People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of 20 Août 1955 of Skikda

Faculty Of Sciences

Department Of Informatics

Automatic Extraction of Knowledge From Al-Quran

Presented by **:**

Boucherkha Ali

Specialty **:**

Master2 RSD

Under the Supervision of:

Dr. Mawloud MOSBAH

2023/2024

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude to my supervisor Dr. *Mawloud MOSBAH* for guiding me through a wonderful learning adventure. Our discussions were always fruitful and insightful, your encouragement to work on challenging yet achievable goals helped me grow both personally and academically. In addition, I want to express my gratitude for your invaluable guidance and mentorship in introducing me to the world of scientific writing and research. Your support in co-writing and submitting our research paper was instrumental in helping me gain practical experience in this field. Thank you for sharing your knowledge and expertise with me and for encouraging me to pursue excellence.

I would like to extend my heartfelt thanks to my parents, and to all family
and my friends, who have been a constant source of support throughout this journey. Without their encouragement and understanding, I would not have been able to reach this milestone.

Thank you all, from the bottom of my heart.

Abstract

Automatic natural language processing and text-mining have emerged as scientific fields helping machines to understand the communication language of a human being. Many researches have been conducted for automatic processing of natural languages some are generic no matter what the considered language, based generally on statistics, and some others, qualified as linguistic approaches, are dedicated to a specific language.

Arabic language is considered among the well spread languages over the world where many humans are interested with because either to their Arabic original or because they are Muslims. Unfortunately, Arabic has, up to now, not received the suitable attention owing to may reasons such as: the complexity of the language where there is a rich repertoire of vocabulary with many indicating for the same indicated. Moreover, in terms of derivation and fluctuation where we need to consider affixes like suffixes and pre-fixes, Arabic is among the scarce languages that include infixes revealed for instance in broken plural. Diacritics which replace vowels in latin and Germanic languages give more difficulty to automatically process of Arabic.

Al-Quran, the holly book, the book of God, or the parole of God, is a valuable book regarding the Muslim background and culture. Indeed, from this book, including 114 Surrah (or chapters), Muslims extract their laws and rules to well live this life considered as a preparation for another continuous life coming later where all humans will be directed forever either to hell or to paradise regarding their actual works.

Unfortunately, Al-Quran, coming in Arabic, as its original language, and translated so far to many human being languages, is not completely interpreted by Muslims. Although its principles and laws are staying unchangeable, its comprehension over years needs to be continuously treated by the novel generations without any deviation from its general context range designated by the last prophet on behalf of God.

Our work is a tentative to extract knowledge from Al-Quran through designating some structural rules using some Arabic signal in order to generate these rules. Although it is difficult to process Al-Quran automatically owing to its sensible aspect ‘Parole of God’, we try to do the job carefully considering the De Saussure vision through processing the language

as it is and not as it should be. So, we process Al-Quran not to check its validity but in order to learn from it and may be considered later as a reference to check any text written in Arabic.

To the best of our knowledge, a lot of works published in the literature is dedicated to build prototypes as Question-Answering systems based only on structure and morphology without addressing the semantic level. By this work, we directly address the meaning level through extracting knowledge and rules from Al-Quran. The results achieved are encouraging, as preliminary results, but they need to be well analysed and interpreted in order to open avenues to ask other questions and so on.

الملخص

يعتبر كل من ميدان المعالجة الآلية للغة و معالجة النصوص إضافة نوعية للآلة بهدف فهم اللغة المستخدمة من قبل الإنسان.

لقد أجريت عدة أبحاث في هذا المجال, بعضها لدراسة اللغة بصفة عامة بالإعتماد على بعض البيانات التي تنتمي إلى عالم الإحصاء, والأخرى لدراسة لغة معينة بالإعتماد على خصوصية اللغة محل الدراسة.

تعتبر اللغة العربية من بين اللغات الرائجة حول العالم بإعتبارها مستعملة من قبل العرب و المسلمين الذين يزداد عددهم كل يوم.

للأسف المعالجة الآلية للغة العربية لا ترتقي إلى أهمية هذه اللغة بإعتبارها لغة القرآن و هذا بسبب عدة عوامل متعلقة باللغة نفسها. بالفعل, تعتبر اللغة العربية لغة يصعب التعامل معها بطريقة آلية وذلك لغني ألفاظها إذ نجد عدة ذلالات لنفس المذلول, بالإضافة إلى كثرة إشتقاقاتها إذ لا توجد لغات كثيرة يكون الإشتقاق فيها من داخل الكلمة نفسها و ليس على الجانبين كجمع التكسير مثلا, الشكل (الضمة, الفتحة, السكون, و الشدة) التي تعوض الحروف المتحركة في اللغات الجرمانية و اللاتينية تساهم بقدر كبير في جعل اللغة العربية صعبة المعالجة بالطريقة الآلية.

القرآن, كتاب الله, أو كلام الله, هو كتاب مقدس عند العرب و المسلمين لما يحمله دلالات و رمزية. بالفعل, من هذا الكتاب يستنبط المسلمون الأحكام الشرعية التي تمكنهم من العيش بسلام و طمأنينة كما أمرهم الله عز و جل و كما يريدهم أن يحيوا للدخول إلى الجنة و رؤية وجهه الكريم.

للأسف, حال القرآن الكريم ليس أحسن من حال اللغة العربية من حيث الدراسة الآلية فهناك بعض الأعمال البحثية فقط التي تعني بالقرآن و ليست جميعها أبحاث عرب و مسلمين.

حسب معرفتنا المتواضعة, كل التطبيقات و النماذج المقترحة يمكن إعتبارها أنظمة (سؤال-جواب) تعتمد فقط على الصيغة الشكلية للكلمة دون الغوص في جانب المعنى.

من خلال هذا العمل, الذي يعتبر مذكرة تخرج لنيل شهادة الماستر في الإعلام الآلي, نحاول جاهدين لدراسة القرآن في لغته الأصلية و هي اللغة العربية بغية إستخراج ما يمكن إستخراجه من معارف و قواعد تعطي بعض التعاريف لبعض المصطلحات و تعزز بعض الأحكام الموجودة في الشريعة. النتائج الأولية تعتبر مشجعة لكنها تحتاج إلى تمحيص و مراجعة أكثر بغرض التوسع في الموضوع و الإجابة على أسئلة أخرى متقدمة.

Contents

List of Figures

General Introduction

Natural Language Processing and text mining, appeared recently as two scientific fields helping to understand by machines human being languages, are addressed in this master dissertation in order to discover knowledge from Al Quran. Indeed, Al Quran, with its 114 chapters, maybe considered as a valuable resource to be mined and analysed. For doing so, the machine needs firstly to understand Arabic language. Although the difficulties tied to this human language, we try then to design and build an automatic application for Al Quran processing. Consulting of the literature reveals that the majority of proposed systems dealing with Al Quran are qualified as question-answering systems. For us, our initial idea is to develop an application allowing to extract concepts and concept-definitions from Al Quran. This first purpose was moving to other additional aspects such as designating the common intersection between the definitions of the same concept.

This manuscript is organised as follows: In the first part titled 'state of the art', we have two chapters: the first one 'Natural Language Processing' gives an overview about NLP aspects and basic concepts. The second chapter 'Arabic NLP and Automatic processing of Al-Quran' presents the specificity of Arabic language to be taken into account, we also talk about Al Quran processing and the various proposals from the literature as well as the different operational systems may exist. The second part titled 'Contribution' includes two chapters: 'Analysis Requirement and Design' and 'Implementation'. The first chapter explains how we conceive our prototype while the second one reveals how we develop it.

Some perspectives are given in the last of this dissertation for ulterior possible improvement.

Part 1:

State of the art

Chapter 1:

Natural Language Processing

1. Introduction

Natural language processing (NLP) is a multidisciplinary field that includes linguistics, Computer science and machine learning with the goal of developing computers

Technology that automatically analyzes, understands and generates human language Content[1]. For example, NLP can be used to translate sentences from one language to another On the other hand, we build conversational systems [2] that automatically converse with users, convert speech to text and create summaries .

The vision of NLP is to create machines that can read, understand and integrate A vast amount of human knowledge helps us complete routine or repetitive tasks [3], such as: Provide personalized summaries of books, news articles or academic papers[4],[5]. In the era of big data, the amount of text information available on the Internet is increasing.

As the spread of NLP rapidly increases on the Internet, it becomes increasingly important, and not just in a general sense knowledge, but also in the scientific literature [4].

Early NLP research focused on developing manual rules to transform linguistic units, but this proved challenging due to high variability and ambiguity of human language. This led to the development of statistical methods, which has become increasingly common in recent decades [6]. Sparse lexical features, such as co-occurrence Word window statistics are used as input to the standard planar linear representation Classification algorithms such as support vector machines or logistic regression. this The availability of large amounts of training data is critical to the success of these measures method.

2. The Evolution and Impact of Natural Language Processing

Natural Language Processing (NLP) has undergone significant evolution, from early rule-based systems to modern advanced deep learning models. This transformation has revolutionized human-computer interaction, enabling machines to understand and generate human language with remarkable accuracy. The impact of NLP spans various industries, enhancing efficiency and providing insightful data analysis. Despite facing challenges like language ambiguity and biases, ongoing research aims to develop more generalizable and

ethical models, integrating NLP with other AI technologies to create comprehensive and context-aware systems.

The journey of NLP began in the 1950s with rule-based systems, which relied on manually crafted linguistic rules for tasks such as syntactic parsing and rudimentary translation. A notable milestone was the Georgetown-IBM experiment in 1954, which demonstrated the potential of automatic translation by converting over sixty Russian sentences into English [5]. This era laid the foundation for NLP as a distinct field of study.

The 1980s and 1990s marked a paradigm shift with the introduction of statistical methods in NLP. Researchers began leveraging probabilistic models and statistical techniques to learn language patterns from large text corpora. This period saw the development of algorithms for part-of-speech tagging, named entity recognition, and machine translation, significantly improving the accuracy and efficiency of NLP tasks [5].

The advent of deep learning in the late 2000s revolutionized NLP further. Deep learning models, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), enabled handling sequential data and complex linguistic structures. However, the introduction of the Transformer architecture by Vaswani et al. (2017) truly transformed the field. Transformers, with their self-attention mechanisms, allowed for parallel processing of text sequences, leading to significant advancements in language understanding and generation [6].

The development of Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) by Google and GPT (Generative Pre-trained Transformer) by OpenAI has set new benchmarks in NLP. BERT, introduced in 2018, brought the concept of bidirectional context, allowing models to understand the meaning of a word based on its context from both left and right [7]. GPT-3, released in 2020, took generative modeling to new heights with its 175 billion parameters, enabling it to perform a wide array of language tasks with minimal fine-tuning [8].

NLP has found applications in numerous fields, profoundly transforming how we interact with technology and process information. In communication, chatbots and virtual assistants like Siri, Alexa, and Google Assistant use NLP to provide intuitive user interfaces, understanding and responding to user queries [9]. In healthcare, NLP analyzes clinical notes,

research literature, and patient records, aiding in disease diagnosis, treatment planning, and extracting valuable insights from unstructured data [10]. The legal industry benefits from NLP through automated document analysis, contract review, and legal research, while in finance, it is employed for sentiment analysis, fraud detection, and automated trading [11].

Despite its advancements, NLP faces challenges such as the inherent ambiguity and variability of human language, which can lead to misinterpretations by machines. Ensuring the ethical use of NLP, particularly in avoiding biases that can be encoded in data, is another critical concern. Bias in NLP models can perpetuate stereotypes and result in unfair treatment of certain groups [12].

Future research aims to develop more generalizable and interpretable models that can understand and reason about language in a human-like manner. Ensuring the ethical use of NLP involves creating models that are transparent, fair, and accountable, minimizing biases and promoting inclusivity. The integration of NLP with other AI fields, such as computer vision and robotics, holds promise for creating more comprehensive and context-aware systems [12].

The continuous evolution of NLP highlights its pivotal role in advancing artificial intelligence and its potential to transform how we interact with and utilize technology in our daily lives and professional activities.

3. NLP Tasks

3.1. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is an important task in natural language processing (NLP). The purpose is to determine whether the mood expressed within the text is positive, negative, or neutral. This task is often used to measure public opinion, customer satisfaction, and brand awareness. When monitoring social media, sentiment analysis can help businesses understand public opinion about their products and services. By analyzing tweets, posts, and comments, companies can gain insight into customer opinions and proactively address issues [13]. In finance, sentiment analysis is used to analyze news articles, blogs, and

social media to predict market trends and help investors make informed decisions[14].

