REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ MATHEMATIQUES ET INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : [**Génie Informatique**]

Par :

[DJATTI Hanane]

Sur le thème

Moteur de Recherche via les Émotions associées

Soutenu publiquement le 13 /10 / 2021 à Tiaret devant le jury composé de :

Mme. Benathmane Lalia	M.A.A	Université IBN-KHALDOUN Tiaret	Président
Mme. LAKHDARI Aicha	M.A.A	Université IBN-KHALDOUN Tiaret	Encadrante
Mr. Adda Boualem.	M.C.B	Université IBN-KHALDOUN Tiaret	Examinateur

2020-2021

Remerciements

Tout d'abord, nous remercions Dieu Tout-Puissant de nous avoir accordé le courage, la patience et la force morale et physique pour pouvoir accomplir ce travail.

Ma gratitude s'adresse à Mme ***LAKHDARI AICHA*** *pour son encadrement, son orientation, ses conseils et la disponibilité qu'elle m'a témoigné pour me permettre de mener à bien ce travail.*

Je tiens à exprimer mess vifs remerciements à Mme Benathmane Lalia qui a accepté de présider le jury de soutenance, pour tout ce qu'elle a pu nous apprendre ; qu'elle trouve ici l'expression de ma profonde et sincère reconnaissance, Mme, pour m'avoir fait l'honneur d'accepter d'examiner ce travail.

J'adresses mes remerciements à tous les professeurs, pour leurs conseils et leurs critiques qui ont guidé mes réflexions durant mes recherches, et je remercie également tous mes collègues et amis du département d'informatique de l'université Ibn Khaldoun Tiaret.

Je souhaite exprimer mes profondes gratitudes à mes parents qui m' ont soutenus tout au long de mon projet, ainsi que toute la famille, les amis,... pour leur soutien indéfectible.

Enfin, Je remercie tous ceux qui m'ont aidé de près ou de loin dans l'élaboration de ce travail.

Dédicaces

*D*u plus profond de mon cœurs, je dédie ce travail à:

*M*es famille, DJATTI et BOUGANDJA, en particulier mes parents MEHANI ET MESSOUDA. Ce mémoire n'aurait pas vu le jour sans leurs encouragements au fil des ans.

À tous mes frères et sœurs surtout BOUTHAINA ET ASSMA.

À mes cousins et cousines, en particulier, sans oublier

Mes collègues BEDRANE SARAH, MACHOU RANIA, BEGOUGUI KHADIDJA Ahlem et Sebti Sara qui m'ont aidés tout au long de nos études.

À tous mes amis et aux personnes qui ont toujours cru en moi.

Résumé :

Notre travail se situe au carrefour des moteurs de recherche et analyse de sentiment qui est la Recherche d'Information orientée émotion. Nous nous intéressons, plus précisément, à l'amélioration des résultats de recherche en exploitant l'émotion dans le processus de recherche. Pour le faire, nous commençons d'abord par une approche de recherche classique bonifiée par une recherche émotionnelle. Cette dernière repose sur l'Analyse de la requête et Appariement document-requête. Cependant, l'émotion est intégrée en tant que source d'information additionnelle pour améliorer la pertinence des résultats.

Notre modèle d'appariement est un modèle probabiliste. Il effectue une correspondance entre la requête saisie et les documents déjà préparés pour un classement (Ranking) pertinent des résultats.

Nos expériences menées ont montré une augmentation du taux de satisfaction meilleur (Pertinence) par rapport aux systèmes de moteur de rechercher classiques

Mots clés: Moteur de recherche, Analyse de sentiment, Requête, Document, Pertinence

Abstract:

Our work is at the crossroads of search engines and sentiment analysis which is emotion-oriented Information Retrieval. More specifically, we are interested in improving search results by harnessing emotion in the search process. To do this, we first start with a classical research approach enhanced by emotional research. The latter is based on the Analysis of the query and Document-query matching. However, emotion is integrated as an additional source of information to improve the relevance of the results. Our matching model is a probabilistic model. It makes a correspondence between the entered query and the documents already prepared for a relevant ranking of the results. Our experiments have shown an increase in the best satisfaction rate (Relevance) compared to conventional search engine systems

Keywords: Search engine, Sentiment analysis, Query, Document, Relevance

Sommaire

Table des Figures **i**

Liste des tableaux **iii**

Introduction générale 1

Chapitre I Le Domaine De La Recherche D'information CLASSIQUE

I.1. Introduction 4

I.2. Définitions 4

I.3. Concepts de base de la RI 5

I.3.1. Collection de documents 5

I.3.2. Document 5

I.3.3. Besoin en information 5

I.3.4. Requête 6

I.3.5. Pertinence 6

I.3.5.1. La pertinence algorithmique (ou système) : 6

I.3.5.2. La pertinence thématique 6

I.3.5.3. La pertinence cognitive 6

I.4. Indexation 7

I.4.1. L'extraction des mots 8

I.4.2. L'élimination des mots vides 8

I.4.3. La lemmatisation 8

I.4.4. La pondération 8

I.4.4.1. TF (Term Frequency) 9

I.4.4.2. IDF (Inverse Document Frequency) 9

I.4.5. Appariement document-requête 9

I.4.5.1. Modèle booléen 10

I.4.5.2. Modèle vectoriel 10

I.4.5.3. Modèle probabiliste 11

I.4.6. Reformulation de la requête 12

I.5. Les applications de la RI 13

I.5.1. Filtrage des e-mails (spam) 13

I.5.2. Application Biomédicale 13

I.5.3. La classification des textes 13

I.5.4. .Moteur de recherche 13

I.6. Moteur de recherche 14

I.6.1. Historique des moteurs de recherche (MRs): 14

Sommaire

I.6.2. Architecture générale des premiers moteurs de recherche. 15
I.6.3. Architecture distribuée et adaptative. 16
I.6.4. Fonctionnement des moteurs de recherche 17
I.6.4.1. L'exploration 17
I.6.4.2. L'analyse et l'indexation 17
I.6.4.3. Le classement 18
I.7. Les moteurs de recherche généralistes classiques 19
I.7.1. Google 19
I.7.1.1. Architecture du moteur de recherche Google 19
I.7.1.2. Fonctionnement de Google : 20
I.7.1.3. Les avantages et les Inconvénients de Google 21
I.7.2. Bing 22
I.7.2.1. Fonctionnement de Bing 22
I.7.2.2. Caractéristiques du moteur Bing 23
I.7.2.3. Avantages et inconvénients de Bing 23
I.7.3. Bing Vs Google 24
I.7.4. Yahoo 24
I.7.4.1. Fonctionnement de Yahoo 24
I.7.4.2. Caractéristiques du moteur Yahoo 25
I.7.4.3. Avantages et inconvénients de Yahoo 25
I.7.5. Yahoo et Bing Vs Google 25
I.7.6. DuckDuckGo 26
I.7.6.1. Fonctionnement de DuckDuckGo 26
I.7.6.2. Caractéristiques du moteur DuckDuckGo 26
I.7.6.3. Avantages et inconvénients de DuckDuckGo 27
I.8. Conclusion 28
Chapitre II Analyse des sentiments
II.1. Introduction 30
II.2. Catégorisation des sentiments 30
II.3. Types d'analyse des sentiments 31
II.4. Analyse fine des sentiments (fine-grained sentiment analysis) 31
II.5. Détection d'émotion (Emotion détection) 32
II.6. Analyse de sentiments à base d'aspects (Aspect-Based Sentiment Analysis ABSA) 32
II.7. Difficultés d'Analyse des Sentiments 32

Sommaire

II.8. Domaines d'application d'analyse des sentiments 33
II.8.1. Politique 33
II.8.2. Prise de décision 33
II.8.3. Les systèmes de recommandations 34
II.8.4. Domaine de Transport 34
II.8.5. Domaine médical 34
II.8.6. Domaine éducation 34
II.8.7. Marketing 34
II.9. Disciplines en relation avec l'analyse de sentiments 34
II.10. Exploration de texte (TextMining) 34
II.11. Méthodes de classification des sentiments 36
II.12. Conclusion 37
Chapitre III Notre service de recherche émot-thématique
III.1. Introduction 39
III.2. L'analyse des sentiments 39
III.3. Processus générale de notre service de recherche orienté émotion : 40
III.3.1 Collection de données (Dataset) 41
III.3.2 Requête 41
III.3.3 Indexation requête-documents 42
III.3.4 Appariement Requête - documents 42
III.3.5 Processus générale de notre méta-service de recherche classique 43
III.4. Analyse et organisation des données (document-requête) 44
III.4.1 Prétraitement des données 44
III.4.2 Suppression de la ponctuation 44
III.4.3 .Suppression de mots vides 45
III.4.4 Lemmatisation 46
III.4.5 Quantification individuelle des polarités 46
III.4.6 Sauvegarde des données 48
III.4.7 Appariement document-requête 48
III.4.8 Résultat final de la recherche d'appariement (Q,D) 48
III.5. Conclusion 50
Chapitre IV : MISE EN ŒUVRE D'UN SERVICE DE RECHERCHE ORIENTE EMOTION
IV.1. Introduction 52
IV.2. Environnement de travail 52

Sommaire

IV.2.1 Outils de développement 52
IV.2.2 Langages de programmation 53
IV.2.3 Bibliothèques principales 53
IV.3. Diagramme d'activité globale du SR orienté émotion 54
IV.4. Diagramme d'activité des évaluations de scores 55
IV.5. Diagramme d'activité du méta-service de recherche standard 55
IV.6. Diagramme de séquence de la collecte et mémorisation des données 56
IV.7. Prétraitement du contenu textuel des données: 56
IV.7.1 Analyse de sentiment : 57
IV.7.2 Analyse des sentiments des documents : 58
IV.7.3 Statistiques des données les plus utilisées 58
IV.7.3.1 Sentiment de requête 61
IV.7.3.2 Appariement document requête 61
IV.7.3.3 Distribution des classes du datas et complet 61
IV.8. Présentation des pages d'interaction de notre service de recherche 62
IV.9. Conclusion 66
Conclusion Générale **67**
Références bibliographiques **70**

Table Des Figures

Table des figures

Chapitre I: LE DOMAINE DE LA RECHERCHE D'INFORMATION CLASSIQUE

Figure I. 1: Processus en U de la recherche d'information 7

Figure I. 2: Taxonomie des modèles de RI 10

Figure I. 3: Représentation algébrique des documents et des requêtes dans l'espace des termes à deux dimensions 11

Figure I. 4: Historique des moteurs de recherche (1990 - 2014) 15

Figure I. 5: Architecture originale du moteur de recherche Altavista 16

Figure I. 6: Architecture du système Harvest. 16

Figure I. 7: Architecture du moteur de recherche Google. 20

Figure I. 8: Comment-fonctionne-Google 21

Chapitre II : ANALYSE DES SENTIMENTS

Figure II.1 : Workflow de l'analyse des sentiments 31

Figure II.2 : Distinction entre la subjectivité et l'objectivité via un exemple 31

Figure II.3 : Domaines d'application d'analyse des sentiments 33

Figure II.4 : Les domaines de l'exploration de texte dans l'exploration de données 35

Figure II.5 : Approches d'analyse des sentiments 37

Chapitre III : NOTRE SERVICE DE RECHERCHE ÉMOT-THEMATIQUE

Figure III. 1: Processus générale de notre service de recherche orienté émotion 40

Figure III. 2: Préparation du dataset 41

Figure III. 3: Prétraitement d'une requête simple ou complexe 42

Figure III. 4: Processus générale du meta-service de recherche classique 43

Figure III. 5: Processus de prétraitement 44

Figure III. 6: Liste de ponctuation 45

Figure III. 7: Algorithme d'élimination de mots vides 46

Figure III. 8: Quantification individuelle de polarités 46

Figure III. 9: Fichier Csv enrichi de polarités 48

Chapitre IV : MISE EN ŒUVRE D'UN SERVICE DE RECHERCHE ORIENTE EMOTION

Figure IV. 1 : Diagramme d'activité du processus de recherche global. 54

Figure IV. 2 : Diagramme d'activité des évaluations de scores ('pos', 'neu', 'neg' et 'comp') 55

Figure IV. 3 : Diagramme d'activité du méta-service de recherche standard 56

Figure IV. 4 : Diagramme de séquence de la collecte et stockage des données 56

Figure IV. 5 : Prétraitement du contenu textuel 57

Table Des Figures

Figure IV. 6 : Code source et résultats de sentiment des documents 58

Figure IV. 7 : Statistiques de chaque mot positif le plus répété 59

Figure IV. 8 : Statistiques ensemble des mots positifs les plus répétés 59

Figure IV. 9 : Statistiques de chaque mot négatif le plus répété 60

Figure IV. 10 : statistiques et l'ensemble des mots négatifs les plus répétés. 60

Figure IV. 11 : Code source du sentiment de Requête 61

Figure IV. 12 : Code source de l'appariement document-requête 61

Figure IV. 13 : Distributions des classes de notre dataset complet 62

Figure IV. 14 : Interface principale du moteur de recherche émotionnel 62

Figure IV. 15 : Interface principale du moteur de recherche émotionnel 63

Figure IV. 16 : Statistiques des résultats de recherche individuelle et groupée de mots 64

Figure IV. 17 : l'interface principale du méta-service de recherche standard 64

Figure IV. 18 : La phase de recherche 65

Figure IV. 19 : Résultats du méta-service de recherche 65

Liste Des Tableaux

Chapitre III: **NOTRE SERVICE DE RECHERCHE ÉMOT-THEMATIQUE**

Tableau. III. 1:Mots vides anglais 45

Tableau. III. 2: Résultats de l'appariement (Q,D) 49

Chapitre IV :**MISE EN ŒUVRE D'UN SERVICE DE RECHERCHE ORIENTE EMOTION**

Tableau. IV. 1 : Statistiques des données raffinées 54

INTRODUCTION

GENERALE

Introduction générale

> **"La connaissance des mots conduit à la connaissance des choses."**
>
> **Platon**

La Recherche d'Information (RI) est un domaine qui s'intéresse à la structure, à l'analyse, à l'organisation, au stockage, à la recherche et à la découverte de l'information. Le défi est, parmi le volume important de documents disponibles, de trouver ceux qui correspondent au mieux aux attentes des utilisateurs. L'opérationnalisation de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d'Information (SRI), ces systèmes ont pour but de mettre en correspondance une représentation du besoin en information de l'utilisateur avec une représentation interne du contenu des documents au moyen d'une fonction d'appariement (correspondance).

L'architecture des outils de RI est généralement caractérisée par l'utilisation d'un index inversé et d'un ensemble de machines fonctionnant en parallèle, de même que Google, Yahoo et Bing. La pertinence des réponses est liée à un système de tri (Ranking) de pertinence construit sur la notion de lien existant entre les pages. Ce principe de recherche et d'évaluation est qualifier aujourd'hui de classique, et les approches en RI se sont orientées vers l'une des nouvelles générations de systèmes de recherche basés comme exemple sur le sentiment ou l'accès émotionnel.

L'analyse des sentiments (Sentiment Analysis ou Opinion Mining) forme une méthode de traitement automatique du langage naturel (Naturel Langage Processing) qui tente de repérer la présence des sentiments ou d'émotions exprimés dans un texte, ou dans une phrase.

Dans ce contexte, nous nous intéressons à développer un service de recherche orienté émotion et à intégrer l'unité informationnel probabiliste au sein de son modèle de recherche pour voir l'impact résultant d'un tel choix sur les résultats de recherche.

Néanmoins, plusieurs questions se posent au sujet de l'amélioration du processus de recherche d'information, et de la manière dont les résultats sont retournés. Les problématiques auxquelles nous cherchent à applicé des solutions dans le cadre de ce PFE sont :

1. Comment peut-on améliorer les résultats de recherche par la prise en compte du sentiment?
2. Comment peut-on assurer une évaluation textuelle et/ou émotionnelle des réponses retournées par les outils de recherche d'information?

Le mémoire est organisé de la manière suivante :

- Le chapitre 1 présente le domaine de la recherche d'information classique (textuelle),
- Le chapitre 2 présente le domaine de l'analyse des sentiments,
- Le chapitre 3 décrit la structure générale du système de recherche orienté émotion proposé, et Le chapitre 4 présente les expérimentations et des résultats obtenus.

Chapitre I

Le Domaine De La Recherche D'information CLASSIQUE

I.1. Introduction

La Recherche d'Information (RI) peut être définie comme une activité dont la finalité est de localiser et de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur.

L'opération de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d'Information (SRI), ces systèmes ont pour but de mettre en correspondance une représentation du besoin de l'utilisateur (requête) avec une représentation du contenu des documents au moyen d'une fonction de comparaison (ou de correspondance). L'essor du web a remis la RI face à de nouveaux défis d'accès à l'information, il s'agit cette fois de retrouver une information pertinente dans un espace diversifié et de taille considérable. Ces difficultés ont donné naissance à une nouvelle discipline appelée Recherche d'Information sur le Web.

Dans ce chapitre, nous nous intéressons au processus de recherche d'informations. Après avoir donné les définitions, nous étudions l'architecture générale des systèmes de recherche d'informations ainsi que les trois principaux modèles : booléen, vectoriel et probabiliste. Les capacités d'un tel système sont étroitement liées à la représentation des documents qui est utilisée, soulignant l'importance des éléments pris en compte lors de l'indexation. Vers la fin, nous prêtons une attention particulière au moteur de recherche classique, qui se trouve au centre de la RI.

I.2. Définitions

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

- La Recherche d'Information peut se définir comme : « Action, méthodes et procédures ayant pour objet d'extraire d'un ensemble de documents les informations voulues. Dans un sens plus large, toute opération (ou ensemble d'opérations) ayant pour objet la recherche, la collecte et l'exploitation d'informations en réponse à une question sur un sujet précis » d'après l'AFNOR
- La recherche d'information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations [1].
- 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 [2].

- La recherche d'information est une discipline de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information [3].
- Ces notions se confondent dans les systèmes actuels du fait notamment de la dématérialisation massive des documents et de l'intérêt toujours grandissant du Web. L'utilisateur peut donc indifféremment réaliser ces types de recherche et ce de façon transparente. Il y a donc un accès direct au contenant et au contenu.
- La plupart des activités quotidiennes font appel à la RI et ce quel que soit le contexte (professionnel, loisirs…) pour combler un manque en information, vérifier voire valider une information

I.3. Concepts de base de la RI

Plusieurs concepts clés s'articulent autour de la définition d'un système de RI [4]:

I.3.1. Collection de documents

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

I.3.2. Document

Le document constitue l'information élémentaire d'une collection de documents. L'information élémentaire, appelée aussi granule de document, peut représenter tout ou une partie d'un document.

I.3.3. Besoin en information

Cette notion est souvent assimilée au besoin de l'utilisateur. Ingwersen [5] a défini trois types de besoins utilisateur:

Besoin vérificatif : l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un besoin de type vérificatif est dit stable, c'est-à-dire qu'il ne change pas au cours de la recherche.

Besoin thématique connu : l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et domaine connus. Un besoin de ce type peut être stable ou variable ; il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche.

Besoin thématique inconnu : pour ce type de besoins, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations hors des sujets ou domaines qui lui sont familiers. Le besoin est intrinsèquement variable et est toujours exprimé de façon incomplète.

I.3.4. Requête

La requête est l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre le système de RI et l'utilisateur. Divers types de langages d'interrogation sont proposés dans la littérature. Une requête est un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique.

I.3.5. Pertinence

La pertinence est une notion fondamentale dans le domaine de la RI. La pertinence peut être définie comme la correspondance entre un document et une requête, ou encore une mesure d'informativité du document à la requête [6]. Parmi les classes de pertinence les plus fréquentes [7] :

I.3.5.1. La pertinence algorithmique (ou système) :

Souvent présentée par un score de l'adéquation du contenu des documents vis-à-vis de celui de la requête. Pour mesurer cette adéquation, le système de RI procède au calcul du degré de similitude du document et de la requête en se basant sur les représentations internes de chacun de ceux-ci. Le but de tout système de RI est de rapprocher la pertinence algorithmique calculée par le système aux jugements de pertinence donnés par des vrais utilisateurs.

I.3.5.2. La pertinence thématique :

Traduit le degré d'adéquation de l'information retrouvée au thème évoqué par le sujet de la requête. C'est la mesure la plus utilisée dans les moteurs de recherche classiques.

I.3.5.3. La pertinence cognitive :

Représente la relation entre l'état de la connaissance intrinsèque de l'utilisateur et l'information portée par les documents telle qu'interprétée par l'utilisateur, cette pertinence se caractérise par une dynamique qui permet d'améliorer la connaissance de l'utilisateur via

Le processus de RI qui permet, à partir d'une requête, d'ordonnancer les documents est appelé "processus en U". Il est décomposé en trois principales étapes, illustrées dans la Figure I-I : Processus en U de la recherche d'information et détaillées ci-dessous:

- L'indexation des documents et des requêtes de l'utilisateur.

- ❖ L'appariement document-requête.
- ❖ La reformulation de la requête.

