NATURAL LANGUAGE PROCESSING
AI 829

IIITB

Disease Diagnosis.

Karnadevsinh Zala(MT2023188)

Somesh Awasthi(MT2023172)

Mandate 3

Brief Description

Update on the project so far:

We began by selecting a standard medical book and with the help of PyPDF loader we loaded the document in this project. The preprocessing of the supplied PDF involved several steps: tokenization of every word in the document, converting all words to lowercase words, removal of stopwords such as 'a', 'an', and 'the', lemmatization, and elimination of punctuation. These tasks were performed using NLTK and SpaCy libraries.

Subsequently, we employed langchain_text_splitters to divide lengthy documents into smaller, semantically meaningful chunks. Recognizing the challenge of fitting all words into the context window of our model, we split the text into manageable chunks, gradually combining them into larger segments until reaching a predefined size, with some overlap to maintain context. We set the parameters chunk_size=1000 and chunk_overlap=100.

The chunks were then stored in a database, considering the inefficiency of storing words directly. We utilized chromadb to store vector embeddings and generated these embeddings using OpenAIEmbeddings from langchain.

Additionally, we integrated a prompt template and implemented similarity search in the vector database, enhancing it with a reranking feature to prioritize relevant data. The responses were generated using OpenAI model 3.5. While we also experimented with huggingface embedding models and the lamma2 model, the results did not meet our expectations.

POS Tagging

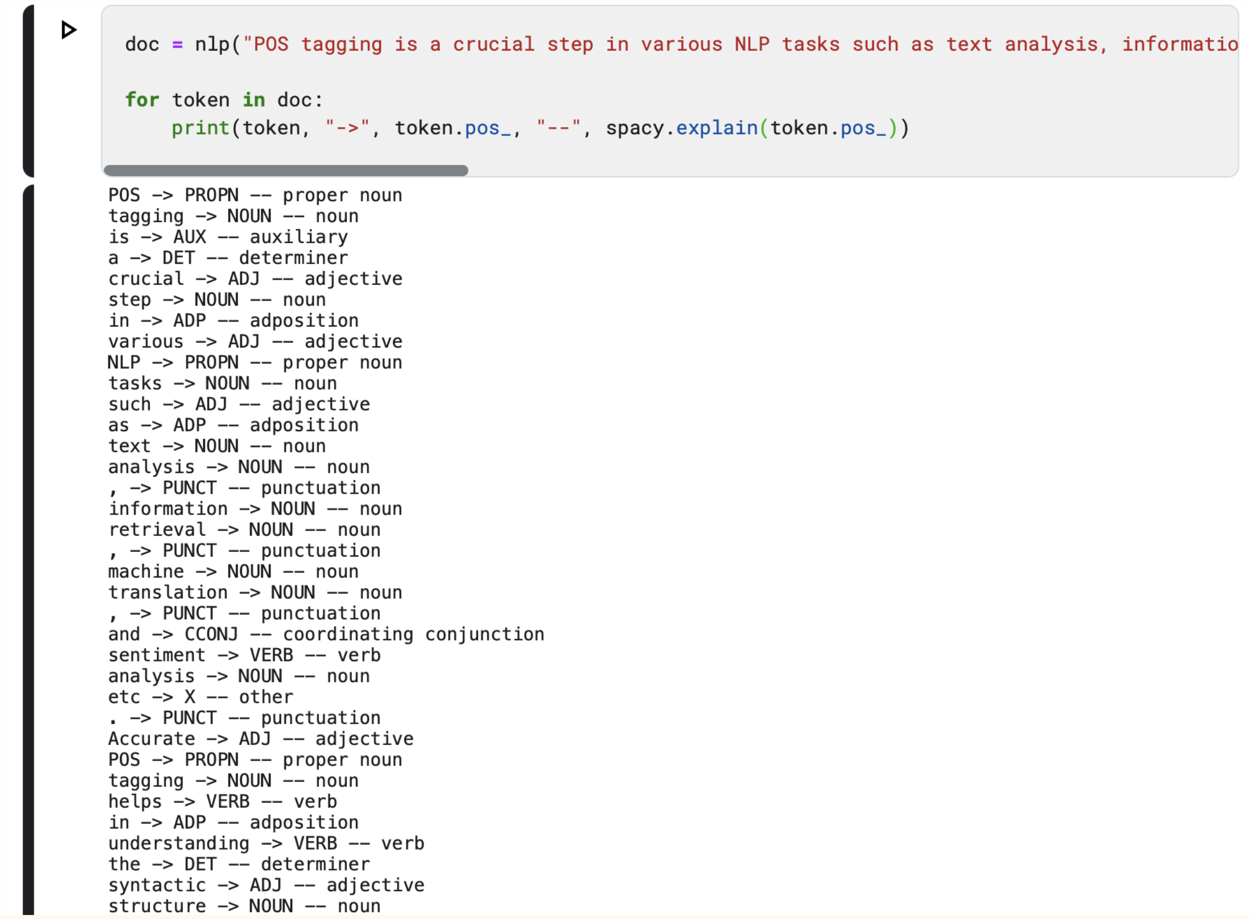
POS tagging, which stands for Part-of-Speech tagging, is a process in natural language processing (NLP) that involves labeling each word in a text with its corresponding part of speech based on its definition and its context in the sentence. The parts of speech include nouns, verbs, adjectives, adverbs, pronouns, conjunctions, prepositions, and interjections.