Various linguistic methods are used in sentiment analysis. One common approach is the dictionary-based approach, which is based on a predefined list of words with known sentiment values. This approach calculates the sentiment score of the text based on the number of occurrences and sentiment strength of these words. Another approach is the machine learning-based approach, where a labeled dataset is used to train classifiers such as support vector machines (SVM), naive Bayes, or deep learning models such as recurrent neural networks (RNN) and convolutional neural networks (CNN) to predict sentiment [50].

There are a variety of systems and tools that use these linguistic approaches to perform sentiment analysis. For example, VADER (Valence-Aware Dictionary and sEntiment Reasoner) is a lexicon-based tool for sentiment analysis on social media. It is well suited for handling informal language commonly found on social platforms[51]. On the other hand, BERT (Bidirectional Encoder Representations from Transformers) is a Transformer-based model that has been optimized for sentiment analysis tasks and has shown state-of-the-art performance on various benchmarks[7].

When it comes to customer feedback, tools such as Sentiment140 analyze tweets to determine customer sentiment towards a product or service. Financial companies use the Thomson Reuters MarketPsych Index to analyze news and social media sentiment to predict stock market movements. These systems integrate lexicon-based and machine learning methods to provide comprehensive sentiment analysis solutions[52].

3.2. Machine Translation

Machine translation (MT) is the task of automatically converting text from one language to another. This task has made significant progress with the development of neural machine translation (NMT) models, especially those based on the Transformer architecture [6]. NMT models like Google's Translate service use large-scale parallel corpora to learn the mappings between languages, resulting in translations that are more accurate and fluent compared to earlier methods. Machine translation is crucial in breaking down language barriers, enabling global communication and access to information, and is widely used in applications from

translating web pages and documents to facilitating multilingual customer support.

The dominant linguistic approach to modern MT is to use neural networks, specifically the Transformer model. This architecture uses a self-attention mechanism to handle dependencies between words in a sentence, allowing for better contextual understanding and, therefore, more accurate translations[53]. Another approach, although more traditional, is phrase-based translation, which breaks sentences into phrases and translates each phrase independently before reassembling them into the target language. Rule-based translation systems based on linguistic rules and dictionaries were once the standard approach, although they are less common today.

Several systems and tools have effectively implemented these approaches. Google Translate is one of the most widely used machine translation services, providing real-time translation for numerous language pairs using advanced NMT models[54].

Machine translation is essential to breaking down language barriers and enabling global communication and access to information. It is used in many different applications, such as translating web pages and documents, enabling multilingual customer support, and real-time communication through translation applications.

3.3. Named Entity Recognition (NER)

Named Entity Recognition (NER) is the task of identifying proper nouns in text and classifying them into predefined categories (e.g., names of people, organizations, locations, dates, etc.). NER is essential for extracting structured information from unstructured text, making it useful in numerous applications. In information retrieval, NER helps improve search engines by identifying the key entities in a document so that it is better indexed and, therefore, gives more accurate and relevant results. For example, if a search engine recognizes the names of people, organizations, or locations in documents, then content can be more precisely indexed, producing more relevant responses to user queries[15]. In the legal and financial industries, NER is used to extract important details from contracts, agreements, and reports, facilitating faster and more accurate document analysis [11].

Early named entity recognition methods relied heavily on rule-based systems and hand-

crafted lexicons. These methods provide high accuracy but require a lot of manual work and are not scalable. For example, hand-crafted grammar rules have been used to identify and classify entities based on language patterns and context[15].

Machine learning methods, especially those based on sequence labeling models such as conditional random fields (CRFs) and hidden Markov models (HMMs), have been widely used for NER tasks. These models learn from annotated data to identify entities and their categories in text. For example, CRFs have been shown to effectively model the sequential nature of text and capture dependencies between labels[55].

3.4. Question Answering

Question answering (QA) systems are designed to automatically respond to questions posed in natural language. These systems can be open-domain, answering questions on a broad range of topics, or closed-domain, focusing on a specific area, such as customer support or medical information. Modern QA systems leverage large-scale pre-trained language models such as BERT and GPT to understand context and generate accurate answers [7]. These systems are used in virtual assistants and chatbots, enhancing their ability to provide relevant and accurate information to users. In educational settings, QA systems help students by providing quick answers to their queries, supporting their learning process.

Traditional QA systems typically rely on information retrieval (IR) techniques to find the most relevant documents that may contain the answer [17]. The system then extracts possible answers from these documents and ranks them. This approach involves multiple steps, including query processing, document retrieval, and response extraction.

Knowledge-based QA systems use structured data from knowledge bases (such as Wikidata or DBpedia) to generate answers [56]. These systems map user queries to entities and relations in the knowledge base, enabling precise and factual answers. For example, a query about the population of a city can be answered by retrieving relevant data from the knowledge base.

Recent advances in neural network models, especially those using Transformers, have revolutionized QA systems [6]. Models such as BERT and GPT-3 are trained on large datasets and can understand query context better than traditional methods. These models use deep

learning techniques to generate coherent and contextually relevant answers. For example, BERT uses a bidirectional attention mechanism to capture context from both directions, resulting in more accurate understanding and response generation.

3.5. Text Summarization

Text summarization is the task of compressing a long text into a shorter version while retaining the necessary information. This can be achieved through extractive methods, which select key sentences from the original text, or abstractive methods, which generate new sentences that convey the main ideas [16]. Text summarization is particularly useful for professionals who need to understand information but do not have enough time to read long reports or articles. For example, news aggregators use summarization to provide concise news briefs, allowing readers to quickly grasp the main points.

3.6. Information Retrieval

Information retrieval (IR) is the task of retrieving relevant information from a collection of documents or data sources in response to a user query. NLP techniques are employed to improve the effectiveness of information retrieval systems by analyzing and understanding the content of documents. This includes techniques such as keyword extraction, which identifies important words or phrases in a document to represent its content and facilitate indexing and searching [17]. Information retrieval systems powered by NLP enable users to quickly access and retrieve relevant information from large document collections, thereby improving productivity and decision-making.

3.7. Cross-language Information Retrieval

Cross-language information retrieval (CLIR) is a key area of information retrieval that focuses on retrieving information written in different languages from a query language. The field is particularly important in the context of global information access, where users seek information from different language backgrounds.

CLIR systems allow users to query in one language and retrieve relevant documents in another. This skill is critical in a multilingual world, as valuable information is often spread across multiple languages. For example, an Arabic-speaking researcher may be interested in retrieving scientific articles written in German or Chinese. CLIR bridges this language barrier and provides a mechanism to access information that would otherwise be inaccessible [18].

3.8. Keyword Extraction

Keyword extraction is the task of identifying important words or phrases in a document that represent its content and main topic. This task is crucial for summarization, indexing, and categorization of documents. NLP techniques are used to analyze the frequency, location, and context of words in a document to determine their importance and relevance. Keyword extraction helps identify key concepts and topics in documents, allowing users to quickly grasp the main ideas and extract valuable insights [19]. In information retrieval and document analysis, keyword extraction helps with indexing and searching, and improves the efficiency and accuracy of information retrieval systems.

4. Language Models - Probabilistic Language

Language models (LMs) are fundamental to natural language processing as they provide a framework for understanding and generating human language. Probabilistic language modeling involves estimating the likelihood of a sequence of words, enabling a variety of NLP tasks such as speech recognition, machine translation, and text generation. These models assign probabilities to sequences of words based on how likely they are in a given context, learned from a large corpus of text. By capturing the statistical properties of language, language models can predict the next word in a sentence, correct grammatical errors, and even generate coherent and contextually appropriate text.

Probabilistic language models function by calculating the joint probability of a sequence of words. For example, given a sequence of words $w_1, w_2, \ldots, w_n$,the model computes the probability$P(w_1, w_2, \ldots, w_n)$.This probability can be decomposed using the chain rule of probability into

$$P(w_1).P(w_2|w_1).P(w_3|w_1,w_2)\dots P(w_n|w_1,w_2,\dots,w_{n-1}).$$

The complexity of this computation increases with the length of the sequence, making it difficult to model long-range dependencies without sophisticated techniques. To manage this complexity, various approaches to language modeling have been developed. These range from simple n-gram models, which consider only a fixed number of previous words, to advanced neural network-based models, which can capture more complex dependencies and longer contexts. The effectiveness of these models is evaluated using metrics such as perplexity, which measures how well a model predicts a sample of text. Lower perplexity indicates better predictive performance, making it a critical measure in developing and refining language models [20].

4.1. The n-gram model

The n-gram model is a prevalent type of language model employed in various natural language processing (NLP) tasks. It operates based on the Markov assumption[21] , which posits that the probability of a word depends only on a finite number of preceding words. Specifically, an n-gram refers to a contiguous sequence of *n* words. For instance, a bigram comprises two words, while a trigram comprises three words (as shown in Figure 1). These n-grams facilitate the estimation of the probability of a word given its preceding context, thereby enabling the prediction of subsequent words in a sequence.

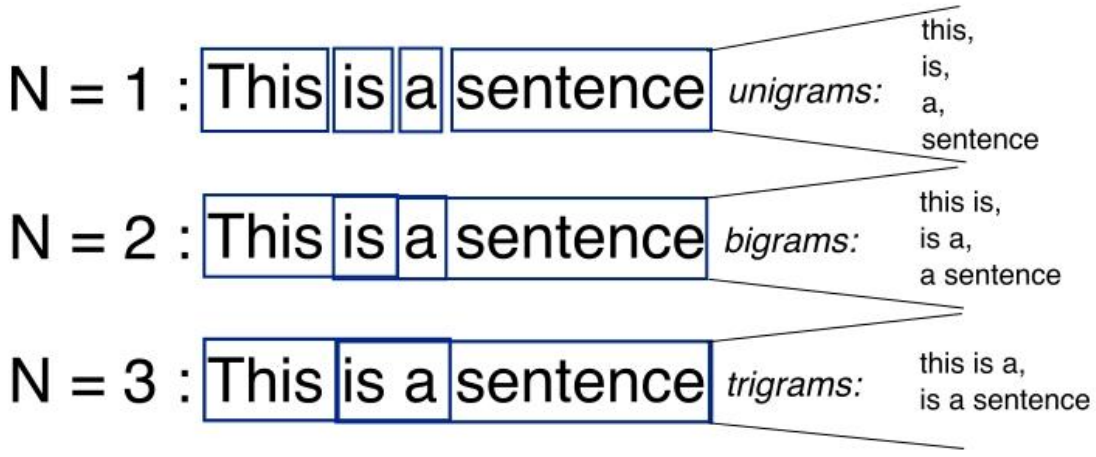

Figure 1: An illustration of n-grams

To calculate the probability of a word in an n-gram model, we count the occurrences of n-grams within a training corpus. The probability of a word given its context is then estimated as the frequency of the n-gram sequence divided by the frequency of the preceding (n-1)-gram sequence. For instance, in a trigram model, to calculate the probability of a word $w3$ given the preceding words $w1$ and $w2$, we count the occurrences of the trigram(w_1, w_2, w_3)This probability can be represented as $(w_3|w_1, w_2)$.

By applying the Markov assumption[21], the n-gram model simplifies the computation of probabilities by considering only a limited context window. However, this approach has limitations. As the value of n increases, the model becomes more sensitive to the sparsity of training data[5]. This sensitivity arises because a large amount of training data is required to accurately estimate the probabilities of less frequent n-grams. To address this issue, various smoothing techniques are employed in n-gram models[5]. Smoothing methods, such as add-one smoothing, backoff, and interpolation, adjust the probability estimates by redistributing probability mass from more frequent n-grams to less frequent ones.

In summary, language models assign probabilities to sentences by leveraging the Chain Rule, the Markov assumption, n-grams, and other techniques. These probabilistic models play a crucial role in various natural language processing tasks, providing valuable insights into sentence structure and aiding in diverse NLP applications.

4.2. Word Embeddings

Word embedding is a method employed in text analysis to represent words. It entails assigning a real-valued vector to each word, capturing its meaning in a way that promotes similarity between words of similar meanings in the vector space. The proximity of two words in the vector space indicates their likelihood to share similar meanings[5]. Techniques such as language modeling and feature learning are utilized to generate word embeddings, converting words or phrases from the vocabulary into numerical vectors. This process facilitates computational analysis and modeling tasks by representing words as real-valued numbers. These vectors contain semantic and syntactic information about words, aiding machines in processing and understanding natural language more effectively[22].