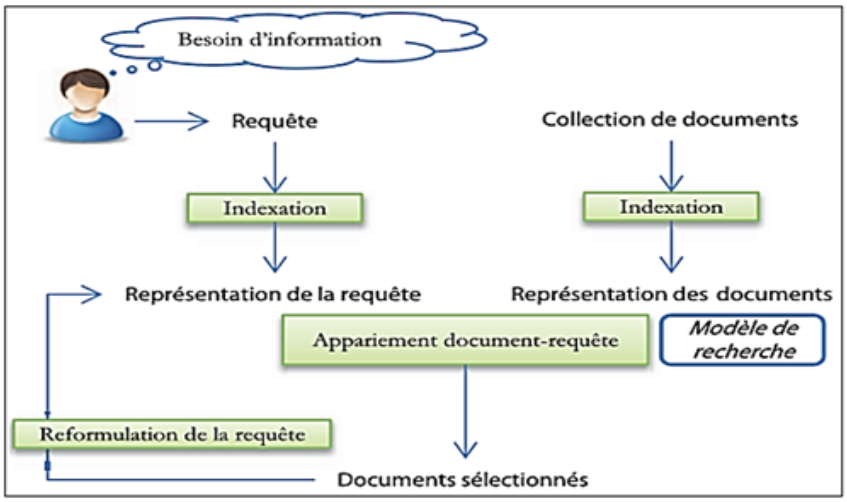

Figure I. 1: Processus en U de la recherche d'information

I.4. Indexation

Les documents à leur état brut sont difficiles à exploiter tels quels lors de la phase de recherche. Ainsi, l'objectif principal de cette étape est de fournir des représentations des documents et des requêtes facilement exploitables par la machine dans la phase de recherche. Cette représentation est souvent une liste pondérée de mots-clés significatifs que l'on nomme descripteurs du document (ou de la requête) [8]. L'indexation peut être :

- **Manuelle :** la représentation du document est réalisée par un expert qui identifie les termes les plus représentatifs du document.

•**Automatique** : le processus d'indexation est entièrement informatisé. Il repose sur une démarche algorithmique qui traite chaque terme selon un processus défini : extraction, suppression des mots vides, normalisation et pondération.

- **Semi-automatique :** est une combinaison des deux précédentes approches où le choix final des termes à indexer revient à l'expert.

A la fin de cette étape, les documents sont représentés dans des fichiers index qui stockent la cartographie des couples terme-document en y associant un poids. La formule de pondération la plus utilisée est celle basée sur la fréquence des termes dans les documents, appelée TF-IDF [9]. L'intuition de cette pondération est de favoriser les termes qui sont à la fois fréquents dans le document et peu fréquents dans la collection. Cette dernière condition est basée sur les propriétés de la loi de Zipf [10] qui étudie la distribution des termes dans une collection de documents.

L'indexation automatique [11] regroupe un ensemble de traitements automatisés sur un document comme :

I.4.1. L'extraction des mots :

Ce processus consiste à analyser le texte d'un document afin d'extraire ses mots en reconnaissant les espaces de séparation des mots, les ponctuations, etc.

I.4.2.L'élimination des mots vides :

Un document contient souvent des mots non significatifs appelés mots vides (pronoms personnels, prépositions).

L'élimination de ces mots se fait à l'aide d'une liste prédéfinie de mots vides ou en supprimant les mots ayant une fréquence dépassant un certain seuil. Éliminer les mots vides permet de réduire la taille de l'index, gagner en espace mémoire et optimiser le temps d'exécution.

I.4.3.La lemmatisation :

Ce traitement consiste à radicaliser les mots restants, c'est à dire réduire les mots à leur forme canonique par exemple, toutes les formes d'un verbe sont regroupées à l'infinitif, tous les mots au pluriel sont ramenés au singulier, etc. Grâce à la lemmatisation, les documents contenant différentes formes d'un même terme auront les mêmes chances d'être restitués ce qui améliore la capacité d'un SRI à retrouver les documents pertinents. Parmi les méthodes utilisées pour la lemmatisation on peut citer l'algorithme de Porter [12] pour les textes en anglais et la troncature [13] pour les autres langues (Français, Italien, Allemand).

I.4.4.La pondération :

Les termes d'un document n'ont pas souvent la même importance. Un terme qui apparaît dans la majorité des documents de la collection aura moins d'importance qu'un terme qui existe dans quelques documents seulement. Plusieurs fonctions de pondération de termes ont été proposées dans la littérature. La plupart de ces fonctions combinent des variantes des facteurs *TF (Term Frequency)* et *IDF (Inverse Document Frequency)* qui mesurent un poids local (dans le document) et global (dans la collection) d'un terme [14].

La mesure *TF-IDF*, font intervenir deux facteurs : fréquence du terme t dans le document d, notée TF (Term Frequency), et la fréquence inverse du document, notée *IDF (Inverse Document Frequency)*

La formule du *TF-IDF* est donnée par le produit des deux fonctions *TF* et *IDF* comme suit:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \ldots \ldots$$

Equation 1:La formule du *TF-IDF*

I.4.4.1. TF (Term Frequency)

Ce facteur prend en compte le nombre d'occurrence d'un terme dans un document. L'idée derrière cette mesure est que plus un terme est fréquent dans un document plus il est important. Elle représente une pondération locale d'un terme dans un document.

I.4.4.2. IDF (Inverse Document Frequency)

Ce facteur mesure la fréquence d'un terme dans toute la collection, c'est la pondération globale. [1]

I.4.5. Appariement document-requête

Le processus d'appariement met en relation la collection de documents, indexée au préalable, avec la requête, également prétraitée, afin d'identifier les documents pertinents. Cette étape permet au SRI de retourner une liste de documents à l'utilisateur.

Dans le processus d'appariement, le système calcule un score de correspondance entre la représentation de chaque document et celle de la requête. Ce score peut être binaire (pertinent ou non pertinent) ou multi value pour exprimer un degré de pertinence système. La pertinence système est calculée à partir d'une fonction de similarité appelée RSV (Q, *D*) (Retrie val statuts Value) où Q est une requête et D un document. Pour une requête donnée, le système retourne des documents en ordre décroissant du score de pertinence.

L'appariement *document-requête* repose sur un cadre théorique défini par un modèle de recherche d'information. Une taxonomie des modèles (**Error! Reference source not found.**) a été présentée par Baez-Yates [15] et présente quatre familles principales. Les modèles reposant sur le texte des documents (modèles de RI classiques et modèles basés sur le texte semi-structuré), les liens entre documents (modèles orientés web) et les documents multimédia (recherche d'images, de musiques, d'audio ou de vidéos).

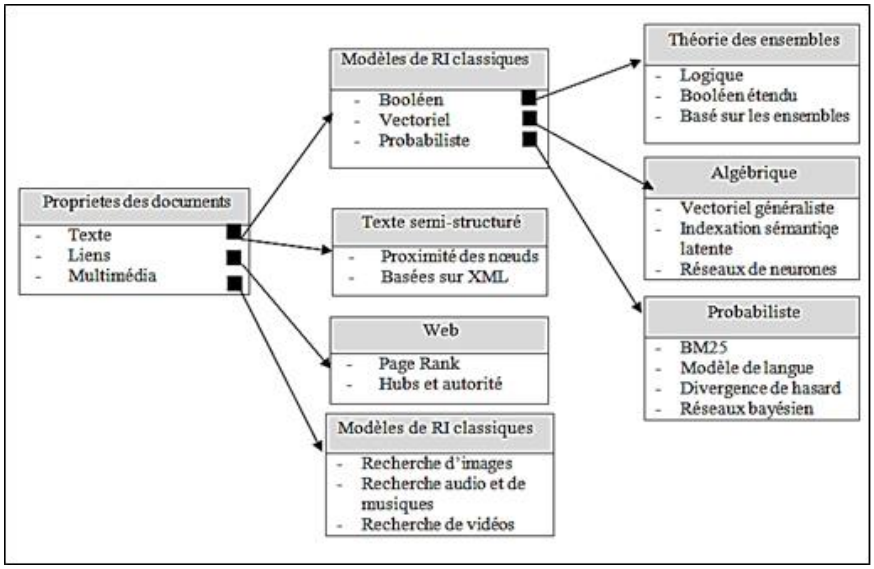

Figure I. 2: Taxonomie des modèles de RI

Un modèle de recherche d'informations est un cadre de calcul qui, à partir d'une représentation des documents et une représentation de la requête, détermine la relation ou le degré de similitude entre le document et la requête.

Ici, nous présentons les trois grands modèles utilisés en recherche d'informations : booléen, vectoriel et probabiliste:

I.4.5.1. Modèle booléen

Le modèle booléen [16] est le modèle le plus ancien dans la recherche d'information. Il est basé sur la théorie des ensembles et l'algèbre de Boole. Le document est représenté par un ensemble de termes. La requête est représentée sous forme d'une expression logique composée de termes reliés par des opérateurs logiques ***ET***, ***OU***, ***NON***. L'appariement (RSV) entre une requête et un document est un appariement exact, autrement dit si un document implique au sens logique de la requête alors le document est pertinent. Sinon, il est considéré non pertinent. Par conséquent, le score de similarité entre un document *d* et une requête *q* est inclus dans l'ensemble {0, 1}:

$$\mathbf{RSV}(\boldsymbol{d}, \boldsymbol{q}) = \begin{cases} \mathbf{1} \ldots\ldots\ldots \text{Si d appartient à l'ensemble décrit par q} \\ \mathbf{0} \qquad\quad \text{Sinon} \end{cases}$$

I.4.5.2. Modèle vectoriel

Initialement proposé par Salton et implémenté dans le système SMART [17]. La pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel. Le modèle vectoriel représente les documents et les requêtes par des vecteurs d'un espace à n dimensions, les dimensions étant constituées par les termes du vocabulaire d'indexation

Les coordonnées d'un vecteur document sont les poids des termes d'index dans ce document.

$$D_i = w_{i1}, w_{i2}, w_{i3} \ldots w_{in} \qquad \textit{pour } i = 1, 2 \ldots m$$

Où w_{ij} est le poids du terme t_j dans le document D_i,
m est le nombre de documents dans la collection,
n est le nombre de termes d'indexation.
On représente aussi la requête par un vecteur de mots-clés défini dans le même espace vectoriel que le document. $Q = w_{Q1}, w_{Q2}, \ldots w_{Qn}$,
Où w_{Qj} est le poids de terme t_j dans la requête Q.

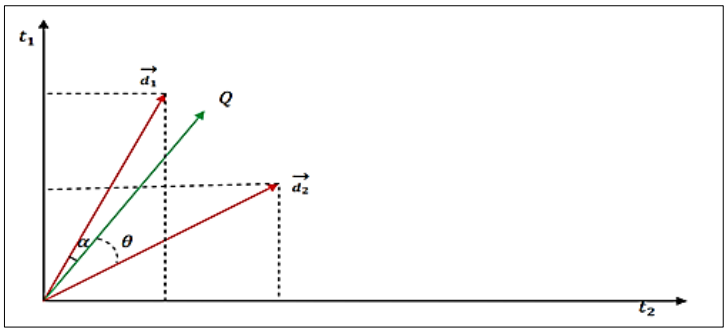

Figure I. 3: Représentation algébrique des documents et des requêtes dans l'espace des termes à deux dimensions

L'appariement document-requête dans le modèle vectoriel, consiste à trouver les vecteurs documents qui s'approchent le plus de vecteur de la requête. Cet appariement est obtenu par l'évaluation de la distance entre les deux vecteurs où plusieurs mesures de similarité ont été définies. La plus fréquente est:

- **La mesure cosinus :**

$$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} w_{ki} \cdot w_{kj}}{\sqrt{\sum_{k=1}^{M} w_{ki}^2} \cdot \sqrt{\sum_{k=1}^{M} w_{kj}^2}}$$

Plus les vecteurs sont similaires, plus l'angle formé est petit, et plus le cosinus de cet angle est grand. A l'inverse du modèle booléen, la fonction de correspondance évalue une correspondance partielle entre un document et une requête, ce qui permet de retrouver des documents qui ne reflètent pas la requête qu'approximativement. Les résultats peuvent donc être ordonnés par ordre de pertinence décroissante [18].

I.4.5.3. Modèle probabiliste

Ce modèle est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête [19] …Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Etant donné une

requête utilisateur Q et un document D, il s'agit de calculer la probabilité de pertinence du document pour cette requête.

Le score de pertinence d'un document d par rapport à la requête q est estimé comme suit :

$$RSV(q, d_j) = \frac{p(p/di)}{p(\bar{p}/di)}$$

Où $P(P|d_i)$ et $P(\bar{P}|d_i)$: La probabilité q'un document d_i soit pertinent (P) vis-à-vis de la requête q (respectivement non pertinent $P(\bar{P}|d_i)$). [20]

Ces probabilités sont estimées par de probabilités conditionnelles selon qu'un terme de la requête est présent, dans un document pertinent ou dans un document non pertinent. Cette mesure de similarité entre la requête et les documents peut se calculer par différentes formules. Ce modèle a donné lieu à de nombreuses extensions. Il est à l'origine du système OKAPI qui est l'un des systèmes les plus performants selon les compagnes d'évaluation TREC[1]. L'inconvénient majeur de ce modèle est que les calculs des probabilités sont complexes et que l'indépendance des variables n'est pas toujours vérifiée voir pas prise en compte [21].

Ces modèles ont une base théorique saine [22] et se sont montrés particulièrement performants dans TREC.

I.4.6. Reformulation de la requête

La reformulation du besoin en information est l'étape qui permet de redéfinir le besoin de l'utilisateur au fur et à mesure de la session de recherche.

Cette étape peut être effectuée:

- Manuellement, dans le cas où l'utilisateur soumet lui-même une nouvelle requête.
- De façon automatique, lorsque le système de RI s'appuie sur les termes importants dans les documents les plus pertinents ou visités par l'utilisateur qui sont réutilisés.

L'une des stratégies de reformulation de requêtes est celle qui est dirigée par l'utilisateur. Le principe de cette stratégie est de construire une nouvelle requête à partir de la structure des documents jugés par l'utilisateur : c'est ce que l'on appelle la réinjection de pertinence « relevance feedback » [23] [24] [25].

[1] WWW.TREC.com

I.5. Les applications de la RI

Le domaine de la RI a donné naissance à plusieurs applications

I.5.1. Filtrage des e-mails (spam)

Un spam est un courrier électronique envoyé en grand nombre, de façon anonyme ou sous une fausse identité, à des destinataires qui ne l'ont pas sollicité et que l'on maintient dans l'incapacité de s'opposer à cette diffusion. La lutte anti-spam est un ensemble de comportements, de systèmes et de moyens techniques et juridiques permettant de combattre le spam, courriers électroniques publicitaires non sollicités.

I.5.2. Application Biomédicale

Dans le domaine biomédical plus encore que dans d'autres domaines, l'emploi de termes spécialisés est la clef de l'accès à l'information. Pour faciliter l'accès à l'information de ce domaine, plusieurs terminologies ont été développées pour une indexation contrôlée des documents dans les portails de santé.

I.5.3. La classification des textes

La classification automatique de textes est un domaine où la fouille de textes et les techniques statistiques produisent des résultats à partir des calculs de fréquence d'occurrence de termes extraits. L'analyse syntaxico-sémantique était considérée, jusqu'à présent, comme pénalisante en raison des limitations des analyseurs eux-mêmes. Elle est peu sensible à la qualité des corpus d'entraînement puisqu'elle se sert de ressources stables (des dictionnaires non variant), alors que les méthodes statistiques y sont très sensibles. [26]

I.5.4. Moteur de recherche

Certains sites web offrent un moteur de recherche comme principale fonctionnalité ; on appelle alors moteur de recherche le site lui-même (Google Vidéo par exemple est un moteur de recherche vidéo). Les sites suivent les liens hypertextes (qui relient les pages les unes aux autres) rencontrés sur chaque page atteinte. Chaque page identifiée est alors indexée dans une base de données, accessible ensuite par les internautes à partir de mots-clés. Les moteurs de recherche ne s'appliquent pas qu'à Internet : certains moteurs sont des logiciels installés sur un ordinateur personnel. Il existe également des métas moteurs, c'est-à-dire des sites web où une même recherche est lancée simultanément sur plusieurs moteurs de recherche (les résultats étant ensuite fusionnés pour être présentés à l'internaute). [27]

I.6. Moteur de recherche

Un moteur de recherche (MR) est, comme son nom l'indique, un outil qui permet de rechercher sur le Web (mais aussi sur un ordinateur personnel) des ressources, des contenus, des documents etc, à partir de mots clés. Il suffit de renseigner les expressions qui forment la requête et le moteur de recherche déniche automatiquement les ressources correspondant à la recherche. Les résultats apparaissent organisés selon une logique propre à chaque moteur. Pour exercer sa mission et offrir les réponses les plus pertinentes à une recherche, le moteur de recherche parcourt le web et référence les pages Internet selon leurs contenus. Cette opération est réalisée par ce que l'on appelle un robot d'indexation, spider ou crawler. Apparaître en tête des résultats sur internet est l'objectif de beaucoup de sites. Pour y parvenir, il faut connaître les facteurs de positionnement du moteur de recherche et optimiser son site pour faciliter le travail d'indexation des robots. Lorsqu'ils doivent indexer de grandes quantités de données, les moteurs de recherche utilisent des algorithmes pour améliorer leurs résultats.

Google est aujourd'hui le moteur de recherche le plus utilisé dans le monde. Plus de 90% des requêtes des internautes passent en effet par l'outil de recherche du géant américain. Yahoo Search, Qwant ou encore Bing sont quelques-uns de ses concurrents. La plupart proposent des fonctionnalités pour affiner les recherches par type de contenu, langue, date de la dernière mise à jour, etc. [28]

I.6.1. Historique des moteurs de recherche (MRs):

Les premiers MRs ont été développés dans les années 90. Le premier moteur baptisé au nom de « Archie » fut développé en tant que projet d'école par Alan Emtage, un étudiant de l'Université McGill à Montréal. Ces premiers moteurs n'étaient cependant que des précurseurs de véritables outils de recherche tels que connus aujourd'hui car ils étaient très limités et ne permettaient de faire des recherches qu'à partir des noms des fichiers disponibles, sans idée d'indexation des contenus. Il fut le premier moteur à déployer des robots d'indexation. L'idée de base, qui était de mesurer la croissance du Web, fut rapidement remaniée pour arriver au premier moteur de recherche à indexation automatique. Par la suite, l'expansion de robots capables de scanner le contenu des pages Web a permis de développer des moteurs de recherche plus puissants. « Yahoo! » créé en 1994, proposa l'idée

D'un annuaire pour la recherche. Les pages étaient alors recensées et triées par l'homme, en fonction de leur pertinence et de leur qualité. Altavista, développé à partir de recherches menées par des scientifiques d'un laboratoire de recherche en Californie en 1995, est jugé efficace et

rapide, et deviendra ainsi la star des moteurs de recherche du moment jusqu'aux années 2000. Enfin, c'est en 1998 que né Google, crée par Sergey Brin et Lawrence Page, étudiants de Stanford. Google va littéralement révolutionner le monde des moteurs de recherche grâce à sa simplicité et son efficacité. Vers 2001-2002, l'éclatement de la bulle internet fait disparaitre les premiers moteurs de recherche, et seuls les plus grands survivent. De manière exhaustive, la figure n°1 ci-dessous fait un récapitulatif sous forme de time line de l'historique de naissance des moteurs de recherches [28]

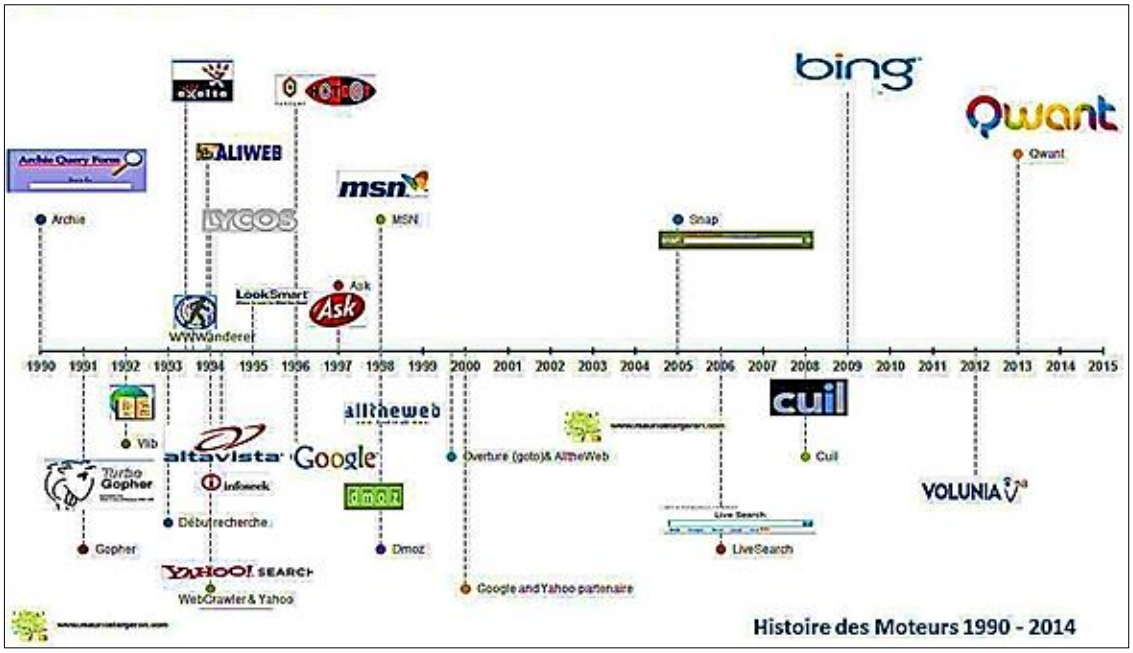

Figure I. 4: Historique des moteurs de recherche (1990 - 2014) - [2]

I.6.2. Architecture générale des premiers moteurs de recherche.

L'architecture originale utilisée par Altavista représente la première catégorie de systèmes. Il s'agit d'une architecture très simple qui se divise en deux parties distinctes. On retrouve d'une part un crawler et d'autre part l'interface d'interrogation du moteur de recherche et le système d'analyse des requêtes proposés par les utilisateurs du système.

