



REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE  
MINISTERE DE L'ENSEIGNEMENT SUPERIEURET DE LA RECHERCHE  
SCIENTIFIQUE

## **UNIVERSITE IBN KHALDOUN-TIARET**

### **MEMOIRE**

Présenté à:

FACULTÉ MATHÉMATIQUES ET INFORMATIQUE  
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

### **MASTER**

Spécialité :

Génie Informatique

Par:

**MELIANI HAMIDA**  
**HAMZI NESRINE**

---

Sur le thème

## **Approche de profilage sur les réseaux sociaux**

---

Soutenu publiquement le.27/06/2022 à Tiaret devant le jury composé de:

Mr.ABID Khaled	Grade M.A.A	U.I.K.Tiaret	Président
Mme.LAKHDARI Aicha	Grade M.A.A	U.I.K.Tiaret	Encadreur
Mr.BENOUDA Habib	Grade M.A.A	U.I.K.Tiaret	Examineur

2021-2022

## **DEDICACES**

*Je dédie ce travail :*

*A mon père;*

*A ma mère, qu'Allah la protège ;*

*A mes frères et sœurs,*

*A tous mes amis sans citer leurs noms, sinon la liste sera très longue.*

**HAMIDA**

## **DEDICACES**

*Je dédie ce travail :*

*A mon père;*

*A ma mère, qu'Allah la protège ;*

*A mes frères et sœurs,*

*A tous mes amis sans citer leurs noms, sinon la liste sera très longue.*

**NESRINE**

## **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 gratitude 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 ABID Khaled 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 BENOUDA Habib 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 Informatique 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

With the increasing number of research works in the field of social information retrieval (SIR), there is a need to build a clearer picture of the objective to be achieved. This is done, mainly, by improving the process of traditional IR, more specifically the ranking of results, by exploiting user-generated content ( UGCs ) from social networks. These UGCs represent added value, whether for the creation of document profiles (raw profile) or for enriching the content and representation of documents (social profile). These UGCs will thus be included in the calculation of relevance.

**Keywords:** SIR, Social networks, UGCs , Profile

# Résumé

La recherche d'informations sociales (RIS) vise à intégrer la dimension sociale dans le processus de recherche. Les approches de la RIS sont diverses. Ils sont principalement basés sur l'identification et l'intégration d'informations sociales dans le processus de recherche. En fait, plusieurs types d'informations sociales sont utilisées dans les travaux de RI sociale. Notre travail s'articule autour de la problématique d'accès et d'identification des informations sociales et de leur exploitation, d'une part, dans la construction d'un profil document enrichi d'une dimension sociale, et d'autre part, dans le calcul de la pertinence.

**Mots clefs :** RIS, information sociale, Profil document, pertinence

## TABLE DES MATIÈRES :

Introduction générale.....	1
CHAPITRE I : De la RI classique à la RISociale.....	3
1-Introduction.....	3
2-Contexte .....	4
2.1-Définition de RI .....	4
2.2-Concepts de base de la RI traditionnelle.....	4
2.3-Processus générale de la RI .....	5
2.4-Indexation .....	6
2.5-Appariement document-requête.....	8
2.6-Reformulation de la requête .....	13
3-Réseaux sociaux.....	14
4-Recherche d'informations sociales.....	14
4.1-Définition .....	15
4.2-Approches RIS .....	16
5-Conclusion .....	20
CHAPITRE II : Généralités sur l'Analyse des Sentiments (AS).....	21
11. Introduction.....	21
2. Notions de base.....	21
2.1-Sentiment.....	21
2.2-Émotion.....	21
2.3-Opinion.....	22
3-Analyse des sentiments SA.....	23
3.1-Définition.....	23
3.2-Taches de SA .....	23
3.3 Types de SA.....	24
4-Niveaux de SA.....	24
4.1-Niveau Document.....	25
4.2-Niveau Phrase.....	25
4.3-Niveau aspect (entité).....	25
5. Domaines d'application d'AS.....	25

5.1-Domaine du commerce .....	26
5.2-Domaine de santé .....	26
5.3-Détection de spam d'opinion.....	26
5.4-Domaine politique.....	26
6-Problèmes de SA.....	26
6.1-Détection de la subjectivité.....	26
6.2-Sentiment implicite.....	27
6.3-Dépendance du domaine.....	27
6.4-Identification d'entité.....	27
6.5-Négation.....	27
7-Algorithmes d'analyse des sentiments.....	28
7.1-L'approche automatique.....	28
7.2-Approche à base de règles (Rule-based) .....	28
7.3-Approche hybride.....	29
8-L'analyse des sentiments avec Twitter .....	29
8.1-Outils d'analyse de sentiments .....	30
9-Conclusion.....	32
CHAPITRE III : Notre processus RIS.....	33
1-Introduction.....	33
2-Architecture générale de notre système .....	33
2.1-Collection des Tweets Covid'19 .....	34
2.2-Prétraitement(Preprocessing).....	35
2.3-Indexation :.....	36
2.4-Appariement document-requête.....	36
3-Dimension sociale .....	37
4-profile tweet .....	37
4.1-Pertinence classique .....	38
4.2-Pertinence sociale.....	39
4.3-Pertinence émotionnelle .....	40
5-Pertinence globale .....	40
6-Conclusion .....	40



CHAPITRE VI : Implémentation.....	41
1-Introduction .....	41
2-Présentation de l'environnement utilisé.....	41
2.1-Outils de développement .....	42
3-Création du dataset .....	43
3.1-Enregistrement des données .....	44
4-Prétraitement des données.....	45
5-Pertinence thématique.....	49
5.1-La similarité BM25 .....	50
6-Résultats et retournes .....	51
6.1-Score thématique .....	51
6.2-Score sociale .....	51
6.3-Score émotionnelle .....	52
6.4-Score globale .....	53
7- conclusion.....	54
8-Conclusion générale .....	55
REFERENCES BIBLIOGRAPHIQUES.....	56

## LISTE DES FIGURES :

Figure 1: Processus en U de la recherche d'information.....	06
Figure 2 : les modèles mathématique de RI.....	09
Figure 3: Représentation algébrique des documents et des requêtes dans l'espace des termes à deux dimensions.....	11
Figure 4 : une taxonomie pour les modèles de RIS .....	15
Figure 5 : Exploitation d'information sociale dans la RI .....	17
Figure 6 : Roue des émotions de Plutchik.....	22
Figure 7 : Classification de l'orientation sémantique.....	24
Figure 8 : les niveaux d'analyse de sentiment .....	25
Figure 9 : l'approche automatique .....	28
Figure 10 : Approche à base de règle .....	29
Figure 11 : Diagramme montrant la structure de SentiWordNet .....	31
Figure 12 : Architecture générale de notre SRI .....	34
Figure 13 : Collecte des tweets via l'API-TWITTER.....	35
Figure 14 : traitement d'un twitte .....	36
Figure 15 : Formule Okapi BM25.....	38
Figure 16 : Formule idf .....	38
Figure 17 : Time line de notre SRI .....	40
Figure 18 : Code source python de création du dataset .....	45
Figure 19 : Fichier csv des tweets collectés.....	45
Figure 20 : prétraitement du texte (Tweet) .....	46
Figure 21 : Prétraitement détaillé de texte .....	46
Figure 22 : suppression des balises de chaine de caractères s.....	47
Figure 23 : Remplacement des caractères de ponctuation par des espaces dans les chaines de caractères s.....	47
Figure 24 : Suppression des caractères d'espacement répétitifs .....	48
Figure 25 : retirer les chiffres de chaine de caractères .....	48
Figure 26 : Suppression des mots d'arrêt de chaine de caractères .....	49

Figure 27 : Suppression de chaîne de caractères des mots dont la longueur est inférieure à minsize .....	49
Figure 28 : Transformation des chaînes de caractères en minuscules .....	50
Figure 29 : dataset indexé .....	50
Figure 30 : Prétraitement du côté module requête .....	51
Figure 31 : Exploitation de la mesure de similarité BM25 Rank .....	51
Figure 32 : Résultat du score thématique via BM25.....	52
Figure 33 : Code source et résultat du classement social .....	52
Figure 34 : utilisation de bibliothèque de textblob .....	52
Figure 35 : Estimation du score émotionnel .....	53
Figure 36 : Code source du score global .....	53
Figure 37 : Résultats des tweets via le score global .....	54

# Liste des abréviations

**IR** Information retrieval  
**SIR** Social Information Retrieval  
**IRS** Information Retrieval System  
**TF** Term Frequency  
**IDF Inverse Document Frequency**  
**UGC** User Generated Content  
**URL** Uniform Resource Locator  
**SA** Sentiments Analysis  
**API** Application Programming Interface  
**BM25** Best Matching 25  
**CSV** Comma Separated Values  
**NLP** Natural Language Processing  
**NLTK** Natural Language Toolkit  
**HTTP** Hypertext Transfer Protocol  
**IMDB** Internet Movies Database  
**JSON** JavaScript Object Notation

# INTRODUCTION

# Introduction générale

La recherche d'information classique est un processus atomique, en un seul coup "soumettre une requête, obtenir des résultats" mais elle est naturellement composée de deux tâches importantes : (1) récupérer les documents pertinents pour une requête (correspondance documents-requête) ; et (2) classer ces documents en fonction de leur pertinence par rapport à la requête. L'ordonnement des résultats de recherche a toujours été un problème important dans la recherche d'informations (RI). La capacité de classer correctement les résultats de recherche a longtemps été préconisée comme un avantage des modèles de RI traditionnels. L'objectif de l'ordonnement est de trier les résultats par ordre décroissant de leur pertinence par rapport à la requête et les résultats situés près du haut de la liste retournée ont une probabilité beaucoup plus élevée d'être examinés par l'utilisateur. En effet, les décisions/jugements de l'utilisateur final sur les résultats récupérés sont significativement influencés par le rang/position dans le classement.

Cependant, les approches de classement traditionnelles, les systèmes de recherche à petite échelle, sont basées uniquement sur le contenu des documents. Dans cette situation, le classement échouait systématiquement à identifier les documents les plus pertinents. C'est un contexte parfait pour passer à exploiter les traces laissées par les internautes du Web 2.0, informations sociales.

Le nouveau processus de recherche « Recherche d'Information Sociale ou RIS » est présenté comme une séquence d'actions à effectuer à la fois par l'utilisateur et le système de recherche : « avant la recherche », « pendant la recherche » et « après la recherche ». Ce processus est généralement considéré selon les deux axes de recherches suivants:

*Axe 1* : La définition des approches et de modèles de RI spécifiques pour rechercher de nouveaux types de contenus (la recherche dans les sites des médias sociaux, la recherche des actualités, des vidéos, d'opinion, ...etc).

*Axe 2* : L'exploitation des contenus générés par les utilisateurs (UGC) des divers

réseaux sociaux pour améliorer la RI. Ces UGC peuvent être intégrés en amont, au sein ou en aval du processus de recherche en tant que source d'information additionnelle pour améliorer la pertinence des résultats.

Les UGCS vont des requêtes et des clics, aux votes et aux balises, aux commentaires et aux liens sociaux, et tout le reste. Leur dénominateur commun est l'utilisation de toutes les traces d'informations laissés par les utilisateurs de divers réseaux sociaux en tant que valeur complémentaire de leurs comportements naturels en RI. Les informations sociales sont apparues donc comme une passerelle capable de connecter le domaine de la RI classique aux réseaux sociaux.

Maintenant, les questions à se poser :

(1) Comment identifier et exploiter les informations sociales (UGC) pour construire le profil document ?,

(2) Comment les quantifier et les intégrer pour récupérer et classer les résultats de recherche ?

et (3) est ce qu'elles influencent le classement des résultats de recherche ?

Dans ce présent travail, nous nous focalisons sur l'impact de l'information sociale sur le processus de RI. L'objectif étant orienté document, l'information sociale est ainsi exploitée, d'une part, pour construire son profil social décrivant les réactions pratiques (Favorite, Retweet) et émotionnelles (sentiment) de l'utilisateur, et d'autre part, pour le classement des résultats de recherche en combinant les pertinences thématique-socio-émotionnelle.

Nos expérimentations sont menées sur une collection de documents bâtie à partir du big-social-data Twitter.

Le reste de ce mémoire est organisé comme suit. L'objectif du chapitre 1 « De la recherche d'informations classique (RI) à la recherche d'information sociale (RIS) » est de présenter les principes de la recherche d'information dans des contenus textuels, puis son application à l'environnement social. Le chapitre 2 présente une revue de

littérature sur « l'analyse des sentiments ». Ceci comprend un survol théorique sur les concepts de base et leurs caractéristiques. Le chapitre 3 décrit notre solution RIS à l'identification des facteurs sociaux-émotionnels pour construire le profil document et son intégration dans le processus de RI. Le chapitre 4 est réservé pour l'implémentation et la mise en œuvre de notre solution RIS. Enfin, la conclusion générale tire des orientations futures .



# Chapitre I

De la RI classique à la  
RI sociale

### 1- Introduction

Avec l'émergence du Web social, le Web est passé d'un Web statique, où les utilisateurs ne pouvaient consommer que de l'information, à un Web où les utilisateurs sont également en mesure de produire de l'information. Cette évolution est communément appelée Web social ou Web 2.0. Ce dernier a introduit une nouvelle liberté d'interactions de l'internaute en facilitant ses relations avec d'autres utilisateurs qui ont des préférences similaires ou partagent les mêmes ressources.

Les réseaux sociaux sont certainement les technologies les plus adoptées dans cette nouvelle ère. Ces plateformes sont couramment utilisées pour communiquer avec d'autres utilisateurs, échanger des messages, partager des ressources (photos et vidéos), commenter des publications, créer et mettre à jour des profils, interagir et jouer à des jeux en ligne, etc. Ces tâches d'interaction qui rendent les utilisateurs plus actifs dans la génération de contenu sont parmi les facteurs les plus importants pour la prolifération croissante des données.

Dans un tel contexte, un problème crucial est de permettre aux utilisateurs de trouver des informations pertinentes en ce qui concerne leurs intérêts et leurs besoins. Cette tâche est communément appelée recherche d'information (R). La RI est effectuée tous les jours d'une manière évidente sur le Web, généralement à l'aide d'un moteur de recherche. Cependant, les modèles traditionnels de la RI ne tiennent pas compte du facteur social du Web. Ensuite, leurs algorithmes de classement sont souvent fondés sur (i) une similarité de contenu de requête et de document et (ii) les liens hypertextes existants qui relient ces pages Web[1]. Par conséquent, les modèles classiques de RI doivent être adaptés au nouveau paradigme afin d'exploiter ce facteur social qui entoure les pages Web et les utilisateurs. En effet, les ressources sociales sont souvent accessibles, car la plupart des réseaux sociaux fournissent des API pour accéder à leurs données (même si souvent, un contrat monétisé doit être établi avant toute utilisation à grande échelle).

Il y a actuellement un certain nombre de travaux de recherche entrepris pour améliorer le processus de RI en exploitant les réseaux sociaux. C'est ce qu'on appelle communément la « recherche d'information sociale (SIR) ». Le présent chapitre vise à fournir une

compréhension claire des divers efforts déployés dans le domaine de SIR .

### **2-RI Classique**

Salton [2] et Baeza et al. [3] ont défini la RI comme suit :

La recherche d'information (RI) est la science qui traite de la représentation, du stockage, de l'organisation et de l'accès aux éléments d'information afin de satisfaire aux exigences des utilisateurs concernant cette information.

Un système de RI (IRS) est évalué en fonction de sa précision et de sa capacité à extraire des informations pertinentes, ce qui maximise la satisfaction des utilisateurs, c.-à-d. que plus les réponses correspondent aux attentes des utilisateurs, meilleur est le système.

