

StudyBuddy

A Question-Answering system for the course Virtual Assistants

Group 4:

Aditya Chauhan - 300321428

Pranshul Dobriyal - 300321165

Project Summary

- Implements a basic question-answering system designed to search a large collection of text documents.
- Extract, clean and chunk textual data from various sources to train models.
- Establish a baseline retrieval augmented QA system utilizing n-grams and TF-IDF to generate contextually relevant answers.
- Establish a retrieval augmented QA system by leveraging semantic information.

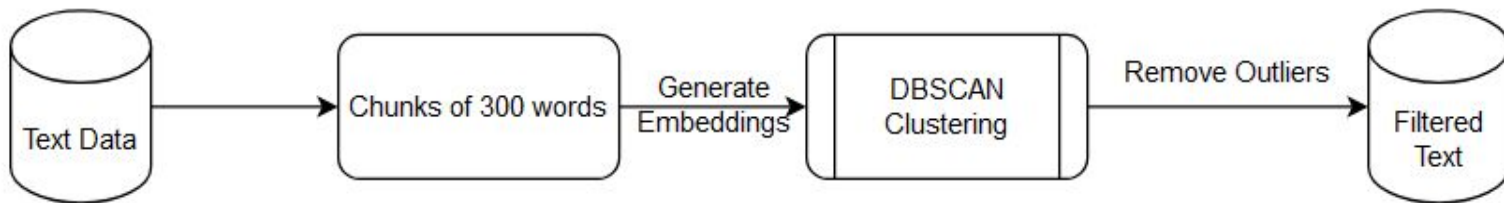
Data Collection

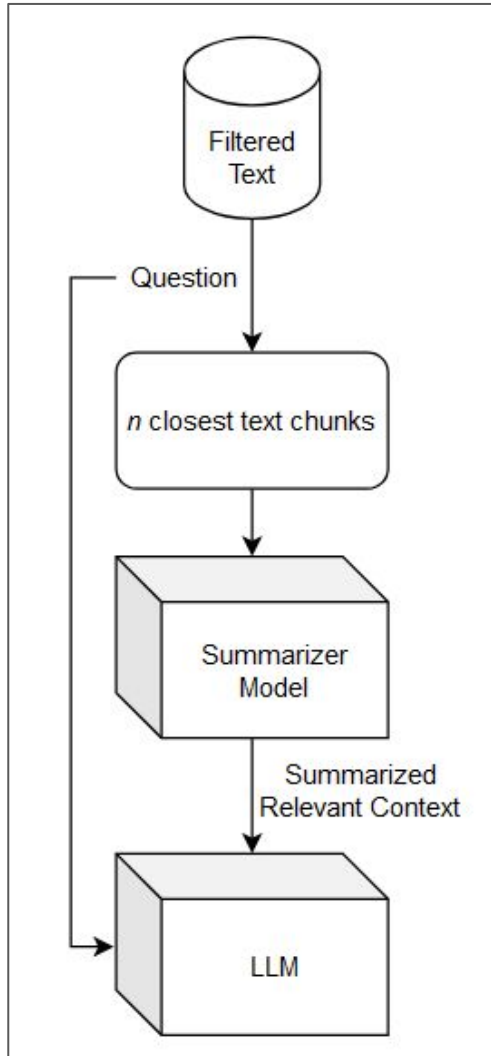
- **Sources:** Combines text data from various sources, including Wikipedia, course PDF files, Web articles, research papers and books, to create a comprehensive knowledge base.
- Create a diverse dataset of text centered around virtual assistants and speech processing by using various sources.
- Language-tool-python

```
text          a virtual assistant va is a software agent tha...
text_length   300
gram_score    20
Name: 0, dtype: object
```

