

به نام خدا



دانشگاه تهران

دانشکده فنی

دانشکده مهندسی برق و کامپیوتر



درس پردازش زبان طبیعی

پاسخ تمرین ۶

نام و نام خانودگی: زهرا ریحانیان

شماره دانشجویی: ۸۱۰۱۰۱۱۷۷

تیر ماه ۱۴۰۳

۳ پاسخ سوال اول
۳ پاسخ بخش اول - دریافت و آماده سازی دادگان
۴ پاسخ بخش دوم - تولید بازنمایی و پایگاه داده برداری
۴ پاسخ بخش سوم - پیاده سازی بازیاب ترکیبی
۵ پاسخ بخش چهارم - پیاده سازی Router chain
۶ پاسخ بخش پنجم - پیاده سازی Search Engine Chain
۷ پاسخ بخش ششم - پیاده سازی Relevancy Check Chain
۸ پاسخ بخش هفتم - پیاده سازی Fallback Chain
۹ پاسخ بخش هشتم - پیاده سازی Generate With Context Chain
۱۰ پاسخ بخش نهم - آماده سازی گراف با استفاده از Langgraph

پاسخ سوال اول

کد مربوط به این بخش در مسیر `codes/chatbot.ipynb` موجود است.

پاسخ بخش اول – دریافت و آماده سازی دادگان

الف) آدرسی که باید از آن لینک pdf ها پیدا شود در یک متغیر ذخیره می کنیم. یک `request` به آن می زنیم و سپس با استفاده از کتابخانه `BeautifulSoup` تمام تگ های `a` در این صفحه که شامل لینک های موجود هستند را پیدا می کنیم و از بین آن ها، هر کدام که با `pdf` تمام شوند را به لیست `pdflinks` اضافه می کنیم. (آن `pdf` ای که تمام فصل ها را دارد حذف می کنیم)

ب) در اینجا با استفاده از `PyPdfLoader` تک تک `pdf` هایی را که در لیست `pdflinks` آدرس شان را ذخیره کرده بودیم، لود می کنیم و محتوای هر کدام از آن ها را به لیست `documents` اضافه می کنیم.

پ) در اینجا با استفاده از `RecursiveCharacterTextSplitter` و مقادیر پیشنهاد شده برای `chunk_size` و `chunk_overlap` اسناد به اندازه های مناسب تبدیل شد.

ت) این تقسیم کننده متن (`text splitter`) با لیستی از کاراکترها پارامتر می شود. سعی می کند به ترتیب روی آنها تقسیم شود تا اینکه تکه ها به اندازه کافی کوچک شوند. فهرست پیش فرض `["\n\n", "\n", " ", ""]` است. نتیجه ی این، تلاش برای نگه داشتن تمام پاراگراف ها (و سپس جملات، و سپس کلمات) در کنار هم تا زمانی که ممکن باشد، است، زیرا به طور کلی به نظر می رسد که آنها قوی ترین قطعات متنی مرتبط از نظر معنایی باشند. `chunk_overlap` تعداد کاراکترهایی که بین تکه های متوالی همپوشانی دارند و `chunk_size` حداکثر تعداد کاراکتر در هر قطعه است.

نحوه تقسیم متن: بر اساس فهرست کاراکترها.

نحوه اندازه گیری اندازه تکه (`chunk size`): با تعداد کاراکترها.

تقسیم متن به قطعات کوچکتر به چند دلیل ضروری است:

۱. اسناد بزرگ می توانند حافظه فشرده داشته باشند. شکستن آنها به قطعات کوچکتر به مدیریت موثرتر حافظه کمک می کند.
۲. بسیاری از مدل های پردازش زبان طبیعی (NLP) محدودیتی در مقدار متنی دارند که می توانند همزمان پردازش کنند. قطعه قطعه شدن متن را در این محدوده ها قرار می دهد.
۳. همپوشانی بین تکه ها به حفظ زمینه کمک می کند که برای درک و پردازش دقیق متن بسیار مهم است.