Some common Parts Of Speech abbreviations & their description:

1. Noun (NN): A word that represents a person, place, thing, or idea. Example: "cat," "Paris," "book."
2. Verb (VB): A word that describes an action, occurrence, or state of being. Example: "run," "eat," "is."
3. Adjective (JJ): A word that describes or modifies a noun. Example: "happy," "blue," "tall."
4. Adverb (RB): A word that describes or modifies a verb, adjective, or another adverb, often indicating manner, frequency, or degree. Example: "quickly," "very," "often."
5. Pronoun (PRP): A word that takes the place of a noun. Example: "he," "she," "it."
6. Conjunction (CC): A word that connects words, phrases, or clauses. Example: "and," "but," "or."
7. Preposition (IN): A word that shows the relationship between a noun or pronoun and other words in the sentence. Example: "in," "on," "under."
8. Interjection (UH): A word or phrase that expresses emotion or sudden feeling. Example: "wow," "oh," "ouch."

POS tagging is a crucial step in various NLP tasks such as text analysis, information retrieval, machine translation, and sentiment analysis. Accurate POS tagging helps in understanding the syntactic structure of sentences, which is essential for many downstream NLP applications.

Practical:



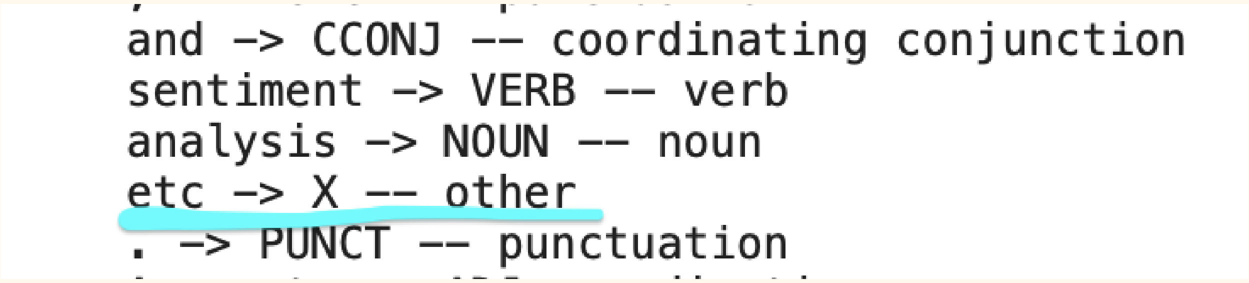
```
doc = nlp("POS tagging is a crucial step in various NLP tasks such as text analysis, informatio

for token in doc:
    print(token, "->", token.pos_, "--", spacy.explain(token.pos_))
```

```
POS -> PROPN -- proper noun
tagging -> NOUN -- noun
is -> AUX -- auxiliary
a -> DET -- determiner
crucial -> ADJ -- adjective
step -> NOUN -- noun
in -> ADP -- adposition
various -> ADJ -- adjective
NLP -> PROPN -- proper noun
tasks -> NOUN -- noun
such -> ADJ -- adjective
as -> ADP -- adposition
text -> NOUN -- noun
analysis -> NOUN -- noun
, -> PUNCT -- punctuation
information -> NOUN -- noun
retrieval -> NOUN -- noun
, -> PUNCT -- punctuation
machine -> NOUN -- noun
translation -> NOUN -- noun
, -> PUNCT -- punctuation
and -> CCONJ -- coordinating conjunction
sentiment -> VERB -- verb
analysis -> NOUN -- noun
etc -> X -- other
. -> PUNCT -- punctuation
Accurate -> ADJ -- adjective
POS -> PROPN -- proper noun
tagging -> NOUN -- noun
helps -> VERB -- verb
in -> ADP -- adposition
understanding -> VERB -- verb
the -> DET -- determiner
syntactic -> ADJ -- adjective
structure -> NOUN -- noun
```

The above code uses pos_ function of spacy library to classify the POS tags.

Now based on our use case we can remove the unwanted words derived from POS analysis.

For example:



```
and -> CCONJ -- coordinating conjunction
sentiment -> VERB -- verb
analysis -> NOUN -- noun
etc -> X -- other
. -> PUNCT -- punctuation
```

We can remove all the tokens to which POS is tagged as X i.e other.

