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

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en informatique

Spécialité : Technologie de l'information et de la communication

THEME

Système question/réponse pour la langue arabe en utilisant le deep learning

Présenté par :

Zerguine Soumia

Bendrimia Amira

Soutenu publiquement le : 20/06/2023

Devant le jury composé de:

Président : Barkat hadj

Examinateur : Badaoui atika

Encadreur : Mme. Saad Saoud Manel

2022/2023

Dédicace

Je dédie ce travail

A l'âme de mon incroyable grand-mère

À l'être le plus cher de ma vie, ma mère

À qui m'a aidé à devenir ce que je suis aujourd'hui, mon père

A ma sœur Roumaissa

Mon frère Abed Samia

À mon mari Oussama pour m'avoir encouragé jusqu'à la fin

A ma princesse Amira chère amie avant d'être binôme

A mon amie Safia, qui m'a toujours encouragée, et à qui je souhaite plus de succès

À tous mes amis de la promotion de 2éme année master en informatique

Zerguine Soumia

Dédicace

Je dédie ce travail

A l'âme de mon incroyable grand-mère

À l'être le plus cher de ma vie, ma mère

À l'homme de ma vie, mon exemple éternel, mon père

À Mes chères sœurs et frère

Rima, Dounia, Ahlem et Youcef

À Soumia, chère amie avant d'être binôme

A mes amies Safa, Safia et Amel qui m'ont toujours encouragés, et à qui je souhaite plus de succès

A Tout personne qui occupe une place dans mon cœur

À tous mes amis de la promotion de 2éme année Master T.I.C en informatique

Bendrimia Amira

Remerciements

Avant tout, on tiens à remercier et exprimer notre profonde gratitude envers le bon dieu tout puissant qui nous a donné la force d'accomplir ce modeste travail.

On exprime notre reconnaissance et nos vif remerciements à notre encadrante Mme. Saad Saoud Manel pour sa présence, ses conseils et toute l'aide qu'elle nous a apporté tout au long de la réalisation de ce projet,

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre projet de fin d'études en acceptant d'examiner ce travail et de l'enrichir par leurs propositions.

Nous souhaitons exprimer notre gratitude envers nos familles pour leur soutien et encouragements tout au long de ce travail.

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

ملخص

معظم الأسئلة الدينية التي يطرحها المسلمون قد تم الرد عليها فعليا من قبل العلماء المتخصصين، ولكن صعوبة الاتصال المباشر معهم حالت دون الإجابة الفورية، مما اوجب البحث عن طرق بديلة للحصول على هذه المعلومات.

هدف هذا البحث هو إنشاء نظام يجيب تلقائيًا على الأسئلة الدينية المطروحة من قبل المسلمين باللغة العربية، باستخدام تقنيات الذكاء الاصطناعي وتحليل اللغة الطبيعية لتوفير إجابات تلقائية وفورية على هذه التساؤلات.

في بحثنا، استخدمنا نموذج التعلم العميق يتميز بقدرته على فهم اللغة العربية الطبيعية واستخلاص المعنى من النصوص المكتوبة.

يتألف النهج المقترح من وحدتين رئيسيتين، الوحدة الأولى هي وحدة إدخال طلبات المسلمين، حيث يتم استقبال أسئلة المستخدمين الدينية وإدخالها في النظام. والوحدة الثانية تستجيب بشكل مختص للطلب من خلال البحث على أساس التشابه الدلالي في قواعد البيانات الموجودة.

أظهرت النتائج التجريبية أن نموذجنا « Sora-QA » المبني على التعلم العميق, قدم نتائج أفضل بالمقارنة مع طريقة TF-IDF

الكلمات المفتاحية : أنظمة الإجابة على الأسئلة ، تصنيف النصوص ، اللغة العربية ،التعلم الآلي ، التعلم العميق ,TF-IDF , BERT

Abstract

Most of the religious questions asked by Muslims have already been answered by specialized scholars, but the difficulty of direct communication with them has hindered immediate responses, necessitating the search for alternative methods to obtain this information.

The goal of this research is to create a system that automatically answers the religious questions posed by Muslims in the Arabic language, using artificial intelligence techniques and natural language processing to provide automatic and instant answers to these inquiries.

In our research, we used a deep learning model that stands out for its ability to understand natural Arabic language and extract meaning from written texts.

The proposed approach consists of two main units. The first unit is the input unit for Muslim inquiries, where users' religious questions are received and entered into the system. and the second unit responds concretely to the request by doing a search based on semantic similarity in existing databases.

The experimental results have shown that our deep learning-based model « Sora-QA » yielded better results compared to the TF-IDF method.

Keywords : Question Answering system, Arabic Language, Text classification, Machine learning, Deep learning, TF-IDF, BERT.

Résumé

La plupart des questions religieuses posées par les musulmans ont déjà reçu des réponses de la part de spécialistes. Cependant, la difficulté de les contacter directement a empêché d'obtenir des réponses immédiates, ce qui a conduit à la recherche de méthodes alternatives pour obtenir ces informations.

L'objectif de ce travail est de créer un système qui répond automatiquement aux questions religieuses posées par les musulmans en langue arabe, en utilisant des techniques d'intelligence artificielle et d'analyse du langage naturel pour fournir des réponses automatiques et instantanées à ces interrogations.

Dans ce manuscrit, nous avons utilisé un modèle d'apprentissage profond, qui se distingue par sa capacité à comprendre la langue arabe naturelle et à extraire le sens des textes écrits.

Les résultats expérimentals ont montré que notre modèle « Sora-QA » basé sur le deep learning a donné de meilleurs résultats par rapport à la méthode TF-IDF.

Mots clés : Systèmes de Question-Réponse, Langue arabe, Classification de textes, Apprentissage automatique, Apprentissage profond, TF-IDF, BERT.

Table des matières

Liste des figures

Liste des tableaux

Liste des Abréviations

ArabiQA Arabic Question Answering.

IA Intelligence Artificielle.

JAWEB web-based Arabic question answering application system.

NLP Natural Language Processing.

QARAB Arabic Question Answering System.

RI Recherche d'Information.

RNA Réseaux de neurones Artificiels.

SQR systèmes de questions-réponses.

SVM Support Vector Machine.

IHM Interaction homme-machine

TALN Traitement Automatique du Langage Naturel.

Tf-Idf Fréquence du terme * Fréquence inverse du document.

BERT Bidirectional Encoder Representations from Transformers

SQuAD Stanford Question Answering Dataset

NSP Next Sentence Prediction

Introduction Générale

Introduction Générale

1.1 Contexte et problématique :

Les gens ont souvent de nombreuses questions et préoccupations et souhaitent trouver des réponses rapidement, en particulier lorsqu'il s'agit de questions religieuses . Parfois, nous avons beaucoup d'idées sur notre religion, en particulier lorsque nous lisons le Coran et y réfléchissons. Par conséquent, l'homme a toujours cherché à développer des moyens et des méthodes pour faciliter l'accès aux réponses, notamment : le système des questions et réponses.

Le système de questions-réponses est un système simple qui aide l'utilisateur à obtenir sa réponse de manière très simple et facile, mais le problème réside dans l'indisponibilité de ces systèmes en arabe, et peut-être car le développement d'un système de questions-réponses en langue arabe présente des défis spécifiques. L'un de ces défis est la difficulté d'analyser la langue arabe en raison de sa structure grammaticale complexe et de sa diversité linguistique ce qui nécessite l'utilisation de techniques avancées dans le domaine du traitement du langage naturel, et on a aussi le manque des ensembles des données dans cette langue ce qui limite l'entraînement des modèles d'apprentissage profond. Le travail présenté dans ce mémoire porte sur la création d'un système de question-réponse pour la langue arabe plus précisément sur les questions du Coran.

1.2 Objectif :

L’objectif de notre travail consiste à implémenter un système question-réponse qu’on a nommé Sora-QA pour répondre aux questions basées sur les informations contenues dans les versets du Coran.

Sora-QA est un système de question-réponse basé sur l'apprentissage profond, développé en Python et fonctionnant via un site web. Il utilise le modèle de BERT et la méthode TF-IDF pour obtenir les meilleures réponses.

1.3 Structure du mémoire :

Le présent mémoire est divisé en trois chapitres :

Chapitre1 : Ce chapitre est consacré à l'exploration des systèmes de questions-réponses. La présentation d'une manière détaillée des éléments importants de l'apprentissage automatique et de la classification du texte.