Le cœur du système repose sur un index inversé permettant d'associer des mots à un ou plusieurs documents. La demande de l'utilisateur est traitée en interrogeant l'index inversé pour connaître les documents dans lesquels apparaissent le plus souvent les mots de la requête.[29]

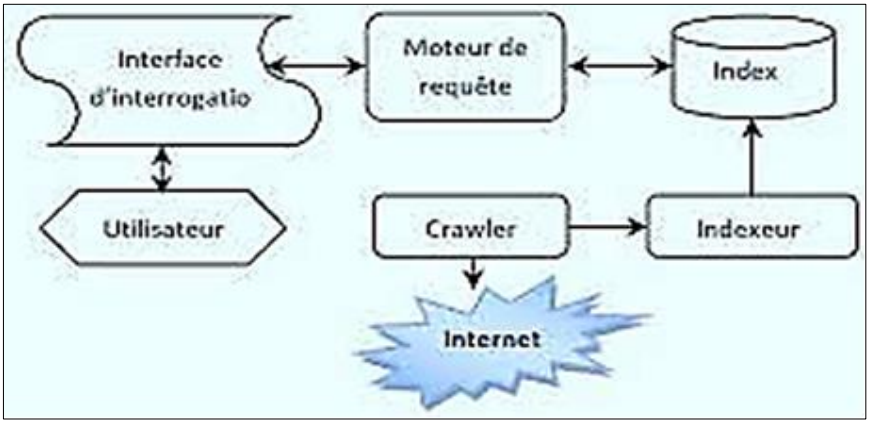

Figure I. 5: Architecture originale du moteur de recherche Altavista [29].

I.6.3. Architecture distribuée et adaptative.

Des variantes de l'architecture précédente, basées sur le modèle indexeur-crawler, ont été imaginées pour gommer les défauts inhérents à sa conception. L'une d'entre elle, appelée Harvest s'est révélée très innovante en matière de distribution des ressources.

Le récolteur : est chargé de collecter et d'extraire périodiquement des informations d'indexation -textes, images - depuis plusieurs sites Web.

Le broker : quant à lui, fournit le mécanisme d'indexation et l'interface d'interrogation sur les données amassées par le récolteur.

On retrouve ici, le mécanisme indexeur-crawler identifié dans la section précédente. Cependant, plusieurs brokers et plusieurs récolteurs peuvent communiquer ensemble, chacun se spécialisant dans un domaine précis. Lorsqu'une requête est émise sur un broker dont le domaine traité ne correspond pas à ses capacités, celui-ci transmet la requête à une autre entité capable de la gérer.

C'est un système totalement adaptatif dans lequel il est possible de configurer les brokers et les récolteurs de manière à répartir le besoin en ressources sur un ou plusieurs domaines particuliers. [29]

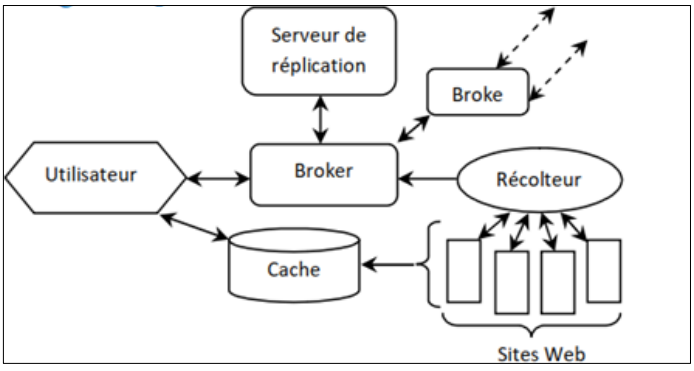

Figure I. 6: Architecture du système Harvest. [29]

I.6.4.Fonctionnement des moteurs de recherche

Avant même de vous permettre de saisir une **requête** et de rechercher sur le Web, les moteurs de recherche doivent réaliser de nombreuses opérations pour qu'il soi en mesure de vous présenter un ensemble de résultats précis et de qualité qui répondent à votre question ou et intentions de recherches.

Ils ont trois fonctions principales :

I.6.4.1. L'exploration :

Des robots, appelés crawleurs, parcourent le web en naviguant de lien en lien. L'araignée se déplace de page en page et de site en site à l'aide de liens. Les moteurs ont un certain nombre de programmes et logiciels appelés crawlers (appelés aussi araignées, bots ou spiders) qui sont chargés de trouver des informations accessibles au public. Le robot le plus connu de Google est appelé GoogleBot. Ils visitent chaque site et en utilisant différentes techniques, que ce soit du contenu texte, des images, des vidéos ou tout autre format (CSS, HTML, javascript…), les site maps et robots.txt. Lors de la visite d'un site, ils suivent tous les liens internes et externes, se déplacent de page en page et de site en site à l'aide des liens. S'ils n'ont pas le temps de tout explorer, les robots reviennent plus tard finir le travail d'exploration. les moteurs de recherche attribue à chaque site, selon de nombreux critères, un budget de crawl qui limite le nombre de pages que ses robots viennent visiter sur une période donnée. Les crawlers gardent également des traces des modifications apportées aux pages de savoir afin de pouvoir actualiser son analyse des contenus. [30]

I.6.4.2. L'analyse et l'indexation :

Stockage et analyse du contenu découvert durant l'exploration. Elle signifie essentiellement qu'ils les enregistrent dans les bases de données. Cette étape est la plus complexe, et par conséquent, celle dont les secrets sont les mieux gardés. Elle diffère d'un moteur à l'autre, mais les grandes lignes de ce processus sont connues et communes à tous les moteurs de recherche L'analyse consiste à décomposer les données textuelles en unités fondamentales de recherche, appelées termes. Pendant l'analyse, le texte subit plusieurs opérations : extraction des mots, suppression des mots les plus courants et de la ponctuation, réduction des mots à leur racine, passage en minuscule, etc. L'analyse du texte est effectuée avant l'indexation et l'analyse des requêtes. Elle convertit les données textuelles en entités lexicales (« tokens », en anglais), et ces okens sont ajoutés en tant que termes à la base d'index. L'indexation est donc le processus d'organisation des données textuelles analysées, pour un stockage dans un format qui facilite une

recherche rapide. Une analogie simple est celle de l'index que l'on trouve à la fin d'un livre : cet index indique la localisation des différents sujets qui sont traités dans ledit livre. [30]

I.6.4.3. Le classement :

Affichage des documents répondant à la requête d'un chercheur en fonction de formules secrètes. Le moteur parcourt sa gigantesque base et les utilise pour filtrer ce qui est pertinent pour vous. Après une bonne indexation, il est primordial de développer un mécanisme efficace d'exploitation des éléments indexés. Il s'agit surtout de fournir aux utilisateurs un outil pour interroger la base de données d'informations collectées, de la façon la plus complète et rapide possible. Cet outil constitue alors le cœur de recherche du moteur que nous désignerons par la suite « serveur de recherche ». Pour retrouver une information précise, l'utilisateur se doit de formuler une requête. On aurait alors pu imaginer utiliser le langage naturel pour la formulation de nos requêtes, mais ce n'est pas toujours aussi évident.

En effet, le langage naturel est tel que très souvent, pour une simple interrogation, il existe une multitude de formulations possibles, pouvant alors induire à des interprétations différentes de la demande ; ceci s'avèrerait encore plus délicat suite à d'éventuelles erreurs que pourrait commettre l'utilisateur (fautes d'orthographes, de frappes, etc.). Certes, avec le temps et les progrès de l'Intelligence Artificielle, le langage naturel finira par s'imposer et sera davantage mieux interprété par les moteurs de recherche. Mais pour le moment, il y a encore d'améliorations importantes à faire pour parvenir à des systèmes relativement satisfaisants. La plupart des moteurs de recherche offrent des syntaxes spécifiques pour l'amélioration de la formulation de ses requêtes. C'est ainsi que, dans la formulation de la requête, on peut par d'exemple exclure les sites comportant certains mots ou même rechercher les sites contenant deux mots juxtaposés et non dispersés dans le texte grâce à des opérateurs dits opérateurs de recherche.

On en dénombre plusieurs, souvent classés par catégories suivantes :

- Opérateurs logiques
- Opérateurs numériques
- Opérateurs de proximité
- Opérateurs de troncature
- Autres opérateurs linguistiques. [30]

I.7. Les moteurs de recherche généralistes classiques

I.7.1. Google :

L'origine du nom Google fait référence au terme mathématique « gogol » qui désigne un nombre commençant par 1 et suivi de 100 zéros. Au tournant des années 2000, après avoir imposé son moteur de recherche en l'espace de quelques années, Google a entamé une vaste diversification qui lui a permis de consolider son modèle économique basé sur la publicité et une facturation au nombre de clics La stratégie de Google consiste à proposer de nombreux services en ligne gratuits tels que la messagerie Gmail, la navigation assistée Maps, la suite bureautique Google Docs, le navigateur Internet Chrome, le service de stockage de fichiers Drive ou encore le service de vidéos YouTube (acquis en 2006 pour 1,6 milliard de dollars) en se rémunérant avec le trafic publicitaire qu'ils génèrent. En 2007, l'entreprise a pris le virage du mobile en lançant son système d'exploitation Android pour smartphones et tablettes tactiles. Ce dernier est en 2017 installé sur 88% des téléphones dans le monde. est un moteur de recherche gratuit et libre d'accès sur le World Wide Web, ayant donné son nom à la société Google. C'est aujourd'hui le moteur de recherche et le site web le plus visité au monde 90 % des internautes l'utilisaient en 2018. [31]

I.7.1.1. Architecture du moteur de recherche Google

L'architecture du moteur de recherche Google est certainement une des plus efficacies actuellement. Elle ne repose pas sur un système monolithique mais sur un grand nombre de machines classiques coopérant ensemble. Ce système peut se décomposer en plusieurs parties comprenant :

- Un sous-système d'exploration d'Internet.
- Un indexeur.
- Un analyseur de la topologie d'Internet formée par les liens hypertextes : et un sous-système de présentation et d'exécution de requêtes.

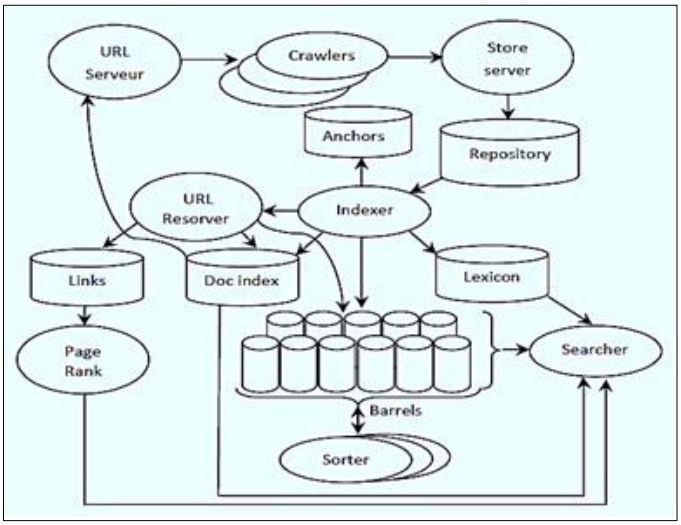

Figure I. 7: Architecture du moteur de recherche Google. [29]

Un serveur d'URL garde la mémoire des liens des pages à visiter. Des robots chargés d'explorer le Web récupèrent ces liens afin de télécharger les documents correspondant et les stocker dans une base de données recensant la totalité des pages indexées. Cette opération est réalisée continuellement et alimente et met à jour en permanence la base de documents du moteur. Périodiquement, cette base est analysée pour réaliser un index inversé reliant des termes aux documents les contenant. D'autres informations sur les termes sont extraites comme leur position dans le document, la taille de la police utilisée ou sa fonte.

Cette analyse permet également d'extraire tous les liens hypertextes des documents rencontrés afin d'alimenter le serveur d'URL. Cette base de liens est utilisée afin de calculer le PageRank permettant de trier les documents de l'index par pertinence décroissante [29].

I.7.1.2. Fonctionnement de Google :

Google fournit une liste de résultats, avec pour objectif de retourner les pages les plus pertinentes possibles pour répondre aux attentes exactes de l'utilisateur.

L'affichage des résultats suit plusieurs étapes:

1. L'utilisateur entre son mot-clé.
2. Google met en place un "index" dans lequel il regroupe des milliards de pages web parcourues et évaluées par ses robots d'indexation (crawlers). Cet index lui servira de base pour sa recherche.
3. Google extrait de cet index les pages qu'il estime répondre le mieux au mot-clé saisi par l'utilisateur lors de la requête.

4. Google calcule et classe les résultats par pertinence, c'est le fameux algorithme, mis à jour très régulièrement.
5. Google affiche les résultats.

Ce moteur de recherche est aussi apprécié pour sa rapidité de recherche et sa sobriété il ne contient ni de Flash, ni de bandeau publicitaire clignotant. Son interface a inspiré celle d'autres moteurs, comme Yahoo!. Cette sobriété, loin d'être anecdotique, est au moins en partie à l'origine du succès du site. À l'époque de son lancement en effet, la mode était aux moteurs de recherche insérés sur des pages très chargées en contenu et en publicité. Ces pages étaient souvent longues à s'afficher et difficiles à lire. Google utilise un système d'AdWords (« publicité de mots ») comme une de ses sources de revenus. Ce système est fondé sur une valeur par mot selon sa demande. Plus le mot sera demandé, plus il sera payé cher par clic. Mais il est toujours possible pour l'utilisateur de bloquer l'affichage de ces publicités grâce à des plugins, le plus populaire étant Adblock Plus avec qui Google a passé un accord financier pour qu'il ne filtre plus ses publicités. Selon Optify, 94 % des clics vont aux dix premiers résultats et Google génère à lui seul 36,4 % des clics. Google utilise des robots nommés Google bot qui visitent à intervalle régulier l'ensemble des sites web n'ayant pas explicitement demandé à ne pas être référencés afin de maintenir à jour la base de données qui fournit les réponses aux requêtes des internautes. Une version bêta est habituellement une mention signifiant qu'un programme est en phase de finition. [32]

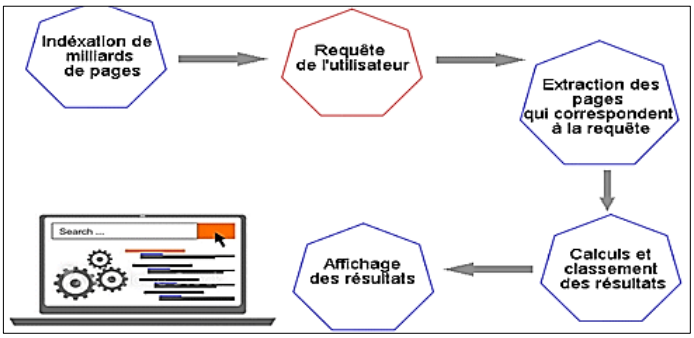

Figure I. 8: Comment-fonctionne-Google [32]

I.7.1.3. Les avantages et les Inconvénients de Google :

Les avantages :

- Prise en charge de la recherche vocale intelligente et des haut-parleurs intelligents
- Comprend Internet derrière votre requête
- Donne la priorité au contenu frais et pertinent plutôt qu'à une présence à long terme
- Accessible sur n'importe quel appareil
- Des secteurs verticaux spécifiques vous aident à trouver ce dont vous avez besoin

- Champ de recherche amusant et attrayant
- Grande quantité de personnalisation disponible
- Possibilité d'obtenir des cartes-cadeaux pour la recherche PPC. [33]

Inconvénients:

- Plateforme publicitaire robuste et compétitive
- Peut-être plus difficile à avancer qu'avec le moteur de recherche de Microsoft
- L'algorithme en constante amélioration est difficile à suivre
- Très sensible au spam potentiel sur Google Maps
- Suit beaucoup de données personnelles pour la recherche [33]

I.7.2.Bing

Bing est un outil de moteur de recherche ingénieux appartenant à Microsoft. Il affiche des résultats précis en fonction de la requête de recherche de l'utilisateur. Il présente quelques services de recherche. Parmi les principales, citons les résultats Web, images, vidéos et cartes. Bing est très connu pour ses possibilités de référencement. Il peut vous donner des résultats complets sur tous les restaurants autour de votre bloc. Bing a été annoncé le X mai 28 par le PDG de Microsoft, Steve Ballmer, et a officiellement été lancé en juillet de la même année. Bing a vu le jour afin d'affronter Live Search et MSN Search. Pour faciliter la transition, Microsoft a acheté Powerset, une société qui développait un moteur de recherche en langage organique pour l'espace Internet. Et pour conclure le marché, Microsoft et Yahoo ont convenu de laisser Bing contrôler Yahoo! Moteur de recherche. [33]

I.7.2.1. Fonctionnement de Bing :

Le moteur de recherche Bing utilise des spiders pour indexer un site Web. Un Spider est un programme automatisé qui est exécuté par un système du moteur. Si vous soumettez votre site au moteur de Microsoft en remplissant un formulaire de soumission ou par tout autre moyen, il enverra automatiquement ses robots pour indexer votre site Web. Ces Spider visiteront ensuite un site et liront son contenu sur les pages visibles, les balises du site et les liens sur un certain site Web qui mènera également à d'autres sites. A mesure que la technologie s'améliore, ces Spider deviennent plus efficaces et plus intelligentes. Ils peuvent en effet indexer et lire des informations

statiques et dynamiques, des informations vidéo et des contenus flash sur un site Web. Lorsque vous demandez au moteur de recherche Bing de rechercher des informations, il effectue une recherche dans les index créés et non dans le Web. Bien que différents moteurs de recherche puissent produire des résultats différents dans les classements, tous les moteurs de recherche n'utilisent pas le même processus pour effectuer des recherches dans leurs index indépendants. [34]

I.7.2.2. Caractéristiques du moteur Bing :

Les résultats obtenus par le moteur de recherche BING sont généralement très utiles. Ils sont également bien présentés avec une liste de résultats apparaissant dans la colonne principale de la page de résultats. Il existe également des sous-titres, qui comportent des mots-clés associés et quelques résultats pour ces mots. Cette fonctionnalité est utile du point de vue de l'utilisateur, car elle peut l'aider à trouver rapidement une grande variété d'informations pour un sujet. La mise en page distingue également de façon claire les résultats des sites sponsorisés. Elle les place dans la barre latérale droite sous l'en-tête de colonne Sites sponsorisés. Les autres mots clés figurant sur la page de résultats sont répertoriés en haut de la Barre latérale à gauche pour faciliter la navigation. Il existe également une liste de recherches connexes auxquelles vous pouvez accéder pour obtenir encore plus d'informations. [34]

I.7.2.3. Avantages et inconvénients de Bing :

Les avantages:

– Accès à des informations en dehors de la recherche, telles que des événements historiques, des images intéressantes et des actualités.

– Indexation vidéo fantastique pour les chercheurs et les utilisateurs webmaster

– Filtre plus facile grâce à la recherche d'images

– Une interface utilisateur esthétique, grâce à la technologie de recherche en direct de Windows

– Plus facile d'attirer des chercheurs plus âgés et plus riches si vous êtes propriétaire d'une entreprise

– Mises à jour fréquentes avec de nouvelles fonctionnalités pour rivaliser avec Google et Google Chrome [33]

Inconvénients:

– Bing classe les sites de forum plus bas en faveur des sites plus anciens et mieux établis

– Bing n'est pas aussi bon pour comprendre l'intention et le contexte

– Bing n'offre pas autant de portée pour les entreprises [33]

I.7.3. Bing Vs Google

On peut soutenir que choisir entre Bing et Google se réduit aux préférences personnelles. Si vous êtes plus intéressé par la recherche d'images et de vidéos, Bing a tendance à être si compétent. D'autre part, il est juste de donner crédit quand c'est dû. Google domine de manière significative puisqu'il possède YouTube. Cela ne veut pas dire que Bing ne donne pas un coup de main dans les recherches vidéo. En fait, sa disposition unique de miniatures vidéo, qui diffuse un aperçu d'un résultat donné, est une caractéristique spectaculaire sur laquelle il faut vraiment compter.

Si un utilisateur souhaite rechercher une image sous licence à utiliser dans son blog ou son projet, Bing le rend beaucoup plus facile. Contrairement à Google, où les paramètres de filtrage des images sont masqués, Bing fournit un filtre de licence comme option principale lors de la r secherche d'une image. [33]

I.7.4. Yahoo :

Yahoo est un moteur de recherche qui est plus vieux que Google. Et quelques internautes l'utilisent encore pour effectuer leurs recherches. Il occupe actuellement la 3e place parmi les moteurs de recherche les plus utilisés dans le monde juste derrière Bing. Yahoo existe depuis encore plus longtemps que Google, et bien que certains le considèrent comme obsolète, c'est toujours le troisième moteur de recherche le plus populaire au monde. C'est même le moteur de recherche par défaut pour Firefox. Une des grandes choses à propos de Yahoo, c'est que c'est beaucoup plus qu'un simple moteur de recherche. Yahoo est l'un des moteurs de recherche les plus connus, principalement parce que c'est aussi un fournisseur de messagerie très populaire. Dans les premiers jours, il était un concurrent proche de Google. Dans le passé, avant 2015 Yahoo recherche a été alimenté uniquement par Bind, mais après que Google est également venu dans l'image de sorte que les résultats ont été fournis par les deux. [35]

I.7.4.1. Fonctionnement de Yahoo :

Si vous souhaitez être bien référencé sur Yahoo, il est donc essentiel de bien choisir les mots-clés, autrement dit les Meta Keywords. Il est à noter que Bing est propulsé par Yahoo. Les deux moteurs de recherche ont beaucoup de similarités mais savent aussi marquer leur différence. [36]

I.7.4.2. Caractéristiques du moteur Yahoo :

Parmi les fonctionnalités qu'il propose :

La recherche multilingue en offrant la possibilité d'une traduction automatique d'une requête en plusieurs langues étrangères et une traduction automatique des pages. un outil de stockage de favoris dans lequel l'utilisateur peut stocker sur le réseau Yahoo! une image conforme de la page au moment où il la voit. Il peut aussi partager un de ces favoris avec des communautés. S'ajoute à cela la possibilité d'annoter des sites et de filtrer ensuite tous ces résultats.