#### **2.1-Concepts de base de la RI traditionnelle**

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

##### **2.1.1-Collection de documents**

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

##### **2.1.2-Document**

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

##### **2.1.3-Besoin en information**

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

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

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

## Chapitre I : De la RI classique à la RI sociale

---

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

### 2.1.4-Requête

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

### 2.1.5-Pertinence

La pertinence est une notion fondamentale dans le domaine de la RI. La pertinence peut être définie comme la correspondance entre un document et une requête, ou encore une mesure d'informativité du document à la requête [6]. Ces différents critères ont amené à la catégorisation de la pertinence utilisateur principalement en 5 classes de pertinence [7] :

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

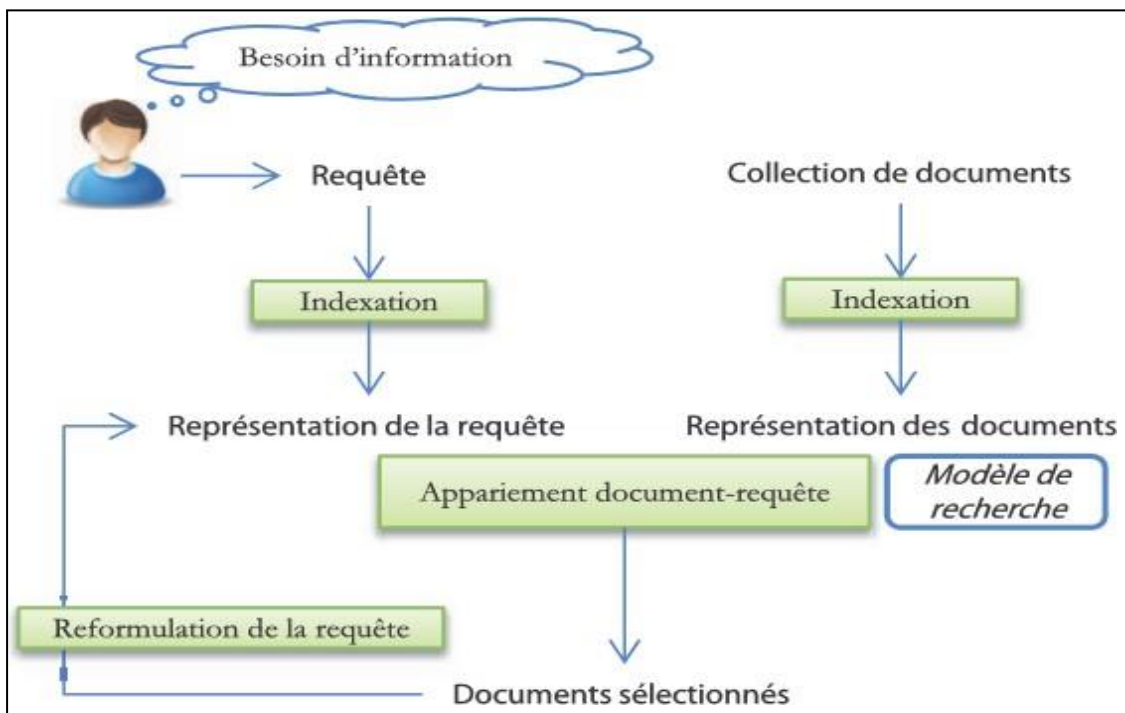
– **la pertinence thématique** : traduit le degré d'adéquation de l'information retrouvée au thème évoqué par le sujet de la requête. C'est la mesure la plus utilisée dans les moteurs de recherche classiques.

– **la pertinence cognitive** : représente la relation entre l'état de la connaissance intrinsèque de l'utilisateur et l'information portée par les documents telle qu'interprétée par l'utilisateur, cette pertinence se caractérise par une dynamique qui permet d'améliorer la connaissance de l'utilisateur via l'information renvoyée le long de sa recherche.

– **la pertinence situationnelle** : est vue comme l'utilité de l'information retrouvée par rapport à la tâche ou le problème posé par l'utilisateur.

– **la pertinence motivationnelle (ou affective)** : décrit la relation entre les intentions, les buts et les motivations de la recherche tels que fixés par l'utilisateur d'une part et les informations retrouvées d'autre part.

### 2.2-Processus générale de la RI :



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

Le processus de RI qui permet, à partir d'une requête, d'ordonner les documents est appelé "processus en U". Il est décomposé en trois principales étapes.

#### 2.2.1-Indexation

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

- **Manuelle** : la représentation du document est réalisée par un expert qui identifie les termes les plus représentatifs du document.
- **Automatique** : le processus d'indexation est entièrement informatisé. Il repose sur une démarche algorithmique qui traite chaque terme selon un processus défini : extraction, suppression des mots vides, normalisation et pondération...
- **Semi-automatique** : est une combinaison des deux précédentes approches où le choix final des termes à indexer revient à l'expert.

A la fin de cette étape, les documents sont représentés dans des fichiers index qui stockent la cartographie des couples terme-document en y associant un poids. La formule de pondération la

## Chapitre I : De la RI classique à la RI sociale

---

plus utilisée est celle basée sur la fréquence des termes dans les documents, appelée TF-IDF [9]. L'intuition de cette pondération est de favoriser les termes qui sont à la fois fréquents dans le document et peu fréquents dans la collection. Cette dernière condition est basée sur les propriétés de la loi de Zipf [10] qui étudie la distribution des termes dans une collection de documents.

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

- l'extraction des mots : Ce processus consiste à analyser le texte d'un document afin d'extraire ses mots en reconnaissant les espaces de séparation des mots, les ponctuations, etc.
- l'élimination des mots vides : Un document contient souvent des mots non significatifs appelés mots vides (pronoms personnels, prépositions).

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

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

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

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

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

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

- **TF (Term Frequency)** :

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

. **IDF (Inverse Document Frequency)** :

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

### 2.2.2-Appariement document-requête

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

Dans le processus d'appariement, le système calcule un score de correspondance entre la représentation de chaque document et celle de la requête. Ce score peut être binaire (pertinent ou non pertinent) ou multivalué pour exprimer un degré de pertinence système. La pertinence système est calculée à partir d'une fonction de similarité appelée RSV (Q, D) (Retrieval Status Value) où Q est une requête et D un document. Pour une requête donnée, le système retourne des documents en ordre décroissant du score de pertinence. L'appariement document-requête repose sur un cadre théorique défini par un modèle de recherche d'information. Une taxonomie des modèles (**Error! Reference source not found.**) a été présentée par Baeza-Yates [15] et présente quatre familles principales. Les modèles reposant sur le texte des documents (modèles de RI classiques et modèles basés sur le texte semi-structuré), les liens entre documents (modèles orientés web) et les documents multimédia (recherche d'images, de musiques, d'audio ou de vidéos).

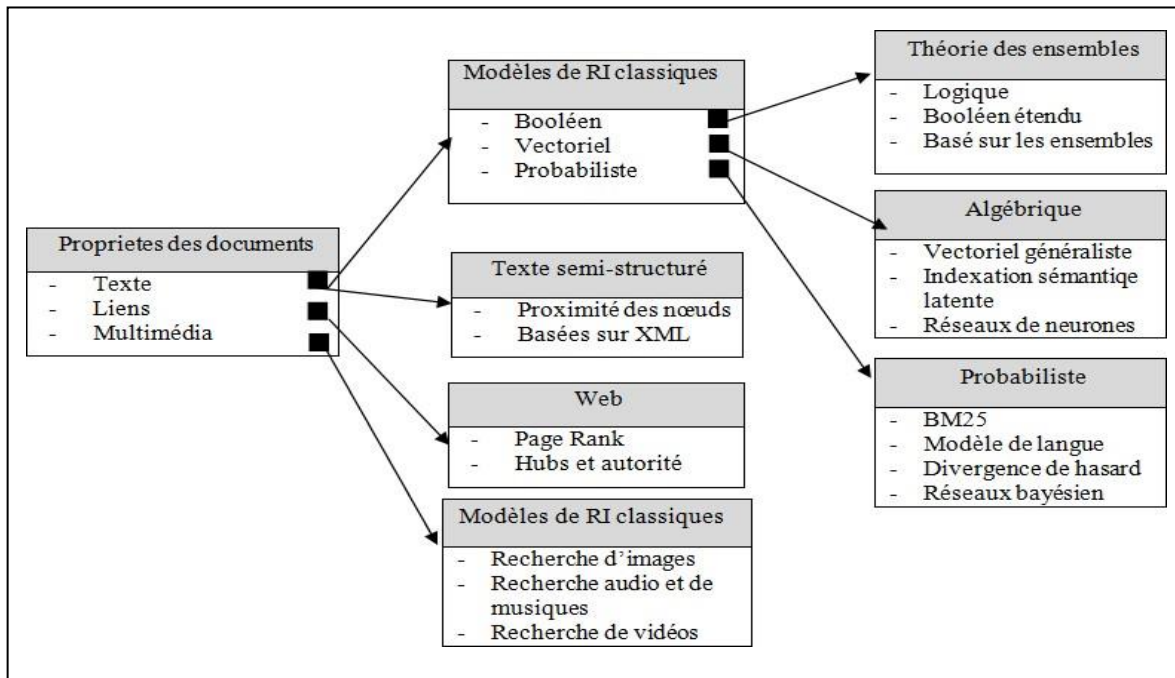


Figure 2 : les modèles mathématique de RI[15].

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

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

**2.2.2.1-Modèle booléen**

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

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

Le modèle booléen a des avantages qui sont présentés par :



## Chapitre I : De la RI classique à la RI sociale

---

- Le modèle de recherche booléen est reconnu pour sa force pour faire une recherche très restrictive et obtenir, pour un utilisateur expérimenté, une information exacte et spécifique.

- La simplicité du modèle le rend aisément compréhensible pour un utilisateur.

- L'efficacité du modèle est due aux spécialistes qui ont explorés le corpus avec une bonne connaissance du vocabulaire.

- La formulation des requêtes devient vite laborieuse quand la requête se fait précise.

Ce modèle n'a pas seulement d'avantages, il a aussi des inconvénients qui sont les suivants :

- La représentation binaire d'un terme dans un document est peu informative, car elle ne renseigne ni sur la fréquence du terme dans le document ni sur la longueur de document, qui peuvent constituer des informations importantes pour la RI.

- Les documents retournés à l'utilisateur ne sont pas ordonnés selon leur pertinence.

- L'impossibilité de rendre compte d'une correspondance partielle d'un document à une requête.

- Il est difficile pour les utilisateurs de formuler de bonnes requêtes. Par conséquent, l'ensemble des documents trouvés est souvent trop grand, pour les requêtes courtes, ou complètement vide dans le cas de requêtes longues.

- Les tests effectués sur des collections d'évaluation standards de RI ont montré que les systèmes booléens sont d'une efficacité de recherche inférieure.

### 2.2.2.2-Modèle vectoriel :

Initialement proposé par Salton et implémenté dans le système SMART [17]. La pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel. Le modèle vectoriel représente les documents et les requêtes par des vecteurs d'un espace à  $n$  dimensions, les dimensions étant constituées par les termes du vocabulaire d'indexation (Figure 3).

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

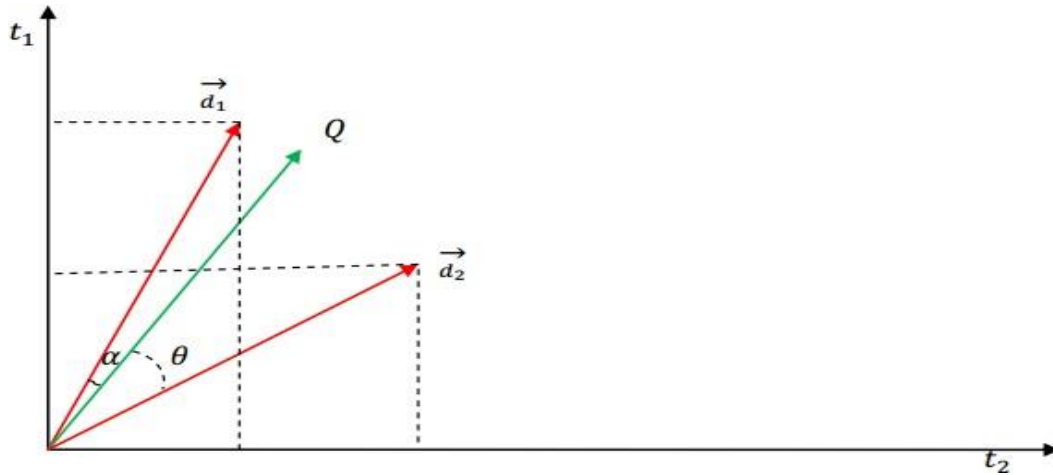
$$D_i = w_{i1}, w_{i2}, w_{i3} \dots w_{in}, \text{ pour } i = 1, 2 \dots m$$

Où  $w_{ij}$  est le poids du terme  $t$  dans le document  $D_i$ ,  $m$  est le nombre de documents dans la collection,

$n$  est le nombre de termes d'indexation.

On représente aussi la requête par un vecteur de mots-clés défini dans le même espace vectoriel que le document.  $Q = w_{q1} \dots, w_{q2}, \dots w_{qn}$ .

Où  $w_{qj}$  est le poids de terme  $t_j$  dans la requête  $Q$ .



**Figure 3:** Représentation algébrique des documents et des requêtes dans l'espace des termes à deux dimensions[17]

L'appariement document-requête dans le modèle vectoriel, consiste à trouver les vecteurs documents qui s'approchent le plus de vecteur de la requête. Cet appariement est obtenu par l'évaluation de la distance entre les deux vecteurs. Plusieurs mesures de similarité ont été définies [18], dont les plus courantes sont décrites ci-dessous.

- **Le produit scalaire :**

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

- **La mesure de Jaccard :**

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

- **La mesure cosinus :**

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

## Chapitre I : De la RI classique à la RI sociale

---

Plus les vecteurs sont similaires, plus l'angle formé est petit, et plus le cosinus de cet angle est grand. A l'inverse du modèle booléen, la fonction de correspondance évalue une correspondance partielle entre un document et une requête, ce qui permet de retrouver des documents qui ne reflètent pas la requête qu'approximativement. Les résultats peuvent donc être ordonnés par ordre de pertinence décroissante [19].

Le modèle vectoriel a des avantages [8] qui sont présentés par :

- -La pondération améliore les résultats de recherche.
- Le modèle permet une correspondance partielle ou approximative entre les documents et les requêtes (best match).
- -La mesure de similarité permet d'ordonner les documents selon leur pertinence vis à vis de la requête.

Ce modèle n'a pas seulement d'avantages, il a aussi des inconvénients qui sont les suivants :

- L'inconvénient majeur de modèle vectoriel est qu'il repose sur l'hypothèse de l'indépendance des termes d'indexation, or ces termes dans les documents sont souvent sémantiquement liés [16].
- Il comporte également plusieurs limitations qui furent, pour certaines, corrigées par des affinements du modèle [20].
- Dans un texte l'ordre des mots, les synonymes, la morphologie des contenus ne sont pas pris en compte [20].

### 2.2.2.3-Modèle probabiliste :

Ce modèle est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête [21] [22] [11]...Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Etant donné une requête utilisateur Q et un document D, il s'agit de calculer la probabilité de pertinence du document pour cette requête.

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

$$RSV(q, d_j) = \frac{P(P|d_i)}{P(\bar{P}d_i)}$$

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

Ces probabilités sont estimées par de probabilités conditionnelles selon qu'un terme de la requête est présent, dans un document pertinent ou dans un document non pertinent. Cette mesure de similarité entre la requête et les documents peut se calculer par différentes formules. Ce modèle a

donné lieu à de nombreuses extensions. Il est à l'origine du système OKAPI qui est l'un des systèmes les plus performants selon les compagnes d'évaluation TREC<sup>1</sup>. L'inconvénient majeur de ce modèle est que les calculs des probabilités sont complexes et que l'indépendance des variables n'est pas toujours vérifiée voir pas prise en compte [23].

Ces modèles ont une base théorique saine [24] et se sont montrés particulièrement performants dans TREC.

### **2.2.3-Reformulation de la requête**

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

Cette étape peut être effectuée :

Manuellement, dans le cas où l'utilisateur soumet lui-même une nouvelle requête.

De façon automatique, lorsque le système de RI s'appuie sur les termes importants dans les documents les plus pertinents ou visités par l'utilisateur qui sont réutilisés.

L'une des stratégies de reformulation de requêtes est celle qui est dirigée par l'utilisateur. Le principe de cette stratégie est de construire une nouvelle requête à partir de la structure des documents jugés par l'utilisateur : c'est ce que l'on appelle la réinjection de pertinence « relevance feedback » [25] [26] [27].

La reformulation est un processus évolutif et interactif. Son principe fondamental est d'utiliser la requête initiale pour amorcer la recherche, puis modifier celle-ci à partir des jugements de pertinence et/ou de non-pertinence de l'utilisateur dans le but de repondérer les termes de la requête initiale, ou y ajouter (respectivement supprimer) d'autres termes contenus dans les documents pertinents (respectivement non pertinents). La nouvelle requête obtenue à chaque itération de feedback, permet de corriger la direction de la recherche dans le sens des documents pertinents. En effet, la simple comparaison du contenu de la requête et des documents de la base ne permet pas d'avoir tous les documents correspondant à une requête donnée. Il reste toujours des documents pertinents non restitués, car ne contenant pas les termes de la requête [23].

## **3-Réseaux sociaux**

Les réseaux sociaux sont au cœur du Web 2.0. Un réseau social est la structure sociale virtuelle, qui émerge des interactions humaines à travers une application en réseau.

---

<sup>1</sup> WWW.TREC.com

## **Chapitre I : De la RI classique à la RI sociale**

---

La structure d'un réseau social peut être construite de deux façons : (i) soit explicitement déclaré par l'utilisateur, comme exemple, les liens d'amitié dans Facebook, ou (ii) implicitement déduit du comportement et des intérêts communs des utilisateurs, par exemple, le réseau social des services Web [28]. Pour comprendre les structures et les phénomènes sociaux sous-jacents, il existe un ensemble de techniques et de méthodes, connues sous le nom de techniques d'analyse des réseaux sociaux [29]. Ces techniques introduisent des méthodes et des mesures (p. ex., centralité et influence) pour analyser un réseau social, par exemple, mesurer le rôle des individus et des groupes d'individus dans un réseau social.

