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 Logiciel

Par :

LARACHI ABDELHADI
DAMINE MOHAMED SAAD

Sur le thème

Impact des interactions sociales sur la recherche d'information

Soutenu publiquement le. 23/ 09 / 2021 à Tiaret devant le jury composé de :

Mr. DAHMANI Youcef	Grade	Professeur	U.I.K. Tiaret	Président
Mme. LAKHDARI Aicha	Grade	M.A.A	U.I.K. Tiaret	Promoteur
Mr. MOKHTARI Ahmed	Grade	M.A.A	U.I.K. Tiaret	Examinateur

2020-2021

DÉDICACE

Du plus profond de nos cœurs, nous dédions ce travail :

A nos chers parents, pour tous leurs sacrifices, leurs amours, leurs tendresses, leurs soutiens et leurs prières tout au long de nos études.

A nos chers frères et sœurs pour leurs appuis et leurs encouragements permanents.

A nos familles **LARACHI** et **DAMINE** pour leurs soutiens tout au long de nos parcours universitaires.

À tous nos amis et aux personnes qui ont toujours cru en nous.

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fuit de votre soutien infaillible.

REMERCIEMENTS

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

Nos vifs remerciements et gratitudes s'adressent à notre encadreur Mme : LAKHDARI AICHA pour avoir accepté de diriger ce travail. Son soutien, ses compétences, sa gentillesse et sa disponibilité qu'elle nous a témoignée pour nous permettre de mener à bien ce travail.

Nous tenons à remercier sincèrement les membres du Jury qui nous font le grand honneur d'évaluer ce travail :

A Mr DAHMANI YOUCEF qui a accepté de présider le jury de soutenance et pour l'intérêt qu'il est porté à notre recherche en acceptant de scruter notre travail et de l'enrichir par ses propositions.

A Mr MOKHTARI AHMED pour nous avoir fait l'honneur d'accepter d'examiner ce travail.

Nos remerciements les plus chaleureux vont à tous nos camarades au Master 2 Génie logiciel de la Faculté mathématique et informatique de Tiaret, ainsi que tous nos autres camarades de cette Université pour leur présence dans les moments difficiles et les excellents moments que nous avons passés avec eux tout au long de cette année.

Enfin, nous tenons également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Merci à tous et à toutes.

ABSTRACT

The field of information retrieval (IR) is considered as a successful application area of natural language processing (NLP). Nowadays, the emergence of social networks has led to the appearance of a new branch of information retrieval, it is social information retrieval which benefits from the exploitation of information from user-user and user-resource interactions on the social Web to improve search results.

Our work is therefore part of the IR problematic where we define an approach based on three essential relevance measures to estimate the global relevance of search results: (a) thematic relevance: translates the degree of adequacy of the retrieved information to the theme evoked by the keywords of the query, (b) social relevance: computed according to a formula exploiting social signals (c) emotional relevance: computed according to a process of sentiment analysis.

Our experiments show a significant improvement in the performance of the classical movie search system by integrating social and emotional criteria from social networks. This exploitation is performed according to measures defined in the IR domain and using a dataset of movies defined and built from the IMDb website.

Key words: information retrieval, natural language processing, social information retrieval, relevance, social signals, sentiment analysis process, information retrieval system, dataset.

RESUME

Le domaine de la recherche d'information (RI) est considéré comme un domaine d'application réussi du traitement de langage naturel (TLN). De nos jours, l'émergence des réseaux sociaux a conduit à l'apparition d'une nouvelle branche de recherche d'information, c'est la recherche d'information sociale qui bénéfice de l'exploitation des informations issus des interactions user-user et user-ressources Web sur les réseaux sociaux pour améliorer les résultats de recherche.

Notre travail rentre donc dans la problématique de la RI où nous définissons une approche basée sur trois mesures de pertinences essentielles pour estimer la pertinence globale des résultats de recherche : (a) pertinence thématique : traduite le degré d'adéquation de l'information retrouvée au thème évoqué par les mots clefs de la requête, (b) pertinence sociale : calculée selon une formule exploitant les signaux sociaux (c) pertinence émotionnelle : calculée selon un processus d'analyse des sentiments.

Nos expériences montrent une amélioration significative de la performance du système de la recherche classique de films en intégrant des critères sociaux et émotionnels provenant des réseaux sociaux. Cette exploitation est réalisée selon des mesures définies dans le domaine de la RI et en utilisant un dataset de films définis et construit à partir du site IMDb.

Mots-clés : recherche d'information, traitement de langage naturel, recherche d'information sociale, pertinence, signaux sociaux, processus d'analyse des sentiments, système de recherche d'information, dataset.

Table des matières

Table de figures

Liste de tableau

LISTE DES ABREVIATIONS

API	**A**pplication **P**rogramming **I**nterface
BM25	**B**est **M**atching 25
CSV	**C**omma **S**eparated **V**alues
HTML	**H**ypertext **M**arkup **L**anguage
HTTP	**H**yper**t**ext **T**ransfer **P**rotocol
IDF	**I**nverse **D**ocument **F**requency
IMDB	**I**nternet **M**ovies **D**ata**B**ase
JSON	**J**ava**S**cript **O**bjet **N**otation
LSI	**L**atent **S**emantic **I**ndexing
LTR	**L**earning **T**o **R**ank
NLP	**N**atural **L**anguage **P**rocessing
NLTK	**N**atural **L**anguage **T**ool**k**it
RI	**R**echerche d'**I**nformation
RIS	**R**echerche d'**I**nformation **S**ociale
SRI	**S**ystème de **R**echerche d'**I**nformation
TF	**T**erm **F**requency
UGC	**U**ser **G**enerated **C**ontent
URL	**U**niform **R**esource **L**ocator
VSM	**V**ector **S**pace **M**odel
XML	E**X**tensible **M**arkup **L**anguage

1. Introduction Générale :

1.1. Contexte :

La recherche d'informations (RI) est généralement une tâche individuelle qui vise à répondre à un besoin d'information exprimé généralement par une requête simple ou complexe.

En effet, en raison de la grande quantité de documents disponibles, le Web est devenu une source d'information privilégiée pour répondre aux exigences quotidiennes des internautes. Pour cela, la RI sur le Web peut se faire selon deux modalités spécifiques : l'interrogation et la navigation [1]. L'interrogation consiste à utiliser un outil spécifique tel qu'un moteur de recherche afin de trouver des informations pertinentes. La navigation consiste à parcourir le Web grâce aux hyperliens qui existent dans les pages Web.

Cependant, la quantité d'informations et sa croissance sont une arme à double tranchant en raison du problème de la surcharge d'informations, le fait que tous les contenus sur le Web ne sont pas pertinents ou de qualité acceptable pour les chercheurs d'informations. Diverses techniques ont été adoptées sur le Web pour résoudre ces problèmes et aider les utilisateurs à satisfaire leurs besoins d'information. De nombreuses études ont montré que les interactions entre utilisateurs et utilisateurs-ressources Web jouent un rôle important dans le processus de recherche et d'utilisation des informations. Il n'est pas rare qu'en recherchant des informations, il faut faire appel à des amis et collègues des réseaux sociaux pour satisfaire ce besoin en information. Ce qui conduit à l'existence d'une nouvelle branche de la RI qui est la recherche d'information sociale (RIS).

Cette dernière désigne une famille de techniques qui aident les utilisateurs à répondre à leurs besoins en informations en exploitant l'intelligence collective d'autres utilisateurs, leurs connaissances spécialisées ou leurs expériences dans la recherche. La RIS est un domaine émergent et une voie prometteuse pour la conception et la mise en œuvre d'une nouvelle génération de systèmes de recherche d'informations (SRI).

1.2. Motivation :

La motivation derrière l'exploitation des contenus sociaux est d'essayer d'améliorer la RI classique par rapport à un besoin en information en exploitant l'information sociale émotionnelle afin de prouver leurs impacts sur la performance des SRI. Le principe de 'Wisdom of Craws' [2] est l'idée que "de grands groupes de personnes sont collectivement plus intelligents que les experts individuels lorsqu'il s'agit de résoudre des problèmes, de prendre des décisions, d'innover et de prévoir". Ce concept se réfère à l'intelligence collective des internautes qui commentent, tagguent et notent ainsi des ressources web via des Wikis, des blogs et des réseaux sociaux.

1.3. Défis et enjeux :

Pour une RI pertinente, l'indexation et la classification sont les facteurs clés. Pour mieux servir les utilisateurs, il doit y avoir des critères et le critère le plus connu est la précision. Mais ce critère n'est pas totalement accepté et respecté dans la RI.

Cependant, Il existe actuellement un certain nombre de travaux de recherche visant à mieux comprendre les problématiques de la RIS, en intégrant dans le processus de la recherche de nouvelles informations provenant des réseaux sociaux. Ces problématiques s'accentuent sur plusieurs facteurs :

- **Volume :** L'émergence des réseaux sociaux a conduit à l'explosion du volume de données et la variété des documents générées par les utilisateurs. Ces données sociales peuvent améliorer l'efficacité des systèmes de recherche d'information. Le problème ici concerne le stockage, l'accès et l'analyse d'énorme quantités d'informations sociales.
- **Structure des réseaux sociaux:** Le monde des réseaux sociaux est en perpétuel changement, Il évolue au fur et à mesure que les gens trouvent de nouveaux moyens de se connecter et de communiquer avec les autres et que la technologie se développe. Chaque réseau social fournit une structure propre à son réseau pour le distinguer de ses concurrents. Cette diversité apporte des difficultés lors de l'extractions des informations dont nous avons besoin.
- **Acteurs sociaux :** L'évaluation des acteurs sociaux comprend l'identification des utilisateurs influenceurs dans les réseaux sociaux, la pertinence sociale d'un acteur dépend de sa popularité sur les réseaux sociaux.

1.4. Questions de recherche :

Les questions de recherches auxquelles nous devront répondre portent sur la définition de la pertinence Socio-Emo-Thématique via une exploitation des contenus sociaux de Twitter :

- Quelle est l'impact des interactions sociales (signaux sociaux, les émotions) sur les performances des SRI ?
- Comment combiner la pertinence sociale, émotionnelle et thématique pour améliorer le processus de la RI ?
- Est-ce que la pertinence Socio-Emo-Thématique est un paramètre essentiel pour évaluer la RI ?

1.5. Objectifs :

L'objectif de notre travail est de proposer une approche basée sur un modèle de recherche probabiliste visant à améliorer les résultats de la recherche d'informations classique en exploitant des critères sociaux-émotionnels tel que : les signaux sociaux et les émotions.

Nous prenons en compte les critères sociaux associés aux ressources Web comme une information additionnelle afin de mesurer l'importance d'une ressource indépendamment de la requête.

Nos expérimentations sont menées sur un Dataset crée et construit à partir du site Web des films IMDb[1] (Internet Movies Data base) en s'appuyant sur une technique d'extraction et de collecte d'informations concernant ces films. L'enrichissement de ce dataset, par la suite, par des signaux sociaux collectés de Twitter va accélérer l'amélioration des résultats de recherches. Cette collection de documents extraite contient 45000 documents de films.

1.6. Organisation du mémoire

Afin de parvenir à notre but, le mémoire est organisé en deux parties principales : Un volet théorique ainsi qu'un volet pratique.

Le volet théorique :

Le Chapitre 1 présente les concepts généraux de la recherche d'information, ainsi que les questions relatives à l'indexation et les modèles de RI.

[1] WWW.IMDb.com

Le Chapitre 2 présente la recherche d'information sociale où nous décrivons l'information sociale dans le Web. Ensuite, la notion de la RI sociale est définie. Et enfin, Nous étudions et par la suite nous synthétisons et discutons les travaux relatifs à l'impact des contenus sociaux sur la recherche d'information.

Le chapitre 3 explore l'essence du processus d'analyse des sentiments qui consiste à extraire et à catégoriser les opinions d'un certain document film.

Le volet pratique :

Le chapitre 4 est consacré à la conception de notre outil de Ranking et Ré-Ranking en exploitant les contenus sociaux du big-social data Twitter et en proposant une amélioration des résultats de recherche via une métrique globale qui est une combinaison linéaire de scores thématique, social et émotionnel.

Le Chapitre 5 présente les outils de développement et les expérimentations réalisés afin de réaliser un ordonnancement pertinent de documents (films) enrichis de signaux sociaux et d'émotions. Nous décrivons notre méthodologie d'expérimentation et les résultats d'évaluations obtenus concernant le score global qui confirme l'impact des interactions sociales sur le processus de la RI classique.

Volet théorique

CHAPITRE 01 :

Recherche d'information classique (RI)

Chapitre 01

2. Recherche d'information classique :

2.1. Introduction :

Pendant que des volumes plus importants de documents comprenant des éléments multimédias sont devenus disponibles sur internet, les utilisateurs ont eu besoin d'outils plus sophistiqués pour la localisation et la recherche d'informations. C'est pourquoi un certain nombre de technologies sont déployées dans diverses applications de gestion de l'information : moteurs de recherche multilingues, traduction automatique, système d'accès vidéo, technologies linguistiques basées sur le contenu pour la gestion de l'information, résumé de documents., traitement de texte, etc...

2.2. Historique :

L'histoire de la recherche d'informations ne commence pas avec l'internet, les moteurs de recherche sur le web sont devenus très répandus et que la recherche a été intégrée dans des systèmes d'exploitation. Les systèmes RI ont été utilisés dans des applications commerciales et de renseignement dès les années 1960 [3]. Les premiers systèmes de RI ont été construits à la fin des années 1940 et se sont inspirés des innovations de la première moitié du 20e siècle. Comme pour de nombreuses technologies informatiques, les capacités des systèmes de recherche se sont accrues avec l'augmentation de la vitesse des processeurs et de la capacité de stockage. Un système de RI localise les informations pertinentes en fonction de la requête d'un utilisateur. Son évolution à travers le temps est retracée comme suit :

- 1940 : des prototypes des dispositifs mécaniques et électromécaniques ont été construits, et qui ont effectué des recherches dans des catalogues générés manuellement, par exemple ``Shaw's rapid selector `` conçus pour chercher dans une bobine de film de 2000 pieds. Il est indiqué que 72000 images étaient stockées sur un film. Selon Shaw, le sélecteur était capable d'effectuer des recherches au rythme de 78000 entrées par minute [4].

- 1950 : d'autres technologies mécaniques ont été examinées. Luhn, par exemple, a fabriqué un sélecteur utilisant des cartes perforées, de la lumière et des cellules photoélectriques,

l'une des principales caractéristiques de ce système était possible de faire correspondre une séquence consécutive de caractères dans une chaîne plus large [5], ce système recherche à la vitesse de 600 cartes par minute. C'est à cette époque que le terme RI a été utilisé pour la première fois.

- 1960 : Les années 1960 ont vu un large éventail d'activités reflétant le passage de la simple question de savoir si la RI était possible sur les ordinateurs à la détermination des moyens d'améliorer les systèmes de RI. Gerard Salton, qui a formé et dirigé un groupe de RI, a produit de nombreux rapports techniques établissant des idées et des concepts utilisés jusqu'à présent. L'un de ces concepts est la formalisation d'algorithmes pour classer les documents par rapport à une requête. Il faut noter en particulier une approche où les documents et les requêtes sont considérés comme des vecteurs dans un espace de dimension N (N étant le nombre de termes uniques dans la collection recherchée) [6]. Salton a suggéré que la similarité entre un document et un vecteur soit mesurée en utilisant le calcul du cosinus [7].

- 1970 : L'un des développements clés de cette période a été que les pondérations de fréquence des termes (TF) de Luhn ont été complétées par les travaux de Spärck Jones sur l'occurrence des mots à travers les documents d'une collection. Un autre moyen de modéliser les systèmes de RI consistait à étendre l'idée de Maron, Kuhns et Ray d'utiliser la théorie des probabilités. Robertson a défini le principe du classement probabiliste [8], qui détermine comment classer de manière optimale les documents sur la base de mesures probabilistes par rapport à des mesures d'évaluation définies.

- 1980 – milieu des années 1990 : sur la base des développements des années 1970, des variations des schémas de pondération TF IDF ont été produites et les modèles formels de recherche ont été étendus. Des avancées sur le modèle vectoriel ont également été développées et la plus connue est probablement l'indexation sémantique latente (LSI), où la dimensionnalité de l'espace vectoriel d'une collection de documents a été réduite par la décomposition des valeurs singulières [9]. Les requêtes ont été mises en correspondance dans l'espace réduit.

- Milieu des années 1990 – présent : L'arrivée du web a initié l'étude de nouveaux problèmes de RI. Les développeurs de moteurs de recherche ont rapidement compris qu'ils pouvaient utiliser les liens entre les pages web pour construire un robot qui traverse et rassemble la plupart des pages web sur l'internet, automatisant ainsi l'acquisition de contenu, les applications de la recherche et le domaine de la récupération d'informations continuent d'évoluer à mesure que l'environnement informatique change, la communauté de la RI a développé la recherche sociale, qui traite de la recherche impliquant des communautés d'utilisateurs et l'échange informel d'informations [3].

2.3. Définition :

La recherche d'informations désigne le processus, les méthodes et les procédures de recherche, de localisation et de récupération de données et d'informations enregistrées dans un fichier ou une base de données. Dans les bibliothèques et les archives, la recherche moderne d'informations se fait par la consultation de bases de données en texte intégral, la localisation d'articles dans des bases de données bibliographiques et la fourniture de documents via un réseau [10].

2.4. Système de recherche d'information :

Un système de recherche d'information est conçu pour analyser, traiter et stocker des sources d'information et récupérer celles qui correspondent aux besoins d'un utilisateur particulier. Les systèmes modernes de recherche d'informations peuvent soit récupérer des éléments bibliographiques, soit le texte exact qui correspond aux critères de recherche d'un utilisateur à partir d'une base de données stockée de documents. A l'origine, SRI désignait les systèmes de recherche de texte, car ils traitaient des documents textuels. Les systèmes modernes de recherche d'information traitent non seulement des informations textuelles mais aussi des informations multimédias comprenant du texte, du son, des images et des vidéos [11].

2.5. Concept de base de la RI :

2.5.1. Collection de documents :

Une collection de document représente l'ensemble de documents de texte authentique de taille importante et rassemblée selon des critères spécifiques. Il est défini aussi comme un ensemble de documents ou corpus manipulé par un SRI.

2.5.2. Document :

Un document est défini comme un ensemble formé par un support et une information, il constitue le potentiel d'information d'une base ou collection documentaire. Un document peut être un texte, un fragment de texte, une audio, une vidéo, une image...etc. C'est-à-dire toute unité qui peut former une réponse à un besoin informationnel de l'utilisateur.