Chapitre 2 : Ce chapitre est consacré à la présentation d'une manière détaillée des éléments importants de l'apprentissage automatique, et présentation quelque travaux connexes.

Chapitre 3 : Le troisième chapitre expose les performances des modèles utilisés et leur déploiement dans un site web.

Enfin, notre travail s'achève par une conclusion générale résumant les grands points qui ont été abordés.

Chapitre 01 :

Systèmes question réponse

1.1. Introduction :

Le système question réponse est un domaine de recherche qui a un grand intérêt ces dernières années, ce qui a conduit à des avancées significatives de la part des chercheurs. Dans ce chapitre, nous expliquerons ce qu'est un système de question-réponse et pourquoi il est utilisé.

1.2. Qu'est-ce qu'une question ?

« Demande faite pour obtenir une information, vérifier des connaissances : Répondre aux questions des enquêteurs. » [1]

Une question est une forme d'expression linguistique utilisée pour demander des informations, rechercher des éclaircissements ou solliciter une réponse. Nous posons généralement des questions pour obtenir des informations manquantes sur un sujet ou résoudre un problème scientifique ou autre. Cependant, il est essentiel de connaître la méthode de formulation des questions, car c'est la clé pour obtenir des informations précises et spécifiques.

Il existe différents types de questions, notamment :

- Fermée : il existe 2 réponses possibles : oui ou non. On dit alors que c'est une interrogation totale.

  Exemple : Possédez-vous une voiture ?

- Ouverte : la question ouverte est un type d'interrogation qui appelle des réponses explicatives. Elle est utilisée pour comprendre, faciliter l'expression, dialoguer et échanger.

  Ce type d'interrogation commence par un adverbe tel que "Pourquoi", "Combien", "Comment", "Où", "Quand", etc. La question contient une information nouvelle qui n'est pas incluse dans la question elle-même.

  Exemple : Pourquoi avez-vous choisi d'étudier en France ?

- Factuelle : permet de se concentrer sur les faits et de décrire une situation en évitant les perceptions ou les jugements. [2]

Exemple : Quelle est la plus Grande Guerre au monde ?

1.3. Qu'est-ce qu'une réponse ?

Une réponse est une information ou une explication fournie en réponse à une question posée. Avec un système de question-réponse, nous devons apporter à l'utilisateur une réponse à la question qu'il a formulée. Or, cette notion qui semble intuitive pose des problèmes de définition. Tout d'abord, la notion de « réponse » qui peut référer à un très court fragment de texte aussi bien qu'a une longue phrase justificative n'est pas clairement définie dans le langage courant. En conséquence, les personnes peuvent répondre d'une manière différente à une même question. Classiquement, la quasi-totalité des systèmes de question-réponse possèdent des architectures communes mais cela ne signifie pas qu'ils soient similaires. La différence principale entre ces systèmes réside dans l'approche proposée pour chacun d'eux. De surcroît, cette différence se survient dans les techniques et les outils utilisés par ces systèmes. Ainsi, pour quelle langue, un tel système est mis en place. [3]

1.4. Système Question-Réponse

« Un système de question-réponse est une application qui cherche dans un corpus de document, une réponse exacte à une question posée en langue naturelle » [4]

La question-réponse est une méthode d'interrogation qui utilise des questions formulées en langage naturel afin d'obtenir des réponses précises, sans nécessiter de renvoi à des pages spécifiques. Cette approche repose sur l'utilisation de techniques avancées telles que le Traitement Automatique du Langage Naturel (TALN), la Recherche d'Informations (RI) et l'Interface Homme-Machine (IHM). Les systèmes de question-réponse sont conçus pour analyser la question posée, comprendre son sens et fournir une réponse pertinente et complète. Cette approche se situe à l'intersection de plusieurs domaines et vise à faciliter l'accès à l'information de manière efficace et conviviale. [5]

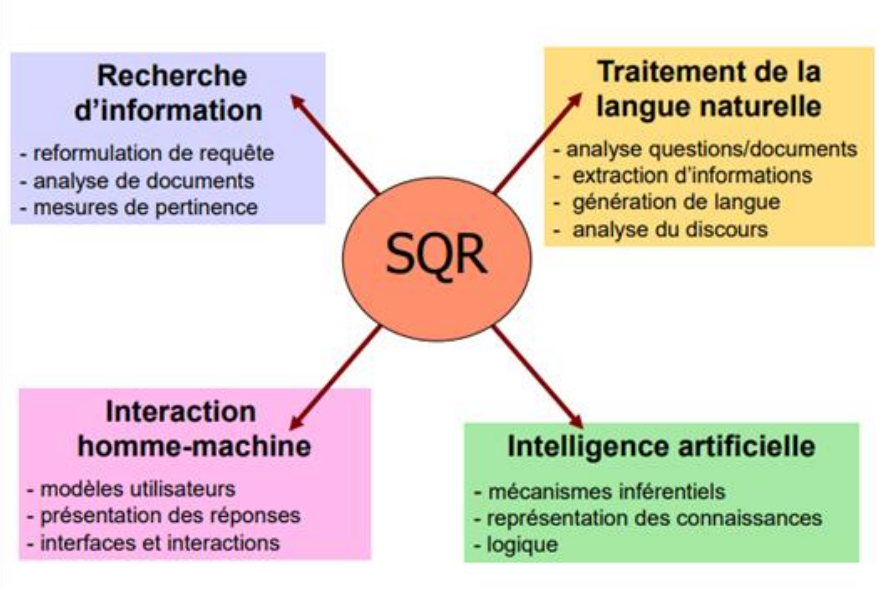

Figure 0-1 Système Question réponse [33]

1.4.1. Le traitement automatique du langage naturel

Le langage naturel désigne la langue « normale » parlée par un être humain, et le traitement automatique du langage naturel, est une discipline s'appliquant au domaine de l'informatique et du langage. Il est utilisé par exemple pour les traductions, la reconnaissance vocale ou encore les réponses automatiques aux questions.

1.4.2. La recherche d'information

« La recherche d'information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations » [6].

La Recherche d'Information (RI) consiste à localiser et fournir un ensemble de documents à un utilisateur en fonction de ses besoins en informations. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur.

1.4.3. L’interaction homme-machine (IHM)

L'interaction homme-machine (IHM) désigne le domaine d'étude et de recherche qui se concentre sur la conception, le développement et l'évaluation des interfaces entre les êtres humains et les machines. Il s'agit de créer des systèmes informatiques et des dispositifs interactifs qui permettent aux utilisateurs de communiquer et d'interagir de manière efficace avec les machines, que ce soit des ordinateurs, des smartphones, des tablettes, des objets connectés ou d'autres types de dispositifs technologiques[7].

1.4.4. L’intelligence artificielle

L'intelligence artificielle est une discipline informatique qui se concentre sur la création de machines intelligentes, en opposition à l'intelligence naturelle des êtres vivants. Ce terme a évolué au fil du temps et englobe maintenant toutes les idées visant à permettre aux machines d'émuler et de surpasser les capacités cognitives humaines[8].

1.5. Classification de systèmes de questions-réponses

Les systèmes de question-réponse peuvent être classifiés selon différentes dimensions. Voici quelques classifications courantes :

1.5.1. Classification selon le type de réponse :

- Réponses courtes : Ces systèmes fournissent des réponses concises et précises à des questions spécifiques.
- Réponses longues : Ces systèmes génèrent des réponses plus détaillées et informatives, souvent en utilisant des techniques de résumé automatique.

1.5.2. Classification selon le domaine d'application :

- Domaine ouvert : ne sont pas limités à un domaine spécifique et fournissent une réponse courte à une question, traitée en langage naturel. [9]

- Domaine fermé : permettent de répondre aux questions relatives à un domaine particulier (médecine, cinématographie, aquariophilie, etc) en se basant sur les connaissances spécifiques aux domaines souvent formalisés dans des ontologies. [9]

1.5.3. Classification selon l'approche de traitement du langage :

- Approche statistique : Ces systèmes utilisent des techniques statistiques et probabilistes pour analyser et traiter les questions et les textes.
- Approche basée sur l'apprentissage automatique : Ces systèmes utilisent des algorithmes d'apprentissage automatique pour améliorer leur performance en s'entraînant sur des données d'entraînement.
- Approche basée sur les règles : Ces systèmes utilisent des règles préalablement définies pour analyser les questions et générer les réponses.

1.6. Architecture générique d'un SQR

L'architecture des systèmes de question-réponse traitant les questions factuelles est similaire dans la plupart des implémentations.