OUR CODE:

```python
def preprocess_text(text):
   # Tokenization and POS tagging using SpaCy
   doc = nlp(text)

   # Filtering out tokens based on POS tags and dependency parsing
   filtered_tokens = []
   for token in doc:
       if token.pos_ not in ["SPACE", "X"]:
           if token.dep_ not in ["det", "punct"]:
               filtered_tokens.append(token.text.lower())

   # Stopword removal
   filtered_tokens = [token for token in filtered_tokens if token not in stopwords.words('english')]

   # Lemmatization
   lemmatized_tokens = [token.lemma_ for token in nlp(" ".join(filtered_tokens))]

   return " ".join(lemmatized_tokens)
```

We are removing the unnecessary tokens which have POS tag of SPACE & OTHER.

Dependency Parsing

Dependency parsing is a technique used in natural language processing (NLP) to analyze the grammatical structure of a sentence by identifying the relationships between words. It focuses on determining the dependency relations between individual words in a sentence and represents these relations as a directed graph, where the nodes correspond to the words and the edges represent the dependencies.

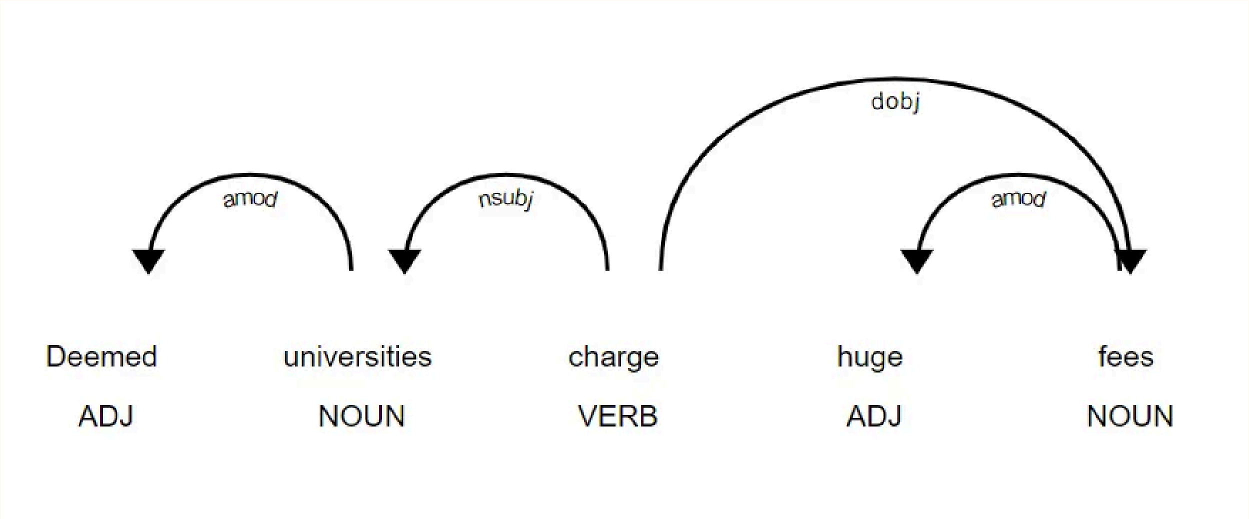
In a dependency parse tree:

- Nodes: Each word in the sentence is represented as a node in the graph.
- Edges: The edges between nodes represent the syntactic relationships or dependencies between the words. The direction of the edge usually points from the dependent word to its head or governor word.

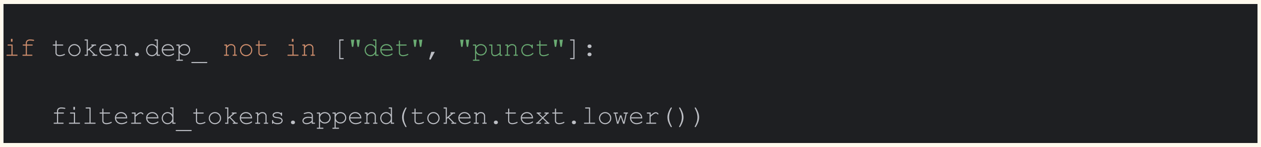
Some common dependency relations:

1. nsubj: Nominal subject - The noun phrase that is the subject of the clause.
2. dobj: Direct object - The noun phrase that is the object of the verb.
3. iobj: Indirect object - The noun phrase that is the recipient of the action.
4. amod: Adjectival modifier - An adjective that modifies a noun.
5. advmod: Adverbial modifier - An adverb that modifies a verb, adjective, or another adverb.
6. conj: Conjunct - A word or phrase that is connected to another word or phrase by a coordinating conjunction like "and," "or."

Dependency parsing is beneficial because it provides a more detailed and accurate representation of the syntactic structure of a sentence compared to other parsing techniques like constituency parsing.