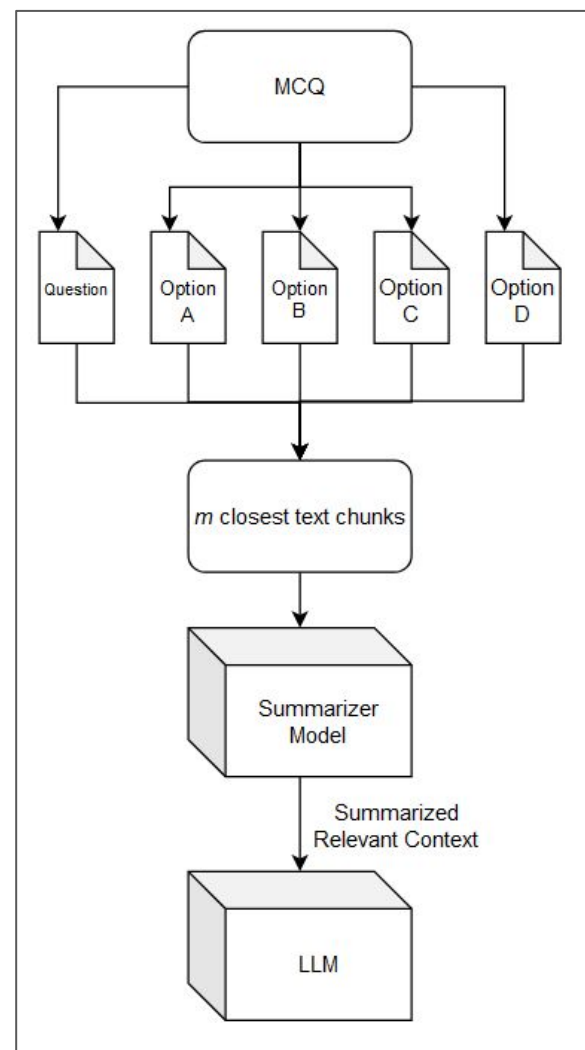
Data Cleaning

- Preprocessing: Applies basic text cleaning steps like removing extra spaces, punctuations, and lemmatization.
- Additionally, remove repeating texts (ex: CSI 5180), non-UTF-8 characters and encoded characters.
- Gram_score/num(words)
- Cleaning Data Further:





RAG Architecture



Methodology (TF-IDF)

- Leverages TF-IDF weighting and n-gram analysis to identify the most relevant terms and phrases, enabling context-aware question answering.
- **N-grams:** Uses bigrams and trigrams to preserve some of the local word order and relationships within the text.
- **TF-IDF:** Algorithm to assign weights to words and n-grams based on their importance within a text chunk and across the entire text corpus.
- **FAISS Indexing:** The TF-IDF vectors representing texts in the corpus are used to build a FAISS index. Efficiently finds the k nearest neighbours (text) based on their L2 distances to the query vector.

Outputs (Baseline Approach: TF-IDF)

Question: Question: MFCCs are a set of ___ coefficients calculated per frame. Fill the blank with the number of coefficients. (a) 39 features (b) 12 features (c) 26 features (d) 13 features

Context Extracted: are used and different phone set use different feature mfccs perceptual linear prediction coefficient lpc and combination of them a result they collected several prediction for the same phone that are later combined in another mlp the best result were achieved by optimizing the number of hidden node and also using information from seven broad class . they are a good example of what they are doing . they are also a great example of how they feel about their work . if you have any questions, please contact us. we need a much denser representation of the acoustic model so we can determine the likelihood of an audio feature vector x given the phone p with mfcc we extract 39 feature from an audio frame let's simplify the picture and assume that there is only one feature for each frame for the state s phone the value of this feature can be modeled with a normal distribution . the following diagram visualizes a bivariate the challenge is to build a system that can discriminate the key feature of the phenomenon without being affected by the other source of signal variability highperformance wellgeneralizing system can be approached with three basic strategy well designed feature robust classifier or end-to-end learning where signal representation features and classifier are jointly learned from the data selection of the approach depends on two primary factors domain knowledge and availability of representative largescale training data fig 3 the basic principle is that well designed or otherwise properly chosen feature the challenge is to build a system that can discriminate the key feature of the phenomenon without being affected by the other source of signal variability highperformance wellgeneralizing system can be approached with three basic strategy well designed feature robust classifier or end-to-end learning where signal representation features and classifier are jointly learned from the data selection of the approach depends on two primary factors domain knowledge and availability of representative largescale training data fig 3 the basic principle is that well designed or otherwise properly chosen feature the challenge is to build a system that can discriminate the key feature of the phenomenon without being affected by the other source of signal variability highperformance wellgeneralizing system can be approached with three basic strategy well designed feature robust classifier or end-to-end learning where signal representation features and classifier are jointly learned from the data selection of the approach depends on two primary factors domain knowledge and availability of representative largescale training data fig 3 the basic principle is that well designed or otherwise properly chosen feature