Un système de question-réponse peut généralement être décomposé en plusieurs étapes : analyse des questions, sélection des passages, et extraction de la réponse.[10] :

- Analyse de question : A Pour fonction de comprendre le sens de la question, Et de déterminer le type d'information recherchée.
- Traitement des documents : Permet de transformer les données textuelles en connaissances utilisables pour répondre aux questions.
  Le prétraitement des documents de référence permet d'appliquer les méthodes d'extractions nécessaires pour identifier l'information essentielle du corpus de Référence et remplir les bases de connaissances.

- Extraction des réponses : Est réalisée à partir d'une représentation de la question et d'un mécanisme d'appariement, permettant d'extraire un ensemble de Réponses

candidates qui sont ensuite évaluées pour déterminer la réponse la Plus vraisemblable.

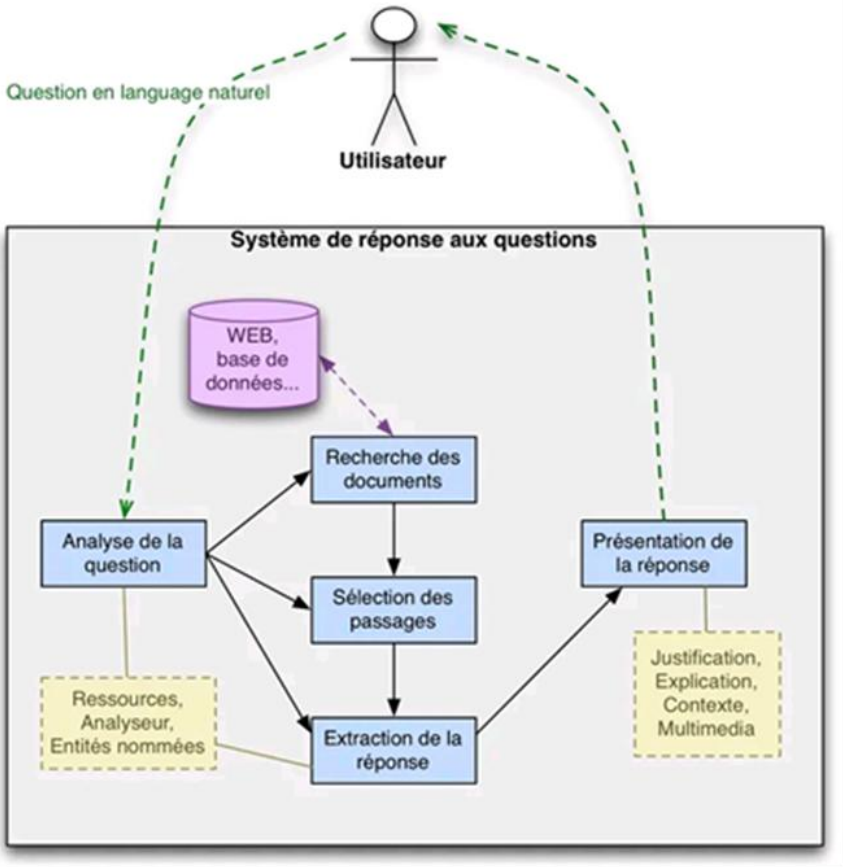

Figure 0-2 Architecture générique d'un SQR [11]

1.7. Système de question-réponse Arabe

Les systèmes de questions-réponses, qui permettent d'obtenir des réponses précises aux questions formulées dans une langue naturelle, ont connu des évolutions majeures dans plusieurs langues telles que l'anglais, le français et d'autres. Cependant, la langue arabe n'a pas bénéficié du même niveau d'attention et de développement. Elle fait face à des défis spécifiques qui rendent difficile l'application et le développement de systèmes de questions-réponses avec le même niveau d'efficacité et de précision. Parmi ces défis :

- La complexité de la grammaire et de la conjugaison : La langue arabe se distingue par ses règles grammaticales et ses conjugaisons complexes et variées, ce qui rend l'analyse et la compréhension des phrases plus compliquées. Cela nécessite le développement de modèles et de techniques robustes pour faire face à ces complexités.
- L'absence de ponctuation et d'espaces : Les textes arabes traditionnels manquent souvent de signes de ponctuation et d'espaces entre les mots, ce qui rend plus difficile l'analyse et l'interprétation des textes, ainsi que la délimitation des phrases et des expressions.
- La polysémie : La langue arabe se caractérise par la multiplicité des significations possibles des mots et des expressions, ce qui constitue un défi pour comprendre le contexte et extraire les bonnes réponses.
- Le manque de ressources et de données : Le manque de ressources et de données disponibles en langue arabe entrave le développement de modèles de questions-réponses. Il peut être difficile de collecter une quantité suffisante de données diverses et fiables pour l'entraînement et l'évaluation de ces modèles.

Malgré tous ces défis, il existe plusieurs facteurs motivants pour choisir la langue arabe, tels que :

- La langue arabe est la sixième langue la plus parlée dans le monde, ce qui en fait une langue d'importance mondiale.
- Elle compte environ 280 millions de locuteurs natifs et environ 250 millions de locuteurs non natifs, ce qui en fait une langue largement répandue.
- La langue arabe est l'une des six langues officielles des Nations Unies, ce qui souligne son importance dans les affaires internationales et la diplomatie.
- On observe une croissance des données textuelles arabes sur le web, ce qui témoigne de l'importance croissante de la présence en ligne de la langue arabe.
- Il y a une demande croissante de logiciels et de technologies de pointe pour la langue arabe, ce qui crée une opportunité pour le développement de systèmes de questions-réponses efficaces et performants.

1.8. Conclusion

Dans ce chapitre, nous avons présenté quelques notions sur les systèmes question réponse, c'est quoi une question, une réponse, l'architecture de système question réponse et ses classifications. Dans le chapitre suivant, nous allons aborder le concept d'apprentissage profond et après nous allons présenter quelques systèmes question-réponse existants, leurs fonctionnements et leurs objectifs, et nous préparerons une description sommaire pour notre propre système tout en identifiant ses principaux objectifs.

Chapitre 02 :

Deep learning pour les systèmes question- réponse

2.1. Introduction

Pendant notre travail, nous avons examiné plusieurs travaux liés à notre recherche sur les systèmes de question-réponse. Dans ce chapitre, nous aborderons également le concept d'apprentissage profond. De plus, nous allons présenter certains de ces travaux pour identifier leurs points forts et leurs faiblesses. L'objectif est d'exploiter les idées provenant de ces travaux pour améliorer notre propre système de question-réponse.

2.2. Apprentissage profond

L'apprentissage profond est un sous-ensemble de l'apprentissage automatique, où les réseaux neuronaux artificiels des algorithmes conçus pour fonctionner comme le cerveau humain émettent à partir de grandes quantités de données. [12]

2.2.1. L’apprentissage automatique :

Le Machine Learning ou apprentissage automatique est un domaine scientifique, et plus particulièrement une sous-catégorie de l’intelligence artificielle. Son objectif est de permettre aux algorithmes de découvrir des motifs récurrents, appelés "patterns", au sein d’ensembles de données. Ces données peuvent prendre différentes formes telles que des chiffres, des mots, des images, des statistiques, et bien d'autres. [13]

2.2.1.1. Types de Machine Learning :

Le Machine Learning comporte deux principaux types d'algorithmes : l'apprentissage supervisé et l'apprentissage non supervisé. La différence entre les deux se définit par la méthode employée pour traiter les données afin de faire des prédictions. [14]

Machine Learning supervisé	Les algorithmes de Machine Learning supervisé sont les plus couramment utilisés. Avec ce modèle, un data scientist sert de guide et enseigne à l'algorithme les conclusions qu'il doit tirer. Tout comme un enfant apprend à identifier les fruits en les mémorisant dans un imagier, en apprentissage supervisé, l'algorithme apprend grâce à un jeu de données déjà étiqueté et dont le résultat est prédéfini. Comme exemples de Machine Learning supervisé, on peut citer des algorithmes tels que la régression linéaire et logistique, la classification en plusieurs catégories et les machines à vecteurs de support.
Machine Learning non supervisé	La Machine Learning non supervisé utilise une approche plus indépendante dans laquelle un ordinateur apprend à identifier des processus et des schémas complexes sans un quelconque guidage humain constant et rigoureux. Le Machine Learning non supervisé implique une formation basée sur des données sans étiquette ni résultat spécifique défini. Pour continuer avec l'analogie de l'enseignement scolaire, le Machine Learning non supervisé s'apparente à un enfant qui apprend à identifier un fruit en observant des couleurs et des motifs, plutôt qu'en mémorisant les noms avec l'aide d'un enseignant. L'enfant cherche des similitudes entre les images et les sépare en groupes, en attribuant à chaque groupe sa propre étiquette. Comme exemples d'algorithmes de Machine Learning non supervisé, on peut citer la mise en cluster de k-moyennes, l'analyse de composants principaux et indépendants, et les règles d'association.

