République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Borj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique


**MEMOIRE**

Présenté en vue de l'obtention du diplôme

**Master en informatique**

Spécialité : Réseaux et Multimédia

# THEME

# Réalisation d'un moteur de recherche avec une approche sémantique

**Présenté par** :

ZIANI Chaima

Soutenu publiquement le : 03/07/2022

**Devant le jury composé de** :

ATTIA Safa

SAIDANI Kaouthar

**Encadreur :** Mme.BELALTA Ramla

**2021/2022**

# Dédicace

Je dédie ce modeste travail, comme preuve de respect, de gratitude et de reconnaissance :

Pour ceux dont la supplication a été le secret de mon succès

A mon cher père qui a travaillé dans le désert toute sa vie pour nous donner une meilleure vie, je suis fier de porter son nom.

A ma chère mère, qui m'a nourri d'amour et de tendresse et qui a été comme une mère et un père en l'absence de mon père, je suis tellement reconnaissante pour ses sacrifices pour nous garantie une vie riche en bonheur.

Il n'y a rien au monde comme les efforts consentis pour mon éducation et mon bien-être et aucune dédicace ne peut exprimer mes sentiments profonds à leur égard.

A mes chers frère « **Mohamed** , **Sidali et Farid**» et mes précieuses sœurs

«**Linda** , **Nousseiba et Rima**» , et ma défunte sœur «**Fahima**»

A mes chers neveux et nièces

«**Abdou**, **Djihad**, **Wail**, **Ritadj**, **Tasnim et Yakoub**»

Et mes chers petits neveux et nièces « **Adam**, **Ghoufran**, **Nouh** »

Que Dieu les protège

À mon fiancé, le plus beau cadeau qu'Allah m'a donné « **Khalil**»

A mes sœurs de Coeur « **Zohra, Sarah et Afnane** »

Au frère que ma mère na pas enfante « **Taib** »

Aux beaux frères « **Ammar, Amine** »

A les beaux sœurs« **Rabia, Houda** »

A ma promotrice Mme **BELALTA Ramla** pour sa patience, sa disponibilité qu'elle a consacré pour l'élaboration de ce travail.

A toute personne qui m'a appris une lettre et m'a fait partager son savoir.

A ceux qui me connaissent de près ou de loin

*Chaïma*

# Remerciement

Avant tous, nous remercions **ALLAH**, pour nous avoir données la capacité

Pour réaliser ce travail.

En second lieu, nous tenons à remercier plus particulièrement Mme **BELALTA Ramla** pour son orientation, sa confiance, sa patience et sa supervision tout au long de notre mémoire

Nous tenons à remercier les membres de jury pour avoir accepté d'évaluer mon modeste travail et remercier tous les enseignants de département d'**informatique** par leurs compétences nous ont soutenu dans la poursuite de nos études et pour leurs efforts fournis pour faire de nous des futurs informaticiens.

J'adresse mes remerciements les plus sincères à toute personne ayant participé de près ou de loin durant mon cursus universitaire.

Mes remerciements les plus sincères à mes très chers parents qui ont été toujours là pour m'aider à réussir et atteindre mes rêves depuis mon premier jour à l'école jusqu'aujourd'hui.

Enfin, nous souhaitons remercier du fond du cœur nos familles et nos amies.

Ce mémoire est le fruit de cinq années d'étude.

Merci à toutes et à tous

# Résumé

Aujourd'hui, si vous voulez interagir avec un ordinateur, vous devez apprendre un langage de programmation, mais au lieu de cela nous voudrions maintenant que la machine comprend ce que nous disons ou plutôt, ce que nous écrivons lorsqu'on cherchons des données, et c'est le but de notre projet « réalisation d'un moteur de recherche avec une approche sémantique » qui essaye de comprendre le sens des requêtes des utilisateurs pour répondre avec des résultats plus pertinents qui reflètent l'intention de l'utilisateur.

Afin de réaliser notre application, nous avons utilisé la technique de Word embedding qui est largement connue et utilisée dans la recherche textuelle qui exploite la similarité sémantique.

Notre application est basée sur le traitement de langage naturel (NLP) et la recherche d'information (IR)

**Mots clés** :

La recherche d'information (IR), moteur de recherche sémantique, apprentissage automatique, recherche d'information textuelle et le traitement du langage naturel (NLP)

# Abstract

Today, if you want to interact with a computer, you have to learn a programming language, but instead we would like the machine to understand what we say or rather, what we write when we search for data, and this is the goal of our project "realization of a search engine with a semantic approach" which tries to understand the meaning of the users' queries in order to respond with more relevant results that reflect the user's intention.

In order to realize our application, we used the Word embedding technique which is widely known and used in textual search that exploits semantic similarity.

Our application is based on natural language processing (NLP) and information retrieval (IR)



**Keywords**:

Information retrieval (IR), semantic search engine, machine learning, textual information retrieval, and natural language processing (NLP)

# ملخص

اليوم، إذا كنت تريد التفاعل مع جهاز كمبيوتر، فعليك تعلم لغة برمجة، ولكن بدلاً من ذلك نود أن تفهم الآلة ما نقوله أو بالأحرى، ما نكتبه عندما نبحث عن البيانات، وهذا هو هدفنا مشروع "تحقيق محرك بحث بنهج دلالي" يحاول فهم معنى استفسارات المستخدمين من أجل الاستجابة لنتائج أكثر صلة تعكس نية المستخدم.

من أجل تحقيق تطبيقنا، استخدمنا تقنية تضمين الكلمات المعروفة والمستخدمة على نطاق واسع في البحث النصي الذي يستغل التشابه الدلالي.

يعتمد تطبيقنا على معالجة اللغة الطبيعية (NLP) واسترجاع المعلومات (IR)

**الكلمات الرئيسية**

استرجاع المعلومات، محرك البحث الدلالي، التعلم الآلي، استرجاع المعلومات النصية، معالجة اللغة الطبيعية، والبرمجة اللغوية العصبية

# Table des matières

# Liste des abréviations

| | |
|---|---|
| IR | Information retrieval |
| NLP | Natural Language Processing |
| SRI | Systèmes de Recherche d'Informations |
| MRI | Modèles de Recherche d'Informations |
| CSV | Comma Separated Values file |
| Glove | Global Vectors for Word Representation |

# Liste des figures

**Chapitre 02**



**Chapitre 03**



**Chapitre 04**

# Chapitre I : Introduction générale

# Chapitre I : Introduction Générale

## I.1. Introduction

Le domaine de recherche d'information remonte au début des années 1950, peu après l'invention des ordinateurs. Comme plusieurs autres domaines informatiques, les pionniers de l'époque étaient enthousiastes à utiliser l'ordinateur pour automatiser la recherche des informations, qui dépassaient la capacité humaine : il y avait une explosion d'information après la deuxième guerre mondiale. **[1]**

L'outil qui utilise souvent la recherche d'information est bien les moteurs de recherches qui permettant de retrouver plus facilement une donnée parmi un très grand nombre de données. Il existe différents moteurs de recherche dans le monde. Tel que : Google, Bing, Yahoo.

## I.2. Contexte

Avec l'augmentation rapide de données dans plusieurs domaines, il devient très difficile d'obtenir les données souhaites dans un temps conviviable et ce comme chercher une aiguille dans une botte de foin **[2]**. Pour remédier à cette cause, notre projet vise à développer un moteur de recherche sémantique qui a un rôle d'un cote, d'automatisé l'accès efficace à cette énorme quantité de données, et de l'autre cote ajouter l'aspect sémantique a la recherche qui renvoie les résultats corrects qui ont une similarité avec la requête de l'utilisateur même si elle était différente en termes d'orthographe.

Dans ce projet « Réalisation d'un moteur de recherche sémantique », nous intéresserons à étudier le domaine de la recherche d'information sémantique qui nous permet de concevoir puis réaliser un moteur de recherche sémantique fonctionne en deux modes en ligne et hors ligne en exploitant les techniques du machine Learning et de l'intelligence artificielle.

## I.4. Problématique

La machine de l'utilisateur contient un ensemble important de données et chacune d'elles nécessitent un processus de recherche d'informations, l'utilisateur a du mal à trouver des données similaires et appropriées à sa requête et passe beaucoup de temps à chercher, donc il a besoin de faire comprendre la machine ce qu'il recherche.

Pour résoudre ce problème, nous proposons dans ce projet un moteur de recherche qui intègre une nouvelle couche aux moteurs de recherche traditionnels, qui est la couche sémantique qui comprend l'intention de l'utilisateur et fournit des résultats plus pertinents qui répondent au mieux à ses requêtes, avec une interface simple et ergonomique, ce qui peut améliorer l'expérience utilisateur et facilitent son travail.

## I.3. Objectifs

Considérant que la plupart des utilisateurs ont des difficultés de rechercher des données dans une base de données, les moteurs de recherche facilitent la recherche de donnée parmis un très grand nombre de données.

Dans le cadre de ce travail, notre objectif consiste à réaliser une application bureau avec **python** dans le but de :

- Effectuer une étude sur la recherche d'information et les méthodes existantes de l'apprentissage automatique qui permettent l'intégration de la couche sémantique dans les moteurs de recherche ;
- Renvoie des résultats pertinents en se basant sur la compréhension sémantique du sens de la requête donnée par l'utilisateur ;
- Minimiser le temp de recherche ;
- Classer dans un ordre décroissant les résultats de plus proche au sens du requête de l'utilisateur jusqu'à la plus loin au sens ;
- Développer une application avec une interface ergonomique.

## I.5. Structure du rapport

Notre rapport est organisé comme suit :

Après une description globale du contexte, de la problématique, des objectifs de notre travail, nous avons focalisé et plus précisément dans le chapitre 2 sur les concepts de base de la recherche d'information, ensuite on a enchainé avec la méthode d'apprentissage machine pour l'intégration de la couche sémantique dans les moteurs de recherche. Au chapitre 3, on a présenté les moteurs de recherche existants et leur fonctionnement. Ensuite au chapitre 4, on a présenté le cycle de développement de notre solution avec les différents processus, les outils de développement utilisés pour la réalisation de notre projet et donner les processus d'une façon détaillée.

A la fin de ce mémoire, nous terminerons par une conclusion générale ainsi que quelques perspectives pour des futures améliorations éventuelles

# Chapitre II :

# Généralités sur la recherche d'information

# Chapitre II : Généralités sur la recherche d'information

## II.1. Introduction

L'immense avancement technologique que le monde voit actuellement a créé des grands défis dans l'informatique d'aujourd'hui en général et dans la recherche d'information en particulier. L'utilisation de la technologie de moteur de recherche sémantique a résolu une partie de ce problème, car elle renvoie des résultats corrects à l'utilisateur sur la base du domaine de recherche d'informations (IR) qui utilise la technologie de NLP.

Le traitement sémantique des requêtes des utilisateurs avec le domaine IR fonctionne par comprendre le sens contextuel de la requête, nous utilisons donc des techniques d'apprentissage automatique et d'intelligence artificielle. Actuellement, plusieurs études dans ce domaine sont menées à l'aide du traitement automatique du langage naturel (NLP). Au cours des dix dernières années, le traitement du langage naturel (NLP) est devenu un élément essentiel de nombreux systèmes de recherche d'information, principalement sous la forme de réponses aux questions, de résumés, de traductions automatiques et de prétraitements tels que le décomptage.

Ce chapitre a pour but de présenté le domaine de la recherche d'information (IR). Ensuite nous introduisons les concepts et les processus associes. Après nous enchainons par la présentation des modèles de recherche. A la fin de ce chapitre, nous allons défini la technique de l'apprentissage automatique NLP.

## II.2. Recherche d'information (Information Retrieval)

L'informatique a permis le développement d'outils pour traiter l'information et établir la représentation des documents au moment de leur indexation **[3],** ainsi que pour rechercher l'information. La recherche d'information est aujourd'hui un champ pluridisciplinaire, intéressant même les sciences cognitives.