Chaque réseau social peut être caractérisé par les relations qui lient ses utilisateurs, par exemple, les relations amis, les relations suiveurs et les relations éditeur-abonné.

### **4-Recherche d'informations sociales**

Une large gamme d'applications et de services rendent l'utilisateur plus interactif avec les ressources Web, et beaucoup d'informations qui concernent à la fois les utilisateurs et les ressources sont constamment générées. Cette information peut être très utile dans les différentes tâches de recherche d'information du côté utilisateur et côté ressources. Cependant, les modèles classiques de RI sont aveugles à ce contexte social qui entoure à la fois les utilisateurs et les ressources.

Par conséquent, les domaines de la RI et des réseaux sociaux ont été fusionnés, ce qui a donné lieu à des modèles de recherche d'information sociale (SIR) [30]. Très souvent, les modèles SIR étendent les modèles RI traditionnels afin d'intégrer l'information sociale.

Le sens du concept de recherche d'information sociale peut être très large :

La recherche d'information sociale est le processus d'exploitation de l'information sociale (à la fois les relations sociales et le contenu social généré (UGC)), pour effectuer une tâche de RI dans le but de mieux répondre aux attentes et besoins des utilisateurs.

RIS vise à fournir un contenu d'informations pertinent aux utilisateurs dans les domaines de la RI, de la recommandation, la recherche collaborative, l'analyse des réseaux sociaux, les systèmes de questions et réponses et le filtrage collaboratif [30].

#### **4.1-Approches SIR:**

Aujourd'hui, chaque travail de recherche dans le domaine de la SIR considère un type de réseau et d'information sociale pour effectuer une tâche de RI [31] (figure 4) :

- 1.Approches de recherche sur le Web sociale : L'information sociale est utilisées pour améliorer le processus classique de RI, comme, le reclassement des documents, la reformulation des requêtes, etc.

## Chapitre I : De la RI classique à la RI sociale

2. Approches de Recherche sociale : Les ressources et les entités sociales sont récupérés à partir des réseaux sociaux.

3. Approches de recommandation sociale : L'information sociale est utilisée pour faire des recommandations, telles que : la recommandation de tags, la recommandation d'utilisateurs, etc. Le réseau social de l'utilisateur est utilisé pour fournir une meilleure recommandation.

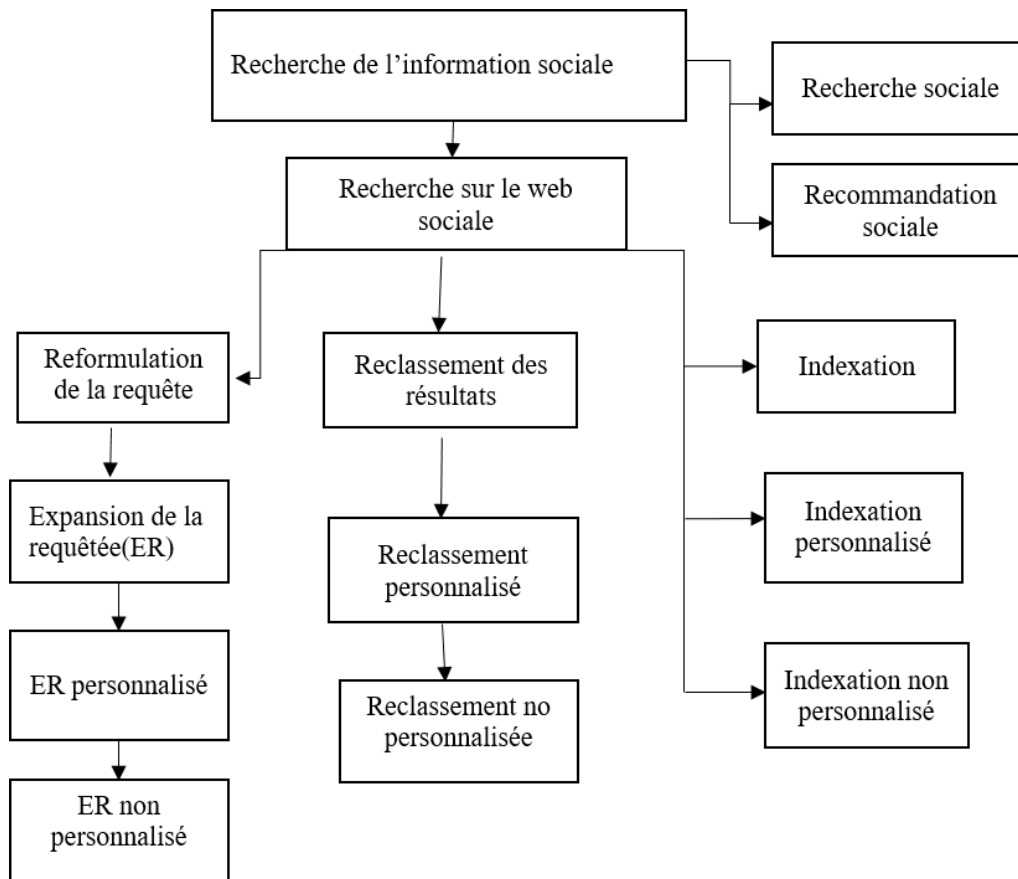


Figure 4 : Taxonomie des approches SIR [31]

### 4.1.1-Recherche sur le Web social :

Cette catégorie d'approches exploite les informations sociales pour améliorer le processus de RI traditionnel dans le Web. Dans les systèmes IR existants, les requêtes sont généralement interprétées et traitées à l'aide d'index de documents, qui sont transparentes pour les utilisateurs. Les documents qui en résultent ne sont pas nécessairement pertinents du point de vue de l'utilisateur final, malgré le classement effectué par le moteur de recherche. [32]

Pour améliorer le processus classique d'IR et réduire le volume de documents non pertinents, il existe principalement trois pistes d'amélioration possibles (figure 5): (i) la reformulation de la

## Chapitre I : De la RI classique à la RI sociale

requête, c'est-à-dire qui inclut l'expansion ou la réduction de la requête, (ii) le post-filtrage ou le reclassement des documents récupérés (en fonction du profil ou du contexte de l'utilisateur), et (iii) l'amélioration du modèle IR, c'est-à-dire la manière dont les documents et les requêtes sont représentés et appariés pour quantifier leurs similitudes.

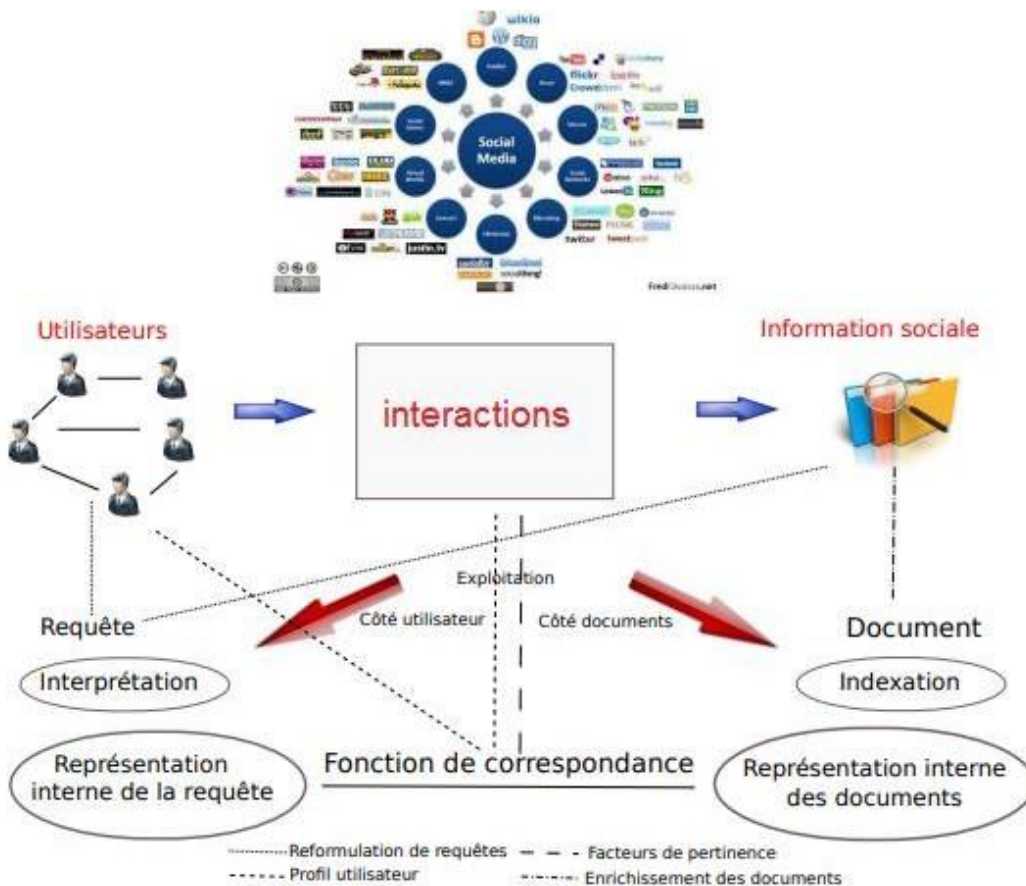


Figure 5 : Exploitation d'information sociale dans la IR

### 4.1.1.1-Reformulation de la requête

Dans les systèmes IR, les utilisateurs expriment généralement leurs besoins à travers un ensemble de termes (mots-clés) qui résument leurs besoins en informations. Ainsi, différents utilisateurs sont censés utiliser différents mots-clés pour exprimer le même besoin (par exemple, des synonymes), et vice versa (c'est-à-dire que les mêmes mots-clés peuvent être utilisés par différents utilisateurs pour exprimer différents besoins d'information). La reformulation des requêtes peut alors apporter une solution à ce problème. Elle est définie comme suit : "La reformulation de la requête est le processus qui consiste à faire passer une requête initiale  $Q$  en une requête finale  $Q'$ ". Cette transformation peut être soit une réduction ou une expansion. [33]

### **A. Expansion non personnalisée des requêtes**

L'espace des termes de recherche (requête) est enrichie d'informations sociales sans aucune personnalisation. Par conséquent, l'idée sous-jacente est de prendre en compte des interactions des utilisateurs avec le système pour construire implicitement l'ensemble de termes. Cet ensemble devrait alimenter ou étendre la requête initiale. [34][35][36][37]

### **B. Expansion personnalisée des requêtes**

Fournir une expansion uniforme à tous les utilisateurs n'est souvent pas vraiment approprié ni efficace car la pertinence des documents est relative à chaque centre d'intérêt de chaque utilisateur [38]. Ainsi, une extension de requête simple et uniforme n'est pas suffisante pour fournir des résultats de recherche satisfaisants pour chaque utilisateur. L'expansion sociale personnalisée des requêtes fait référence au processus d'expansion de la même requête différemment pour chaque utilisateur en utilisant des informations sociales. [39][40][41]

#### **4.1.2.1-Classement social des résultats de recherche :**

En SIR, le classement social des résultats consiste en la définition d'une fonction qui permet de quantifier les correspondances entre les documents et les requêtes. Deux catégories de classement des résultats sociaux, qui diffèrent dans la manière dont ils utilisent les informations sociales, sont à considérer : La première catégorie utilise les informations sociales en ajoutant une pertinence sociale au processus de classement, tandis que la seconde les utilise pour personnaliser les résultats de recherche. [31]

#### **A. Classement par pertinence sociale (non personnalisé)**

La pertinence sociale fait référence à l'information générée par les internautes qui caractérisent un document par sa popularité et son importance sociale. [37]

#### **B. Classement social personnalisé**

Ces approches personnalisent le classement des résultats de recherche pour chaque utilisateur en fonction de son profil [39 ,31 42]. Pour la plupart de ces approches, la fonction de correspondance personnalisée est une combinaison de deux scores de similarité : la similarité entre le document et la requête et la similarité entre le document et l'utilisateur. [31]

#### **4.1.1.3-Indexation sociale**

Dans le Web social, un contexte social est souvent associé aux pages Web ou plus spécialement à leur contenu pour construire leur descripteur. À titre d'exemple, ce contexte social comprend des annotations, des commentaires, des mentions similaires, etc.

L'information sociale a été principalement utilisée de deux manières pour améliorer la représentation des documents : (i) soit en ajoutant des métadonnées sociales au contenu des



documents, par exemple, l'expansion des documents, ou (ii) en personnalisant la représentation des documents.

### A. Enrichissement du document (indexation non personnalisée).

Certains travaux étudient l'utilisation des métadonnées sociales (annotations) pour enrichir le contenu des documents. Le framework consiste à représenter par exemple un document Web sous deux représentations : (i) représentation de contenu textuel enrichi et (ii) représentation de contenu social enrichi. Chaque composante étant enrichie à partir de l'autre [43][44]

### B. Indexation personnalisées des documents

Étant donné un document, chaque utilisateur a sa propre compréhension et son propre point de vue sur le contenu. Par conséquent, chaque utilisateur emploie un vocabulaire et des mots différents pour décrire, commenter et annoter ce document. Chaque document est indexé en fonction de son utilisateur.[45]

#### 4.1.2-Recherche sociale

Les services sociaux comme Twitter et Facebook permettent aux utilisateurs de partager et de publier des informations avec leurs amis et souvent avec le grand public. Dans la recherche sociale, les réseaux sociaux sont considérés comme des moteurs de recherche dédiés à la gestion des interactions sociales comme le contenu social (les commentaires, les tweets, ...) et les relations sociales des utilisateurs.[46].

La recherche sociale est le processus de recherche d'informations uniquement avec l'aide de réseaux sociaux.

La recherche sociale est divisée en trois catégories principales :

- (1) les outils de questions/ réponses,
- (2) la recherche de contenu
- (3) la recherche collaborative.

#### 4.1.3-Recommandation sociale

La deuxième catégorie de modèles SIR considère les domaines de filtrage et de recommandation (par exemple, le filtrage basé sur le contenu, le filtrage collaboratif, les systèmes de recommandation). Fondamentalement, la recommandation vise à prédire l'intérêt que les utilisateurs accorderaient à un élément/entité qu'ils n'avaient pas encore envisagé explicitement. Il existe deux principales méthodes de recommandation : [31]

- une approche basée sur la recommandation d'éléments similaires et proches aux éléments favorisés par cet utilisateur dans le passé, dite « basé sur le contenu ».

## **Chapitre I : De la RI classique à la RI sociale**

---

- une approche qui vise à recommander des articles à l'utilisateur en fonction d'autres personnes qui s'avèrent avoir des préférences ou des goûts similaires, qui est connue sous le nom de «filtrage collaboratif».

La recommandation sociale est un ensemble de techniques qui tentent de suggérer : (i) des éléments (par exemple, des films, de la musique, des livres, des actualités, des pages Web), (ii) des entités sociales (par exemple, des personnes, des événements, des groupes), ou (iii) les sujets d'intérêt (par exemple, le sport, la culture et la cuisine) qui sont susceptibles d'intéresser l'utilisateur grâce à l'utilisation d'informations sociales.

Les systèmes de recommandation sociale sont catégorisés en fonction du type de sortie qu'ils ont l'intention de recommander : (i) recommandation d'éléments, (ii) recommandation d'utilisateurs et (iii) recommandation de sujets.

### **5-Conclusion :**

Dans ce chapitre nous avons présenté les fondements de base de la recherche d'information traditionnelle et sociale où plusieurs champs d'exploration ont été développés. Ces champs sont catégorisés en : (i) la recherche sur le Web social, (ii) la recherche sociale et (iii) la recommandation sociale. Ces trois catégories sont fondamentalement différentes dans la manière dont elles exploitent et utilisent les informations sociales. Cependant, dans le chapitre suivant, nous montrons l'intégration de l'information sociale dans la représentation du profil document et dans l'amélioration du processus RI classique.

# Chapitre II :

## Généralités sur l'analyse des sentiments

### 1. Introduction

Nous aborderons dans ce chapitre les détails sur le sujet de l'analyse des sentiments notamment : termes, applications, approches, ...

### 2. Notions de base

Lorsqu'on aborde le domaine de l'analyse de sentiments, l'une des premières questions à se poser pourrait être la suivante : Quelle est la différence entre un sentiment et une opinion et la relation avec l'émotion ?

Pour y répondre, nous allons tout d'abord définir le sentiment, puis l'émotion et enfin l'opinion.

#### 2.1-Sentiment

Larousse <sup>2</sup> définit le sentiment comme étant « un état affectif durable lié à certaines émotions ou représentations ».

Et selon [47 ], le sentiment est défini comme «la composante de l'émotion qui implique les fonctions cognitives de l'organisme et la manière d'apprécier. Le sentiment est à l'origine d'une connaissance immédiate ou d'une simple impression ».

[48] présentent le sentiment d'un point de vue psychologique comme un concept sociologique basique utile pour analyser le lien des sensations corporelles, la gestuelle et les relations sociales.