Word embeddings are typically learned from large text corpora using unsupervised learning methods like Word2Vec[23], GloVe[24], or FastText[25]. These methods create vector representations by considering the context in which words appear, assuming that words with similar meanings will occur in similar contexts. One of the primary advantages of word embeddings is their ability to capture semantic relationships between words. For example, in a well-trained word embedding space, the vectors for "king" and "queen" would be close to each other, indicating their semantic similarity. Similarly, the vector for "man" might be closer to the vector for "woman" than to the vector for "car."

Word embeddings have proven valuable in various NLP tasks, enhancing the performance of applications like text classification, sentiment analysis, named entity recognition, machine translation, and information retrieval[21]. By representing words as continuous vectors, word embeddings enable algorithms to better capture word meaning and context, leading to improved performance in these tasks. Moreover, word embeddings can capture analogical relationships between words. For instance, in the embedding space, the vector resulting from the equation "king" - "man" + "woman" would be close to the vector representation of "queen," enabling computational models to perform analogical reasoning tasks.

In practice, pre-trained word embeddings are often used, leveraging large-scale corpora and sophisticated training algorithms. These pre-trained embeddings can be readily employed in various NLP applications, saving time and computational resources. Word embeddings provide a compact and meaningful representation of words in NLP tasks, capturing semantic relationships and contextual information to enable algorithms to understand and process natural language more effectively. The use of word embeddings has become standard practice in many NLP applications, driving advancements in the field.

5. Text mining

Text mining, also known as text data mining or text analytics, is the process of extracting meaningful information and insights from unstructured text data. It uses various techniques such as NLP, machine learning, and information retrieval to transform raw text into structured data suitable for analysis and decision making. With the increasing availability of digital texts

such as emails, social media posts, online reviews, and research articles, the importance of text mining has grown exponentially.

The main goal of text mining is to discover patterns, trends, and relationships in large amounts of text data. This is achieved through several key tasks, including information extraction, sentiment analysis, topic modeling, document clustering, and text classification [26].

5.1. Text mining tasks

5.1.1 Information extraction

Information extraction is a fundamental task in text mining, which involves identifying specific information from text, such as names of people, organizations, places, dates, and other entities. Named entity recognition is a common method for accomplishing this task, which allows the system to accurately locate and classify named entities in a text[27]. This is particularly useful for a variety of applications, such as creating structured databases from unstructured text, improving search engine functionality, and improving information retrieval systems .

5.1.2. Topic modeling

Topic modeling is another important task in text mining, which involves discovering underlying themes in a collection of documents. Techniques such as Latent Dirichlet Allocation (LDA) are often used to identify topics based on the co-occurrence patterns of words in a text[28]. Topic modeling helps organize and summarize large text corpora and facilitates the understanding of major themes and trends .

5.1.3. Document clustering

Document clustering is the process of grouping similar documents based on their content. This task is very useful for organizing large amounts of text, making document retrieval

easier, and improving search engine performance. Clustering algorithms such as K-Means and hierarchical clustering are used to identify and group related documents [29].

5.1.4. Text classification

Text classification assigns predefined categories or labels to documents based on their content. This task is crucial for various applications such as spam detection, message classification, and sentiment classification. Machine learning algorithms such as support vector machines (SVM), decision trees, and neural networks are widely used for text classification tasks [30].

6. External Semantic Resources

External semantic resources play a key role in improving the capabilities of NLP systems by providing structured information that can be used to better understand and analyze text. These resources provide valuable contextual and semantic information that can improve various NLP tasks [31].

6.1. Ontologies

Ontologies are formal representations of knowledge in a given domain, consisting of a set of concepts and the relationships between them. They are used to model the structure of information and enable machines to understand the meaning of terms and their relationships. Ontologies enable more accurate information extraction and retrieval by providing a semantic framework that guides the interpretation of text [32].

6.2. Dictionaries

Dictionaries, which provide definitions and contextual usage of words, are another important semantic resource. They support various NLP tasks by providing accurate word meanings and linguistic properties, helping with tasks such as disambiguation and part-of-speech tagging [33].

6.3. Citations

Citations are references to other texts that provide additional context, background information, and confirmation of facts. In academic and research contexts, citations are critical to understanding the source and credibility of information. NLP systems can use citation networks to analyze the influence and relevance of scientific papers, enabling more effective literature retrieval and knowledge discovery [34]. Citation analysis helps identify important works and trends within a field and facilitates the retrieval of relevant documents based on citation patterns[35] .

6.4. Wikipedia

As a collaborative online encyclopedia, Wikipedia is a comprehensive and up-to-date knowledge base covering a wide range of topics. Due to its rich semantic content and network structure, it is widely used in NLP tasks such as entity association, text classification, and knowledge extraction [36]. Wikipedia provides a rich repository of structured information, including infoboxes, categories, and hyperlinks, which can be used to improve the performance of NLP systems by providing detailed and diverse context for various terms and entities [37].

6.5. WordNet

WordNet is a popular lexical database that categorizes English words into sets of synonyms called synsets, it also has a variety of semantic relations between these synsets, including hypernyms, hyponyms, and meronyms. WordNet is beneficial for enhancing the semantic understanding of words and their relationships in text, this supports tasks like word sense disambiguation, semantic similarity measurement, and information retreival [38]. By providing a comprehensive network of word definitions and connections, WordNet increases the complexity and context of text processing.

7. conclusion

In this chapter, we explored the field of Natural Language Processing (NLP),This overview demonstrates NLP's crucial role in enabling machines to understand and generate human language, fostering advancements across multiple domains.

Chapter 2:

Arabic NLP and Automatic processing of Al-Quran

1. Introduction

Arabic Natural Language Processing (ANLP) concerns itself with the processing and understanding of Arabic language using computational methods. This particular area in NLP deals with the difficulties in processing Arabic texts due to the complexity of its syntax, richness of morphology as well as diverse dialects involved. The implication of this scenario is that Arabic NLP has become increasingly important due to the high amount of Arabic digital content that is now available online as well as the necessity for more sophisticated tools to assist with its processing and analysis.

Al-Quran processing is an Arabic NLP subfield that defines analyzing, interpreting, and extracting meaningful information from Al-Quran, the Holy Book of Islam. Most of the approaches of study may involve different techniques and methodologies pertinent to the Quranic Arabic unique linguistic, stylistic, and structural characteristics.

Arabic of Al-Quran shows several characteristic linguistic traits that make it differ from Modern Standard Arabic (MSA) and other Arabic dialects: the distinct syntactic constructions, overuse of the classical vocabularies, and stylistic features such as rhyme and rhythm; the language of Al-Quran demonstrates a very high degree of morphological complexity. Words are often richly inflected and derived. These characteristics make this type of task computationally difficult and resource-intensive.

2. Arabic NLP

2.1 Linguistic Challenges in Arabic NLP

Arabic presents multiple linguistic challenges which compared with other languages, makes natural language processing tasks more complicated. Illustrations of these challenges are in the following list:

1. **Morphological Complexity:** Arabic is a derivational-intensive language with rich morphology. Words are often formed by combining roots, prefixes, suffixes, and infixes, leading to a good deal of possible word forms. For example, the root "ك-ت-ب" can generate

several words like "كتب" (he wrote), "يكتب" (he writes), "مكتب" (office), and "كتب" (books) [39]. This great morphological richness calls for insensitive and sophisticated morphological analysis of stemming techniques to process and interpret Arabic text accurately.

2. **Dialectal Variation:** Arabic has numerous dialects, and they may greatly vary from the Modern Standard Arabic (MSA) in which writing and media are done. Every dialect is characterized by vocabulary, grammar, and even pronunciation differences, further complicating the development of NLP tools requiring handling different varieties of Arabic. For instance, the word for "how" is "kayfa" in MSA, "kif" in the Levantine Arabic, and "izzay" in the Egyptian Arabic[40]. Developing NLP systems which can process those dialects accurately, switch between them and the MSA, is a significant challenge.

3. **Ambiguity and Lack of Vowels:** By nature, Arabic script is written without its short vowels, so high ambiguity is one of the hallmarks of reading it. For example, the word "كتب" could mean "kataba" (he wrote), "kutub" (books), among others, depending on the context. This ambiguity requires advanced disambiguation techniques to accurately identify the intended meaning of words and sentences [39] like context information, part-of-speech tagging, and morphological analysis.

4. **Grammar and Sentence Structure:** Arabic grammar is complex and flexible, with varying word order and sentence structure. While MSA generally follows a Verb-Subject-Object (VSO) order, Subject-Verb-Object (SVO) is also common. In addition, Arabic offers considerable flexibility in sentence construction, which can result in multiple valid syntactic interpretations of the same sentence. This variability requires powerful syntactic analysis techniques to accurately analyze and generate Arabic text [41].

5. **Orthographic variation:** The Arabic script itself can vary in spelling and encoding. For example, the letter "ء" (همزة) can appear in different shapes and positions, which can affect the interpretation of words. In addition, variations in spelling and diacritics can introduce additional complexity, requiring a normalization process to ensure consistent treatment of the script [39].

2.2. Key Arabic NLP Tasks

1. **Tokenization and Morphological Analysis:** Tokenization is part of segmenting text into words, and morphological analysis is necessary to break down the morphemes comprising a word into its constituents such as roots, prefixes or suffixes. The Buckwalter Arabic Morphological Analyzer is one tool that people use extensively and another popular option is Farasa segmenter[42].
2. **Part-of-Speech Tagging:** This is a process in which one allocates grammatical classes, for example, noun, verb, adjective, to every word in a sentence. Part-of-speech tagging in Arabic needs special models that take into consideration the morphological complexity of the language[43].
3. **Named Entity Recognition:** NER is an AI technology that identifies and classifies entities like names of people, organizations, locations, and dates that are present within a text. For Arabic to be effective, its NER systems would need to cater for the languages rich morphology as well as its diverse dialects[44].
4. **Machine Translation:** The translation of text between Arabic and other languages is a critical application of Arabic NLP. The modern systems for machine translation use neural network-based models, such as the Transformer architecture, which attains significant improvements in translation quality[43].
5. **Sentiment analysis:** This involves identifying the sentiment or emotional tone that is communicated in Arabic text. The systems for Arabic sentiment analysis must deal with variations in dialect and nuances of language in expressing feelings[45].

2.3. Applications of Arabic NLP

Arabic NLP has many uses in different areas, and here we discuss specific applications, their importance, and challenges.

2.3.1. Information Retrieval

IR systems for Arabic are designed to fetch relevant documents in response to user queries. These systems must handle the morphological richness and syntactic complexity of Arabic. Techniques such as stemming, which reduces words to their root forms, and advanced ranking algorithms are employed to improve retrieval accuracy [46]. For example, search engines tailored for Arabic content need to understand different word forms and syntactic structures to provide accurate search results.

2.3.2. Question Answering

QA systems aim to provide precise answers to user queries posed in natural language. For Arabic, these systems must handle the language's complexity and variability. QA systems often use information retrieval techniques to find relevant passages and natural language understanding models to extract answers. Recent advancements involve deep learning models that understand context and nuances in the Arabic language, providing more accurate answers [47].

2.3.3. Text Summarization

Text summarization involves creating concise summaries of longer Arabic texts while retaining the essential information. This task can be extractive, where key sentences are selected, or abstractive, where new sentences are generated to summarize the content. The morphological and syntactic diversity of Arabic makes this task challenging. Advanced models like sequence-to-sequence neural networks have been used to generate coherent and contextually appropriate summaries[48].

2.3.4. Keyword Extraction

Keyword extraction identifies significant words or phrases within Arabic texts that capture the main topics. This task is crucial for indexing, tagging, and information retrieval. Techniques include statistical methods, such as TF-IDF (Term Frequency-Inverse Document Frequency),

and more sophisticated machine learning approaches. The richness of Arabic morphology requires these systems to effectively disambiguate and accurately identify keywords [49].

3. Automatic Processing of Al-Quran

3.1. Applications of Al-Quran Processing

3.1.1. Information Retrieval

Information retrieval systems designed for Quranic texts aim to provide accurate and relevant responses to queries posed in natural language. These systems are particularly useful for scholars, researchers, and students who seek to find specific verses or themes within the Quran. For instance, the Quranic Arabic Corpus provides a searchable database of the Quranic text, enabling users to locate verses based on keywords or phrases [59]. Advanced IR systems also incorporate semantic search capabilities, allowing users to search for concepts rather than exact word matches, thus improving the retrieval of relevant information .

3.1.2. Cross-Language Information Retrieval

Cross-language information retrieval systems facilitate accessing Quranic information across different languages. These systems allow users to query in one language and retrieve relevant Quranic texts in another, bridging language barriers and enhancing accessibility. Such systems are particularly valuable in multilingual contexts, enabling broader access to Quranic knowledge for non-Arabic speakers[60] .