### II.2.1. Définition

Plusieurs définitions de la recherche d'information est apparu, nous citons les définitions suivantes :

La recherche d'informations (RI ou Information Retrieval en anglais) est une branche de l'informatique qui porte sur la représentation, le stockage, l'organisation et l'accès à des informations telles que des documents, des pages Web, des images, des données structurés et semi-structurés. La représentation et l'organisation des éléments d'information devraient être telles qu'elles permettent aux utilisateurs d'accéder facilement aux informations qui les intéressent **[4]**

En d'autres termes, la recherche d'information (RI ou Information Retrieval en anglais) est le domaine qui consiste à trouver un objet dans une base de données disponible pour bien répondre à la requête d'un utilisateur.

### II.2.2. Le système de la recherche d'information (SRI)

Un système de recherche d'information (SRI) ou appelé aussi moteur de recherche est un ensemble de programmes informatiques qui se communiquent ensemble pour satisfaire les requêtes données par les utilisateurs.

Le rôle d'un système de recherche d'information SRI est de mettre en œuvre des techniques et des moyens permettant de retourner les documents pertinents d'une collection en réponse à un besoin en information d'un utilisateur, exprimée par un langage de requête qui peut être le langage naturel.

### II.2.3. Concepts et processus de RI

#### II.2.3.1. Processus de RI

Le processus U ou modèle en U de la recherche d'information est une approche pratique pour la construction des moteurs de recherche robustes, qui a pour but d'améliorer la pertinence des résultats ainsi que les performances des systèmes de recherches. Il est décomposé en trois principales étapes : l'indexation, l'appariement document-requête et la reformulation de la requête. Le schéma ci-dessous (figure 1) représente mieux le processus de la RI en utilisant le modèle U :

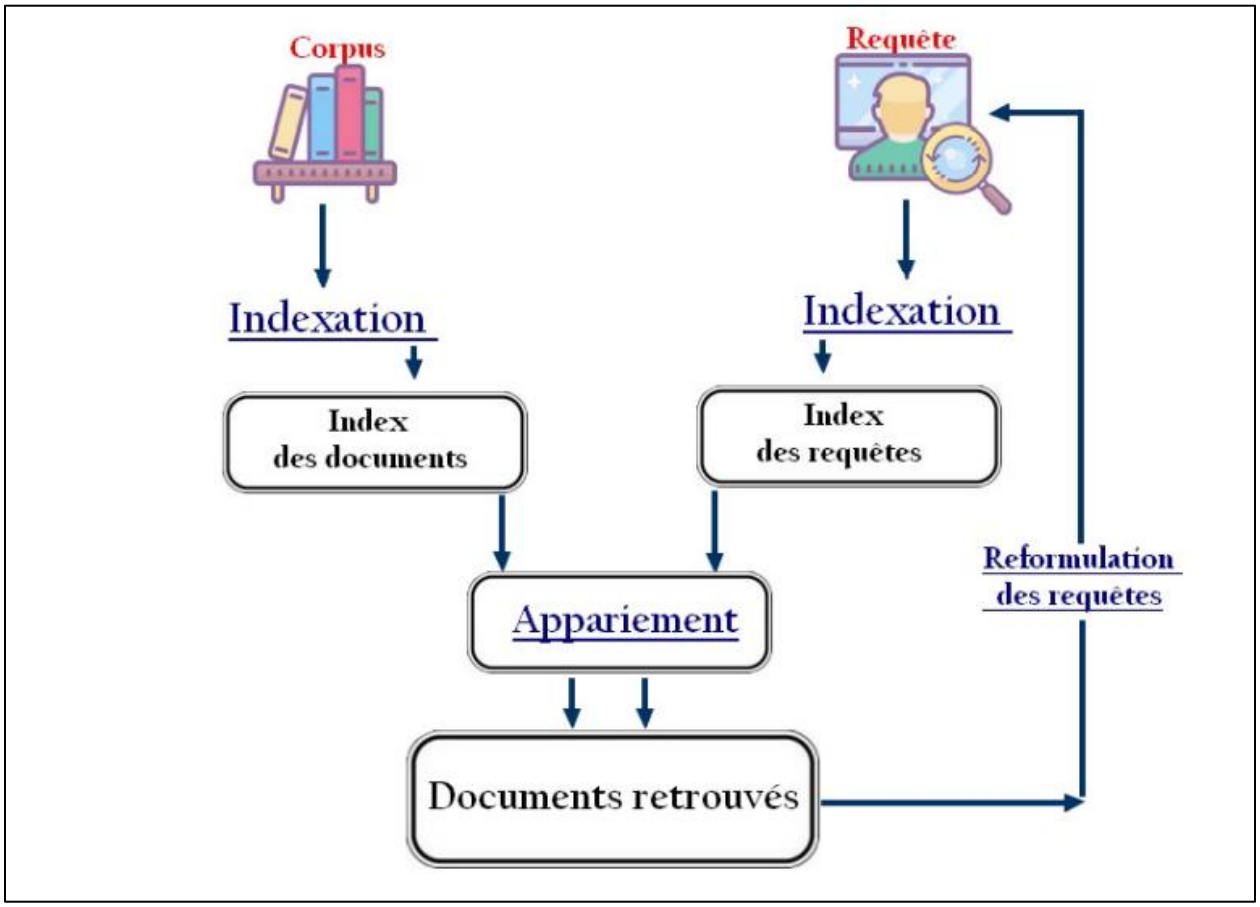


**Figure 1 :** Architecture générale d'un Système de recherche d'information **[5]**.

Nous détaillons les principales étapes dans un processus de recherche d'informations dans ce qui suit.

1. L'indexation consiste à extraire et à représenter le contenu des documents de manière interne sous forme d'index. Cette structure d'index permet de retrouver rapidement les documents contenant les mots clés de la requête.

2. Appariement document/requête, une fois les documents sont représentés sous forme interne d'index. Suite à une requête utilisateur, le système calcule la pertinence de chaque document vis-à-vis de la requête utilisateur selon une mesure de correspondance du modèle de RI, et retourne la liste des résultats à l'utilisateur.

3. 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.

II.2.3.1.1. Indexation

Cette étape consiste à identifier pour chaque document les termes importants, puis à exploiter ces termes comme index pour accéder rapidement aux documents. Un des objectifs de l'indexation est donc de permettre de retrouver rapidement les documents contenant les termes (mots-clés) de la requête. L'indexation peut être : manuelle, semi-automatique ou automatique.

Le processus d'indexation des documents passe par des étapes qui permettent d'avoir les descripteurs de l'ensemble des documents. Ces étapes sont comme suit

***a)*** **Extraction des mots**

Cette phase consiste à extraire/segmenter le texte du document en mots. La segmentation (Tokenization) du texte est une première étape importante dans ce processus qui permet de reconnaître les espaces de séparation des mots, les chiffres, les ponctuations, etc.

***b)*** **Élimination des mots vides (Stop words)**

Les textes contiennent souvent des termes non significatifs appelés mots vides (pronoms personnels, prépositions, etc.). Ce traitement a pour but d'enlever les mots grammaticaux, ainsi que rejeter les mots dépassant un certain nombre d'occurrences dans la collection. La suppression de

ces termes peut réduire de manière considérable la taille des indexes. Selon le modèle de recherche d'information utilisé, la suppression de ces mots n'a généralement pas d'impact sur l'efficacité du moteur de recherche, et peut même l'améliorer.

***c)*** **Normalisation**

Cette phase est liée à la lemmatisation (ou racinisation), il s'agit d'un traitement morphologique des mots permettant de regrouper les variantes d'un mot. En effet, dans un texte, il peut y avoir différentes formes d'un mot désignant le même sens. Le but de ce processus est de les représentés par un seul mot qui porte un concept commun (ex. biologie, biologiste, biologique ? par : biologie). Grâce à la lemmatisation, les documents contenant différentes formes d'un même mot auront les mêmes chances d'être restitués. Par conséquent, elle réduit la taille de l'index et améliore le rappel

***d)*** **Pondération des mots**

Cette étape vient après l'identification des termes des documents et leur normalisation. Les termes qui représentent un document n'ont pas la même importance. Donc, la pondération est une phase primordiale puisqu'elle traduit l'importance des termes en indices qui reflètent le poids relatif des mots dans les documents.

II.2.3.1.2. Appariement document-requête

Une fois les documents indexés et la requête analysée, le système de RI procède à la mesure de pertinence de chaque document vis-à-vis du besoin d'information (requête) selon une fonction de correspondance relative au modèle de recherche, et à renvoyer ensuite à l'utilisateur une liste de résultats qui obtiennent une valeur de correspondance élevée Cette mise en correspondance génère un score de pertinence reflétant le degré de similarité entre la requête et le document.

### II.2.4. Modèles de base de la recherche d'information MRI

Les systèmes de recherche d'information se basent sur des modèles théoriques différents, qui déterminent comment les tâches d'indexation et de correspondance sont réalisées. Il existe un grand nombre de modèles théoriques de recherche d'information dans la littérature. Ils sont généralement classés dans trois familles notamment :

- Modèle basé sur la théorie des ensembles comme le modèle booléen ;
- Modèle algébrique comme le modèle vectoriel ;
- Modèle probabiliste comme le modèle probabiliste classique.

Les autres modèles existants sont principalement des formes mixtes des trois types. La figure suivante représente les modèles de recherche d'information :

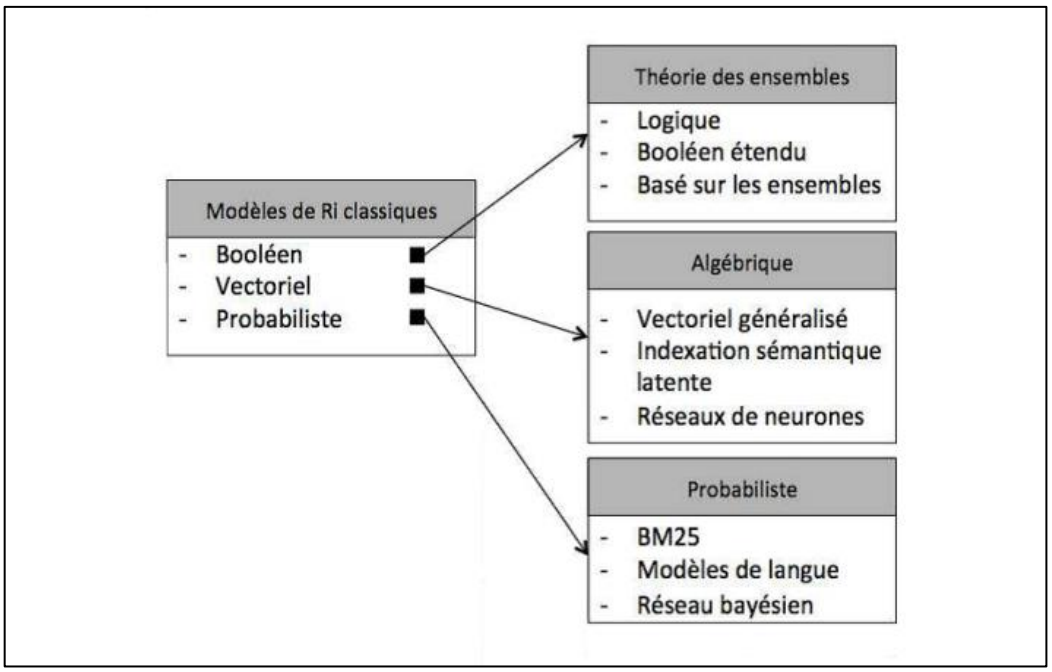


**Figure 2 :** Modèles de recherche d'information **[6]**.

Nous représentons par la suite en détail ces **trois** fameux modèles :

### II.2.4.1. Les modèles ensemblistes

Ce sont les modèles basés sur la théorie des ensembles et l'algèbre de Bool, dans le représentant le plus connu est le modèle booléen.

Le modèle booléen

Le modèle booléen est le premier modèle qui a émergé dans le monde de la RI. Il est basé sur la manipulation des ensembles et l'algèbre booléenne.