#### 2.2-Émotion

L'auteur de [49] définit l'émotion comme étant une expérience psychophysologique complexe et intense avec un début brutal et une durée relativement brève.

L'auteur, était un leader d'opinion dans l'étude des émotions, il a conçu la théorie psycho-évolutionnaire de l'émotion (Figure 6), ce qui permet de catégoriser les émotions en émotions primaires et les réponses afférentes [49]

Il a soutenu que les émotions primaires sont un développement évolutif et que la réponse à chacune de ces émotions est celle qui est susceptible d'offrir le plus haut niveau de survie possible:

---

<sup>2</sup> <https://www.larousse.fr/>

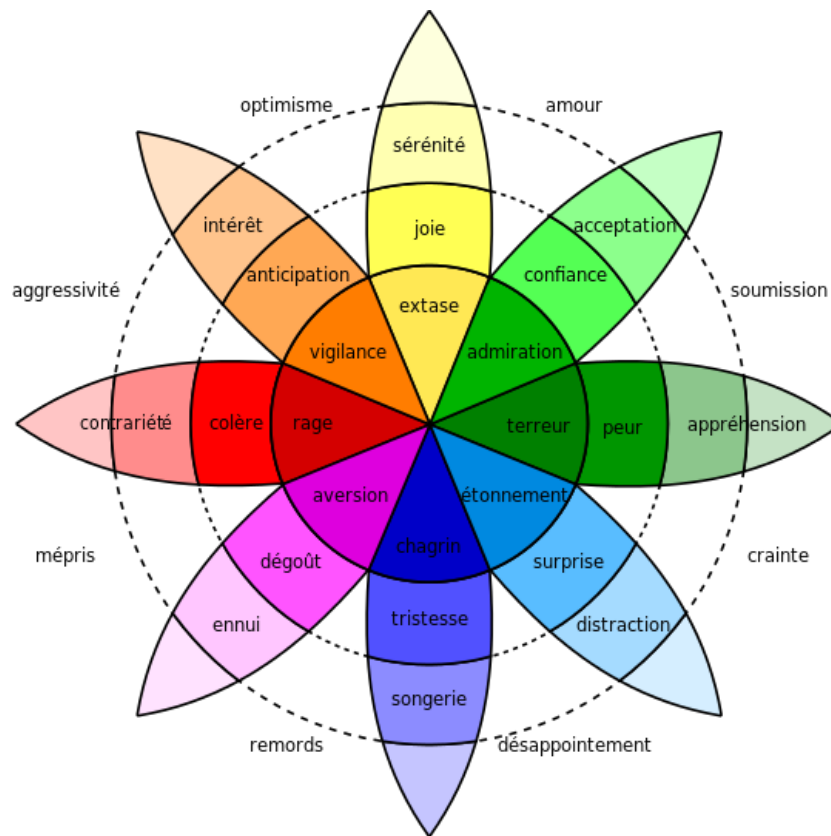


Figure 6 : Roue des émotions de Plutchik [49]

### 2.3-Opinion

Selon LAROUSSE , l'opinion est « Jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense», ou encore comme «Ensemble des idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc.»

L'auteur de [50] définit une opinion par le quintuple suivant, dont les composants sont liés les uns aux autres :

- $e_i$  : entité  $i$  cible de l'opinion.
- $a_i$  : aspect  $j$  de l'entité  $i$ .
- $s_{ijkl}$  : sentiment exprimé sur l'aspect  $j$  de l'entité  $i$  par la source  $k$  dans le temps  $l$ .
- $h_k$  : source de l'opinion.
- $t_l$  : moment de l'opinion.

#### 2.3.1-Type d'opinions

Les types d'opinions selon [50] :

##### 2.3.1.1-Opinion ordinaire

Ce type est tout simplement appelé opinion dans la littérature, on peut cependant distinguer deux

## **Chapitre II : Généralités sur l'analyse des sentiments**

---

types d'opinions :

- **Opinion directe** : désigne une opinion exprimée directement sur une entité ou un aspect d'une entité (exemple : L'écran de ce téléphone est impressionnant).

- **Opinion indirecte** : désigne une opinion exprimée indirectement sur une entité ou un aspect d'une entité basé sur d'une autre entité (exemple : Après avoir changé de type de carburant, la voiture roulait difficilement).

### **2.3.1.2-Opinion comparative**

L'opinion comparative exprime une relation de similitude ou de différence entre plusieurs entités, il existe deux types d'opinions comparatives :

- **Comparaison évaluée** : dans ce type de comparaison, il existe une préférence évidente entre les entités (exemple : la BMW est plus rapide que la Renault 4)

- **Comparaison non évaluée** : dans ce cas, il existe une différence entre les entités, cependant, on ne peut déterminer laquelle le détenteur de l'opinion préfère (exemple : La vitesse de cette BMW est différente de la Renault4)

## **3-Analyse des sentiments SA**

### **3.1-Définition**

Dans la littérature, sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, sont des termes utilisés pour désigner des technologies d'analyse automatique des discours, écrits

Ou parlés, afin d'en extraire des informations subjectives comme des jugements, des évaluations ou des émotions. [51]

Les auteurs de [52] ont présenté une définition de SA comportant les domaines d'application ainsi que sa relation avec le TALN «l'analyse des sentiments est le domaine d'étude qui analyse les opinions, les sentiments, les évaluations, les attitudes et les émotions des gens vers des entités telles que des produits, des services, des organisations, des particuliers, des problèmes, des événements, des sujets, et leurs attributs».

### **3.2-Taches de SA**

Les tâches de SA peuvent être différentes la détection de la subjectivité, la classification de l'orientation et les type d'analyse :

#### **3.2.1-Détection de subjectivité**

Les documents qui expriment le point de vue de l'auteur sont subjectifs, et ceux qui sont factuels sont objectifs. La détection de la subjectivité consiste à classer les textes comme étant subjectifs ou objectifs. [53]

### 3.2.2-Classification de l'orientation

La classification de subjectivité (Subjectivity classification) (Figure 7) est la tâche qui distingue les phrases exprimant des informations objectives des phrases exprimant des vues et opinions subjectives.

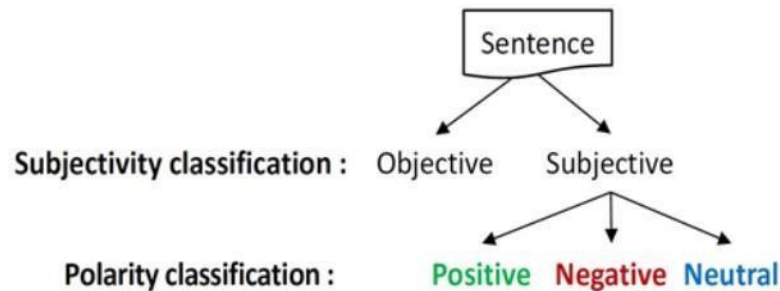


Figure 7 : Classification de l'orientation sémantique [53]

### 3.3 Types de SA

Il existe de nombreux types d'analyses de sentiments allant des systèmes qui se concentrent sur la classification de la polarité (positif, négatif, neutre) aux systèmes qui détectent des émotions (en colère, heureux, triste, etc.) ou identifient des intentions (par exemple, intéressé, pas intéressé). Dans la section suivante, nous aborderons les types les plus importants. [54]

#### 3.3.1 Analyse fine des sentiments

Au lieu de parler de phrases positives, négatives ou neutres, on considère les catégories suivantes :

«Très positive, Positive, Neutre , Négative ,Très négative»

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

#### 3.3.2-Détection d'émotions

La détection des émotions vise à détecter des émotions telles que le bonheur, la frustration, la colère, la tristesse, etc.

De nombreux systèmes de détection d'émotions sont basés sur l'utilisation de lexiques de sentiments (c'est-à-dire des listes des émotions) ou sur des algorithmes d'apprentissage automatique complexes.

### 4- Niveaux de SA

L'étude de l'analyse de sentiment peut s'effectuer à différents niveaux (Figure 8). La décomposition de niveau de granularité permet de simplifier l'analyse de sentiments du document global. [55]

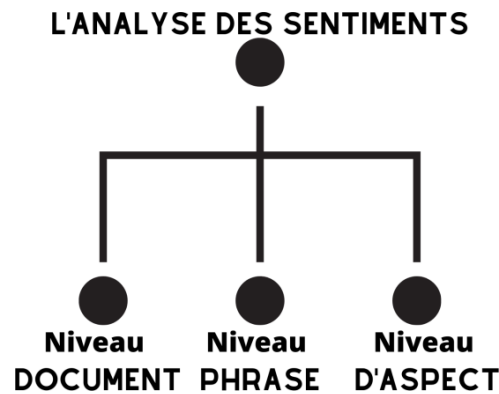


Figure 8 : les niveaux d'analyse de sentiment[55]

#### 4.1-Niveau Document

L'analyse des sentiments au niveau du document suppose que chaque document exprime des opinions sur une seule entité (ex : un seul produit)

#### 4.2-Niveau Phrase

Le niveau phrase est une direction principale dans la fouille d'opinion, [56], il est considéré comme une étape intermédiaire dans un processus globale pour déterminer l'orientation sémantique du document entier.

La polarité du document est obtenue en synthétisant les scores de polarité des phrases subjectives avec un poids représentant l'importance de la phrase.

La subjectivité d'une phrase est obtenue par la présence de mots subjectifs d'un dictionnaire de langue, et aussi par un ensemble de règles de langage qui indiquent l'intensification.[57]

#### 4.3-Niveau aspect (entité)

Selon [55], le sentiment global que comporte un document n'est pas suffisant, et on cherche les aspects (caractéristiques ou parties) d'un objet (produits, sujet, ...) qui sont ciblés par les commentaires.

En se basant sur ce niveau d'étude, il est possible d'avoir une structuration des opinions sur les entités et leurs différents aspects.

Dans l'exemple suivant cité dans [50] "The iPhone's call quality is good, but its Battery life is short" l'évaluation concerne deux aspects ; la qualité d'appel (call quality) sur laquelle l'opinion est



Positive, et la durée de vie de batterie (battery life) dont un sentiment négatif est exprimé, et cela Pour une même entité iPhone.

### **5. Domaines d'application d'AS**

L'importance de AS est présente dans plusieurs domaines ainsi plusieurs applications ont vu le jour dans ce contexte.

Quelques applications sont mentionnées brièvement ci-dessous

#### **5.1-Domaine du commerce**

L'analyse de sentiments est appliquée dans le but d'extraire le sentiment du marché.

Les vendeurs veulent avoir des informations sur l'opinion des clients à travers les commentaires sur les produits, pour utiliser cette information dans leur stratégie de marketing afin de combattre leurs concurrents. Les clients, à leurs tours, veulent avoir les opinions des autres clients pour guider leurs choix (produits, hôtels, vacances, ...) [55].

#### **5.2-Domaine de santé**

Le domaine de la santé est l'un des domaines les moins explorés dans les travaux d'analyse de sentiment.

Dans ce domaine, il est question de savoir que pensent les gens envers leurs médecins, leurs médicaments prescrits, leurs maladies et les traitements qu'ils subissent ?

SentiHealth-Cancer (SHC-pt) est un outil d'analyse de sentiments qui permet la détection de l'état émotionnel des malades de Cancer dans des communautés brésiliennes à travers leurs postes Facebook en langue portugaise. [58]

#### **5.3-Détection de spam d'opinion**

Tout le monde peut mettre n'importe quoi sur Internet, ce qui augmente les risques de spam sur le Web. Les internautes peuvent écrire ou envoyer des spams pour induire les utilisateurs en erreur.

#### **5.4-Domaine politique**

De nos jours, les médias et les populations s'intéressent bien avant le début des élections à connaître l'élu du peuple et le promis à un siège important du gouvernement.

L'avis populaire étant longuement caché dans la presse et dans les cafés voit maintenant le jour avec la virilité (diffusion rapide) dans réseaux sociaux

### **6-Problèmes de SA**

Le domaine de l'AS est au départ lié au traitement du langage naturel TALN, ce domaine présente quelques problèmes d'analyses :

### 6.1-Détection de la subjectivité

Pour améliorer les performances du système, un module de détection de subjectivité est inclus pour filtrer les faits objectifs, mais cela est souvent difficile à faire. Il s'agit donc de différencier le texte avec opinion et sans opinion.

A considérer les exemples suivants :

- Je déteste les histoires d'amour.
- Je n'aime pas le film "Je déteste les histoires".

Le premier exemple présente un fait objectif tandis que le deuxième exemple représente l'opinion sur un film particulier.

### 6.2-Sentiment implicite

Une phrase peut avoir un sentiment implicite même sans la présence de tout sentiment porteur de mots, considérez les exemples suivants :

Comment peut-on s'asseoir à travers ce film ?

Il faut s'interroger sur la stabilité d'esprit de l'auteur qui a écrit ce livre.

Les deux phrases ci-dessus ne portent pas explicitement de mots avec un sentiment bien que les deux soient des phrases négatives.

Ainsi, l'identification de la sémantique est plus importante en AS que la détection de syntaxe

### 6.3-Dépendance du domaine

Il existe de nombreux mots dont la polarité change d'un domaine à l'autre.

Considérez les exemples suivants :

- L'histoire était imprévisible.
- La direction de la voiture est imprévisible.
- Allez lire le livre.

Dans le premier exemple, le sentiment véhiculé est positif alors que le sentiment véhiculé dans le second est négatif. Le troisième exemple a un sentiment positif dans le domaine du livre mais un sentiment négatif dans le domaine du film (où le réalisateur est invité à aller lire le livre).

### 6.4-Identification d'entité

Un texte ou une phrase peut avoir plusieurs entités, il est extrêmement important de connaître l'entité vers laquelle l'avis est dirigé.

Considérez les exemples suivants :

- Samsung est meilleur que Nokia

## **Chapitre II : Généralités sur l'analyse des sentiments**

---

- Ram a battu Hari au football.

Les exemples sont positifs pour Samsung et Ram respectivement mais négatifs pour Nokia et

Hari.

### **6.5-Négation**

La gestion de la négation est une tâche difficile en SA, elle peut être exprimée de manière subtile même sans l'utilisation explicite d'un mot négatif.

Une méthode souvent suivie pour gérer explicitement la négation dans des phrases comme : « Je n'aime pas le film », consiste à inverser la polarité de tous les mots apparaissant après l'opérateur de négation (comme pas), mais cela ne fonctionne pas pour "je n'aime pas le jeu mais j'aime la mise en scène".

### **7-Algorithmes d'analyse des sentiments**

Dans la littérature, il existe de nombreuses méthodes et algorithmes pour mettre en œuvre des systèmes d'analyse des sentiments, que l'on peut classer comme suit [60] :

- ✓ Approche automatique
- ✓ Approche à base de règles
- ✓ Approche hybride

#### **7.1-L'approche automatique**

Les approches automatiques reposent sur des techniques d'apprentissage automatique (Machine Learning). La tâche d'analyse des sentiments est généralement modélisée comme un problème de classification dans lequel un classificateur est alimenté avec un texte et renvoie la catégorie correspondante, par ex. positif, négatif ou neutre (en cas d'analyse de polarité) [59] (Figure 9).

Il existe différentes classifications de méthodes d'apprentissage supervisé utilisées pour analyser les sentiments, certaines d'entre elles sont les suivantes :

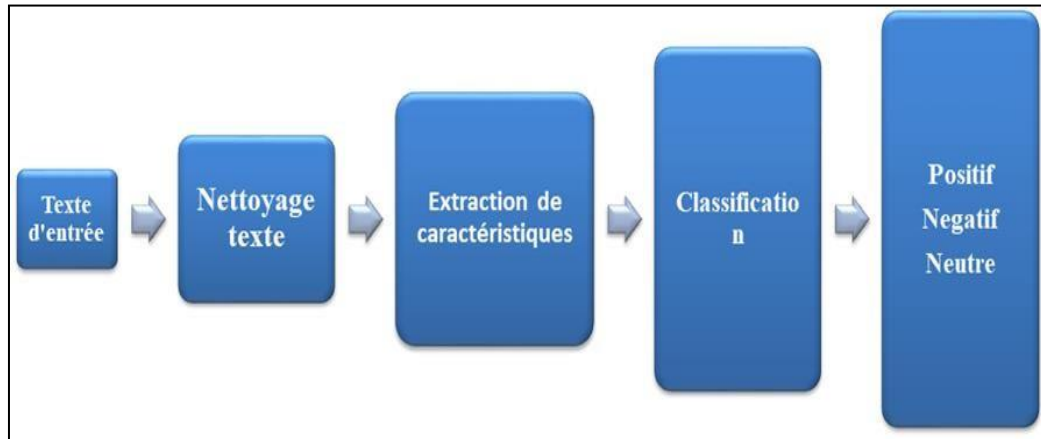


Figure 9 : l'approche automatique [60].

### 7.2 Approche à base de règles (Rule-based)

L'approche à base de règles (ou l'approche lexicale) définit un ensemble de règles dans un type de langage de programmation (script) qui identifie la subjectivité, la polarité ou le sujet d'une opinion. Cette approche peut utiliser diverses entrées, telles que [59](figure 10) :

- Techniques classiques de NLP, telles que la racinisation, tokenisation, POS-tagging et Chunking.
- Autre opération basée sur le lexique, ils utilisent le dictionnaire des sentiments avec des mots d'opinion et les faire correspondre avec les données pour déterminer la polarité.