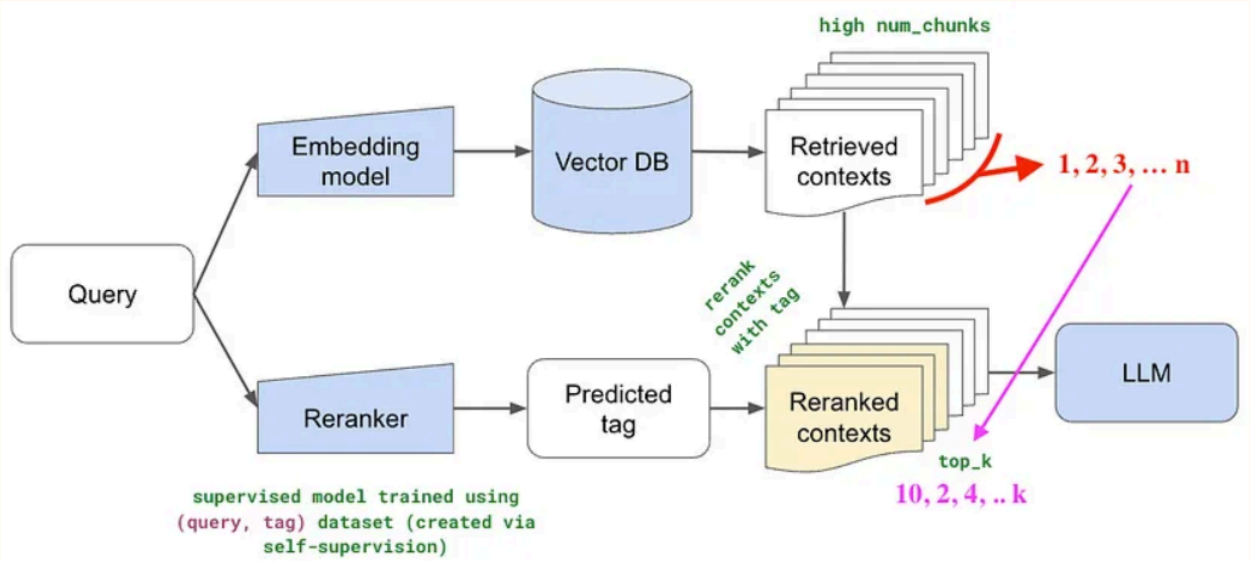
OUR CODE:



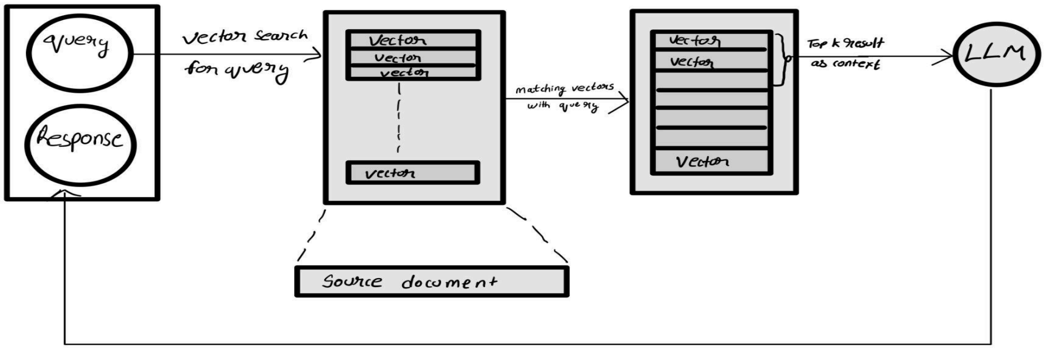
```
if token.dep_ not in ["det", "punct"]:

   filtered_tokens.append(token.text.lower())
```

Again we are using Spacy library for implementing Dependency parsing. Basically we are finding relations for Determinants & Punctuations and after finding them we are removing those tokens as they are not necessary & we can trim the data.

Long-Context Reorder (Re-Ranking)

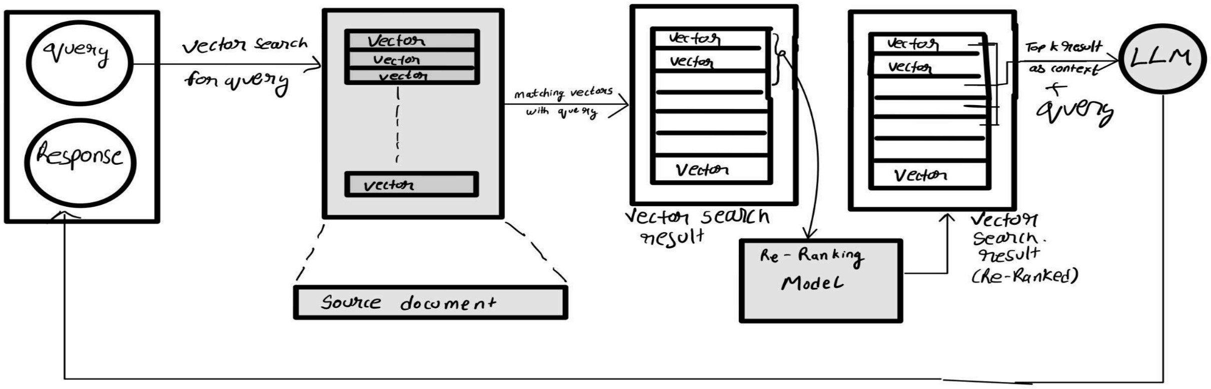


Reranking in RAG aims to enhance the relevance of responses generated by large language models (LLMs) by providing additional context along with a query. While setting up a basic RAG system is relatively simple, it often fails to deliver highly accurate responses because it may not offer precise context to the LLM.