L'organisation de ses favoris via des tags, des dossiers ou des vignettes pour un usage plus personnel que communautaire

Un moteur de recherche spécialement réservé aux vidéos qui permet d'accéder à l'ensemble de toutes les vidéos référencées sur une requête précise. Ces vidéos peuvent être hébergées chez Yahoo! mais aussi provenir d'autres fournisseurs de vidéos. Il est possible de charger des vidéos personnelles mais aussi de partager des vidéos trouvées via ce service.

Le moteur de recherche de Yahoo! est totalement gratuit. Vous pouvez l'utiliser via n'importe quel navigateur web ou via application mobile. [37]

I.7.4.3. Avantages et inconvénients de Yahoo :

Avantages

- ✓ Véritable portail d'information tourné vers l'actualité
- ✓ Sur la page d'accueil : possibilité d'aimer, commenter ou de partager sur les réseaux sociaux les différentes actualités
- ✓ Possibilité d'effectuer sa recherche dans les « questions/réponses » [38]

Inconvénients

- ✓ Tracke les internautes
- ✓ Résultats de recherche pas très lisible
- ✓ Ne propose pas de catégorie de recherche « carte » (contrairement à Google Maps et autres concurrents). Beaucoup trop de publicité [38]

I.7.5. Yahoo et Bing Vs Google :

Ayant été lancé en 1995, Yahoo! a été dans le jeu de recherche le plus longtemps, comparé à Google et Bing. Cependant, au cours de ses premières années, Yahoo! utilisé les résultats de recherche d'autres robots d'exploration Web, il a commencé sa propre indexation et son exploration en 2003. Bien que cela ne semble pas si menaçant, Yahoo! est le plus grand concurrent de Google

dans l'industrie des moteurs de recherche. Surtout que Yahoo! a formé un partenariat avec Bing en 2009, les forces combinées de Yahoo! et Bing constituent une menace sérieuse pour Google. Dernièrement, Yahoo! a apporté de nombreux changements et mises à niveau à ses algorithmes de moteur de recherche. Lorsqu'il s'agit d'indexer des sites Web, Yahoo! est assez bon pour ça. En ce qui concerne la pertinence des résultats fournis aux utilisateurs lorsqu'ils tapent leur requête dans le champ de recherche, il s'avère que Yahoo! se concentre davantage sur la correspondance des textes, plutôt que sur la numérisation du contenu pour déterminer sa pertinence. Leur point étant que, bien que Yahoo! place les anciens sites en haut du système de classement - ce n'est pas près de Google, qui place beaucoup plus haut les anciens sites sur son SERP.

Ainsi, pour eux, les nouveaux propriétaires de sites ont de bien meilleures chances d'apparaître sur le SERP de Yahoo que sur celui de Google.

I.7.6.DuckDuckGo :

Lancé en 2008, DuckDuckGo se présente comme un méta-moteur de recherche à la fois pertinent, en compilant les résultats de nombreux autres moteurs (Bing, Yahoo! ou Wikipédia, par exemple), mais aussi respectueux de la vie privée des internautes ne collectant aucun cookies ou adresses IP. Récemment, DuckDuckGo a présenté un nouvel outil baptisé "Tracker Radar" recensant en temps réel l'ensemble des trackers publicitaires circulant sur le Net. "Le moteur de recherche qui ne vous retrace pas" est de plus en plus prisé, notamment chez les jeunes internautes adeptes de confidentialité. [39].

I.7.6.1. Fonctionnement de DuckDuckGo :

DuckDuckGo est construit principalement sur les API de recherche de différents fournisseurs majeurs (tels que Yahoo! BOSS, embed.ly, Wolfram Alpha, Entire Web et Bing)47. Tech Crunch caractérise même « d'hybride » le service de moteur de recherche48,49. Dans le même temps, le moteur de recherche produit ses propres pages de contenu et est également similaire à des sites comme Mahalo, Kos mix et SearchMe. Le moteur de recherche DuckDuckGo est écrit en Perl et JavaScript et fonctionne avec nginx sur FreeBSD et Ubuntu50. [40]

I.7.6.2. Caractéristiques du moteur DuckDuckGo :

DuckDuckGo offre quelques Caractéristiques importants que les moteurs de recherche:

- **Confidentialité de l'utilisateur** – Aucune information personnelle n'est collectée et stockée sur les utilisateurs. Il n'y a pas d'historique de recherche stocké pour vous en tant qu'utilisateur.

Par conséquent, si des gouvernements ou des institutions comme la police demandent à DuckDuckGo des données vous concernant, aucune donnée ne leur sera communiquée.

- **– Protection des fuites de recherche**. Les sites Web que vous visitez depuis DuckDuckGo ne savent pas ce que vous avez cherché, pour y accéder.

- **Contact direct avec l'équipe de développeurs**. DuckDuckGo a une petite équipe derrière elle et elle compte sur sa communauté d'utilisateurs pour s'améliorer régulièrement. Si vous le souhaitez, vous pouvez fournir des commentaires à DuckDuckGo en vous rendant sur le site Web de leur communauté. Vous pouvez contribuer en soumettant des traductions, des idées de fonctionnalités, des rapports de bugs, etc.
- **Transparence sur son fonctionnement et son traitement par les utilisateurs**. Si vous voulez connaître leur histoire et leur évolution, où se trouve leur équipe, comment accéder à leur communauté et lire leur politique de confidentialité à jour, tout cela est accessible depuis le web. [34]

I.7.6.3. Avantages et inconvénients de DuckDuckGo

Avantages

- ✓ Moteur de recherche sécurisé car il ne collecte/stocke pas les informations personnelles
- ✓ Ne piste pas l'internaute
- ✓ Possibilité de personnaliser le thème, l'apparence… dans les paramètres
- ✓ Les « ! Bang » : commandes pour effectuer une requête redirigeant vers d'autres moteurs de recherche ou sites. Exemples : « ! g » recherche sur Google, « ! wfr » redirection vers le site de Wikipédia, « ! amz » recherche sur Amazon …
- ✓ Les résultats de recherche s'affichent sur une seule page (il suffit de scroller pour avoir les suivants)
- ✓ Design très minimaliste qui fait penser à Google [39]

Inconvénients

- ✓ Résultats de recherche pas toujours pertinents
- ✓ Obligation de valider sa recherche pour voir les options de tri (Web, Images, Vidéos et Actualités) car ils ne sont pas proposés sur la page d'accueil [39]

I.8. Conclusion

Dans la première partie de ce premier chapitre du volet théorique, nous avons présenté les principales notions et concepts de la RI. Nous y avons ainsi développé les principales étapes d'un processus de RI, à savoir, la représentation ou indexation de l'information, les principaux modèles existants, et autres.

Dans la seconde partie, nous avons exposé les moteurs de recherche classique (architecture, fonctionnement, comparaison...) tels que Google, Yahoo Bing et DuckDuckGo.

Et ce que nous avons retenu de cette description qu'il n'y a pas de différence significative entre ces MRs, car la plupart d'entre eux dépendent de la même méthode de recherche sur Internet. Mais, il y a une différence dans la vitesse de recherche des champs de recherche et comment les résultats sont organisés pendant le processus de recherche.

Chapitre II
Analyse des sentiments

II.1. Introduction

La construction d'un seul écosystème mondial d'objets qui communiquant entre eux est pratiquement impossible aujourd'hui. Il n'existe pas de protocole d'application unique et universel pour l'Internet des objets qui puisse fonctionner sur les nombreuses interfaces de réseau disponibles aujourd'hui. Pour être franc, l'IoT d'aujourd'hui est essentiellement une collection d'Intranets isolés d'objets qui ne peuvent réellement interagir les uns avec les autres.

Un domaine utilisant les techniques de IR (Recherche d'Informations), TC (Catégorisation de Texte), ML (Apprentissage Automatique) ou Fouille de Texte est notamment le domaine de l'Analyse des Sentiments, connu sur le nom de (ang : Opinion Mining). Les recherches dans ce domaine couvrent plusieurs sujets, notamment l'apprentissage de l'orientation sémantique des mots ou des expressions, l'analyse sentimentale de documents et l'analyse des opinions et attitudes à l'égard de certains sujets ou produits. Cependant, L'analyse de sentiments est l'étude computationnelle et sémantique des parties de textes en fonction des opinions, des sentiments et des émotions exprimés dans le texte.

Généralement l'expression « analyse des sentiments» est utilisée pour désigner la tâche de classification automatique des unités de texte en fonction de leur polarité (positive, négative, neutre).

Si l'on distingue d'habitude les sentiments des émotions, la difficulté à les différencier est de taille. Selon Jacques Cosnier [COSNIER, J. (1994). Psychologie des émotions et des sentiments. Retz.]:

- Les émotions sont des processus dynamiques qui ont un début et une fin et une durée relativement brève. Ces phénomènes "phasiques" sont causés par des événements précis et inattendus,
- Les "sentiments" tels que l'amour, la haine, l'angoisse, entre autres, se distinguent nettement des précédents par leurs causes plus complexes, par leur durée plus longue ("tonique"), et leur intensité plus basse. Ces définitions montrent bien le caractère évolutif des émotions.

II.2. Catégorisation des sentiments

Les phrases sont objectives ou subjectives. Lorsqu'une phrase est objective, aucune autre tâche fondamentale n'est requise. Lorsqu'une phrase est subjective, ses polarités (positive, négative ou neutre) doivent être estimées.

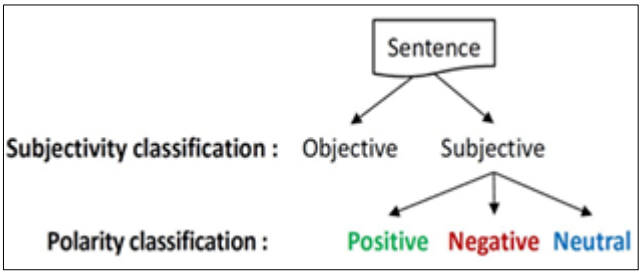

Figure II.1 : Workflow de l'analyse des sentiments [44]

La classification de subjectivité (Subjectivity classification) est la tâche qui distingue les phrases exprimant des informations objectives, des phrases exprimant des vues et opinions subjectives.

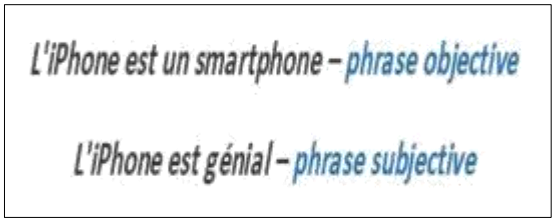

Figure II.2 : Distinction entre la subjectivité et l'objectivité via un exemple [44]

La classification de polarité et intensité de sentiment (Polarity classification) est la tâche qui distingue les phrases qui expriment des polarités positives, négatives ou neutres. [44]

L'intensité décrit à quel point la polarité d'une opinion est forte. Par exemple, dans une opinion à polarité positive, aimer est plus intense qu'apprécier, ou encore, dans une opinion de polarité négative, haïr est plus intense que de ne pas aimer [45].

II.3. Types d'analyse des sentiments

Il existe de nombreux types d'analyses de sentiments allant des systèmes qui se concentrent sur la classification de la polarité (positif, négatif, neutre) aux systèmes qui détectent des émotions (en colère, heureux, triste, etc.) ou identifient des intentions (par exemple, intéressé, pas intéressé). Nous abordons donc les types les plus importants [44]:

II.4. (fine-grained sentiment analysis)

Au lieu de parler de phrases positives, négatives ou neutres, nous considérons les catégories suivantes :

- Très positive
- Positive
- Neutre
- Négative
- Très négative

Certains systèmes offrent également différentes classifications de polarité en identifiant si le sentiment positif ou négatif est associé à un sentiment particulier, tel que la *colère*, la *tristesse* ou des inquiétudes (*sentiments négatifs*) ou du *bonheur*, de l'*amour* ou de l'*enthousiasme* (*sentiments positifs*).

II.5. Détection d'émotion (Emotion détection)

La détection des émotions vise à détecter des émotions telles que le bonheur, la frustration, la colère, la tristesse, etc. De nombreux systèmes de détection d'émotions sont basés sur l'utilisation de lexiques de sentiments (c'est-à-dire des listes des émotions) ou sur des algorithmes d'apprentissage automatique complexes.

II.6. Analyse de sentiments à base d'aspects (Aspect-Based Sentiment Analysis ABSA)

Au lieu de classer le sentiment général d'un texte en positif ou en négatif, l'analyse de sentiments à base d'aspects permet d'analyser le texte afin d'identifier différents aspects et de déterminer le sentiment correspondant pour chacun. Les résultats sont plus détaillés, intéressants et précis car l'analyse à base d'aspects examine de manière précise les informations contenues dans un texte.

II.7. Difficultés d'Analyse des Sentiments

Parmi les difficultés d'analyse [42]:

- Ambiguïté de certains mots positifs ou négatifs selon les contextes et qui ne peut pas être toujours levée.
- Difficulté due aux structures syntaxiques et sémantiques d'une phrase et l'expression de l'opinion: Par exemple " l'histoire du film est intéressante mais les acteurs étaient mauvais ". Dans ce cas la polarité de la deuxième partie est opposée à la première.
- Difficulté due au contexte : la nécessite d'une bonne analyse syntaxique du texte, analyse qui peut se révéler particulièrement difficile dans des cas de coordination entre plusieurs parties d'une phrase. Par exemple "ma tonte a bien préparé le gâteau, son décor est bonne mais je n'ai pas aimée le goût", l'opinion de la dernière partie de la phrase est la plus importante.
- Difficulté due à l'analyse de la phrase par " paquets de mots ": Les deux phrases suivantes contiennent les mêmes paquets de mots sans pour autant exprimer les mêmes sentiments. La première phrase contient un sentiment positif alors que la deuxième est négative : " Je

l'ai apprécié pas seulement à cause de ...", " Je l'ai pas apprécié seulement à cause de ... " ou se présente la gestion de négation.

- Difficulté due au langage qu'utilisent les internautes pour s'exprimer: Les ponctuations ne sont pas forcément utilisées pour marquer les fins de phrases, des mots spécifiques sont utilisés tel que : «ha haha», «Goood», «super».
- Difficulté de déterminer un lexique adapté à l'analyse de l'ensemble des textes d'opinion.

II.8. Domaines d'application d'analyse des sentiments

L'importance de l'analyse des sentiments est présentée dans plusieurs domaines, ainsi que plusieurs applications ont vu le jour dans ce contexte. Nous mentionnons brièvement quelques applications ci-dessous [45]:

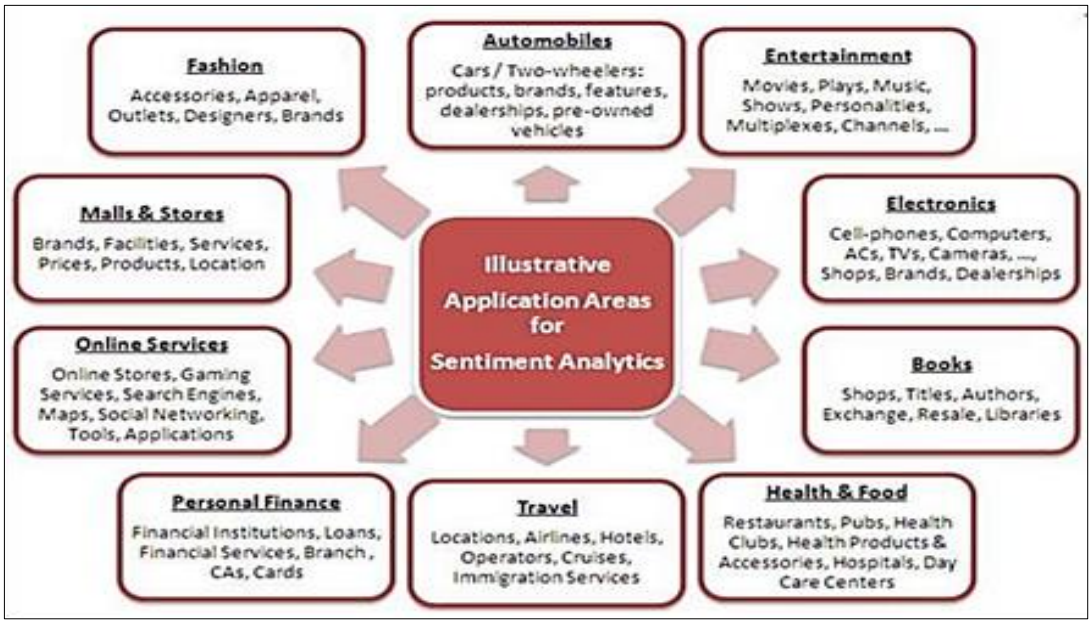

Figure II.3 : Domaines d'application d'analyse des sentiments

II.8.1. Politique :

Grace à l'analyse de sentiments et via généralement des sondages, les décideurs de politique prennent en considération les avis et opinions de leurs citoyens sur certaines politiques. Cette solution politique va les aider à améliorer ou à créer de nouvelles politiques qui conviennent à leurs sociétés.

II.8.2. Prise de décision :

L'opinion et l'expérience des gens sont un élément très utile dans le processus de prise de décision.

II.8.3. Les systèmes de recommandations :

À travers l'analyse des sentiments, on peut classer les opinions des individus en positive ou négative, le système définit qui devrait prendre la recommandation et qui ne devrait pas prendre la recommandation.

II.8.4. Domaine de Transport:

Pour assembler et analyser les opinions de public sur le statut de transport. Exemple : Système de transport intelligent moderne.

II.8.5. Domaine médical :

Analyse d'opinions des médecins, patients, médicaments, services hospitalier....

II.8.6. Domaine éducation :

Développer le niveau d'enseignement à travers l'analyse et l'interprétation de l'opinion de l'étudiant à travers les méthodes d'enseignement (.i.e améliorer l'enseignement et l'apprentissage).

II.8.7. Marketing :

Du côté entreprises, permet au fournisseur plus de connaissances à propos des besoins des consommateurs, du côté client il peut donner son opinion, s'inspirer des opinions d'autres clients pour l'aidé à sa décision et aussi comparer les produits avant de les acquérir.

II.9. Disciplines en relation avec l'analyse de sentiments :

Plusieurs disciplines ont une relation directe ou moins directe avec l'analyse de sentiments et l'opinion mining. Intelligence artificiel, traitement automatique du langage naturel, texte mining et même data mining offrent des outils algorithmiques indispensables pour le traitement et la classification des sentiments [44]

II.10. Exploration de texte (TextMining)

La fouille de textes peut être définie comme le processus au cours duquel le texte est transféré dans des données pouvant être analysées. Cela entraîne les procédures de création d'un index pour les termes individuels, basé sur l'emplacement du terme dans le texte d'origine, ou basé sur d'autres techniques ou protocoles. Les mots et les index peuvent ensuite être utilisés pour diverses méthodes d'analyse.

L'idée derrière l'indexation n'est pas de stocker l'emplacement du mot abstrait. Le processus d'indexation dans l'exploration de texte consiste plutôt à stocker la signification ou le concept du

mot dans le contexte donné. Les concepts sont ensuite stockés dans la base de données. Cette dernière est utilisée pour une analyse et un traitement plus poussés. Les concepts sont extraits du texte donné via une série de techniques linguistiques pour les processus d'extraction. [41]

L'exploration de données (Data mining en anglais), contrairement à l'exploration de texte, est le processus de découverte de modèles et de tendances au sein de grands ensembles de données. L'exploration de données est un domaine interdisciplinaire, impliquant des constructions issues des domaines de l'apprentissage automatique, de l'intelligence artificielle, des systèmes de bases de données et des méthodes statistiques. Les deux opérations les plus essentielles du processus d'exploration de données sont le regroupement et la classification. [41]

Parmi les domaines de l'exploration de texte dans l'exploration de données [42]:

Figure II.4 : Les domaines de l'exploration de texte dans l'exploration de données

A. Recherche d'informations (RI)

La RI est considérée comme une extension de la recherche de documents. Que les documents renvoyés soient traités pour se condenser. Ainsi, la recherche documentaire suit une étape de synthèse de texte. Cela se concentre sur la requête posée par l'utilisateur. Les systèmes IR aident à restreindre l'ensemble des documents qui sont pertinents pour un problème particulier. Comme l'exploration de texte implique l'application d'algorithmes très complexes à de grandes collections de documents. En outre, IR peut accélérer considérablement l'analyse en réduisant le nombre de documents.

B. Exploration de données (DM)

L'exploration de données peut être décrite vaguement comme la recherche de modèles dans les données. Il peut davantage se caractériser par l'extraction de données cachées. Les outils d'exploration de données peuvent prédire les comportements et les tendances futures. En outre, il

permet aux entreprises de prendre des décisions positives fondées sur les connaissances. Les outils d'exploration de données peuvent répondre aux questions commerciales. En particulier, cela a traditionnellement pris trop de temps à résoudre. Ils recherchent des bases de données pour des modèles cachés et inconnus.

C. Traitement du langage naturel (NLP)

TALN est l'un des problèmes les plus anciens et les plus difficiles. C'est l'étude du langage humain. La recherche TALN poursuit la vague question de savoir comment nous comprenons le sens d'une phrase ou d'un document? Quelles sont les indications que nous utilisons pour comprendre qui a fait quoi à qui? Le rôle de TALN dans l'exploration de texte est de fournir le système dans la phase d'extraction d'informations en tant qu'entrée.