2.5.3. Besoin d'information :

C'est une représentation de ce que l'utilisateur souhaite chercher, selon Ingwersen [12], il existe trois types de besoins :

- **Besoin thématique connu :** L'utilisateur cherche à exploiter de nouvelles notions dans un sujet et un domaine connu, il existe deux types de ce besoin : stable ou variable.
- **Besoin thématique inconnu** : L'utilisateur vise à chercher de nouvelles notions ou relations différentes des domaines qui lui sont familiers, il est toujours exprimé de façon incomplète.
- **Besoin vérificatif :** L'utilisateur est censé de faire une comparaison entre le texte à vérifier et les données connus qu'il possède déjà, il recherche donc une donnée précise, et sait comment y accéder. La recherche d'un article sur internet et sa date de publication à partir d'une adresse connue est un exemple d'un tel besoin.

2.5.4. Requête :

La requête est l'expression du besoin en information de l'utilisateur, elle représente l'interface entre le SRI et l'utilisateur, elle peut être exprimé en langage :

- **Naturel :** Sous forme de question (exemple : c'est quoi la RI, RIS).
- **Booléen :** Utilisation des opérateurs logiques (RI ET RIS).
- **Graphique :** à partir d'une interface graphique.

2.5.5. Pertinence :

C'est un concept élémentaire dans le domaine de la RI, présenté dans ces modèles comme une mesure appelé score, cherchant à évaluer la correspondance entre un document et une requête. Il existe deux types [13]:

- La pertinence système : Présenté par un score attribué par le SRI.

- La pertinence utilisateur : correspond à l'identification d'ensemble de documents restitués par le SRI.

2.6. Processus de RI :

2.6.1. Indexation :

Les systèmes de recherche fiables ne travaillent pas directement avec les documents (ou les requêtes). Ils utilisent différentes techniques et stratégies pour représenter les principaux aspects sémantiques des documents et des requêtes. Ce processus est appelé indexation.

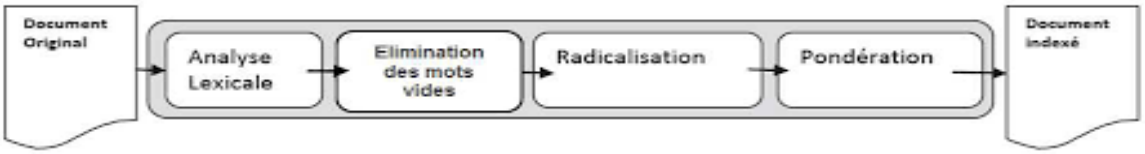

Figure 1 : Processus d'indexation.

- **Indexation des dimensions :** Présente la représentation du contenu sémantique des documents et leurs caractéristiques externes.
- **Processus d'indexation :** La plupart des systèmes de RI existants se basent sur l'indexation automatique des documents et des requêtes. Développé dès le début des années 60 [14], ce type d'indexation permet de traiter les énormes quantités d'informations disponibles en ligne. Un algorithme d'indexation automatique simple est composé de quatre étapes [15]:
  - **Extraction des mots :** Dans cette première étape, les documents sont analysés afin de reconnaître leur structure. Pour chaque structure logique pertinente, le système segmente ensuite les phrases en un ensemble de termes où un terme est un radical ou une unité lexicale. Cette analyse a pour but de reconnaitre les espaces de séparation, des chiffres, des mots et des ponctuations.
  - **Elimination des mots vides :** Dans la deuxième étape, les formes de mots très fréquentes (telles que les déterminants, les prépositions, les conjonctions, les pronoms, etc.) apparaissant dans une liste appelée "liste des mots vides" sont généralement supprimés. Cette suppression est généralement guidée par deux considérations. Premièrement, elle effectue la correspondance entre les requêtes et les documents en se

basant sur les mots porteurs de contenu uniquement, deuxièmement, parce que la suppression des mots vides aide à réduire de la taille de stockage de la collection indexée.

- **Normalisation :** A cette troisième étape, la procédure d'indexation utilise un certain type de normalisation pour tenter de regrouper les variantes de mots dans le même radical ou la même racine. Cette phase qui vise à identifier la racine d'un mot et à l'utiliser à la place du mot lui-même, peut augmenter le taux de réussite lors de la mise en correspondance des documents avec une requête. Cette démarche automatique peut donc être un outil essentiel pour améliorer l'efficacité de la recherche.
- **Pondération des mots :** Comme mentionné précédemment, le SRI divisera automatiquement la phrase donnée en mots en éliminant ainsi les mots les plus courants et supprimant les suffixes pour générer un ensemble d'unités d'index. On distingue deux types de pondération [16]:

I. Pondération locale : représenté par la fonction :

TF (Terme Frequency) : établi en comptant le nombre d'occurrences d'un terme dans un document. Soit le document $\boldsymbol{d_j}$et le terme $\boldsymbol{t_i}$, alors la fréquence $\boldsymbol{TF_{ij}}$ du terme dans le document peut être donnée selon l'une des formulations suivantes :

$$\boldsymbol{TF_{ij}} = 1 + log\ (\boldsymbol{td_{ij}})\ ou\ bien :\ \boldsymbol{TF_{ij}} = \frac{tdij}{\sum_k tdkj}$$

Avec td_{ij} représente le nombre d'occurrences du terme t_i dans le document d_j. Le dénominateur est la somme des occurrences de tous les termes dans le document d_j.

II. Pondération globale : représenté par la fonction :

IDF (Inverse Document Frequency) : est calculé en divisant le nombre total de documents par le nombre de documents de la collection contenant le terme. Il est utile pour réduire le poids des termes qui sont communs dans une collection de documents Cette mesure est exprimée selon l'une des formulations suivantes :

$$IDF_i = log_n\ (\frac{N}{n_i})\ ou\ bien :\ IDF_i = log_n\ (\frac{N}{n_{i+1}})$$

Avec N représente le nombre de documents de la collection et n_i est le nombre de documents dans lesquels le terme t_i apparaît. De manière générale, la méthode de pondération la plus utilisée est construite par la combinaison de ces deux facteurs (TFIDF) :

$$TFIDF_{td} = TF_{t,\ d} * IDF_t$$

2.6.2. Requêtage et appariement :

La recherche s'intéresse à sélectionner les documents pertinents qui couvrent les besoins d'information de l'utilisateur, à cet effet un score de pertinence entre la requête indexée et les descripteurs des documents de la collection est calculée à partir d'une valeur appelée RSV (q, d) (Retrieval Status Value), où q représente une requête et d d'un document. Seuls les documents dont la similitude dépasse un seuil prédéfini sont sélectionnés par le système de RI.

2.7. Modèle de RI :

Un modèle de recherche d'information sélectionne et classe les documents pertinents par rapport à la requête d'un utilisateur, Le texte du document et la requête sont exprimés de la même manière, de sorte que la sélection et le tri des documents peuvent être formalisés via la fonction de correspondance, qui renvoie la valeur de l'état de recherche de chaque document de la collection [17]. On distingue trois principales catégories de modèles : modèles booléens, modèles vectoriels et modèles probabilistes :

2.7.1. Modèle Booléen :

Le modèle booléen est le premier modèle de la recherche d'information [18]. Dans ce modèle, les documents et les requêtes sont représentés sous forme d'un ensemble de termes. Boole a défini trois opérateurs de base, la conjonction appelée AND, la disjonction appelée OR et la négation appelée NOT [19]:

- **La Conjonction (∧) :** Les termes doivent être présents simultanément dans la description d'un document.
- **La Disjonction (V) :**au moins un des termes soit présent dans la description d'un document à retourner.
- **La Négation (¬) :** Utilisée pour écarter les documents qui contiennent un terme.

2.7.2. Modèle Vectoriel :

Le modèle d'espace vectoriel (VSM) pour la recherche d'information représente les documents et les requêtes comme des vecteurs incorporés dans un espace euclidien de M dimensions [20], Chaque vecteur est une mesure de l'importance d'un terme d'indexation dans un document ou une requête. Les poids des termes d'indexation sont calculés sur la base de la fréquence des termes indexé dans le document, la requête ou la collection. Le mécanisme de recherche consiste donc à retrouver les vecteurs documents qui s'approchent le plus du vecteur requête.

Les principales mesures de similarité utilisées sont [21] :

Mesures	Formules
Le produit scalaire	$\mathrm{RSV}(q_i, d_j) = \sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)$
La mesure de consinus	$\mathrm{RSV}(q_i, d_j) = \frac{\sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)}{\sum_{k=1}^{M} \left(w_{k_i}{}^2\right) + \sum_{k=1}^{M} \left(w_{k_j}{}^2\right) - \sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)}$
La mesure de Dice	$\mathrm{RSV}(q_i, d_j) = \frac{2 * \sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)}{\sqrt{\sum_{k=1}^{M} \left(w_{k_i}{}^2\right) + \sum_{k=1}^{M} \left(w_{k_j}{}^2\right)}}$
La mesure de Jaccard	$\mathrm{RSV}(q_i, d_j) = \frac{\sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)}{\sqrt{\sum_{k=1}^{M} \left(w_{k_i}{}^2\right)} * \sqrt{\sum_{k=1}^{M} \left(w_{k_j}{}^2\right)}}$

Table 1 : Mesures de similarité dans le modèle vectoriel.

Prenant l'exemple de la mesure de cosinus au moment de la recherche, les documents sont classés en fonction du cosinus de l'angle entre les vecteurs du document et le vecteur de la requête. Pour chaque document et chaque requête, le cosinus de l'angle est calculé comme le rapport entre le vecteur du document et le vecteur de la requête, et le produit de la norme du vecteur du document par la norme du vecteur de la requête.

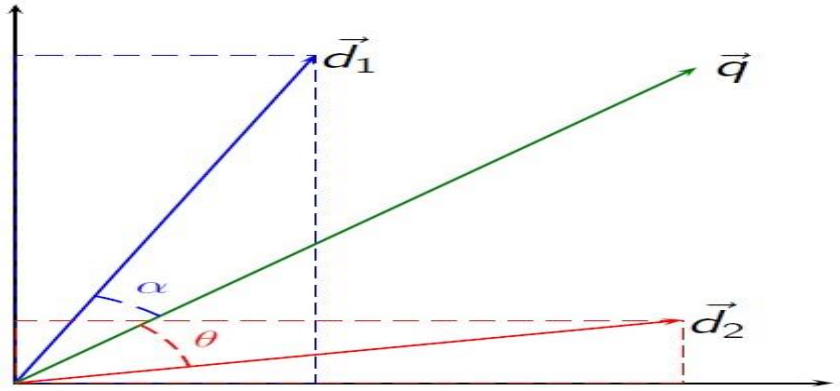

Figure 2 : Représentation vectorielle de deux documents et une requête.

2.7.3. Modèle Probabiliste :

Dans les modèles probabilistes, la recherche est considérée comme un processus de classification. Pour chaque requête, le système doit former deux classes : pertinente et non pertinente. Ainsi, pour un document donné Di, on doit estimer la probabilité qu'il appartient à la classe pertinente (classe notée R) ou à la classe non pertinente (notée $\overline{R}$). Avec deux classes, la règle de décision est assez simple : récupérer Di si Prob [R | Di] > Prob [$\overline{R}$ | Di]. Le principal fondement théorique de ce modèle est donné par le principe suivant [22] Le principe du classement probabiliste énonce que la pertinence a une interprétation probabiliste. Selon ce principe, les documents sont classés selon une probabilité P (Rel | d, q), où Rel désigne l'éventualité qu'un document d soit pertinent pour une requête q.

2.8. Evaluation :

Pour juger qu' un modèle de RI est meilleur qu'un autre, une méthodologie d'évaluation doit être adoptée dans ce domaine, celle-ci doit être appliquée au processus de recherche dans son ensemble (paradigme de l'utilisateur), ce qui signifie une évaluation par des utilisateurs réels avec leurs besoins réels en information, il est intéressant d'analyser une série de caractéristiques telles que la vitesse de réponse, sa qualité, l'effort nécessaire à l'utilisateur pour écrire une requête, l'interface du système de recherche, la couverture de la collection, etc. Tous ces aspects sont certainement importants mais les études sur les utilisateurs sont coûteuses, et certaines caractéristiques sont difficiles à mesurer objectivement. Ainsi, les approches traditionnelles d'évaluation de la RI se limitent généralement à la performance de la qualité de réponse (efficacité de la recherche). Pour mesurer la performance de la SRI [23], nous avons d'abord

besoin d'une collection de test contenant un ensemble d'unités d'information (par exemple, des documents), et un ensemble de formulations de requêtes avec leurs évaluations de pertinence. En conclusion, afin de trouver des documents pertinents à une requête, et donc utiles pour l'utilisateur, la qualité d'un système doit être mesurée en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir. Plus les réponses du système correspondent à celles que l'utilisateur espère, mieux est le système.

2.9. Conclusion :

Dans ce chapitre nous avons passé en revue les principaux concepts de la RI classique. Nous avons, particulièrement, introduit des notions de base, telles que le besoin en information, la requête, le document et la pertinence. Nous avons aussi décrit le processus de base de la RI, à savoir l'indexation, l'appariement requête-document et la reformulation de la requête. Ensuite, nous avons étudié les différents modèles de la RI. Enfin, l'évaluation des systèmes de recherche d'information est traitée. Avec l'avènement du web et surtout l'émergence des technologies web 2.0, les utilisateurs sont devenus des producteurs de l'information, et donc le document présente de nouvelles dimensions (information sociale) autre que le contenu textuel.

Ces nouvelles sources d'information sur le document doivent être intégrées dans les modèles de RI, afin d'améliorer les performances de la recherche d'information. Dans le chapitre suivant, nous abordons la recherche d'information sociale (RIS).

CHAPITRE 02 :

Recherche d'information sociale (RIS) :

Chapitre 02

3. Recherche d'information sociale:

3.1. Introduction :

Dans le contexte du web 2.0 et avec l'émergence des blogs, des wikis et des réseaux sociaux, l'utilisateur ne passe plus sans laisser de traces. Il est producteur d'information et non plus consommateur uniquement. L'information générée est dite « *information sociale ou contenu social* », qui est donc toute information fournie par l'usage du web 2.0 : les tags utilisés pour l'annotation, les traces de l'utilisateur (boutons sociaux, clics, navigation sur le web, visualisation des pages web et documents...), les relations entre les utilisateurs et les profils des utilisateurs. [24] En outre, les réseaux sociaux et les sites collaboratifs [25] (tels que Facebook, LinkedIn, Twitter, YouTube, etc.) sont les sources les plus courantes et les plus populaires de contenu interactif. Le nombre d'utilisateurs de ces réseaux ne cesse de croître et le nombre de leurs utilisateurs actifs est très élevé.

3.2. Information sociale dans le web :

3.2.1. Media sociaux : Les médias sociaux sont une technologie informatique qui permet de simplifier le partage d'idées, de pensées et d'informations par la création de réseaux et de communautés virtuels. Les médias sociaux sont des outils d'information sur Internet qui permettent aux utilisateurs de partager rapidement du contenu par voie électronique [26].

3.2.2. Web social : Le web social désigne les services, structures et interfaces web qui favorisent les interactions sociales entre les humains. Il s'agit notamment des plateformes de médias sociaux, des forums et même des portails de commerce électronique. Tous ces éléments sont liés aux différentes façons dont les humains utilisent la technologie pour interagir les uns avec les autres en ligne [27].

3.2.3. Réseaux sociaux: Les réseaux sociaux sont apparus comme une plate-forme permettant d'afficher des profils individuels, de partager des informations, des photos, des vidéos et des expériences entre les internautes, ainsi que de créer des liens d'amitié et de s'envoyer des messages [28].

Types des réseaux sociaux : Il existe plusieurs services de réseaux sociaux basés sur le Web, tels que Facebook, Twitter, LinkedIn, YouTube, etc., qui offrent une interface interactive et facile à utiliser pour entrer en contact avec des internautes à travers le monde.

- **Les réseaux personnels (Facebook) :** Facebook[2] est un réseau social qui permet de partager des textes, des photos, des vidéos, etc. Depuis sa création dans une chambre d'étudiant à Harvard en 2004 par Mark Zuckerberg, Facebook est devenu le plus grand réseau social du monde avec plus d'un milliard d'utilisateurs [29].
- **Les réseaux de médias (YouTube) :** YouTube[3] est un site web conçu pour le partage de vidéos. Des millions d'utilisateurs dans le monde ont créé des comptes sur le site qui leur permettent de télécharger des vidéos que tout le monde peut regarder. YouTube racheté par Google en 2006, classé en 2018 comme la deuxième destination en ligne la plus populaire [30].
- **Les réseaux d'actualité (Twitter) :** Twitter[4] est un réseau social que les gens utilisent pour communiquer entre eux par de courts messages, appelés tweets. Twitter a été créé à San Francisco au sein de la startup Odeo Inc. fondée en 2006 par Noah Glass et Evan Williams [31].
- **Les réseaux professionnels (LinkedIn) :** LinkedIn[5] est un réseau social axé sur le réseautage professionnel et le développement de carrière. Fondé en 2002, le site est un lieu où les professionnels peuvent augmenter le nombre de leurs relations d'affaires, créer un réseau au sein de leur secteur, discuter d'idées commerciales, rechercher des emplois et chercher de nouveaux recrutements [32].

3.2.4. Contenus générés par les utilisateurs :

Les données publiées par les utilisateurs sur le Web ont les caractéristiques suivantes : elles sont publiquement accessibles aux autres utilisateurs (ou au moins à un cercle d'amis), elles contiennent une certaine part d'effort créatif et sont créées par le grand public en dehors de leurs activités professionnelles. Ces données sont appelées contenu généré par l'utilisateur (CGU), également appelé contenu créé par l'utilisateur ou média généré par le consommateur

[2] www.facebook.com

[3] www.youtube.com

[4] www.twitter.com

[5] www.Linkdln.com

(MGC) [33]. Nous suivons la définition de l'Organisation pour la coopération et de développement économiques (OCDE) [34] : "Content is considered as UGC when the content is published on a publicly accessible Web site or on a page on a social networking site that is accessible to a select group of people. There is some creative effort in putting together the work or adapting existing works in order to create a new one, which is done by one person or as a collaborative effort. Finally, there is the requirement that UGC is created outside professional routines and practices. UGC is currently distinguished from professionally generated content because a well-established class of professional content providers currently have their presence on the Internet".

3.2.5. Signaux sociaux :

Les signaux sociaux constituent l'un des plus populaires UGC (User Generated Content) sur le web. En effet, les pages web incluent des boutons de différents réseaux sociaux où les utilisateurs peuvent indiquer s'ils soutiennent, recommandent ou n'aiment pas un contenu (texte, image, vidéo, etc.). [35] Ces boutons qui indiquent des actions liées aux activités sociales (par exemple, aimer, partager) sont associées à des réseaux sociaux spécifiques (par exemple, Facebook et twitter) avec des compteurs montrant le taux d'interaction avec la ressource web.