RAG With Vector Search

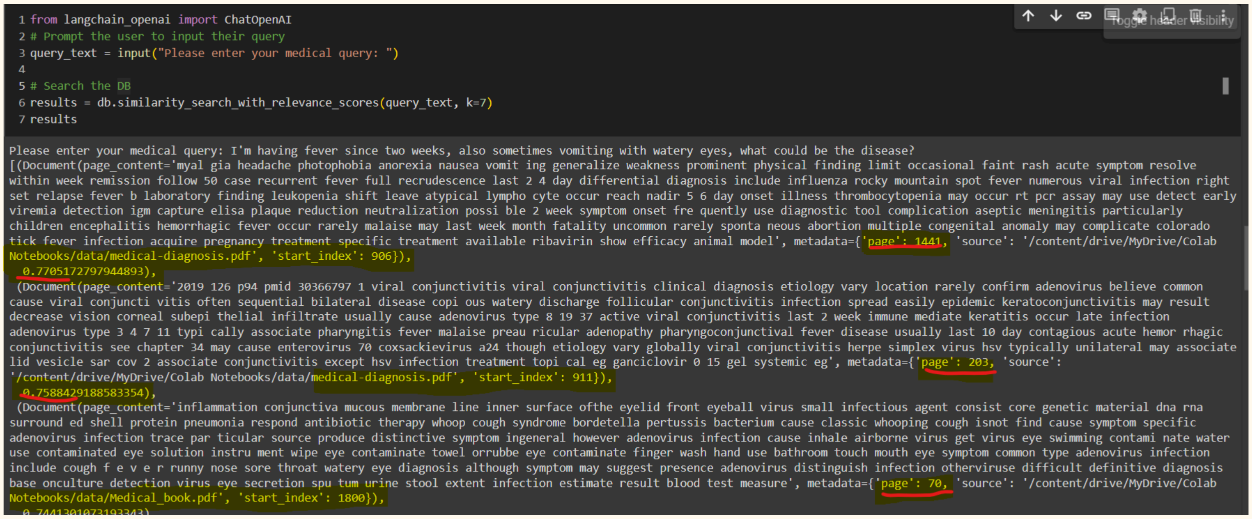
The current approach typically involves passing only the top_k responses from the vector search to the LLM as context. However, this may overlook other returned vectors containing more relevant information related to the query, resulting in less accurate responses from the LLM. To address this issue, a crucial step in refining RAG implementation involves re-ranking.



RAG with vector search and re-ranking

A re-ranking model evaluates matching scores for query-document pairs, allowing for the rearrangement of vector search results to prioritize the most relevant ones at the top of the list.