Dans ce modèle, une requête est représentée par une expression logique composée de termes séparés par des opérateurs logiques (AND, OR et NOT). Les fréquences des termes dans l'index, c'est-à-dire la matrice terme-document, sont toutes binaires, c'est-à-dire que les termes sont présents ou absents dans le document (wij{0,1}).

Le modèle booléen utilise la méthode de correspondance exacte, c'est-à-dire qu'il ne prend en compte que les documents par rapport à la requête décrite. La similarité entre un document et une requête est définie par :

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

La fonction de similarité du modèle est booléenne. Par conséquent, il n'y aurait pas de correspondance partielle, pas de classement pour les documents récupérés et cela peut être gênant pour les utilisateurs, ce qui empêche le modèle d'avoir de bonnes performances.

### II.2.4.2. Les modèles algébriques

Ce sont les modèles basés sur la théorie d'algèbre, comme le modèle vectoriel

Modèle vectoriel

Ce modèle est venu à cause des inconvénients du modèle booléen, en plus est plus précis et plus proche du langage naturel. Il se base sur des calculs mathématiques (vectoriels) pour extraire les documents les plus pertinents. Son principe consiste à représenter les documents et les requêtes dans un espace vectoriel engendré par tous les termes de la collection de document, puis calculer la similitude et retourner une liste ordonnée des résultats par pertinence. la figure ci-dessus illustre bien le model vectoriel :

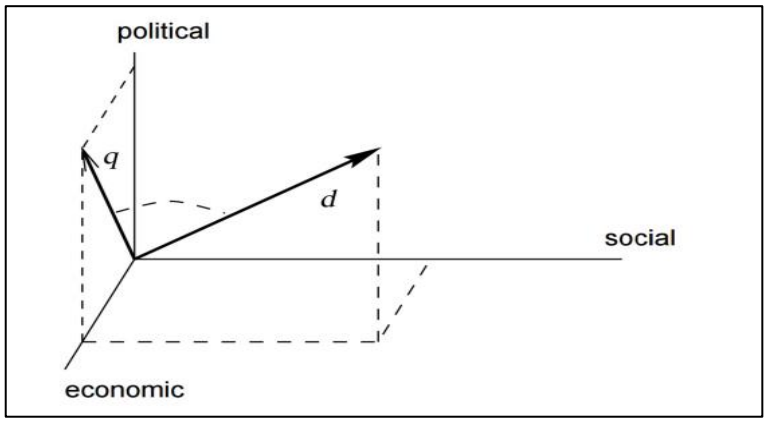


**Figure 3** : Représentation de requête et document dans l'espace des termes à 3D.

La pertinence est traduite en une similarité vectorielle : un document est plus pertinent à une requête, si le vecteur associé est similaire à celui de la requête. Cette similarité est calculée à l'aide de plusieurs fonctions, la plus utilisé est celle de la similarité cosinus (cosine similarité en anglais) qui consiste à calculer le cosinus entre les deux vecteurs des documents et les requêtes. Plus l'angle est petit, plus la similarité est grande **(F. Fkih, 2016).** L'équation suivante illustre bien la similarité cosinus :

**Mesure de cosinus :**

$$Sim(D_j, Q_k) = \frac{\sum_{i=1}^{N} (wd_{ij} * wq_{ik})}{\sqrt{\sum_{i=1}^{N} wd_{ij}^2 * \sum_{i=1}^{N} wq_{ik}^2}}$$

**Sim** : similarité

**D** : le vecteur de document

**Q** : le vecteur de requête

**Sim (D, Q ) = 1**, lorsque d et q n'a aucune similarité ;

**Sim (D, Q ) = 0**, lorsque d = q.

Plus la valeur est proche de zéro, plus les mots sont similaires

Les avantages de ce modèle d'espace vectoriel sont nombreux : il permet la pondération des termes, ce qui augmente les performances du système et permet de renvoyer les documents qui correspondent approximativement à la requête et de trier efficacement les documents correspondant

à cette requête. Les documents peuvent être retournés dans l'ordre décroissant de leur degré de similarité avec la requête. Plus le degré de similarité d'un document est élevé, plus le document correspond à la requête.

Théoriquement, le modèle vectoriel présente le principal inconvénient de l'indépendance mutuelle des termes d'indexation. Aujourd'hui, le modèle vectoriel est le plus populaire dans la recherche d'information, malgré sa simplicité ; il donne de bons résultats par rapport les autres modèles

#### II.2.4.3. Les modèles probabilistes

Le principe de base de ce modèle consiste à présenter les résultats d'un SRI dans un ordre basé sur la probabilité de pertinence d'un document vis-à-vis d'une requête. L'idée de base de cette fonction est de sélectionner les documents ayant à la fois une forte probabilité d'être pertinents et une faible probabilité d'être non pertinents à la requête

La fonction de classement de ce modèle est exprimée ainsi :

$$\mathrm{RSV}(q, d) = \frac{\mathrm{P}(\mathrm{Perlq}, \mathrm{di})}{\mathrm{P}(\mathrm{NPerlq}, \mathrm{d})}$$

## II.3. Traitement de langage naturel (NLP)

Pour bien améliorer les résultats de la recherche d'information, la technique nlp a été développée pour rajouter la couche sémantique au RI. Comme nous le savons, deux phrases peuvent avoir des structures très différentes avec des mots différents, mais elles peuvent avoir le même sens, C'est pour cette raison la technique NLP conçus pour le but de capturer le sens des phrases et renvoie les résultats similaires en sens par des méthodes d'apprentissage automatique.

### II.3.1. Définition NLP

Le NLP est l'acronyme de Natural Language Processing, ce qui signifie Traitement Naturel du Langage ou Traitement Automatisé du Langage Naturel. Cette technologie est une forme d'intelligence artificielle et une branche de la Data Science. Elle permet aux ordinateurs d'analyser, de comprendre le langage humain et de générer des interactions, en transformant de la data brute en conversation intelligente. Ce système permet aux humains et aux machines de parler le même langage. Concrètement, grâce au NLP, les entreprises peuvent analyser automatiquement des phrases issues d'un humain, pour prendre les décisions les plus adaptées.**[7]**

Le NLP regroupe trois domaines principaux : l'informatique, le langage humain et l'intelligence artificielle. Dans ce dernier, le NLP combine des approches linguistiques, des techniques d'apprentissage automatique (ML ou Machine Learning) et d'apprentissage profond (DL ou Deep Learning). Comme montre la figure en bas :

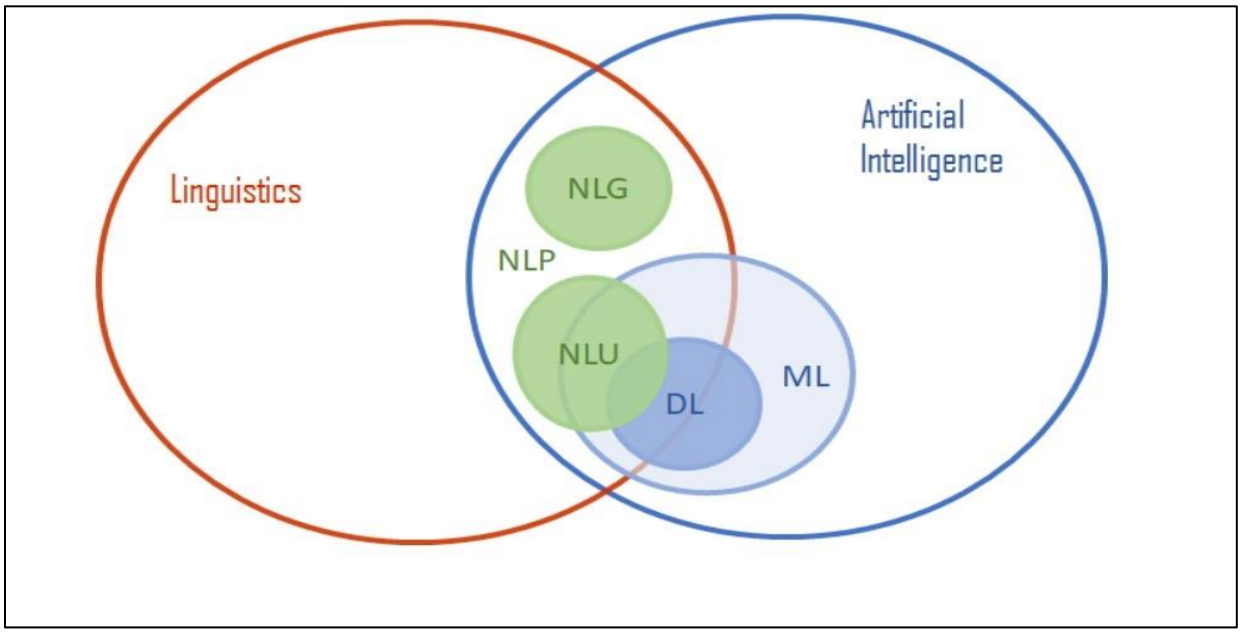


**Figure 4** : Le Positionnement du NLP dans l'écosystème de l'intelligence artificielle **[8]**.

Le traitement du langage naturel (NLP) est divisé en deux catégories clés :

1. La compréhension du langage naturel (NLU ou Natural Language Understanding)
2. La génération du langage naturel (NLG ou Natural Language Generation).

Il peut y avoir plusieurs façons d'effectuer une recherche d'information. Le NLP améliore les résultats obtenus en servant de plusieurs techniques, Mais nous avons choisir à utiliser la technique de Word Embedding.

### II.3.2. Word Embedding

Une méthode de représentation des termes et des phrases dans l'espace des vecteurs multidimensionnels. Cette représentation permet de prendre en considération la sémantique des mots en se basant sur leurs contextes dans la phrase**.** Ce qui permet par la suite de déduire les relations entre les mots et calculer la similarité sémantique entre les phrases. Le développement de cette technique a ouvert la porte pour autres travaux de recherches plus avancés dans le domaine du traitement sémantique. (Voir la figure 5)

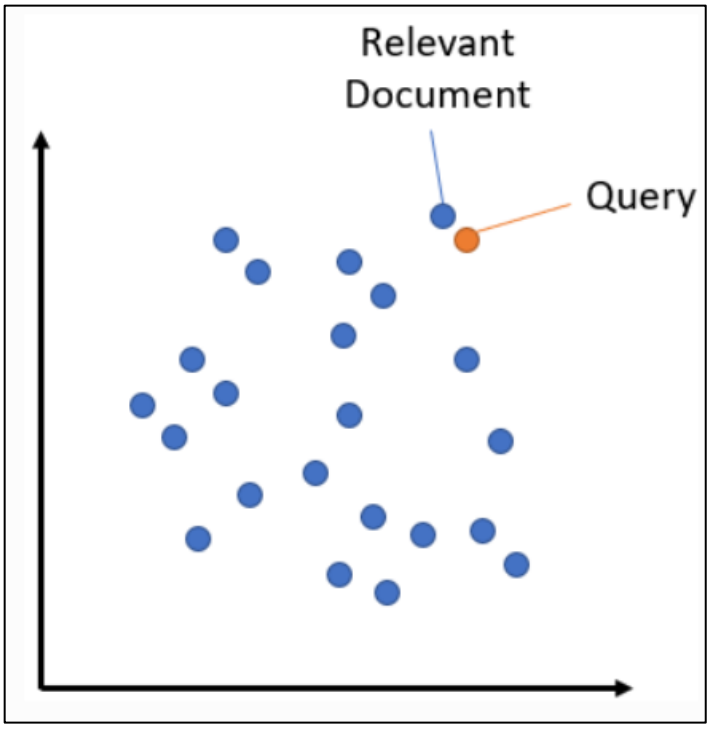


**Figure 5 :** Example vectoriel utilise la technique « Word Embedding » **[9]**.

La caractéristique la plus importante des encastrements de Word Embedding est que les mots similaires dans un sens sémantique ont une distance (euclidienne, cosinus ou autre) plus petite entre eux que les mots qui n'ont pas de relation sémantique. Par exemple, des mots comme "Queen " et "King " devraient être plus proches que les mots "Queen" et "ketchup" ou "King" et "beurre".