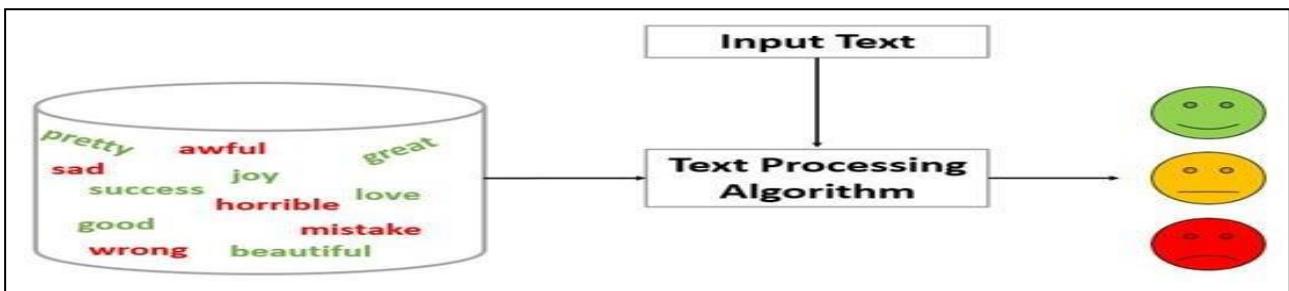


Figure 10 : Approche à base de règle [59].

### 7.3-Approche hybride

Le concept de méthodes hybrides est très intuitif : combiner simplement le meilleur des deux approches, celui basé sur des règles et celui automatique. Généralement, en combinant les deux approches, les méthodes peuvent améliorer la précision [59].

### 8-L'analyse des sentiments avec Twitter

Parmi les réseaux sociaux, Twitter est l'un des plus importants services de microblogging. L'architecture de Twitter fait de la question « Que se passe-t-il ? » La pierre angulaire de l'échange d'informations. Cela a inspiré l'idée d'utiliser les utilisateurs de Twitter en tant que capteurs distribués [61].

Les plates-formes des médias sociaux telles que Twitter sont de plus en plus courantes et fournissent des informations précieuses (générées par les utilisateurs via la publication et le partage de contenus) [62].

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 [63].

Twitter est une plate-forme multimédia qui permet de partager facilement des opinions en utilisant diverses formes de contenu, notamment du texte, des images et des liens contrairement à de nombreuses autres plates-formes de médias sociaux (tel que Instagram).

De plus, en fournissant un accès en temps quasi réel aux publications publiques via l'API, Twitter est une plateforme appropriée pour l'exploration d'opinion à grande échelle en temps quasi réel.

### 8.1-Outils d'analyse de sentiments

#### 8.1.1-SentiWordNet

SentiWordNet [64] a été généré à l'aide du dictionnaire WordNet<sup>3</sup>. Chaque synset (c'est-à-dire un groupe de termes synonymes sur une signification particulière) dans WordNet est associé à trois scores numériques indiquant le degré d'association du synset avec du texte positif, négatif et objectif. En générant le lexique, les synsets initiaux (positifs et négatifs) ont été étendus en exploitant les relations de synonymie et d'antonymie dans WordNet, grâce à quoi la synonymie se

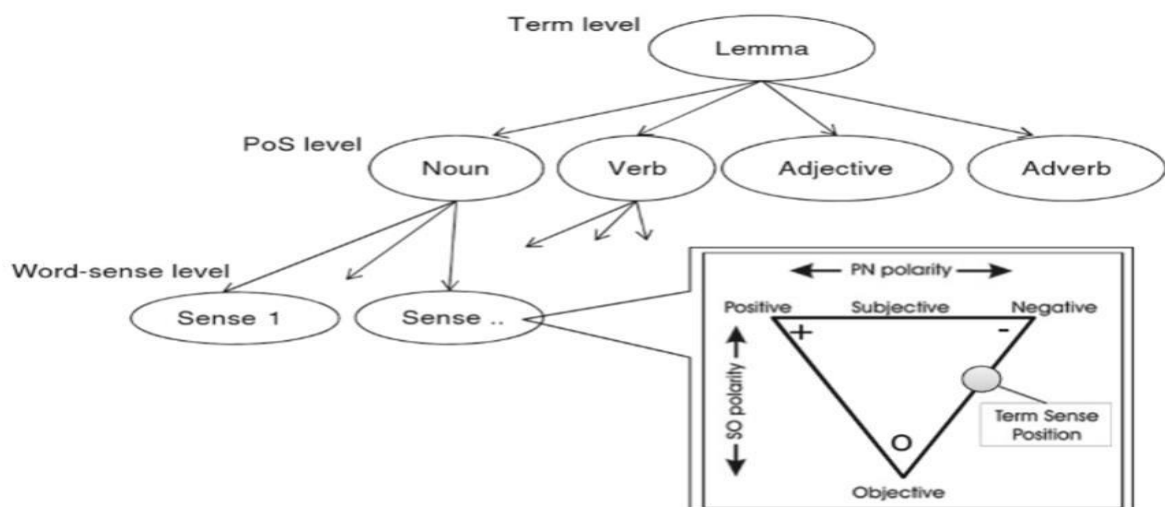
---

<sup>3</sup><https://wordnet.princeton.edu/wordnet/>

## Chapitre II : Généralités sur l'analyse des sentiments

Conserve tandis que l'antonymie inverse la polarité avec un synset donné. Puisqu'il n'y a pas de relation synonyme directe entre les synsets dans WordNet, les relations : `see_also`, `similar_to`, `pertains_to`, `derived_from` et `attribute` ont été utilisées pour représenter la relation de synonymie tandis que la relation antonyme directe a été utilisée pour l'antonymie. Des glosses (c.-à-d. Les définitions textuelles des ensembles élargis de synsets ainsi que celle d'un autre ensemble supposé être composé de synsets objectifs) ont été utilisés pour former huit classificateurs ternaires. Les classificateurs sont utilisés pour classer chaque synset et la proportion de classification pour chaque classe (positive, négative et objective) ont été considérés comme les scores du synset. Dans une version améliorée du lexique (SentiWordNet 3.0), les scores ont été optimisés par une marche aléatoire en utilisant l'approche PageRank. Avec les synsets sélectionnés manuellement, puis propage la polarité des sentiments (positive ou négative) à un synset cible en évaluant les synsets qui se connectent au synset cible à travers l'apparence de leurs termes dans le brillant du synset cible. [65]

SentiWordNet peut être vu comme ayant une structure arborescente comme indiqué sur la (figure 8). Le nœud racine de l'arbre est un terme dont les nœuds enfants sont les quatre balises PoS de base dans WordNet (c.-à-d. nom, verbe, adjectif et adverbe). Chaque POS peut avoir plusieurs sens du mot en tant que nœuds enfants. Les scores de sentiment illustrés par un point dans l'espace triangulaire du diagramme sont attachés aux sens des mots. La subjectivité augmente (tandis que l'objectivité diminue) de bas en haut, et la positivité augmente (tandis que la négativité diminue) de droite à gauche du triangle.



**Figure 11** : Diagramme montrant la structure de SentiWordNet [65].

### 8.1.2-Sentiment140

Sentiment140<sup>4</sup> (anciennement connu sous le nom "Twitter sentiment") est un outil en ligne gratuit qui a été créé par trois étudiants en computer science de Stanford, donc c'est un projet académique. Cet outil, contrairement à la plupart des autres sites d'analyse de sentiments, n'utilise pas de listes de mots positifs ou négatifs mais est fondé sur les algorithmes d'apprentissage automatique. Sentiment140 permet de découvrir des sentiments des tweets d'une marque, un produit ou un sujet sur Twitter.

### 8.1.3-Tweetfeel

Le service Tweetfeel<sup>5</sup> est un outil en ligne d'analyse du sentiment sur Twitter. Il propose une version gratuite et une version payante. Il s'appuie sur les capacités temps réels de Twitter qui donne des sentiments positifs et négatifs des tweets sur des choses. L'évaluation de TweetFeel se fait sur la base de présence de mots clés précis dans les tweets tels que Good, Bad, etc. (Uniquement en anglais pour l'instant). Ensuite un pourcentage est calculé selon le nombre de tweets positifs ou négatifs.

### 8.1.4-Twitrratr

Twitrratr<sup>6</sup> est un outil en ligne gratuit, qui a émergé à partir d'un projet Startup Weekend. Twitrratr fonctionne à partir d'une liste de mots positifs et d'une liste de mots négatifs [66]. Cet outil classe une opinion sur le mot clé de la requête s'il est capable de le croiser avec un mot d'une des deux listes. Les mots positifs et négatifs qui servent à classer les tweets sont surlignés dans l'interface.

### 8.1.5-Tweet Sentiments Analyses

Tweet Sentiments Analyses est un outil en ligne gratuit et open source d'analyse du sentiment sur Twitter. Il peut donner des sentiments positifs, négatifs et neutres des tweets sur le mot clé lancé dans la requête. Il peut travailler sur 12 langues. Il donne les résultats sous forme graphique. [67]

## 9- Conclusion

Dans ce chapitre, nous avons présenté la revue de littérature sur l'analyse des sentiments. Ceci comprend un survole théorique sur les concepts de base et leurs caractéristiques.

---

<sup>4</sup> <http://www.sentiment140.com/>

<sup>5</sup> <http://www.tweetfeel.com>

<sup>6</sup> <http://twitrratr.com/>

# CHAPITRE III

## Notre processus RIS



# CHAPITRE III : Notre processus RIS

---

## 1-Introduction

Afin de répondre aux besoins et aux attentes des internautes, les travaux classiques en recherche d'information (RI) se sont basés essentiellement sur la notion contenu qui est la correspondance classique entre les documents et les termes de la requête. Avec l'émergence du Web social (Web 2.0) et le volume important des contenus sociaux générés(UGCs), les travaux de RI ont commencé à s'intéresser d'avantage à l'enrichissement social et plus spécialement du côté document. En sens, nous proposons d'étendre la dimension de l'espace document par ces UGCs : On parle de Profil document.

Dans ce chapitre, nous présentons notre solution de profilage de document qui vise à exploiter la dimension social pour améliorer les résultats de recherche retournés par l'approche classique.

## 2-Architecture générale de notre système :

Notre approche de profilage commence par la construction et l'expansion socio-émotionnelle de la Dimension document, passe par un classement(Ranking) thématique via le modèle de référence BM25 et se termine par un reclassement (re-ranking) thematique-Socio-Emotionnel des résultats de recherche retournés.

Notre processus de RI, qui est la procédure fondamentale du système, a pour but la mise en relation, d'une part, des informations disponibles dans la collection des données( dataset) construit à partir de tweets collectés du -social data "Twitter" et précisément sont des tweets parlant sur la pandémie COVID'19, et de l'autre part des besoins de l'utilisateur traduit par une requête (Query).

Notre module de recherche classique basée sur le modèle de référence BM25 est constitué de Module de représentation des documents pour l'indexation et le stockage d'information, Module de traitement des requêtes qui interprète les requêtes des utilisateurs et Module de recherche d'information qui tri les résultats de recherche via une fonction de correspondance RSV (Document, Query). A l'étape finale, le système renvoi l'information de tout un ordonnancement thematique de résultats pertinents.

L'utilisateur exprime son besoin en informations en soumettant une requête (Query) et en interrogeant la base des tweets (Document) enrichie de données sociales. Cette valeur ajoutée formé de signaux sociaux et de polarités de sentiment (positif, négatif ou neutre) constitue le profil document.

## CHAPITRE III : Notre processus RIS

Donc, l'objectif principal de notre processus RI est de combiner les trois pertinences Socio-Emot-Thématique d'une ressource en exploitant les UGCs sociaux-émotionnelles associés et quantifiés, où chaque dimension textuelle, sociale et émotionnelle représente un facteur de pertinence.

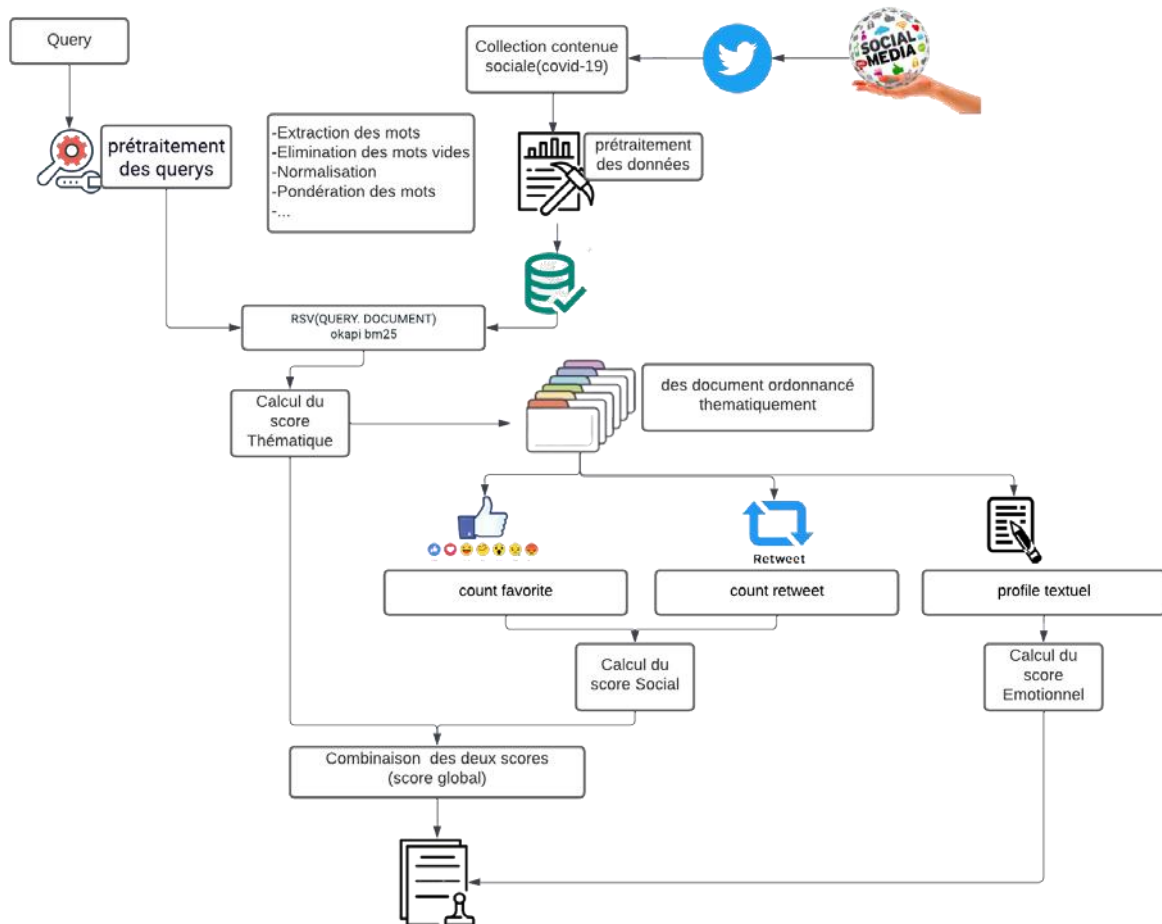
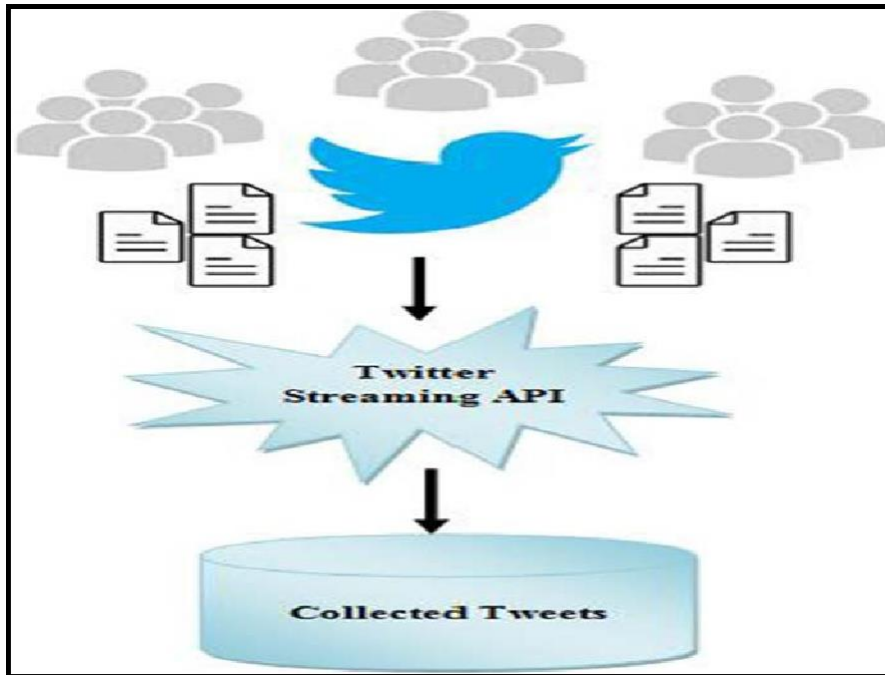


Figure 12 : Architecture générale de notre SRI

### 2.1-Collection des Tweets Covid'19 :

Twitter est rapidement devenu l'un des médias sociaux les plus populaires depuis son lancement, il compte plus de 313 millions d'utilisateurs actifs qui produisent plus de 6000 tweets sur Twitter chaque seconde et chaque année. En faveur de la collecte des données relatives à la Covid'19, chaque semaine, notre système continue à surveiller un ensemble de tweets en streaming qui inclut les mots proches à «COVID'19» dans toute la communauté Twitter.