Figure 3 : Exemple d'une page web avec des boutons des signaux sociaux.

Un signal social est une évaluation de l'activité sur les médias sociaux. Il s'agit d'une interaction sociale d'une personne avec une ressource sur le web à travers les fonctionnalités proposées par les réseaux sociaux. Comme pour les hyperliens, les actions sociales (par exemple, j'aime, commenter, partager) peuvent être interprétées comme une appréciation de la ressource, ce qui peut contribuer à améliorer son classement dans les moteurs de recherche. Les médias sociaux permettent aux utilisateurs d'obtenir et de partager des informations et de donner leur avis en appuyant sur les boutons "J'aime" et "Je n'aime pas" ou en commentant les informations disponibles [36]. Le tableau 4 [16] résume les signaux sociaux les plus populaires sur les réseaux sociaux :

Activités sociales	Boutons de signaux sociaux	Réseaux sociaux
Vote	Like +1	Facebook, LinkedIn,youtube Google+, StumbleUpon
Message	Tweet Post	Facebook, Google+, LinkedIn, Twitter
Partage	Share Re-tweet	Google+,Twitter, Buffer,youtube Facebook, LinkedIn
Tag	Bookmark Pin	Delicious, Diigo, Digg Pinterest
Commentaire	Comment Reply	Facebook, Google+,youtube LinkedIn, Twitter
Emotion	Love, Haha, Wow Sad, Angry	Facebook
Reaction-evenement	Pride (Pride month) Thankful (Mother's day)	Facebook
Relation d'amis	Followers	Facebook, Twitter

Table 2 : Types de signaux sociaux.

3.2.6. Signaux sociaux et moteurs de recherche :

Malgré l'absence de relation claire entre les signaux sociaux et les moteurs de recherche les plus connus (par exemple, Google et Bing), il existe de nombreuses raisons pour lesquelles les signaux ne peuvent pas être ignorés. Plutôt que de considérer les signaux et le classement des résultats par les moteurs de recherche comme deux éléments distincts. Il est utile de les considérer comme des processus interconnectés qui tendent vers un objectif global : augmenter la visibilité en ligne. Depuis la création de Facebook ou d'autres réseaux sociaux, les signaux sociaux sont également devenus une information importante pour le référencement (Search Engine Optimization) [37]. Ils fournissent des informations sur l'interactivité sociale, le comportement social et les relations sociales. La corrélation entre le signal social et la position de classement d'une URL est extrêmement élevée. Les travaux de Lewoniewski et al. [37] Conduisent à l'hypothèse que les résultats des signaux sociaux sont également en corrélation avec la qualité des articles de Wikipédia. Par conséquent, les signaux sociaux peuvent être un indicateur de la pertinence des ressources web. Google reste encore ambiguë sur la façon d'exploiter les signaux sociaux pour classer ses résultats de recherche, mais certaines études menées annuellement depuis 2016, searchmetrics [38] ont montré qu'il existe une forte corrélation entre les signaux sociaux et les classements fournis par les moteurs de recherche tels que Google.

Cependant, le degré auquel les signaux sociaux jouent un rôle dans le référencement n'est pas clair. John Mueller (Webmaster Trends Analyst at Google) a déclaré: "Do social media signals have an impact on organic rankings in Google? Not directly. No. So it's not that there's any kind of a ranking effect there. To a large part social networks also have a no follow on the links that they provide when they post this content, so it's not the case that would give you any kind of a ranking boost there. What you do sometimes see however is that the social posts show up in the search results." [39].

- **GOOGLE :** Bien que Google ne dispose pas d'un partenariat avec Facebook, il a quand même l'accès aux données publiques de Facebook et peut en utiliser certaines pour mieux comprendre la popularité des pages web. En 2015, Google a signé un accord avec Twitter pour indexer les tweets en temps réel, ce qui permet d'y effectuer des recherches. L'accès libre à la base de données de Twitter signifie que toutes les informations sur Twitter sont

disponibles pour Google automatiquement. Les algorithmes de Google se concentrent également sur les profils Twitter qui tweetent et retweetent du contenu, mais comment le faire reste une boîte noire.

- **BING :** Bing, le deuxième plus grand moteur de recherche après Google, utilise des signaux tels que les "tweets" de twitter et les "j'aime" de Facebook ainsi que d'autres signaux sociaux comme facteurs de classement [39, 40, 41]. Bing s'est associé à Facebook pour la recherche sociale. [42]Les algorithmes de Bing se concentrent sur le contenu des médias sociaux, les liens, la popularité des différents réseaux sociaux qui sont considérés comme des facteurs importants par Bing pour définir le classement des résultats. Bing est un exemple typique de l'exploitation d'images et d'informations (messages, signaux) provenant des médias sociaux pour fournir des résultats plus fiables aux utilisateurs [41]. L'activité des médias sociaux est aussi présentée sur les pages de résultats de Bing de façon plus apparente que dans les autres moteurs de recherche. Le contenu social tel que les tweets et les données de Facebook contenant des mots clés pertinents, sont généralement intégrés aux résultats de recherche Bing.

Ainsi, publier du contenu visuel (images, vidéos, etc.) sur les réseaux sociaux est un excellent moyen d'augmenter la visibilité sur Bing. Par ailleurs, Bing met en place un service appelé "Social Sidebar" qui exploite Facebook pour améliore une recherche [43]. Il s'agit de la troisième colonne de la page de résultats, permettant aux utilisateurs connectés de commenter et d'aimer les résultats Facebook associés sans quitter la page de recherche. Ce service n'est disponible qu'aux États-Unis.

3.3. Notion de la RI sociale :

La recherche d'information sociale se réfère à une sorte de technique de recherche d'information qui aide les utilisateurs à obtenir des résultats qui répondent à leurs besoins en informations en exploitant les connaissances ou l'expérience de recherche des autres utilisateurs. Les utilisateurs sont reliés par des systèmes d'information en réseau tels que l'Internet.

Alors que la RI classique concerne l'interaction d'un individu avec un système d'information lors de la recherche d'informations (les autres utilisateurs ne sont pas pris en considération), l'approche collaborative de la recherche d'information fait appel à l'expertise d'autres utilisateurs lors de la recherche d'informations. Alors que la première approche peut être assimilée à un effort individuel, la seconde est plus un effort d'équipe pour la recherche [44].

3.4. Exploitation des informations sociales pour la RI :

L'utilisation des réseaux sociaux nous aide à améliorer nos recherches en se basant sur de nouvelles ressources sociales (documents, utilisateurs, etc.) :

- La réputation d'un document annoté par plusieurs utilisateurs qui ont une influence sur leur public dans les réseaux sociaux peut la rendre fiable et intéressante.
- Un utilisateur référencé peut être reconnu comme une bonne source d'information grâce à son impact à son entourage.
- Un utilisateur est plus intéressé par des recommandations des documents venant de son entourage que par des documents qui lui sont fournis par des individus inconnus.
- A l'aide de l'identification de profil de l'utilisateur et son centre d'intérêts à partir des informations au sein de son réseau social, le SRI est capable de récupérer des résultats de recherche répondant le mieux à ces besoins [45].

3.5. Travaux relatifs à la RIS :

Certaines approches de la recherche sociale reposent sur l'exploitation des "tags". Abel et al. [46] et Hotho et al. [47] ont proposé différents algorithmes basés sur les folksonomies. Par analogie FolkRank est basé sur la relation entre le tag, la ressource et l'utilisateur., alors qu'avec GFolkRank un groupe de ressources est identifié par un tag unique dans ce contexte.

Dans le même contexte, Bao et al. [48] ont proposé deux algorithmes, SocialPageRank et SocialSimRank, en collaboration avec Yanbe et al. [49], ils proposent l'algorithme SBRank. Ces algorithmes sont motivés par le constat qu'il existe une forte interdépendance entre la popularité des utilisateurs, des tags et des ressources dans une folksonomie. Ils se concentrent sur la recherche sociale collective et ne respectent pas les différents types d'engagement ou les différents niveaux de confiance. SocialSimRank calcule la similarité entre deux tags de folksonomie et déclare que les tags similaires sont généralement assignés à des ressources similaires, les tags sont parfois plus fiables que les métadonnées fournies par le producteur de contenu. Cependant, un seul tag peut difficilement couvrir un sujet entier et est plus ambiguë pour l'utilisateur qu'une phrase contextuelle.

Plusieurs travaux récents se concentrent sur la manière d'améliorer l'efficacité de la RI en exploitant les actions des utilisateurs sur leur réseau social. Chelaru et al. [50]ont étudié l'impact des signaux (like, dislike, comment) sur l'efficacité de la recherche sur YouTube. Ils ont indiqué, bien qu'il a une efficacité pour la recherche des vidéos en basant sur l'utilisation du concept de base de similarité entre les requêtes, les annotations et les titres de vidéos. Les critères sociaux sont également utiles et améliorent le classement des résultats de recherche pour 48% des requêtes. Ils ont utilisé des algorithmes de sélection d'attribut et des algorithmes d'apprentissage du classement.

Karweg et al. [51] ont suggéré une approche combinant le score thématique et le score social en fonction de deux facteurs : a) l'intensité de l'engagement de l'utilisateur quantifie l'effort fourni par un utilisateur au cours d'une interaction avec le document, mesuré par le nombre de clics, le nombre de votes, le nombre d'enregistrements et la recommandation. b) le degré de confiance mesuré à partir du graphe social pour chaque utilisateur en fonction de sa popularité, en utilisant l'algorithme PageRank. Ils ont constaté que les résultats sociaux sont disponibles pour la plupart des requêtes et conduisent généralement à des résultats plus satisfaisants.

De même, Khodaei et al. [52, 40]ont proposé une approche de classement basée sur plusieurs facteurs sociaux, notamment les relations entre les propriétaires de documents et l'utilisateur demandeur, l'importance de chaque utilisateur et leur actions (playcount : nombre de fois qu'un utilisateur écoute un morceau sur last.fm) effectuées sur les documents Web. Ils ont réalisé un vaste ensemble d'expériences sur des données de last.fm. Ils ont montré une amélioration considérable du classement socio-textuel par rapport aux approches textuelles et sociales.

Sur Twitter, Hong et al. [53] ont utilisé les retweets comme mesure de la popularité et ont appliqué des techniques d'apprentissage automatique pour prédire la fréquence à laquelle les nouveaux messages seront retweetés. Ils ont exploité différentes caractéristiques : le contenu des messages les informations temporelles, les métadonnées des messages et des utilisateurs, et le graphe social de l'utilisateur. Cependant, des tweets sans intérêt peuvent être très répandus comme ceux concernant des célébrités, qui ont généralement un grand nombre de followers.

Chan et al. [54] ont proposé un système appelé Postscholar, un service qui augmente les résultats retournés par Google Scholar, un moteur de recherche pour les références académiques. Postscholar détecte l'activité Twitter liée à un article et affiche cette information sur la page de résultats de recherche renvoyée par Google Scholar. Un hyperlien additionnel apparaît dans les résultats pour chaque article auquel est associée une activité Twitter (le nombre de tweets trouvés pour cet article, la date du tweet le plus récent). Ces tweets sont triés en fonction de leur score de sentiments.

Albishre et al. [55] ont proposé un mécanisme innovant pour sélectionner automatiquement les documents de retour d'information utiles en utilisant une technique de modélisation thématique pour améliorer l'efficacité des modèles de retour d'information de pseudo-pertinence. L'idée principale du modèle qu'ils proposent est de découvrir les latentes dans les documents les mieux classés qui permettent l'exploitation de la corrélation entre les termes des sujets pertinents. Afin de capturer les termes sélectifs pour l'expansion des requêtes, ils ont incorporé des caractéristiques thématiques dans un modèle de pertinence qui se concentre sur les informations temporelles dans l'ensemble des documents sélectionnés. Les résultats expérimentaux sur les ensembles de données de microblogs TREC 2011-2013 ont montré que le modèle proposé surpasse de manière significative tous les modèles de référence de l'état de l'art. Il existe d'autres travaux lancés par des chercheurs de Microsoft Bing [56, 41] qui indiquent l'utilité de différents comportements sociaux générés par le réseau d'amis de l'utilisateur sur Facebook. Pantel et al. [57]ont étudié l'effet de l'annotation sociale sur la qualité des résultats de recherche. Ils ont observé que les annotations sociales peuvent bénéficier à la recherche sur le web sous deux aspects : 1) les annotations sont généralement de bons résumés des pages Web correspondantes. 2) le nombre d'annotations indique la popularité des pages Web.

Hecht et al. [58] Ont présenté un système appelé 'Searchbuddies' basé sur n'importe quel réseau social en particulier l'information autour de l'utilisateur et ce que ses amis ont aimé et partagé sur des pages Facebook.

Gou et al. [59] Ont proposé une approche de classement en tenant en compte le contenu des documents et la similarité entre l'utilisateur et les documents qu'il possède dans le réseau social. Ils ont utilisé un algorithme à plusieurs niveaux pour mesurer la similarité entre les acteurs. Les résultats expérimentaux basés sur les données de YouTube montrent que, par rapport

à l'algorithme tf-idf, la méthode SNDocRank renvoie des documents plus pertinents. D'après ces résultats, un utilisateur peut améliorer la recherche en rejoignant des réseaux sociaux plus importants, en ayant plus d'amis et en se connectant à de plus grandes communautés.

Dans ce qui suit, nous présentons quelques travaux de Badache et al dans l'objectif d'exploiter les signaux sociaux pour améliorer la précision et la pertinence de la recherche textuelle conventionnelle sur le web. Ils ont exploité divers signaux extraits de différents réseaux sociaux. De plus, au lieu de considérer les caractéristiques sociales séparément comme le font les travaux précédents, ils ont proposé de les combiner pour mesurer des propriétés sociales spécifiques, à savoir la popularité et la réputation d'une ressource. Ils ont essayé également de mesurer l'impact de la fraîcheur du signal sur la performance du système de RI. L'approche proposée est complètement supervisée en exploitant les signaux sociaux collectés à partir de différents réseaux sociaux comme critères de pertinence. Elle est évaluée sur différents types de données de test standard (INEX Social Book Research) [39]:

Dans [60]Ils ont présenté la première étude préliminaire de base sous la forme d'un poster qui propose une problématique sur les besoins en information des utilisateurs provenant du web social, cette problématique est liée à ce que l'on appelle la recherche sociale ou la recherche d'information sociale.

Dans [61, 62] ils ont utilisé certaines caractéristiques telles que les commentaires et les j'aimes pour classer les ressources Web et étudier l'impact du nombre de partages et de j'aime sur la pertinence de la recherche. Dans ces études, ils sont particulièrement intéressés à : premièrement, montrer l'impact de la diversité des signaux associés à une ressource sur les performances de recherche d'information ; deuxièmement, étudier l'influence de l'origine de leurs réseaux sociaux sur leur qualité. Ils ont proposé de modéliser ces caractéristiques sociales comme des éléments intégré dans un modèle de langage non supervisé.

Dans [63, 64], l'objectif est de déterminer l'influence des nouveaux signaux sociaux émotionnels, appelés réactions Facebook (love, haha, angry, wow, sad) dans la recherche. Ces réactions permettent aux utilisateurs d'exprimer des émotions plus nuancées par rapport aux signaux classiques (par exemple, j'aime, je suis en colère, je suis triste) aux signaux classiques (par exemple, aimer, partager). Dans un premier temps, Ils ont analysé ces réactions et montré

comment les utilisateurs utilisent ces signaux pour interagir avec les publications. Ensuite, Ils ont évalué l'impact de chacune de ces réactions dans la recherche en les comparant à la fois au modèle textuel sans caractéristiques sociales et au premier signal classique.

Dans [65, 66], l'objectif est de montrer comment ces indices sociaux peuvent jouer un rôle essentiel dans l'amélioration de la recherche en arabe sur Facebook. Tout d'abord, ils ont identifié les polarités (positives ou négatives) portées par les signaux textuels (par exemple les commentaires) et non-textuels (par exemple les réactions love et sad) pour un post Facebook donné. Par conséquent, la polarité de chaque commentaire exprimé sur une publication Facebook donnée, est estimée sur la base d'un modèle neuronal de sentiment en langue arabe. Deuxièmement, ils ont regroupé les signaux en fonction de leur complémentarité en utilisant des algorithmes de sélection. Troisièmement, ils ont appliqué des algorithmes d'apprentissage au classement (LTR) pour classer les résultats de recherche de Facebook en fonction des groupes de signaux sélectionnés. Enfin, des expériences sont menées sur 13 500 publications Facebook, recueillies à partir de 45 sujets en langue arabe. Dans [67], ils ont mené une étude comparative sur l'impact des traces d'utilisateurs sur la recherche en arabe et en anglais sur Facebook.

3.6. Conclusion :

Dans ce chapitre, nous avons donné un aperçu des médias sociaux, en particulier la diversité des réseaux sociaux et l'importance de leur contenu dans la recherche sociale. Ensuite, nous avons parlé du concept de recherche d'information sociale et de ses principales tâches pour améliorer le processus de RI. Et enfin, Nous avons étudié, synthétisé et discuté les travaux relatifs à l'impact des contenus sociaux sur la recherche d'information.

Dans le dernier chapitre du volet théorique nous monterons la relation entre les émotions et la RI et comment analyser des sentiments extraits de contenus sociaux textuels depuis Twitter.

CHAPITRE 03 :

Analyse des sentiments:

Chapitre 03

4. Analyse des sentiments :

4.1. Introduction :

Plusieurs types d'analyse sont apparus pour tirer profit des informations sur le web, l'un de ces types est l'analyse des sentiments qui est l'étude informatique des opinions, des sentiments, des émotions et des attitudes des personnes. Ce problème fascinant est de plus en plus important dans les affaires et la société. Il présente de nombreux défis de recherche, mais promet un aperçu utile à toute personne intéressée par l'analyse des opinions et des médias sociaux.

4.2. Historique :

Un premier brevet sur la classification de textes incluait le sentiment, l'adéquation, l'humour et de nombreuses d'autres concepts comme labels de classe possibles [68]. Étant donné que les recherches existantes et les applications de l'analyse des sentiments se sont principalement concentrées sur le texte écrit, un domaine de recherche actif du traitement du langage naturel (NLP). Cependant, le sujet a également été largement étudié dans les domaines de la fouille de données et de la recherche d'informations car de nombreux chercheurs dans ces domaines traitent des données textuelles.