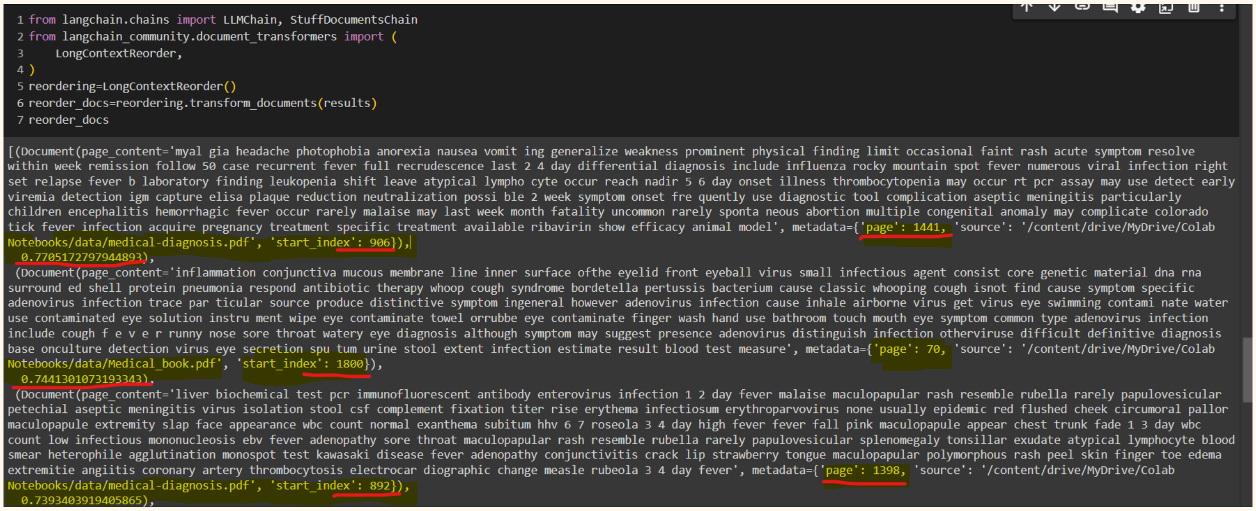
Code Snippet



```
from langchain_openai import ChatOpenAI
# Prompt the user to input their query
query_text = input("Please enter your medical query: ")

# Search the DB
results = db.similarity_search_with_relevance_scores(query_text, k=7)
results
```

```
Please enter your medical query: I'm having fever since two weeks, also sometimes vomiting with watery eyes, what could be the disease?
[(Document(page_content='myal gia headache photophobia anorexia nausea vomit ing generalize weakness prominent physical finding limit occasional faint rash acute symptom resolve
within week remission follow 50 case recurrent fever full recrudescence last 2 4 day differential diagnosis include influenza rocky mountain spot fever numerous viral infection right
set relapse fever b laboratory finding leukopenia shift leave atypical lympho cyte occur reach nadir 5 6 day onset illness thrombocytopenia may occur rt pcr assay may use detect early
viremia detection igm capture elisa plaque reduction neutralization possi ble 2 week symptom onset fre quently use diagnostic tool complication aseptic meningitis particularly
children encephalitis hemorrhagic fever occur rarely malaise may last week month fatality uncommon rarely sponta neous abortion multiple congenital anomaly may complicate colorado
tick fever infection acquire pregnancy treatment specific treatment available ribavirin show efficacy animal model', metadata={'page': 1441, 'source': '/content/drive/MyDrive/Colab
Notebooks/data/medical-diagnosis.pdf', 'start_index': 906}),
  0.7705172797944893),
 (Document(page_content='2019 126 p94 pmid 30366797 1 viral conjunctivitis viral conjunctivitis clinical diagnosis etiology vary location rarely confirm adenovirus believe common
cause viral conjuncti vitis often sequential bilateral disease copi ous watery discharge follicular conjunctivitis infection spread easily epidemic keratoconjunctivitis may result
decrease vision corneal subepi thelial infiltrate usually cause adenovirus type 8 19 37 active viral conjunctivitis last 2 week immune mediate keratitis occur late infection
adenovirus type 3 4 7 11 typi cally associate pharyngitis fever malaise preau ricular adenopathy pharyngoconjunctival fever disease usually last 10 day contagious acute hemor rhagic
conjunctivitis see chapter 34 may cause enterovirus 70 coxsackievirus a24 though etiology vary globally viral conjunctivitis herpe simplex virus hsv typically unilateral may associate
lid vesicle sar cov 2 associate conjunctivitis except hsv infection treatment topi cal eg ganciclovir 0 15 gel systemic eg', metadata={'page': 203, 'source':
'/content/drive/MyDrive/Colab Notebooks/data/medical-diagnosis.pdf', 'start_index': 911}),
  0.7588429188583354),
 (Document(page_content='inflammation conjunctiva mucous membrane line inner surface ofthe eyelid front eyeball virus small infectious agent consist core genetic material dna rna
surround ed shell protein pneumonia respond antibiotic therapy whoop cough syndrome bordetella pertussis bacterium cause classic whooping cough isnot find cause symptom specific
adenovirus infection trace par ticular source produce distinctive symptom ingeneral however adenovirus infection cause inhale airborne virus get virus eye swimming contami nate water
use contaminated eye solution instru ment wipe eye contaminate towel orrubbe eye contaminate finger wash hand use bathroom touch mouth eye symptom common type adenovirus infection
include cough f e v e r runny nose sore throat watery eye diagnosis although symptom may suggest presence adenovirus distinguish infection otherviruse difficult definitive diagnosis
base onculture detection virus eye secretion spu tum urine stool extent infection estimate result blood test measure', metadata={'page': 70, 'source': '/content/drive/MyDrive/Colab
Notebooks/data/Medical_book.pdf', 'start_index': 1800}),
  0.7441301073193343),
```

