضة	Opposition assumptions
Q068	الفساد المالي	Financial corruption
Q069	الفضة	Silver
Q070	القرآن الكريم	The Holy Quran
Q071	القواطع الغازية	Gas Breakers
Q072	الكاثوليك والشيعة	Catholics and Shiites
Q073	الكثافة	Density
Q074	الكهروكيميائية	Electrochemical
Q075	اللغة العربية	Arabic

Q076	المبادلات الحرارية	heat exchangers
Q077	المباني الإدارية	Administrative buildings
Q078	المبيدات الحشرية	Pesticide
Q079	المتغيرات العشوائية اللوغارثمية	Logarithmic random variables
Q080	المجال الكهربائي	Electric field
Q081	المجتمع القطري	Qatari society
Q082	المحروقات	Fuels
Q083	المرافق الطبية	Medical Facilities
Q084	المركبات العضوية الخفيفة	Light organic compounds
Q085	المسح الضوئي	Photo scan
Q086	المسلم السني	The Sunni Muslim
Q087	المصارف الإسلامية	Islamic Banks
Q088	المصفوفات العشوائية	Random arrays
Q089	المعادلات التراجعية	Equation equations
Q090	المعادلات التفاضلية	Differential Equations
Q091	المعادلات العشوائية	Random equations
Q092	المعالجة البيولجية	Biological treatment
Q093	المقاولون	Contractors
Q094	المقاومة الكهربائية	electrical resistance
Q095	المكثفات	Capacitors
Q096	المملكة العربية السعودية	Kingdom of Saudi Arabia
Q097	المناخ الحار	Hot climate
Q098	المناخ الرطب	Wet climate
Q099	المنكشفات الصخرية	Rocky Expositions
Q100	المنهاج القرآني	Quranic curriculum
Q101	الموجات الصوتية	Sound waves
Q102	المؤسسات المالية	financial institutions
Q103	النظريات النفسية	Psychological theories
Q104	النفط الخام	Crude Oil
Q105	الوثائق العربية	Arabic Documents
Q106	الوقود البترولي	Petroleum fuels
Q107	أنابيب الكربون	Carbon Pipes
Q108	انظمة الاتصالات	Communication Systems
Q109	أنظمة الاتصالات اللاسلكية	Wireless Communication Systems
Q110	أنظمة التحكم في الرياح	Wind Control Systems
Q111	أنظمه البرمجيات	Software systems
Q112	اهتزازات الآلات الدوارة	Vibrating rotary machines
Q113	أوبك	OPEC
Q114	اول أكسيد الكربون	Carbon Monoxide
Q115	بروتوكول الانترنت	Internet Protocol
Q116	تاريخ الأديان المعاصر	History of Contemporary Religions
Q117	تاكل الحديد	Corrosion of iron
Q118	تحلية مياه البحر	Seawater desalination
Q119	تحليل البلاستيك الصناعى	Analysis of industrial plastics
Q120	تحويل نماذج البرامج	Converting software models

Q121	ترشيد إستهلاك الطاقة الكهربائية	Rationalize electricity consumption
Q122	تركيبة جيولوجية معقدة	Complex geological composition
Q123	تصنيف الأديان	Classification of religions
Q124	تطوير البرمجيات	Software development
Q125	تطوير محرك بحث	Search engine development
Q126	تغير المناخ	Climate change
Q127	تقسيم الشيفرة	Divide the code
Q128	تنسيق ترجمة	Translation format
Q129	ثاني أكسيد الكربون	Carbon Dioxide
Q130	حركة المرور	traffic movement
Q131	حسابات تحليلية	Analytical calculations
Q132	حماية البيانات	Data protection
Q133	حوادث الطرق	Road accidents
Q134	خدمة الهاتف	Phone service
Q135	خوارزميات فعالة	Effective algorithms
Q136	خوارزمية مستعمرة النمل	Ant colony algorithm
Q137	دول مجلس التعاون الخليجي	GCC countries
Q138	دولة الإمارات العربية المتحدة	State of United Arab Emirates
Q139	رأس المال الإنتاجي و البشري	Productive and human capital
Q140	سرعة الصوت	The speed of sound
Q141	شبكات	Networks
Q142	شبكات نقل الطاقة	Power transmission networks
Q143	شبكة البيانات	Data Network
Q144	فرضيات دولة العقيدة	Hypotheses of the state of faith
Q145	كتب التفسير	Books Interpretation
Q146	كتلة الجاذبية	Gravity mass
Q147	كشف الأخطاء	Detection of errors
Q148	مثبطات الحبوب	Cereal inhibitors
Q149	محطة كهرباء	electricity station
Q150	مرضى السكر	Diabetics
Q151	مشاريع بناء	Building projects
Q152	مكافحة التلوث النفطي البحري	Oil Pollution Control
Q153	مكيفات الهواء	Air conditioners
Q154	ملفات متعددة الابعاد	Multidimensional files
Q155	ملوثات الهواء	Air pollutants
Q156	منظومة إدارة المبنى	Building Management System
Q157	موجات الضوء	Light waves
Q158	مياه الصرف الصحي	Sewage
Q159	نظام الأرصاد الجوية الآلية	Automated Meteorological System
Q160	نظام الشبكات اللاسلكية	Wireless Networking System
Q161	نمو الحبوب	Grain growth