Le premier article sur le sujet a été publié sur la fouille de données en 2004 [69]. Cet article définissait le cadre de l'analyse et du résumé des sentiments basés sur les aspects, ainsi que certaines idées et algorithmes de base pour résoudre le problème qui sont couramment utilisés dans les systèmes industriels et de recherche aujourd'hui [70].

4.3. Différence entre l'analyse des sentiments et fouille d'opinions :

Il n'est pas surprenant qu'il y ait eu une certaine confusion parmi les chercheurs sur la différence entre le sentiment et l'opinion et sur la question de savoir si le domaine devrait être appelé analyse de sentiment ou fouille d'opinion. Comme le domaine est issu de l'informatique plutôt que de la linguistique, la différence entre les deux mots dans le dictionnaire Merriam-Webster, le sentiment est défini comme une attitude, une pensée ou un jugement motivé par un sentiment, tandis que l'opinion est définie comme un point de vue, un jugement ou une appréciation formée dans l'esprit sur une question particulière [71].

4.4. Niveaux d'analyse du sentiment :

La recherche sur l'analyse des sentiments a été principalement menée à trois niveaux : le niveau du document, le niveau de la phrase et le niveau de l'aspect. Nous les présentons brièvement ici [71] :

4.4.1. Niveau du document :

La tâche au niveau du document est de classer si un document d'opinion complet exprime un sentiment positif ou négatif. Elle est donc connue sous le nom de classification de sentiments au niveau du document. Par exemple, Dans le cas d'un avis sur un produit, le système détermine si l'avis exprime une opinion globalement positive ou négative sur le produit. Ce niveau d'analyse suppose implicitement que chaque document exprime des opinions sur une seule entité (par ex, un seul produit ou service). Il n'est donc pas applicable aux documents qui évaluent ou qui comparent plusieurs entités, pour lesquels une analyse plus précise est nécessaire.

4.4.2. Niveau de la phrase :

Le niveau suivant consiste à déterminer si chaque phrase exprime une opinion positive, négative ou neutre "opinion neutre" signifie généralement "pas d'opinion". Ce niveau d'analyse est étroitement lié à la classification de la subjectivité, qui distingue les phrases qui expriment des informations factuelles (appelées phrases objectives) des phrases qui expriment des vues et des opinions subjectives (appelées phrases subjectives).

4.4.3. Niveau de l'aspect :

Ni les analyses au niveau du document ni celles au niveau de la phrase ne permettent de découvrir ce que les gens aiment et n'aiment pas exactement. En d'autres termes, elles ne permettent pas de savoir sur quoi porte chaque opinion, c'est-à-dire la cible de l'opinion. Si nous pouvons classer une phrase comme positive, tout ce qui se trouve dans la phrase peut prendre une opinion. Cependant, cela ne fonctionnera pas non plus, car une phrase peut avoir plusieurs opinions, Au lieu d'examiner les unités linguistiques (documents, paragraphes, phrases ou expressions), l'analyse au niveau des aspects examine directement l'opinion et sa cible (appelée cible d'opinion). Réaliser l'importance des cibles d'opinion nous permet d'avoir une meilleure compréhension du problème de l'analyse des sentiments.

4.5. Analyse des sentiments et NLP :

L'analyse des sentiments est généralement considérée comme un sous-domaine du traitement du langage naturel (NLP). Depuis sa création, l'analyse des sentiments a élargi la

recherche en NLP de manière significative car elle a introduit de nombreux problèmes de recherche complexes qui n'avaient pas été étudiés auparavant.

Cependant, les recherches menées au cours des quinze dernières années semblent montrer que, plutôt que d'être un sous-problème de la NLP, l'analyse des sentiments est en fait une mini version de la NLP complète ou un cas particulier de la NLP complète. C'est-à-dire que chaque sous-problème de la NLP est également un sous-problème de l'analyse des sentiments, et vice versa. La raison est que l'analyse des sentiments touche tous les domaines essentiels de la NLP, tels que la sémantique lexicale, la désambiguïsation du sens des mots, l'analyse du discours, l'extraction d'informations et l'analyse sémantique [70].

4.6. Analyse des sentiments et Twitter

Twitter est une plate-forme de communication basée sur le Web, qui permet à ses abonnés de diffuser des messages appelés « tweets » de 280 caractères maximum, leur permettant de partager des pensées, des liens ou des images. Par conséquent, Twitter est une source riche de données pour l'exploration d'opinion et l'analyse de sentiment. La simplicité d'utilisation et les services offerts par la plate-forme Twitter lui permettent d'être largement utilisée dans le monde entier et en particulier en Algérie. Cette popularité nous donne accès à une mine riche d'informations qui peuvent servir comme base de données à l'analyse des tweets, qui nous fournissent des informations précieuses [72].

4.6.1. Twitter et tweet

Twitter est un réseau social et un microblog qui permet aux utilisateurs de publier des messages en temps réel, appelés tweets. Les tweets sont des messages courts, limités à 280 caractères. En raison de la nature de ce service de microblogging (messages rapides et courts), les gens utilisent des acronymes, commettent des erreurs d'orthographe, utilisent des émoticônes et d'autres symboles qui expriment des significations particulières [73].

Twitter est actuellement l'une des plates-formes de micro-blogage les plus populaires. Son premier slogan était : « Que faites-vous ? » néanmoins, son utilisation a pris une autre piste où les utilisateurs s'échangent des avis et des informations, le slogan devient « Quoi de neuf ? ».

Plusieurs célébrités utilisent Twitter, on y trouve même des chefs d'Etat.

Selon les derniers chiffres 2 :

- ✓ Twitter a plus que 645 millions d'utilisateurs inscrits.
- ✓ 58 millions de tweets envoyés chaque jour.

Dans le cadre de l'analyse des sentiments, la taille minime du message (280 caractères) formule l'hypothèse que ce dernier ne renferme, à priori, plus d'une seule idée, ce qui facilite l'identification de l'opinion exprimée. Certains tweets apparaissent comme des messages codés à cause de l'usage des hashtags, abréviations en tout genre, argot, et émoticons.

Les termes à connaître pour bien utiliser Twitter [74] :

- ✓ Followers : les personnes qui vous suivent.
- ✓ Followings : les personnes que vous suivez.
- ✓ Friends : les personnes que vous suivez et qui vous suivent.

4.6.2. Caractéristiques d'un tweet

On peut se sentir un peu perdu dans le vocabulaire utilisé dans les tweets, notamment, à cause du langage et des symboles spécifiques à l'utilisation de Twitter. A quoi sert le @ et # ?

C'est quoi RT ? Toutes ces abréviations peuvent paraître un peu floues. Dans une perspective de classement, un petit lexique des principaux mots et signes Twitter est présenté [75, 76] :

- ✓ Mention @ : se présente sous la forme @NomUtilisateur. Il cible un utilisateur de Twitter dans le tweet posté. Exemple : salut à vous de la part de @Hanae et @Moncef.
- ✓ Hashtag # : se présente sous la forme #mot-clé. Il identifie le mot-clé en question comme important et peut en faire un sujet populaire. Exemple : #gouvernement.
- ✓ RT (ReTweet) : se présente sous la forme RT Nom_Utilisatuer. Il permet de partager le tweet d'un utilisateur. Exemple : RT Hanae Excellente.
- ✓ URL (Lien) : se présente sous la forme https :// ou http ://www. Twitter, permet à l'utilisateur de rejoindre les liens dans son tweet. Exemple : https://www.fstf.com.
- ✓ VIA : s'utilise pour mentionner votre source d'information, dans votre tweet.

Example: Via YouTube, Via Facebook.

4.6.3. Analyse qualitative Twitter

L'objectif du Sentiment Score est de calculer à quel point un rapport est positif ou négatif en général. Il est important de prendre en compte le nombre d'impacts (ou d'impressions) Twitter et d'utilisateurs qui ont tweeté de manière positive, négative ou neutre. Par conséquent, le Sentiment Score prend en considération, en général, les variables suivantes [77] : Nombre de tweets positifs, négatifs ou neutres, quantité d'utilisateurs qui ont participé au rapport, nombre d'impressions (impacts) positives, négatives ou neutres, ...

L'analyse des sentiments de Twitter Binder sur Twitter considère « plus positif » que plus d'utilisateurs tweetent de manière positive plutôt qu'un seul. Cela signifie que si 500 utilisateurs ont tweeté positivement à propos de notre hashtag, c'est mieux que si seulement 1 utilisateur envoie 500 tweets. Par exemple : 500 tweets positifs envoyés par 1 utilisateur = mauvais (probablement un spammeur ou quelqu'un associé à la campagne ou à l'événement Twitter), 500 tweets positifs envoyés par 100 utilisateurs = mieux (cela signifie que beaucoup de gens conviennent que l'hashtag est génial)

Même chose arrive avec des tweets négatifs. Plusieurs fois, les trolls tenteront de saboter un événement ou une campagne. Ces utilisateurs envoient généralement des tonnes de tweets mais ils ne sont pas un grand groupe d'utilisateurs. Cela signifie qu'une petite quantité d'utilisateurs enverra une grande quantité de tweets (et probablement ils ont un petit nombre d'abonnés). Ce n'est pas très négatif pour le rapport que ce petit nombre d'utilisateurs envoie des tonnes de tweets plutôt que s'ils provenaient d'une énorme quantité de comptes. Par exemple :

- 500 tweets négatifs envoyés par 500 utilisateurs = mauvais. [77]

4.7. Conclusion :

Dans ce chapitre, nous avons présenté la revue de littérature de l'analyse des sentiments qui comprend un survole sur l'historique, la différence entre un sentiment et une opinion, les domaines d'application et sa relation avec le NLP. Ces 03 chapitres du volet théorique seront dédiés au volet pratique afin de concevoir et mettre en œuvre un système d'amélioration de la recherche classique des films.

Volet pratique

CHAPITRE 04 :

Conception de notre système de recherches de films :

Chapitre 04

5. Conception de notre système de recherches de films:

5.1. Introduction :

Dans le volet théorique sur l'état de l'art de la RIS, nous avons passé en revue de nombreux travaux existants qui montrent et prouvent que la RIS se trouve au carrefour de la RI et des réseaux sociaux. Cependant, cette RIS est ainsi appréhendée selon 3 axes :

1) le premier axe concerne la recherche d'information de nature sociale. Il s'agit de trouver des informations sociales qui répondent à l'utilisateur. On distingue par exemple la recherche d'information dans les blogs, microblogs et la recherche de conversations ;

2) la deuxième porte sur l'exploitation des contenus sociaux pour améliorer la RI, dans laquelle l'information sociale est utilisée afin d'améliorer le processus de recherche d'information, par exemple, les tags pour améliorer la recherche Web et la recherche personnalisée, le reclassement (re-ranking) des résultats de recherche et

3) le troisième paradigme concerne la recherche d'information effectuée par plusieurs personnes, recherche collaborative.

Les travaux les plus liés et proches à notre projet de fin d'étude (PFE) incluent le deuxième axe. Ces travaux s'intéressent à l'exploitation des caractéristiques sociales pour améliorer la RI sur le Web.

Notre PFE portent donc sur la proposition d'une approche RI améliorée via Twitter qui se focalise sur l'impact des interactions sociales sur le processus de recherche et qui permet d'allier efficacité et simplicité de mise en œuvre. Notre approche s'inspire de l'une des contributions de I. Badache [16] que nous espérons enrichir et étendre. Nous partons donc de l'hypothèse qu'un document doit être classé et reclassé (re-ranking) en fonction de ses pertinences thématique, sociale et émotionnelle. L'information sociale est employée du côté des documents. Elle est utilisée donc pour enrichir la représentation des ressources documentaires.

5.2. Architecture générale de notre système :

Cette section est consacrée à la description du processus général de la méthodologie de Ranking thématique et re-ranking Socio-Emotionnel des résultats de recherche en exploitant les contenus sociaux du big-social data "Twitter".

Notre processus de RI, qui est la procédure fondamentale du système, a pour but la mise en relation des informations disponibles (dataset IMDb) d'une part, et les besoins de l'utilisateur (Query) d'autre part.

Ce processus de recherche peut se résumer dans la **Error! Reference source not found.**5 ci-dessous :

- L'indexation des documents (Document) et des requêtes (Query).

- L'appariement (Document, Query), qui permet de comparer la requête et le document.

L'utilisateur exprime son besoin en information par une requête (Query) et interroge la base des films (Document) brute d'une part et augmentée de données sociales d'autres part, via une interface qui assure la communication entre la base des films et l'utilisateur. Ces données sociales représentent ainsi une valeur ajoutée (des méta-données), pour enrichir le contenu et la représentation des documents originaux.

Notre SRI est constitué de Module de traitement des documents pour l'indexation et le stockage d'information, Module de traitement des requêtes qui a pour objectif de représenter les requêtes des utilisateurs et Module de recherche d'information en effectuant une correspondance RSV (Document, Query) entre requête utilisateur et documents de la base. A l'étape final, le système renvoi l'information pertinente via un outil de visualisation.

Le processus de RI que nous proposons consiste à estimer l'importance Socio-Emot-Thématique d'une ressource en exploitant ses signaux sociaux associés et quantifiés, où chaque contenu social textuel représente un facteur de pertinence.

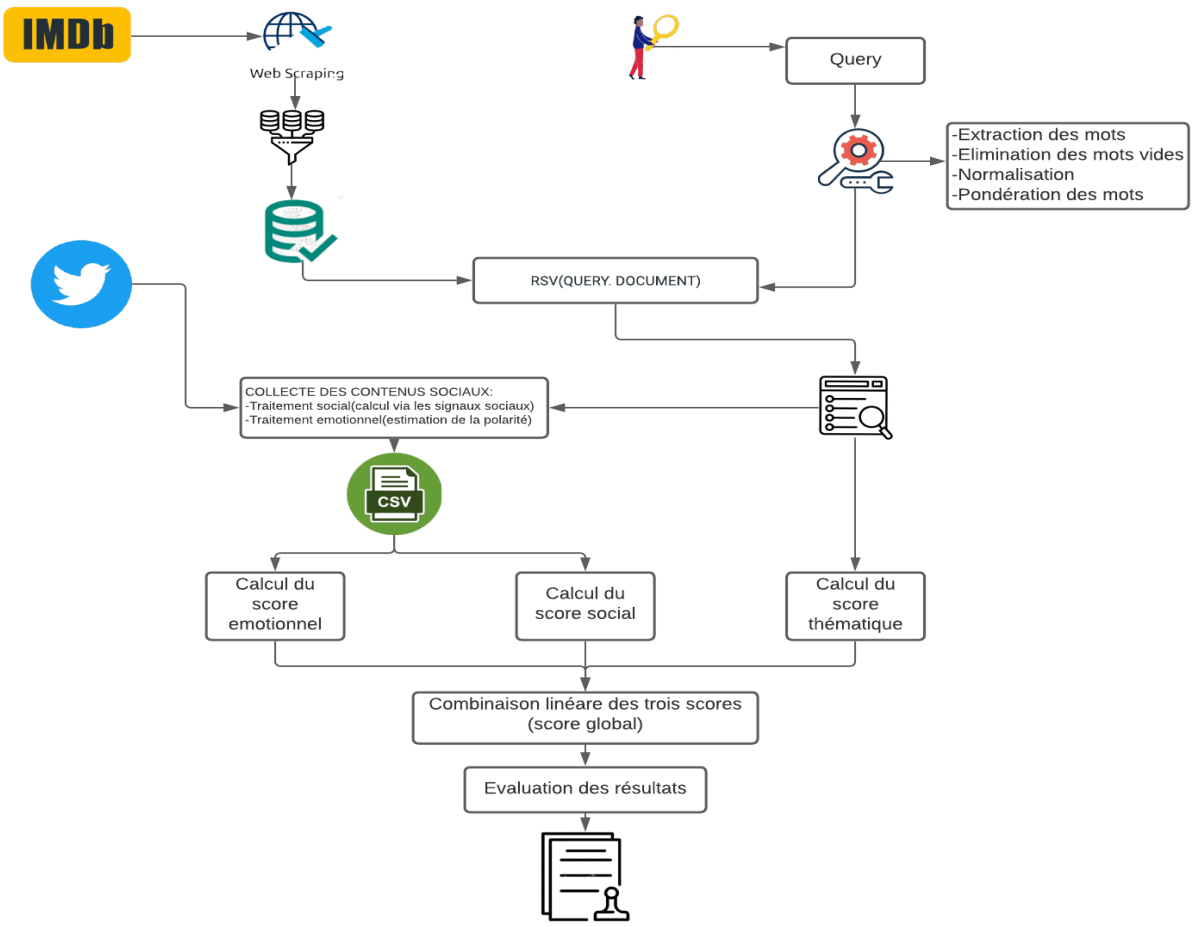

Figure 4 : Architecture générale de notre SRI

Les étapes à suivre :

5.2.1. Collecte des documents de films d'IMDB : Extraire du site Web IMDb les informations détaillées qui caractérisent les documents de films et définir le dataset IMDb brut. L'abréviation en IMDb (littéralement « Base de données cinématographiques d'Internet ») est une base de données sur le cinéma mondial, sur la télévision, et plus secondairement les jeux vidéo. IMDb restitue un grand nombre d'informations concernant les films, les acteurs, les réalisateurs, les scénaristes et toutes personnes et entreprises intervenant dans l'élaboration d'un film, d'un téléfilm, d'une série télévisée ou d'un jeu vidéo. L'accès aux informations publiques au niveau du site Web d'IMDb est gratuit. Un service payant, IMDbPro, donne accès aux informations supplémentaires susceptibles d'intéresser les professionnels.

Créé le 17 octobre 1990 par l'Anglais Col Needham, c'est un site visité par plus de 57 millions d'usagers uniques chaque mois, au 39e rang des sites les plus visités au monde. Il appartient depuis 1998 à Amazon[6].

5.2.2. Prétraitement (Preprocessing) : Étant donné que les données brutes extraites du site Web IMDb sont biaisées et bruyante, leur nettoyage est notre primordial première tâche. Ce prétraitement consiste à structurer et à faciliter l'utilisation de la collection de données originale. Cependant, Le prétraitement du dataset est le processus de nettoyage et de préparation de son contenu pour l'indexation.