D. Extraction d'informations (IE)

L'extraction d'informations est la tâche d'extraire automatiquement des informations structurées/non structurées. Dans la plupart des cas, cette activité comprend le traitement de textes en langage humain au moyen de la PNL.

II.11. Méthodes de classification des sentiments

Dans la littérature, il existe de nombreuses méthodes et algorithmes pour mettre en œuvre des systèmes d'analyse des sentiments, qu'on peut les classer comme suit [44]:

- **Approche d'apprentissage automatique :** systèmes qui s'appuient sur des techniques d'apprentissage automatique à partir de données.
- **Approche basée sur le lexique(TALN) :** systèmes qui effectuent une analyse des sentiments basée sur un ensemble de règles.
- **Approche hybride :** systèmes combinant à la fois des approches basées sur des règles et des approches automatiques.

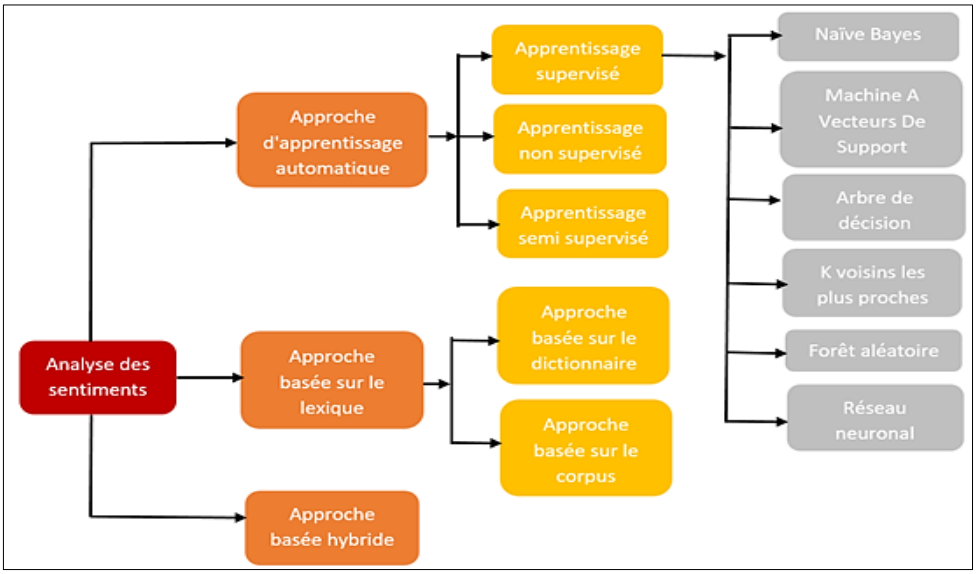

Figure II.5 : Approches d'analyse des sentiments

II.12. Conclusion

Dans ce chapitre, nous avons présenté la revue de littérature pour l'analyse des sentiments. Ceci comprend un survole sur les principaux fondements, méthodes et techniques d'analyse et classification des sentiments. Le chapitre suivant sera dédié à étudier la conception et réalisation de notre service de recherche de ressources via les émotions associées.

Chapitre III

Notre service de recherche émot-thématique

III.1. Introduction

Les moteurs de recherche sont des services qui ne fonctionnent pas strictement de la même manière, ils réalisent au moins les actions basiques, qui sont :

1) La collecte d'information,

2) l'analyse des mots à chercher et leur indexation et

3) la mise à disposition d'un processus de recherche aidant à retrouver des mots ou des combinaisons de mots dans la base d'index constituée.

Maintenant, si à ces actions basiques sont intégrés d'autres critères spécifiques comme une valeur additionnelle, comme le sentiment ou l'émotion. Est-ce que les résultats de recherche seront meilleurs? ET est qu'il s'agit du même processus d'analyse pour le sentiment que pour l'émotion?

III.2. L'analyse des sentiments

ne considèrent que la polarité (positive, négative, neutre) dans la tâche de classification. La majorité des études existants dans ce domaine, appelé aussi fouille d'opinions, consistent à classifier un mot, une phrase ou un document dans une des polarités en fonction des termes qui les composent ;

Dans **l'analyse des émotions**, l'objectif consiste à classer des énoncés (mots, phrases, documents) selon une liste de catégories émotionnelles joie, peur, surprise, . . .

Aux émotions peuvent également être associées des polarités (aussi appelée *valence* ou *orientation sémantique*) :

- Positive (attirance, joie, soulagement, . . .),
- Négative (colère, tristesse, peur, inquiétude, . . .),
- Neutre (étonnement, surprise, anticipation, . . .).

En plus, dans TAL (Traitement Automatique de Langues), les travaux sur l'analyse des émotions sont moins nombreux que ceux consacrés à l'analyse de sentiments.

Cependant, dans ce PFE, nous adoptons la notion large d'émotions et nous considérons également les sentiments comme des marqueurs émotionnels.

Notre objectif est de développer un service Émo-Thématique dédié à la recherche de News de la chaine BBC_News se focalisant sur les tâches d'expression des requêtes et documents et sur leurs correspondances au niveau du modèle d'appariement. Les résultats d'une requête soumise à

ce service seraient certainement plus pertinents s'il prend en compte ou intègre certaines spécificités liées à la requête et au document. Par spécificités, nous faisons allusion à l'exploitation des sentiments/émotions dans le processus de recherche. Nous partons donc de l'hypothèse qu'un document doit être classé (ranking) en fonction de sa pertinence thématique et émotionnelle.

III.3. Processus générale de notre service de recherche orienté émotion :

Notre système de recherche (SR) (figure 16) possède trois fonctions fondamentales qui définissent le modèle de recherche : représenter le contenu des documents, représenter le besoin de l'utilisateur et comparer ces deux représentations.

La représentation des documents et de la requête se fait à l'issue d'une phase appelée indexation qui consiste à choisir les termes représentatifs des documents et à les ajouter à un index qui à chaque terme associe le document dans lequel il se trouve. En effet, la préparation des documents ainsi que les requêtes sont l'étape la plus importante de notre SR. Elle se base sur une représentation simplifiée des données. Enfin, le modèle effectue un appariement entre ces deux représentations pour retourner le classement pertinent des résultats.

La pertinence Emo-textuelle entre la requête et le document améliore le calcul traditionnel de la similarité requête-document. Le mécanisme de recherche consiste donc à retrouver les documents qui s'approchent le plus de la requête en s'appuyant sur le texte augmenté par l'émotion et sur l'unité de mesure BM25. L'émotion va être ainsi intégrée dans le modèle de classement (ranking) afin d'en améliorer les résultats de recherche.

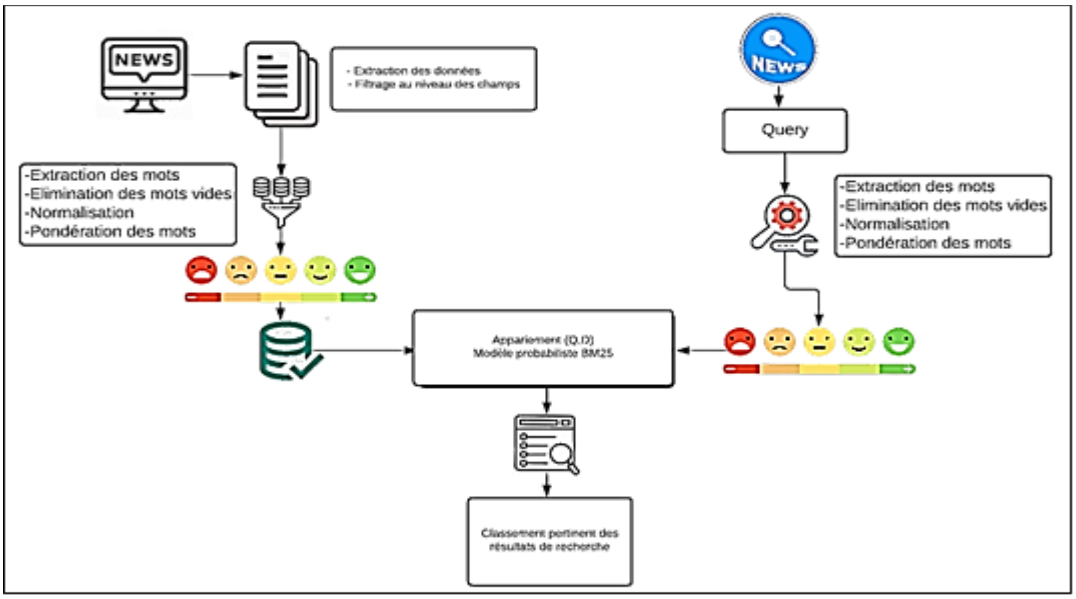

Figure III. 1: Processus générale de notre service de recherche orienté émotion

Le processus général de notre service de recherche est détaillé comme suit :

III.3.1 Collection de données (Dataset) :

Pour évaluer notre proposition, nous avons mené une série d'expérimentations sur la collection de chaines d'information BBC NEWS. Le choix de cette collection se justifie par des raisons techniques. Les ressources sont stockés au format CSV, qui est un format texte ouvert représentant des données tabulaires sous forme de colonnes séparées par des virgules.

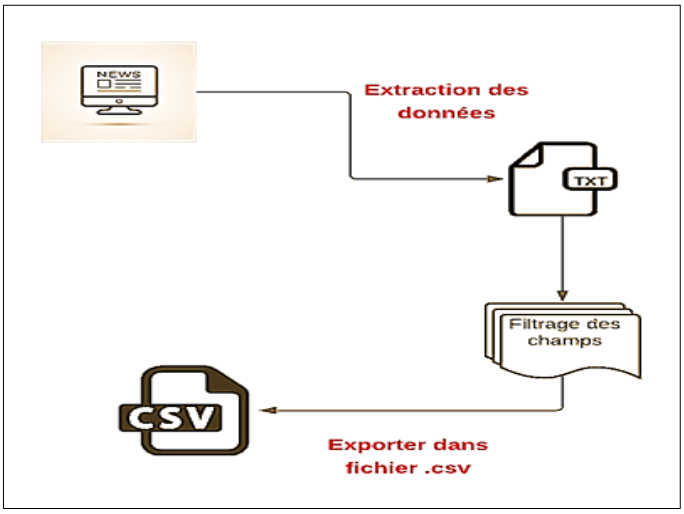

Figure III. 2: Préparation du dataset

III.3.2 Requête :

il s'agit des différentes formes que peuvent prendre les termes élémentaires de la requête, et comment ils sont articulés. En ce qui concerne la forme des termes, il faut considérer d'une part la gestion des variations linguistiques (lemmatisation et réduction dérivationnelle en morphologie et syntaxe, réduction conceptuelle en sémantique), et d'autre part la capacité de faire des recherches sur des syntagmes (mots composés, expressions).

Quant à l'articulation des termes composant la requête, l'utilisation d'une syntaxe formelle (par exemple booléenne) peut être vue comme contraignante (par les tenants des interfaces en langue naturelle) ou au contraire comme permettant une interrogation plus précise et plus puissante (combinaison d'opérateurs, pondérations). Un équilibre est à trouver, pour respecter les habitudes d'interrogation sans pour autant niveler par le bas (en s'interdisant des options plus complexes mais efficaces).

Notre principal objectif est de rechercher les documents pertinents à un sujet précis proposé par l'utilisateur. Ce sujet est représenté par une requête composé soit par un seul mot ou par une composition de mots. C'est pour cela, la requête est représentée directement comme au moins un token.

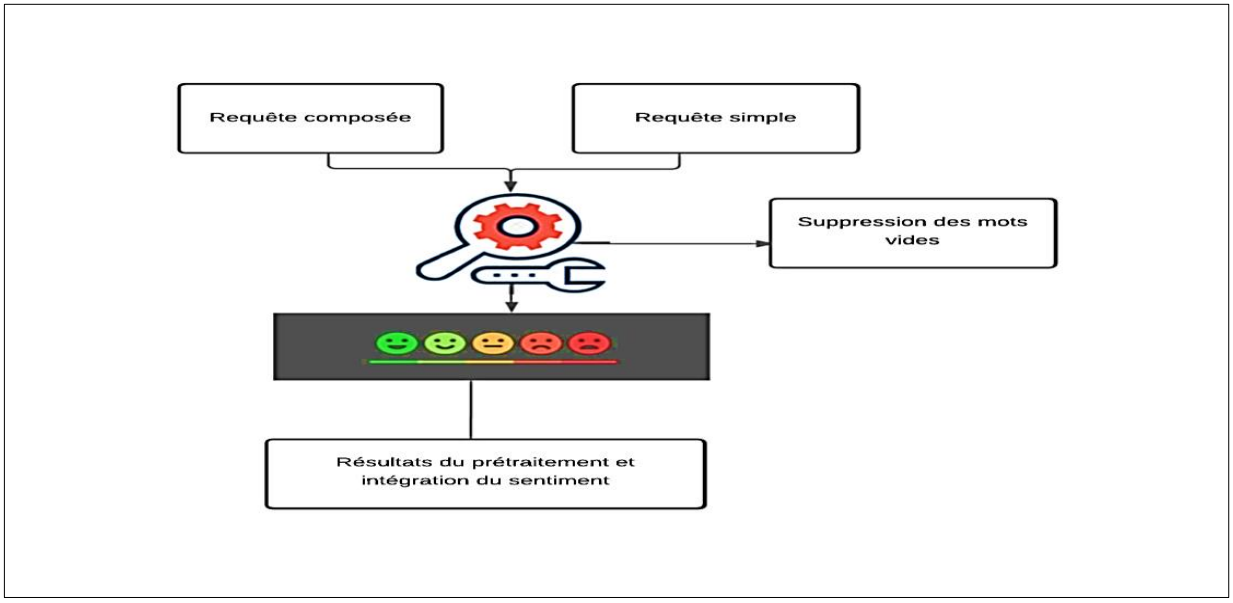

Figure III. 3: Prétraitement d'une requête simple ou complexe

III.3.3 Indexation requête-documents

Permet de représenter les documents (ou les requêtes), par un ensemble de termes clés afin de les identifier facilement par la suite. L'indexation recherche les mots qui modélisent le mieux possible le contenu informationnel d'un document. Les mots les plus représentatifs d'un document sont ceux qui apparaissent souvent dans ses textes. Mais, puisque les mots les plus fréquents sont des mots fonctionnels, qui ne représentent aucun intérêt informationnel (par exemple of, the, ...), il est nécessaire de faire appel à un mécanisme de filtrage de ces mots à partir d'une liste des mots appelées anti-dictionnaire, anti-lexiques ou 'stoplist'. Suite à une opération d'indexation, le corpus est représenté par une matrice composée de la liste des documents ainsi que par le poids des termes lui correspondant. Le poids des termes est déterminé selon des formules mathématiques. C'est ce qu'on appelle l'unité d'information ou de mesure. Plusieurs paramètres sont à considérer selon l'unité d'information, comme, la fréquence des termes, la longueur des documents, ...

III.3.4 Appariement Requête - documents:

Est un mécanisme qui s'appuie sur une relation de similarité entre les termes de la requête et ceux d'une collection afin de déterminer l'ensemble des documents pertinents et ordonnés selon leur degré de correspondance. Plusieurs modèles ont déjà été proposés et développés pour réaliser cette fonction d'appariement. On trouve par exemple: le modèle vectoriel, le modèle booléen, ...etc. L'unité de mesure BM25 (BM signifie: 'Best Match') est une méthode de pondération des termes dans les documents et les requêtes selon le modèle probabiliste de pertinence développé par

Robertson et Sparck Jones [Robertson, S.E. et Sparck Jones, K. (1976) "Relevance Weighting of Search Terms", Journal of the American Society for Information Science, (Vol. 27, pp. 129-146)]:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

- f (qi, D) est la fréquence du terme qi dans le document D,
- |D| est la longueur du document D en nombre de mots,
- avgdl est la longueur moyenne des documents dans la collection (ou corpus),
- k1 et un paramètre d'optimisation généralement situé entre 1,2 et 2. Dans notre cas nous avons pris la moyenne i.e. : 1.6,
- b est un autre paramètre généralement fixé à 0,75,
- IDF(qi) est la fréquence inverse de document déterminée par le calcul suivant :

$0 < \text{IDF} < 1$ $\qquad\qquad$ $\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$

III.3.5 Processus générale de notre méta-service de recherche classique:

La recherche basique est celle des moteurs de recherche classique ou l'internaute fait entrer une requête et attend un ensemble de résultats dépendants. La requête sera soumise au site web Stackoverflow en exploitant en ligne la base de données de MongoDB et l'Api Google.

Stack Overflow est un site web proposant des questions et réponses sur un large choix de thèmes concernant la programmation informatique. Il fait partie du réseau de sites StackExchange. Et MongoDB est une base de données orientée documents. En clair, nous bénéficions de la scalabilité et de la flexibilité que nous voulons, avec les fonctions d'interrogation et d'indexation qu'il nous faut.

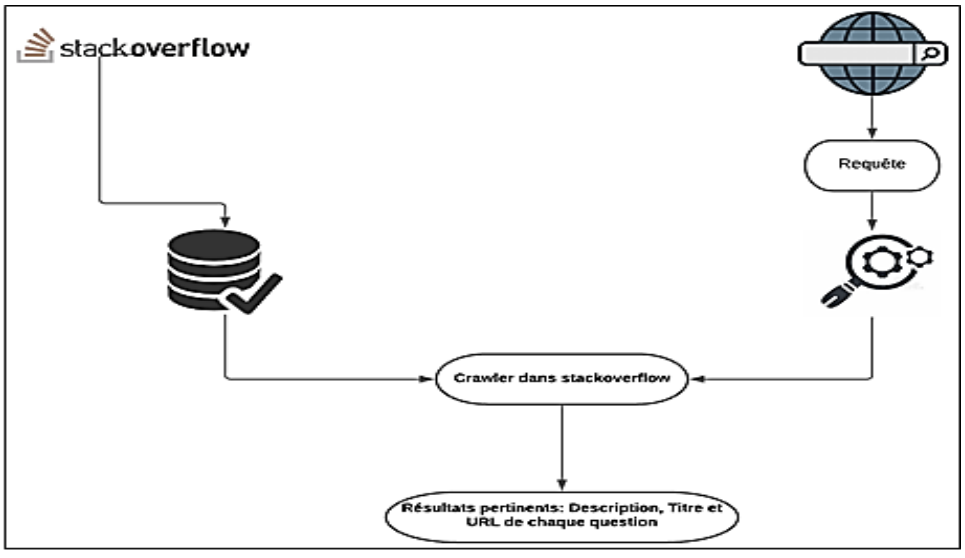

Figure III. 4: Processus générale du meta-service de recherche classique

III.4. Analyse et organisation des données (document-requête):

III.4.1 Prétraitement des données :

Étant donné que les données brutes sont biaisées et bruyantes, leurs nettoyages sont notre première tâche. Ce prétraitement consiste à structurer et à faciliter l'utilisation de ces données d'entrées. Cependant, le prétraitement est le processus de nettoyage et de préparation des différents contenus pour l'indexation.

Tout d'abord, notre SR identifie les données inutiles et gênantes pour le processus de la recherche. L'étape de prétraitement consiste soit à les supprimer pour réduire la complexité du système, ou bien à les mieux organiser pour donner une structure simple et lisible aux documents qui y sont exploités. Enfin, les données prétraitées représentent les données prêtes à être utilisées dans le processus de recherche.

La phase de prétraitement suit directement la sélection du dataset, cette étape est très importante et critique dans le processus de l'analyse de sentiments.

Les mots sont souvent dupliqués et peuvent contenir des caractères spéciaux ou des mots indésirables. Tout le nettoyage se fait dans l'optique d'avoir un corpus avec un vocabulaire minimal d'une part, d'autre part, avec un maximum de mots informatifs, parlants, qui portent de l'information pertinente concernant la polarité.

Le processus de prétraitement est comme suivi:

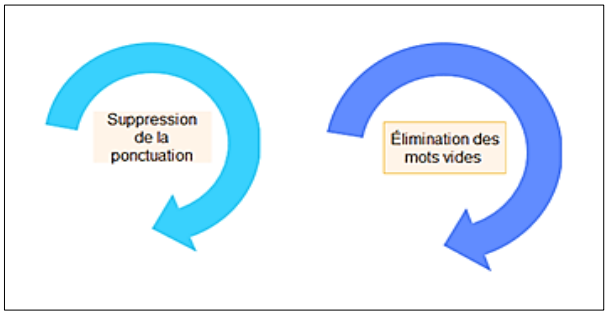

Figure III. 5: Processus de prétraitement

III.4.2 Suppression de la ponctuation

Dans de nombreux travaux, il est courant de supprimer les signes de ponctuation en prétraitement. Il s'agit encore une fois d'un processus de normalisation de texte qui aidera à traiter « hourra » et « hourra ! » de la même manière. Une façon de le faire est de parcourir la série avec la compréhension de la liste et de conserver tout ce qui n'est pas dans string. Ponctuation, une liste

de toute la ponctuation que nous avons importée au début avec import string [46]. Cet ensemble de symboles sont : [!"#$%&'()*+,./:;<=>?@[\]^_`{|}~]: [47]

Entrée: [massacre""? or ""attack"" ""?an incident that never happened]

Sortie: [massacre or attack an incident that never happened]

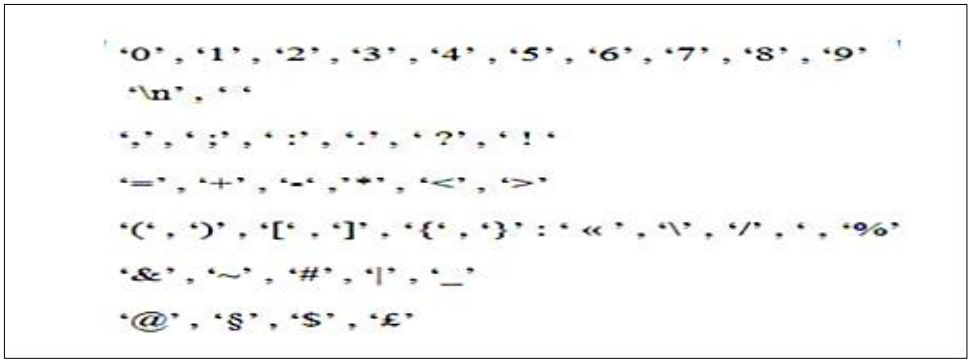

Figure III. 6: Liste de ponctuation

III.4.3 . Suppression de mots vides

Les mots vides sont des mots courants dans une langue comme « in», «all »,... etc. Ils peuvent être supprimés du texte la plupart du temps, car ils ne fournissent pas d'informations précieuses pour l'analyse en aval. Dans des cas tels que le balisage Part of Speech, nous ne devons pas les supprimer car ils fournissent des informations très précieuses sur le point de vente.

