An Academic AI Chatbot

Table of Contents

Executive Summary

In today's era of Big Data and AI, data lies at the heart of everything. It is identified in various forms, from organized structured data to more unstructured data. With the exponential growth of unstructured data sources such as social media, emails, and customer feedback, organizations face the challenge of extracting valuable insights from these data streams. The Unstructured Data Analysis focuses on utilizing several data sources available and processing them for understanding. One of the revolutionary advantages of this unstructured data was research and use cases in Natural Language Processing (NLP).

For our project, AI chatbot, we leveraged the NLP techniques that can extract, process, and understand the meaning and context of unstructured textual data. The NLP chatbot addresses this challenge by providing a scalable and efficient solution for analyzing unstructured data in real time.

Project Background

A Natural Language Processing Chatbot is a human simulation to "**Question and Answer**" powered by AI. The use of chatbots is ever evolving in various fields like Customer Support, Marketing, Education, Healthcare, and Entertainment to be precise. In this paper, we have targeted the "Education" domain to analyze various NLP techniques to get the answer in adherence to Natural Language Understanding. Chatbots mimic human conversation, and we wanted to explore the best use-case of AI, where our chatbot is a friend to the students for their exam preparations. One can contact our chatbot to revise the topics and have multiple conversations with the bot without interruption.

Features: The Academic AI Chatbot provides educators and students with a variety of ways to be supported. Students can utilize the "Personalized Assistance" function to have a virtual learning partner that is customized to meet their specific needs. The chatbot can be used to get help with assignments, find further resources, or get clarification on difficult subjects. It is a dependable and approachable resource. Additionally, the chatbot continuously improves its

responses by examining user interactions, guaranteeing that every student receives a customized and adaptive learning experience.

Apart from aiding students, the chatbot offers educators priceless "Driven Insights". The chatbot produces meaningful insights on student learning practices and performance trends by compiling and evaluating user interactions. Teachers can utilize this data to pinpoint problem areas, monitor student development and adjust teaching methods as necessary. Additionally, by facilitating communication between teachers and students, the chatbot enhances the quality of education by promoting a collaborative learning environment where feedback is given in real-time.

In this paper, we start with the History of chatbot - Traditional chatbots and how NLP has revolutionized it. Next follows the Project Introduction, Project scope and Project Technical Flow.

History of Chatbot

News from CNN: In the 1960's a miraculous computer program called Eliza attempted to simulate human conversation. It was capable of understanding human emotions and responding - "I am sorry to hear you are depressed."

Eliza is the widely recognized "First chatbot", for sure was not as versatile as today. It relied on NLU and reacted to keywords on which it was trained on. The approach back then was more on Rule-based ones, where the answers were trained on specific or complete sentences like Hi and How are you?

But now, nearly 60 years later, the market is flooded with chatbots. From tech companies to banks and airlines, everyone delegates many of the customer queries to chatbot. The ability of machine personified to "chat" with humans is an enormous success and with advancement in NLP and ML (Machine Learning), the market has risen to advanced standards.

Today, the chatbot derives the context of the question and provides an answer based on it. And this has made Chatbot a needful "disruptive" technology.

An Academic AI chatbot developed by Team 4

Technological improvements are causing a revolution in education. One innovative approach that has the potential to completely change the educational scene is the Academic AI Chatbot. The chatbot's objective is to close the gap between conventional teaching approaches and contemporary student requirements by smoothly integrating into already-existing educational platforms. The chatbot, with its capacity to provide individualized support and insights, is a revolutionary tool that will change the way educators and students interact with instructional materials and track and adjust to each other's development.

Project Objectives

1. Develop an NLP-powered chatbot capable of processing unstructured textual data.
2. Implement advanced NLP techniques like Data Pre-processing, Topic modeling, Vectorization and Vector Search.
3. Enable the chatbot to extract actionable insights for "Academic Questions" and provide intelligent responses to student queries.
4. Ensure Optimization and efficiency in processing large volumes of unstructured data streams.

Future Scope: Integrate the chatbot with existing Education Business Domain with seamless deployment for utilization.