Tableau 0-1 Types d'algorithmes de Machine Learning

2.2.2. L'apprentissage profond :

Le Deep Learning ou apprentissage profond est l'une des technologies principales de la Machine Learning. Avec le Deep Learning, nous parlons d'algorithmes capables de mimer les actions du cerveau humain grâce à des réseaux de neurones artificielles. Les réseaux sont

composés de dizaines voire de centaines de « couches » de neurones, chacune recevant et interprétant les informations de la couche précédente. [15]

L'apprentissage profond présente un avantage majeur : il peut traiter de grandes quantités de données brutes et apprendre automatiquement des caractéristiques pertinentes, sans avoir besoin de spécifier explicitement ces caractéristiques à extraire. Les réseaux de neurones profonds sont constitués de plusieurs couches, où chaque couche effectue des transformations sur les données et les transmet à la couche suivante. Cela permet de créer des modèles complexes et de saisir des représentations de plus en plus abstraites à mesure que l'information se propage à travers les différentes couches du réseau.

2.2.3. Les réseaux de neurones profonds

« Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau. » [16]

Les principes fondamentaux du réseau de neurones comprennent le parallélisme, les poids synaptiques et l'apprentissage. Le parallélisme se réfère à la capacité des neurones à fonctionner en parallèle, ce qui permet un traitement du signal simultané. Les poids synaptiques sont les connexions entre les neurones et ils déterminent l'intensité de l'interaction entre chaque paire de neurones. L'apprentissage se réfère à la modification des poids synaptiques lors du processus d'apprentissage, afin que le réseau puisse accomplir la fonction souhaitée. [16]

2.2.3.1. Architectures des réseaux de neurones

Les réseaux neuronaux peuvent varier en fonction des données, de leur complexité et de la méthode de traitement utilisée. Chaque architecture présente ses propres avantages et limites, et les combiner permet d'optimiser les résultats en fonction de l'objectif. [17]

Pour adapter un réseau neuronal à un problème spécifique, il est nécessaire de choisir sa topologie et les poids des connexions entre les neurones.

La topologie des réseaux neuronaux peut être très variée. Il est possible de concevoir différents types de réseaux en modifiant les règles de connexion. Parmi ces types, on retrouve les réseaux à une seule couche, les réseaux multicouches et les réseaux récursifs. [18]

2.2.3.2. Les types des réseaux de neurones

Les réseaux à une seule couche : sont des réseaux neuronaux simples composés d'une seule couche de neurones. Chaque neurone est connecté directement aux entrées et génère une sortie. Ces réseaux sont souvent utilisés pour des tâches de classification binaire où ils peuvent

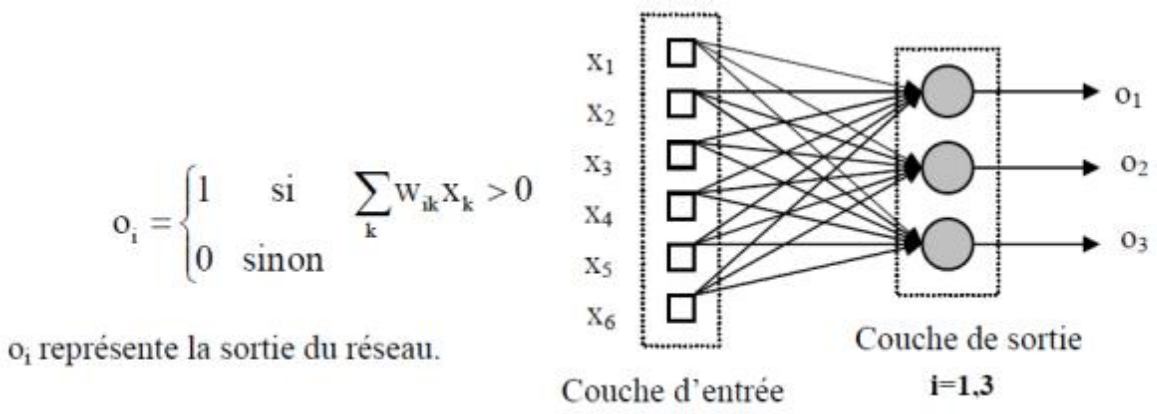

Figure 0-4 Exemple d'un réseau d'une seule couche [16]

apprendre à distinguer deux classes en fonction des caractéristiques d'entrée.

- Les réseaux multicouches : Les réseaux multicouches, également connus sous le nom de réseaux neuronaux à plusieurs couches, sont une forme plus avancée de réseaux neuronaux. Ils sont composés de plusieurs couches de neurones, généralement une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Les neurones d'une couche sont connectés à ceux de la couche suivante, formant ainsi un flux d'information du début à la fin du réseau.

Les réseaux multicouches peuvent traiter des données complexes et effectuer des tâches avancées comme classifier plusieurs classes, faire de la régression ou générer du contenu.

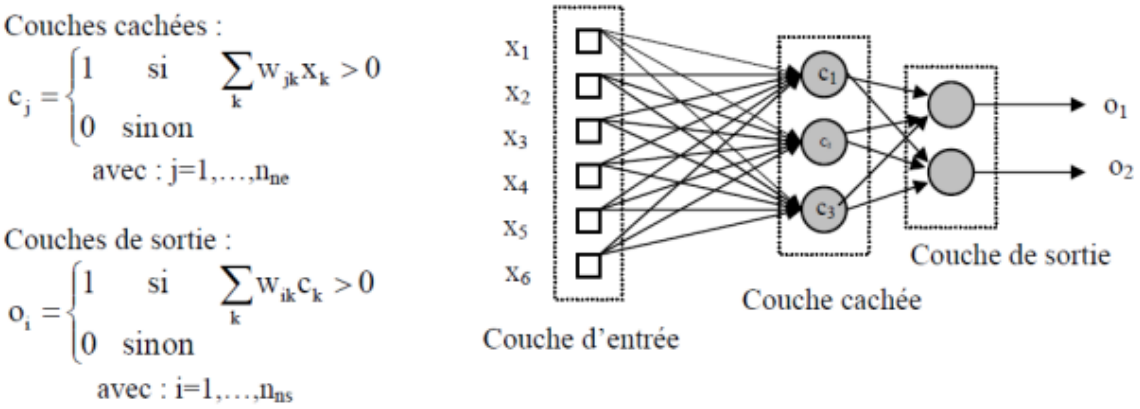

Figure 0-5 Exemple d'un réseau multicouche [16]

- Les réseaux récursifs : Les réseaux récursifs sont des réseaux neuronaux spéciaux qui peuvent modéliser des structures hiérarchiques ou séquentielles, comme les arbres ou les phrases. Ils prennent en compte les relations entre les éléments en utilisant des connexions récurrentes qui bouclent sur elles-mêmes. Cela leur permet de comprendre le contexte et de traiter des données avec une dépendance temporelle ou une structure complexe. Les réseaux récursifs sont souvent utilisés dans le traitement du langage naturel, la traduction automatique et la reconnaissance vocale.

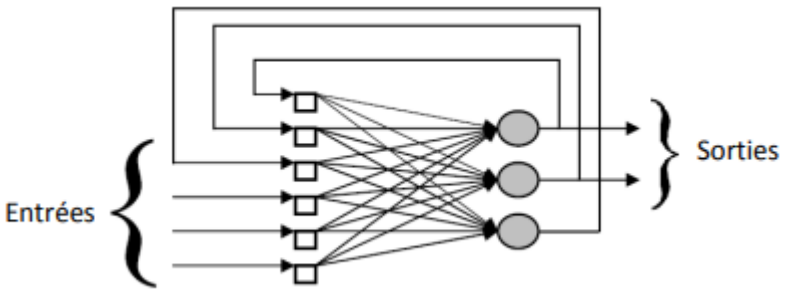

Figure 0-6 Exemple d'un réseau récursif [16]

2.2.4. Le modèle BERT ;

BERT (Bidirectional Encoder Representations from Transformers) est un modèle de langage pré-entraîné qui aide à comprendre le sens des mots dans une phrase en analysant leur contexte. Il a été développé par Google et est largement utilisé dans le traitement du langage naturel. BERT utilise une architecture spéciale appelée transformer, qui lui permet de capturer les informations à la fois avant et après chaque mot, ce qui lui donne une bonne compréhension du texte. Grâce à son entraînement sur de grandes quantités de données, BERT peut fournir des représentations de mots riches en informations, ce qui est utile pour des tâches comme la classification de texte, la génération de texte et les questions-réponses.