Time Taken = 99.55172824859619 seconds

System Output:

Answer: a) 39 features

Explanation: MFCCs are a set of 39 features calculated per frame to represent the spectral envelope of the speech signal.

Question: What are transformers?

Context Extracted: transformer block each of which is a multilayer network that map sequence of input vector x_1 into sequence of output vector z_1 of the same length these block are made by combining simple linear layer feedforward network and selfattention layer . transformer selfattention allows a network to directly extract and use information from arbitrarily large context we'll start by describing how self attention work . et al introduced a diffusion transformer dit which facilitates use of the transformer architecture for diffusionbased image production also google released a transformercentric image generator called muse based on parallel decoding and masked generative transformer technology transformer played a lesscentral role with the transformer . xiong et al this is called preln transformer in 2023 . transformer predecessor of attention mechanism was added to gated recurrent .

System output:

Sure, here's a summary of the context about transformers:

Transformers are a type of neural network architecture that has gained significant attention in recent years. They are based on the idea of self-attention, a mechanism that allows a network to directly extract and use information from arbitrarily large context. This allows transformers to achieve state-of-the-art performance on various natural language processing (NLP) tasks, such as machine translation, text summarization, and sentiment analysis.

Time Taken = 131.39818572998047 seconds

=====
=====

Question: How is this course graded?

Context Extracted:

System output:

Sure, here's the answer to your question:

The course is graded based on the following percentages:

- Assignments: 24%
- Student presentations: 10% each
- Final project: 26%
- Final exam: 30%

Time Taken = 75.17999696731567 seconds

Methodology (Semantic Distance)

- Generate Embeddings using a Sentence Transformer (all-MiniLM-L6-v2)
- Cosine Distance measures semantic similarity
- Retrieve text chunks whose embeddings have the least cosine distance from that of the Question (and Options in case of MCQs).
- Text Summarisation: text_summarization, Fine-tuned T5 small model, by Falconsai.

Outputs (Approach 2: Semantic Distance)

Question: What is the dense representation of text in the form of numerical vectors known as ?
Options: ['Word Embeddings ', 'BERT ', 'Latent Embeddings ', 'Pre-trained Embeddings']