Tout d'abord, ce module identifie les données inutiles et gênantes (comme les liens URL, les mots vides, les signes de ponctuation, les symboles...) pour le processus de la RI. 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, le dataset final représente les données prêtes à être utilisées dans le processus de recherche. Nous présentons l'ensemble des problèmes détectés des données extraites de « Twitter » : Le problème des espaces et des nouvelles lignes, qui occupent un espace important, qui réduit des performances du système et augmente sa complexité, De plus, Le problème des lignes vides et des contenus sans texte (par exemple, une image seule, une vidéo sans statut, etc.) et donc, ces données ne seront pas des documents pertinents aux requêtes d'utilisation, ainsi que tous leurs signaux sociaux deviennent inutiles car ils ne sont pas considérés comme des critères pour décrire l'importance de la ressource. Le problème du trop grand nombre de champs dans l'ensemble de données Twitter, dont nous n'avons pas besoin pour mesurer l'importance des documents telles que : le type, le lien et la date de la publication des statuts. Nous devons donc garder que les données utiles (id de statut, message de statut, nombre de réactions comme les montres la figure19.

5.2.3. Indexation : L'indexation consiste à choisir les termes représentatifs à partir des 45000 documents de films de notre dataset crée 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 est l'étape la plus importante de notre SRI. Un index est un ensemble de documents analysés et traités suivant schema.xml défini. Un document est un ensemble de champs (Fields) auxquels sont associées des valeurs.

[6] www.Amazon.com

5.2.4. 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 le vecteur requête et les vecteurs documents selon un score de correspondance thématique entre ces deux représentations.

L'approche thématique consiste à rechercher parmi les mots de document (tokens) celui qui a une syntaxe identique à la requête. Le résultat de cet algorithme (Figure 6) est un score de pertinence entre le document et la requête. S'il existe un token (Représentation en sac de mots) dont la syntaxe correspond à la syntaxe de la requête, alors le score égale à 1 et donc le document est pertinent à cette requête. Sinon le score de pertinence sera de zéro ce qui signifie que le document n'est pas pertinent. Ce score de pertinence thématique sert à classer et à reclasser les top premiers documents à l'aide du modèle de recherche probabiliste qui aide à prendre des décisions de classement assez précise.

Entrée : tableau de tokens du document
chaine de caractères : requête
Sortie : entier : score de similarité
Début :
Si le tableau de tokens contient la requête
Alors score de similarité ⟵ 1.0
Sinon score de similarité ⟵ 0.0
Fin.

Figure 5 : Algorithme d'appariement thématique

5.2.5. Collecte des contenus sociaux via Twitter : Afin d'améliorer les résultats de recherche du processus thématique, les contenus sociaux ("favoris" et "retweet") sont collectés et incorporés comme champs additionnels au sein du dataset original. D'une part, cette information sociale est exploitée pour le calcul du score Social. Et d'autre part, le contenu social textuel (tweet) est extrait et analyser afin de quantifier sa polarité émotionnelle ([-1,1]) qui va être exploité dans le modèle RI comme une valeur ajoutée pour le calcul du score global.

5.2.6. Evaluation des résultats : Pour juger que le modèle de recherche proposé est meilleur que le modèle thématique, une série d'expérimentation d'évaluation est présentée via des tableaux comparatifs en s'appuyant sur les scores calculés et combinés au préalable. Ceci signifie qu'une évaluation manuelle doit être effectuée par des utilisateurs réels selon leurs besoins réels en information.

5.3. Contenu social

Le contenu social peut être représentée par le quadruplet <Utilisateurs(U), Ressources(R), Actions(A), Twitter >:

- ***Ressources :*** R = $\{D_1, D_2, \ldots D_n\}$ est une collection de n ressources et une ressource D est une page Web ou une ressource Web 2.0 (Tweets). Une ressource D peut être représentée à la fois comme un ensemble de mots-clés textuels, soit D = $\{w_1, w_2, \ldots w_Z\}$ où w est un terme, et comme un ensemble de caractéristiques sociales réalisées sur cette ressource, D = $\{a_1, a_2, \ldots a_m\}$ où a est une action relevant d'activité sociale.
- ***Signaux sociaux :*** un ensemble, S = $\{s_1, s_2, \ldots s_m\}$, représente m contenus sociaux que les utilisateurs peuvent effectuer sur les ressources. Ces contenus représentent la relation entre l'ensemble des utilisateurs U = $\{u_1, u_2, \ldots u_h\}$ et l'ensemble des ressources R. Par exemple sur Twitter, les utilisateurs peuvent effectuer des actions relevant d'activités sociales comme : publier, aimer, partager ou commenter.
- ***Twitter*** : Twitter est un réseau social qui contient un ou plusieurs signaux sociaux spécifiques réalisés sur une ressource D.

5.4. Approche thématique :

5.4.1. TF-IDF similarité :

TF-IDF (term frequency-inverse document frequency) (antérieurement étudié au paragraphe 2.6.1) est une mesure statistique qui évalue la pertinence d'un mot par rapport à un document d'un ensemble de documents [78].

5.4.2. BM25 similarité :

BM25 est une fonction de classement qui classe un ensemble de documents en fonction des termes de recherche apparaissant dans chaque document, indépendamment de l'interrelation entre les termes de recherche à l'intérieur d'un document (p. ex. leur proximité relative). Ce n'est pas une fonction unique, mais en fait toute une famille de fonctions de notation, avec des composants et des paramètres légèrement

différents. Il est utilisé par les moteurs de recherche pour classer les documents correspondants en fonction de leur pertinence pour une recherche donnée et est souvent appelé 'Okapi BM25' [79].

$$score(q,d) = \sum_{i=1}^{|q|} idf(q_i) \cdot \frac{tf(q_i,d) \cdot (k_1+1)}{tf(q_i,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl}}$$

Figure 6 : Formule Okapi BM25.

- tf (qi,d) est en corrélation avec la fréquence du terme, définie comme le nombre de fois que le terme est interrogé qi figure dans le document d.
- | d | est la longueur du document d en mots (terms). Dans notre implémentation |d|est défini par :

  | d | = 1/ (norm*norm), où norm est le facteur de score utilisé par la fonction de similitude par défaut de Lucene.
- Avgdl est la longueur moyenne des documents sur tous les documents de la collection.
- K1 et b sont des paramètres libres, habituellement choisis comme k1 = 2,0 et b = 0,75.
- Idf (qi) est le poids de fréquence inverse du terme de requête qi. Il est calculé par:

$$idf(q_i) = \log \frac{N - df(q_i) + 0.5}{df(q_i) + 0.5}$$

Figure 7 : Formule idf.

- N est le nombre total de documents de la collection.
- df (qi) est le nombre de documents contenant le terme d'interrogation qi.

5.5. Approche sociale :

Comme le marketing des médias sociaux lui-même, le référencement social est plus d'un marathon qu'un sprint au rang de recherche élevé. Bien que les liens dans nos publications ne nous aident pas à nous classer plus haut sur les moteurs de recherche, notre présence globale dans les médias sociaux à une incidence sur le classement de notre contenu. Les profils et messages individuels sur les médias sociaux peuvent être recherchés, apportant plus de visibilité aux réseaux de notre marque. Notre approche de l'analyse du contenu social, consiste à exploiter les

signaux sociaux pour définir les propriétés sociales à prendre en compte dans le modèle de recherche, ces deux propriétés sont :

5.5.1. La popularité : Est un phénomène social qui indique qui est le plus connu par le public, la popularité d'une ressource web peut être estimée en fonction du taux de partage de cette ressource entre les utilisateurs par le biais d'actions sociales (signaux sociaux), Nous supposons que la popularité provient des activités sociales des utilisateurs sur le réseau social 'Twitter', c'est-à-dire qu'une ressource est dite populaire si elle a été partagée par plusieurs utilisateurs au point qu'elle devient très connue du grand public, Nous décrivons la formule de probabilité suivante pour la popularité :

$$Probabilité\ (popularité) = \frac{nombre\ de\ retweet}{\underset{(favoris,comment,retweet)}{nombre\ total\ des\ signaux\ sociaux}}$$

5.5.2. La réputation: La réputation est une opinion sur une ressource. En effet, elle dépend du degré d'appréciation des utilisateurs sur les réseaux sociaux. Nous pensons que l'estimation de cette propriété peut être calculée en se basant sur l'activité sociale favori sur twitter. Nous décrivons la formule de probabilité suivante pour la réputation :

$$Probabilité\ (réputation) = \frac{nombre\ de\ favoris}{\underset{(favoris,comment,retweet)}{nombre\ total\ des\ signaux\ sociaux}}$$

5.5.3. Score social :

Le score social prend en compte les propriétés sociales, en calculant le rapport entre la probabilité de réputation et la probabilité de popularité selon la formule suivante :

$$Score\ sociale = Probabilité\ (réputation)\ *\ Probabilité\ (popularité)$$

5.6. Approche émotionnelle :

Le classement dans l'analyse des sentiments est un score qui examine la façon dont les commentaires négatifs et positifs sont représentés. En général, il est représenté sur une échelle de -1 à 1, le bas de l'échelle indique les réponses négatives et le haut de l'échelle les réponses positives et le 0 indique les réponses neutres.

5.7. Approche globale :

Cette solution présente des idées méthodologiques en intégrant des analyses thématiques et statistiques à l'analyse des médias sociaux, pour cela nous avons proposé la formule suivante qui représente le score global de notre expérimentation :

$$Score\ globale = \alpha.score\ thématique + \beta.score\ social + \gamma.score\ émotionnel$$

D'où : $\alpha + \beta + \gamma = 1$

Et que : $\alpha = 0.5$ $\beta = 0.25$ $\gamma = .025$

Pour résumer notre modèle :

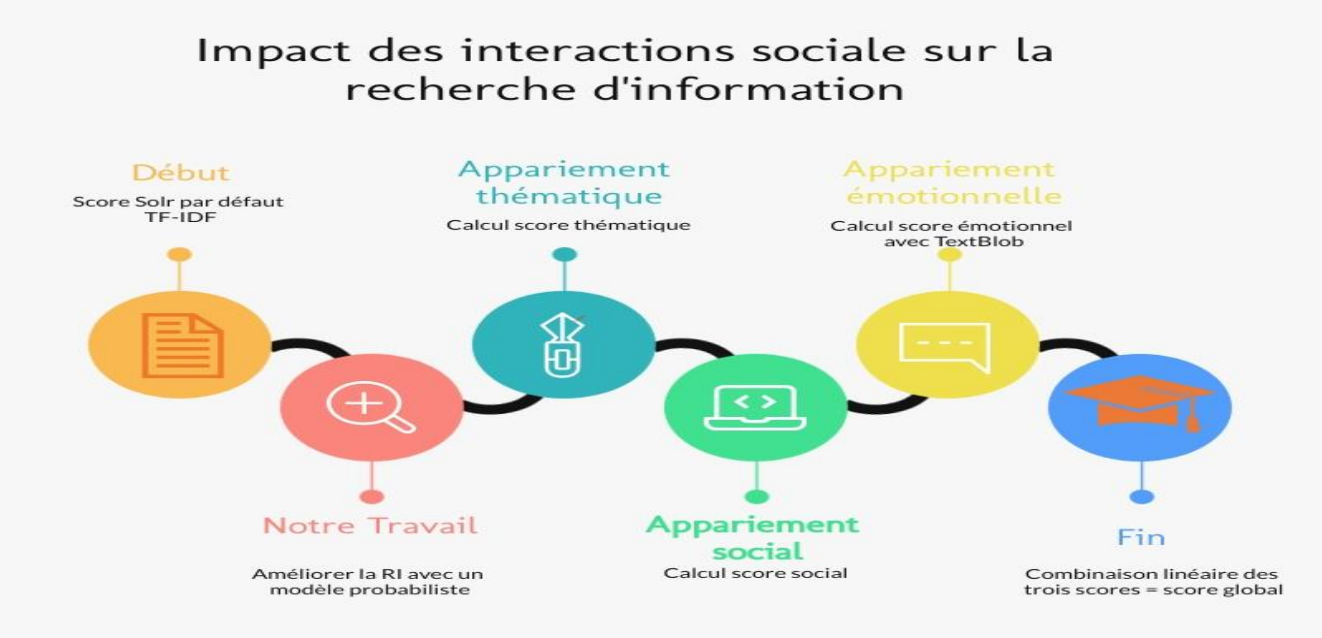

Figure 8 : Timeline de notre SRI

5.8. Conclusion :

Dans ce chapitre, nous avons présenté une solution RIS afin d'améliorer les résultats de recherche en intégrant des critères sociaux-émotionnelles où nous nous sommes partis de l'hypothèse qu'une ressource web doit être classé en fonction d'une combinaison linéaire de ses pertinences thématique, sociale et émotionnelle.

Le rang d'une ressource dans le classement global des résultats est donc déterminé par son score global, qui est calculé par la combinaison des valeurs des différents scores. Notre but est donc de fournir des résultats aussi pertinents que possible et aussi frais que possible. Dans le chapitre suivant, nous présentons l'implémentation de notre SRI en diffusant et évaluant les résultats obtenus.

CHAPITRE 05 :

Implémentation :

Chapitre 05

6. Implémentation :

6.1. Introduction :

Ce chapitre est consacré à la présentation de notre SRI développé et mis en œuvre pour répondre aux besoins informationnels des utilisateurs et au choix de certains outils de développement (collections, logiciels, langages, ...). Le système retourne donc par la suite les documents potentiellement pertinents par ordre décroissant d'appariement vis-à-vis de la requête.

6.2. Présentation de l'environnement utilisé :

Dans cette partie nous allons détailler les différents outils (collections, logiciels, langages...) utilisés pour la réalisation de notre projet :

6.2.1. Outils de développement :

- Apache Solr : Solr[7] est une plateforme open-source de recherche d'entreprise, écrite en Java. Ses principales caractéristiques incluent la recherche en texte intégral, le surlignage, la recherche facettée, l'indexation en temps réel, le clustering dynamique, l'intégration de bases de données, les fonctionnalités NoSQL et la gestion riche de documents [80].

- **Apache ZooKeeper :** Apache ZooKeeper[8] est un logiciel open source d'Apache Software Fondation, il s'agit d'un logiciel de gestion de configuration pour systèmes distribués. ZooKeeper est un sous projet de Hadoop mais il est un projet top-level à part entière [81].

- **Anaconda** : Anaconda[9] est une distribution des langages de programmation Python et R pour le calcul scientifique,Qui vise à simplifier la gestion et le déploiement des paquets. La distribution comprend des paquets de science des données compatibles avec Windows, Linux et MacOs [82].

[7] www.solr.apache.org

[8] www.zookeeper.apache.org

[9] www.anaconda.com

- **Pycharm Community :** Pycharm[10] est un environnement de développement intégré utilisé en programmation informatique, en particulier pour le langage Python.Il est développé par la société tchèque JetBrains [83].

- **Jupyter notebook :** Project Jupyter[11] est un projet et une communauté dont le but est de développer des logiciels libres, Des standards ouverts et des services pour l'informatique interactive dans des dizaines de langages de programmation. Il a été tiré d'I Python en 2014 par Fernando Pérez [84].

6.2.2. Langages de programmation :

- **Python :** Python[12] est un langage de programmation polyvalent interprété de haut niveau. Sa philosophie de conception met l'accent sur la lisibilité du code avec son utilisation d'indentation significative. Ses constructions de langage ainsi que son approche orientée objet visent à aider les programmeurs à rédiger un code clair et logique pour des projets de petite et de grande envergure [85].

6.2.3. Bibliothèques principales :

- Django : Django[13] est un Framework web libre et open-source basé sur Python qui suit le schéma architectural model-Template-Views. Il est maintenu par la Django Software Fondation, une organisation américaine indépendante fondée en tant que 501 à but non lucratif [86].

- Pysolr :Pysolr[14] est un client Python léger pour Apache Solr. Il fournit une interface qui interroge le serveur et renvoie les résultats basés sur la requête [87].

- Tweepy : Tweepy[15] comprend un ensemble de classes et de méthodes qui représentent les modèles et les terminaux API de Twitter, et il gère de manière transparente différents détails d'implémentation, tels que l'encodage et le décodage des données [88].

[10] www.jetbrains.com/fr-fr/pycharm

[11] www.jupyter.org

[12] www.python.org

[13] www.djangoproject.com

[14] www.pypi.org/project/pysolr

- Pandas : Pandas[16] est une librairie logicielle écrite pour le langage de programmation Python pour la manipulation et l'analyse de données. En particulier, il propose des structures de données et des opérations de manipulation de tableaux numériques et de séries chronologiques. Il s'agit d'un logiciel libre distribué sous la licence BSD à trois clauses [89].

- Numpy : NumPy[17] est une librairie pour le langage de programmation Python, ajoutant la prise en charge de grands tableaux et matrices multidimensionnels, ainsi qu'une grande collection de fonctions mathématiques de haut niveau pour fonctionner sur ces tableaux [90].

- Nltk : Le Natural Language Toolkit[18], ou plus communément NLTK, est une suite de bibliothèques et de programmes pour le traitement symbolique et statistique Du langage naturel pour l'anglais écrit en langage de programmation Python [91].

- Rank_bm25 : BM25[19] est un package Python simple et peut être utilisé pour indexer les données, tweets dans notre cas, en fonction de la requête de recherche [79].

- Tfidfvectorizer : TfidfVectorizer[20] – Transforme le texte en vecteurs pouvant servir d'entrée à l'estimateur. Vocabulaire_ Est un dictionnaire qui convertit chaque jeton (mot) en index de fonction dans la matrice, chaque jeton unique reçoit un index de fonction [78].

- Textblob : TextBlob[21] est une librairie python et offre une API simple pour accéder à ses méthodes et effectuer des tâches NLP de base. Une bonne chose à propos de TextBlob est qu'ils sont comme des chaînes de python. Donc, vous pouvez vous transformer et jouer avec comme nous l'avons fait en python [92].

---

[15] www.tweepy.org

[16] www.pandas.pydata.org

[17] www.numpy.org

[18] www.nltk.org

[19] www.pypi.org/project/rank-bm25/

[20] www.scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[21] www.textblob.readthedocs.io/en/dev/

- Curl : CURL[22] est un projet de logiciel informatique fournissant une bibliothèque et un outil en ligne de commande pour le transfert de données en utilisant différents protocoles réseau. Le nom signifie « Client URL », qui a été publié pour la première fois en 1997 [93].
- BeautifulSoup [23] : Est un paquetage Python pour l'analyse de documents HTML et XML. Il crée une arborescence d'analyse pour les pages analysées qui peut être utilisée pour extraire des données du HTML, ce qui est utile pour le scraping web [94].
- Requests : Est une bibliothèque[24] HTTP pour le langage de programmation Python. Le but du projet est de rendre les requêtes HTTP plus simples et plus conviviales. La version actuelle est la 2.26.0. Requêtes est publié sous la licence Apache 2.0. Requêtes est l'une des bibliothèques Python les plus populaires qui n'est pas incluse avec Python [95].