Ces listes de mots vides sont déjà compilées pour différentes langues et nous pouvons les utiliser en toute sécurité.

Par exemple, la liste des mots vides pour la langue anglaise du package nltk peut être vue ci-dessous :

Entrée : [Betsy DeVos Confirmed as Education Secretary, With Pence Casting Historic Tie-Breaking Vote']

Sortie : ['Betsy', 'DeVos', 'Confirmed', 'Education', 'Secretary', ',', ' With', 'Pence', 'Casting', 'Historic', 'Tie-Breaking', 'Vote']

Langue	Nombre de mot vides	Exemple
English	**635**	**All, also, for, in…**

Tableau. III. 1: Mots vides anglais [82]

Puisque nous nous intéressons aux mots vides en anglais, l'algorithme suivant illustre les étapes à suivre pour l'élimination de ces mots vides :

```
Entrée : tableau de tokens
         liste des mot vides
Sortie :  tableau de tokens
Début :
      Pour chaque token du tableau faire
            Si  le token appartient à la liste des mots vides
               Alors  supprimer le token du tableau
            Finsi
      Finpour
Fin.
```

Figure III. 7: Algorithme d'élimination de mots vides

III.4.4 Lemmatisation

La lemmatisation est l'étape qui désigne l'analyse lexicale chargée de faire regrouper les mots d'une même famille qui partagent le même suffixe lexical. Chacun des mots du texte se trouve ainsi réduit en une entité appelée « Lemme ». Ce lemme désigne la forme canonique des mots. La lemmatisation regroupe les différentes formes que peut avoir un mot.

Par exemple, un nom en pluriel va être réduit au singulier, un verbe à son infinitif, etc.

La lemmatisation est étroitement liée à la racinisation. La différence est qu'un stemmer fonctionne sur un seul mot sans connaissance du contexte, et ne peut donc pas faire la distinction entre des mots qui ont des significations différentes selon la partie du discours. Cependant, les stemmers sont généralement plus faciles à implémenter et à exécuter plus rapidement, et la précision réduite peut ne pas avoir d'importance pour certaines applications.

À la fin de cette étape d'indexation, une représentation du document et de la requête est prête à être utilisée au sein du modèle d'appariement "Emo_thematique"

III.4.5 Quantification individuelle des polarités

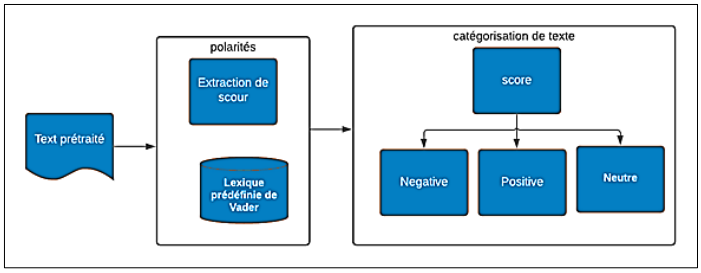

Figure III. 8: Quantification individuelle de polarités

L'approche d'AS adoptée est basée sur un lexique (Lexicon based approach) c-à-d elle utilise un lexique prédéfini de Vader, appelé lexique ou dictionnaire. Cette approche aide à identifier la polarité d'un texte en se servant de catégories de mots prédéfinis et pondérés. Elle identifie tous les mots positifs, neutre ou négatifs au sein d'un texte. Le dictionnaire est constitué d'un petit ensemble de mots d'opinion subdivisés généralement en catégories. L'une des catégories comporte des mots dont la terminologie est plus positive, tandis que l'autre catégorie regroupe les mots associés à un sentiment plus négatif. Pour établir un dictionnaire, il commence généralement par les mots les plus intuitifs et qui expriment un sentiment positif, le nombre de ces mots est amplifié par leurs synonymes constitue la catégorie positive. Une catégorie de mots négatifs peut être automatiquement constituée à partir des antonymes de mots de la catégorie positive, puis elle est amplifiée par d'autres mots jugés négatifs. À chaque mot est associée une pondération (W). La somme des mots (S) est soit positive, négative ou neutre et représente une évaluation globale du sentiment dans le texte.

L'algorithme utilisé dans l'approche basée sur un lexique génère des scores de sentiment dans 4 classes de sentiments et il est exprimé comme suit :

Algorithme ***Extraction Individuelle des polarités***

Variables s: note totale, w: pondération du token

1. Initialiser la note totale du sentiment s=0
2. Pour chaque token du texte, vérifier sa présence dans le lexique :
    a. Si le token est présent :
        i. Si le token est positif, alors : s=s + w
        ii. Si le token est négatif, alors : s=s – w
3. Evaluer le sentiment total du texte
    a. Si s >= 0.5, alors le sentiment du texte est positif (pos)
    b. Si - 0.5 < s < 0.5, alors le sentiment du texte est neutre (neu)
    c. Si s <=- 0.5, alors le sentiment est négatif (neg)

d. Sentiment composé (c'est à dire score agrégé)

Le score composé est calculé en additionnant les scores de valence de chaque mot du lexique, ajusté selon les règles, puis normalisé pour être compris entre -1 (plus extrême négatif) et +1 (plus extrême positif). C'est la métrique la plus utile si vous voulez une seule mesure unidimensionnelle du sentiment pour une phrase donnée.

La normalisation la plus utilisée est: $$x = \frac{x}{\sqrt{x^2+a}}$$

Où x = somme des scores de valence des mots constitutifs, et a = constante de normalisation (la valeur par défaut est 15)

III.4.6 Sauvegarde des données

Une fois la catégorisation des sentiments terminés. Les données seront stockées dans un fichier.CSV. En l'ouvrant, vous constaterez que l'ID, le texte et le sentiment correspondant sont séparés en 3 colonnes. Avec cette sortie, Nous aurons un ensemble de données de requête et de documents et leurs sentiments correspondant filtrés par des mots-clés (Pos., nég., et neu.).

	publish_date	article_source_link	title	text	scores	compound	comp_score
0	2017-02-07 00:00:00	http://abcnews.go.com/Politics/pence-break-tie...	Betsy DeVos Confirmed as Education Secretary W...	Michigan billionaire education activist Betsy ...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	pos
1	2017-02-07 00:00:00	http://abcnews.go.com/Politics/wireStory/melan...	Melania Trump Says White House Could Mean Mil...	First lady Melania Trump has said little about...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	pos
2	2017-02-07 00:00:00	http://abcnews.go.com/Politics/wireStory/trump...	As Trump Fears Fraud GOP Eliminates Election C...	A House committee voted on Tuesday to eliminat...	{'neg': 0.524, 'neu': 0.476, 'pos': 0.0, 'comp...	-0.7650	neg
3	2017-02-07 00:00:00	http://abcnews.go.com/Politics/appeals-court-d...	Appeals Court to Decide on Challenge to Trumps...	This afternoon three federal judges from the 9...	{'neg': 0.0, 'neu': 0.885, 'pos': 0.115, 'comp...	0.0772	pos
4	2017-02-07 00:00:00	http://abcnews.go.com/US/23-states-winter-weat...	At Least 4 Tornadoes Reported in Southeast Lou...	At least four tornadoes touched down in Louisi...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	pos

Figure III. 9: Fichier Csv enrichi de polarités

III.4.7 Appariement document-requête:

La relation d'appariement consiste à rechercher parmi les documents prétraités, ceux qui répondent le mieux à la requête c.-à-d. calculer un score de pertinence entre la requête et le document selon un score de correspondance entre ces deux représentations augmentées de polarités. Pour notre système, le calcul du score d'appariement s'effectue en se basant sur l'unité de mesure BM25 et l'émotion.

✍ Exemple de classification et de polarités calculées d'une requête en 'neg', 'neu', 'pos', 'compound':

Requête soummise = "love education"

Sentiment_Requête_renvoyé = {'neg': 0.0, 'neu': 0.192, 'pos': 0.808, 'compound': 0.6369}

III.4.8 Résultat final de la recherche d'appariement (Q,D) :

Le résultat final de cette étape d'appariement est un tableau des scores de pertinence requête-document:

Requête soummise = "Mocks"

Résultats "Mocks" , "mocks":

	article_source_link	title	text	score émotionnelle	comp_score
0	http://abcnews.go.com/International/wireStory/...	EU to Britain Pay Up for What You Ordered Befo...	The European Union is warning Britain that any...	-0.1027	neg
1	http://abcnews.go.com/International/wireStory/...	Irans Top Leader Mocks Newcomer Trump	Irans supreme leader said Tuesday that newcome...	-0.2960	neg
2	http://abcnews.go.com/International/wireStory/...	Mother of Backpacker Slain in Australia Critic...	The mother of a backpacker slain in an Austral...	-0.3400	neg
3	http://www.cnn.com/2017/02/07/politics/yemen-r...	Al Qaeda leader mocks Trump after Yemen raid	In an 11minute recording AQAP leader Qassim al...	-0.4588	neg
4	http://www.cnn.com/2017/02/07/europe/romanian-...	People appalled by corruption bill Romanian Pr...	President Klaus Iohannis said it is time for R...	-0.4588	neg
5	http://abcnews.go.com/Politics/times-kellyanne...	2 Other Times Kellyanne Conway Referred to Bow...	Last week senior Trump adviser Kellyanne Conwa...	-0.4767	neg
6	http://www.cnn.com/2017/02/07/politics/white-h...	Journalists call out White House claims on ter...	On Monday the White House issued a list of 78 ...	-0.5267	neg
7	http://abcnews.go.com/Politics/homeland-securi...	Homeland Security Secretary John Kelly Defends...	In his first appearance before Congress as sec...	-0.5927	neg
8	http://abcnews.go.com/US/multi-state-manhunt-s...	MultiState Manhunt in Southeast Intensifies fo...	A manhunt is intensifying in the Southeast for...	-0.6808	neg

Tableau. III. 2: Résultats de l'appariement (Q,D)

III.5. Conclusion

Nous avons fait beaucoup d'effort dans deux volets important dans ce PFE. Le premier est l'annotation et analyse émotionnelle de la requête et du document par positive, négative, neutre ou composé. Le deuxième est le calcul du score de pertinence final entre la requête et les documents en exploitant un modèle de probabilité.

Dans le chapitre suivant, nous allons présenter notre expérimentation et discuter les résultats obtenus**.**

Chapitre IV :

MISE EN ŒUVRE D'UN SERVICE DE RECHERCHE ORIENTE EMOTION

IV.1. Introduction

Après avoir établi une étude conceptuelle de notre service de recherche (SR) spécifique et dédié à la recherche de News de la chaine BBC_News, nous passons à l'implémentation de ce dite service de recherche tout en présentant les outils utilisés et en expliquant les étapes de de recherche. Cette implémentation est basée sur la recherche d'information orientée émotion et réalisée avec le langage de programmation Python.

IV.2. Environnement de travail :

Nous présentons brièvement notre environnement de travail en montrant les principaux outils, langages et bibliothèques utilisées pour la mise en œuvre de notre moteur de recherche orienté émotion:

IV.2.1 Outils de développement

- **Anaconda :** est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique, qui vise à simplifier la gestion des paquets et de déploiement. Les versions de paquetages sont gérées par le système de gestion de paquets Conda. [48]

- **Notebook Jupyter :** Le notebook Jupyter est un environnement HTML pour Python, similaire à Mathématico ou à Maple. Il fournit un environnement organisé en cellules interactives qui peuvent être exécutées, permettant l'organisation et la documentation de calculs de façon structurée. Le notebook jupyter peut être invoqué en se plaçant dans le répertoire contenant les fichiers notebook (.ipynb) et en tapant dans un terminal la commande :jupyter notebook[49]

- **VS code :** (Visual Studio Code) est un éditeur de code open-source, gratuit et multi-plateforme (Windows, Mac et Linux), développé par Microsoft, à ne pas confondre avec Visual Studio, l'IDE propriétaire de Microsoft. Principalement conçu pour le développement d'application avec JavaScript, TypeScript, PHP et Node.js, l'éditeur peut s'adapter à d'autres types de langages grâce à un système d'extension bien fourni. [50]

- **XAMPP :** est un ensemble de logiciels permettant de mettre en place un serveur Web local, un serveur FTP et un serveur de messagerie électronique. Il s'agit d'une distribution de logiciels libres (X (cross) Apache MariaDB Perl PHP) offrant une bonne souplesse d'utilisation, réputée

pour son installation simple et rapide. Il permet de configurer un serveur de test local avant la mise en œuvre d'un site internet. [36]

IV.2.2 Langages de programmation

- **Python 10:** est un langage de programmation open source créé par le programmeur Guido van Rossum en 1991. Il s'agit d'un langage de programmation interprété, qui ne nécessite donc pas d'être compilé pour fonctionner. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser. [51]

IV.2.3 Bibliothèques principales

- **NLTK:** (Natural Language Toolkit) est une bibliothèque logicielle en Python permettant un traitement automatique des langues, développée par Steven Bird et Edward Loper du département d'informatique de l'université de Pennsylvanie. En plus de la bibliothèque, NLTK fournit des démonstrations graphiques, des données-échantillon, des tutoriels, ainsi que la documentation de l'interface de programmation (API). [52]
- **Pandas :** est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. Pandas est un logiciel libre sous licence BSD. [53]
- **BM25 *:*** est une fonction de récupération de sacs de mots qui classe un ensemble de documents en fonction des termes de requête apparaissant dans chaque document, quelle que soit leur proximité dans le document. Il s'agit d'une famille de fonctions de notation avec des composants et des paramètres légèrement différents. [54]
- **Matplotlib :** est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques. [55]

VADER : (Valence Aare Dictionary and Sentiment Reasoner) est un lexique et un outil d'analyse des sentiments basé sur des règles, spécifiquement adapté aux sentiments exprimés dans les médias sociaux. Il est entièrement open-source et est disponible dans le package NLTK qui peut être appliqué directement à des données texte non étiquetées. VADER est capable de détecter la polarité et l'intensité des émotions. VADER est sensible à la fois à la polarité (positive / négative) et à l'intensité (force) de l'émotion. ... L'analyse sentimentale VADER s'appuie sur un dictionnaire qui mappe les caractéristiques lexicales aux intensités d'émotions appelées scores de sentiment. [56]

IV.3. Diagramme d'activité globale du SR orienté émotion

Le diagramme d'activité de la figure (27) montre le fonctionnement de notre MR et en particulier la phase de matching document-requête. Tout d'abord, l'utilisateur doit sélectionner la collection de test (dataset) dans laquelle il souhaite effectuer une recherche, puis il introduit le sujet de recherche (requête). Cependant, le système commence son processus de RI par une annotation automatique des documents et de la requête pré-traités et pré-indexés (tokens). Si la requête est satisfaite dans ce document, alors son score de similarité Emo-Thématique est calculé (la valeur du score est nécessaire afin de classer les documents), et donc le document est pertinent à la requête. Sinon, la requête n'est pas satisfaite et le document est non pertinent.

Les statistiques des données exploitées dans le processus de recherche sont comme suit:

	Titre	Mots	Vocabulure
Data set	3824	33866	11210

Tableau. IV. 1 : Statistiques des données raffinées

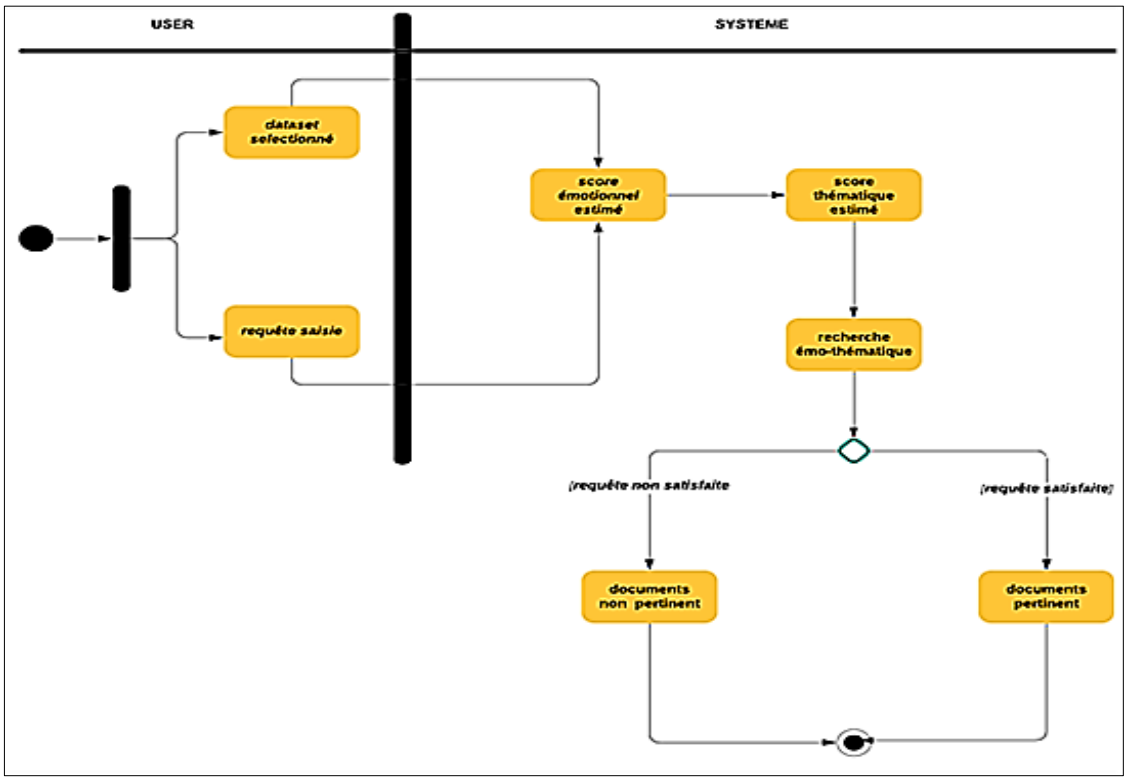

Figure IV. 1 : Diagramme d'activité du processus de recherche global.

IV.4. Diagramme d'activité des évaluations de scores

Une fois les tokens (mots) sont prétraités, pré-indéxé et étiquetées en fonction de leur orientation sémantique comme positives, négatives ou neutres, la métrique du score composé (compound) calcule la somme de toutes ces évaluations qui ont été normalisées entre -1 (le plus extrême négatif) et +1 (le plus extrême positif):

Sentiment positif: (score (composé) > 0)

Sentiment neutre: (score (composé) = 0)

Sentiment négatif: (score (composé) < 0)

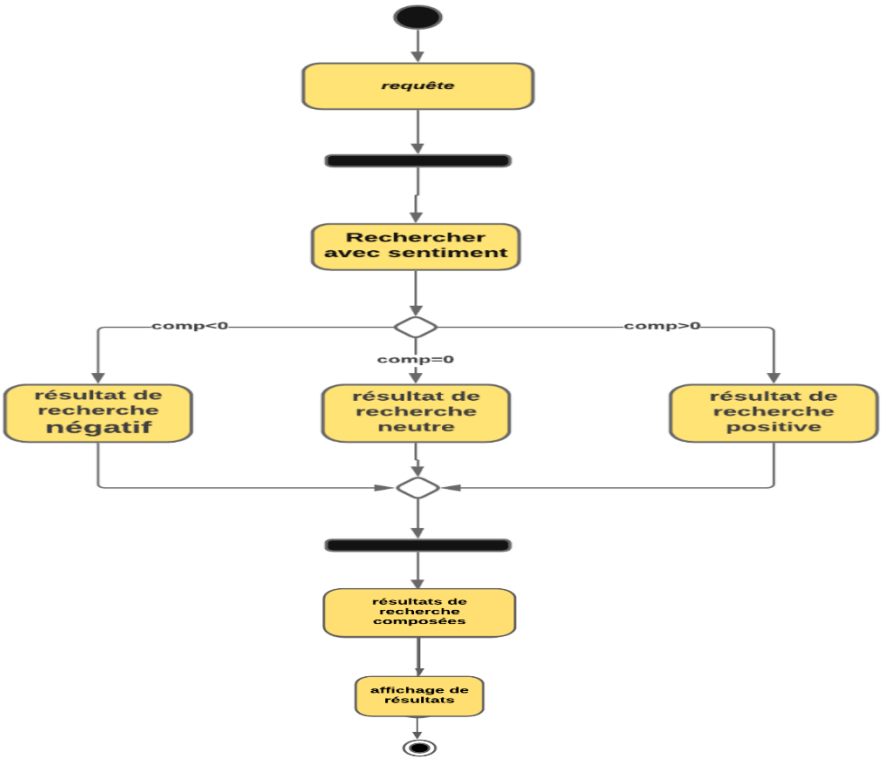

Figure IV. 2 : Diagramme d'activité des évaluations de scores ('pos', 'neu', 'neg' et 'comp')

IV.5. Diagramme d'activité du méta-service de recherche standard

Cette recherche passe par la séquence suivante:

- Soumission de la requête de recherche
- Correspondance entre la requête et les questions posées dans le stackoverflow
- Affichage des résultats de recherche pertinents (Description, titre et URL de chaque question)

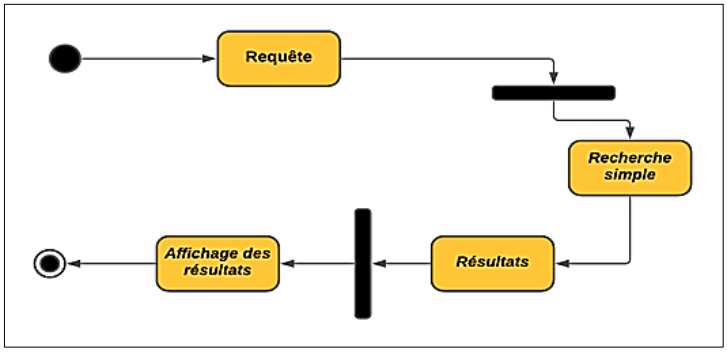

Figure IV. 3 : Diagramme d'activité du méta-service de recherche standard

IV.6. Diagramme de séquence de la collecte et mémorisation des données

Une fois la collection de données est choisie, la version brute est sauvegardée sous forme de fichier texte. Txt. Ce dernier subira un filtrage au niveau des rubriques c'est à dire à ne garder que les champs indispensables au processus de recherche. Le document .txt se transforme par la suite en .csv.