Output Without re-ranking



```
from langchain.chains import LLMChain, StuffDocumentsChain
from langchain_community.document_transformers import (
    LongContextReorder,
)
reordering=LongContextReorder()
reorder_docs=reordering.transform_documents(results)
reorder_docs
```

```
[(Document(page_content='myal gia headache photophobia anorexia nausea vomit ing generalize weakness prominent physical finding limit occasional faint rash acute symptom resolve
within week remission follow 50 case recurrent fever full recrudescence last 2 4 day differential diagnosis include influenza rocky mountain spot fever numerous viral infection right
set relapse fever b laboratory finding leukopenia shift leave atypical lympho cyte occur reach nadir 5 6 day onset illness thrombocytopenia may occur rt pcr assay may use detect early
viremia detection igm capture elisa plaque reduction neutralization possi ble 2 week symptom onset fre quently use diagnostic tool complication aseptic meningitis particularly
children encephalitis hemorrhagic fever occur rarely malaise may last week month fatality uncommon rarely sponta neous abortion multiple congenital anomaly may complicate colorado
tick fever infection acquire pregnancy treatment specific treatment available ribavirin show efficacy animal model', metadata={'page': 1441, 'source': '/content/drive/MyDrive/Colab
Notebooks/data/medical-diagnosis.pdf', 'start_index': 906}),
  0.7705172797944893),
 (Document(page_content='inflammation conjunctiva mucous membrane line inner surface ofthe eyelid front eyeball virus small infectious agent consist core genetic material dna rna
surround ed shell protein pneumonia respond antibiotic therapy whoop cough syndrome bordetella pertussis bacterium cause classic whooping cough isnot find cause symptom specific
adenovirus infection trace par ticular source produce distinctive symptom ingeneral however adenovirus infection cause inhale airborne virus get virus eye swimming contami nate water
use contaminated eye solution instru ment wipe eye contaminate towel orrubbe eye contaminate finger wash hand use bathroom touch mouth eye symptom common type adenovirus infection
include cough f e v e r runny nose sore throat watery eye diagnosis although symptom may suggest presence adenovirus distinguish infection otherviruse difficult definitive diagnosis
base onculture detection virus eye secretion spu tum urine stool extent infection estimate result blood test measure', metadata={'page': 70, 'source': '/content/drive/MyDrive/Colab
Notebooks/data/Medical_book.pdf', 'start_index': 1800}),
  0.7441301073193343),
 (Document(page_content='liver biochemical test pcr immunofluorescent antibody enterovirus infection 1 2 day fever malaise maculopapular rash resemble rubella rarely papulovesicular
petechial aseptic meningitis virus isolation stool csf complement fixation titer rise erythema infectiosum erythroparvovirus none usually epidemic red flushed cheek circumoral pallor
maculopapule extremity slap face appearance wbc count normal exanthema subitum hhv 6 7 roseola 3 4 day high fever fever fall pink maculopapule appear chest trunk fade 1 3 day wbc
count low infectious mononucleosis ebv fever adenopathy sore throat maculopapular rash resemble rubella rarely papulovesicular splenomegaly tonsillar exudate atypical lymphocyte blood
smear heterophile agglutination monospot test kawasaki disease fever adenopathy conjunctivitis crack lip strawberry tongue maculopapular polymorphous rash peel skin finger toe edema
extremitie angiitis coronary artery thrombocytosis electrocar diographic change measle rubeola 3 4 day fever', metadata={'page': 1398, 'source': '/content/drive/MyDrive/Colab
Notebooks/data/medical-diagnosis.pdf', 'start_index': 892}),
  0.7393403919405865),
```