ANNEXE 2

Document de la collection sous format de XML :

Comme nous avons dit précédemment que nous avons construit notre collection en 5 formats, aussi nos documents et nos requêtes sont de 5 formats et nous voulons donner un exemple sur le format de XML (*Extensible Markup Language)* pour un document, nous avons fait par ce langage de balise 11 informations sur chaque document : id, titre en arabe, titre en anglais, université, année, résumer en arabe, résume en anglais, mots clés en arabe et en anglais, domaine et finalement sous domaine.

Nous présentons maintenant le fichier XML du document numéro 534 :

<doc>

<id> 534 </id>

<title_ar> تأثير الطفل النانوي على خواص اسمنت آبار النفط للتطبيقات عالية الضغط و الحراره </title_ar>

<titre_eng> Effect of Nano clay on oilwell cement properties for high pressure and temperature applications" </titre_eng>

<university>KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN , SAUDI ARABIA </university>

<year>2013</year>

<abstract_ar>مع الزيادة المتواصلة في الطلب على البترول، إتجهت الشركات للبحث في أعماق بعيدة ذات ضغط و حرارة عاليين،

و في ظروف مماثلة، يتعرض التغليف الاسمنتي للآبار لإجهادات عالية ، و التي قد تؤثر سلباً على ترابط الاسمنت ولذلك، فإن التصميم الأمثل للتركيبة الاسمنتيه أصبح بغاية الأهمية بحيث أنه يضمن إستمرارية و تماسك التغليف الإسمنتي لأطول فترة ممكنة، و للقيام بهذه المهمة، تم دراسة و إستخدام عدد مقدر من المواد المضافة في الآبار ذات الظروف المماثلة.

في الآونة الأخيرة، أثبتت المواد الدقيقة الحجم، أو ما يعرف بالمواد النانوية، فعاليتها في مختلف المجالات، كالنسيج و الدفاع و الفضاء و الطاقة، و قد اصبحت الان مواد تجارية واسعة الإنتشار، لها القدرة على إيجاد الحلول للعديد من

المصاعب التقنية، و في صناعة البترول، وجدت هذه المواد إنتشاراً نسبة لتميزها بكِر مساحة السطح النوعية و صغر حجمها.

الهدف من هذه الدراسة هو تقييم الخواص الفيزيائية للاسمنت السعودي فئة جي بوجود المواد النانوية و مواد مضافة اخرى كالمواد المانعة لفقدان السوائل و المثبطات و المواد الموسعة، إضافة الى مانعات الإحتكاك، كل ذلك تحت درجات عالية من الحرارة و الضغط العالى. في هذه الدراسة، تم إضافة المواد النانوية بنسب مختلفة من التركيبة الكلية للخليط الاسمنتي المستخدم في آبار النفط و الغاز في المملكة العربية السعودية، و قد تمت دراسة مختلف الخواص المتعلقة بالخليط مخبرياً، كالخواص المتعلقة بزمن زيادة السماكة وفصل الماء الحر و الخواص الريولوجية المختلفة و قوة الضغط و الكثافة و النفازية و المسامية، إضافة الى بعض الإختبارات المجهرية كالمسح الإلكتروني و الحيود السيني.

من خلال النتائج المخبرية، تم إختيار افضل تركيبة للخليط الإسمنتي المحتوية على نسبة معينة من المواد النانوية. من خلال الإختبارات السابق ذكرها، وجد أن المواد النانوية المضافة لها المقدرة على تحسين الخواص الريولوجية و الميكانيكية، إضافة الى التقليل من المحتوى المائي والتثبيط من زيادة السماكة و تقليل الكثافة. كل هذه التحسينات تصب في تقوية التركيبة الدقيقة للإسمنت و تساعد على تصميم نظام إسمنتي جديد له القدرة على تحمل ظروف الضغط و الحرارة العاليين و قابل للتطبيق حقلياً بدون آثار عكسية.

</abstract_ar>

<abstract> With the continued growth of petroleum demand, the oil and gas companies are searching in new or unexplored areas. The search is getting extreme in terms of depth, temperature and pressure. In deeper wells, high temperature and pressure and post cementing operations put extreme stresses on the cement sheath and affect the integrity of the cement. In such conditions, the design of cement slurry is very critical and it must possess properties which ensure the durability and long term integrity of cement sheath.

To resolve this problem, many additives are being investigated and used in high pressure and temperature wells. Recently, Nano Materials have demonstrated effectiveness across a wide variety of industries, from textiles and defense to aerospace and energy. They are now being used as commercially worthwhile solutions to technical challenges faced many industries. Having high surface area and small size, attracts them to petroleum industry. Their application in petroleum industry is progressing in a last few years.