**Figure 13** : Collecte des tweets via l'API-TWITTER.

Afin d'améliorer les résultats de recherche retournés par le processus classique, les UGCs (comme "favoris" et "retweet") sont déjà collectés et incorporés en tant que champs additionnels au sein du profil Tweet brut et sont exploités par la suite pour le calcul du score Social. Ensuite, le contenu textuel du tweet (texte brute) est identifier et analyser afin de quantifier sa polarité émotionnelle (positive, négative ou neutre) qui va bonifier le calcul global du score Socio-Emot-Thématique. Le tout reconstitue le profil brut : profil Socio-Emot-Thématique du document.

**2.2-Prétraitement (Preprocessing)** : Le prétraitement du Dataset est le processus de nettoyage et de préparation de son contenu pour l'indexation. 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.

```
print('le text avant traitement')
print('-----')
print(tdf['text'][1])
print('-----')
print('le text après traitement')
print('-----')
tdf.loc[:, "text"] = tdf["text"].apply(cleanTxt)
print(tdf['text'][1])
```

```
le text avant traitement
-----
@ffsebg19 Hi there 🤖, I am sorry for the delay in response. Entry rules and Covid-19 testing is a government specif...
-----
le text après traitement
-----
19 Hi there I am sorry for the delay in response Entry rules and Covid19 testing is a government specif...
```

Figure 14 : traitement d'un twitter

### 2.3-Indexation :

L'indexation consiste à choisir les termes représentatifs des tweets de notre dataset créée 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 processus de recherche. Un index est un ensemble de tweets analysés et traités. Un document est un ensemble de champs (Fields) auxquels sont associées des valeurs.

### 2.4-Appariement document-requête :

Le processus d'appariement Tweet-requête permet de mesurer la pertinence d'un Tweet vis-à-vis d'une requête. De manière générale, à chaque réception d'une requête, le système de recherche calcule un score de pertinence thématique (via modèle BM25). Ce score de pertinence est calculé à partir d'une fonction ou d'une mesure de similitude, notée RSV (Q;T) (Retrieval Status Value) où Q est une requête et T un Tweet de la collection. Le processus d'appariement est étroitement lié au processus d'indexation et de pondération des termes. La relation d'appariement consiste à rechercher parmi les documents prétraités, ceux qui répondent le mieux à la requête.

### 3-Contenu social :

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

- **Ressources** :  $R = \{D_1, D_2, \dots, D_n\}$  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  $D = \{w_1, w_2, \dots, w_3\}$  où w est un terme, et

## CHAPITRE III : Notre processus RIS

---

comme un ensemble de caractéristiques sociales réalisées sur cette ressource,  $D = \{a_1, a_2, \dots, a_m\}$  où  $a$  est une action relevant d'activité sociale.

- **Signaux sociaux** : un ensemble,  $S = \{s_1, s_2, \dots, s_m\}$ , représente  $m$  contenus sociaux que les utilisateurs peuvent effectuer sur les ressources. Ces contenus représentent la relation entre l'ensemble des utilisateurs  $U = \{u_1, u_2, \dots, u_h\}$  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.
- **Emotion** :  $E = \{-1, 0, 1\}$  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.
- **Twitter** : Twitter est un réseau social qui contient un ou plusieurs signaux sociaux spécifiques réalisés sur une ressource.

### 4-profile document :

Le modèle utilisée pour créer le profil de document est présenté par un profil thématique, un profil social et un profil émotionnel :

- **Profil document thématique :**

Le profil thématique du document consiste en un ensemble de termes qui représente un Tweet  $d$ . L'ensemble des termes est prédéfini et a une longueur fixe à 140 caractères. Les termes du document comprennent un ensemble de mots-clés représentant son contenu.

- **Profil document social :**

Le profil social du document consiste en un ensemble de clics qui a réagi au document  $d$ . On note cet ensemble  $A_i = \{a_1; a_2; \dots; a_{n2}\}$ , le nombre d'actions qui reflètent les interactions sur le document  $d$  indique sa popularité dans les réseaux sociaux

- **Profil document émotionnel :**

Le sentiment d'un tweet est traduit par l'un des indicateurs (-1, 0, +1) qui exprime sentiment positif, négatif ou neutre.

### 4.1-Pertinence classique :

#### 4.1.1-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

## CHAPITRE III : Notre processus RIS

---

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'<sup>1</sup> [68].

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

**Figure 15 :** Formule Okapi BM25

- $tf(q_i, d)$  est en corrélation avec la fréquence du terme, définie comme le nombre de fois que le terme est interrogé  $q_i$  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.
- $k_1$  et  $b$  sont des paramètres libres, habituellement choisis comme  $k_1 = 2,0$  et  $b = 0,75$ .
- $idf(q_i)$  est le poids de fréquence inverse du terme de requête  $q_i$ . Il est calculé par :

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

**Figure 16 :** Formule idf.

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

---

<sup>1</sup> : <https://pypi.org/project/rank-bm25/>

## CHAPITRE III : Notre processus RIS

---

### 4.2-Pertinence sociale :

Le profil social du Tweet consiste en un ensemble de traces  $A_i = \{a_1 ; a_2; \dots; a_n\}$  laissées par les utilisateurs. Le nombre d'interactions sur le Tweet T indique sa popularité sur Twitter.

L'enrichissement du classement social des Tweets consiste à intégrer ces  $A_i$  pour définir le facteur social à prendre en compte dans le calcul de la pertinence.

Le facteur de pertinence sociale se calcule comme suit :

$$Pertinence_{sociale} = Pertinence (favorite) \cup Pertinence (retweet) = P(F) \cup P(RT)$$

avec :

$$\text{Score de favorite : } P(F) = \frac{ccccccccc(\text{ favorite }) * 111111}{\sum_{k=1}^M ccccccccc (fffffccfffcff) + \sum_{k=1} ccccccccc (fffcrrfffc)}$$

$$\text{Score de retweet : } P(RT) = \frac{ccccccccc (fffcrrfffc) * 100}{\sum_{k=1}^M ccccccccc (fffffccfffcff) + \sum_{k=1} ccccccccc (fffcrrfffc)}$$

### 4.3-Pertinence émotionnelle :

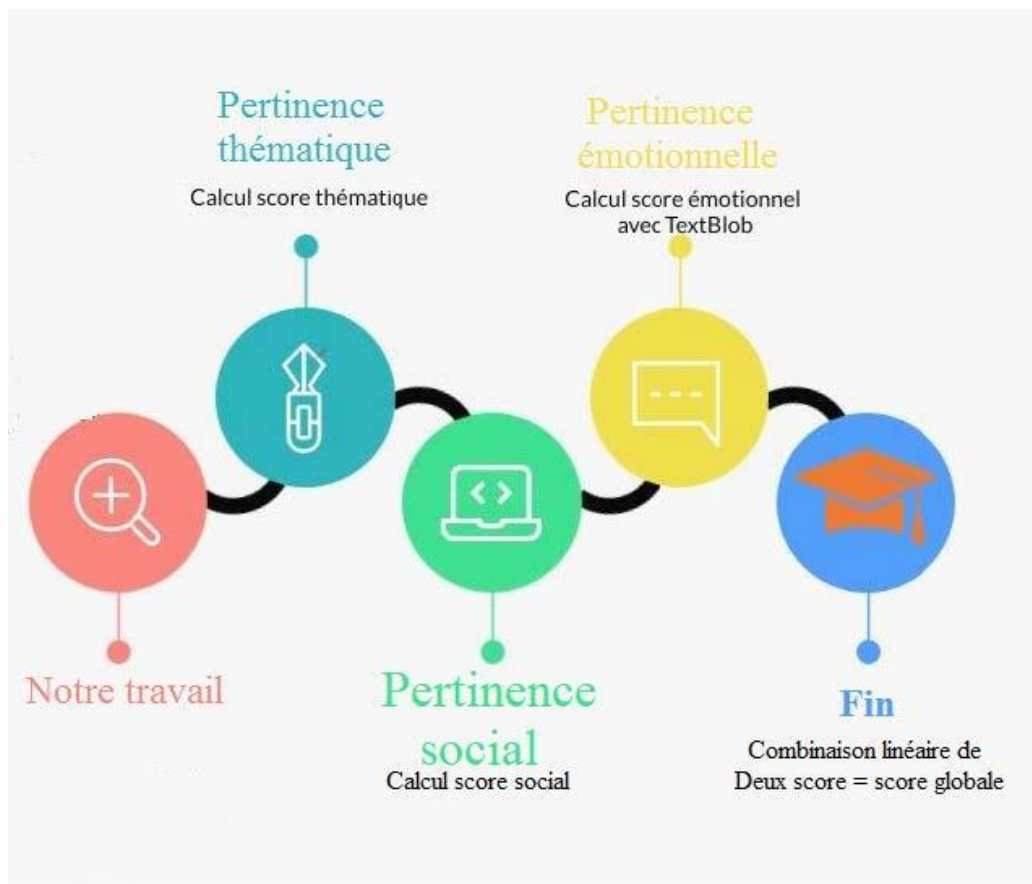
La détermination de la polarité des tweets est représentée sur une échelle de -1 à 1, le bas de l'échelle indique les réponses négatives, le haut de l'échelle les réponses positives et le 0 indiquent les réponses neutres.

$P(E)$  = Score émotionnelle qui exprime le sentiment des réactions individuelles et le point de vues général sur un tweet

### 5-Pertinence globale :

La pertinence globale est une combinaison linéaire des pertinences Socio-Emot-Thématique :

$$Pertinence_{globale} = P(T) \cup P(S) \cup P(E)$$



**Figure 17 :** Time line de notre SRI

### 6- Conclusion :

Afin d'améliorer les résultats de recherche retournés selon un critère classique, nous avons intégré des critères sociaux-émotionnelles au sein du modèle de base où nous nous sommes partis de l'hypothèse Qu'une ressource web doit être reclassée en fonction d'une combinaison linéaire de trois 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 cette combinaison des valeurs des différentes pertinences. Le chapitre suivant expose le volet pratique de notre processus RI amélioré en diffusant et évaluant les résultats obtenus.



# CHAPITRE VI :

# IMPLEMENTATION

# CHAPITRE VI : Implémentation

---

## 1- Introduction :

Ce chapitre est consacré à la partie implémentation et mise en œuvre de notre processus de RI amélioré où nous présentons, d'une part, les choix effectués de certains outils de développement (collections, logiciels, langages, ...), et d'autre part, les résultats collectés.

## 2-Présentation de l'environnement utilisé :

Dans cette partie nous allons détailler les différents outils utilisés pour la réalisation de notre PFE:

### 2.1-Outils de développement :

- **Pycharm Community :**

Pycharm<sup>10</sup> est un environnement de développement intégré utilisé en programmation informatique, en particulier pour le langage Python. Il est développé par la société tchèque JetBrains<sup>8</sup> [69]

- **Jupyter notebook :**

Project Jupyter<sup>11</sup> 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<sup>9</sup> [70].

- **Colaboratory :**

**Colab**, est un produit de Google Research. permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. C'est un environnement particulièrement adapté à la machine Learning, à l'analyse de données et à l'éducation<sup>10</sup>. [71]

---

<sup>8</sup> [www.jetbrains.com](http://www.jetbrains.com)

<sup>9</sup> <https://jupyter.org>

<sup>10</sup> <https://research.google.com › colaboratory>

## CHAPITRE VI : Implémentation

---

### 2.2-Langages de programmation :

- **Python** : Python 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<sup>12</sup> [73]

### 2.3-Bibliothèques principales :

- **Tweepy** :

Tweepy 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<sup>13</sup> [74]

- **Pandas** :

Pandas 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<sup>14</sup> [75].

- **Numpy** :

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<sup>15</sup> [76]

- **Rank\_bm25** :

BM25 est un package Python simple et peut être utilisé pour indexer les données, tweets dans notre cas, en fonction de la requête de recherche<sup>16</sup> [77].

---

<sup>12</sup> <https://docs.python.org>

<sup>13</sup> <https://docs.tweepy.org>

<sup>14</sup> <https://pandas.pydata.org>

<sup>15</sup> <https://numpy.org>

<sup>16</sup> <https://pypi.org/project/rank-bm25>

## CHAPITRE VI : Implémentation

---

- **Textblob** : TextBlob 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<sup>17</sup> [78]

- **Nltk** :

Natural Language Toolkit<sup>18</sup>, 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<sup>18</sup> [79].

- **Matplotlib** :

Est une bibliothèque multiplateforme de visualisation de données et de traçage graphique pour Python et son extension numérique NumPy. En tant que tel, il offre une alternative open source viable à MATLAB. Les développeurs peuvent également utiliser les API (Application Programming Interfaces) de matplotlib pour intégrer des tracés dans des applications GUI<sup>19</sup>. [80].

- **Seaborn** :

Est une bibliothèque qui utilise Matplotlib en dessous pour tracer des graphiques. Il sera utilisé pour visualiser des distributions aléatoires<sup>20</sup>. [81]

### 3-Création du dataset :

La figure ci dessous (figure 16) présente une portion du code source python pour la description et la définition du dataset brut.

---

<sup>17</sup> <https://textblob.readthedocs.io/en/dev>

<sup>18</sup> <https://www.nltk.org/>

<sup>19</sup> <https://matplotlib.org>

<sup>20</sup> <https://michaelwaskom.medium.com>

# CHAPITRE VI : Implémentation

```
# We create a tweet list as follows:
#,since="2020-01-01",until=datetime.date(datetime.now())
#tweets = tweepy.Cursor(api.search,q="algeria",lang="en").items(1000)

#search_term = "#covid-19 follow -filter:retweets filter:verified"
search_term = "#covid-19 follow -filter:retweets"
tweets = tweepy.Cursor(api.search,
                        q=search_term,
                        since="2019-01-1",
                        until=datetime.date(datetime.now()),
                        lang='en').items()
```

Figure 18 : Code source python de création du dataset.

### 3.1-Enregistrement des données :

La figure 19 présente le stockage dans une structure tabulaire CSV des tweets collectés avant son nettoyage des données indésirables et inutiles (biaisées et bruyante).

	created_at	id	id_str	text	truncated	entities	metadata	source
0	Mon May 23 23:53:54 +0000 2022	1528886971946237952	1528886971946237952	Watch Out! Here Comes a Wave Of Infections, Co...	True	{'hashtags': [{'text': 'COVID19', 'indices': [...	{'iso_language_code': 'en', 'result_type': 're...	<a href="https://mobile.twitter.com" rel="nofo...
1	Mon May 23 23:40:36 +0000 2022	1528883626015870976	1528883626015870976	@KBonasera72 @wbz You're wrong. In national sur...	True	{'hashtags': [], 'symbols': [], 'user_mentions...	{'iso_language_code': 'en', 'result_type': 're...	<a href="http://twitter.com/download/android" ...
2	Mon May 23 23:33:37 +0000 2022	1528881869651349504	1528881869651349504	@UnchainedCub Hello man, Covid-19 is now suppr...	False	{'hashtags': [], 'symbols': [], 'user_mentions...	{'iso_language_code': 'en', 'result_type': 're...	<a href="http://twitter.com/download/android" ...
3	Mon May 23 23:00:01 +0000 2022	1528873414701658114	1528873414701658114	Here's what you need to know about #COVID19 se...	True	{'hashtags': [{'text': 'COVID19', 'indices': [...	{'iso_language_code': 'en', 'result_type': 're...	<a href="https://mobile.twitter.com" rel="nofo...
4	Mon May 23 22:59:04 +0000 2022	1528873172749037570	1528873172749037570	https://t.co/YGcJ5uCA48 https://t.co/rBRJMPZ5n...	False	{'hashtags': [], 'symbols': [], 'user_mentions...	{'iso_language_code': 'en', 'result_type': 're...	<a href="https://d1vrit.com" rel="nofollow">d...
...	...	...	...	...	...	...	...	...
660	Mon May 16 10:00:09	1526140438081316611	1526140438081316611	"Without trust, politicians struggle to	True	{'hashtags': [], 'symbols': [], 'user_mentions...	{'iso_language_code': 'en', 'result_type': 're...	<a href="https://www.hootsuite.com" ...