Output With re-ranking

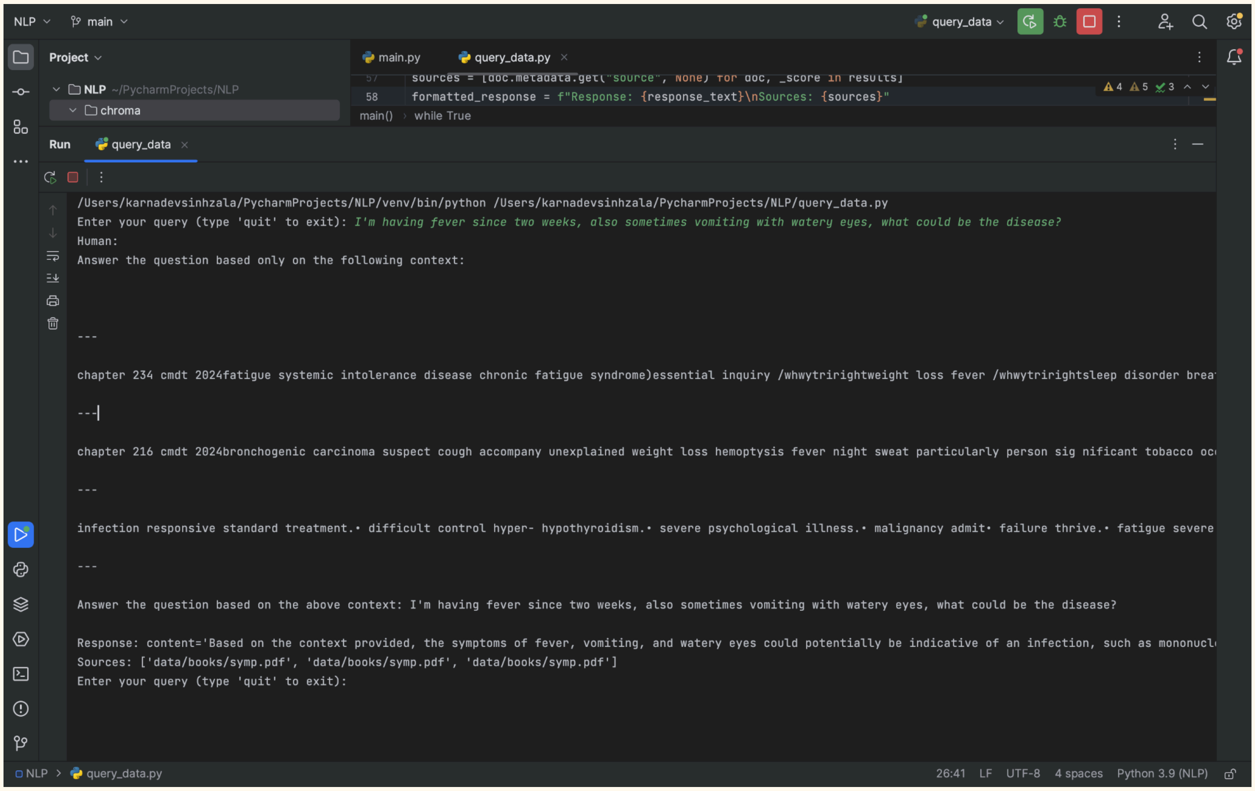
Persisting Context

Previously we were just providing a single shot of query to model but now we are giving option to ask multiple queries & to facilitate the quality of answer, we're appending the context everytime when a new query is fired.

```python
def main():
    # Prepare the DB.
    embedding_function = OpenAIEmbeddings()
    db = Chroma(persist_directory=CHROMA_PATH, embedding_function=embedding_function)
    context_text = ""
    while True:
        query_text = input("Enter your query (type 'quit' to exit): ")

        if query_text.lower() == 'quit':
            break

        # Search the DB.
        results = db.similarity_search_with_relevance_scores(query_text, k=3)
        if len(results) == 0 or results[0][1] < 0.5:
            print(f"Unable to find matching results.")
            return

        new_context_text = "\n\n---\n\n".join([doc.page_content for doc, _score in results])
        context_text += "\n\n---\n\n" + new_context_text

        prompt_template = ChatPromptTemplate.from_template(PROMPT_TEMPLATE)

        prompt = prompt_template.format(context=context_text,
question=query_text)

        print(prompt)


        model = ChatOpenAI()

        response_text = model.invoke(prompt)


        # Load the model

        sources = [doc.metadata.get("source", None) for doc, _score in results]

        formatted_response = f"Response: {response_text}\nSources: {sources}"

        print(formatted_response)
```

Here context_text is appended everytime for the new query.

Final Output



```
sources = [doc.metadata.get("source", None) for doc, _score in results]
formatted_response = f"Response: {response_text}\nSources: {sources}"
```

```
/Users/karnadevsinhzala/PycharmProjects/NLP/venv/bin/python /Users/karnadevsinhzala/PycharmProjects/NLP/query_data.py
Enter your query (type 'quit' to exit): I'm having fever since two weeks, also sometimes vomiting with watery eyes, what could be the disease?
Human:
Answer the question based only on the following context:



---

chapter 234 cmdt 2024fatigue systemic intolerance disease chronic fatigue syndrome)essential inquiry /whwytrirightweight loss fever /whwytrirightsleep disorder brea

---

chapter 216 cmdt 2024bronchogenic carcinoma suspect cough accompany unexplained weight loss hemoptysis fever night sweat particularly person sig nificant tobacco oc

---

infection responsive standard treatment.• difficult control hyper- hypothyroidism.• severe psychological illness.• malignancy admit• failure thrive.• fatigue severe

---

Answer the question based on the above context: I'm having fever since two weeks, also sometimes vomiting with watery eyes, what could be the disease?

Response: content='Based on the context provided, the symptoms of fever, vomiting, and watery eyes could potentially be indicative of an infection, such as mononucl
Sources: ['data/books/symp.pdf', 'data/books/symp.pdf', 'data/books/symp.pdf']
Enter your query (type 'quit' to exit):
```

Answer the question based on the above context: I'm having fever since two weeks, also sometimes vomiting with watery eyes, what could be the disease?

Output: Response: content='Based on the context provided, the symptoms of fever, vomiting, and watery eyes could potentially be indicative of an infection, such as mononucleosis or sinusitis. It is important to seek medical attention for a proper diagnosis and treatment.'

Sources: ['data/books/symp.pdf']

Reference book for creating vector database

1. Dr Pedagogy One Touch Medicine
2. PJ Mehta’s Practical Medicine
3. Harrison's Principles of Internal Medicine, 21e

Plan for Mandate-4

Developing the UI(Front-End) section of the project.

Also to improve accuracy, we will try to combine multiple models together to yield output.

Github link: click here
NoteBook link: click here

REFERENCES

1. DAY-12 | End to End Medical Chatbot Project | Part -1
2. LangChain official website
3. Part Of Speech POS Tagging: NLP Tutorial For Beginners - S1 E11
4. Dependency Parsing (Towards Data Science)
5. RAG + Langchain Python Project: Easy AI/Chat For Your Docs