3.1.3. Question Answering

Question answering systems for the Quran aim to provide accurate responses to user queries by understanding the semantic context of the questions and retrieving relevant passages. These systems leverage advanced NLP techniques, including deep learning models, to interpret questions and provide precise answers. Such systems are integrated into educational platforms and virtual assistants, helping users find answers to their Quranic queries

efficiently[47] .

3.1.4. Keyword Extraction

Keyword extraction systems identify key terms and phrases within the Quranic text, facilitating tasks such as indexing, tagging, and thematic analysis. These systems help in highlighting the most significant concepts and entities within the text, aiding in deeper analysis and research. Such tools are essential for creating thematic indices and supporting topic-based studies of the Quran [61].

3.2. Linguistic Challenges

Processing the Quranic text presents several unique linguistic challenges that distinguish it from other types of text processing. These challenges arise due to the text's historical, religious, and linguistic context.

- **Classical Arabic**

The Quran is written in Classical Arabic, a language form that differs significantly from Modern Standard Arabic and various contemporary Arabic dialects. Classical Arabic features complex grammatical structures, rich morphology, and a vast lexicon that includes many words and expressions not commonly used today. For instance, the morphological richness of Arabic means that a single root can generate numerous word forms through various inflections and derivations [39]. This complexity necessitates specialized linguistic resources and tools tailored to Classical Arabic to handle tasks such as tokenization, part-of-speech tagging, and syntactic parsing accurately.

- **Diacritics and Orthography**

The Quranic text includes diacritics, which are critical for correct pronunciation and meaning but are often omitted in other forms of Arabic text. Diacritics can change the meaning of words significantly; hence, their correct handling is essential in text processing. Additionally, the Quran employs a specific orthography with unique characters and symbols that must be accurately processed to maintain the text's integrity[45].

- **Semantic Complexity**

The Quranic text is semantically rich and often uses metaphorical and idiomatic expressions. Understanding these expressions requires deep semantic analysis beyond simple lexical matching. For instance, the text frequently employs metaphors, allusions, and allegories, which can be challenging for computational models to interpret accurately. Researchers must develop advanced semantic models that can capture these nuances and provide meaningful interpretations [57].

- **Contextual Interpretation**

Verses in the Quran (ayahs) are often context-dependent, meaning their interpretation can vary based on their placement within a chapter (surah) or in relation to other verses. This interdependence requires systems to consider broader contextual information to accurately interpret individual verses. Contextual dependencies make tasks such as machine translation and semantic role labeling particularly challenging, as these tasks must account for the broader discourse context to maintain accuracy [57].

- **Dialectal Variations**

Although the Quran is written in Classical Arabic, understanding its content and applying it to contemporary contexts often involves bridging the gap between Classical Arabic and modern dialects. This challenge is significant in applications such as information retrieval and question answering, where users might pose queries in Modern Standard Arabic or dialects, necessitating robust translation and normalization mechanisms [40].

- **Ambiguity and Polysemy**

The Quran contains a high degree of lexical ambiguity and polysemy, where words can have multiple meanings depending on their context. This characteristic poses a challenge for word sense disambiguation (WSD), an essential task for accurate text processing. Effective WSD in Quranic processing requires leveraging extensive lexical resources and contextual information to resolve ambiguities correctly [58].

4.Conclusion

In this chapter, we explored the unique challenges and recent advancements in Arabic natural

language processing, emphasizing its linguistic complexity and the progress in machine learning application, we also explored various aspects of Al-Quran processing, highlighting the unique linguistic challenges posed by the Quranic text and the key areas of research and application.

Part 2 :
contribution

Chapter1:

Requirement analysis and design

1. Introduction

In this chapter we focus on the analysis and the design of our prototype with more details on the development process without forgetting the functional architecture of the system as well as the different diagrams helping to well define our future system.

Our aim is then to design and implement a system for automatic asking of Q'uran book trying to extract the definitions of the essential concept maybe figured in the holly book.

We start our chapter then with presenting the motivation of our system in order to pass later to the architectural and detailed conception.

2. UML

UML (Unified Modelling Language) is the language of graphical modelling.It appeared in the software engineering realm in the context of object oriented paradigm.

UML is used to specify, visualise, modify and build documents required to well develop object oriented applications through offering modelling standard to represent the architecture of our futur application.

2.1. The advantages of UML

- UML is a formal and standardized language:
  - a gain in precision
  - a guarantee of stability
  - the use of tools
- UML is a powerful communication medium
  - It frames the analysis and facilitates the understanding of complex abstract representations.
  - Its versatile nature and flexibility make it a universal language.

2.2. The disadvantages of UML

- Learning and adjustment period.
- The process (not covered by UML).

3. Requirement analysis

3.1. Our approach

Our aim is to design and build an NLP system in order to extract the definitions of some important concepts from Al-Quran. Indeed, this Islamic holly book contains a lot of religious concepts to be defined and to be well understood by Muslims. Consulting of literatures reveals that Qur'an has not been well processed. Indeed, there are just some works that deal with this book of God and the majority of them focuses on syntactical and structural aspects rather than the semantic one. Moreover, the majority of works may be considered as question-answering systems employing some keywords. Few works have addressed the meaning and the semantics but in the major of cases with considering some external semantic resources either specific one such as EL-Hadith or general one like Arabic dictionaries.

We can classify our future system to two utilization modes: Off-Line using, aiming to extract concept definitions and general rules from the whole of Al-Quran and On-Line using where there is a user who comes and demands the service from the system. He/she demands the definitions for a specific concept that he/she submits as well as asking a question and the application has to answer.

To note that On-Line using is based on Off-Line one which extracts knowledge and save it as a structured information to be asked later when the user either ask a question or demand definitions for a submitted concept.

The key question to ask here is how to implement our idea of extracting definitions of concepts from Qur'an. The answer for this question is that we consider some Arabic signals like: «أولئك هم», «ذلك هو», «وأولئك هم»«فأولئك هم», «فذلك هو», «وذلك هو», «ألا ذلك هو» «أولئك الذين», «أولئك».The definitions come before these signals where the associated concepts come later.

We can observe that there are two types of concepts: (i) a concept as a singular term such as «الكافرون», (to note that the majority of concepts are in plural form) and (ii) a concept as an entire phrase like «الذين هدى الله». The last one may be considered also as a question whose the answer is before the definition signal.

Moreover, for the same concept, we can find many definitions. Our next work is to ask the question what the difference is between the definitions of the same concept, what is the plus added in someone which does not exist in the other and what are the verses of common vocabulary. Moreover, there are some advanced concepts that are built on some other basic concepts with more details of definition such as: what comes in the «LOKMAN» chapter, verses (3)-(4)-(5) :

'للمحسنين الذين يقيمون الصلاة ويؤتون الزكاة وهم بالآخرة هم يوقنون أولئك على هدى من ربهم وأولئك هم'

In this instance, we have three concepts: two advanced concepts (maybe synonyms), namely: " على هدى من ربهم " (who are truly guided by their lord) and " المفلحون " (successful) and one basic concept, namely " المحسنين " (good-doers)

Going in this context, we can master the Arabic language and discover the secrets of this rhetorical language as linguistics do. Qur'an as a valuable resource does not require to be checked but it is analysed and processed to learn from it.

Definition signal is not the only kind of signals that may exist in Qur'an, there are surely other types of signals to generate synonymy relationship (or to answer a question «Question-Answer phrase»). Unlike definition signal, the synonymy relationship usually considers two separated signals such as:

- (to be read from right to left)«concept2+هم+concept1+وأن», example:

" وأن الكافرين هم أصحاب النار"

- (to be read from right to left) «concept2+هي/هو+concept1+إن», example:

"إن ربك هو أعلم بمن ضل عن سبيله"

"إن ناشئة الليل هي أشد وطئا"

There are some signals to indicate general rules and laws such as:

- (to be read from right to left) «phrase2+ف+phrase1+ومن», given that: phrase1 is a premise (or a condition) and phrase2 is a conclusion, example:

"ومن يؤت الحكمة فقد أوتي خيرا كثيرا"

These signals to generate rules maybe found also with concept-definition signals like:

"ومن يتعد حدود الله فأولئك هم الظالمون"

3.2. Class diagram

During the analysis phase, the diagram represents the entity (the information) that the user .utilizes. In the design phase, it represents the object structure of object-oriented development

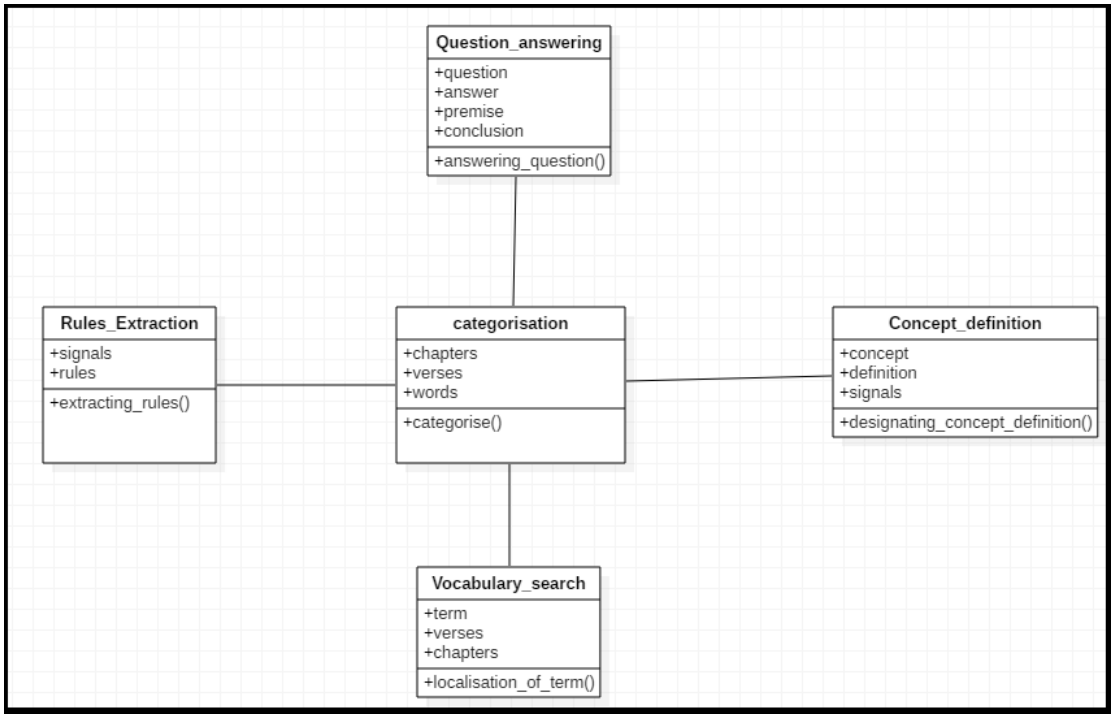

Figure 2: class diagram

4. Conclusion

In this chapter of requirement analysis and design, we focused on presenting in details our approach aiming to extract knowledge from Al-Quran based on some rules we have designated using some Arabic signals. Indeed, we have concept-definition association, synonymy relationship, and some laws to be generated.

Chapter2:
implementation

1. Introduction

In this chapter, we focus on how to implement our prototype starting firstly with the presentation of the considered development environment then we show in details the various used tools as well as the different execution scenarios. We conclude the chapter with presenting some preliminary results.

2. Development environment

In this section, we talk about *JAVA* as programming language we have used as well as Eclipse as the development environment we have employed.

- **JAVA**

Java is a programming language belonging to the oriented object paradigm. Created by Sun MicroSystems enterprise in 1995 and developed since 2009 by Oracle company, *JAVA* technology is highly considered into the web owing to its portability aspect allowing to the applications developed in this language to be deployed into networks no matter what the considered machine architecture and the operating system.

JAVA has received a great attention from the software development community owing to the following reasons:

- It is an oriented object language having common keywords with the C and C++ languages.
- JAVA has a rich library and it is extensible in that it is enough to archive an application as a jar file to be considered as a library in the ulterior.
- Multi-thread programming, exceptions' management, and the access to files and networks are some advantages of the JAVA language.
- It is a multi-platform language ie. the applications are executed without any modification no matter what the considered environment owing to the JVM machine.