2.2.4.1. Architecture de BERT :

BERT de base est constitué de 12 couches d'attention. Chaque couche d'attention permet au modèle de comprendre les relations entre les mots à différentes distances dans la séquence. Chaque couche est composée de plusieurs têtes d'attention, qui se concentrent sur différentes parties de la séquence lors de la représentation des mots. Les couches d'attention successives améliorent progressivement la compréhension du modèle et extraient des représentations de mots de plus en plus informatives. Il existe des variantes de BERT avec un nombre différent de couches, comme BERT-large avec 24 couches, qui offrent des capacités de représentation plus puissantes mais nécessitent plus de ressources de calcul. [19]

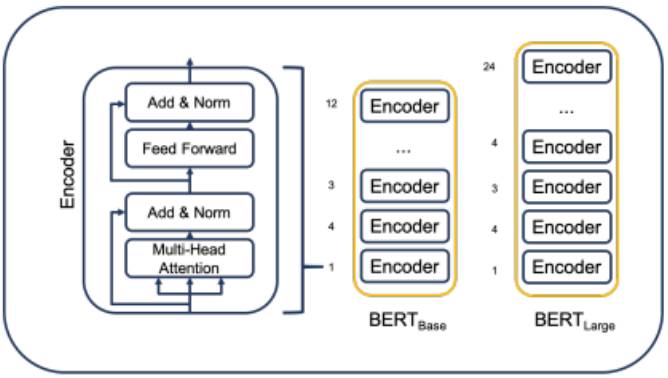

Figure 0-7 Architecture de BERT[20]

2.2.4.2. Pré-entraînement de BERT

BERT se différencie de ses prédécesseurs (modèles de NLP pré-entraînés), par la façon dont il est pré-entraîné. Ce pré-entraînement est non-supervisé c'est-à-dire qu'il ne nécessite pas de jeu de données labellisé. BERT est pré-entraîné sur un grand jeu de données constitué de textes des pages Wikipédia en anglais (2 500 millions de mots) ainsi qu'un ensemble de livres (800 millions de mots).

Ce pré-entraînement est fait sur deux tâches :

1. **Masked Language Modeling (MLM) :** Le Language Modeling est une tâche habituelle de NLP qui consiste à prédire le mot suivant étant donnée le début de la phrase.

   Exemple :

- Séquence initiale : « Le lion ne s'associe pas avec le cafard »
- Séquence donnée : « Le lion ne s'associe pas avec le »
- Prédiction du modèle de LM : « cafard »

2. **Next Sentence Prediction (NSP) :** Cette tâche consiste à prédire si oui ou non, une certaine séquence A est suivie par une certaine séquence B. L'entraînement pour cette tâche se fait avec deux séquences à chaque itération et dans 50% des cas la phrase A est vraiment suivie par B.

   Exemple :

- Séquence A : Il va pleuvoir.
- Séquence B : Je prends mon parapluie.
- Prédiction : IsNext.

2.3. Travaux connexes

2.3.1. YouTaQA

YouTaQA est un système de questions-réponses intelligent basé sur le Deep Learning et la recherche d'information, Il est conçu pour répondre aux questions des utilisateurs dans

différents domaines, Il s'appuie sur la base de connaissances de Wikipédia et est entraîné sur l'ensemble de données SQuAD (Stanford Question Answering Dataset), qui est une collection de questions-réponses basées sur des extraits de texte provenant de Wikipédia.

Il utilise un moteur de recherche d'Information (MRI) pour trouver les informations pertinentes et générer des réponses appropriées à partir de la base de connaissances de Wikipédia. [21]

De plus, le système intègre des modules d'apprentissage approfondi, implémentés à l'aide du modèle pré-entraîné BERT. Ce modèle BERT améliore la compréhension du langage naturel et l'analyse des textes.La figure suivante présente l'application web « YouTaQA » :

Figure 0-8 YouTaQA[21]

2.3.2. QARAB

QARAB est un système qui traite des questions exprimées en langue arabe et essaie de fournir des réponses courtes. Le système a pour principale source de connaissance une collection de journaux arabes extraits d'Al-Raya2, un journal publié au Qatar. QARAB ne procède pas à une analyse sémantique de la question. [22]

Il utilise une méthode de recherche d'informations basée sur "sac de mots" pour construire la requête et interroger le système IR. Le document le mieux classé qui correspond étroitement à la requête est ensuite utilisé pour extraire la réponse appropriée. Comme indiqué dans les figures suivantes :

Token		Stem	Part of Speech	Stop Word
هو	he	هو	Pronoun	Yes
محافظ	governor	محافظ	Noun	
البنك	bank	بنك	Noun	
المركزي	central	مركز	Noun	
الكويتي	Kuwaiti	كويت	Noun	
و	and	و	Conjunction	Yes
الذي	that	الذي	Pronoun	Yes
قال	said	قال	Verb	
بأن	that	بأن	Particle	Yes
بلاده	his country	بلاد	Noun	
ليس	not	ليس	Verb	Yes

لديها	have	لدى	Particle	Yes
النية	intention	نية	Noun	
لخفض	devaluation	خفض	Noun	
قيمة	value	قيمة	Noun	
الدينار	dinar	دينار	Noun	
للحد	restriction	حد	Noun	
من	from	من	Preposition	Yes
عجز	inability	عجز	Noun	
الميزانية	budget	ميزانية	Noun	
؟	؟	؟	Punctuation	Yes

Figure 0-9 Traitement des requêtes[22]

محافظ
بنك
مركز
كويت
بلاد
نية
خفض
قيمة
دينار
حد
عجز
ميزانية

Figure 0-10 Sac de mots[22]

2.3.3. ArabiQA

Le système ArabiQA est utilisé pour répondre à des questions en arabe dans différents domaines. Il repose sur deux parties principales : un module d'extraction de texte et un système de reconnaissance d'entités nommées (NER) [23]. Le module d'extraction de texte collecte et prépare les informations provenant de différentes sources, comme des documents ou des articles en ligne. Le système de reconnaissance d'entités nommées identifie les noms de personnes, de lieux, de dates, d'organisations, etc., dans le texte. Ces éléments aident à comprendre le contexte et à trouver les informations nécessaires pour répondre de manière précise aux questions posées.

2.4. Etude comparative des systèmes de question-réponse en arabe

Système	Domaine	Langage D'implémentation	Source	Réponse	Approche
QARAB	Ouvert	Non mentionné	Des données nonstructurées (corpus du journal Al-Raya)	Passage court	Traite la questioncomme un "sac demots". Le module de recherche d'information est basé sur le modèle d'espace vectorielde Salton. Reconnaît les entités nommées
AQAS	Fermé	Non mentionné	Des données structurées	Phrase	Utilise un modèle basé sur la Connaissance. Recherche dans des basesde données structurées
ArabiQA	Ouvert	Java	Corpus	Phrase	Catégorise la question on(nom, date, laquantité, et ladéfinition) selon les mots d'interrogationet attribue unrang plusélevé pour lespassages qui ont une plus petite distance entre les mots clés : densité de la distance.

JAWEB	Ouvert	Java	Corpus	Phrase	Jaweb analyse les questions et extrait les informations importantes pour récupérer les réponses lesplus pertinentes à partir d'un corpus arabe. Il fournit une interface d'utilisateur.
Al-Bayan	Fermé : Coran	Non mentionné	Coran et ses livres d'interprétation (tafsir)	Phrase	Al-Bayan Comprend la sémantique du Coran et répond aux questions des utilisateurs en utilisant le Coran et ses livres d'interprétation (tafsir).

Tableau 0-2 Etude comparative des systèmes de question-réponse en arabe

2.5. Avantages et limites des travaux connexes

Système	Avantages	Limites
QARAB	QARAB se base sur une approche favorisant une liaison entre un système de recherche d'information et un système de TALN qui réalise l'analyse linguistique.	Le système fournit des réponses à des questions factuelles, mais ne supporte pasles autres types de questions.
AQAS	AQAS est considéré le premier prototype pour a question-réponse arabe	L'analyse morphologique utilise un dictionnaire de taille limitée.