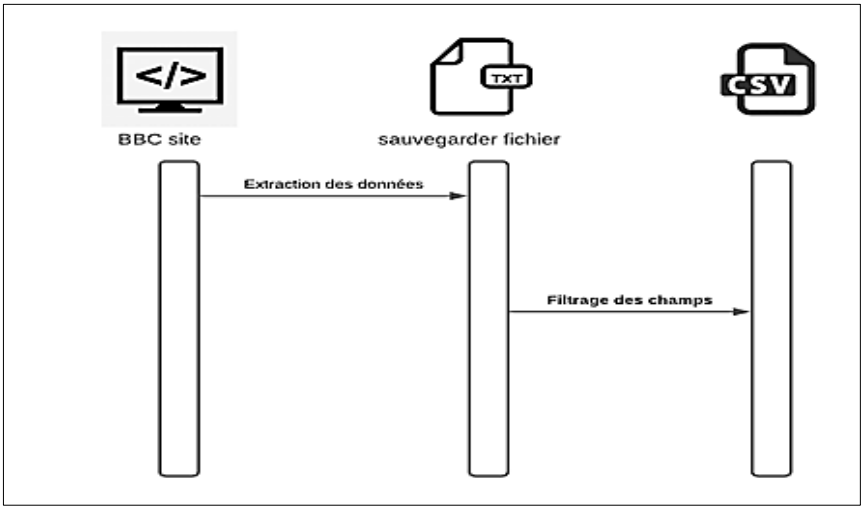

Figure IV. 4 : Diagramme de séquence de la collecte et stockage des données

IV.7. Prétraitement du contenu textuel des données:

Le prétraitement est l'un des éléments les plus importants du processus d'analyse. Il reformate les données non structurées en une forme uniforme et normalisée. Les caractères, les mots et les phrases identifiés à ce stade sont les unités fondamentales transmises à toutes les étapes ultérieures du traitement. La qualité du prétraitement a une grande influence sur le résultat final de l'ensemble du processus ***:***

```
Entrée [12]:
import string
def remove_punctuations(text):
    for punctuation in string.punctuation:
        text = text.replace(punctuation, '')
    return text
# Remove punctuations in pandas
df1["title"] = df1['title'].apply(remove_punctuations)
df1["text"] = df1['text'].apply(remove_punctuations)

Entrée [13]:
df1["text"]

Out[13]:
0       Michigan billionaire education activist Betsy ...
1       First lady Melania Trump has said little about...
2       A House committee voted on Tuesday to eliminat...
3       This afternoon three federal judges from the 9...
4       At least four tornadoes touched down in Louisi...
                              ...
3819    The tribunal set up to investigate an alleged ...
3820    Plans to tilt the wreckage of Coast Guard Resc...
3821    Children in Northern Ireland are eating three ...
3822    Hundreds of people with disabilities and their...
3823    All ten victims of a fire at a halting site in...
```

Figure IV. 5 : Prétraitement du contenu textuel

IV.7.1 Analyse de sentiment :

L'analyse du sentiment est un processus d'analyse et de classification des données en fonction des besoins en informations :

IV.7.2 Analyse des sentiments des documents :

```
Entrée [3]:
from nltk.sentiment.vader import SentimentIntensityAnalyzer
SIA = SentimentIntensityAnalyzer()

Entrée [4]:
SIA.polarity_scores(df.loc[0]['text'])

Out[4]: {'neg': 0.076, 'neu': 0.881, 'pos': 0.043, 'compound': -0.9249}

Entrée [6]:
df['scores'] = df['text'].apply(lambda review: SIA.polarity_scores(review))
df.head()
```

Out[6]:

	publish_date	article_source_link	title	text	scores
0	2017-02-07 00:00:00	http://abcnews.go.com/Politics/pence-break-tie...	Betsy DeVos Confirmed as Education Secretary W...	Michigan billionaire education activist Betsy ...	{'neg': 0.076, 'neu': 0.881, 'pos': 0.043, 'co...
1	2017-02-07 00:00:00	http://abcnews.go.com/Politics/wireStory/melan...	Melania Trump Says White House Could Mean Mill...	First lady Melania Trump has said little about...	{'neg': 0.097, 'neu': 0.811, 'pos': 0.092, 'co...
2	2017-02-07 00:00:00	http://abcnews.go.com/Politics/wireStory/trump...	As Trump Fears Fraud GOP Eliminates Election C...	A House committee voted on Tuesday to eliminat...	{'neg': 0.093, 'neu': 0.832, 'pos': 0.075, 'co...
3	2017-02-07 00:00:00	http://abcnews.go.com/Politics/appeals-court-d...	Appeals Court to Decide on Challenge to Trumps...	This afternoon three federal judges from the 9...	{'neg': 0.094, 'neu': 0.801, 'pos': 0.105, 'co...

```
Entrée [7]:
df['compound'] = df['scores'].apply(lambda score_dict: score_dict['compound'])
df.head()
```

Out[7]:

	publish_date	article_source_link	title	text	scores	compound
0	2017-02-07 00:00:00	http://abcnews.go.com/Politics/pence-break-tie...	Betsy DeVos Confirmed as Education Secretary W...	Michigan billionaire education activist Betsy ...	{'neg': 0.076, 'neu': 0.881, 'pos': 0.043, 'co...	-0.9249
1	2017-02-07 00:00:00	http://abcnews.go.com/Politics/wireStory/melan...	Melania Trump Says White House Could Mean Mill...	First lady Melania Trump has said little about...	{'neg': 0.097, 'neu': 0.811, 'pos': 0.092, 'co...	-0.3484
2	2017-02-07 00:00:00	http://abcnews.go.com/Politics/wireStory/trump...	As Trump Fears Fraud GOP Eliminates Election C...	A House committee voted on Tuesday to eliminat...	{'neg': 0.093, 'neu': 0.832, 'pos': 0.075, 'co...	-0.9029
3	2017-02-07 00:00:00	http://abcnews.go.com/Politics/appeals-court-d...	Appeals Court to Decide on Challenge to Trumps...	This afternoon three federal judges from the 9...	{'neg': 0.094, 'neu': 0.801, 'pos': 0.105, 'co...	-0.1333
4	2017-02-07 00:00:00	http://abcnews.go.com/US/23-states-winter-weat...	At Least 4 Tornadoes Reported in Southeast Lou...	At least four tornadoes touched down in Louisi...	{'neg': 0.057, 'neu': 0.937, 'pos': 0.006, 'co...	-0.9729

Figure IV. 6 : Code source et résultats de sentiment des documents

IV.7.3 Statistiques des données les plus utilisées

Les figures 7 et 8 dans ce chapitre représentent les statistiques des mots et l'ensemble des mots positifs et négatifs les plus utilisés dans notre dataset après l'application du processus d'analyse de sentiments.

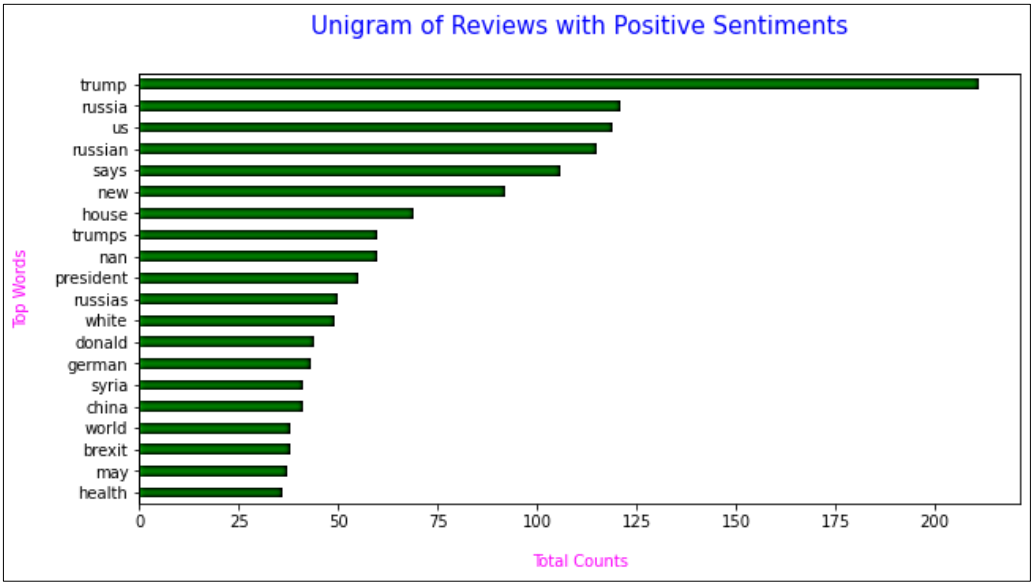

Figure IV. 7 : Statistiques de chaque mot positif le plus répété

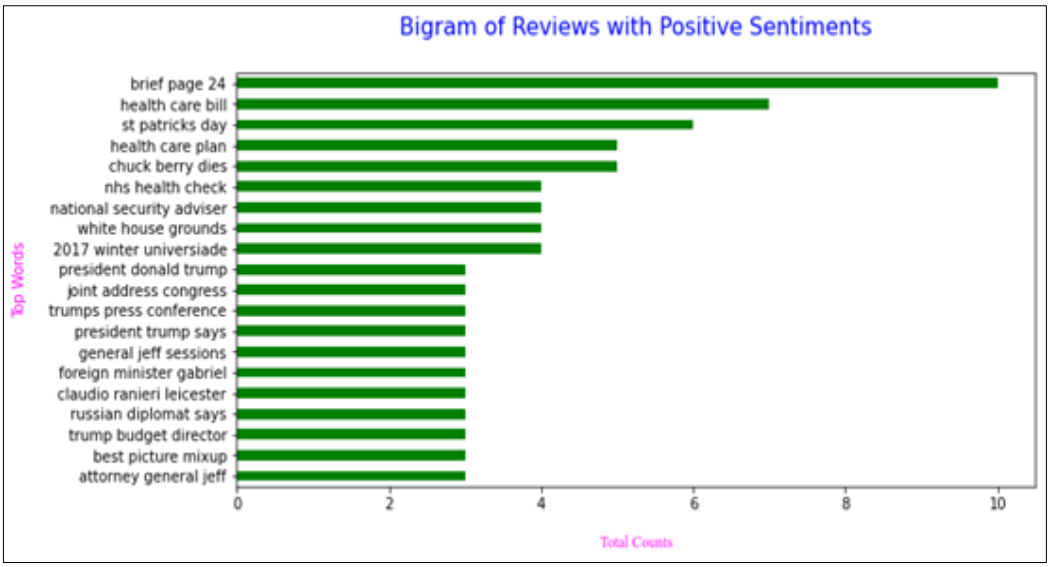

Figure IV. 8 : Statistiques ensemble des mots positifs les plus répétés

Les figures ci-dessous représente les statistiques des mots et l'ensemble des mots négatifs les plus utilisés dans notre dataset après l'application de notre programme de l'analyse des sentiments sur le Data-set.

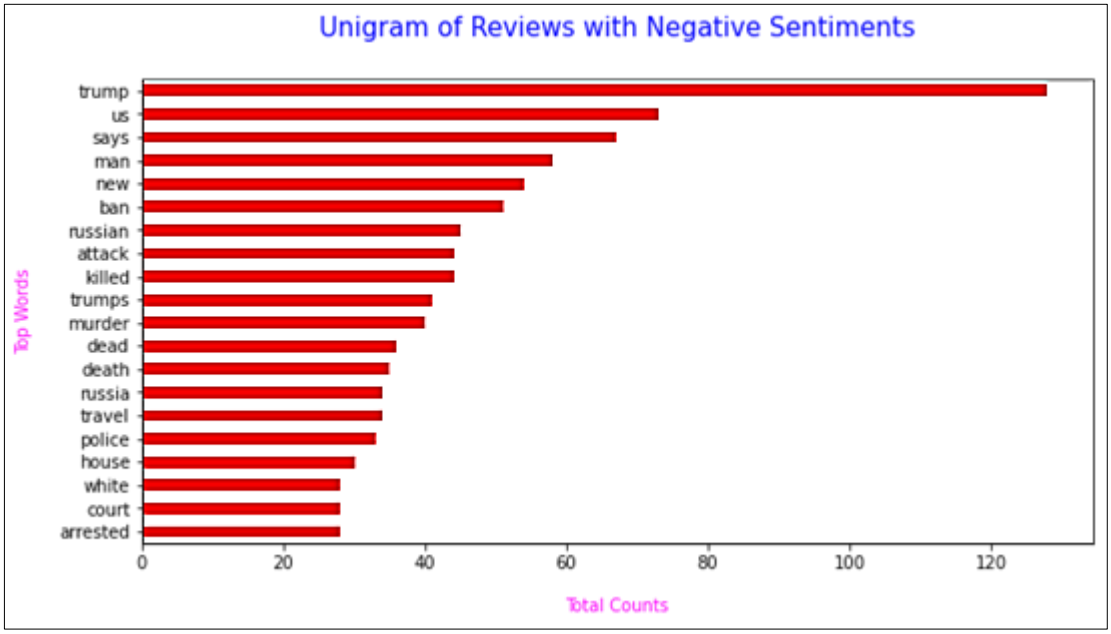

Figure IV. 9 : Statistiques de chaque mot négatif le plus répété

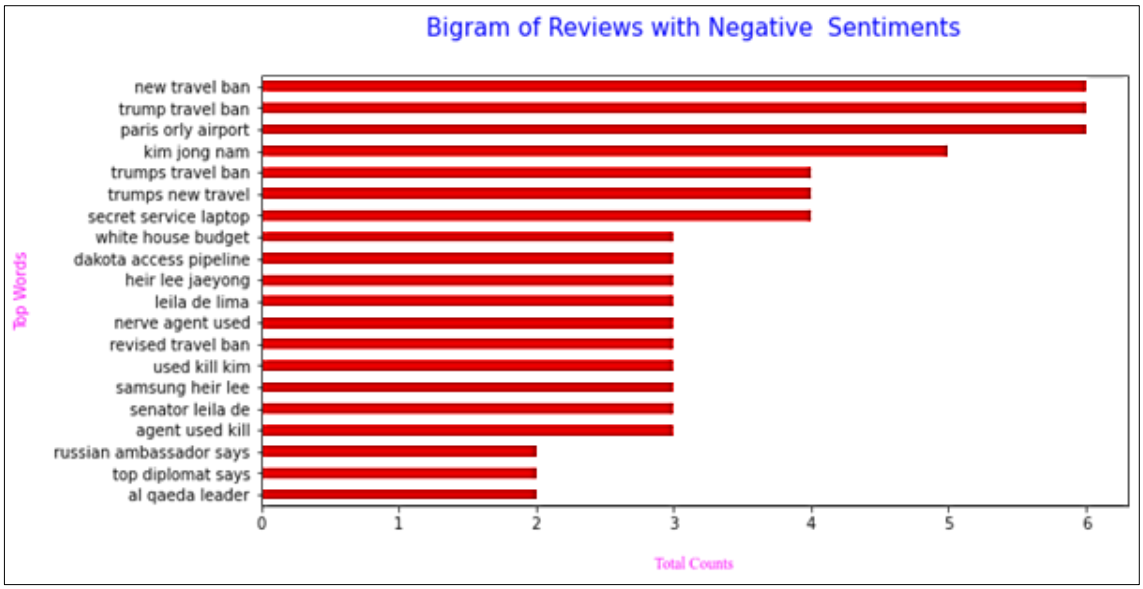

Figure IV. 10 : statistiques l'ensemble des mots négatifs les plus répétés.

IV.7.3.1 Sentiment de requête :

```
Entrée [22]:
# query side

query = "love education"
query_tok = simple_tok(query)
query_en= ""


from nltk.sentiment.vader import SentimentIntensityAnalyzer
SIA = SentimentIntensityAnalyzer()

print(SIA.polarity_scores(query))

if(SIA.polarity_scores(query)['compound']>=0):
    query_en = "pos"
else :
    query_en = "neg"

print(query_en)

{'neg': 0.0, 'neu': 0.192, 'pos': 0.808, 'compound': 0.6369}
```

Figure IV. 11 : Code source du sentiment de Requête

IV.7.3.2 Appariement document requête :

```
Entrée [23]:
scores = bm25.get_scores(query_tok)

article_source_link = []
title = []
text = []
score_thématique = []
score_émotionnelle = []
comp_score = []

best_docs = sorted(range(len(scores)), key=lambda i: scores[i], reverse=True)[:30]
for i, b in enumerate(best_docs):
    if (query_en == df['comp_score'][i]):
        article_source_link.append(df['article_source_link'][i])
        title.append(df['title'][i])
        text.append(df['text'][i])
        score_thématique.append(scores[i])
        score_émotionnelle.append(df['compound'][i])
        comp_score.append(df['comp_score'][i])

dfObj = pd.DataFrame({
    'article_source_link': article_source_link,
    'title': title ,
    'text': text,
    'score émotionnelle': score_émotionnelle,
    'comp_score': comp_score
})
dfObj.sort_values(by=['score émotionnelle'],ascending=False , ignore_index = True)
```

Out[23]:

	article_source_link	title	text	score émotionnelle	comp_score
0	http://www.cnn.com/2017/02/07/politics/trumpj...	Trump teeing up a softer diplomatic approach w...	After a combative round of telephone diplomacy...	0.9965	pos
1	http://abcnews.go.com/Sports/tom-brady-missing...	Texas Rangers to Help Search for Tom Bradys St...	The Texas Rangers are joining the search for t...	0.9963	pos

Figure IV. 12 : Code source de l'appariement document-requête

IV.7.3.3 Distribution des classes du datas et complet

La figure 23 ci-dessous représente des statistiques sur le pourcentage des mots négatifs et positifs existants après l'analyse de sentiments de tout le data-set.

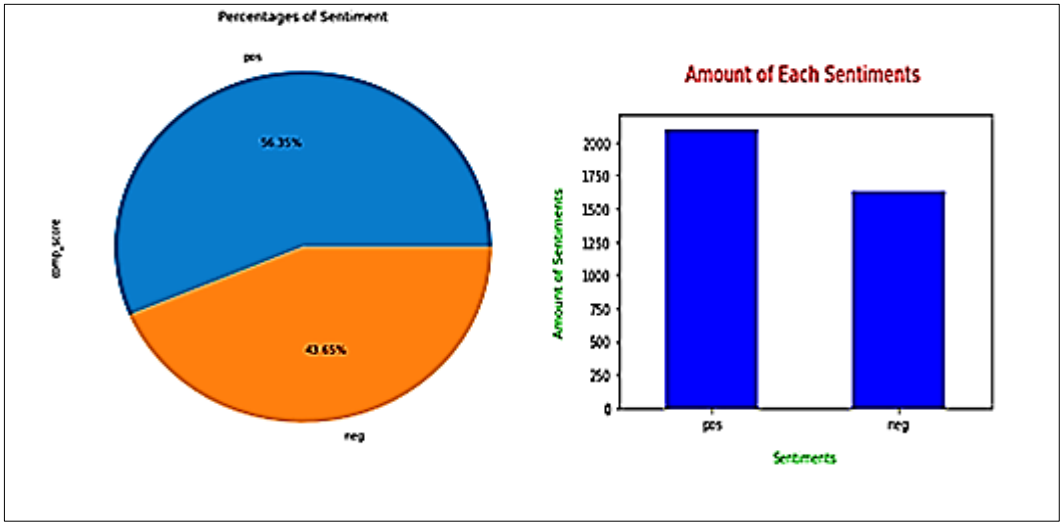

Figure IV. 13 : Distributions des classes de notre dataset complet

IV.8. Présentation des pages d'interaction de notre service de recherche :

Page de recherche :

La page de recherche permet à l'utilisateur de formuler la requête, en lui présentant une interface graphique conviviale, qui permet la saisie des critères de recherche.

Pour lancer la recherche il suffit de saisir les critères de recherche et de valider le bouton libellé par "Search by name or quote"(Figure 35). Une fois la saisie effectuée et validée par l'utilisateur, le navigateur web extrait les informations demandées.