- **XML**

XML, or Extensible Markup Language, is a versatile format designed for storing and transporting data. It utilizes tags to structure information hierarchically, making it both human-readable and machine-understandable. XML's flexibility allows it to represent a wide range of data types and is widely used in web development, data interchange, and configuration files. Its adoption is driven by its simplicity, interoperability across different platforms, and the ability to define custom tags tailored to specific data structures, making XML a foundational technology in modern information systems.

- **JDOM**

JDOM, or Java Document Object Model, is a Java-based API for processing XML documents. It provides an intuitive way to represent and manipulate XML data using Java objects, allowing developers to easily navigate through XML structures, modify content, and create new XML documents programmatically. JDOM simplifies XML parsing and generation tasks by abstracting the complexities of low-level XML handling, making it a popular choice for Java developers working with XML data in various applications, from web services to data integration tasks. Its object-oriented approach enhances readability and maintainability of XML processing code compared to traditional DOM or SAX parsers in Java.

- **Eclipse**

Eclipse is a popular integrated development environment (IDE) used primarily for JAVA programming, although it supports various other languages through plug-ins.It provides developers with a compressive set of tools for writing, debugging, and testing JAVA applications efficiently. With features like code completion, syntax highlighting, and project management tools, Eclipse simplifies the development process and enhances productivity. Additionally, its extensibility allows developers to customise their IDE with plug-ins for specific tasks or technologies.

3. Some Screen-Shots of our Prototype

In this section, we present some screen-shots of our prototype. Indeed, Figure 3 depicts the home page of our system where there are some functionalities to ensure.

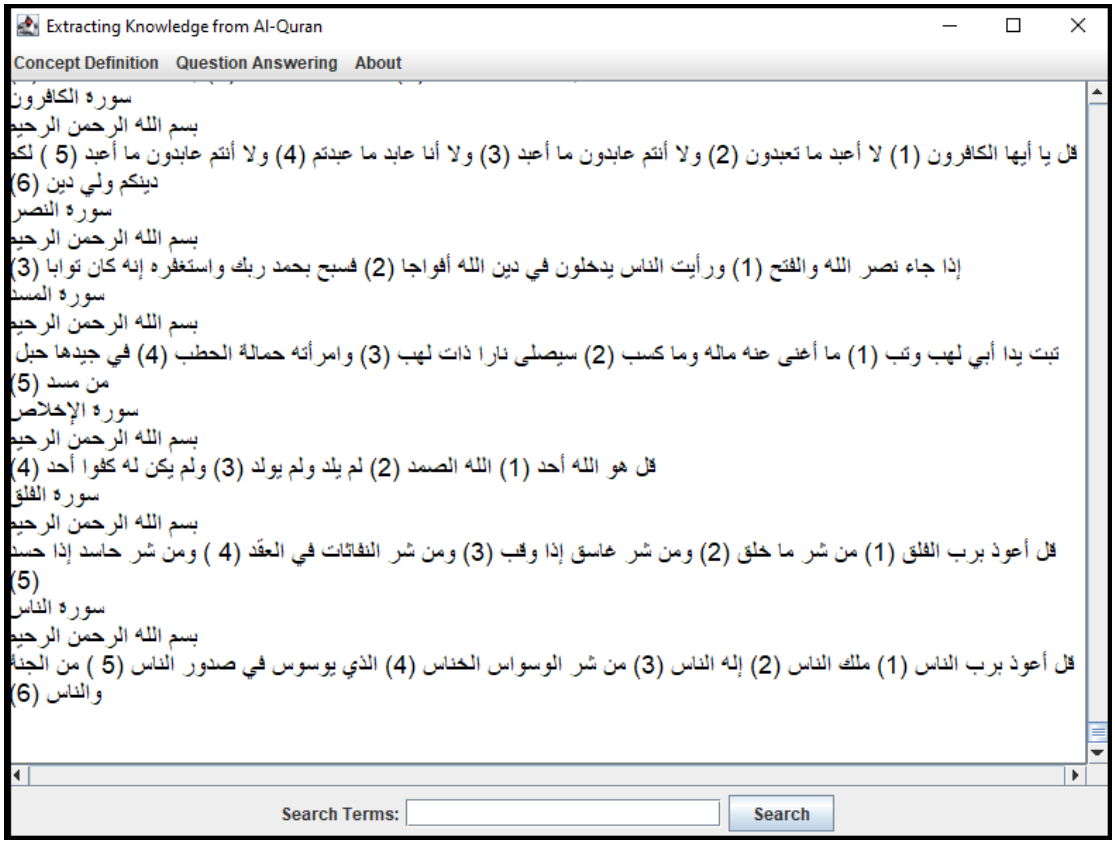

Figure 3. The home Page.

Figure 4 gives some results answered by our system, as a basic function, for the word ‘ربك‘. These results are all the verses, along with 114 chapters of Al-Quran, that include the word.

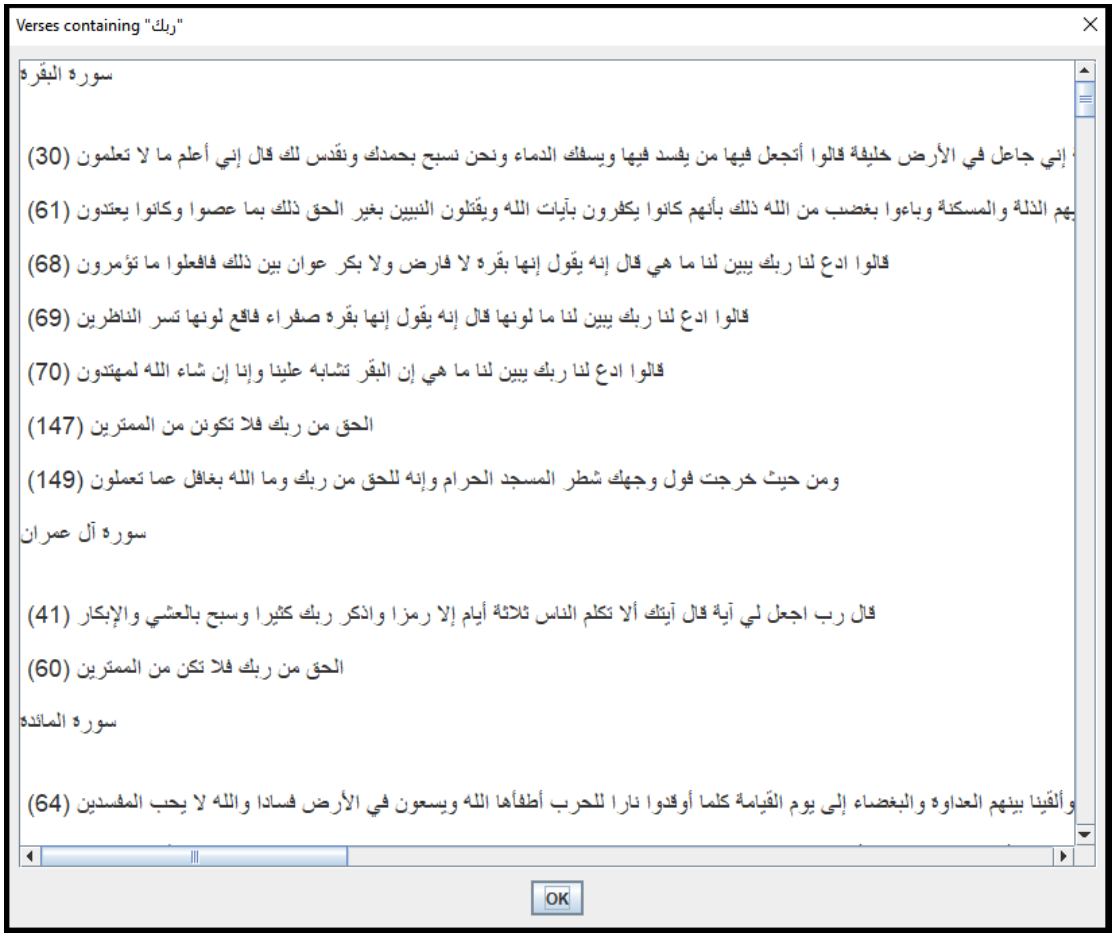

Figure 4. Results of searching for the word ' ربك'.

Figure 5 shows the results given by our system according to the concept '**المفلحون**'. These results are all the definitions may exist in Al-Quran. For each definition, the system gives the number of the associated verse and chapter as well as the signal considered for extracting this definition.

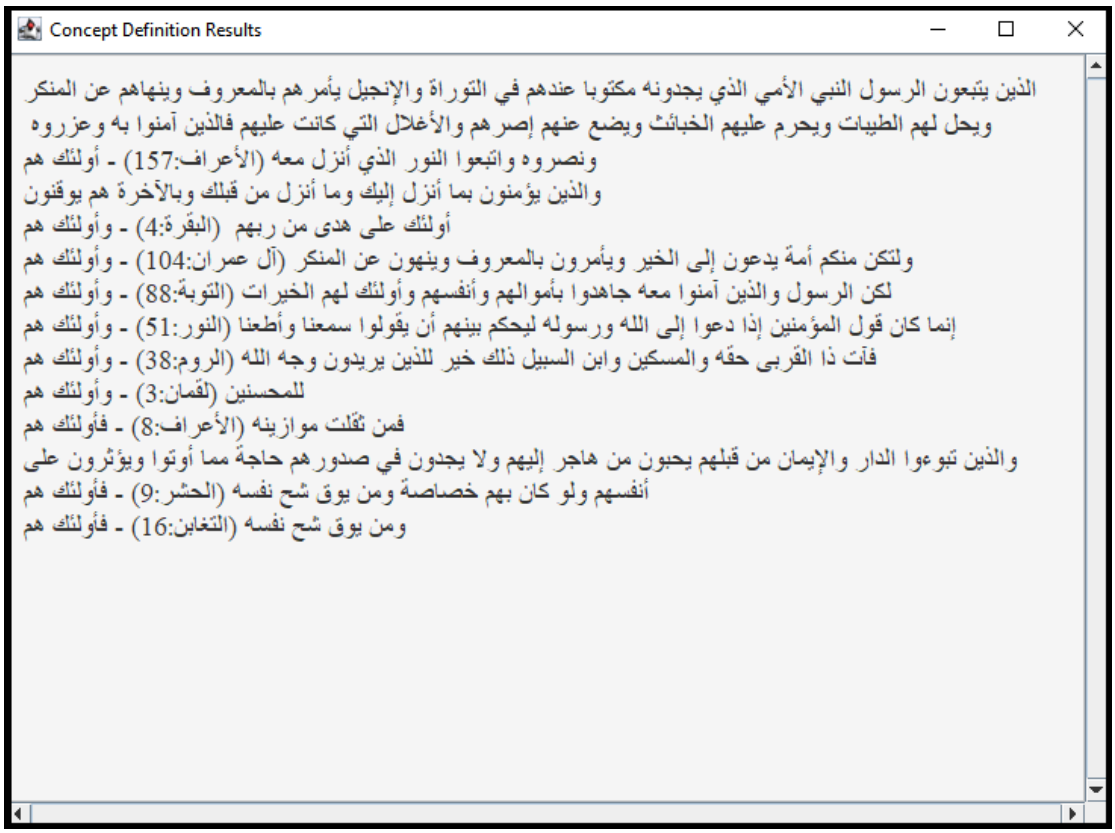

Figure 5. Results of definitions for the concept 'المفلحون'.

Figure 6, associated to the menu 'About' gives detailed information about our application.

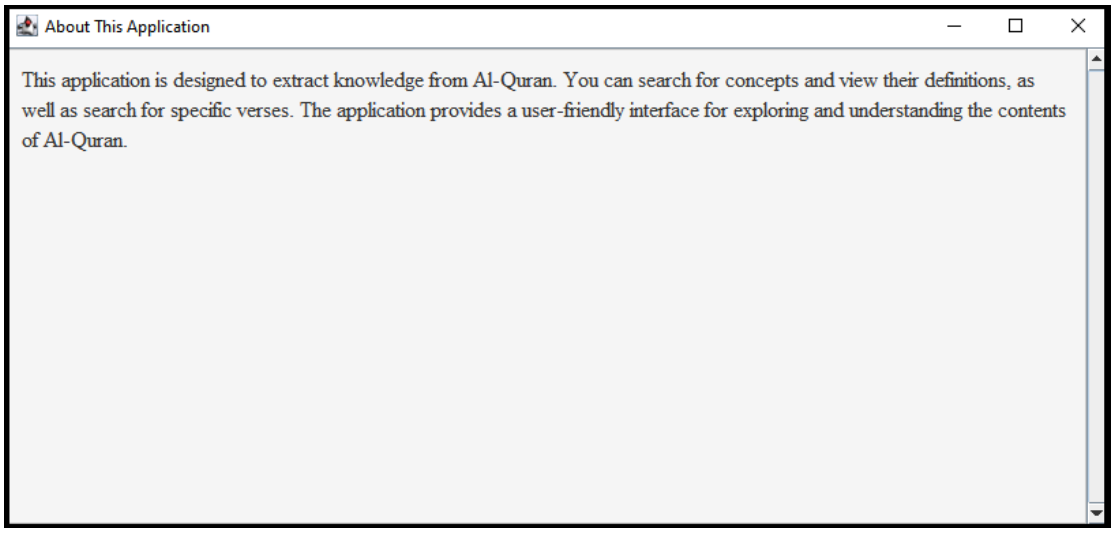

Figure 6. the About Frame.