ArabiQA	L'approche proposée se base sur une liaison entre un système de recherche d'information et un système de TALN qui réalise L'analyse linguistique	L'implémentation du système n'a pas été achevée.
JAWEB	Une approche à base du Web qui fournit une interface pour l'utilisateur.	Le système ne fonctionne que pour certains types de questions (questions factuelles), mais ne supporte pas les autres types de questions.
Al-Bayan	L'approche présentée construit un modèle de recherche d'information sémantique	Les auteurs ont révisés manuellement les1200 concepts et leurs versets.

Tableau 0-3 Avantages et limites des travaux connexes

2.6. Description sommaire de notre système

Après avoir étudié les travaux connexes, on a essayé de trouver et sélectionner que les points jugés très bénéfiques dans chaque système afin de les rassembler dans le nôtre avec l'ajout de plusieurs nouvelles idées.

Notre système est basé sur la création d'un site web de système question-réponse pour la langue arabe plus précisément sur les questions du Coran, il est basé sur les principes du Deep Learning. Ce site nommé « Sora-QA ».

Parmi les objectifs de notre système :

- Fournir des réponses précises et fiables aux questions posées sur le Coran en langue arabe.
- Faciliter l'accès à la connaissance religieuse.
- Promouvoir la compréhension et la diffusion du savoir religieux.

2.7. Conclusion

Dans ce chapitre, nous avons abordé le concept d'apprentissage profond. Et après nous avons présenté quelques systèmes basés sur l'idée de question et réponse, leurs fonctionnements et leurs objectifs. Après l'analyse de ces systèmes, nous avons préparé une description sommaire

pour notre propre système et nous avons identifié ces principaux objectifs. Dans le chapitre suivant, nous allons faire une description détaillée de notre système « Sora-QA ».

Chapitre 03 :

Développement & Implémentation

3.1. Introduction

Dans ce chapitre nous allons décrire le processus de réalisation de notre système Question-Réponse « SORA-QA ». Ceci en spécifiant l'environnement de développement, l'implémentation de la base de données et un aperçu sur les différentes interfaces de notre système.

3.2. Outils de développement

Afin de développer des applications sous Android, un ensemble d'outils est nécessaire. Les différents outils matériels et logiciels utilisés pour la réalisation de notre application sont présentés dans ce qui suit :

3.2.1. Environnement Matériel

Pour réaliser notre projet, nous avons utilisé deux PC marque Lenovo et Hp,

- Processeur: Intel i3
- Ram: 8 GO
- Disque Dure: 500 GO & 250 GO SSD
- Système d'exploitation : Windows 10 64 Bit

3.2.2. Environnement Logiciel

3.2.2.1. Le langage python :

Pour atteindre notre but, nous avons utilisé le langage de programmation Python, version 3.10. Python est relativement simple à prendre, open source, gratuit, interprété et le langage le plus employé par les informaticiens récemment. Il a été créé par Guido van Rossum et sa première version a été publiée en 1991.

Pour se focaliser sur notre projet et tirer profit des puissances du langage Python, nous avons utilisé les outils suivants :

Nous avons utilisé Visual studio code comme un éditeur et de divers packages comme :

- pandas : Utilisée pour la manipulation et l'analyse des données.

- numpy : Bibliothèque fondamentale pour le calcul numérique en Python.

- matplotlib : Bibliothèque de visualisation de données en Python.

- sklearn : Utilisé pour les tâches d'apprentissage automatique, y compris la classification et l'évaluation.

- nltk : Utilisé pour les tâches de traitement du langage naturel, telles que la tokenisation et la suppression des mots vides.

- PyTorch : Est conçu pour être efficace et flexible, offrant des outils puissants pour la création de modèles d'apprentissage automatique, l'optimisation des réseaux neuronaux et la manipulation des données. Il est largement utilisé dans le domaine de la recherche en intelligence artificielle, ainsi que dans l'industrie pour le développement de modèles d'apprentissage profond.

- Transformers : est une bibliothèque open source publiée par HuggingFace , elle fournit des API et des outils qui permettent de télécharger et d'utiliser les modèles pré-entraînés de Traitement Automatique de Langage (NLP).

La bibliothèque Transformers prend en charge l'interopérabilité des Frameworks entre PyTorch, et TensorFlow , ce qui permet d'offrir une flexibilité dans l'utilisation de plusieurs Framework dans l'entraînement d'un modèle.

3.2.2.2. Le langage HTML

Le langage HTML (Hypertext Markup Language) est le langage de balisage utilisé pour structurer et présenter le contenu des pages web. Il est composé d'éléments HTML qui permettent de décrire la structure logique d'une page web en utilisant des balises. Les balises HTML définissent les différents éléments tels que les titres, les paragraphes, les liens, les images, les tableaux, les formulaires, etc.

3.2.2.3. Le langage CSS

Le langage CSS (Cascading Style Sheets) est utilisé pour définir la présentation et l'apparence des éléments HTML sur une page web. CSS permet de spécifier les styles, les

couleurs, les polices, les marges, les alignements, les effets visuels et autres propriétés visuelles des éléments HTML. En séparant la structure (HTML) de la présentation (CSS), CSS permet de rendre les pages web plus flexibles, cohérentes et faciles à gérer.

1.3 La conception générale du système :

Figure 0-11 La conception générale du système

3.3. La conception détaillée du système

Notre système « Sora-QA » soit basé sur 2 méthodes la première est une méthode d'apprentissage de similarité « TF-IDF » et le deuxième algorithme prée-entraine des réseaux neuronaux « BERT ».

Les étapes qu'on a suivies pour le développer sont les suivantes :

1- La collection des données

2- Le prétraitement des données

3- La vectorisation des donnes

4- Le développement de système

5- Le déploiement de système

3.3.1. La collection des données

La première étape consiste en fait à collecter les données nécessaires pour alimenter les modèles d'apprentissage. Dans notre cas les données sont déjà collectées et sauvegardées par les anciens étudiants dans un fichier CSV qui s'appelle " AAQQAC.csv". Cet ensemble de données compris 1224 Question et réponse. La Figure représente la base de données collectée.

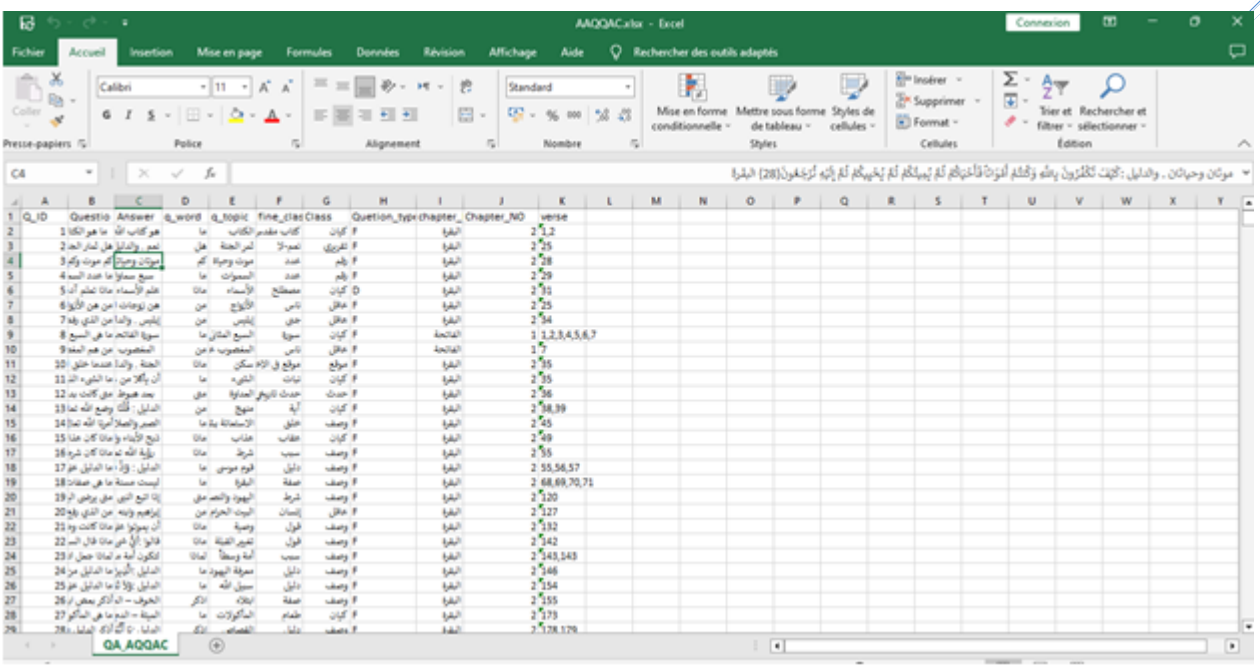

Figure 0-12 La collection des données

3.3.2. Le prétraitement des données :

Le prétraitement des données est une technique d'exploration de données utilisée pour transformer les données brutes dans un format utile et efficace.