L'un des exemples fréquemment donnés est l'équation " king-men + women = queen ". Ce qui se passe ici, c'est que la valeur vectorielle obtenue à la suite de la soustraction et de l'addition des vecteurs les uns aux autres est égale au vecteur correspondant à l'expression "reine". On peut comprendre que les mots "roi" et "reine" sont très semblables l'un à l'autre, mais que les différences vectorielles ne se produisent qu'en raison de leur sexe. Comme montre la figure ci-dessus :

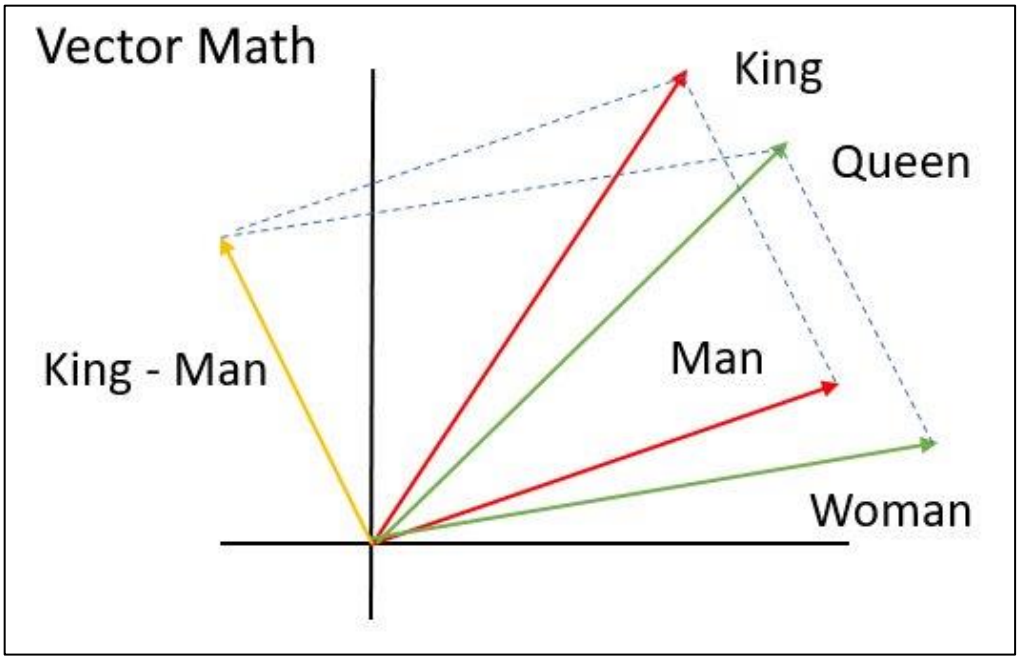


**Figure 6**: Exemple du Word Embedding **[10]**.

Certaines des techniques du Word Embedding les plus prometteuses dans le domaine du traitement automatique des langues (Word2Vec, Glove, Fast Text) pour estimer la similarité sémantique des phrases. Mais la technique la plus connu et la plus utilisé est la technique Glove, qui représente un algorithme d'apprentissage non supervisé permettant d'obtenir des représentations vectorielles des mots. L'apprentissage est effectué sur des statistiques globales agrégées de cooccurrence mot - mot d'un corpus, et les représentations résultantes présentent des sous-structures linéaires intéressantes de l'espace vectoriel des mots.

## II.4. Conclusion

A travers ce chapitre, nous avons pu se familiariser avec la notion de la recherche d'information et construire une connaissance approfondie dans ce domaine. L'étude des modèles de la RI, leurs processus et leurs architectures de base. Ainsi nous pouvons conclure que l'intégration de la sémantique dans les moteurs de recherche améliore les résultats obtenus efficacement et d'une manière très importante en utilisant le NLP. L'avancement actuel du domaine NLP ainsi que les approches sémantiques du traitement du texte ont rendu cet axe de recherche plus facile et plus exploitable qu'avant. Dans le chapitre suivant, nous allons présenter les moteurs de recherche existants ainsi leur fonctionnement.

# Chapitre III : les Moteurs de recherche

# Chapitre III : Les moteurs de recherche

## III.1. Introduction

Il existe de nombreux moteur de recherche qui devenus familiers et parfois indispensables, ainsi, leur usage s'est banalisé dans des situations très diverses de la vie quotidienne, que ce soit en contexte professionnel ou privé.

Dans ce chapitre, nous commençons d'abord par une définition simple des moteurs de recherche en focalisant sur les moteurs de recherche sémantique. Ainsi, nous présentons une liste non exhaustive de moteur de recherche offline et en ligne utiles pour l'objectif de répondre aux besoins des utilisateurs. Ensuite, le fonctionnement d'un moteur de recherche en ligne.

## III.2. Définition du moteur de recherche

Un moteur de recherche représente un outil qui permet d'obtenir des informations utiles grâce à l'utilisation des mots clés et des termes de recherches. Il faut juste renseigner ces derniers en formant une requête et le moteur de recherche se chargera de faire ressortir de manière automatique les informations en rapport avec celle-ci. Ces informations sont proposées dans un ordre défini en fonction de chaque moteur de recherche. En diffère deux types de moteur de recherche à savoir moteur de recherche traditionnel et moteur de recherche sémantique qui est né du moteur de recherche traditionnel. **[11]**

### III.2.1. Moteur de recherche sémantique

Dans le langage, la définition de la sémantique est liée au sens linguistique. Ainsi, en termes de recherche, un moteur de recherche sémantique une application permettant à un utilisateur d'effectuer une recherche **offline** ou **en ligne** qui effectue une étude de sens en se concentrant sur la signification des termes de recherche saisis **[12]**.

Essentiellement, la recherche sémantique fonctionne en établissant des liens entre des mots et des phrases, il est capable d'interpréter le contenu numérique d'une manière plus "humaine". Lorsque cela est réalisé, il peut offrir au chercheur des résultats de recherche plus personnalisés

et plus précis. Les technologies de la recherche sémantique jouent un rôle crucial dans l'amélioration de la recherche traditionnelle, car elles contribuent à créer des données lisibles par machine

Un moteur de recherche est constitué de deux éléments principaux :

- Une base de données, qui contient un ensemble d'informations sur les contenus.
- Un algorithme, chargé de classer les résultats de manière pertinente selon la requête utilisateur

## III.3. Modèles existants

Comme nous avons cites précédemment, Il existe deux modèles de moteur de recherche diffèrent essentiellement en termes de connectivite avec l'Internet :

### III.3.1. Les moteurs de recherche offline

C'est un moteur de recherche qui ne nécessite pas la connexion Internet, Donc les résultats retenus existe localement sur une base de donnes locale ou bien sur de divers types de documents sur la machine. Le fameux moteur de recherche offline est l'application « Google Desktop ».

Google Desktop

Google Desktop est un moteur de recherche offline de bureau qui facilite la recherche de documents sur votre ordinateur. Cet outil n'étant pas limité aux noms de fichiers, il est en mesure d'effectuer des recherches dans le contenu de divers types de documents, Word, PDF, PowerPoint, Excel, Images, Vidéos, musiques ainsi que vos e-mails et leurs pièces jointes, Google Desktop fonctionne sur les systèmes d'exploitation Windows. **[13],** La figure ci-dessus représente l'interface de google desktop:

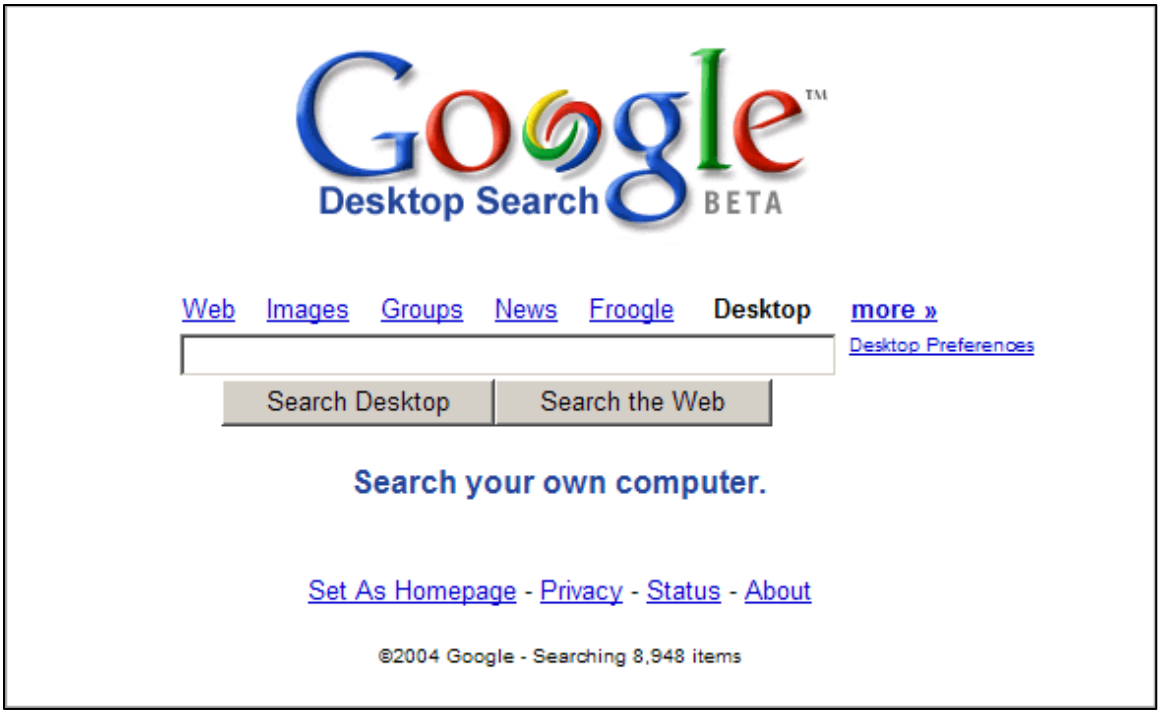


**Figure 7** : Moteur de recherche offline Google Desktop **[14]**.

### III.3.2. Les moteurs de recherche en ligne

C'est un service en ligne permettant de trouver facilement des résultats souhaites grâce à un ou plusieurs mots-clés renseignés dans un formulaire de recherche. Ce type de moteur est le plus utilises dans le monde, par exemple Google, Yahoo et Bing.

#### III.3.2.1. Le fonctionnement d'un moteur de recherche en ligne

Un moteur de recherche fonctionne à l'aide de robots (appelés aussi "spiders" ou "crawlers") chargés de parcourir tout le contenu présent sur internet et de le stocker dans d'immenses bases de données. Le contenu est ensuite analysé puis trié avant d'être mis à disposition des utilisateurs. Il est impossible de parcourir en une journée tous les fichiers stockés sur internet. Les moteurs de recherche ont donc chacun leur propre fréquence et leur propre manière de mettre à jour leurs données.

Ils reposent sur un algorithme complexe qui va identifier la nature des documents mis en ligne (musiques, photos, vidéos, textes, logiciels etc.). L'objectif d'un moteur de recherche est de fournir le plus rapidement possible une réponse pertinente par rapports aux recherches de ses utilisateurs.