Methodology

The project followed a systematic methodology with following steps:

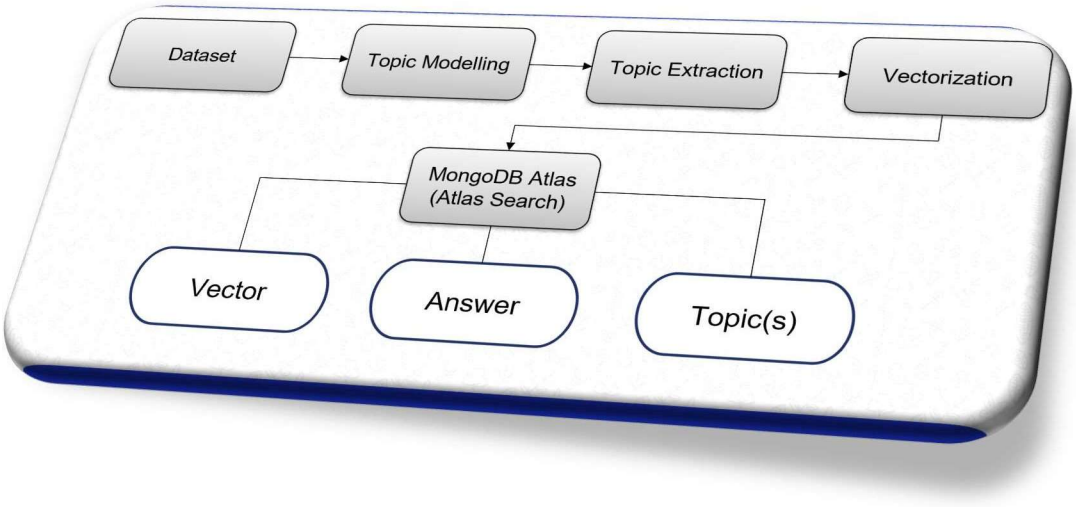
1. **Data Collection**: We aim to gather diverse sources of academic data textual data in the form of "Question" and "Answer" limited to Data Science, Statistics and AI subjects.
2. **Preprocessing**: Cleaning and preprocessing the raw textual data which are the "questions" to remove noise, tokenize text, and normalize formats.
3. **NLP Pipeline**: Developing an NLP pipeline consisting of various modules for
    1) Tokenization

2) Stop Word removal.
3) Lemmatization
4) Topic Modelling
5) Vectorization
6) Storage.

4. **Chatbot Development:** Designing and implementing a conversational interface for the chatbot, integrating it with the NLP pipeline to process user questions and provide intelligent responses.
5. **Deployment:** Deploying the NLP chatbot in environments, ensuring reliability, and seamless integration with existing systems.

Project Architecture

Artificial intelligence (AI) is being increasingly incorporated into education in the modern digital era. One such cutting-edge tool is this chatbot. To train and maximize the chatbot's potential, this project makes use of state-of-the-art natural language processing (NLP) techniques and datasets from websites such as Kaggle. An overview of the NLP methodologies, data sources, and capabilities used in the creation of the Academic AI Chatbot are provided in this paper.



Let us deep dive into the components:

Data Sources

The Academic AI Chatbot's reliance on top-notch datasets obtained from websites like **Kaggle** is essential to its operation. These datasets cover a wide range of topics in statistics, databases, data science, and mathematics. Through the utilization of extensive datasets, the chatbot acquires a profound comprehension of several subject areas, hence facilitating the provision of precise and contextually appropriate answers to user inquiries. The quantity of data also makes it possible for the chatbot's NLP models to be rigorously trained and validated, guaranteeing top performance in real-world situations.

Pre-processing Data

The Academic AI Chatbot's capacity to efficiently analyze and comprehend natural language inputs will determine how well it performs. The project uses a variety of NLP approaches for data pre-processing. To enable further analysis, incoming text is tokenized—that is, broken down into individual words or tokens. Lemmatization makes ensuring that words are reduced to their dictionary or base form, which improves the chatbot's capacity to identify word variants. By removing frequently used terms with limited semantic significance, stop word elimination lowers noise in the input data. Furthermore, prior to the input data being fed into the chatbot's training pipeline, regex processing is used to find and extract pertinent patterns or entities from text.