Figure 7 gives a screen-shot tied to question-answering part of the system where the user gives a question (as a conclusion or a premise of the extracted rule) and the system answer with a response (as the premise or the conclusion for the addressed rule) in this case we ask

about ‘الله’ and the system answers with the conclusions in the figure below. These responses are extracted from XML document of Figure 9.

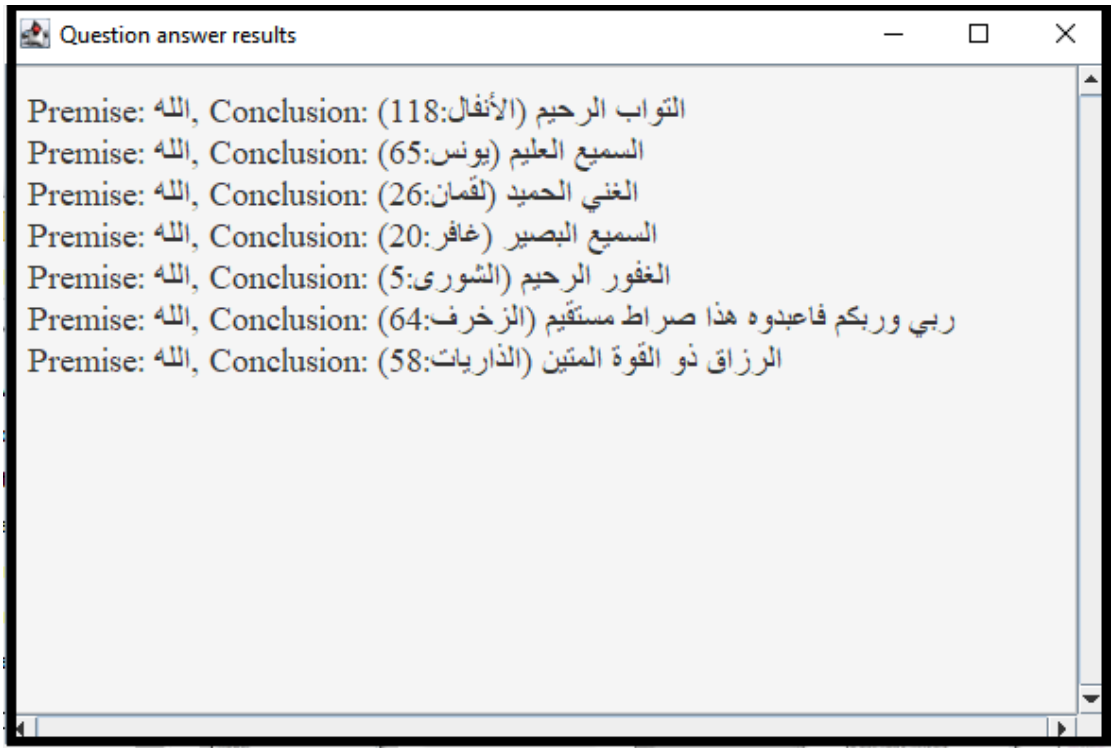

Figure 7. Question-answering part of the system.

Figures 8 and 9 give the XML files that save the knowledge extracted considering the Off-Line using mode. The first document contains knowledge about concept-definition while the second one includes knowledge about premise-conclusion of the different found rules.

Figure 8. XML document which contains the concepts and their associated definitions.

Figure 9. XML document which contains premises and conclusions of the found rules

4. Results

Our application allows to generate the following results:

Concept	Signal	Definition	Chapter and verse
الخاسرون	أولئك هم	الذين ينقضون عهد الله من بعد ميثاقه ويقطعون ما أمر الله به أن يوصل ويفسدون في الأرض	سورة البقرة (27)
		ليميز الله الخبيث من الطيب ويجعل الخبيث بعضه على بعض فيركمه جميعا فيجعله في جهنم	سورة الأنفال (37)
		قل كفى بالله بيني وبينكم شهيدا يعلم ما في السماوات والأرض والذين آمنوا بالباطل وكفروا بالله	سورة العنكبوت (52)

سورة الزمر(63)	والذين كفروا بآيات الله		
سورة التوبة(69)	كالذين من قبلكم كانوا أشد منكم قوة وأكثر أموالا وأولادا فاستمتعوا بخلاقهم فاستمتعتم بخلاقكم كما استمتع الذين من قبلكم بخلاقهم وخضتم كالذي خاضوا أولئك حبطت أعمالهم في الدنيا والآخرة	وأولئك هم	
سورة البقرة(121)	الذين آتيناهم الكتاب يتلونه حق تلاوته أولئك يؤمنون به ومن يكفر به	فأولئك هم	
سورة الأعراف(178)	ومن يضلل		
سورة المنافقون(9)	يا أيها الذين آمنوا لا تلهكم أموالكم ولا أولادكم عن ذكر الله ومن يفعل ذلك		
سورة النساء(150)-(151)	إن الذين يكفرون بالله ورسله ويريدون أن يفرقوا بين الله ورسله ويقولون نؤمن ببعض ونكفر ببعض ويريدون أن يتخذوا بين ذلك سبيلا	أولئك هم	الكافرون حقا
سورة الأعراف (157)	الذين يتبعون الرسول النبي الأمي الذي يجدونه مكتوبا عندهم في التوراة والإنجيل يأمرهم بالمعروف وينهاهم عن المنكر ويحل لهم الطيبات ويحرم عليهم الخبائث ويضع عنهم إصرهم والأغلال التي كانت عليهم فالذين آمنوا به وعزروه ونصروه واتبعوا النور الذي أنزل معه		المفلحون
سورة البقرة(4)-(5)	والذين يؤمنون بما أنزل إليك وما أنزل من قبلك وبالآخرة هم يوقنون أولئك على هدى من ربهم	وأولئك هم	
سورة آل عمران(104)	ولتكن منكم أمة يدعون إلى الخير ويأمرون بالمعروف وينهون عن المنكر		
سورة التوبة(88)	لكن الرسول والذين آمنوا معه جاهدوا بأموالهم وأنفسهم وأولئك لهم الخيرات		
سورة النور(51)	إنما كان قول المؤمنين إذا دعوا إلى الله ورسوله ليحكم بينهم أن يقولوا سمعنا وأطعنا		
سورة الروم(38)	فآت ذا القربى حقه والمسكين وابن السبيل ذلك خير للذين		

	يريدون وجه الله		
سورة لقمان(3)-(4)-(5)	للمحسنين الذين يقيمون الصلاة ويؤتون الزكاة وهم بالآخرة هم يوقنون أولئك على هدى من ربهم		
سورة الأعراف(8)-سورة المؤمنون(102)	فمن ثقلت موازينه	فأولئك هم	
سورة الحشر(9)	والذين تبوءوا الدار والإيمان من قبلهم يحبون من هاجر إليهم ولا يجدون في صدورهم حاجة مما أوتوا ويؤثرون على أنفسهم ولو كان بهم خصاصة ومن يوق شح نفسه		
سورة التغابن(16)	ومن يوق شح نفسه		
سورة الأعراف(179)	ولقد ذرأنا لجهنم كثيرا من الجن والإنس لهم قلوب لا يفقهون بها ولهم أعين لا يبصرون بها ولهم آذان لا يسمعون بها أولئك كالأنعام بل هم أضل	أولئك هم	الغافلون
سورة النحل(108)	أولئك الذين طبع الله على قلوبهم وسمعهم وأبصارهم	وأولئك هم	
سورة الأنفال(74)	والذين آمنوا وهاجروا وجاهدوا في سبيل الله والذين آووا ونصروا	أولئك هم	المؤمنون حقا
سورة الأنفال(3)-(4)	الذين يقيمون الصلاة ومما رزقناهم ينفقون		
سورة المؤمنون(8)/(9)	والذين هم لأماناتهم وعهدهم راعون والذين هم على صلواتهم يحافظون		الوارثون
سورة النور(50)	أفي قلوبهم مرض أم ارتابوا أم يخافون أن يحيف الله عليهم ورسوله		الظالمون
سورة البقرة(229)	ومن يتعد حدود الله	فأولئك هم	
سورة آل عمران(94)	فمن افترى على الله الكذب من بعد ذلك		
سورة المائدة(45)	ومن لم يحكم بما أنزل الله		
سورة التوبة(23)	يا أيها الذين آمنوا لا تتخذوا آباءكم وإخوانكم أولياء إن		

	استحبوا الكفر على الإيمان ومن يتولهم منكم		
سورة الحجرات(11)	ومن لم يتب		
سورة الممتحنة(9)	إنما ينهاكم الله عن الذين قاتلوكم في الدين وأخرجوكم من دياركم وظاهروا على إخراجكم أن تولوهم ومن يتولهم		
سورة الحشر(19)	ولا تكونوا كالذين نسوا الله فأنساهم أنفسهم	أولئك هم	الفاسقون
سورة النور(4)	والذين يرمون المحصنات ثم لم يأتوا بأربعة شهداء فاجلدوهم ثمانين جلدة ولا تقبلوا لهم شهادة أبدا	وأولئك هم	
سورة المائدة(47)	وليحكم أهل الإنجيل بما أنزل الله فيه ومن لم يحكم بما أنزل الله	فأولئك هم	
سورة النور(55)	ومن كفر بعد ذلك		
سورة عبس(40)-(41)	ووجوه يومئذ عليها غبرة ترهقها قترة	أولئك هم	الفجرة الكفرة
سورة الحجرات(15)	إنما المؤمنون الذين آمنوا بالله ورسوله ثم لم يرتابوا وجاهدوا بأموالهم وأنفسهم في سبيل الله		الصادقون
سورة الحشر(8)	للفقراء المهاجرين الذين أخرجوا من ديارهم وأموالهم يبتغون فضلا من الله ورضوانا وينصرون الله ورسوله		
سورة الحجرات(7)	واعلموا أن فيكم رسول الله لو يطيعكم في كثير من الأمر لعنتم ولكن الله حبب إليكم الإيمان وزينه في قلوبكم وكره إليكم الكفر والفسوق والعصيان		الراشدون
سورة الزمر(33)	والذي جاء بالصدق وصدق به		المتقون
سورة البقرة(177)	ليس البر أن تولوا وجوهكم قبل المشرق والمغرب ولكن البر من آمن بالله واليوم الآخر والملائكة والكتاب والنبيين وآتى المال على حبه ذوي القربى واليتامى والمساكين وابن السبيل والسائلين وفي الرقاب وأقام الصلاة وآتى الزكاة والموفون بعهدهم إذا عاهدوا والصابرين في البأساء والضراء وحين البأس أولئك الذين صدقوا	وأولئك هم	
سورة الحديد(19)	والذين آمنوا بالله ورسله أولئك هم الصديقون والشهداء عند	أولئك	أصحاب الجحيم