Context Extracted: a visualization 24 chapter 6 v ector semantics and embeddings method rohde et al 2006 probably the most common visualization method how ever is to project the 100 dimension of a word down into 2 dimension fig 61 showed one such visualization a doe fig 616 using a projection method called tsne van der maaten and hinton 2008 610 semantic property . the data is coming from immediately nearby word when vector are computed from long context window the highest co the word embedding appears to have described the mapping from the space of sparse count vector to the latent space of svd dense vector in lsa . the word thus originally meant mapping from one space to another it ha metonymically shifted to mean the resulting dense vector . it is in this sense that we currently use the word by the next decade bengio et al 2003 . the word embedding appears to have described the mapping from the space of sparse count vector to the latent space of svd dense vector in lsa . the word thus originally meant mapping from one space to another it ha metonymically shifted to mean the resulting dense vector . it is in this sense that we currently use the word by the next decade bengio et al 2003 . a visualization 24 chapter 6 v ector semantics and embeddings method rohde et al 2006 probably the most common visualization method how ever is to project the 100 dimension of a word down into 2 dimension fig 61 showed one such visualization a doe fig 616 using a projection method called tsne van der maaten and hinton 2008 610 semantic property . the data is coming from immediately nearby word when vector are computed from long context window the highest co bert is the greatest search engine ever able to find the answer to any question we pose it in part 1 of this post notebook i ll explain what it really mean to apply bert to qa and illustrate the details part 2 contains example code we ll be downloading a model that s already finetuned for question answering . bert performs well on text that wasn t in the squad dataset links contents by chris mccormick . Variational autoencoders learn to represent the input just in a compressed form called the latent space or the bottleneck . a variational auto Encoder would construct latent attribute in the following manner . the decoder reconstructing the input can be very easily understood via statistical expression . v 20 chapter 6 v ector semantics and embeddings 1wcaardvark zebra zebra aardvark apricot vv12v target words context noise words 1d figure 613 . the algorithm store two embeddings for each word . wand the context noise embeddings c 682 learning skipgram embeddings the learning algorithm . for each of these wcostraining instance we ll create knegative sample each consist v 20 chapter 6 v ector semantics and embeddings 1wcaardvark zebra zebra aardvark apricot vv12v target words context noise words 1d figure 613 . the algorithm store two embeddings for each word . wand the context noise embeddings c 682 learning skipgram embeddings the learning algorithm . for each of these wcostraining instance we ll create knegative sample each consist fig 1012 show the idea x composite embeddings word position transformer block janet will 12 back 3 janet will back the bill the 4 bill 5 . position embeddings are both of size 1d so their sum is also 1d . approach is to choose a static function that map integer input to real valued vector in a way that capture the inherent relationship among the position that is it capture the fact that position 4 in an input is more closely related to

System output:

***: Word Embeddings

Therefore, the correct answer is ***: Word Embeddings.

Time Taken = 99.18942499160767 seconds

Question: Compare text-dependent and text-independent speaker verification

Context Extracted: speaker recognition is the task of identifying a speaker using their voice recognition is classified into two part speaker identification and speaker verification . speaker identification is the process of determining which voice in a group of known voice best match the speaker speaker verification is computationally le complex than speaker identification system because they require a comparison between only one or two model whereas speaker identification requires comparison of one model to n speaker modelsspeaker verification method are divided into textdependent and textindependent mf ccs12 models hidden markov model are predominantly used type of hmm depends on application fixed phrase word or phrase models prompted phrase utterance model built from phone models gaussian mixture model gmm13 textindependent speaker verification the speaker model is built using speaker adaptation relatively small amount of speech map adaptation from the imposter model14 decision the decision is a 2class hypothesis test H_0 the speaker is automatic speaker verification2 outline introduction speaker identification v speech recognition speech recognition recognitionwordsspeaker identity4 speaker identifier . determine the speaker identity selection between a set of known voices the user does not claim an identity closed set . assume that speaker is not among the speaker known to the system eg prompted phrase password phrase xed phrase knowing the text can improve system performance prompting may reduce risk of jf a modeling defines two distinct space the speaker space defined by the eigenvoice matrix and the channel space represented by the channel matrix . the channel factor estimated using factor analysis a feature extractor that defines only a single space instead of two separate space in this new space a given speech recording is represented by a new vector called total factor a it contains the speaker and channel variability simultaneously speaker recognition based on the ivector framework . this browser is no longer supportedupgrade to microsoft edge to take advantage of the latest feature security update and technical supportspeaker recognition can help determine who is speaking in an audio clip . the service can verify and identify speaker by their unique voice characteristic by using voice biometryyou can then crosscheck audio voice sample against this profile to see if it matches any profile in the group s speaker identificationimportantmicrosoft limit access to speaker recognition you can apply for access through the azure Time Taken = 254.26394295692444 seconds

System Output:

Sure, here's the comparison between text-dependent and text-independent speaker verification:

****Text-Dependent Speaker Verification:****

- * The system uses a pre-recorded database of speech samples from known speakers.
- * The system compares the speech sample to the database and returns the identity of the speaker with the highest similarity score.
- * This method is simple to implement but can be less accurate than text-independent verification.