Tokenization is the process of dividing a text sequence into more manageable chunks, usually words or tokens. Tokenization is essential to the Academic AI Chatbot project because it transforms user input into a format that the chatbot's natural language understanding (NLU) engine can comprehend. This entails dividing the input text into discrete tokens according to preset guidelines, like punctuation or spaces. The text "The lazy dog is jumped over by the quick brown fox" can be tokenized into the following tokens, for instance: ["the", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog"]. The chatbot can evaluate and comprehend the meaning of each word in the context of the user query by dissecting the input text into tokens.

Removal of Stop Words: During text processing, stop words—common terms with limited semantic significance—are frequently filtered out to lower noise and boost computational effectiveness. "The", "and" "is", "in", "of", and so on are a few examples of stop words. Stop word removal in the context of the Academic AI Chatbot project is locating and eliminating such terms from the input text prior to conducting additional analysis. This enhances the accuracy of natural language interpretation and helps to concentrate on the most pertinent stuff. By eliminating stop words from user inquiries, the chatbot can more accurately identify the main ideas and concepts being asked, providing more detailed and accurate answers.

Regex Processing: To find and extract patterns or entities from text, regex (regular expression) processing approaches use pattern-matching methods. Regex processing is used in the Academic AI Chatbot project to identify and extract pertinent data from user questions, such as dates, numbers, or named things. A regex pattern, for instance, could be used to locate and retrieve numerical values from user queries pertaining to mathematical issues. The chatbot's ability to comprehend the structure and context of user inquiries is enhanced by the application of regex processing, which facilitates more precise interpretation and response production.

Lemmatization: The process of reducing words to their dictionary-based or basic form, or lemmas, is called lemmatization. This makes it simpler to recognize and evaluate the semantic meaning of various inflected word forms, such as plurals and verb conjugations, by standardizing them. Lemmatization is used in the Academic AI Chatbot project to make sure that words with similar meanings are regarded as interchangeable. The terms "running", "runs", and "ran" can all be reduced to their fundamental form, "run". The chatbot can more precisely match user inquiries with pertinent responses by breaking down words into its lemma forms, which enhances the quality of interaction and user experience overall.

Details and Visualization of our research

Topic Modelling

Topic modeling is an ***unsupervised*** classification of documents, similar to classification / clustering of supervised data, but this NL technique uses machine learnt way to identify some natural groups of items (topics) even when we are not sure what we are looking for.

As a human, we can easily identify the topic from a sentence, but it's pretty interesting to understand how machine does it for us when we don't provide the labels.

Why topic modeling? Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large textual archives. It can:

Discover the hidden topics/subjects in the collection.

Classify them into the discovered topics.

For AI chatbot, let us say a question belongs to the topics *"Data," "ai"* and *"statistics."* If a user queries "*Can you define Data Science*," they might find the above-mentioned document relevant because it belongs to the topic mentioned above (among other topics). We are able to identify the relevance of questions with respect to topics which can optimize the search.

LDA: Latent Dirichlet Allocation: It is one of the most popular algorithms used for Topic Modelling method.

Caveats:

- LDA works unsupervised where the documents with the same topic will have a lot of words in common.
- LDA is a bag of words model meaning that it only considers individual tokens and not their relationships in a sentence.