6.3.Constitution du dataset :

Cette application doit créer et extraire les données IMDB à partir d'un grand nombre d'URL IMDB. Nous utilisons l'API simultanée de Python pour rendre le processus parallèle et homogène. Il s'agit de notre fonction principale qui sera responsable de l'itération des divers attributs des données IMDb. Nous fournirons à cette fonction des URL pour différentes pages IMDB pour extraire leurs informations. L'algorithme de constitution du dataset est comme suit

```
Entrée : IMDB_URLS
Sortie : ficher csv
Début :
    Pour chaque page de imdb_pages faire :
        Pour chaque movie de movies de page faire :
            Url_movie ← getMovieURL()
            data_movie ←getData(url_movie())
        FinPour
    FinPour
Fin
```

[22] www.curl.se

[23] www.crummy.com/software/BeautifulSoup/bs4/doc/

[24] www.docs.python-requests.org/en/latest/

Figure 9 : Algorithme de constitution de notre dataset.

Le code source python de constitution du dataset :

```
# Movies in English
headers = {"Accept-Language": "en-US, en;q=0.5"}

# Request contents of the URL
# https://www.imdb.com/search/title/?release_date=1926-01-01,2021-12-
31&ref_=adv_prv
url = "https://www.imdb.com/search/title/?groups=top_1000&sort=moviemeter,desc"
results = requests.get(url, headers=headers)

# Using BeautifulSoup
soup = BeautifulSoup(results.text, "html.parser")

# Find all lister-item mode-advanced divs
movie_div = soup.find_all('div', class_='lister-item mode-advanced')

# Storing each of the urls of 50 movies
for page in pages:
    # Getting the contents from the each url
    page = requests.get('https://www.imdb.com/search/title/?
groups=top_1000&start=' + str(page) + '&ref_=adv_nxt', headers=headers)
    soup = BeautifulSoup(page.text, 'html.parser')
    # Aiming the part of the html we want to get the information from
    movie_div = soup.find_all('div', class_='lister-item mode-advanced')
    for container in movie_div:
        movie_data.append(get_data(container))

#save to csv file
movies = pd.DataFrame(movie_data)
movies.to_csv('movies.csv', index=False)
```

Figure 10 : Code source python de constitution du dataset.

Les étapes essentielles concernant la construction du dataset sont comme suit:

- Extraction des données HTML : Analyser d'abord la page et son code HTML sous-jacent pour obtenir les données requises.
- Enregistrement des données : Une fois les données collectées du site Web, elles seront immédiatement stockées dans un fichier en tant qu'objet JSON. Ce dernier regroupe tout ce qui concerne un film donné en termes de champs comme (titre, genre, description, date de

réalisation, compagnie de production). Par la suite, une migration de type de données est effectuée de l'objet. Json vers .csv. La figure 14 présente le fichier brut.csv des films de IMDb où un film est caractérisé par un ID, budget, genre, langue d'origine, description,

	id	budget	genre	original_language	overviews	release_date	runtime	titles	movieStars	movieDirector	
0	tt7131622	$142.50M	[Comedy, Drama]	en-US	A faded television actor and his stunt double ...	2019	161 min	Once Upon a Time... In Hollywood	[Leonardo DiCaprio, Brad Pitt, Margot Robbie, ...	Quentin Tarantino	http://www.imdb.com/title
1	tt4154796	$858.37M	[Action, Adventure, Drama]	en-US	After the devastating events of Avengers: Infi...	2019	181 min	Avengers: Endgame	[Robert Downey Jr., Chris Evans, Mark Ruffalo,...	Directors:Anthony Russo, Joe Russo	http://www.imdb.com/title
2	tt12361974	-	[Action, Adventure, Fantasy]	en-US	Determined to ensure Superman's ultimate sacri...	2021	242 min	Zack Snyder's Justice League	[Henry Cavill, Ben Affleck, Gal Gadot, Amy Adams]	Zack Snyder	http://www.imdb.com/title/
3	tt0111161	$28.34M	[Drama]	en-US	Two imprisoned men bond over a number of years...	1994	142 min	The Shawshank Redemption	[Tim Robbins, Morgan Freeman, Bob Gunton, Will...	Frank Darabont	http://www.imdb.com/titl
4	tt0068646	$134.97M	[Crime, Drama]	en-US	An organized crime dynasty's aging patriarch t...	1972	175 min	The Godfather	[Marlon Brando, Al Pacino, James Caan, Diane K...	Francis Ford Coppola	http://www.imdb.com/title
...	...	...	...	...	...	...	...	...	...	...	

Figure 11 : Objet .csv

6.4. Prétraitement des données :

Le prétraitement peut faire référence au filtrage des données avant leur utilisation. Le prétraitement constitue une étape importante du processus d'exploration des données.

- Nettoyage des données:

Étant donné que les données recueillies à partir des tweets sont biaisées et bruyante, le nettoyage des tweets est notre première tâche (Figure 12). Ce prétraitement consiste à structurer et à faciliter l'utilisation des données textuelles originales.

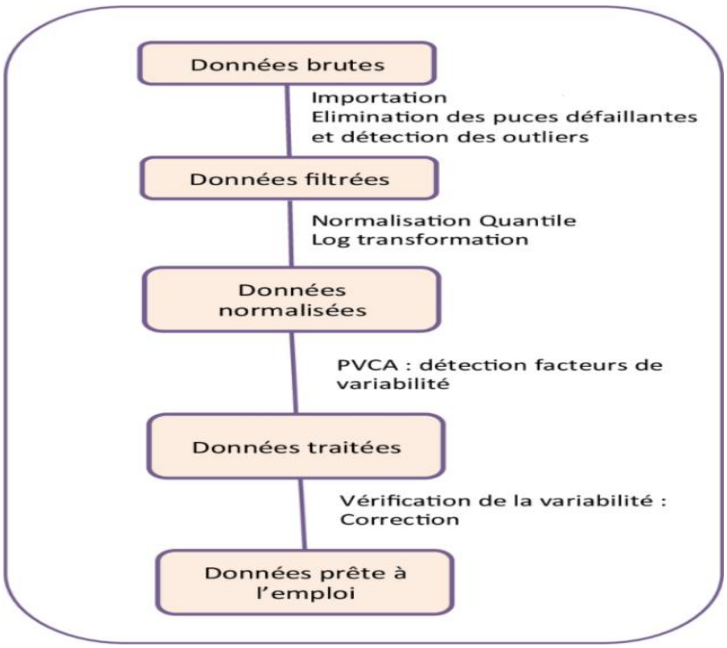

Figure 12 : Etapes de prétraitement des données brutes.

6.5. Création d'un dataset vide :

Créer une collection ou un dataset de films vide au départ étend le concept d'index nommé, géré et configuré de manière unique à un index divisé en plusieurs fragments et distribué sur plusieurs serveurs. Pour notre cas d'étude, nous avons configuré 02 serveurs chacun d'eux est composé de 02 fragments :

```
solr create -c movies -s 2 -rf 2
```

Figure 13 : Création d'un dataset de films vide.

- A cette étape nous avons définis manuellement les champs qui caractérisent un film de notre collection. Les champs sont définis dans l'élément Fields de schéma XML. La figure 20 présente les champs crées d'un film en détails sous un format .XML :

```
<field name="_nest_path_" type="_nest_path_"/>
  <field name="_root_" type="string" docValues="false" indexed="true" stored="false"/>
  <field name="_text_" type="text_general" multiValued="true" indexed="true"
stored="false"/>
  <field name="_version_" type="plong" indexed="false" stored="false"/>
  <field name="budget" type="plong" multiValued="false" stored="true"/>
  <field name="genres" type="text_general" multiValued="true" stored="true"/>
  <field name="id" type="string" multiValued="false" indexed="true" required="true"
stored="true"/>
  <field name="original_language" type="text_general" multiValued="false" stored="true"/>
  <field name="overview" type="text_general" multiValued="false" stored="true"/>
  <field name="popularity" type="pfloat" multiValued="false" stored="true"/>
  <field name="production_companies" type="text_general" multiValued="true"
stored="true"/>
  <field name="release_date" type="pdate" multiValued="false" stored="true"/>
  <field name="revenue" type="plong" multiValued="false" stored="true"/>
  <field name="runtime" type="pint" multiValued="false" stored="true"/>
  <field name="tagline" type="text_general" multiValued="false" stored="true"/>
  <field name="title" type="text_general" multiValued="false" stored="true"/>
  <field name="vote_average" type="pfloat" multiValued="false" stored="true"/>
  <field name="vote_count" type="pint" multiValued="false" stored="true"/>
```

Figure 14 : Création des champs de notre dataset .

- Ajouter un champ de saisie : la recherche est basée sur le champ saisie, ce qui signifie que le format de recherche d'une requête contiendra des mots-clés composés de champs. Lorsque nous ne spécifions pas le champ dans la chaîne de requête, la recherche ne s'effectuera jamais. Là, nous pouvons attacher tous les champs nécessaires à chercher au champ copyField. Lorsque le champ copyField est recherché par défaut, tous les champs qui le composent seront également recherchés :

```
curl -X POST -H "Content-type:application/json" --data-binary "{\"add-copy-field\" :
{\"source\":\"*\",\"dest\":\"_text_\"}}" http://localhost:8983/solr/movies/schema
```

Figure 15 : Ajout d'un champ de saisie.

- Définir la composante des suggestions automatiques pour les termes de la requête.

```
curl -X POST -H "Content-type:application/json" --data-binary "{\"add-searchcomponent\":
{\"name\":\"suggest\",\"class\":\"solr.SuggestComponent\",\"suggester\":
{\"name\":\"fuzzySuggester\",\"lookupImpl\":\"FuzzyLookupFactory\",\"storeDir\":\"fuzzy_s
uggestions\",\"dictionaryImpl\":\"DocumentDictionaryFactory\",\"field\":\"title\",\"weigh
tField\":\"popularity\",\"exactMatchFirst\":\"true\",\"suggestAnalyzerFieldType\":\"text_
general\",\"buildOnStartup\":\"false\",\"buildOnCommit\":\"false\"}}}"
http://localhost:8983/solr/movies/config
```

Figure 16 : Ajout d'un champ de suggestion automatique.

- Charger les données : Le service qui se charge de mettre à jour un index se situe à l'URL http://localhost:8983/solr/update (si Solr est sur votre machine locale bien sûr). La requête HTTP pour transmettre un document XML à indexer et s'attend à recevoir un document de type Content-type :application/xml, avec un codage binaire pour éviter d'interpréter un codage utf-8 comme de l'ASCII :

```
java -jar -Dc=movies -
Dparams="f.genres.split=true&f.production_companies.split=true&f.genres.separator=|&f.pro
duction_companies.separator=|" -Dauto ..\exampledocs\post.jar ..\movies\movies.csv
```

Figure 17 : Chargement des données des films.

- Construire un modèle LTR : Il faut d'abord modifier /chemin/vers/solr-<version>/solr/serveur/solr/test/conf/solrconfig.Xml ensuite Copier et collez le texte ci-dessous n'importe où entre les balises <config> et </config> (en haut et en bas du fichier, respectivement) :

```
<lib dir="${solr.install.dir:../../../..}/contrib/ltr/lib/" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-ltr-\d.*\.jar" />

<queryParser name="ltr" class="org.apache.solr.ltr.search.LTRQParserPlugin"/>

<cache name="QUERY_DOC_FV"
       class="solr.search.LRUCache"
       size="4096"
       initialSize="2048"
       autowarmCount="4096"
       regenerator="solr.search.NoOpRegenerator" />

<transformer name="features"
class="org.apache.solr.ltr.response.transform.LTRFeatureLoggerTransformerFactory">
  <str name="fvCacheName">QUERY_DOC_FV</str>
</transformer>
```

Figure 18 : Construction d'un modèle LTR.

- Les caractéristiques de notre modèle : les caractéristiques de notre modèle sont définies à l'aide d'un fichier formaté JSON:

```
[
   {
      "store" : "my_efi_feature_store",
      "name" : "tfidf_sim_a",
      "class" : "org.apache.solr.ltr.feature.SolrFeature",
      "params" : { "q" : "{!dismax qf=text_tfidf}${text_a}" }
   },
   {
      "store" : "my_efi_feature_store",
      "name" : "tfidf_sim_b",
      "class" : "org.apache.solr.ltr.feature.SolrFeature",
      "params" : { "q" : "{!dismax qf=text_tfidf}${text_b}" }
   },
   {
      "store" : "my_efi_feature_store",
      "name" : "bm25_sim_a",
      "class" : "org.apache.solr.ltr.feature.SolrFeature",
      "params" : { "q" : "{!dismax qf=text}${text_a}" }
   },
   {
      "store" : "my_efi_feature_store",
      "name" : "bm25_sim_b",
      "class" : "org.apache.solr.ltr.feature.SolrFeature",
      "params" : { "q" : "{!dismax qf=text}${text_b}" }
   },
   {
      "store" : "my_efi_feature_store",
      "name" : "max_sim",
      "class" : "org.apache.solr.ltr.feature.SolrFeature",
      "params" : { "q" : "{!dismax qf='text text_tfidf'}${text}" }
   },
   {
      "store" : "my_efi_feature_store",
      "name" : "original_score",
      "class" : "org.apache.solr.ltr.feature.OriginalScoreFeature",
      "params" : {}
   }
]
```

Figure 19 : caractéristiques de notre modèle.

- Enregistrer la spécification du modèle suivante : Les caractéristiques du modèle défini doivent être toujours spécifiés. Pour commencer, nous allons utiliser le modèle linéaire, qui prend simplement une somme pondérée des valeurs des caractéristiques pour générer un score. Ici, nous attribuons une pondération de 0,0 à chaque fonctionnalité sauf le score original, qui reçoit une pondération de 1,0. Ce système de pondération garantira que les résultats sont retournés dans leur ordre d'origine :

```
{
    "store" : "my_efi_feature_store",
    "name" : "my_efi_model",
    "class" : "org.apache.solr.ltr.model.LinearModel",
    "features" : [
        { "name" : "tfidf_sim_a" },
        { "name" : "tfidf_sim_b" },
        { "name" : "bm25_sim_a" },
        { "name" : "bm25_sim_b" },
        { "name" : "max_sim" },
        { "name" : "original_score" }
    ],
    "params" : {
        "weights" : {
            "tfidf_sim_a" : 0.0,
            "tfidf_sim_b" : 0.0,
            "bm25_sim_a" : 0.0,
            "bm25_sim_b" : 0.0,
            "max_sim" : 0.0,
            "original_score" : 1.0
        }
    }
}
```

Figure 20 : Enregistrement de la spécification du modèle défini.

6.6. Utilisation de l'API python:

Pysolr est un client Python léger pour Apache Solr. Il fournit une interface qui interroge le serveur et renvoie les résultats basés sur la requête. Son utilisation de base est comme suit :

```
import pysolr

# Setup a basic Solr instance. The timeout is optional.
solr = pysolr.Solr('http://localhost:8983/solr/', timeout=10)

zookeeper = pysolr.ZooKeeper("localhost:2181")

solr = pysolr.SolrCloud(zookeeper, "movies")

# Later, searching is easy. In the simple case, just a plain Lucene-style
# query is fine.
results = solr.search('my query',fl="*,score",rows="30")
```

Figure 21 : Utilisation de l'API Pysolr.

6.7. Caractéristiques de pysolr (client-serveur):

- Opérations de base telles que la sélection, la mise à jour et la suppression.
- Optimisations d'index.
- Prise en charge de suggestion automatique et la correction d'orthographe automatique
- Sensibilisation au Cloud :

```
# Do a health check.
solr.ping()
# How you'd index data.
solr.add([
    {
        "id": "doc_1",
        "title": "A test document",
    },
])
# query is fine.
results = solr.search('bananas')
# Just loop over it to access the results.
for result in results:
    print("The title is '{0}'.".format(result['title']))
# Finally, you can delete either individual documents,
solr.delete(id='doc_1')
#  ...or all documents.
solr.delete(q='*:*')
```

Figure 22 : Usage de base.

- **Usage de base:**
  - Faire un bilan de contrôle d'activité via ping.
  - Ajouter un document.
  - Chercher un document.
  - Accéder aux résultats.

    Choisir les documents à supprimer entre un ou plusieurs documents.

6.8. 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.

- **Algorithmes les plus utilisés :** Les modèles fréquents fonctionnent autour de la fréquence du terme recherché et des documents contenant le terme recherché. Cependant, le concept et l'algorithme utilisé pour calculer le score diffèrent.

  Les algorithmes de classement implémentés :

6.8.1. La mesure de similarité TF-IDF

```
from sklearn.feature_extraction.text import TfidfVectorizer

indices = pd.Series(metadata.index, index=metadata['title']).drop_duplicates()

# Create TfidfVectorizer object
vectorizer = TfidfVectorizer()

# Generate matrix of word vectors
tfidf_matrix = vectorizer.fit_transform(ted)

# compute and print the cosine similarity matrix
cosine_sim = cosine_similarity(tfidf_matrix, tfidf_matrix)

# Generate recommendations
print(get_recommendations("Harry potter", cosine_sim, indices))
```

Figure 23 : Code source de la mesure de similarité TF-IDF.

6.8.2. La similarité BM25:

```
from rank_bm25 import BM25Okapi

corpus = [
    "Hello there good man!",
    "It is quite windy in London",
    "How is the weather today?"
]

tokenized_corpus = [doc.split(" ") for doc in corpus]

bm25 = BM25Okapi(tokenized_corpus)

query = "windy London"
tokenized_query = query.split(" ")

doc_scores = bm25.get_scores(tokenized_query)

bm25.get_top_n(tokenized_query, corpus, n=1)
```

Figure 24 : code de source de la mesure de similarité BM25 Rank.

6.8.3. Calcul du score thématique :

La notation dépend beaucoup de la façon dont les documents sont indexés. Il est donc important de comprendre l'indexation et savoir comment utiliser la fonctionnalité Searcher.Explain (Query query, int doc) qui est très utile pour expliquer pourquoi une partition est retournée.

```
# Later, searching is easy. In the simple case, just a plain Lucene-style
# query is fine.
# , sort=["release_date desc","vote_average desc"]
results = solr.search('harry potter Adventure',fl="*,score",rows="30")
docs = pd.DataFrame(results.docs)
docs['score']=docs['score']/results.raw_response['response']['maxScore']
```

Figure 25 : Calcul du score thématique.