Figure 19 : Fichier csv des tweets collectés

## CHAPITRE VI : Implémentation

### 4-Prétraitement des données :( preprocess string )

Le prétraitement (figure 20) dans sa globalité est une référence pratique au filtrage des données avant leurs utilisations.

Elle constitue une étape importante du processus d'analyse des données.

```
print('le text avant traitement')
print('-----')
print(tdf['text'][1])
print('-----')
print('le text après traitement')
print('-----')
tdf.loc[:, "text"] = tdf["text"].apply(cleanTxt)
print(tdf['text'][1])
```

le text avant traitement  
-----  
@ffsebg19 Hi there 🤔, I am sorry for the delay in response. Entry rules and Covid-19 testing is a government specif...  
-----  
le text après traitement  
-----  
19 Hi there I am sorry for the delay in response Entry rules and Covid19 testing is a government specif...

Figure 20 : prétraitement du texte (Tweet)

Le prétraitement 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 (figure 21). La qualité du prétraitement a une grande influence sur le résultat final de l'ensemble du processus<sup>21</sup>(figure 21).

```
# 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(' \# ', ' ', text) # Removing ' \# ' hash tag
    text = re.sub('#', '', text) # Removing '#' hash tag
    text = re.sub('RT[\s]+', '', text) # Removing RT
    text = re.sub('https?:\/\/\S+', '', text) # Removing hyperlink
    text = remove_emojis(text) #removing emojis
    text = text.replace('\n', ' ') # removing \n
    text = text.translate(str.maketrans('', '', string.punctuation)) # Remove Punctuation
    return text

tdf.loc[:, "text"] = tdf["text"].apply(cleanTxt)
tdf['text'][0]
```

'Watch Out Here Comes a Wave Of Infections Courtesy of COVID19 COVID19 CoronaVirus...'

Figure 21 : Prétraitement détaillé de texte

Parmi les fonctionnalités permises pour le nettoyage des données brutes , nous avons:

<sup>21</sup> <https://radimrehurek.com>

## CHAPITRE VI : Implémentation

---

**`strip_tags()`**, Supprimer les balises de chaîne de caractères `s` en utilisant `RE_TAGS`



`gensim.parsing.preprocessing.strip_tags(s)`

Remove tags from `s` using `RE_TAGS`.

Parameters

- `s (str)` –

Returns

Unicode string without tags.

Return type

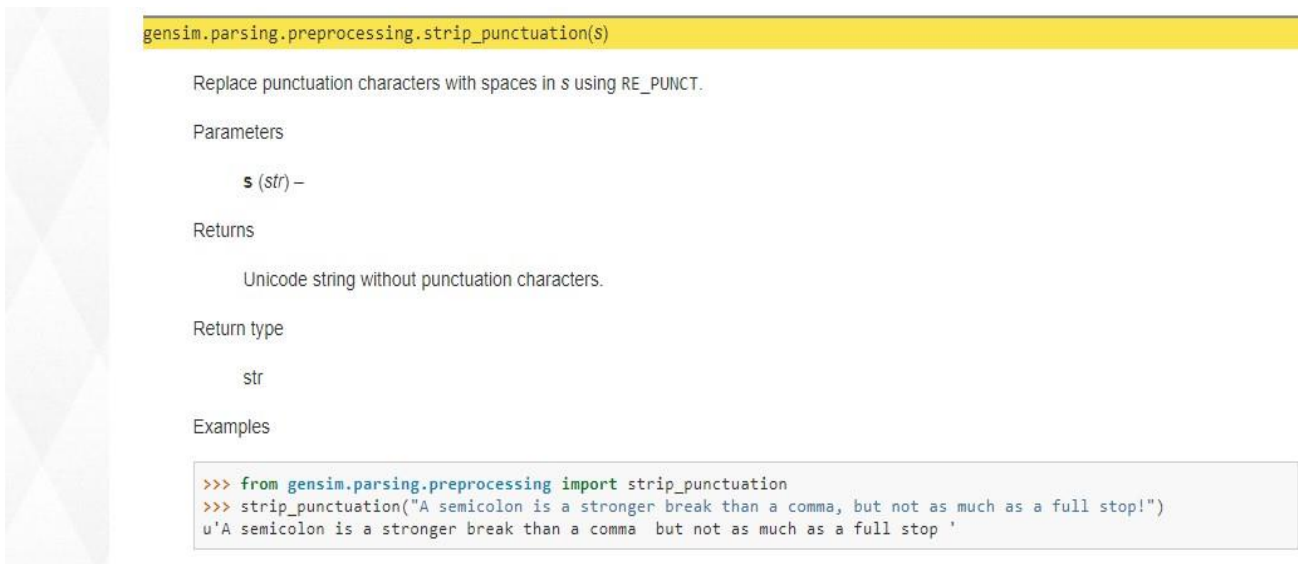
`str`

Examples

```
>>> from gensim.parsing.preprocessing import strip_tags
>>> strip_tags("<i>Hello</i> <b>World</b>!")
u'Hello World!'
```

**Figure 22** : suppression des balises de chaîne de caractères `s`

- **`strip_punctuation()`** Remplacer les caractères de ponctuation par des espaces dans string `s` en utilisant `RE_PUNCT`.



`gensim.parsing.preprocessing.strip_punctuation(s)`

Replace punctuation characters with spaces in `s` using `RE_PUNCT`.

Parameters

- `s (str)` –

Returns

Unicode string without punctuation characters.

Return type

`str`

Examples

```
>>> from gensim.parsing.preprocessing import strip_punctuation
>>> strip_punctuation("A semicolon is a stronger break than a comma, but not as much as a full stop!")
u'A semicolon is a stronger break than a comma but not as much as a full stop '
```

**Figure 23** : Remplacement des caractères de ponctuation par des espaces dans les chaînes de caractères `s`

## CHAPITRE VI : Implémentation

---

**`strip_multiple_whitespaces()`**, Supprime les caractères d'espacement répétitifs (espaces, tabulations, sauts de ligne) de string `s` et transforme les tabulations et les sauts de ligne en espaces en utilisant `RE_WHITESPACE`.

```
gensim.parsing.preprocessing.strip_multiple_whitespaces(s)
```

Remove repeating whitespace characters (spaces, tabs, line breaks) from `s` and turns tabs & line breaks into spaces using `RE_WHITESPACE`.

Parameters

- `s (str)` –

Returns

Unicode string without repeating in a row whitespace characters.

Return type

`str`

Examples

```
>>> from gensim.parsing.preprocessing import strip_multiple_whitespaces
>>> strip_multiple_whitespaces("salut" + '\r' + " les" + '\n' + "      loulous!")
u'salut les loulous!'
```

**Figure 24** : Suppression des caractères d'espacement répétitifs

- **`strip_numeric()`**, Retirer les chiffres de chaîne de caractères `s` en utilisant `RE_NUMERIC`.

```
gensim.parsing.preprocessing.strip_numeric(s)
```

Remove digits from `s` using `RE_NUMERIC`.

Parameters

- `s (str)` –

Returns

Unicode string without digits.

Return type

`str`

Examples

```
>>> from gensim.parsing.preprocessing import strip_numeric
>>> strip_numeric("0text24gensim365test")
u'textgensimtest'
```

**Figure 25** : retirer les chiffres de chaîne de caractères



## CHAPITRE VI : Implémentation

---

- **remove\_stopwords()**, Supprimer les mots d'arrêt de chaîne de caractères *s*.

```
gensim.parsing.preprocessing.remove_stopwords(s)
```

Remove STOPWORDS from *s*.

Parameters

- ***s*** (*str*) –

Returns

Unicode string without STOPWORDS.

Return type

*str*

Examples

```
>>> from gensim.parsing.preprocessing import remove_stopwords
>>> remove_stopwords("Better late than never, but better never late.")
u'Better late never, better late.'
```

**Figure 26** : Suppression des mots d'arrêt de chaîne de caractères .

- **strip\_short()**, Enlever de string *s* des mots dont la longueur est inférieure à *minsize*

```
gensim.parsing.preprocessing.strip_short(s, minsize=3)
```

Remove words with length lesser than *minsize* from *s*.

Parameters

- ***s*** (*str*) –
- ***minsize*** (*int, optional*) –

Returns

Unicode string without short words.

Return type

*str*

Examples

```
>>> from gensim.parsing.preprocessing import strip_short
>>> strip_short("salut les amis du 59")
u'salut les amis'
>>>
>>> strip_short("one two three four five six seven eight nine ten", minsize=5)
u'three seven eight'
```

**Figure 27** : Suppression de chaîne de caractère *s* des mots dont la longueur est inférieure à *minsize*

## CHAPITRE VI : Implémentation

- **stem\_text()**. Transformer le string `s` en minuscule et le mettre en tige.

```
gensim.parsing.preprocessing.stem_text(text)
```

Transform `s` into lowercase and stem it.

Parameters

**text** (*str*) –

Returns

Unicode lowercased and porter-stemmed version of string `text`.

Return type

`str`

Examples

```
>>> from gensim.parsing.preprocessing import stem_text
>>> stem_text("While it is quite useful to be able to search a large collection of documents almost instantly.")
u'while it is quit us to be abl to search a larg collect of document almost instantly.'
```

**Figure 28** : Transformation des chaînes de caractères en minuscules

### 5-Pertinence thématique :

Cette pertinence fonctionne autour de la fréquence du terme recherché et des tweets indexés, figure ci dessous..

```
[ ] meta_df_tokens
```

```
0      [china, call, covid, 'lab, leak', theori, lie,...
1      [nationalsafetymonth, great, remind, covid, sa...
2      [follow, link, covid, travel, updat]
3      [japan, fridai, eas, border, foreign, tourist,...
4      [japan, fridai, eas, border, foreign, tourist,...
...
125     [middaynew, centr, ask, state, monitor, cluste...
126     [centr, ask, state, monitor, cluster, covid, c...
127     [remind, student, selfscreen, covid, symptom, ...
128     [covid, ensur, healthi, tomorrow, follow, covi...
129     [coronaviru, new, case, addit, test, report, u...
Name: text, Length: 130, dtype: object
```

**Figure 29** : dataset indexé

La figure 27 présente le prétraitement d'une requête soumise au système de recherche:

## CHAPITRE VI : Implémentation

---

```
[29] #query bruit
      query = 'study is looking for covid-19 '

      #cleand query
      preprocess_string(query)

      ['studi', 'look', 'covid']
```

Figure 30 : Prétraitement du coté module requête

### 5.1-La similarité BM25 :

Il est donc important de passer par l'indexation pour utiliser la fonction Search (Query, doc)..

```
from rank_bm25 import BM25Okapi
import numpy as np

bm25_index = BM25Okapi(meta_df_tokens.tolist())

def search(search_string, bm25_index):
    # clean queru
    search_tokens = preprocess_string(search_string)
    # calculate The thematique score
    tdf['score_thématique'] = pd.Series(bm25_index.get_scores(search_tokens))
    # ordering the list accending
    top_indexes = np.argsort(bm25_index.get_scores(search_tokens))[:, :-1][:]
    return top_indexes
```

Figure 31 : Exploitation de la mesure de similarité BM25 Rank.

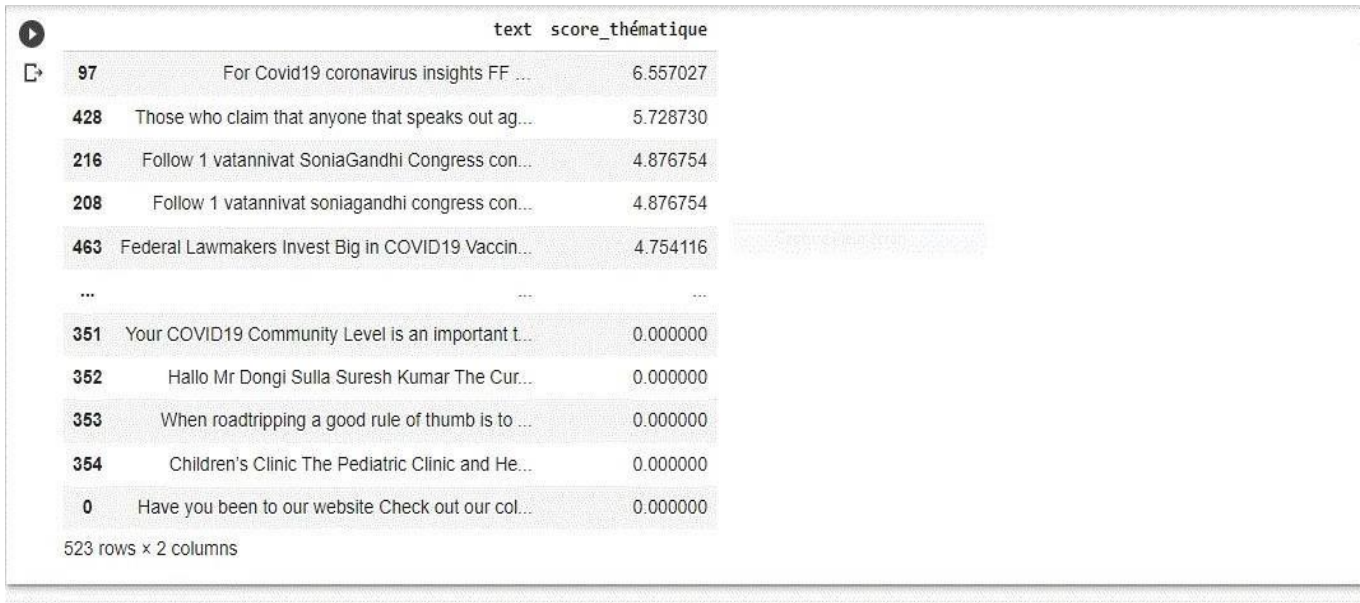
## 6-Résultats retournés :

### 6.1-Score thématique :

NOUS avons mené des expérimentations avec le modèle de référence BM25.

Une liste de résultats est retournée et classée dans un ordre décroissant, du tweet plus pertinent au moins pertinent, en fonction de la valeur de son score thématique.

# CHAPITRE VI : Implémentation



	text	score_thématique
97	For Covid19 coronavirus insights FF ...	6.557027
428	Those who claim that anyone that speaks out ag...	5.728730
216	Follow 1 vatannivat SoniaGandhi Congress con...	4.876754
208	Follow 1 vatannivat soniagandhi congress con...	4.876754
463	Federal Lawmakers Invest Big in COVID19 Vaccin...	4.754116
...	...	...
351	Your COVID19 Community Level is an important t...	0.000000
352	Hallo Mr Dongi Sulla Suresh Kumar The Cur...	0.000000
353	When roadtripping a good rule of thumb is to ...	0.000000
354	Children's Clinic The Pediatric Clinic and He...	0.000000
0	Have you been to our website Check out our col...	0.000000

523 rows x 2 columns

Figure 32 : Résultat du score thématique via BM25.

## 6.2-Score sociale :

Le classement sociale est réalisé suite au cumul de comptage de favorite et retweet .



```
#favorite percentage
indexes_df['favorite_percentage']=indexes_df['favorite_count']/((indexes_df['favorite_count'].sum()+indexes_df['retweet_count'].sum()))
#retweet percentage
indexes_df['retweet_percentage']=indexes_df['retweet_count']/((indexes_df['favorite_count'].sum()+indexes_df['retweet_count'].sum())*100

indexes_df['score_social']=indexes_df['favorite_percentage']+indexes_df['retweet_percentage']

indexes_df.sort_values(by=['score_social'],ascending=False)[['favorite_percentage','retweet_percentage','score_social']]
```

	favorite_percentage	retweet_percentage	score_social
355	17.685851	16.594724	34.280576
346	8.729017	2.026379	10.755396
307	4.544365	0.407674	4.952038
491	3.824940	0.959233	4.784173
4	4.052758	0.407674	4.460432
...	...	...	...
402	0.000000	0.000000	0.000000
156	0.000000	0.000000	0.000000
155	0.000000	0.000000	0.000000
405	0.000000	0.000000	0.000000
0	0.000000	0.000000	0.000000

Activer Windows  
Accédez aux paramètres pour activer Windows.

## CHAPITRE VI : Implémentation

Figure 33 : Code source et résultat du classement social.

### 6.3-Score émotionnelle :

TextBlob est une bibliothèque Python qui renvoie le taux de polarité globale de chaque tweet. Cette polarité reflète un sentiment traduit sur un tweet.

```
#Sentiment Analysis using TextBlob
from textblob import TextBlob

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

tdf["score_emotional"]=tdf["text"].apply(getPolarity)
tdf
```

Figure 34 :utilisation de bibliothèque de textblob.



	text	score_emotional
0	Have you been to our website Check out our col...	0.050000
1	22221 gabai My claim that it wasn't approved...	0.000000
2	Follow precautions amp COVID19 guide line PI ...	0.000000
3	This is something that Dr Luc Montagnier was p...	0.083333
4	This morning I tested positive for COVID19 For...	0.240152
...	...	...
518	Gavin Newsom has had 4 shots He is allowed to ...	0.700000
519	Covid 19 isnt over so follow precautionary mea...	0.400000
520	Clinical Support through Telemedicine in Heart...	-0.316667
521	Has Covid19 got you feeling down This bot aim...	-0.155556
522	national Honestly if the messaging was clear t...	0.233333

523 rows x 2 columns

Figure 35 : Estimation du score émotionnel.