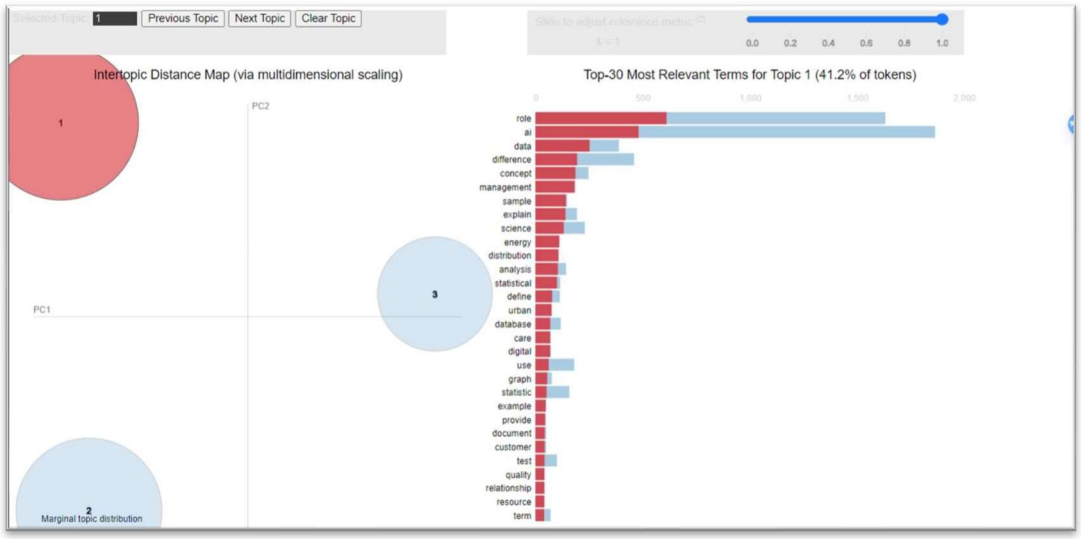
For chatbot, we trained corpus with LDA model to understand extract the topics. These topics were added to every question that was trained and was saved along with the topic model in database. Hence, when the user asked a question, it was passed from the same trained LDA model, to classify in one of the topics identified and hence optimized our search. Below are the input parameters with which we trained our model:

- Number of topics: 3
- Top words: 3 (to save in the database).

pyLDAvis for chatbot:

- Topic 1: role, ai, data
- Topic 2: machine, difference, learning

- Topic 3: model, database, concept



Explanation: The three topics are quite far from each other hence the topics are different from each other. We chose 3 topics, adhering to limited subjects in our raw data. We came up with this number, so that we do not find too many topics. This helped us to reach our goal, where we got our Mongo queries optimized as explained in the Mongo Atlas Vector Search section.

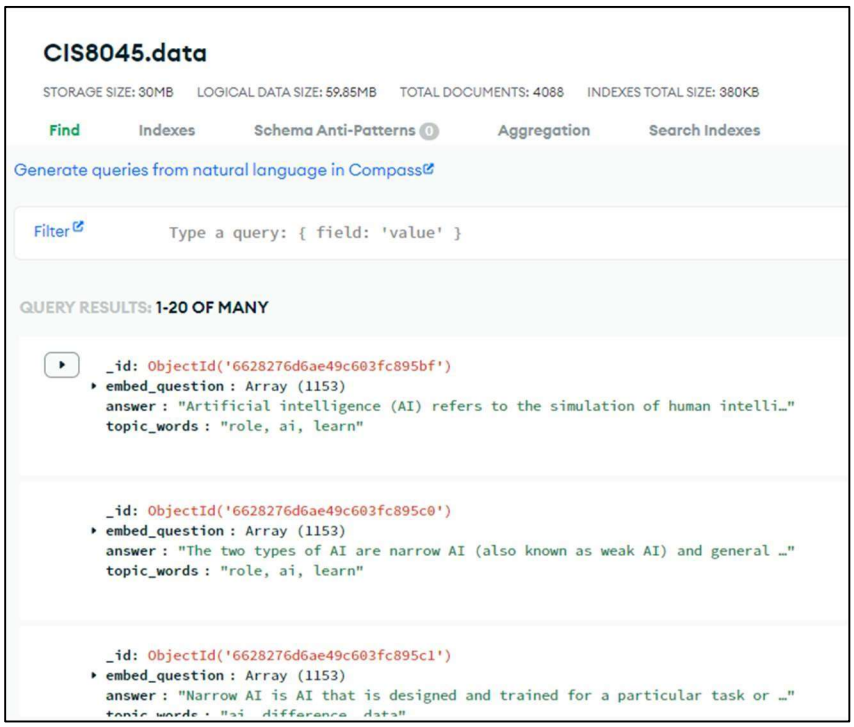
MongoDB

Overview: MongoDB Atlas is the premier cloud database service that offers a fully automated cloud service engineered by MongoDB. It's a database solution for modern applications that not only simplifies operations but also provides robust capabilities for data distribution and management.

How we use it: Within the project, we employ MongoDB Atlas to store our data in the "data" collection, which contains approximately 4088 documents. These documents are JSON-like structures that permit us to represent complex hierarchies, store arrays, and more, providing us with a rich data model that can hold varied data types.