6.9. Résultats et discussions :

6.9.1. Score thématique :

Nous avons mené des expériences avec des modèles basés sur le contenu textuel des documents (TF*IDF et BM25) :

- Le score thématique calculé via TF*IDF est inséré comme un nouveau champ au fichier film en .csv :

	id	release_date	title	overview	genres	vote_average	vote_count	popularity	score
0	tt2033193	2009	A Very Potter Musical	In April 2009, a group of University of Michig...	[Music, Comedy]	7.7	20	0.027608	1.000000
1	tt0241527	2001	Harry Potter and the Philosopher's Stone	Harry Potter has lived under the stairs at his...	[Adventure, Fantasy, Family]	7.5	7188	1.000000	0.996119
2	tt0295297	2002	Harry Potter and the Chamber of Secrets	Ignoring threats to his life, Harry returns to...	[Adventure, Fantasy, Family]	7.4	5966	0.778832	0.947918
3	tt0304141	2004	Harry Potter and the Prisoner of Azkaban	Harry, Ron and Hermione return to Hogwarts for...	[Adventure, Fantasy, Family]	7.7	6037	0.745282	0.938174
4	tt0330373	2005	Harry Potter and the Goblet of Fire	Harry starts his fourth year at Hogwarts, comp...	[Adventure, Fantasy, Family]	7.5	5758	0.652130	0.918509
5	tt0417741	2009	Harry Potter and the Half-Blood Prince	As Harry begins his sixth year at Hogwarts, he...	[Adventure, Fantasy, Family]	7.4	5435	0.499741	0.901549
6	tt2403029	2013	The Starving Games	A spoof movie that references The Hunger Games...	[Comedy]	4.1	219	0.006941	0.900713
7	tt1201607	2011	Harry Potter and the Deathly Hallows: Part 2	Harry, Ron and Hermione continue their quest t...	[Family, Fantasy, Adventure]	7.9	6141	0.654426	0.896435
8	tt0373889	2007	Harry Potter and the Order of the Phoenix	Returning for his fifth year of study at Hogwa...	[Adventure, Fantasy, Family, Mystery]	7.4	5633	0.559462	0.883261
9	tt0926084	2010	Harry Potter and the Deathly Hallows: Part 1	Harry, Ron and Hermione walk away from their l...	[Adventure, Fantasy, Family]	7.5	5708	0.610161	0.836526
10	tt2040264	2010	A Very Potter Sequel	Harry and his pals are back for more adventure...	[Music, Comedy]	7.2	12	0.012343	0.776245

Figure 26 : Résultats du score thématique via TF-IDF.

- Le score thématique calculé via BM25 est ajouté comme un nouveau champ au fichier film en .csv

	id	release_date	title	overview	genres	vote_average	vote_count	popularity	score
0	tt0295297	2002	Harry Potter and the Chamber of Secrets	Ignoring threats to his life, Harry returns to...	[Adventure, Fantasy, Family]	7.4	5966	0.778832	1.000000
1	tt0304141	2004	Harry Potter and the Prisoner of Azkaban	Harry, Ron and Hermione return to Hogwarts for...	[Adventure, Fantasy, Family]	7.7	6037	0.745282	0.960986
2	tt0417741	2009	Harry Potter and the Half-Blood Prince	As Harry begins his sixth year at Hogwarts, he...	[Adventure, Fantasy, Family]	7.4	5435	0.499741	0.951407
3	tt0241527	2001	Harry Potter and the Philosopher's Stone	Harry Potter has lived under the stairs at his...	[Adventure, Fantasy, Family]	7.5	7188	1.000000	0.946900
4	tt0330373	2005	Harry Potter and the Goblet of Fire	Harry starts his fourth year at Hogwarts, comp...	[Adventure, Fantasy, Family]	7.5	5758	0.652130	0.925095
5	tt1201607	2011	Harry Potter and the Deathly Hallows: Part 2	Harry, Ron and Hermione continue their quest t...	[Family, Fantasy, Adventure]	7.9	6141	0.654426	0.917727
6	tt0373889	2007	Harry Potter and the Order of the Phoenix	Returning for his fifth year of study at Hogwa...	[Adventure, Fantasy, Family, Mystery]	7.4	5633	0.559462	0.897980
7	tt0926084	2010	Harry Potter and the Deathly Hallows: Part 1	Harry, Ron and Hermione walk away from their l...	[Adventure, Fantasy, Family]	7.5	5708	0.610161	0.874419
8	tt2033193	2009	A Very Potter Musical	In April 2009, a group of University of Michig...	[Music, Comedy]	7.7	20	0.027608	0.643423
9	tt2403029	2013	The Starving Games	A spoof movie that references The Hunger Games...	[Comedy]	4.1	219	0.006941	0.579540
10	tt0482546	2006	Miss Potter	The story of Beatrix Potter, the author of the...	[Drama, Family, Romance]	6.3	143	0.225189	0.509306

Figure 27 : Résultat du score thématique via BM25.

6.9.2. Score sociale :

Les médias sociaux et la façon dont ils se rapportent au rang de recherche ont longtemps été un sujet attirant. Bien que le marketing sur les médias sociaux n'ait pas d'impact direct sur les résultats de recherche, il est indéniable que votre présence sur les médias sociaux peut vous aider à obtenir un meilleur classement dans les recherches :

- **Constitution du dataset :** Accéder aux données des médias sociaux via *l'API quintly* et traitez-les d'une manière compatible avec nos objectifs de marketing social et d'affaires. Se connecter par la suite à n'importe quelle application qui bénéficie de ce que nous voulons réaliser. La plupart des plateformes ont pris une longueur d'avance pour exposer leurs API afin que les utilisateurs tirent le meilleur parti des médias sociaux. Cela permet aux développeurs d'exploiter de vastes fonctionnalités de médias sociaux et d'intégrer diverses fonctionnalités dans leurs applications. Dans cette figure, nous examinons une API API.getData des médias sociaux Parmi les plus populaires du marché :

Entrée : tableau de paramètres

Sortie : fichier csv

Début :

Connecter l'API

Data ◄—— API.getData (paramètres)

Ouvrir le fichier csv

Pour chaque ligne de Data **faire**

Sélectionner les signaux sociaux utiles

Ajouter la ligne au fichier csv

Finpour

Fermer le fichier csv

Fin.

Figure 28 : Exploitation de l'API pour la collecte des signaux sociaux.

- Collecte des actions sociales « Twitter » : Pour un film donné, un dataset regroupent les tweets avec les chiffres indicateurs de leurs signaux sociaux:

	post_id	text	favourite_count	retweet_count
0	1428829949285314567	[Police are searching for a gunman who fired s...	25	19
1	1429035562942574596	[How to save big on a 65 LG OLED 4K TV PS Plus...	323	17
2	1428736428461019146	[All these NFT projects are fun but wait till ...	1495	163
3	1426718655337930758	[Happy Birthday to the Bilingual Queen and the...	30421	3401
4	1428733587684220928	[Attention Wizarding World fans Be amazed by t...	49	16
5	1428734637862383620	[the best part abt this video is that chand ka...	96	18
6	1428754155582205956	[Years before Harry Potter Jill Murphy wrote t...	242	23
7	1428285357926268932	[When formed back in 2011 the world was a diff...	1119	374

Figure 29 : Indicateurs des signaux sociaux d'une liste de tweets d'un film donné.

Implémentation de la formule du score sociale en fonction de deux critères sociaux, la réputation et la popularité est comme suit :

```
score_rep = df['favourite_count'].sum() / (df['favourite_count'].sum()+df['retweet_count'].sum())
score_pop = df['retweet_count'].sum() / (df['favourite_count'].sum()+df['retweet_count'].sum())
score_t = score_rep * score_pop
df0bj = df0bj.append({'title': 'harry potter', 'score_s': score_t }, ignore_index=True)
df0bj
```

	title	score_s
0	harry potter	0.095266

Figure 30 : Code source du score social et son résultat.

6.9.3. Score émotionnelle :

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

- **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:

```
# Create a function to clean the tweets
def cleanTxt(text):
    text = re.sub('@[A-Za-z0-9]+', '', text) #Removing @mentions
    text = re.sub('#', '', text) # Removing '#' hash tag
    text = re.sub('RT[\s]+', '', text) # Removing RT
    text = re.sub('https?:\/\/\S+', '', text) # Removing hyperlink
    return text

def getPolarity(text):
    return TextBlob(text).sentiment.polarity

for post in posts:
    tdf["text"] = tdf["text"].apply(cleanTxt)
    tdf["score_e"]=tdf["text"].apply(getPolarity)
    score_e = tdf["score_e"].mean()
```

Figure 31 : Code source du prétraitement du contenu textuel des données en python.

- **Analyse du sentiment avec TextBlob :**

TextBlob est une bibliothèque Python (2 et 3) pour le traitement de données textuelles. Il fournit une API simple pour se plonger dans des tâches courantes de traitement du langage naturel (NLP) telles que le balisage partiel,l'extraction de phrases, l'analyse sentimentale, la classification, la traduction, et plus encore…

```
for post in posts:
    tdf["text"] = tdf["text"].apply(cleanTxt)
    tdf["score_e"]=tdf["text"].apply(getPolarity)
    score_e = tdf["score_e"].mean()
```

	title	score_e
0	harry potter	0.101128

Figure 32 : Estimation du score émotionnel.

6.9.4. Score globale :

Les résultats finaux de notre système RIS ont fourni des preuves théoriques de l'importance de la RI dans le contexte des médias sociaux. Selon les implications de cette étude, les praticiens amélioreraient l'efficacité des RIS:

```
df['score_total'] = (df['score']*0.5)+(df['score_s']*0.25)+(df['score_e']*0.25)
df.sort_values(by=['score_total'],ascending=False)
```

Figure 33 : Code source du score global.

La figure 36 présente les résultats de recherches de films de IMDb classés en ordre décroissant du film le plus pertinent au moins pertinent via une combinaison linéaire de 03 scores (scoreT, scoreS, scoreE):

	id	release_date	title	overview	genres	vote_average	vote_count	popularity	score	score_s	score_e	score_total
0	tt0295297	2002	Harry Potter and the Chamber of Secrets	Ignoring threats to his life, Harry returns to...	[Adventure, Fantasy, Family]	7.4	5966	0.778832	1.000000	0.172840	0.092638	0.566369
2	tt0417741	2009	Harry Potter and the Half-Blood Prince	As Harry begins his sixth year at Hogwarts, he...	[Adventure, Fantasy, Family]	7.4	5435	0.499741	0.951407	0.234375	0.035088	0.543069
4	tt0330373	2005	Harry Potter and the Goblet of Fire	Harry starts his fourth year at Hogwarts, comp...	[Adventure, Fantasy, Family]	7.5	5758	0.652130	0.925095	0.185181	0.098154	0.533381
1	tt0304141	2004	Harry Potter and the Prisoner of Azkaban	Harry, Ron and Hermione return to Hogwarts for...	[Adventure, Fantasy, Family]	7.7	6037	0.745282	0.960986	0.035075	0.094883	0.512982
6	tt0373889	2007	Harry Potter and the Order of the Phoenix	Returning for his fifth year of study at Hogwa...	[Adventure, Fantasy, Family, Mystery]	7.4	5633	0.559462	0.897980	0.073614	0.117324	0.496725
5	tt1201607	2011	Harry Potter and the Deathly Hallows: Part 2	Harry, Ron and Hermione continue their quest t...	[Family, Fantasy, Adventure]	7.9	6141	0.654426	0.917727	0.072587	0.036144	0.486047
3	tt0241527	2001	Harry Potter and the Philosopher's Stone	Harry Potter has lived under the stairs at his...	[Adventure, Fantasy, Family]	7.5	7188	1.000000	0.946900	0.011416	0.028391	0.483402
7	tt0926084	2010	Harry Potter and the Deathly Hallows: Part 1	Harry, Ron and Hermione walk away from their l...	[Adventure, Fantasy, Family]	7.5	5708	0.610161	0.874419	0.095679	-0.193524	0.412748

Figure 34 : Résultats des top films via le score global.

6.10. Page web de recherche du côté client :

Une grande partie de l'Internet est basée sur le modèle client-serveur. Dans ce modèle, les périphériques utilisateurs communiquent par l'intermédiaire d'un réseau avec des serveurs centralisés pour obtenir les données dont ils ont besoin, au lieu de communiquer entre eux. Les appareils de l'utilisateur final tels que les ordinateurs portables, les téléphones intelligents et les ordinateurs de bureau sont considérés comme des clients des serveurs, comme s'ils recevaient des services d'une entreprise.

Les appareils clients envoient aux serveurs des demandes de pages Web ou d'applications, et les serveurs envoient les réponses:

Figure 35 : Modèle client-serveur.

6.10.1.Aperçu de la solution :

Le modèle MVC est composé de trois composantes importantes à savoir le contrôleur, le modèle, et vue. Le composant Controller est responsable de la gestion de l'ensemble de la requête provenant de la vue ou interface utilisateur.

La vue est chargée d'afficher les données obtenues à partir du modèle, et aussi pour prendre l'entrée de l'utilisateur et la transmettre au contrôleur. Le modèle représente un objet entité qui assure le stockage persistant des données obtenues à partir de la couche d'accès aux données.

6.10.2.Exemple de requête :

La figue ci-dessous présente le code source d'une syntaxe de requête détaillée:

```
solr_tuples = [
        # text in search box
        ('q', "Harry Potter"),
        # how many products do I want to return
        ('rows', current_query['rows_per_page']),
        # offset for pagination
        ('start', current_query['start_row'] * current_query['rows_per_page']),
        # example of a default sort,
        # for search phrase leave blank to allow
        # for relevancy score sorting
        ('sort', 'price asc, popularity desc'),
        # which fields do I want returned
        ('fl', 'product_title, price, code, image_file'),
        # enable facets and facet.pivots
        ('facet', 'on'),
        # allow for unlimited amount of facets in results
        ('facet.limit', '-1'),
        # a facet has to have at least one
        # product in it to be a valid facet
        ('facet.mincount', '1'),
        # regular facets
        ('facet.fields', ['gender', 'style', 'material']),
        # nested facets
        ('facet.pivot', 'brand,collection'),
        # edismax is Solr's multifield phrase parser
        ('defType', 'edismax'),
        # fields to be queried
        # copyall: all facets of a product with basic stemming
        # copyallphonetic: phonetic spelling of facets
        ('qf', 'copyall copyallphonetic'),
        # give me results that match most fields
        # in qf [copyall, copyallphonetic]
        ('tie', '1.0')
        # format response as JSON
        ('wt', 'json')
    ]
```

Figure 36 : Exemple de requête Solr crée.

6.10.3.Interface des résultats de notre SRI via la requête "Harry Potter" :

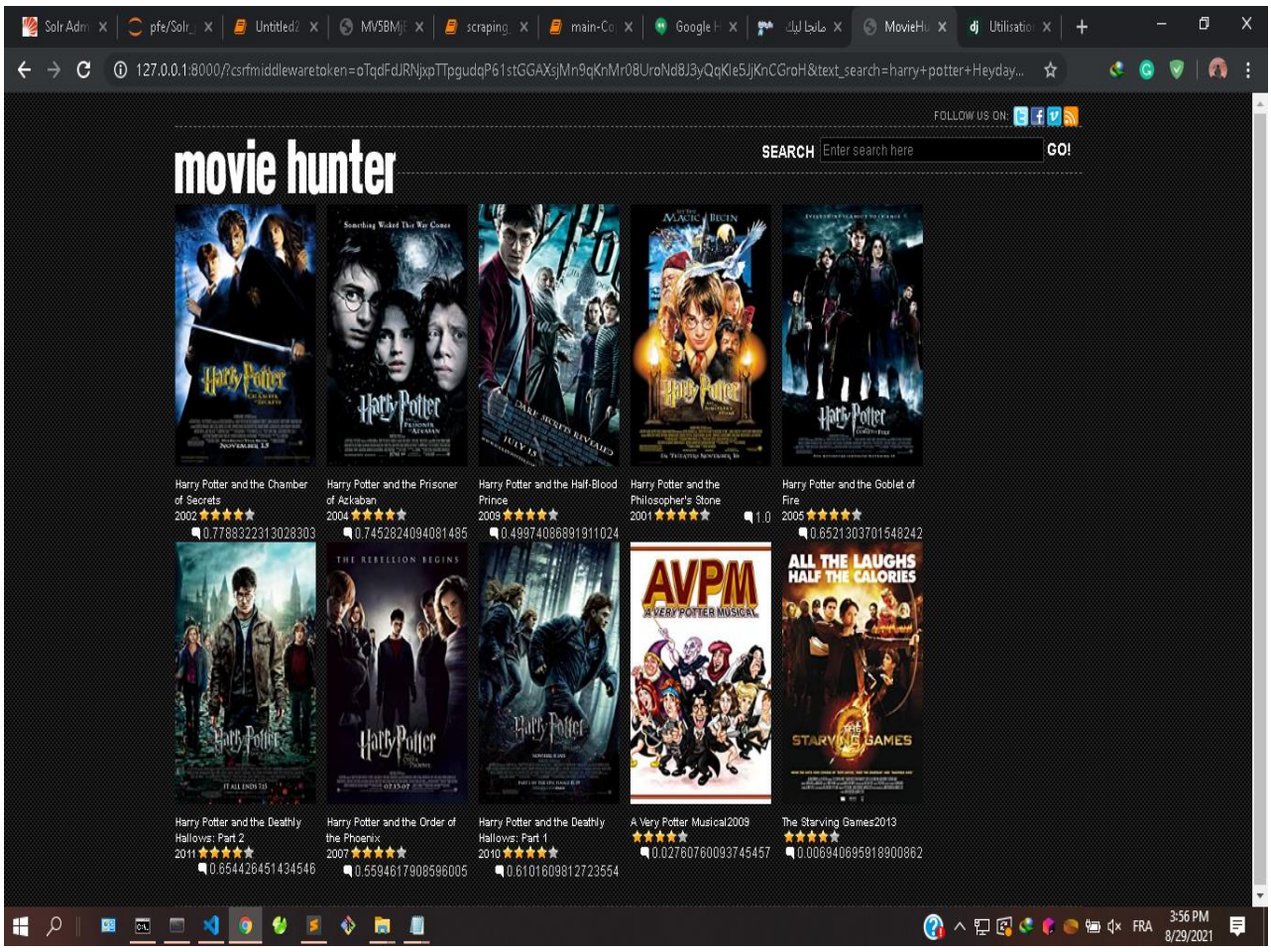

Figure 37 : Résultats de la recherche du film de Harry Potter.

6.10.4.Evaluation des résultats de recherche de notre SRI :

La table 03 ci-dessous représente les résultats de recherche des tops 10 meilleurs films classés en ordre décroissant en fonction de 02 mesures de similarité : TF*IDF et BM25 pour le classement thématique et score global combiné pour le classement Socio-Emo-Thématique. Nous concluons que les contenus sociaux qui représentent les likes, retweets et polarité des tweets ont bien reclassé les documents, où nous constatons une différence dans l'ordre des films restitués pour les deux approches : la RI classique et la RIS.