Pour cela, il doit non seulement interpréter les termes recherchés par ses utilisateurs mais également afficher instantanément ses meilleurs résultats. Si les résultats affichés ne satisfont pas ses utilisateurs, il est probable qu'ils arrêteront d'utiliser ce moteur de recherche.**[15]**

Google search

Google est un moteur de recherche gratuit et libre d'accès sur le World Wide Web, ayant donné son nom à la société Google. C'est aujourd'hui le moteur de recherche et le site web le plus visité au monde : Plus 90 % des internautes l'utilisaient. Son utilisation est assez simple : en tapant une requête dans la barre de recherche, l'outil dresse un classement des sites web les plus pertinents pour apporter une réponse la plus précise possible à l'utilisateur avec un lien pour y accéder. En plus de la liste de liens vers des sites internet, les réponses du moteur de recherche peuvent être présentées sous différents formats. **[16],** La figure suivante (voir Figure 8) représente l'interface de google search :

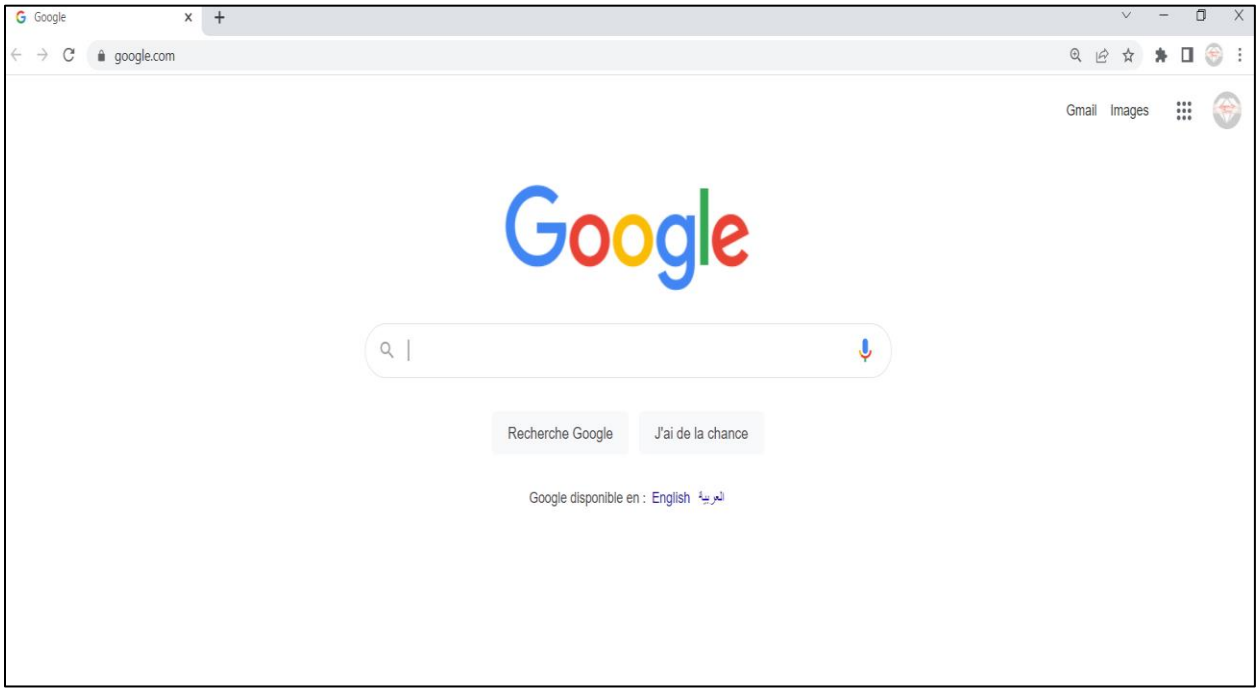


**Figure 8** : Moteur de recherche en ligne Google **[17]**.

Microsoft Bing

Microsoft Bing est le moteur de recherche développé par Microsoft. Son objectif principal est de fournir aux utilisateurs les pages web les plus pertinentes selon les requêtes entrées. Très similaire à Google, Microsoft Bing est une bonne alternative au moteur de recherche le plus populaire.

Lorsque vous entrez des mots clés sur Bing, le moteur de recherche analyse automatiquement toutes les pages web associées à ces mots clés, pour créer un index des URL, et afficher un ensemble de résultats de recherche pertinents, divisés en plusieurs pages. Les classements de Bing prennent en compte beaucoup de critères : mots sur la page, titre de la page, texte d'ancrage des pages renvoyant vers une autre page, emplacement de l'utilisateur (pays, ville, etc.), langue, interactions précédentes. Le moteur de recherche est totalement gratuit, et accessible depuis tous les navigateurs web. **[16],** La figure suivant illustre l'interface de Microsoft bing :

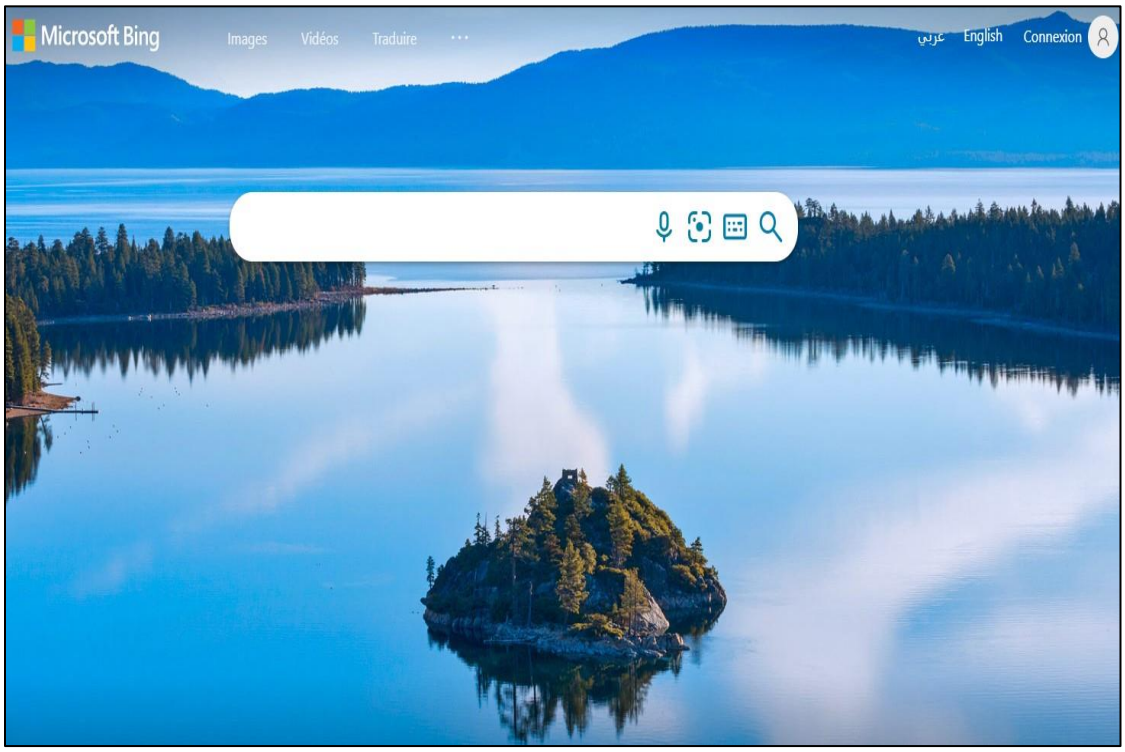


**Figure 9** : Moteur de recherche en ligne Bing **[18]**.

Yahoo search

Yahoo Search, plus connu sous le nom de Yahoo!, est l'un des plus anciens moteurs de recherche. Il est aujourd'hui le troisième moteur de recherche le plus utilisé dans le monde. Il base son fonctionnement sur un aspect communautaire en donnant aux internautes la possibilité de sélectionner et d'annoter les pages et sites web qui leur paraissent intéressants pour les partager avec d'autres. Le moteur de recherche de Yahoo! est totalement gratuit. Vous pouvez l'utiliser via n'importe quel navigateur web ou via application mobile **[16]**. La figure ci-dessus représente l'interface de yahoo search :

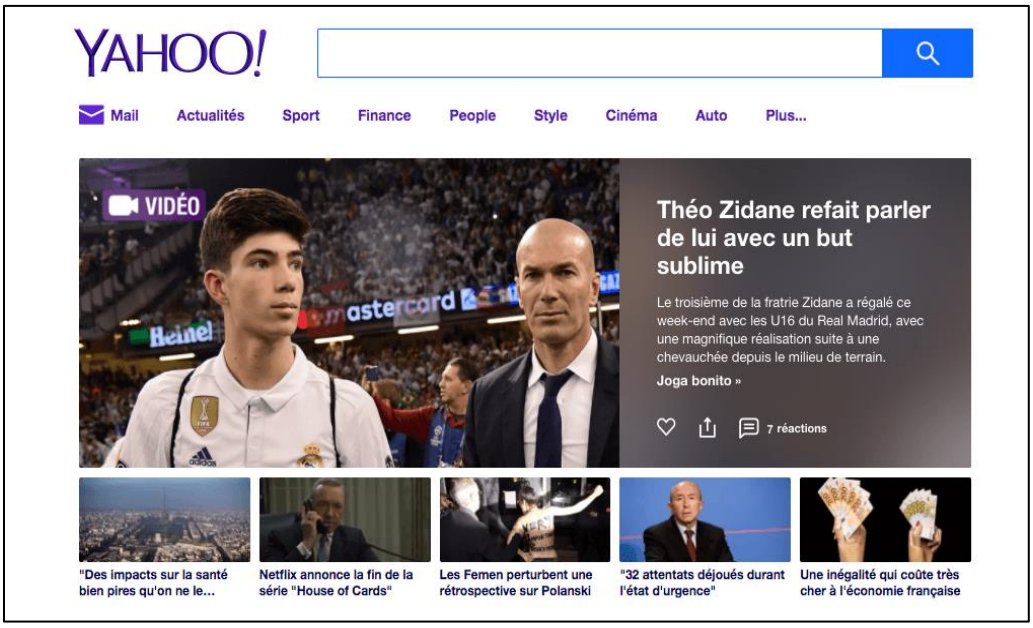


**Figure 10** : Moteur de recherche en ligne Bing **[19]**.

## III.4. Conclusion

Le moteur de recherche est un outil informatique permettant de réaliser des recherches précises à partir de requêtes, aussi appelées mots clés. Dans ce chapitre, nous avons définisse les moteurs de recherche d'une manière general, ensuite nous avons présenté quelques modèles existants de moteurs de recherche actuels, leurs fonctions et leurs objectifs.

Ce chapitre vise à présenter les moteurs de recherche et focalise sur les moteurs de recherche sémantique qui font l'actualité à nos jours. Et en terminant ce chapitre par la domination de GOOGLE, YAHOO et BING du marché. Ensuite au chapitre 4, nous présenterons donner une description sommaire de notre projet avec une brief présentation des outils de développement utilisés.

# Chapitre IV :

# Développement et Implémentation de l'application

# Chapitre IV : Implémentation et développement de l'application

## IV.1. Introduction

Après avoir exposé les différents axes nécessaires à la réalisation du moteur de recherche et quelques modèles de ce dernier. Nous proposerons notre moteur de recherche sémantique qui fonctionne en mode en ligne ainsi qu'en mode offline, et nous répondrons à cette question :

« Comment le moteur recherche-t-il, sélectionne-t-il et affiche-t-il les résultats d'une requête »

On commence par un résumé de notre proposition en décrivant le processus de recherche depuis introduire du requête utilisateur jusqu'au l'affichage des résultats, et pour mener à bien notre projet informatique, il est nécessaire de choisir des technologies permettant de simplifier sa réalisation. Pour cela, Dans ce chapitre nous présentons la description des environnements matériels et logiciels qui nous ont permis de réaliser le projet, des technologies et des langages de programmation que nous avons utilisée. Ensuite nous expliquons le fonctionnement de notre application bureau nome 'Infinity search' en présentant leur interface qui permettent l'interaction entre l'utilisateur et l'application.

## IV.2. Description sommaire du moteur proposé

Après avoir étudié les modèles existants dans ce qui précède, on a essayé de construire une description sommaire de notre application « Application bureau » dédiée aux utilisateurs pour reprendre à leur requête en fournissant une option supplémentaire qui est la possibilité de recherche offline et en ligne selon le choix de l'utilisateur.

Notre application est basée sur la recherche de groupes de données pour une donnée spécifique en comprenant le sens du contenu de la requête, même si les deux entités (requête et document) ont des structures différentes ou ont des compositions des mots différents d'orthographe, principalement notre modeste application fournit des avantages tel que :

- Faciliter la recherche de l'utilisateur offline ou en ligne et obtenir les résultats les plus proches qui ont la même similarité de sens avec leur requête ;
- L'application offre la fonctionnalité de recherche offline, a travers une base de données locale ;
- Interface simple et ergonomique ;
- Réduire le temps de recherche.