Une réponse est constituée de documents classés d'après l'ordre des scores des valeurs dans l'index. Les résultats sont affichés par ordre décroissant de scores Emo-Thematique (Figure 36).

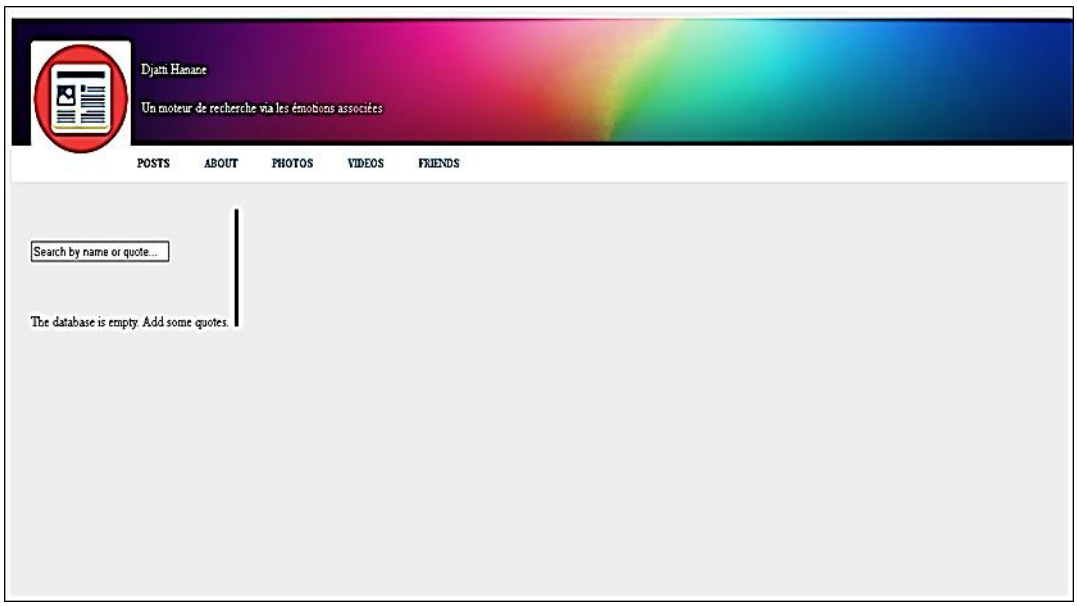

Figure IV. 14 : Interface principale du moteur de recherche émotionnel

Cette page est constituée de boutons menus : post, vidéos, photos et about.

About : affiche des informations sur notre service de recherche.

Photos : Affiche que des photos comme résultats de recherche.

Vidéos : affiche que des vidéos comme résultats de recherche.

Chaque page contient deux parties.

Plage de recherche : ou l'internaute formule sa requête.

Espace d'affichage : affiche les résultats de recherche.

Page des résultats de recherche

Figure IV. 15 : Interface principale du moteur de recherche émotionnel

Page des résultats de recherche

La figure 38 ci-dessous présente les statistiques des résultats de recherche de notre moteur de recherche lors d'une recherche individuelle (un seul mot) et groupée de mots. On remarque que le pourcentage de recherche via un seul mot est inférieur à celui d'une recherche d'une combinaison de mots (phrase) qui contient le même mot (figure 38).

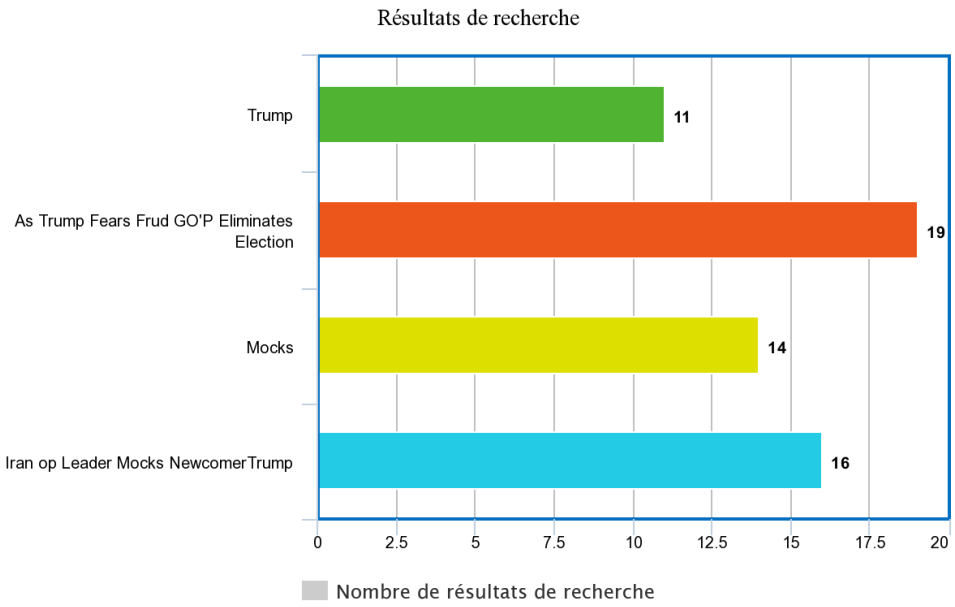

Figure IV. 16 : *S*tatistiques des résultats de recherche individuelle et groupée de mots

Page de recherche du méta-service standard :

La figure 16 présente l'interface principale du méta-service de recherche standard ou apparait le libellé de requête et le bouton « Search » pour lancer la recherche

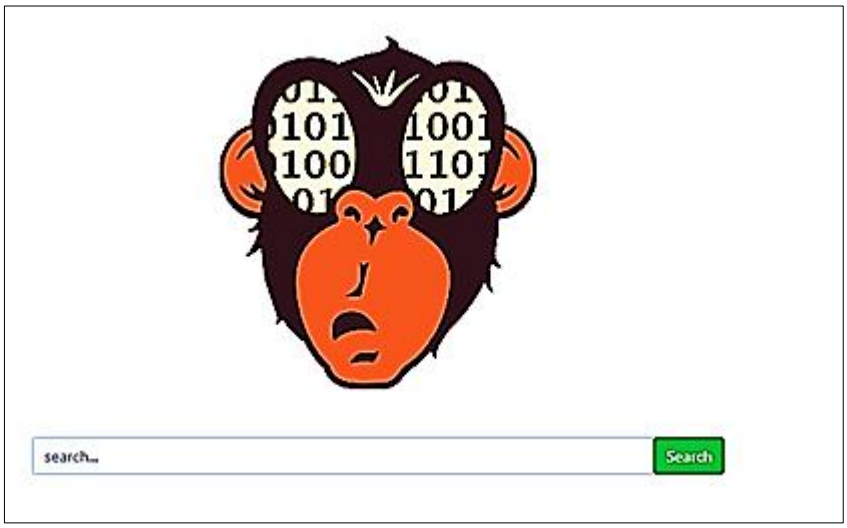

Figure IV. 17 : l'interface principale du méta-service de recherche standard

Cette figure présente la recherche via la soumission assistée par l'outil de recherche qui nous assiste à formuler la bonne requête en afficute une liste de propositions.

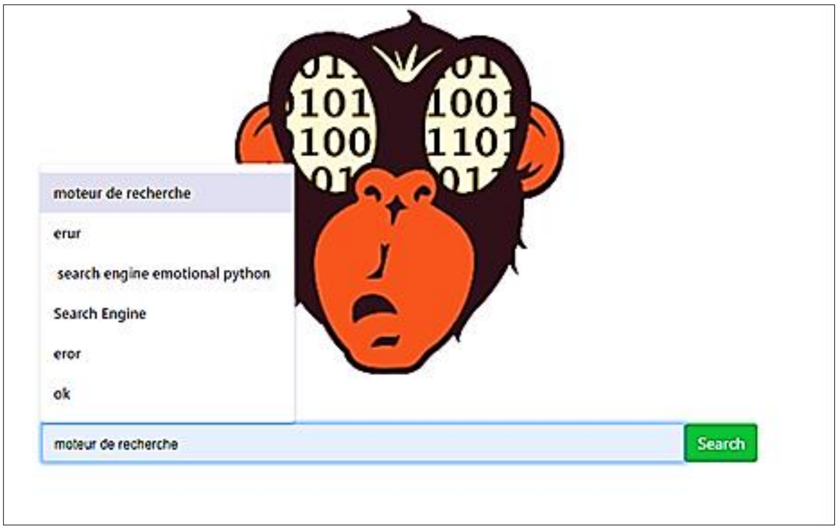

Figure IV. 18 : La phase de recherche

Dans cette figure, Les résultats de la recherche classique sont sous forme de liens de questions déjà posés sur Stackoverflow

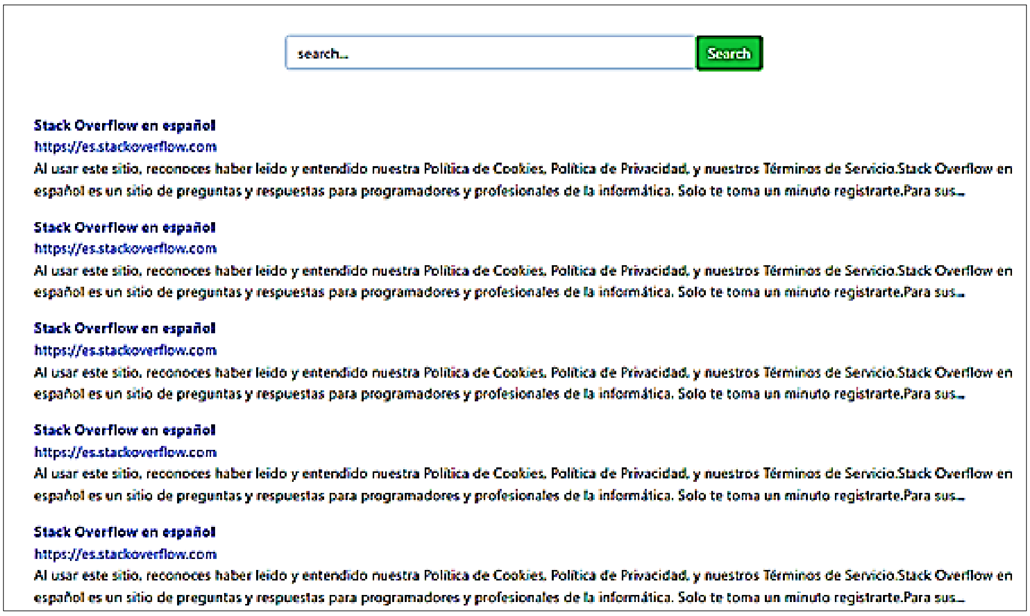

Figure IV. 19 : Résultats du méta-service de recherche

IV.9. Conclusion :

Les outils de recherche actuels ne parviennent pas à absorber la formidable croissance des sites sur le web puisque l'ensemble des moteurs ne référence que la moitié des documents estimés sur Internet à cause de grandes quantités d'informations trouvés. Notre service de recherche orienté émotions essaie au maximum de faire retourner des résultats très proches des résultats manuels et de la pensée humaine.

Finalement, notre service de recherche émotionnel est équipé de meilleures techniques de développement tel que la détection et analyse de sentiment et la pertinence des résultats de recherche.

.

Conclusion Générale

Conclusion générale

La RI est un domaine qui cherche à étudier les différents processus afin de répondre à des besoins informationnels. Dans le cadre d'une recherche textuelle, ces besoins peuvent se traduire comme étant l'ensemble des documents d'une collection (noté : Di) qui répondent le plus au besoin exprimé par un utilisateur sous forme d'une requête (noté: Qi).

Un SRI est composé essentiellement de deux processus fondamentaux : L'indexation des requêtes et des documents et L'appariement de la requête avec les documents de la collection. Et Les logiciels de recherches sont des systèmes qui implémentent les deux processus définis auparavant afin d'assurer l'ensemble des fonctions nécessaires à la recherche d'information.

Le travail présenté dans ce PFE s'intéresse à la recherche d'information et plus particulièrement à l'influence du sentiment sur les résultats de recherche. Le cadre de cette étude se focalise donc sur l'intégration de l'unité informationnel textuel transportant un certain sentiment/émotion au sein du processus de RI. Car elle influence les résultats produits par les modèles de recherche. Dans ce PFE, un effort considérable à été nécessaire afin d'implémenter l'unité d'information BM25 et de voir l'impact résultant d'un tel choix sur les résultats de recherche.

Cependant, comme perspectives, nous proposons d'améliorer le processus de recherche aux niveaux de l'exploitation du dataset et de la requête, de l'indexation et de correspondance au sein du modèle recherche. Il est ainsi intéressant de poursuivre l'exploration du modèle probabiliste en utilisant d'autres variantes qui augmentent la représentativité des ensembles des termes similaires.

Comme, Il est aussi fort intéressant de poursuivre ces recherches avec des collections de documents plus grandes.

REFERENCES

BIBLIOGRAPHIQUES

Références bibliographiques

[1] Hernandez, N. Ontologie de domaine pour la modélisation du contexte en recherche d'information. thèse de doctorat en informatique. s.l. : Université Paul Sabatier, 2006.

[2] Boubekeur, F. Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets. thèse de doctorat en informatique. s.l. : Université Paul, 2008.

[3]. Daoud, M. Accès personnalisé à l'information : approche basée sur l'utilisation d'unprofil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique. thèse de doctorat en informatique. s.l. : Université Paul Sabatier, 2009.

[4] Bouramoul, Abdelkrim. RECHERCHE D'INFORMATION. these de doctorat. Constantine : Université MENTOURI , 2011.

[5.] Ingwersen, Peter. Information retrieval interaction. London : Taylor graham Publishing, 1992.

[6.] Boughanem, M et Savoy, J. Recherche d'information états des lieux et perspectives. s.l. Hermès Sience Publications, 2008.

[7]. Tamine-Lechani, Lynda et Calabretto, Sylvie. Recherche d'information contextuelle et web.

[8]. Damak, Firas. Etude des facteurs de pertinence dans la recherche de microblogs. these de doctorat. Toulouse : s.n., 2014.

[9]. Buckley, C et Salton, G. Term weighting approches in automatic text retrieval. 1988.

[10]. Zipf, George K. Hman Behaviour and the principal of least effort. USA : s.n., 1949.

[11] Maron, M. E et Kuhns, J. L. On relevance, probabilistic indexing and information retrieval. 1960.

[12]. Porter, M. An algorithm for suffix stripping . 1980.

[13] Mayfield, J et McNamee, P. Single n-gram stemming. In proccedings of the 26th annual internationnal ACM SiGIR Conference on research and development in information retrieval. New York : s.n., 2003.

[14] Manning, C, Raghavan, P et Schtze, H. Introduction to information retrieval. New York : Cambridge university press, 2008.

[15] Baeza- Yates, R. A et Ribeiro-Neto, B. A. Modern information retrieval. England : Pearson education Ltd, 2011.

[16] Hammache, Arezki. Recherche d'information: un modèle de langue combinant mot simples et mots composés. these de doctorat. Tizi-Ouzou : s.n.

[17] Salton, G. The SMART Retrieval System - Experiments in Automatic Document Processing. New Jersey : s.n., 1971.

[18] Van Rijsbergen, C.J. Information retrieval. London : Butterworth, 1979.

[19] Abbassi, Meftah. Un modèle de reformulation des requêtes pour la recherche d'information sur le Web. these de master.

[20] Hachemi, Hadjira et Rimouche, Nour El Houda. Moteur de recherche sémantique. memoire de licence. 2013.

[21] Robertson, S.E et Sparch, Jones. Relevance Weighting for Search Terms. s.l. : Journal of The American Society for Information, 1976.

[22] Robertson., S. The probability ranking principle in information retrieval. 1977.

[23] Zemirli, Nesrine. Vers le développement d'un système de recherche d'information personnalisé integrant profile d'utilisateur. thèse de doctorat. Université Paul Sabatier : s.n., 2004.

[24] Bruce Croft, W, Turtle, Howard R et Lewis, David D. The Use of Phrases and Structured Queries in Information Retrieval. 1991.

[25] Rocchio, J. Relevance feedback information retrieval. . 1971.

[26] Harman, D. Relevance feedback revisited. Proceedings of ACM SIGIR : s.n., 1992.

[27] Boughanem, M, Chrisment, C et Soule-dupuy, C. Query modification based on Relevance Back-propagation in ad-hoc environment. s.l. : IPM: Information Process and Management, 1998.

[28] https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203539-moteur-de-recherche-definition-traduction-et-acteurs/

[29] SIMONNOT, Brigitte. Moteurs de recherche. Usages et Enjeux. Questions de communication, 2008, no 14, p. 7-15.

[30]: Abdelkrim, B. O. U. R. A. M. O. U. L. (2017). Recherche d'Information Contextuelle et Sémantique sur le web.

[31] : https://www.netoffensive.blog/referencement-naturel/premier-sur-google/fonctionnement/

[32] https://www.livepepper.fr/academy/internet/referencement/moteurs-de-recherche

[33] https://ecommerce-platforms.com/fr/glossary/bing

[34] https://www.louismaitreau.fr/definition/bing/

[35] https://alejandrorioja.com/fr/moteurs-de-recherche-haut-de-gamme/

[36] https://www.redacteur.com/blog/seo-7-differences-entre-moteurs-recherche/

[37] https://www.blogdumoderateur.com/tools/yahoo-search/

[38] https://www.keacrea.com/moteurs-de-recherche-alternatifs-a-google

[39] https://www.laprovence.com/actu/en-direct/6011434/le-moteur-de-recherche-duckduckgo-defie-le-geant-google.html

[40] https://fr.wikipedia.org/wiki/DuckDuckGo [Accès le 01 10 2021].

[41]. COSNIER, J. (1994). Psychologie des émotions et des sentiments. Retz.

[41].https://textblob.readthedocs.io/en/dev/

[42].https://data-flair.training/blogs/text-mining/

[43] https://arxiv.org/pdf/1707.02919.pdf

[44].https://medium.com/@mehdihadji/analyse-des-sentimentsg%C3%A9n%C3%A9ralit%C3%A9s-99ab87503a5e

[45].http://dspace.univkm.dz/xmlui/bitstream/handle/123456789/2625/Une%20approche%20Deep%20Learning%20pour%20l%E2%80%99analyse%20des%20Sentiments%20Sur%20Twitter.pdf?sequence=1&isAllowed=y

[46] https://towardsdatascience.com/nlp-for-beginners-cleaning-preprocessing-text-data-ae8e306bef0f

[47] https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908

[48] Anaconda (Python distribution),» 12 05 2021. [En ligne]. Available:

https://fr.wikipedia.org/wiki/Anaconda_(Python_distribution). [Accès le 19 09 2021].

[49] https://www.lptmc.jussieu.fr/user/barbi/ENSEIGNEMENT/M2/Python1617/TD01-Exercices/TD01-Introduction_%D0%95_Python.html

[50] « Visual studio code,» 14 09 2021. [En ligne]. Available:

https://edutechwiki.unige.ch/fr/Visual_studio_code#Installation. [Accès le 19 09 2021].

[51] « XAMPP » 07 05 2021. [En ligne]. Available: https://fr.wikipedia.org/wiki/XAMPP. [Accès le 19 09 2021]

[52] « LeBigData» [En ligne]. Available: https://www.lebigdata.fr/python-langage-definition. [Accès le 20 09 2021]

[53] « Natural Language Toolkit,» 28 03 2020. [En ligne]. Available:

https://fr.wikipedia.org/wiki/Natural_Language_Toolkit. [Accès le 21 09 2021].

[54] « Pandas, » 20 06 2021. [En ligne]. Available: https://fr.wikipedia.org/wiki/Pandas. [Accès le 21 09 2021]

[55] https://fr.wikipedia.org/wiki/Okapi_BM25 [Accès le 21 09 2021]

[56]https://fr.wikipedia.org/wiki/Matplotlib#:~:text=Matplotlib%20est%20une%20biblioth%C3%A8que%20du,une%20licence%20de%20style%20BSD[Accès le 21 09 2021].

[57] https://ichi.pro/fr/simplifier-l-analyse-des-sentiments-a-l-aide-de-vader-en-python-sur-le-texte-des-medias-sociaux-274770204542255

Résumé :

Notre travail se situe au carrefour du moteur de recherche et analyse de sentiment qui est la Recherche d'Information émotionnelle. Nous nous intéressons, plus précisément, à la recherche de sentiment textuels pertinents dans des documents. Pour le faire, nous commençons d'abord par une approche de recherche classique bonifiée à la suite par une recherche émotionnelle. Cette dernière repose sur deux mesures Analyse de la requête et Appariement document-requête. Ensuite, nous intégrons, d'une manière séparée et combinée, toutes les interactions des utilisateurs textuelles laissées sur leurs contenus textuels au sein du processus de recherche en tant que source d'information additionnelle pour améliorer la pertinence des résultats.

Notre modèle de moteur de recherche se base sur une représentation probabiliste des données. Il effectue un appariement entre la requête saisie et les documents déjà préparés pour retourner la recherche de ces derniers selon leur pertinence globale.

Nos expériences menées ont montré une augmentation du taux de satisfaction meilleur (Pertinence) par rapport aux systèmes de moteur de rechercher classiques

Abstract:

Our work sits at the crossroads of search engine and sentiment analysis which is Emotional Information Retrieval. We are interested, more precisely, in the search for relevant textual sentiment in documents. To do this, we start with a classical research approach first, followed by emotional research. The latter is based on two measures Analysis of the request and matching document-request. Then, we integrate, in a separate and combined way, all textual user interactions left on their textual contents within the search process as an additional source of information to improve the relevance of the results.

Our search engine model is based on a probabilistic representation of data. It performs a match between the entered query and the documents already prepared to return the search for the latter according to their overall relevance.

Our experiments have shown an increase in the best satisfaction rate (Relevance) compared to conventional search engine systems