	ربهم لهم أجرهم ونورهم والذين كفروا وكذبوا بآياتنا		
سورة المائدة(10)-(86)	والذين كفروا وكذبوا بآياتنا		
سورة الحج(51)	والذين سعوا في آياتنا معاجزين		
سورة البينة(6)	إن الذين كفروا من أهل الكتاب والمشركين في نار جهنم خالدين فيها	أولئك هم	شر البرية
سورة البينة(7)	إن الذين آمنوا وعملوا الصالحات		خير البرية
سورة البقرة(157)	أولئك عليهم صلوات من ربهم ورحمة	وأولئك هم	المهتدون
سورة آل عمران(10)	إن الذين كفروا لن تغني عنهم أموالهم ولا أولادهم من الله شيئا		وقود النار
سورة آل عمران(90)	إن الذين كفروا بعد إيمانهم ثم ازدادوا كفرا لن تقبل توبتهم		الضالون
سورة التوبة(10)	لا يرقبون في مؤمن إلا ولا ذمة		المعتدون
سورة التوبة(20)	الذين آمنوا وهاجروا وجاهدوا في سبيل الله بأموالهم وأنفسهم أعظم درجة عند الله		الفائزون
سورة المؤمنون(52)	ومن يطع الله ورسوله ويخش الله ويتقه	فأولئك هم	
سورة الروم(39)	وما آتيتم من ربا ليربو في أموال الناس فلا يربو عند الله وما آتيتم من زكاة تريدون وجه الله		المضعفون
سورة المعارج(31)	فمن ابتغى وراء ذلك		العادون
سورة البقرة(85)-(86)	ثم أنتم هؤلاء تقتلون أنفسكم وتخرجون فريقا منكم من ديارهم تظاهرون عليهم بالإثم والعدوان وإن يأتوكم أسارى تفادوهم وهو محرم عليكم إخراجهم أفتؤمنون ببعض الكتاب وتكفرون ببعض فما جزاء من يفعل ذلك منكم إلا خزي في الحياة الدنيا ويوم القيامة يردون إلى أشد العذاب وما الله بغافل عما تعملون	أولئك الذين	اشتروا الحياة الدنيا بالآخرة فلا يخفف عنهم العذاب ولا هم ينصرون
سورة البقرة(174)-(175)	إن الذين يكتمون ما أنزل الله من الكتاب ويشترون به ثمنا قليلا أولئك ما يأكلون في بطونهم إلا النار ولا يكلمهم الله يوم		اشتروا الضلالة بالهدى والعذاب

	القيامة ولا يزكيهم ولهم عذاب أليم		بالمغفرة فما أصبرهم على النار
سورة البقرة(177)	ليس البر أن تولوا وجوهكم قبل المشرق والمغرب ولكن البر من آمن بالله واليوم الآخر والملائكة والكتاب والنبيين وآتى المال على حبه ذوي القربى واليتامى والمساكين وابن السبيل والسائلين وفي الرقاب وأقام الصلاة وآتى الزكاة والموفون بعهدهم إذا عاهدوا والصابرين في البأساء والضراء وحين البأس		صدقوا وأولئك هم المتقون
سورة آل عمران(21)-(22)	إن الذين يكفرون بآيات الله ويقتلون النبيين بغير حق ويقتلون الذين يأمرون بالقسط من الناس فبشرهم بعذاب أليم		حبطت أعمالهم في الدنيا والآخرة وما لهم من ناصرين
سورة النساء(51)-(52)	ألم تر إلى الذين أوتوا نصيبا من الكتاب يؤمنون بالجبت والطاغوت ويقولون للذين كفروا هؤلاء أهدى من الذين آمنوا سبيلا		لعنهم الله ومن يلعن الله فلن تجد له نصيرا
سورة النساء(62)-(63)	فكيف إذا أصابتهم مصيبة بما قدمت أيديهم ثم جاءوك يحلفون بالله إن أردنا إلا إحسانا وتوفيقا		يعلم الله ما في قلوبهم فأعرض عنهم وعظهم وقل لهم في أنفسهم قولا بليغا
سورة المائدة(41)	يا أيها الرسول لا يحزنك الذين يسارعون في الكفر من الذين قالوا آمنا بأفواههم ولم تؤمن قلوبهم ومن الذين هادوا سماعون للكذب سماعون لقوم آخرين لم يأتوك يحرفون الكلم من بعد مواضعه يقولون إن أوتيتم هذا فخذوه وإن لم تؤتوه فاحذروا ومن يرد الله فتنته فلن تملك له من الله شيئا		لم يرد الله أن يطهر قلوبهم لهم في الدنيا خزي ولهم في الآخرة عذاب عظيم
سورة الأنعام(70)	وذر الذين اتخذوا دينهم لعبا ولهوا وغرتهم الحياة الدنيا وذكر به أن تبسل نفس بما كسبت ليس لها من دون الله ولي ولا شفيع وإن تعدل كل عدل لا يؤخذ منها		أبسلوا بما كسبوا لهم شراب من حميم وعذاب أليم بما كانوا يكفرون

سورة الأنعام(88)-(89)	ذلك هدى الله يهدي به من يشاء من عباده ولو أشركوا لحبط عنهم ما كانوا يعملون		آتيناهم الكتاب والحكم والنبوة فإن يكفر بها هؤلاء فقد وكلنا بها قوما ليسوا بها بكافرين
سورة الأنعام(89)-(90)	أولئك الذين آتيناهم الكتاب والحكم والنبوة فإن يكفر بها هؤلاء فقد وكلنا بها قوما ليسوا بها بكافرين		هدى الله فبهداهم اقتده قل لا أسألكم عليه أجرا إن هو إلا ذكرى للعالمين
سورة هود(15)-(16)	من كان يريد الحياة الدنيا وزينتها نوف إليهم أعمالهم فيها وهم فيها لا يبخسون		ليس لهم في الآخرة إلا النار وحبط ما صنعوا فيها وباطل ما كانوا يعملون
سورة هود(20)-(21)	أولئك لم يكونوا معجزين في الأرض وما كان لهم من دون الله من أولياء يضاعف لهم العذاب ما كانوا يستطيعون السمع وما كانوا يبصرون		خسروا أنفسهم وضل عنهم ما كانوا يفترون
سورة الرعد(5)	وإن تعجب فعجب قولهم أإذا كنا ترابا أإنا لفي خلق جديد		كفروا بربهم وأولئك الأغلال في أعناقهم وأولئك أصحاب النار هم فيها خالدون
سورة النحل(108)-(109)	ذلك بأنهم استحبوا الحياة الدنيا على الآخرة وأن الله لا يهدي القوم الكافرين		طبع الله على قلوبهم وسمعهم وأبصارهم وأولئك هم الغافلون
سورة الإسراء(56)-(57)	قل ادعوا الذين زعمتم من دونه فلا يملكون كشف الضر عنكم ولا تحويلا		يدعون يبتغون إلى ربهم الوسيلة أيهم أقرب ويرجون

			رحمته ويخافون عذابه إن عذاب ربك كان محذورا
سورة الكهف(104)-(105)	الذين ضل سعيهم في الحياة الدنيا وهم يحسبون أنهم يحسنون صنعا		كفروا بآيات ربهم ولقائه فحبطت أعمالهم فلا نقيم لهم يوم القيامة وزنا
سورة مريم(56)-(57)-(58)	واذكر في الكتاب إدريس إنه كان صديقا نبيا ورفعناه مكانا عليا		أنعم الله عليهم من النبيين من ذرية آدم وممن حملنا مع نوح ومن ذرية إبراهيم وإسرائيل وممن هدينا واجتبينا إذا تتلى عليهم آيات الرحمن خروا سجدا وبكيا
سورة النور(62)	إنما المؤمنون الذين آمنوا بالله ورسوله وإذا كانوا معه على أمر جامع لم يذهبوا حتى يستأذنوه إن الذين يستأذنونك		يؤمنون بالله ورسوله فإذا استأذنوك لبعض شأنهم فأذن لمن شئت منهم واستغفر لهم الله إن الله غفور رحيم
سورة النمل(4)-(5)	إن الذين لا يؤمنون بالآخرة زينا لهم أعمالهم فهم يعمهون		لهم سوء العذاب وهم في الآخرة هم الأخسرون
سورة الزمر(18)	الذين يستمعون القول فيتبعون أحسنه		هداهم الله وأولئك هم أولو الألباب
سورة الأحقاف(15)-	ووصينا الإنسان بوالديه إحسانا حملته أمه كرها ووضعته		نتقبل عنهم أحسن

(16)	كرها وحمله وفصاله ثلاثون شهرا حتى إذا بلغ أشده وبلغ أربعين سنة قال رب أوزعني أن أشكر نعمتك التي أنعمت علي وعلى والدي وأن أعمل صالحا ترضاه وأصلح لي في ذريتي إني تبت إليك وإني من المسلمين		ما عملوا ونتجاوز عن سيئاتهم في أصحاب الجنة وعد الصدق الذي كانوا يوعدون
سورة الأحقاف(17)-(18)	والذي قال لوالديه أف لكما أتعداننني أن أخرج وقد خلت القرون من قبلي وهما يستغيثان الله ويلك آمن إن وعد الله حق فيقول ما هذا إلا أساطير الأولين		حق عليهم القول في أمم قد خلت من قبلهم من الجن والإنس إنهم كانوا خاسرين
سورة محمد(16)	ومنهم من يستمع إليك حتى إذا خرجوا من عندك قالوا للذين أوتوا العلم ماذا قال آنفا		طبع الله على قلوبهم واتبعوا أهواءهم
سورة محمد(22)-(23)	فهل عسيتم إن توليتم أن تفسدوا في الأرض وتقطعوا أرحامكم		لعنهم الله فأصمهم وأعمى أبصارهم
سورة الحجرات(3)	إن الذين يغضون أصواتهم عند رسول الله		امتحن الله قلوبهم للتقوى لهم مغفرة وأجر عظيم
سورة التوبة(72)	وعد الله المؤمنين والمؤمنات جنات تجري من تحتها الأنهار خالدين فيها ومساكن طيبة في جنات عدن ورضوان من الله أكبر	ذلك هو	الفوز العظيم
سورة يونس(64)	لهم البشرى في الحياة الدنيا وفي الآخرة لا تبديل لكلمات الله		
سورة الدخان(57)	فضلا من ربك		
سورة الحديد(12)	يوم ترى المؤمنين والمؤمنات يسعى نورهم بين أيديهم وبأيمانهم بشراكم اليوم جنات تجري من تحتها الأنهار خالدين		
سورة التوبة(111)	إن الله اشترى من المؤمنين أنفسهم وأموالهم بأن لهم الجنة يقاتلون في سبيل الله فيقتلون ويقتلون وعدا عليه حقا في	وذلك هو	

	التوراة والإنجيل والقرآن ومن أوفى بعهده من الله فاستبشروا ببيعكم الذي بايعتم به		
سورة غافر(9)	وقهم السيئات ومن تق السيئات يومئذ فقد رحمته		
سورة إبراهيم(18)	مثل الذين كفروا بربهم أعمالهم كرماد اشتدت به الريح في يوم عاصف لا يقدرون مما كسبوا على شيء	ذلك هو	الضلال البعيد
سورة الحج(12)	يدعو من دون الله ما لا يضره وما لا ينفعه		
سورة إبراهيم(18)	مثل الذين كفروا بربهم أعمالهم كرماد اشتدت به الريح في يوم عاصف لا يقدرون مما كسبوا على شيء		
سورة الحج(11)	ومن الناس من يعبد الله على حرف فإن أصابه خير اطمأن به وإن أصابته فتنة انقلب على وجهه خسر الدنيا والآخرة	ذلك هو	الخسران المبين
سورة الزمر(15)	فاعبدوا ما شئتم من دونه قل إن الخاسرين الذين خسروا أنفسهم وأهليهم يوم القيامة	ألا ذلك هو	
سورة فاطر(32)	ثم أورثنا الكتاب الذين اصطفينا من عبادنا فمنهم ظالم لنفسه ومنهم مقتصد ومنهم سابق بالخيرات بإذن الله	ذلك هو	الفضل الكبير
سورة الشورى(22)	ترى الظالمين مشفقين مما كسبوا وهو واقع بهم والذين آمنوا وعملوا الصالحات في روضات الجنات لهم ما يشاءون عند ربهم		
سورة الجاثية(30)	فأما الذين آمنوا وعملوا الصالحات فيدخلهم ربهم في رحمته		الفوز المبين

5. Discussion, Conclusion and Perspectives

In this chapter of implementation, we have presented the tools we considered to implement our prototype for knowledge extraction from Al-Quran as well as some screen-shots of the application. Some valuable results have been given which open the avenue for other aspects to be addressed. Unfortunately, we cannot compare the results with what exists in the literature for the reason that we have processed Al-Quran differently. Our work has focused mainly on concept-definition aspect where many definitions, for the same concept, have been extracted. As a perspective, we think to post-process the obtained results in order to

distinguish the difference and the common points may exist between the returned definitions. Many ideas in our minds impose their selves as future issues to be later addressed such as considering Al-Quran style as a model correction in order to evaluate the quality of other texts written in Arabic.

General Conclusion

In this work, we address Arabic natural language and especially Al-Quran. The majority of the works may exist in the literature in this context of Al-Quran processing may be qualified as question_answering systems. Indeed, Al-Quran is a valuable resource to be considered for answer questions but the difficulty is how to process it in order to extract knowledge.