****Text-Independent Speaker Verification:****

- * The system uses a machine learning algorithm to train a model to distinguish between speakers.
- * The model can be trained on a large dataset of speech samples from known speakers.
- * This method is more accurate than text-dependent verification but can be more complex to implement.

=====

Results

Semantic Retrieval

Context	<i>Relevant</i>	<i>Partially Relevant</i>	<i>Irrelevant</i>
Subjective Questions	Q1, Q2, Q3, Q4, Q6	Q5	N/A
Multiple Choice Questions (10 questions)	Q1, Q2, Q4, Q9	Q3, Q8, Q10	Q5, Q6, Q7

TF-IDF Retrieval

Context	<i>Relevant</i>	<i>Partially Relevant</i>	<i>Irrelevant</i>
Subjective Questions	Q2, Q3, Q4, Q6	Q1, Q5	N/A
Multiple Choice Questions (10 questions)	Q1, Q8	Q2, Q4, Q5, Q7	Q3, Q6, Q9, Q10

Subjective Questions

1. What are transformers?
2. How is this course graded?
3. Who is the professor for this course?
4. Compare text-dependent and text-independent speaker verification.
5. What are the components of a Virtual Assistant?
6. How can we convert audio data into a learnable form?

Activity Table

Total Hours: 56 hours

Activity	Time Spent	Deliverable
Literature Review	8h	Research prior work done in the field; Includes review of the new libraries employed in the project
Data Extraction	8h	Datasets of text extracted from course PDFs
Data Scraping	6h	Datasets of text scraped from articles referenced in the PDFs, Wikipedia articles, and various other sources.
Dataset Creation	6h	Final datasets with cleaned text data in the format: [Text, Length of the text (chunked with 150 words), Grammar Score]
RAG baseline approach	8h	A basic QA system that utilizes n-grams to retain contextual information & TF-IDF to find relevance of a text to the question.
RAG approach 2	8h	A retrieval augmented QA system that utilizes text embeddings to calculate relevance of a text chunk with respect to the question.
Evaluation	6h	Answers generated by both the methods
Presentation	6h	A video presentation of the QA system and its application along with a Pitch presentation.

Challenges

- Lack of computational resources to train Large Language Models (LLMs) on domain-specific data.
- Text extraction from PDFs using Python is tricky.
- Scraping articles from URLs requires several safeguards to tackle forbidden access for restricted content, unresponsive servers, and encoding mismatch.
- Web scraping is prone to noisy data like ads, unrelated headers/footers, and outdated content sections.
- Since the slides generally contain information in the form of images, the text extracted from them is very noisy.

Learnings

- DO NOT underestimate data cleaning.
- A simple QA system can be developed using basic contextual relations from n-grams.

Future Works

- Extend the domain knowledge by incorporating new sources and cleaning techniques,
- Support for speech-to-text and text-to-speech,
- Reimplement with bigger models and new retrieval strategies,
- Develop methods to automatically identify keywords from the lecture slides.

Conclusion

- This project introduces two strategies to conduct retrieval augmented generation/question answering.
- This system can be easily used in conjunction with any other domain knowledge.

References

Github Link

- [1] Slides curated by Professor Caroline Barrière.
- [2] [What Is Retrieval-Augmented Generation, aka RAG?](#)
- [3] [HuggingFace Gemma 2B](#)
- [4] [TF-IDF with Scikit-Learn](#)
- [5] [Sentence Transformers](#)
- [6] [Faiss: A library for efficient similarity search](#)
- [7] [Introduction to Speech Processing, Tom Bäckström et al.](#)
- [8] [Speech and Language Processing. Daniel Jurafsky & James H. Martin.](#)