### 6.4-Score globale :

Les résultats retournés ont fourni des preuves théoriques de l'importance de la RI dans le contexte des médias sociaux. Selon ces résultats, nous aurons un réordonnement plus intéressant:

## CHAPITRE VI : Implémentation

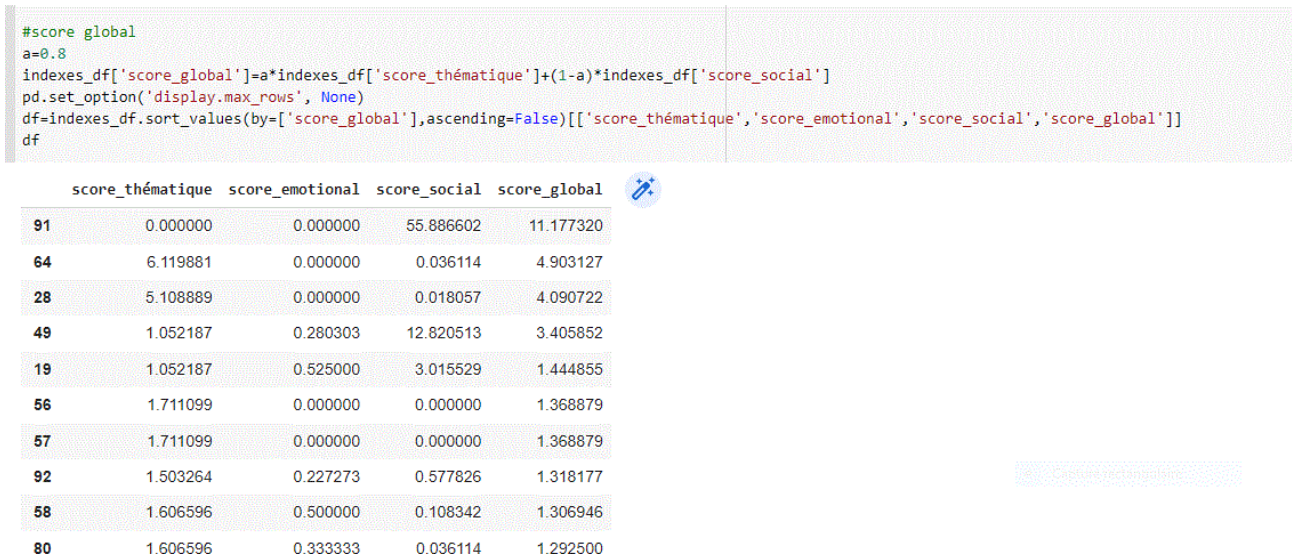


Figure 36 : Code source et résultats du score global.

La figure 36 présente le classement global qui réorganise les tweets:

	score_thématique	score_emotional	score_social	score_global
33	1.061776	0.100000	29.623824	30.685600
126	0.890060	0.000000	14.890282	15.780342
127	0.968364	0.240152	13.871473	14.839838
42	0.000000	0.193182	8.960293	8.960293
107	8.006968	0.000000	0.208986	8.215954
99	8.006968	0.000000	0.182863	8.189831
98	8.006968	0.000000	0.052247	8.059215
90	4.764154	0.250000	0.026123	4.790277
83	1.061776	0.500000	3.709509	4.771284
31	1.454679	0.000000	2.795193	4.249873
11	3.801050	0.037273	0.052247	3.853297
60	0.927562	0.000000	1.671891	2.599454
4	0.000000	0.125000	2.246604	2.246604
88	1.115582	0.800000	1.123302	2.238884
123	1.175132	0.250000	0.679206	1.854338

Figure 37 : Résultats des tweets via le score global.

## **CHAPITRE VI : Implémentation**

---

### **7-Conclusion :**

Dans ce chapitre, nous avons décrit la mise en œuvre des expérimentations réalisées pour valider notre travail.

## Conclusion générale

Dans ce mémoire, nous avons abordé une thématique qui a vu le jour avec l'arrivée des réseaux sociaux, il s'agit de la recherche d'information sociale.

Pour notre objectif, La RIS consiste à exploiter la dimension sociale en deux temps. Dans un premier temps, nous utilisons le contenu, les réactions(clics) et le sentiment de chaque Tweet pour en définir son profil correspondant. Deuxièmement, nous exploitons ce profil dans le classement des résultats de recherche d'information en combinant la pertinence thématique et la pertinence sociale dans le calcul du score global d'un Tweet. Dans notre cas, le Tweet, n'est plus un simple texte indépendant de son contenu. Il est complètement y lié via les réactions générées (clics) par les utilisateurs.

Pour la création de la collection des données, Le problème majeur rencontré lors de la phase de collecte est la non disponibilité et non facilité d'accès à certaines API de quelques réseaux sociaux comme l'API de Facebook et l'accès pénible à l'api Twitter en temps réel.

Comme travaux futurs, nous comptons:

- Elargir le contexte social que ce soit du coté document ou coté utilisateur pour en contenir d'autres interactions sociales à exploiter autres que les clics et qui pourraient être significatives pour d'autres domaines et d'autres réseaux sociaux.
- Tester d'autres modèles de RI (modèle de correspondance) autres que le BM25
- Evaluer notre solution RIS sur des collections de test standard conçues dans le cadre de la RI.



### Bibliographie

- [1]-**N.Hernandez**, Ontologie de domaine pour la modélisation du contexte en recherche d'information. (Université Paul Sabatier, 2006)
- [2 ]-**G.Salton** , Automatic Information Organization and Retrieval, (McGraw Hill Text, 1968)
- [3]-**R.A. Baeza-Yates, B. Ribeiro-Neto**, Modern Information Retrieval, (Longman Publishing C, Inc, Boston, USA, 2010)
- [4]-**Bouramoul Abdelmalek**, recherche d'information, these de doctorat, (Constantine :Université Mentouri , 2011)
- [5]-**Ingwersen Peter**, Information retrieval interaction,(London : Taylor graham Publishing, 1992)
- [6]-**Boughanem M et Savoy J**,Recherche d'information états des lieux et perspectives. , (Hermès Science Publications, 2008)
- [7]-**Tamine Lechani, Lynda et Calabretto Sylvie**,Recherche d'information contextuelle et web
- [8]-**Damak Firas**,Etude des facteurs de pertinence dans la recherche de microblogs. (Toulouse , 2014)
- [9]-**Buckley C et G Salton**,Term weighting approches in automatic text retrieval, 1988
- [10]-**Zipf et George K**,Hman Behaviour and the principal of least effort. (USA , 1949)
- [11]-**Maron M, E et Kuhns**, On relevance, probabilistic indexing and information retrieval,1960.
- [12] **Porter M**, An algorithm for suffix stripping , 1980
- [13]-**Mayfield, J et McNamee P**,Single n-gram stemming in proccedings of the 26th annual international ACM SiGIR Conference on research and development in information retrieval,( New York , 2003)
- [14]-**Manning, C Raghavan et Schtze H**,Introduction to information retrieval,( New York : Cambridge university press, 2008)

## Bibliographie

---

- [15]-**Baeza Yates R A et Ribeiro Neto**, Modern information retrieval ,(England , Pearson education Ltd, 2011)
- [16]-**Hammache Arezki**, un modèle de langue combinant mot simples et mots composés,These de doctorat (Tizi-Ouzou )
- [17]-**Salton G**,The Smart Retrieval System experiments in Automatic Document Processing, (New Jersey , 1971)
- [18]-**Van Rijsbergen**, Information retrieval, (London :Butterworth, 1979)
- [19]-**Abbassi Meftah**, Un modèle de reformulation des requêtes pour la recherche d'information sur le Web,(S.D)
- [20]-**Hachemi Hadjira et Rimouche et Nour El Houda**,Moteur de recherche sémantique, 2013
- [21]-**Robertson Se et Sparch Jones**,«Journal of the American Society for Information», 1976
- [22]-**Robertson, S**,The probability ranking principle in information retrieval, 1977
- [23]-**Zemirli Nesrine**,Vers le développement d'un système de recherche d'information personnalisé intégrant profile d'utilisateur, these de doctorat ,( Université Paul Sabatier , 2004)
- [24]-**Bruce Croft et Turtle, Howard R et Lewis David D**,The Use of Phrases and Structured Queries in Information Retrieval, 1991.
- [25]-**Rocchio J**,Relevance feedback information retrieval, 1971
- [26]-**Harman D**,Relevance feedback revisited. Proceedings of ACM SIGIR , 1992
- [27]-**BoughanemM et Chrisment C et Soule-dupuy C**, Query modification based on Relevance Back-propagation in ad-hoc environment, (IPM: Information Process and Management, 1998)
- [28]-**A Maaradji et H Hacid et R Skraba**, A Vakali.Social web mashups full complete via frequent sequence mining , 2011

[29] **W Stanley et F Katherine**, Social Network Analysis: Methods and Applications (Structural Analysis in the Social Science , (New York: Kambridge University , 1994)

[30] **D Goh et S Foo**, Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively, Information Science Reference , (IGI Publishing, Hershey, PA)

[31] **Mohamed Reda Bouadjenek et Hakim Hacid et Mokrane Bouzeghoub** ,Social networks and information retrieval. how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms , (France :University of Montpellier ,2015)

[32] **MR Morris J Teevan et K Panovich**, Que demandent les gens à leurs réseaux sociaux, et pourquoi ? Une étude d'enquête sur le comportement des questions et réponses sur les messages d'état, dans : Actes de la conférence SIGCHI sur les facteurs humains dans les systèmes informatiques, CHI '10, ACM, (New York, États-Unis, 2010)

[33]-**G Kumaran, VR Carvalho**, Réduire les requêtes longues à l'aide de prédicteurs de la qualité des requêtes, dans : Actes de la 32e conférence internationale ACM SIGIR sur la recherche et le développement dans la recherche d'informations, SIGIR '09, (ACM, New York, NY, États-Unis, 2009)

[34]-**C Lioma et R Blanco et M-F Moens**, Une approche d'inférence logique pour l'expansion des requêtes avec des balises sociales ,( ICTIR, 2009)

[35]-**S Jin, H. Lin et S Su**, Query expansion based on folksonomy tag cooccurrence analysis, ( IEEE International Conference on Granular Computing, 2009)

[36]-**A Mantrach, J-M. Renders**, a general framework for people retrieval in social media with multiple roles, (Berlin: Heidelberg, 2012)

[37]**Y. Lin et H. Lin, S. Jin et Z. Ye** , Annotation sociale dans l'expansion des requêtes : une approche d'apprentissage automatique, dans : Actes de la 34e conférence internationale ACM SIGIR sur la recherche et le développement dans la recherche d'informations, SIGIR ' 11, ACM,( New York :NY, États-Unis, 2011)

- [38]-**J Pitkow et H Schütze et T. Cass et R. Cooley, D. Turnbull, A. Edmonds et E Adar, T. Breuel** . Recherche personnalisée, 2002
- [39]-**M Bender, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, JX Parreira, R. Schenkel, G. Weikum**, Exploiting social relations for query expansion and result ranking ,( ICDE Workshops, 2008 )
- [40]-**R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, JX Parreira, G. Weikum**, Efficient top-k querying over social-tagging networks, Actes de la 31e édition annuelle de l'ACM international Conférence SIGIR sur la recherche et le développement dans la recherche d'informations, SIGIR '08, ACM, (New York, NY, États-Unis, 2008)
- [41]-**M. Bertier, R. Guerraoui, V. Leroy, A.-M. Kermarrec**, Vers une expansion personnalisée des requêtes SNS, 2009
- [42]-**X. He, M. Gao, M.-Y. Kan, Y. Liu et K Sugiyama**, Predicting the popular of web 2.0 items based on user comments, dans : Actes de la 37e conférence internationale ACM SIGIR sur la recherche et le développement en recherche d'information, SIGIR '14, ACM, (New York, NY , États-Unis, 2014)
- [43]-**PA Dmitriev, N. Eiron, M. Fontoura, E. Shekita**, Utilisation des annotations dans la recherche d'entreprise, dans : Actes de la 15e Conférence internationale sur le World Wide Web, WWW '06, ACM, (New York, NY, USA, 2006)
- [44]-**DH Dalip, MA Gonçalves, M. Cristo, P. Calado**, Exploiter les commentaires des utilisateurs pour apprendre à classer les réponses dans les forums de questions-réponses : une étude de cas avec débordement de pile : Actes de la 36e conférence internationale ACM SIGIR sur la recherche et le développement dans Information Retrieval, SIGIR '13, ACM,( New York, NY, États-Unis, 2013)
- [45]-**S Amer-Yahia, M Benedikt, LVS Lakshmanan, J. Stoyanovich**, Recherche efficace et consciente du réseau dans les sites de marquage collaboratif, (Proc. Dotation VLDB, 2008 )

## Bibliographie

---

[46]**Teevan,D.Ramage,M.R.Morris**, twittersearch: a comparison of microblog search and web search,in:Proceedings of the Fourth ACM International conference on Web Search and Data Mining, WSDM '11,ACM,(NewYork,NY,USA,2011)

[47]- **Shalev-Shwartz, Shai ET Shai Ben-David**. Understanding machine learning: From theory to algorithms,( Cambridge university press, 2014)

[48]-**Rosenberg, Morris et Ralph H, Turner** Social psychology: Sociological perspectives, (Transaction Publishers, 1990)

[49]-**Myers, David G**.Theories of emotion : Psychology , (New York, NY: Worth Publishers 500,2004)

[50]-**Liu, Bing**,Sentiment analysis and opinion mining: Synthesis lectures on human language technologies 5.1, 2012

[51]-**Liu, Bing et al**,Sentiment analysis and subjectivity: Handbook of natural language processing 2010

[52]-**LeCun, Yann, Yoshua Bengio et Geoffrey Hinton** , Deep learning,2015

[53]-**Kumar, Akshi et al**.Sentiment Analysis : A Perspective on its Past : Present and Future, 2012

[54]-**Hadji, Mehdi**, Analyse des sentiments,2019

[55]-**Rahab, Hichem** ,Fouille des données d'opinion appliquée à la classification des commentaires en arabe dans la presse en ligne.( Université de Constantine 2-Abdelhamid Mehri),(S.D)

[56]-**Chen, Tao et al**,Improving sentiment analysis via sentence type classification using, 2017

[57]-**Zhang, Changli et al**,Sentiment analysis of Chinese documents : From sentence to document level. (Journal of the American Society for Information Science and Technology, 2009)

## Bibliographie

---

- [58]-**Ramírez-Tinoco, Francisco Javier et al**,Use of sentiment analysis techniques in health care domain: Current Trends in Semantic Web Technologies: Theory and Practice. (Springer,2019)
- [61]-**Kumar Ela.**,Natural Language Processing,(India:I.K.International Publishing House Pvt. Ltd, 2011).
- [62] **Jean Véronis**,Natural Language Processing, [<http://sites.univ-provence.fr/veronis>], 2001
- [63]-**Meena Rambocas and Joo Gama**.The Role of Sentiment Analysis. FEP Economics and Management, 2013
- [64] The International Conference on Advanced Machine Learning Technologies (S.D)

## Bibliographie

---

### Webgraphie :

[59] [en ligne] <https://medium.com/@mehdihadji/analyse-des-sentiments>, (consulté juillet 2020)

[60] [en ligne] <https://biblio.univ-annaba.dz/wp-content/uploads/2020/01/These-Ziani-Amel.pdf>

[65] [en ligne] <http://smm.streamcrab.com> (consulté le juin. 26, 2020)

[66] [en ligne] <https://www.vneuron.com/2018/01/23/introduction-a-lapprentissage-automatique-avec-scikit-learn/>, consulté( Octobre 2020)

[67] [en ligne] <http://www.sentiment140.com>

[69] [en ligne] <https://www.jetbrains.com/help/pycharm/quick-start-guide.html>

[70] [en ligne] <https://jupyter.org/documentation>

[71] [en ligne] <https://research.google.com/colaboratory>

[72] [en ligne] <https://docs.python.org/3/tutorial>

[73] [en ligne] <https://docs.tweepy.org/en/stable>

[74] [en ligne] <https://pandas.pydata.org/docs>

## Bibliographie

---

[75] [en ligne] <https://numpy.org/doc>

[76] [en ligne] <https://pypi.org/project/rank-bm25>

[77] [en ligne] <https://textblob.readthedocs.io/en/dev>

[78] [en ligne] <https://www.nltk.org>

[79] [en ligne] <https://matplotlib.org>

[80] [en ligne] <https://michaelwaskom.medium.com>

[81] [en ligne]

[https://radimrehurek.com/gensim\\_3.8.3/parsing/preprocessing.html#gensim.parsing.preprocessing](https://radimrehurek.com/gensim_3.8.3/parsing/preprocessing.html#gensim.parsing.preprocessing)