Les étapes impliquées dans le prétraitement des données :

1- **Le nettoyage des données :** le nettoyage des données consiste de supprimer et éliminer les hashtags, les ponctuations et tout que ne pas nécessaires dans cette dataset.

Tableau 0-4 Eliminer les ponctuations

Avant	Après
من الذي رفع قواعد البيت الحرام (الكعبة المشرفة) ؟	من الذي رفع قواعد البيت الحرام الكعبة المشرفة

2- **La tokenisation** : également appelée segmentation Cette étape diviser une chaîne de caractères en mots appelés « Tokens ».

Tableau 0-5 La tokenisation

Avant	Après
ما هو الكتاب الوحيد الذي لا يوجد أي ريب فيه ؟	'_ما', '_هو', '_الكتاب', '_الوحيد', '_الذي', '_لا', '_يوجد', '_أي', '_ريب', '_فيه',

3- **Suppressions des mots d'arrêt** : les mots d'arrêt sont les mots que n'ont aucune signification dans le corpus, sont des mots très fréquents tel que "على", "في", "و", "ال" et d'autres prépositions, conjonctions et articles.

Tableau 0-6 Suppressions des mots d'arrêt

Avant	Après
ما هو الكتاب الوحيد الذي لا يوجد أي ريب فيه ؟	['الكتاب' 'الوحيد' 'يوجد' 'ريب']

4- **Normalisation :** normalisation des données est une étape nécessaire dans le prétraitement des données, permet de réduire la complexité des modèles et facilite l'utilisation et la classification des données. Dans ce cas on a premièrement normalisé les caractères arabes que sans le même caractère mais ont une déférente écriture.

Tableau 0-7 Normalisation

Avant	Après
[ا أ آ إ]	[ا]

5- **Stemming et Lemmatisation :** Le Stemming d'un mot est de suprimmer les préfixes et suffixes et garde juste la racine des mots, Il permet de regrouper les mots ayant la même racine même s'ils sont écrits de différentes manière.

La Lemmatisation d'un mot est de retourner le à sa forme canonique, appelée lemme. Contrairement au stemming, la lemmatisation tient compte du contexte et de la signification des mots.

Tableau 0-8 Stemming et Lemmatisation

Avant	Après
ما هو الكتاب الوحيد الذي لا يوجد أي ريب فيه ؟	كتب, وحد, وجد, ريب.

6- La division d'ensemble de données : On a Divisé les données en ensembles d'entraînement et de test. L'ensemble d'entraînement est utilisé pour entraîner le modèle (75%), tandis que l'ensemble de test est utilisé pour évaluer ses performances (25%).

3.3.3. La vectorization des données

La vectorisation des données est une étape cruciale dans le traitement du langage naturel (NLP) pour représenter le texte sous une forme numérique, compréhensible par les algorithmes d'apprentissage automatique. Deux méthodes populaires de vectorisation des données sont Count Vectorizer et Tfi-df Vectorizer.

CountVectorizer convertit un corpus de documents en une représentation matricielle où chaque ligne correspond à un document et chaque colonne correspond à un terme unique extrait de tous les documents. Les valeurs de la matrice indiquent le nombre d'occurrences de chaque terme dans chaque document. Ainsi, chaque document est représenté par un vecteur de fréquences de termes.

Tf-idf Vectorizer (Term Frequency-Inverse Document Frequency) est une autre méthode de vectorisation des données qui prend en compte à la fois la fréquence des termes dans un document spécifique (fréquence des termes) et leur importance globale dans le corpus (fréquence inverse du document). TF-IDF attribue des poids plus élevés aux termes qui apparaissent fréquemment dans un document spécifique tout en étant rares dans l'ensemble du corpus. Par conséquent, les termes les plus importants sont mis en évidence.

Par exemple:

On a la question [ما هو الكتاب الذي لا شك فيه]

```
Matrice des comptages:
[[1 1 1 1 1]]
Matrice TF-IDF:
[[0.4472136  0.4472136  0.4472136  0.4472136  0.4472136 ]]
Vocabulaire:
['الذي', 'الكتاب', 'فيه', 'لا', 'ما', 'شك']
```

3.3.4. Entrainement des modèles

L'entraînement du système de question-réponse est une étape cruciale pour développer un modèle performant. Pendant cette phase, le modèle est exposé à un ensemble de données d'entraînement contenant des paires de questions et de réponses. Le but est d'apprendre les relations et les schémas entre les questions posées et les réponses appropriées.

3.3.5. Le développement de système

Dans notre système de questions-réponses en arabe, nous avons utilisé à la fois la méthode TF-IDF et le modèle BERT pré-entraîné.

La méthode TF-IDF (Term Frequency-Inverse Document Frequency) est une technique utilisée pour évaluer l'importance d'un terme dans un document au sein d'un corpus de documents plus large. Elle se base sur deux composantes principales : la fréquence du terme dans le document (TF) et son taux d'apparition dans l'ensemble des documents (IDF).

Fréquence du terme dans le document (TF) : La fréquence du terme dans le document est une mesure de combien de fois un terme spécifique apparaît dans un document donné. Elle est souvent calculée en comptant simplement le nombre d'occurrences du terme dans le document. Un terme qui apparaît fréquemment dans un document est susceptible d'être important pour ce document en particulier.

Taux d'apparition dans l'ensemble des documents (IDF) : Le taux d'apparition dans l'ensemble des documents est une mesure de la rareté d'un terme dans l'ensemble du corpus de documents. Il est calculé en prenant le logarithme inverse de la proportion du nombre total de documents sur le nombre de documents contenant le terme donné. Un terme qui est rare et n'apparaît que dans quelques documents est considéré comme plus important car il peut aider à distinguer ces documents des autres.

Le modèle BERT (Bidirectional Encoder Representations from Transformers) est un modèle d'apprentissage profond largement utilisé dans le domaine du traitement du langage naturel (NLP). Il a été introduit par Google en 2018 et est basé sur l'architecture des Transformers.

L'architecture des Transformers est conçue pour capturer les relations à longue distance entre les mots dans une phrase ou un texte. Contrairement aux modèles de langage précédents, tels que les réseaux de neurones récurrents (RNN) ou les réseaux de neurones convolutionels (CNN), BERT utilise une approche bidirectionnelle pour traiter le contexte.

Après cette étude et test de notre système, nous avons constaté que la méthode TF-IDF ne fournit pas les bonnes réponses lors des tests, mais donne plusieurs réponses possibles en fonction des scores obtenus.

D'un autre côté, nous avons constaté que le modèle BERT nécessite à la fois la question et le contexte pour fournir une réponse précise.

Pour améliorer les performances de notre système, nous avons eu une idée intéressante : utiliser les résultats de la méthode TF-IDF comme contexte pour le modèle BERT. Cela signifie que nous pouvons utiliser les réponses possibles générées par le TF-IDF comme entrées contextuelles du modèle BERT, ce qui peut aider le modèle à fournir une réponse plus précise et pertinente.

Processus de notre système question-réponse :

- Lorsque l'utilisateur pose une question, nous utilisons la méthode TF-IDF pour rechercher les questions similaires dans votre ensemble de données. Les questions correspondantes avec les scores TF-IDF les plus élevés sont récupérées.
- Nous utilisons ces questions similaires comme contexte pour le modèle BERT. Nous alimentons le modèle avec la question posée par l'utilisateur et chaque question similaire en tant que contexte individuel.
- Le modèle BERT générera une réponse en se basant sur le contexte fourni et sa compréhension globale du langage. La réponse finale est alors renvoyée à l'utilisateur.

3.4. Interface du développement

Pour rendre notre travail apercevable, nous avons développé une interface dans Visual Studio Code pour fournir un environnement générateur de questions, cet outil contient deux rubriques (Question, Réponse)

Dans ce qui suit, nous illustrons des captures d'écran de notre interface.

3.4.1. La fenêtre «الصفحة الرئيسية »

La figure suivante présente le premier lancement de notre système, la première fenêtre qui s'affiche est la fenêtre « الصفحة الرئيسية » suivante :

Figure 0-13 La fenêtre «*الصفحة الرئيسية* »

3.4.2. La fenêtre « Fenêtre de présentation des résultats »

Dans cette fenêtre, l'utilisateur saisi sa question, puis le modèle BERT fournit la réponse correcte, Comme indiqué dans la figure suivante.