The objective of this study is to evaluate the physical properties of Saudi cement class G with Nano clay addition and admixed with different additives such as fluid loss agent, retarder, expanding agent and friction reducing agent under high temperature and pressure conditions. Nano clay is added at different percentages to the cement slurry xx mixture currently used for oil/gas well cementing in Saudi Arabia. Experimental investigations addresses different cement slurry properties such as thickening time, free water separation, rheological properties, compressive strength, density, porosity, permeability, and SEM and XRD analysis. The best slurry design is selected on the basis of results of different percentages of Nano clay

used in experiments. From the laboratory tests, it is investigated that addition of Nano clay improves the rheological and mechanical properties, reduces the free water contents, retards the thickening time, reduces the density, makes the strong microstructure of cement and helps in designing the new cement system which sustains the high pressure and temperature conditions and gives good results. This new cement design can be prepared and implemented in the field without adverse effects."

</abstract>

<keywords_ar> الحراره , الضغط , النفط , اسمنت </keywords_ar>

<keywords> Cement, Oil, Pressure, Heat </keywords>

<topic> ميدان الفيزياء و الكيمياء </topic>

<under_topic> هندسة البترول </under_topic>

</doc>

Références bibliographiques

[Baccini, 2010]
Alain Baccini, Analyse des critères d'évaluation des systèmes de recherche d'information, *Institut de Mathématiques de Toulouse*, *UMR 5219 - Université de Toulouse et CNRS, 2010.*

[Balipa, 2015]
Mamatha Balipa, Search Engine using Apache Lucene, *International Journal of Computer Applications (0975 – 8887) Volume 127 – No.9*, October 2015.

[Beigbeder, 2006]
Michel Beigbeder, évaluation des systèmes de recherche d'information, 23 octobre 2006.

[Bernard, 2017]
Emmanuel Bernard, Hibernate Search 5.7.3.Final, October 2017.

[Boucham, 2009]
Souhila BOUCHAM, Une approche basée Ontologies pour l'indexation automatique et la recherché d'information Multilingues (RIM), *Université m'hamed Bougara de Boumerdes*, Magister en informatique, 2009.

[Bouramoul, 2011]
Abdelkrim BOURAMOUL, Recherche d'information contextuelle et sémantique sur le web, *Université MENTOURI de Constantine*, 09 / 2011.

[Chapelle, 2012]
OLIVIER CHAPELLE, Large-Scale Validation and Analysis of Interleaved Search Evaluation, 2012.

[Doty, 2008]
Python Basics S.R. Doty August 27, 2008.

[El Hachani, 1998]
Mabrouka EL HACHANI, Les descripteurs complexes en indexation automatique arabe : Etude des structures invectivent extraites d'un corpus économique, *Ecole Nationale Supérieure des Sciences, Université Jean Moulin Lyon3*, mai 1998.

[Ferro et Peters, 2008]
Nicola Ferro et Carol Peters, From CLEF to Treble CLEF: the Evolution of the Cross-Language Evaluation Forum, *Proceedings of NTCIR-7 Workshop Meeting*, December 16–19, 2008, Tokyo, Japan.

[Hacene et Taleb, 2015]
Hacene Abdessamed et Taleb Mustapha Younes, Conception et réalisation d'un éditeur de texte Coranique authentique, *Alger : Ecole national supérieur de l'informatique*, 2015.

[Kadri, 2008]
Youssef Kadri, Recherche d'Information Translinguistique sur les Documents en Arabe, *Université de Montréal*, 2008.

[kharitonov, 2013]
Eugene Kharitonov, Using Historical Click Data to Increase Interleaving Sensitivity Yandex, *Moscow, Russia, School of Computing Science, University of Glasgow, UK*, 2013.

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

[Lemaire, 2014]
V. Lemaire, Apprentissage incrémental anytime d'un classifieur Bayésien naïf pondéré, *Orange Labs Lannion, 2 avenue Pierre Marzin, 22300 Lannion ,Conference Paper*, January 2014.

[Leray, 2002]
Philippe Leray, Réseaux Bayésiens pour la Classification Méthodologie et Illustration dans le cadre du Diagnostic Médical, *Université 76801 St-Etienne du Rouvray Cedex*, 2002.

[Rajani, 2013]
Rajani Maski. Using Apache solr for Ecommerce Search Applications, Happiest Minds, IT Services, 2013.

[Salton, 1971]
G. Salton, A comparison between manual and automatic indexing methods. Journal of American Documentation, 20(1): 61–71, 1971.

[Salton, 1983]
G. Salton, Introduction to Modern Information Retrieval, *McGraw-Hill, New York*, 1983.

[Smeaton, 1989]
A. F. Smeaton, Information retrieval and natural language processing, In proceedings of a conference jointly sponsored by ASLIB, *University of York,* page 2, Mars 1989.

[Strzalkowski, 1993]
T. Strzalkowski, Natural language processing in large-scale text retrieval tsks, in text Retrieval Conference (TREC-1), page 137, 1993.