Ces Résultats confirment l'impact des interactions sociales sur la RI classique.

RI		RIS
Tf-idf	BM25	
A Very Potter Musical	Harry Potter and the Chamber of Secrets	Harry Potter and the Chamber of Secrets
Harry Potter and the Philosopher's Stone	Harry Potter and the Prisoner of Azkaban	Harry Potter and the Half-Blood Prince
Harry Potter and the Chamber of Secrets	Harry Potter and the Half-Blood Prince	Harry Potter and the Goblet of Fire
Harry Potter and the Prisoner of Azkaban	Harry Potter and the Philosopher's Stone	Harry Potter and the Prisoner of Azkaban
Harry Potter and the Goblet of Fire	Harry Potter and the Goblet of Fire	Harry Potter and the Order of the Phoenix
Harry Potter and the Half-Blood Prince	Harry Potter and the Deathly Hallows: Part 2	Harry Potter and the Deathly Hallows: Part 2
The Starving Games	Harry Potter and the Order of the Phoenix	Harry Potter and the Philosopher's Stone
Harry Potter and the Deathly Hallows: Part 2	Harry Potter and the Deathly Hallows: Part 1	Harry Potter and the Deathly Hallows: Part 1
Harry Potter and the Order of the Phoenix	A Very Potter Musical	A Very Potter Musical
Harry Potter and the Deathly Hallows: Part 1	The Starving Games	Miss Potter
A Very Potter Musical	Miss Potter	The Starving Games

Table 3 : Top 10 des films IMDB

6.11.Conclusion:

Dans ce chapitre, nous avons décrit la mise en œuvre des expérimentations menées pour valider notre travail. Nos expérimentations nous ont permis d'apprécier la performance et l'amélioration de la recherche d'information sociale en incorporant l'émotion.

CONCLUSION GÉNÉRALE

Le Web 2.0 a conduit à l'émergence des contenus sociaux générés par les utilisateurs (UGC) dans les services sociaux sur Internet. Ces UGC sont généralement évolutifs et de nature différente : des annotations sociales, des clics, des tweets, des commentaires, des relations sociales, des actions relevant d'activités sociales telles que le j'aime, le partage, le +1, le rating, etc. Les utilisateurs interagissent de plus en plus entre eux et/ou avec les ressources. Ces interactions associées aux ressources peuvent être considérées comme une des sources que l'on peut également exploiter pour améliorer la RI.

Pour ce présent mémoire de Master2, nous avons proposé un modèle de recherche d'information basé sur les contenus sociaux de Twitter. Ces contenus sont considérés comme une information additionnelle permettant de mesurer la pertinence globale de la ressource à laquelle ils sont associés. Cette pertinence globale améliore les résultats de recherche en combinant linéairement trois pertinences : thématique, sociale et émotionnelle. En effet, Ces résultats de recherche ont confirmé l'impact des interactions sociales sur le processus de RI classique.

Cependant, notre travail présente quelques limites. D'abord, nous avons considéré que les signaux sont tous de même importance. Ils ne se différencient que par leur nombre vis-à-vis de la ressource correspondante. Selon nos résultats, il semblerait que certains soient plus important que d'autres pour la recherche d'information. Ensuite, nous n'avons pas pu évaluer l'impact d'autres signaux et les auteurs des différents signaux, sur le processus de RI. La récupération de ces informations n'est pas accessible via les APIs d'autres réseaux sociaux actuels.

Une autre limite de notre travail réside dans la non prise en compte des facteurs temporels (temporalité des signaux, date de publication de la ressource, date du signal) ou des facteurs imagerie (images, émojis...). Nous pensons qu'un comptage simple de la quantité des signaux associés à une ressource privilégieront les ressources anciennes.

Nous traitons plus finement ces aspects dans le futur Inchaâllah où nous envisagerons inclure le domaine de la « *Learning Machine* » dans la recherche d'information sociale.

Références

[1] C. F. M. M. Agosti M., «Design and implementation of a tool for the automatic construction of hypertexts for information retrieval,» *Information Processing & Management,* vol. 32, n° %14, pp. 459-476, 1996.

[2] W. o. crowds, «investopedia,» [En ligne]. Available: https://www.investopedia.com/terms/w/wisdom-crowds.asp.

[3] M. Sanderson et C. W. Bruce, The History of Information Retrieval Research, 1940.

[4] R. R. Shaw, «The Rapid Selector,» *Journal of Documentation,* vol. 5, n° %13, pp. 164 - 171, 1949.

[5] «New Tools for the Resurrection of Knowledge,» *Chemical and Engineering News,* vol. 32, n° %19, pp. 866-869, 01-1954.

[6] P. Switzer, «Vector Images in Document Retrieval,» *Harvard University,* n° %14, 01-1963.

[7] G. Salton, Automatic information organization and retrieval, McGraw Hill Text, 1968.

[8] S. E. Robertson, «The probability ranking principle in IR,» *Journal of documentation,* vol. 33, n° %14, pp. 294-304, 1977.

[9] S. T. D. G. W. F. T. K. L. a. R. H. S. Deerwester, «Indexing by latent semantic analysis,» *Journal of the American society for information science,* vol. 41, n° %16, pp. 391-407, 1990.

[10] «librarianshipstudies,» [En ligne]. Available: https://www.librarianshipstudies.com/2020/02/information-retrieval.html.

[11] G. Chowdhury, Introduction to Modern Information Retrieval, 2017.

[12] P. Ingwersen, «Poly representation of information needs and semantic entities: elements of cognitive theory for information retrieval interaction,» *In proceedings of the Seventeenth Annual International ACM SIGIR Conference of Research and Development in Information Retrieval,* pp. 101-110, 1994.

[13] K. Mechach, Etude de l'impact des méthodes de localisation dans les systèmes d'information distribués, Oran : Université 1 Ahmed Ben Bella, 2016.

[14] G. (. Salton, The SMART Retrieval System. Experiments in Automatic Document Processing. Prentice-Hall, Englewood Cliffs (NJ), Prentice-Hall: Englewood Cliffs (NJ), 1971.

[15] E. G. J Savoy, Information Retrieval, Indurkhya: N., & Damerau, FJ (Eds.), 2010.

[16] I. BADACHE, *Exploitation des Signaux Sociaux pour Am´eliorer la Recherche d'information,* Toulouse: Université Paul Sabatier, 2016.

[17] M. d. RI, «link.springer,» [En ligne]. Available: https://link.springer.com/referenceworkentry/10.1007%2F978-1-4614-8265-9_916. [Accès le 18 08 2021].

[18] C. Cleverdon, «Optimizing convenient on-line access to bibliographic databases,» *Information Service & Use,* vol. 4, pp. 37-47, 1984.

[19] N. Ismail, Contribution à l'analyse et à la recherche d'information en texte intégral : Application de la transformée en ondelettes pour la recherche et l'analyse de texte, 2010.

[20] S. G. a. J. M. Michael., Introduction to modern information retrieval, 1983.

[21] I. Bourbie, La recherche des commentaires pertinentes, Tiaret: Université Ibn Khaldoun, 2020.

[22] S. Robertson, «The Probability Ranking Principle in IR,» *Journal of Documentation,* n° %133, pp. 294-304, 1977.

[23] B. C. V. E.M:, «Retrieval system evaluation.,» *In E.M. Voorhees The MIT Press,* pp. 53-75, 2005.

[24] A. H. e. H. Dridi, Recherche d'Information Sociale Reclassement des résultats de recherche à base de pertinence sociale, 2012.

[25] F. K. H. G. Abir Gorrab, «Social Information Retrieval and Recommendation,» *Revue Africaine de la Recherche en Informatique et Mathématiques,* vol. 27, 2019.

[26] S. media, «investopedia,» [En ligne]. Available: https://www.investopedia.com/terms/s/social-media.asp. [Accès le 08 08 2021].

[27] S. web, «techopedia,» [En ligne]. Available: https://www.techopedia.com/definition/30514/social-web. [Accès le 18 08 2021].

[28] R. sociaux, «sciencedirect,» [En ligne]. Available: https://www.sciencedirect.com/topics/social-sciences/social-networking-sites. [Accès le 18 08 2021].

[29] Facebook, «GCFGlobale,» [En ligne]. Available: https://edu.gcfglobal.org/en/facebook101/what-is-

facebook/1/. [Accès le 18 08 2021].

[30] Youtube, «GCFGlobale,» [En ligne]. Available: https://edu.gcfglobal.org/en/youtube/what-is-youtube/1/.

[31] Twitter, «techno science,» [En ligne]. Available: https://www.techno science.net/glossairedefinition/Twitter.html.

[32] Linkedin, «businessinsider,» [En ligne]. Available: https://www.businessinsider.fr/us/what-is-linkedin.

[33] J. L. a. T.-S. C. Marie-Francine Moens, Mining User Generated Content, CRC press, 2014.

[34] S. W.-V. a. G. Vickery., «Participative Web: User created content. Technical report, Directorate for Science,» *Technology and Industry: Committee for Information, Computer and Communications Policy,* April 2007.

[35] O. A. a. V. Kandylas., A study on placement of social buttons in web pages, 2014.

[36] D. R. a. W. C. D.R. Pierce, Social influences on online political information search and evaluation. Political Behavior, vol. 39, 2017, p. 651–673.

[37] R. H. K. W. C. R. a. W. A. W. Lewoniewski, Application of seo metrics to determine the quality of Wikipedia articles and their sources. In Robertas Damaševicius and Giedr Vasiljevien editors, Information and Software Technologies, Springer International Publishing éd., 2018, p. 139–152.

[38] «searchmetrics,» [En ligne]. Available: https://www.searchmetrics.com/knowledge-base/rankingfactors-infographic-2016/ .

[39] I. Badache., Learning social signals for predicting relevant search results. Web Intelligence and Agent Systems, IOS Press, 2020, pp. 15-33.

[40] S. S. a. C. S. A. Khodaei, Personalization of Web Search Using Social Signals, pages 139–163. Springer International Publishing, CHAM, 2015.

[41] M. Raza., A new level of social search: Discovering the user's opinion before he make one. Technical report, Cambridge,, UK: Mi-crosoft Research Cambridge..

[42] S. Z. a. N. Cabage., Does SEO matter? Increasing classroom blog visibility through search engine optimization. In 46th Hawaii International Conference on System Sciences, HI, USA,: HICSS, 2013, p. 1610–1619.

[43] J. T. M. M. a. D. L. B. Hecht, «Searchbuddies: Bringing search engines into the conversation.,» *In*

Proceedings of the Sixth International Conference on Weblogs and Social Media, p. 138–145, 2012.

[44] S. F. N. T. U. S. Dion Goh, Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively.

[45] C. Bouhini., Impact des réseaux sociaux sur le processus de recherche d'information. Réseaux sociaux et d'information [cs.SI], Saint-Etienne: Université Jean Monnet, 2014.

[46] N. H. D. K. a. M. K. F. Abel, On the Effect of Group Structures on Ranking Strategies in Folksonomies, Springer Berlin Heidelberg, Berlin , p. 275–300.

[47] G. a. J. D. Sure York Stumme, «Information retrieval in folksonomies: Search and ranking . In The Semantic Web: Research and Applications,» *3rd European Semantic Web Conference,* p. 411–426, June 11-14, 2006.

[48] G. X. X. W. Y. Y. B. F. a. Z. S. S. Bao, «Optimizing web search using social annotations,» *In Proceedings of the 16th international conference on World Wide Web,* p. 501–510, 2007.

[49] A. J. N. S. a. T. K. Y. Yusuke, «Can social bookmarking enhance search in the web?,» *In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries,* p. 107–116, 2007.

[50] C. O.-R. a. I. A. S. Chelaru, How useful is social feedback for learning to rank YouTube videos? World Wide Web, vol. 17, Sep 2014, p. 997–1025.

[51] C. H. a. K. B. B. Karweg, «Evolving social search based on bookmarks and status messages from social networks,» *In Proceedings of the 20th ACM International Conference on Information and Knowledge Management,* vol. 11, p. 1825–1834, 2011.

[52] A. K. a. C. Shahabi., «Social-textual search and ranking,» *In Proceedings of the First International Workshop on Crowdsourcing Web Search,* p. 3–8, 2012.

[53] O. D. a. B. D. L. Hong, «Predicting popular messages in twitter,» *In Proceedings of the 20th International Conference Companion on World Wide Web, WWW ,* vol. 11, p. 57–58, 2011.

[54] S. J. a. S. S. P. L. Chan, «Surfacing social signals in google scholar search,» *In Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion,* n° %117, p. 17–20, 2016.

[55] Y. L. a. Y. X. K. Albishre, «Query-based automatic training set selection for microblog retrieval,» *In Advances in Knowledge Discovery and Data Mining,* p. 325–336, 2018.

[56] M. M. a. J. Teevan, «Exploring the complementary roles of social networks and search engines,» *In Microsoft Research, Human-Computer Interaction Consortium Workshop (HCIC),* pp. 1-10, 2012.

[57] M. G. O. A. a. K. H. P. Pantel, «Social annotations: Utility and prediction modeling,» *In Proceedings of the 35th SIGIR Conference on Research and Development in Information Retrieval,* p. 285–294, 2012.

[58] J. T. M. M. a. D. L. B. Hecht, «Searchbuddies: Bringing search engines into the conversation,» *In Proceedings of the Sixth International Conference on Weblogs and Social Media,* p. 138–145, 2012.

[59] X. Z. H. C. J. K. a. C. G. L. Gou, «Social network document ranking,» *In Proceedings of the 10th Annual Joint Conference on Digital Libraries,* vol. 10, p. 313–322, 2010.

[60] I. Badache, «intégration de propriétés sociales dans un modèle de recherche,» *n Conférence francophone en Recherche d'Information et Applications,* p. 1–6, 2013.

[61] I. B. a. M. Boughanem, «Exploitation de signaux sociaux pour estimer la pertinence a priori d'une ressource,» *In Conférence francophone en Recherche d'Information et Applications,* p. 163–178, 2014.

[62] I. B. a. M. Boughanem, «Social priors to estimate relevance of a resource,» *In Proceedings of the 5th Information Interaction in Context Symposium,* n° %114, p. 106–114.

[63] I. B. a. M. Boughanem, «Emotional social signals for search ranking,» *In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval,* n° %117, p. 1053–1056, 2017.

[64] I. B. a. M. Boughanem, Les Signaux Sociaux émotionnels : Quel impact sur la recherche d'information ?' In Conférence en Recherche d'Information et Applications, Marseille, France, March 2017.

[65] I. Badache, «'Users' traces for enhancing Arabic Facebook search,» *In Proceedings of the 30th ACM Conference on Hypertext and Social Media,* n° %119, p. 241–245, 2019.

[66] I. Badache, «A. Abu-Thaher, M. Hamdan, and L. Abu-Jaish. 'Social Information Retrieval in Arabic Language: Case of Facebook,» *In Conférence en Recherche d'information et Applications,* March 2019.

[67] I. Badache, «Exploring differences in the impact of users' traces on Arabic and English Facebook search,» *In IEEE/WIC/ACM International Conference on Web Intelligence,* n° %119, p. 225–232, 2019.

[68] C. Elkan, « The foundations of cost-sensitive learning. In International joint conference on artificial intelligence,» *Lawrence Erlbaum Associates Ltd.,* vol. 17, n° %11, pp. 973-978, 2001, August).

[69] M. &. L. B. Hu, «Mining and summarizing customer reviews,» *n Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining ,* pp. 168-177, (2004,

August).

[70] B. Liu, Sentiment Analysis Mining Opinions, Sentiments, and Emotions, Chicago: University of Illinois, 2015.

[71] A. Ziani, La recommandation via l'analyse d'opinions, Annaba : Université Badji Mokhtar, 2018.

[72] M. R. a. J. Gama, «The Role of Sentiment Analysis,» *FEP Economics and,* 2013.

[73] B. X. I. V. O. R. a. R. P.-. n. Apoorv Agarwal, «Sentiment analysis of Twitter data,» *Proceedings of the Workshop on Languagesin Social Media,* vol. 11.

[74] L. Dijoux, Boostez votre business avec Twitter, Almabic, 2009.

[75] F. Colantonio, Communication professionnelle en ligne: comprendre et exploiterles médias et réseaux sociaux, Edipro, 2011.

[76] T. O. a. S. Milstein, The Twitter Book, UK: Angham B3 2PB, UK, 2.

[77] S. E. H. e. K. SAIFIA, Analyse des sentiments cas Twitter, Ghardaia: Université de Ghardaia, , 2015.

[78] TF-IDF_Vectorizer, «scikit-learn,» [En ligne]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

[79] «pypi,» [En ligne]. Available: https://pypi.org/project/rank-bm25/.

[80] «Solr Guide,» [En ligne]. Available: https://solr.apache.org/guide/.

[81] «Apache ZooKeeper Guid,» [En ligne]. Available: https://zookeeper.apache.org/doc/r3.3.3/zookeeperStarted.html.

[82] «Anaconda,» [En ligne]. Available: https://docs.anaconda.com/anaconda/user-guide/index.html.

[83] «Pycharm,» [En ligne]. Available: https://www.jetbrains.com/help/pycharm/quick-start-guide.html.

[84] «Jupyter,» [En ligne]. Available: https://jupyter.org/documentation.

[85] «Python,» [En ligne]. Available: https://docs.python.org/3/tutorial/.

[86] «Django,» [En ligne]. Available: https://www.djangoproject.com/start/ .

[87] «Pysolr,» [En ligne]. Available: https://pypi.org/project/pysolr/.

[88] «Tweepy,» [En ligne]. Available: https://docs.tweepy.org/en/stable/.

[89] «Pandas,» [En ligne]. Available: https://pandas.pydata.org/docs/ .

[90] «numpy,» [En ligne]. Available: https://numpy.org/doc/.

[91] «nltk guid,» [En ligne]. Available: https://www.nltk.org/.

[92] «TextBlob,» [En ligne]. Available: https://textblob.readthedocs.io/en/dev/.

[93] «Curl,» [En ligne]. Available: https://curl.se/docs/manpage.html.

[94] «BeautifulSoup,» [En ligne]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[95] «Requests,» [En ligne]. Available: https://docs.python-requests.org/en/master/.

[96] J. R. S. a. C. J. S. Matthew Eric Glassman, «Social Net- workingand Constituent Communications: Members Use of Twitter and Facebook Dur- ing aTwo-Month Period in the 112th Congress,» *Congressional Research Service,* 2009.

[97] K. Ela, Natural Language Processing, India: I.K.International Publishing HousePvt. Ltd , 2011.

[98] J. Véronis, «Natural Language Processing,» [En ligne]. Available: http://sites.univ-provence.fr/veronis.