## IV.3. Présentation générale de notre moteur de recherche

Comme tout moteur de recherche, notre processus de recherche offline est composé de quatre étapes :

- Étape d'acquisition du requête utilisateur : L'utilisateur entre son mot-clé ;
- Étape de recherche des données correspondent à la requête : notre application « Infinity search » extrait les données qu'il estime répondre le mieux au mot-clé saisi par l'utilisateur lors de la requête ;
- Étape de classement des résultats obtenus : calcule la distance cosinus et classer dans un ordre décroissant les résultats de plus proche au sens du requête de l'utilisateur jusqu'à la plus loin au sens ;
- Étape d'affichage des résultats : les résultats affichés sont des données similaires en sens de la requête.

Le schéma suivant illustre le processus de recherche offline pour la recherche de donnée :

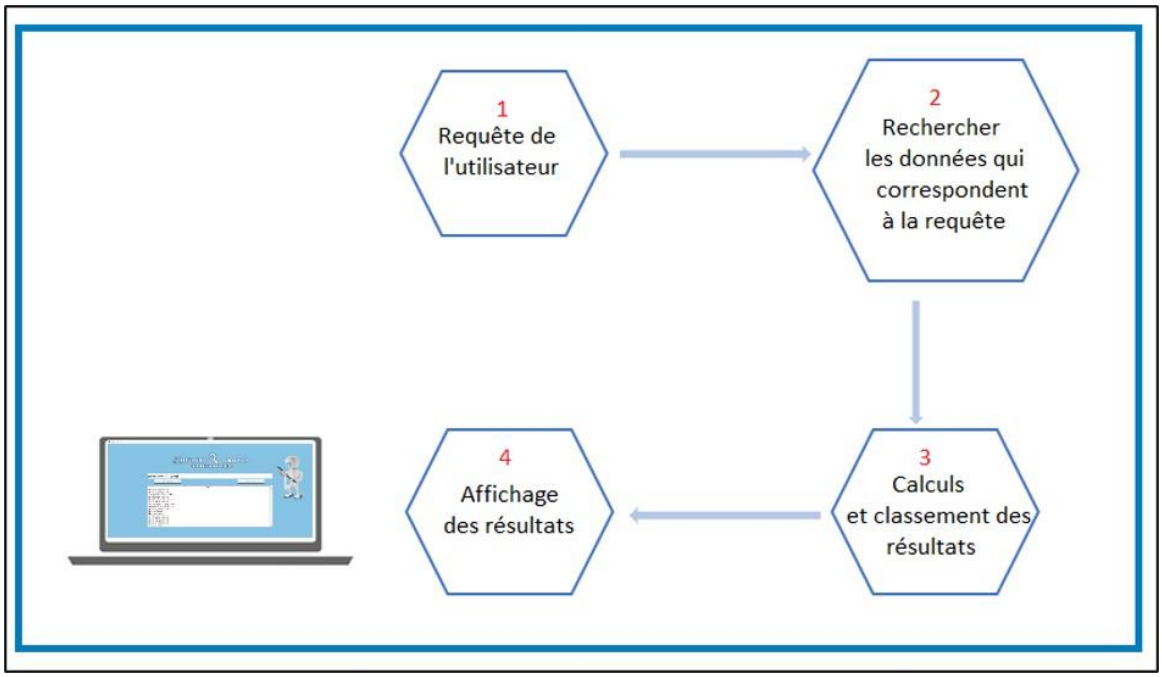


**Figure 11** : Les étapes de processus de recherche offline de l'application

Le processus de recherche en ligne, est composé de trois étapes :

- Étape d'acquisition du requête utilisateur : L'utilisateur entre son mot-clé ;
- Étape d'obtention des liens vers les sites à partir du Google,
- Étape d'affichage des résultats : les résultats obtenus sont des liens vers des sites correspondant à la requête établie

Le schéma suivant illustre le processus de recherche en ligne pour la recherche de donnée :

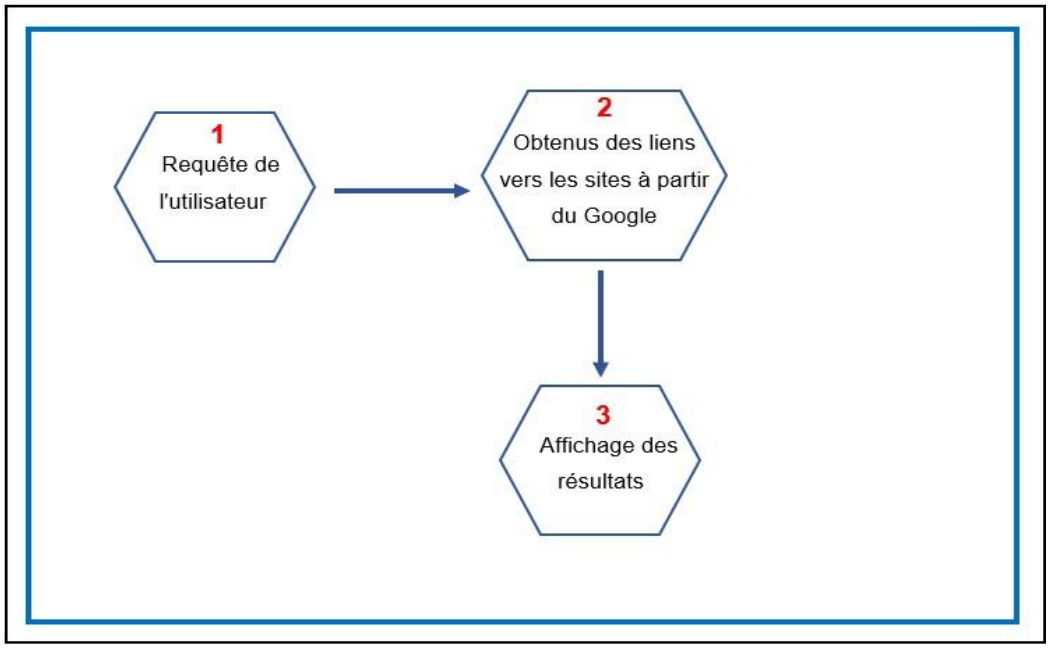


**Figure 12 :** Les étapes de processus de la recherche en ligne de l'application

## IV.4. Environnement de développement

C'est l'ensemble d'outils que nous avons utilisés pour développer notre application. Le choix des outils était base sur le plus utilises et le plus répondus dans le monde, et on distingue environnement matériel et environnement logiciel.

### IV.4.1. Environnement matériel

Pour la réalisation de notre projet, nous avons utilisé un ordinateur DELL caractérisé par :

- Système d'exploitation : Windows 10 21H2.
- Processeur : Intel(R) Core (TM) i7-6820HQ CPU @ 2.70GHz 2.70 GHz Mémoire vive : 32Go.
- Disque Dur : 256Go SSD.

**Figure 13** : PC utilise pour réaliser le travail **[20]**.

### IV.4.2. Environnement logiciel

Dans ce projet, on a utilisé beaucoup de logiciel pour bien abouti a un moteur de recherche répond à nos besoins à savoir le langage de programmation Python pour le développement de l'application, la plateforme Kaggle pour la base de données et d'autre logiciel que nous avons détaillés en ce qui suit.

#### IV.4.2.1. Langage Python

Python est le langage de programmation le plus utilisé dans le domaine du Machine Learning, du Big Data et de Traitement Naturel du Langage, Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels.

En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages **[21]**. L'éditeur qu'on a utilisé pour le Python est **Pycharm**. Dans notre application, nous avons utilisé python pour réaliser les tâches suivantes :

- Processus d'acquisition de données : ce processus est développé entièrement sous le langage python pour la réception de données « recherche d'information » ;
- L'entrainement et la construction du modèle : nous avons utilisé aussi Python pour la construction du modèle « **word embedding** » pour la recherche sémantique.

Pycharm 

PyCharm est un environnement de développement intégré (IDE) utilisé pour programmer en Python. Il permet l'analyse de code et contient un débogueur graphique. Il permet également la gestion des tests unitaires, l'intégration de logiciel de gestion de versions, et supporte le développement web avec Django. Développé par l'entreprise tchèque JetBrains, c'est un logiciel multi-plateforme qui fonctionne sous Windows, Mac OS X et GNU/Linux. Il est décliné en édition professionnelle, diffusé sous licence propriétaire, et en édition communautaire diffusé sous licence Apache.**[22]**

La plateforme Kaggle 

Kaggle est une plateforme web qui accueille la plus grande communauté de Data Science au monde, avec plus de 536 000 membres actifs dans 194 pays et reçoit près de 150 000 soumissions par mois et qui lui fournit des outils et des ressources puissants pour aider à atteindre tous les progrès de science des données, il contient une variété de datasets sur le site. Il existe une variété Jupyter Notebooks personnalisable et sans configuration, vous trouverez tout le code et les données dont vous avez besoin pour réaliser vos projets de science des données. Il y a plus de 50 000 jeux de données publics et 400 000 notebooks publics disponibles pour tous. **[23]**

Et Comme une alternative de la base de données logiciel. Nous avons choisi d'utiliser un fichier csv prêt et de le télécharger depuis le site Kaggle.

Un fichier CSV (en anglais, comma separated values) est le fichier de base des données recueillies - sans formatage particulier. Chaque champ est séparé par une virgule. Les fichiers CSV servent de format universel permettant de voir vos données dans une variété d'applications, comme Microsoft Excel, Numbers, le tableur Google ou autres.**[24]**

### IV.4.2.2. Bibliothèques logicielles

Un ensemble de fonctions utilitaires, regroupées et mises à disposition afin de pouvoir être utilisées sans avoir à les réécrire. Les fonctions sont regroupées de par leur appartenance à un même domaine conceptuel (mathématique, graphique, tris, etc.) **[25]**. L'intérêt des bibliothèques

réside dans le fait qu'elles contiennent du code utile que l'on ne désire pas avoir à réécrire à chaque fois. Dans notre projet on a utilisé plusieurs bibliothèques, qui sont :

Pandas 

Pandas est l'une des bibliothèques de science des données les plus populaires. Elle a été développée par des Data Scientists habitués au Python, elle offre de nombreuses fonctionnalités natives très utiles. Il est notamment possible de lire des données en provenance de nombreuses sources **[26]**.

NumPy 

NumPy est une bibliothèque mathématique du langage Python pour effectuer des calculs sur les matrices et les tableaux multidimensionnels. Nous avons utilisé NumPy dans notre projet pour effectuer des opérations sur les vecteurs des mots dans notre modèle sémantique (à partir de numpy.org) **[27]**.

Nltk 


Nltk est une bibliothèque Python dédiée au traitement naturel du langage ou Natural Language Processing. Nltk est une plateforme de pointe pour la création de programmes Python destinés à travailler avec des données sur le langage humain. Elle fournit des interfaces faciles à utiliser pour plus de 50 corpus et ressources lexicales telles que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la tokenisation, l'étymologie, l'étiquetage, l'analyse syntaxique et le raisonnement sémantique, des wrappers pour des bibliothèques NLP industrielles.**[28]**

Scipy 

Scipy est une bibliothèque pour les calculs techniques et scientifiques. Elle regroupe des modules pour les tâches de science des données et d'ingénierie telles que l'algèbre, l'interpolation, le FFT, ou le traitement de signaux et d'images, Scipy utilise les tableaux et matrices du module NumPy utilisez le module Scipy pour calculer la similarité de cosinus entre deux listes en Python.**[29]**

Tkinter 

Tkinter (de l'anglais Tool kit interface) est la bibliothèque graphique libre d'origine pour le langage Python, permettant la création d'interfaces graphiques. Elle vient d'une adaptation de la bibliothèque graphique Tk écrite pour Tcl **[30]**.