Figure 0-14 Fenêtre de présentation des résultats

3.5. Resultat et interprétation :

Le tableau 3-6 présente les métriques de performance pour différents classificateurs : le classificateur de vecteur de support linéaire, la régression logistique, le naïf bayes multinomiaux et le classificateur de forêt aléatoire.

Tableau 3 6 Le résultat d'exactitude des classificateurs avec le TF-IDF

Classificateur	Métrique				
	Précision	Rappel	Score F1	Exactitude	Le taux d'erreur
Classificateur de vecteur de support linéaire	0.90	0.93	0.99	0.98	2%
Régression logistique	0.89	0.92	0.94	0.78	22%
Naïf bayes multinomiaux	0.87	0.95	0.91	0.78	22%
Classificateur de forêt aléatoire	0.62	0.97	0.76	0.63	37%

Précision : Cela mesure la proportion de résultats positifs correctement prédits par le modèle. On peut observer que le classificateur de vecteur de support linéaire a la plus haute précision (0,90), suivi de près par la régression logistique (0,89) et le naïf bayes multinomiaux (0,87). Le classificateur de forêt aléatoire a la plus faible précision (0,62), ce qui indique qu'il prédit moins précisément les résultats positifs.

Rappel : Cela mesure la proportion de vrais résultats positifs identifiés par le modèle. Le naïf bayes multinomiaux a le plus haut rappel (0,95), suivi par le classificateur de forêt aléatoire (0,97), le classificateur de vecteur de support linéaire (0,93) et la régression logistique (0,92).

Score F1 : C'est une mesure globale qui combine à la fois la précision et le rappel. Le classificateur de vecteur de support linéaire obtient le score F1 le plus élevé (0,99), suivi de près par le naïf bayes multinomiaux (0,91), la régression logistique (0,94) et le classificateur de forêt aléatoire (0,76).

Exactitude : Cela mesure la proportion de prédictions correctes par rapport à l'ensemble des prédictions. Le classificateur de vecteur de support linéaire a la plus haute exactitude (0,98), suivi de près par la régression logistique (0,78) et le naïf bayes multinomiaux (0,78). Le classificateur de forêt aléatoire a la plus faible exactitude (0,63).

Taux d'erreur : Il est calculé en soustrayant l'exactitude de 1. On peut observer que le classificateur de vecteur de support linéaire a le taux d'erreur le plus faible (2%), suivi par la régression logistique (22%) et le naïf bayes multinomiaux (22%). Le classificateur de forêt aléatoire a le taux d'erreur le plus élevé (37%).

3.5 Conclusion

Dans ce chapitre, nous avons présenté les différentes étapes que nous avons suivies pour développer et faire fonctionner notre système de question-réponse. Nous avons également présenté un ensemble d'interfaces qui font partie de notre site web, avec interprétations et d'explications des résultats obtenus.

Conclusion générale

Conclusion générale :

La recherche dans le domaine des systèmes de questions-réponses en langue arabe est devenue une nécessité urgente de nos jours. Cela est dû à la demande croissante d'avoir des méthodes efficaces et innovantes pour interagir avec les textes et fournir des réponses rapides et précises aux questions posées par les utilisateurs.

Dans le contexte des textes religieux, nous avons besoin de systèmes de question-réponse efficaces dans la langue naturelle. En effet, la communauté religieuse s'attend à disposer d'outils technologiques avancés pour répondre à leurs besoins religieux et fournir des réponses correctes et fiables.

Ce travail a été relativement réussi dans la construction d'un système qui répond automatiquement aux questions religieuses, grâce à l'utilisation des techniques modernes d'intelligence artificielle et d'analyse du langage naturel, permettant ainsi de fournir des réponses rapides et précises aux questions religieuses.

En ce qui concerne les difficultés que nous avons rencontrées et qui ont entravé notre progression, la première était l'incapacité de trouver une base de données appropriée en langue arabe à utiliser. Nous avons fait de nombreux efforts pour obtenir une base de données plus vaste et plus complète sur des sujets religieux, mais malheureusement, elle n'était pas disponible et nous n'avions pas suffisamment de temps pour collecter un plus grand nombre de données. Cela a été le premier obstacle. Le deuxième obstacle était le manque de temps. Cependant, à l'avenir, nous essaierons d'améliorer l'application en créant une base de données plus exhaustive dans d'autres domaines des sciences religieuses.

Cependant, il est important de noter que ce système automatisé ne remplace pas les experts religieux, car ils possèdent une expérience suffisante et peuvent traiter des questions complexes ou ayant des contextes spécifiques nécessitant une compréhension plus approfondie et la fourniture de conseils personnalisés.

Références :

[1] W. BAKARI : Une approche vers la compréhension automatique des textes arabes destinée pour les systèmes de question-réponse. *Thèse de doctorat*, 2018.

[2] B. ABDELGHANI : Exploitation des données liées : système questionréponse. *these de doctorat*, 2019.

[3] W. BAKARI : Une approche vers la compréhension automatique des textes arabes destinée pour les systèmes de question-réponse. *Thèse de doctorat*, 2018.

[4] M. M. F. KOLOMIYETS, O. : A survey on question answering technology from an information retrieval perspective. *Information Sciences*, p. 5412–5434, 2011.

[5] . M. M. F. KOLOMIYETS, O. : A survey on question answering technology from an information retrieval perspective. *Information Sciences*, p. 5412–5434, 2011.

[6] W. BAKARI : Une approche vers la compréhension automatique des textes arabes destinée pour les systèmes de question-réponse. *Thèse de doctorat*, 2018.

[7] Thevenin, D. (2001). Adaptation en Interaction Homme-Machine: le cas de la Plasticité (Doctoral dissertation, Université Joseph-Fourier-Grenoble I).

[8] Sadin, É. (2018). L'intelligence artificielle. L'échappée, Paris.

[9] Gouillart, E. (2007). Etude de l'advection chaotique dans des mélangeurs à tiges, en écoulements ouverts et fermés (Doctoral dissertation, Paris 6).

[10] Pageaud, S. (2019). SmartGov: architecture générique pour la co-construction de politiques urbaines basée sur l'apprentissage par renforcement multi-agent (Doctoral dissertation, Université de Lyon).

[11] Ligozat, A. L. (2006). Exploitation et fusion de connaissances locales pour la recherche dřinformations précises (Doctoral dissertation, Ph. D. thesis, Université Paris-Sud 11, Orsay, France).

[12] Deep learning ou apprentissage profond : définition, concept. URL https://www.lebigdata.fr/deep-learning-definition

[13] Deep learning ou apprentissage profond : définition, concept. URL https://www.lebigdata.fr/deep-learning-definition

[14] B. SILVA, C. et Ribeiro : Inductive inference for large scale text

classification : Kernel approaches and techniques. 255, 2009.

[15] Deep learning ou apprentissage profond : définition, concept. URL https://www. lebigdata.fr/deep-learning-definition

[16] Y. DJERIRI : Les réseaux de neurones artificiels. *Journal of Theoretical and Applied Information Technology*, 2017.

[17] Y. DJERIRI : Les réseaux de neurones artificiels. *Journal of Theoretical and Applied Information Technology*, 2017.

[18] Y. DJERIRI : Les réseaux de neurones artificiels. *Journal of Theoretical and Applied Information Technology*, 2017.

[19] J. Devlin, M.-W. Chang, K. Lee, et K. Toutanova, « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ». arXiv, 24 mai 2019. Consulté le: 3 juillet 2022.

[20] « Fine-tune BERT Model for Sentiment Analysis in Google Colab », *Analytics Vidhya*, 28 décembre 2021. https://www.analyticsvidhya.com/blog/2021/12/fine-tune-bert-model-forsentiment-analysis-in-google-colab/(consulté le 3 juillet 2022).

[21] Younes, A. R. Youtaqa: système de questions-réponses intelligent basé sur le deep learning et la recherche d'information.

[22] Hammo, B., Abuleil, S., Lytinen, S., & Evens, M. (2004). Experimenting with a question answering system for the Arabic language. Computers and the Humanities, 38(4), 397-415.

[23] Benajiba, Y., Rosso, P., & Lyhyaoui, A. (2007, April). Implémentation of the ArabiQA question answering systèmes components. In Proc. Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morroco, April (pp. 3-5)