Each document in our "data" collection comprises a set of fields, which are key-value pairs. Here's an overview of the data structure:

- ***_id***: A unique identifier for each document, automatically generated by MongoDB to ensure document uniqueness.
- ***embed_question***: An array with a length of 1153, which holds the numerical representation of questions processed by our NLP model, facilitating semantic search.
- ***answer***: A string that contains the answer for the embedded question that will be retrieved when we call the query.
- ***topic_words***: A list of keywords or tags associated with each question, allowing for a more traditional keyword-based search when necessary.



These fields form the core of each document's structure, enabling the nuanced and dynamic storage of our question data set. The document model's agility is essential, accommodating varying fields and data types without the rigidity of a fixed schema.

Atlas Search

Overview

Atlas Search expands the possibilities of MongoDB Atlas by incorporating sophisticated text search functions into the database. Atlas Search provides a variety of search features that are driven by the industry-standard Apache Lucene search engine.

How we use it

To provide advanced search capabilities in our database, we configured the "vector-search-

question" index. Here's a breakdown of the components:

- *Index Name*: “vector-search-question” - The designated name for the search index, which is used for referencing within search queries.
- *Field Entries*: Each field in the index is designed to serve a specific role in the search process.
- *embed_question* (Vector Field):
    - *Path*: The field in the document where the vector data is stored.
    - *Type*: Set as "vector" to indicate that this field contains vector data for similarity searching.
    - *numDimensions*: Specifies the length of the vector, which in this case is 1153. This represents the feature space for our NLP model.

    - *Similarity*: Uses the cosine similarity measure to evaluate the closeness of the query vector to document vectors.
- *topic_words* (Filter Field):
    - *Path*: Points to the field containing a list of tags in the document.
    - *Type*: Set as "filter" which allows this field to be used for filtering query results based on keyword matching.

The "vector-search-question" index uses the “embed_question” field to get semantically similar questions, converting user searches into vectors and discovering documents that are closely related to the database. Simultaneously, the “topic_words” field refines searches using keyword filters to improve result precision. This index provides both deep semantic matching and basic keyword filtering, resulting in complete and relevant search results.

Conclusion

The NLP chatbot successfully achieved the following outcomes:

1. Topic Modeling: The chatbot performed topic modeling to understand themes and topics of the question, enabling categorization.
2. Intelligent Responses: It provided intelligent responses to user questions adhering to academic limitation, utilizing insights extracted from the NLP pipeline to deliver relevant and informative answers. It also was able to “NOT” provide an answer where it was not trained.
3. Scalability and Efficiency: MongoDB Atlas Search helped us unveil the efficiency in processing large volumes of unstructured data streams, ensuring real-time analysis and response.

The NLP Chatbot offers a revolutionary solution for the education industry seeking to unlock insights from unstructured data available over the internet. By harnessing the capabilities of NLP and machine learning, the chatbot enables efficient analysis, extraction, and interpretation of valuable information, leading to informed customer experiences, and competitive advantages in today's data-driven landscape.

Future - To further enhance the effectiveness and usability of the NLP chatbot, the following recommendations are proposed:

1. Continuous Improvement: Implement a feedback loop mechanism to continuously improve the chatbot's performance based on user feedback and evolving data patterns.
2. Multilingual Support: Extend the chatbot's capabilities to support multiple languages to cater to diverse user demographics and global markets.
3. Integration with Analytics Tools: Integrate the chatbot with advanced analytics and visualization tools to empower users with interactive insights and actionable recommendations.
4. Collaboration with Domain Experts: Collaborate with domain experts and subject matter specialists to fine-tune the chatbot's understanding of industry-specific terminology and contexts.

Appendix

- https://www.kaggle.com/datasets/yapwh1208/chatbot-ai-q-and-a -
- https://www.kaggle.com/datasets/thedevastator/multilingual-conversation-dataset -
- https://www.mongodb.com/developer/products/atlas/articles/
- https://ieeexplore.ieee.org/abstract/document/7375527
- https://dl.acm.org/doi/abs/10.1145/1031171.1031285
- https://www.researchgate.net/publication/334667298_Topic_Modeling_A_Comprehensive_Review