Google search 

Googlesearch est une bibliothèque Python permettant de faire des recherches faciles sur Google sans utiliser leur API. Il utilise des requêtes et BeautifulSoup4 pour scrapper Google. Et pour obtenir des résultats pour un terme de recherche, il suffit d'utiliser la fonction de recherche dans googlesearch **[31]**. Dans notre application, nous avons utilisé Google search pour obtenus des liens vers des sites correspondant à la requête établie à partir du moteur de recherche Google

### IV.4.2.3. Outils de gestion et de collaboration

Il s'agit de solutions qui permettent un meilleur partage de l'information en temps réel, un meilleur partage des documents et une communication simplifiée entre les équipes. Trois outils étaient utilisés :

Stackoverflow 

StackOverflow est un site web proposant des questions et réponses sur un large choix de thèmes concernant la programmation informatique. Il fait partie du réseau de sites Stack Exchange.**[32]**

Google Colab 

Google Colab est un service cloud pour l'écriture et l'exécution du code développé par Google. Ce service donne des capacités matérielles nécessaires pour l'exécution des grandes tâches que l'ordinateur ordinaire n'arrive pas à les résoudre. **[33]**

Google drive 

Google drive est un service de stockage et de partage de fichiers dans le cloud lancé par la société Google. Google Drive, qui regroupe Google Docs, Sheets, Slides et Drawings, est

une suite bureautique permettant de modifier des documents, des feuilles de calcul, des présentations, des dessins, des formulaires, etc. **[34]**

## IV.5. Implémentation de code source du moteur de recherche sémantique

Comme nous avons spécifiés ci dessus, le processus de recherche compose de quatre étapes, nous détaillons en ce qui suit les fonctions implémentées pour chaque étape :

### IV.5.1. Acquisition des données brutes

L'application lit les donnes à partir d'une base de données des fichiers au format csv téléchargé auparavant depuis la plateforme Kaggle, on a choisi comme exemple un fichier csv de titre poste de travail « job post » comme indique la figure ci-dessous :

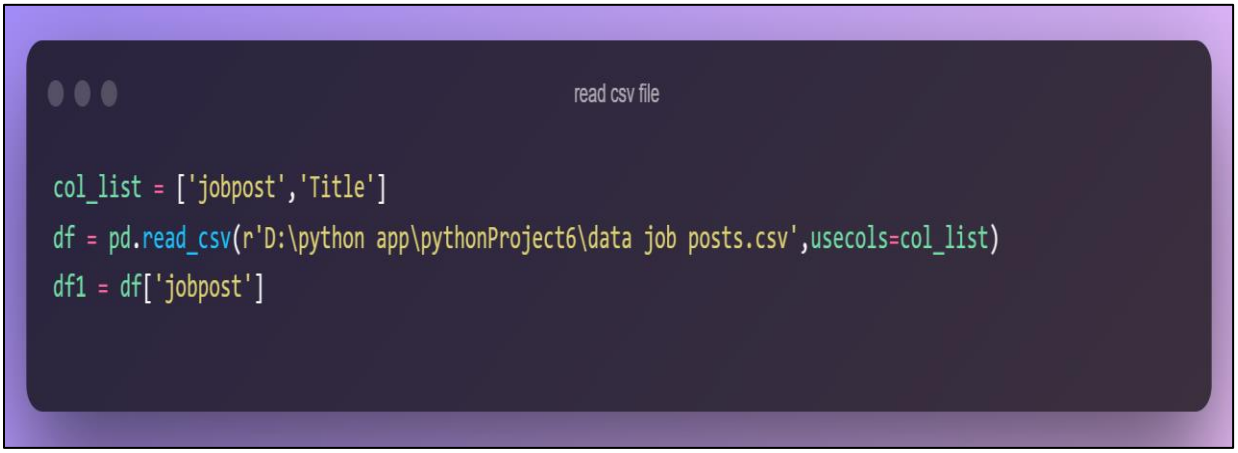


**Figure 14** : Acquisition des données brutes à partir de fichier Csv.

### IV.5.2. Nettoyage des données

En NLP, le nettoyage des données permet toujours de promettre de meilleurs résultats. C'est toujours une bonne pratique d'effectuer le nettoyage des données après l'avoir lu. La fonction de nettoyage des données implémenté est présenté dans la figure ci-dessous :

```python
def remove_stopwords(text, is_lower_case=False):
    global cleaned_tokens
    pattern = r'[^a-zA-z0-9\s]'
    text = re.sub(pattern," ",''.join(text))
    tokens = tokenizer.tokenize(text)
    tokens = [tok.strip() for tok in tokens]
    if is_lower_case:
        cleaned_token = [tok for tok in tokens if tok not in stopword_list]
    else:
        cleaned_tokens = [tok for tok in tokens if tok.lower() not in stopword_list]
    filtered_text = ' '.join(cleaned_tokens)
    return filtered_text
```

**Figure 15 :** Fonction de Nettoyage des données.

IV.5.3. Implémentation Word Embedding

Notre choix est porté sur le modèle vectoriel, nous allons utiliser un vecteur de mots pré-entraîné « glove.6B.300d » qui est disponible en 300 dimensions. La figure ci-dessous représente un exemple d'un vecteur sur 300 D de quelque mot.

```
biomedical 0.34995 0.30508 -0.11671 0.089834 0.019347 -0.36163 -0.31558 -0.55625
0.54188 -0.10815 1.0364 0.12096 0.46901 0.15138 0.25135 0.22477 -0.24121 -0.06975
6 -0.19755 -0.59435 -0.19492 0.14596 0.45069 0.66629 -0.22728 -0.89973 -0.035011
firearm 0.49801 0.31199 0.014536 0.18338 -0.15444 0.27855 -0.063074 0.013543 -0.0
57919 0.4628 -0.16704 0.66892 0.069474 -0.011189 -0.10402 0.19359 -0.063198 -0.78
0.12356 -0.38404 0.86834 0.39648 0.27258 0.13637 -0.18598 0.25354 0.061675 0.2902
apt -0.41618 -0.37872 0.4119 0.015433 0.62914 0.22766 -0.08726 0.069638 0.075714
33536 0.4838 0.20458 -0.099239 -0.068316 0.36846 0.42168 -0.011569 -0.26586 0.113
712 0.034142 -0.3353 0.13945 -0.16584 0.45119 0.1117 0.17954 -0.044742 -0.34739 0
surgeries -0.63834 -0.21662 0.33912 -0.015374 -0.0027064 -0.41182 -0.18348 -0.677
0.21641 -0.65872 0.38399 0.14228 0.21354 -0.10064 0.1587 0.23321 0.12248 -0.7519
6912 0.030972 -0.14674 -0.36171 -0.40152 0.29033 0.71847 -0.44491 -0.34128 0.3652
airliners 0.10745 -0.3197 0.15337 -0.32666 0.18776 1.0215 -0.24665 0.76073 0.1993
```

**Figure 16 :** Vecteur conçus par la technique « Word Embedding »

IV.5.4. Remplissage de dictionnaire « Glove »

**Glove_vector** C'est un dictionnaire contenant des mots en tant que key et les vecteurs de caractéristiques en tant que des valeurs, ce dictionnaire est rempli à partir du fichier retenu de l'étape précédente. La figure suivante montre la fonction implémentée :

```
glove_vectors = dict() #Global Vector for word representation
file = open(r"C:\Users\DELL\Downloads\glove.6B.300d.txt" , encoding = 'utf-8')
for line in file:
    values = line.split()
    word = values[0]
    vectors = np.asarray(values[1:])
    glove_vectors[word] = vectors
file.close()
```

**Figure 17 :** Remplissage de dictionnaire « Glove »

Nous créons une fonction qui prend une phrase et renvoie le vecteur de 300 dimensions à l'aide de « Word Embedding », la figure ci-dessous représente cette fonction :

```
vec_dimension = 300
def get_embedding(x):
    arr  = np.zeros(vec_dimension) # # Creating the placeholders
    text = str(x).split()
    for t in text:
        try:
            vec = glove_vectors.get(t).astype(float)
            arr = arr + vec
        except:
            pass
    arr = arr.reshape(1,-1)[0]
    return(arr/len(text))
```

**Figure 18 :** Fonction de traitement des phrases avec « Word Embedding »

IV.5.5. Obtention du contexte et sens de donnée

Une donnée contient de nombreuses phrases, et une phrase possède un grand nombre de vecteurs basés sur le nombre de mots présents. Pour déterminer le sens général d'une donnée, la fonction calcul la moyenne de tous les vecteurs. Comme indique la figure ci-dessous :

```
out_dict = {}
for sen in df1:
    average_vector = (np.mean(np.array([get_embedding(x) for x in
 nltk.word_tokenize(remove_stopwords(sen))]), axis=0))
    dict = { sen : (average_vector) }
    out_dict.update(dict)
```

**Figure 19 :** Obtention du contexte et sens de donnée.

IV.5.6. Comparaison des requêtes avec les données

La fonction implémenté calcul la similarité entre la donnée et la requête en calculons la distance cosinus entre leurs vecteurs. Si la similarité est plus élevée, plus proche de 1, nous pouvons dire qu'ils ont presque le même sens, si la similarité est proche de 0, on peut dire que leur sens est différent. Selon la fonction représentée dans la figure en bas :

```
def get_similarity(query_embedding, average_vector_doc):
    similarity = (1 - scipy.spatial.distance.cosine(query_embedding,average_vector_doc))
    return similarity
```

**Figure 20 :** Comparaison des requêtes avec les données

IV.5.7. Classement des résultats

A Cette étape, la fonction implémentée fait un classement décroissant selon les valeurs de similarité calculés auparavant en prenons que les vecteurs ou la valeur de la similarité supérieur ou égale à 0.5, c'est-à-dire les plus proche en sens à la requête, et enfin affiche les résultats obtenus**.** En utilisant la fonction illustrée dans la figure ci-dessous :

```
def Ranked_result(query):
    query_words = (np.mean(np.array([get_embedding(x) for x in
 nltk.word_tokenize(query.lower())],dtype=float), axis=0))
    resultats=[]
    for k,v in out_dict.items() :

        if get_similarity(query_words, v) >= 0.5:
            resultats.append([k, get_similarity(query_words, v)])
            resultats = sorted(resultats, key=lambda t: t[1], reverse=True)

    rank = []

    for ch in resultats:

        rank.append(ch[0])

    return rank
```

**Figure 21 :** Classement des résultats selon la similarité

IV.5.8. Recherche offline:

La fonction obtient les résultats à partir du moteur de recherche Google, comme illustré la figure en bas :

```
def searching():
    k = Entry1.get()
    for item in tree.get_children():
        tree.delete(item)


    for j in search(k):

        tree.insert('',END, values=j)
```

**Figure 22** : Obtenus des liens vers les sites à partir du Google

## IV.6. Présentation de l'interface de l'application

L'interface graphique de l'application est très importante, car elle permette de faciliter l'interaction entre l'utilisateur et la machine. L'interface de notre application est ergonomique et facile à utiliser comme montre la figure ci-dessous :


**Figure 23 :** Présentation de l'interface de l'application

### IV.6.1. Introduire la requête de l'utilisateur

Lorsque l'utilisateur ouvre l'application pour effectuer une recherche, il suffit de taper le mot clé à rechercher dans le champ de la recherche comme montre la figure en bas :

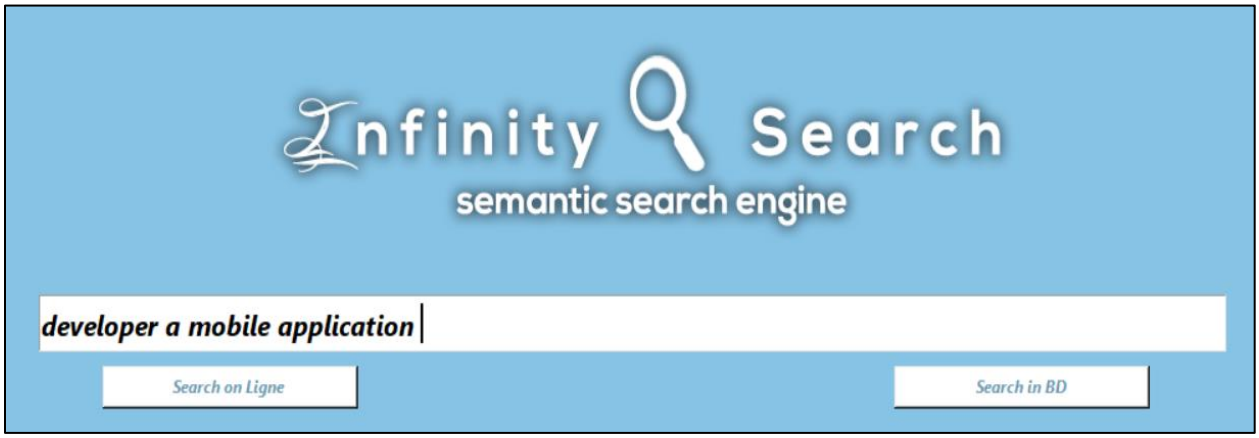


**Figure 24 :** Introduire la requête de l'utilisateur

L'application offre la possibilité de recherche offline c'est à dire à partir de la base de données locale ou bien en ligne directement obtient les résultats à partir du moteur de recherche Google. les résultats affichés comme montre la figure ci-dessous:

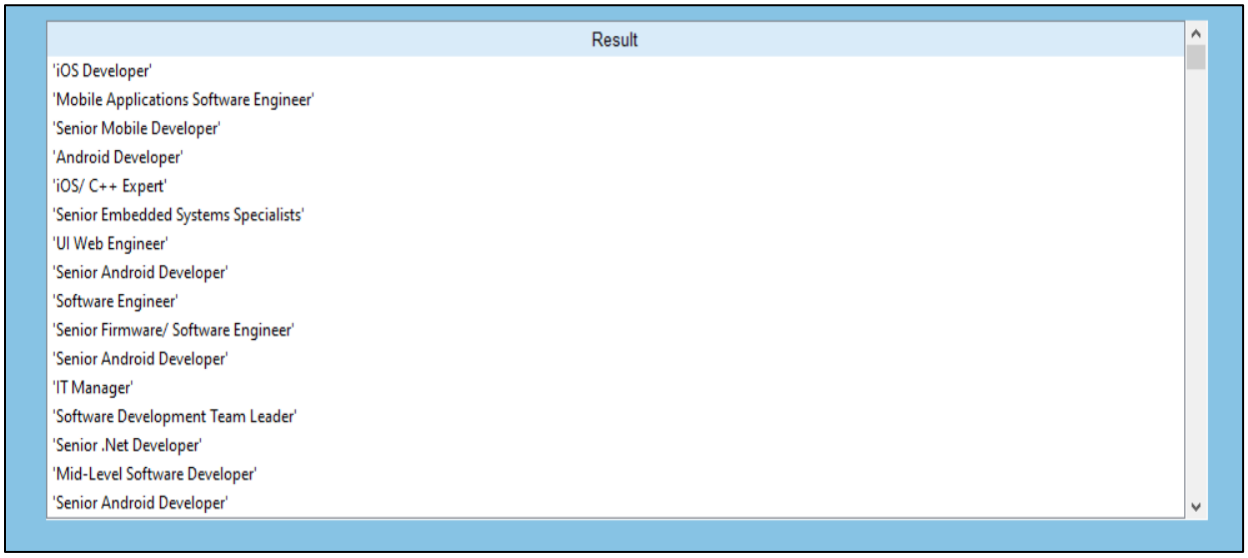


**Figure 25 :** Affichage des résultats depuis la base de donne locale.

Lorsque l'utilisateur choisis de chercher via l'option en ligne, les résultats obtenus sont des liens vers des sites correspondant à la requête établie. La figure ci-dessous représente les résultats obtenus :

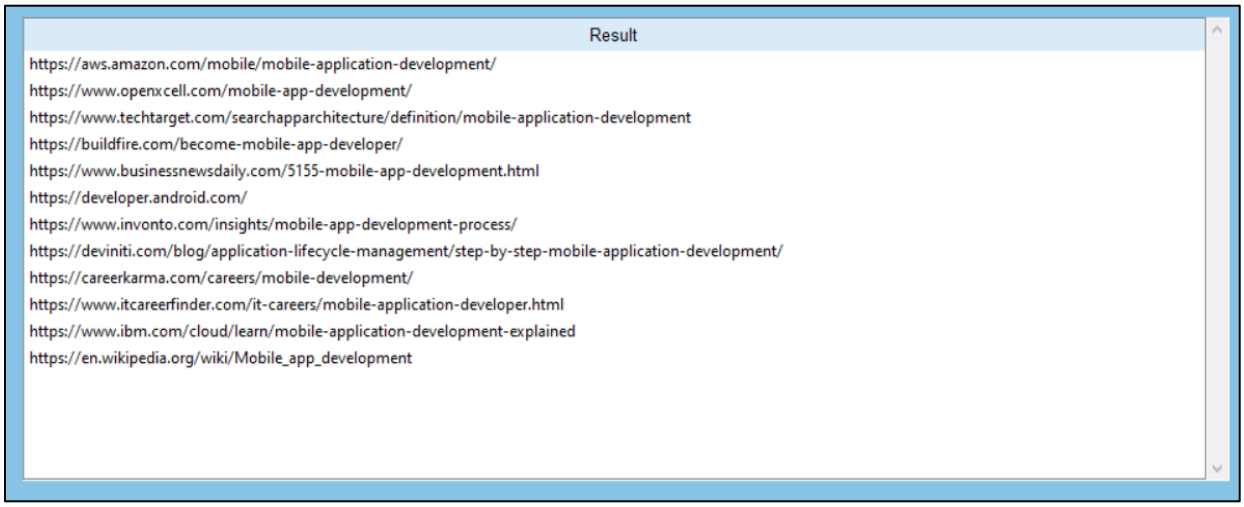


**Figure 26 :** Affichage des résultats en ligne.

## IV.7. Conclusion

L'étape de réalisation représente l'étape la plus importante du cycle de vie de l'application. Dans ce chapitre, nous avons décris brièvement le processus de création de notre application en spécifiant l'environnement, les outils et les langages de développement associés à notre système. Les résultats que nous avons pris ont démontré l'efficacité de notre approche sémantique.

# Conclusion générale

# Conclusion générale

Le rôle d'un moteur de recherche est très simple : aider les utilisateurs à trouver la réponse la plus pertinente à leur requête. Aujourd'hui, les moteurs de recherche sont devenus incontournables surtout avec le développement accéléré de l'informatique, ils deviennent de plus en plus utilisés dans pratiquement tous les secteurs. En effet, ils aident les internautes à trouver facilement et très rapidement de nombreuses informations. Par ailleurs, sans les moteurs de recherche, Internet n'aurait presque plus aucun sens. Tout de même, les moteurs de recherche constituent le fil conducteur qui relie un internaute aux sites Web.

Notre projet s'inscrit dans ce cadre, il s'agit de développer une application bureau dédiée à l'utilisation générale nommée « **Infinity Search** » qui représente un moteur de recherche qui porte une méthode de sélection de résultats basées sur la recherche sémantique en utilisant les techniques du domaine recherche d'information.

Ce qui différencie notre application « **Infinity Search** » des applications existantes, c'est que la nôtre utilise une base de données locale pour la recherche offline ce qui donne la possibilité de recherche même dans le cas d'absence de connectivite avec le réseau Internet.

Pour cela, nous avons en premier lieu présenté les 2 domaines : NLP, Recherche d'information, ainsi une description sommaire des moteurs de recherche existe actuellement. Nous avons établi par la suite, une étude préliminaire pour identifier les différents outils et les langages de développement ainsi les bibliothèques logicielles que nous avons utilisées.

Enfin, les fonctions et les procédures implémenter dans notre application ont été exposés avec illustration en figures.

Il est essentiel pour un étudiant-chercheur de pouvoir avoir un regard critique sur la littérature scientifique et de pouvoir communiquer de manière claire, concise et précise à propos de thématiques scientifiques. Le but de ce projet est de montrer l'impact de l'ajout de la couche sémantique a la recherche d'information malgré que les techniques utilises dans notre application ne sont pas assez développées puisque c'est un domaine d'actualité qui ne cesse pas d'avancer.

Ce projet nous a été très bénéfique, car nous avons enrichi nos connaissances sur les deux plans : théorique et pratique. Il nous a aussi permis de découvrir et d'acquérir de nouvelles connaissances en matière de recherche d'information et de traitement de langage naturel. Finalement on peut imaginer de nombreuses perspectives pour améliorer ce moteur, on peut citer par exemple :

- Utilisation de tous les modèles de recherche d'information existent pour l'obtention des meilleurs résultats en essayons de combiné plusieurs modèles ;
- Offrir la possibilité de la recherche vocale ;
- L'ajout de la fonctionnalité de la recherche Multi-langues ;
- Le stockage des résultats fréquemment recherches pour le besoin de la réutilisation ;
- Le filtrage périodique des donnes par élimination des résultats erronées.

Avec l'application de toutes ces suggestions, la fiabilité du moteur de recherche va augmenter, et si on arrive à ce stade on peut envisager l'utilisation de cette application pour un excellent résultats pour les besoins de l'utilisateur.

# Webographies

**[1]** http://www.iro.umontreal.ca/~nie/IFT6255/historique-RI.html

**[2]** https://www.linternaute.fr/expression/langue-francaise/564/autant-chercher-une-aiguille-dans-une-botte-de-foin/

**[3]** https://fr.wikipedia.org/wiki/Recherche_d%27information

**[4]** Karen Sauvagnat. Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structures. Informatique [cs].

- Université Paul Sabatier - Toulouse III, 2005, CJ Rijsbergen. Van : Information retrieval. London : Butterwoths, 1979

**[5]** https://www.researchgate.net/figure/Processus-de-recherche-dinformation-15-Modeles-de-recherche-dinformation-Des-exemples_fig1_327395681

**[6]** https://dac.lip6.fr/wp-content/uploads/2019/01/RITAL_2019-1.pdf

**[7]** https://blog.smart-tribune.com/fr/definition-nlp

**[8]** https://www.aqsone.com/blog/2020/data-science-fr/quest-ce-que-le-natural-language-processing

**[9]** https://www.sbert.net/examples/applications/semantic-search/README.html

**[10]** https://twitter.com/toshi2fly/status/911306344376012800/photo/2

**[11]** https://www.twaino.com/definition/m/moteur-de-recherche

**[12]** https://fr.wikipedia.org/wiki/Moteur_de_recherche

**[13]** https://www.01net.com/telecharger/windows/Internet/moteur_rech/fiches/31542.html

**[14]** http://googlepress.blogspot.com/2004/10/google-announces-desktop-search.html

**[15]** https://mylittlebigweb.com/comment-fonctionne-un-moteur-de-recherche/

**[16]** https://www.blogdumoderateur.com/tools/microsoft-bing/

**[17]** https://www.google.com/

**[18]** https://www.bing.com/

**[19]** https://search.yahoo.com/

**[20]** https://www.ncis-dz.com/produit/dell-algerie-vostro-15-586284-156-i5-5200u/

**[21]** https://www.lebigdata.fr/python-langage-definition

**[22]** https://fr.wikipedia.org/wiki/PyCharm

**[23]** https://datascientest.com/kaggle-tout-ce-quil-a-savoir-sur-cette-plateforme

**[24]** https://www.bibl.ulaval.ca/geostat/statistiques/CSV-guide.pdf

**[25]** https://www.techno-science.net/definition/1470.html

**[26]** https://www.linkedin.com/pulse/les-meilleures-biblioth%C3%A8ques-pour-applications-big-data-ben-rhouma?trk=public_profile_article_view

**[27]** https://datascientest.com/numpy

**[28]** https://fr.quish.tv/natural-language-processing-with-python

**[29]** https://intelligence-artificielle.com/top-bibliotheque-python-ia-machine-learning

**[30]** https://fr.wikipedia.org/wiki/Tkinter

**[31]** https://pypi.org/project/googlesearch-python/

**[32]** https://fr.wikipedia.org/wiki/Stack_Overflow

**[33]** https://ledatascientist.com/google-colab-le-guide-ultime/

**[34]** https://fr.wikipedia.org/wiki/Google_Drive