For us, we focused in extracting concept definitions from Al-Quran as an important way for knowledge discovery and extraction. We have considered two using modes: Off-Line, in order to extract knowledge and saving it in structured manner through an XML document, and On-Line mode where our application may supply a service for a user through giving the definitions for a submitted concept and answer some questions based on the found rules.

The obtained results are encouraging which leads us to explore other avenues in the context of knowledge extraction from Al-Quran in our future works.

References

[1] "Jumping NLP Curves: A Review of Natural Language Processing Research [Re view Article] | IEEE Journals & Magazine | IEEE Xplore." (), [Online]. Available: https://ieeexplore.ieee.org/document/6786458.

[2] D. Adiwardana, M.-T. Luong, D. R. So, *et al.* "Towards a Human-like Open-Domain Chatbot." arXiv: 2001 . 09977 [cs, stat]. (Feb. 27, 2020), [Online]. Available: http://arxiv.org/abs/2001.09977.

[3] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Sci ence*, vol. 349, no. 6245, pp. 261–266, Jul. 17, 2015, issn: 0036-8075, 1095-9203. doi: 10.1126/science.aaa8685. [Online]. Available: https://www.science.org/ doi/10.1126/science.aaa8685.

[4] M. Yasunaga, J. Kasai, R. Zhang, *et al.* "ScisummNet: A Large Annotated Cor pus and Content-Impact Models for Scientific Paper Summarization with Citation Networks." arXiv: 1909.01716 [cs]. (Sep. 15, 2019), [Online]. Available: http:

//arxiv.org/abs/1909.01716.

[5] Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.). Prentice Hall.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). "Attention is All You Need." In Advances in Neural Information Processing Systems.

[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of NAACL-HLT.

[8] Brown, T. B., Mann, B., Ryder, N., et al. (2020). "Language Models are Few-Shot Learners." In Advances in Neural Information Processing Systems (NeurIPS).

[9] Hoy, M. B. (2018). "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants." Medical Reference Services Quarterly, 37(1), 81-88.

[10] Wang, Y., Kung, L., & Byrd, T. A. (2018). "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations." Technological Forecasting and Social Change, 126, 3-13.

[11] Lacity, M. C., & Willcocks, L. P. (2018). Robotic Process Automation and Risk Mitigation: The Definitive Guide. SB Publishing.

[12] Binns, R. (2018). "Fairness in Machine Learning: Lessons from Political Philosophy." In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT*).

[13] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2017). "New Avenues in Opinion Mining and Sentiment Analysis." IEEE Intelligent Systems, 28(2), 15-21.

[14] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). "Text mining for market prediction: A systematic review." Expert Systems with Applications, 41(16), 7653-7670.

[15] Nadeau, D., & Sekine, S. (2007). "A survey of named entity recognition and classification." Lingvisticae Investigationes, 30(1), 3-26.

[16] Nenkova, A., & McKeown, K. (2011). "Automatic Summarization." Foundations and Trends in Information Retrieval, 5(2-3), 103-233.

[17] Salton, G., Fox, E. A., & Wu, H. (1975). "Extended Boolean Information Retrieval." Communications of the ACM, 18(11), 613-620.

[18] Nie, J.-Y. (2010) Cross-language information retrieval. Cham, Switzerland: Springer.

[19] Hulth, A. (2003). "Improved Automatic Keyword Extraction Given More Linguistic Knowledge." In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[20] Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). "Perplexity—a measure of the difficulty of speech recognition tasks." The Journal of the Acoustical Society of America, 62(S1), S63-S63.

[21] Paek, T. and Chickering, D.M. (2006) Evaluating the Markov assumption in Markov decision processes for spoken dialogue management - language resources and evaluation, SpringerLink. Available at: https://link.springer.com/article/10.1007/s10579-006-9008-2

[22] Almeida, F. and Xexéo, G. (2023) Word embeddings: A survey, arXiv.org. Available at: https://arxiv.org/abs/1901.09069

[23] Mikolov, T. et al. (2013) Efficient estimation of word representations in vector space, arXiv.org. Available at: https://arxiv.org/abs/1301.3781.
[24] Pennington, J., Socher, R. and Manning, C.D. (2014) Glove: Global vectors for word representation, ACL Anthology. Available at: https://aclanthology.org/D14-1162/.
[25] Joulin, A. et al. (2016) Bag of tricks for efficient text classification, arXiv.org. Available at: https://arxiv.org/abs/1607.01759v3.
[26] Feldman, R. and Sanger, J. (2013) The text Mining Handbook: Advanced Approaches in analyzing unstructured data. Cambridge: Cambridge University Press. Available at: https://tinyurl.com/56eke7dh.
[27] Nadeau, D. and Sekine, S. (1970) A survey of named entity recognition and classification, Latest TOC RSS. Available at: https://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002.
[28] Blei, D., Ng, A. and Jordan, M. (2003) Latent dirichlet allocation David M. Blei, Andrew Y. Ng and ... Available at: https://ai.stanford.edu/~ang/papers/nips01-lda.pdf.
[29] Steinbach, M., Karypis, G. and Kumar, V. (2000) A comparison of document clustering techniques, University Digital Conservancy Home. Available at: https://conservancy.umn.edu/handle/11299/215421.
[30] Sebastiani, F. and Ricerche, C.N. delle (2002) Machine learning in automated text categorization, ACM Computing Surveys. Available at: https://dl.acm.org/doi/10.1145/505282.505283.
[31] Navigli, R. and Velardi, P. (2004) Learning domain ontologies from document warehouses and dedicated web sites, MIT Press. Available at: https://direct.mit.edu/coli/article/30/2/151/1851/Learning-Domain-Ontologies-from-Document.
[32] Thomas R. Gruber (2002) A translation approach to portable ontology specifications, Knowledge Acquisition. Available at: https://www.sciencedirect.com/science/article/pii/S1042814383710083.
[33] Miller, G.A. et al. (1990) Introduction to wordnet: An on-line lexical database*, OUP Academic. Available at: https://academic.oup.com/ijl/article/3/4/235/923280.
[34] Garfield, E. (1979) Citation indexing, its theory and application in science, technology, and humanities. Philadelphia: ISI Press.
[35] Bornmann, L. and Daniel, H. (2008) What do citation counts measure? A review of studies on citing behavior, Journal of Documentation. Available at: https://www.emerald.com/insight/content/doi/10.1108/00220410810844150/full/html.
[36] Milne, D. and Ian H. Witten (2008) Learning to link with wikipedia: Proceedings of the 17th ACM Conference on Information and Knowledge Management, ACM Conferences. Available at: https://dl.acm.org/doi/10.1145/1458082.1458150.
[37] Zesch, T., Müller, C. and Gurevych, I. (2008) Extracting lexical semantic knowledge from Wikipedia and wiktionary, ACL Anthology. Available at: https://aclanthology.org/L08-1139/.
[38] Miller, G.A. (1995) WordNet: A lexical database for English: Communications of the ACM: Vol 38, no 11, Communications of the ACM. Available at: https://dl.acm.org/doi/10.1145/219717.219748.

[39] Habash, N.Y. (2010) Introduction to arabic natural language processing. Cham, Switzerland: Springer.
[40] Habash, N., Diab, M. and Rambow, O. (2014) Conventional orthography for dialectal Arabic, ACL Anthology. Available at: https://aclanthology.org/L12-1328/.
[41] Mohamed Maamouri et al. (2004) (PDF) the penn arabic treebank: Building a large-scale annotated Arabic corpus. Available at: https://www.researchgate.net/publication/228693973_The_penn_arabic_treebank_Building_a_large-scale_annotated_arabic_corpus.
[42] Pasha, A. et al. (2014) Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic, ACL Anthology. Available at: https://aclanthology.org/L14-1479/.
[43] Shaalan, K. (2014) A survey of Arabic named entity recognition and classification, MIT Press. Available at: https://direct.mit.edu/coli/article/40/2/469/1475/A-Survey-of-Arabic-Named-Entity-Recognition-and.
[44] Benajiba, Y., Mona Diab and Paolo Rosso (2008) Arabic named entity recognition using optimized feature sets. Available at: https://aclanthology.org/D08-1030.pdf.
[45] Samhaa R. El-Beltagy and Ahmed Ali (2016) Open issues in the sentiment analysis of Arabic social media: A case study | IEEE conference publication | IEEE xplore. Available at: https://ieeexplore.ieee.org/abstract/document/6544421.
[46] Darwish, K. and Magdy, W. (2014) Arabic information retrieval, Foundations and Trends® in Information Retrieval. Available at: https://www.nowpublishers.com/article/Details/INR-031.
[47] Alwaneen, T.H. et al. (2021) Arabic question answering system: A survey - artificial intelligence review, SpringerLink. Available at: https://link.springer.com/article/10.1007/s10462-021-10031-1.
[48] Asmaa Elsaid et al. (2022) IEEE Xplore. Available at: https://ieeexplore.ieee.org/Xplore/home.jsp.
[49] Ebtehal H.Omoush and Venus W. Samawi (2016) Arabic Keyword Extraction using SOM Neural Network. Available at: http://www.ijascse.org/volume-5-theme-based-issue-7/NEURAL_NETWORK_DEMODULATOR.pdf.
[50] Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature.
[51] Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).
[52] AR Alaei, S Becken and B Stantic (2019) Sentiment Analysis in tourism: Capitalizing on big data - Ali Reza Alaei, Susanne Becken, Bela Stantic, 2019. Available at: https://journals.sagepub.com/doi/abs/10.1177/0047287517747753.
[53] Koehn, P. (2012) Statistical machine translation Philipp Koehn. Cambridge: Cambridge University Press.
[54] Wu, Y. et al. (2016) Google's Neural Machine Translation System: Bridging the gap between human and machine translation, arXiv.org. Available at: https://arxiv.org/abs/1609.08144.

[55] Lafferty, J., McCallum, A., & Pereira, F. (2001, June). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Icml (Vol. 1, No. 2, p. 3).

[56] Bollacker, K. et al. (2008) Freebase: Proceedings of the 2008 ACM SIGMOD international conference on management of data, ACM Conferences. Available at: https://dl.acm.org/doi/abs/10.1145/1376616.1376746.

[57] Sharaf, A. B. M., & Atwell, E. (2012, May). QurSim: A corpus for evaluation of relatedness in short texts. In LREC (pp. 2295-2302).

[58] Musahar, S. J., Talib, H., Musahar, R., Azmi, F., & Zakaria, M. Z. (2019). Ambiguity in Holy Quran Commentaries: The Use of Polysemic Words "Imam & Ummah". In Proceedings of the Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016) Social Sciences (pp. 443-454). Springer Singapore.

[59] Dukes, K., Atwell, E., & Habash, N. (2013). Supervised collaboration for syntactic annotation of Quranic Arabic. Language resources and evaluation, 47, 33-62.

[60] Elayeb, B., & Bounhas, I. (2016). Arabic cross-language information retrieval: a review. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 15(3), 1-44.

[61] Al-Taani, A. T., & Al-Gharaibeh, A. M. (2010). Searching concepts and keywords in the Holy Quran. Jordan: Department of Computer Science, Yarmouk University.

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليـــــــــم العالـــــــــي و البحث العلمـــــــــي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جـامعة 20 أوت 1955- سكيكدة
كـــــلية العلـــــوم
مكتبـة الكليـة

Université 20 Août 1955 - Skikda
Faculté des Sciences
Bibliothèque de la Faculté

سكيكدة في / 2024

بطاقة معلومات خاصة بمذكرة التخرج

اسم و لقب الطالب : Boucherkha Ali *

رقم التسجيل : 191936001898 *

اسم و لقب المشرف على المذكرة : Mouloud Mosbah

عنوان المذكرة : Automatic Extraction of Knowledge From Al-Quran

القسم : Informatique

المستوى : Master 2

التخصص : RSD

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire
وزارة التعليـــــــــم العالـــــــــي و البحث العلمـــــــي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université 20 Août 1955- skikda-

Faculté des Sciences

Département d'Informatique

جـامعة 20 أوت 1955 - سكيكدة

كـــــلية العلـــــوم

قسم الاعلام الالي

الرقم : ..../ ق/ ا/ آ/ ك.ع/2024

Autorisation de Dépôt de Mémoire de Master

Je soussigné: ..Dr. Mawloud Mosbah.....

Certifie que l'étudiant(e) :..Ali Boucherkha.....

Spécialité : .....RSD.....

Ayant soutenu le projet intitulé :..Automatic Extraction of Knowledge from Al-Quran.....
.....

A apporté les corrections nécessaires sur son manuscrit de Master

Signature de l'encadreur