ث) اگر «chunk_size» خیلی بزرگ باشد، ممکن است از محدودیت‌های پردازش مدل‌های NLP فراتر رود که منجر به خطا یا ناکارآمدی شود. اگر «chunk_size» خیلی کوچک باشد، ممکن است تکه‌ها خیلی تکه‌شوند که باعث از بین رفتن متن و پردازش معنادار متن می‌شود.

اگر «chunk_overlap» خیلی بزرگ باشد، افزونگی قابل توجهی بین تکه‌ها ایجاد می‌شود که می‌تواند منجر به پردازش ناکارآمد و افزایش استفاده از حافظه شود. اگر «chunk_overlap» خیلی کوچک باشد، ممکن است اطلاعات زمینه‌ای مهم بین تکه‌ها از بین برود که منجر به عدم دقت احتمالی در پردازش متن می‌شود.

انتخاب مقادیر مناسب برای «chunk_size» و «chunk_overlap» برای متعادل کردن کارایی و دقت بسیار مهم است. یک رویکرد معمولی آزمایش با مقادیر مختلف و ارزیابی عملکرد برای یافتن تنظیمات بهینه برای یک برنامه خاص است.

پاسخ بخش دوم – تولید بازنمایی و پایگاه داده برداری

الف) در این قسمت store برای ذخیره سازی بازنمایی‌ها روی local تعریف شد. برای مدل جانمایی از HuggingFaceEmbeddings استفاده شد. سپس این مقادیر به همراه namespace مدل به CacheBackedEmbeddings متد from_bytes_store داده شد. در نهایت این embedder به همراه chunked_documents تولید شده در قسمت قبل، به FAISS داده شد و vectorstore ساخته شد.

ب) استفاده از embedder مناسب در NLP برای به دست آوردن ظرایف نحوی، معنایی و واژگانی منحصر به فرد یک زبان خاص، حصول اطمینان از نمایش دقیق و معنادار جملات بسیار مهم است. اگر برای جاسازی جملات فارسی از مدلی که بر روی داده‌های فارسی آموزش ندیده است، استفاده کنیم، با چندین مشکل مواجه خواهیم شد:

۱. مدل دستور زبان فارسی، واژگان یا بافت فرهنگی را درک نمی‌کند که منجر به جاسازی‌های نادرست می‌شود.
۲. مدل ممکن است متن فارسی را به اشتباه نشانه گذاری کند یا کلمات را بر اساس داده‌های آموزشی خود به اشتباه تفسیر کند.
۳. این مدل عبارات و اصطلاحات فارسی منحصر به فرد را به تصویر نمی‌کشد و جنبه‌های مهم معنا را از دست می‌دهد.
۴. برنامه‌های پایین دستی مانند تجزیه و تحلیل احساسات و ترجمه از جاسازی‌های نادرست رنج می‌برند و خوب آموزش داده نمی‌شوند.

به طور خلاصه، استفاده از یک مدل آموزش داده شده بر روی زبان مقصد برای جاسازی جملات دقیق و معنی دار ضروری است.

پاسخ بخش سوم – پیاده سازی بازیاب ترکیبی

الف) - بازیاب واژگانی: بر روی تطبیق دقیق کلمات تمرکز می کند، سریعتر و ساده تر است، اما فاقد درک زمینه و معنی است. برای برنامه هایی که نیاز به جستجوی کلیدواژه ساده دارند، بهترین است.

- بازیاب معنایی: معنا و زمینه را درک می کند، پیچیده تر و نیازمند منابع است، اما نتایج مرتبط تر و آگاهانه تری ارائه می دهد. این برای موتورهای جستجوی پیشرفته، سیستم های توصیه و برنامه های NLP ایده آل است.

ب) ابتدا دو بازیاب خواسته شده تعریف شد سپس با استفاده از EnsembleRetriever بازیاب ترکیبی ساخته شد.

پ) برای این قسمت، دو مقدار ۰ و ۱ را برای هر یک از بازیاب ها امتحان کردم. یعنی یک بار بازیاب FAISS وزن ۰ و BM25 وزن ۱ و برعکس. در حالتی که وزن بازیاب BM25 برابر ۱ است، اسنادی که برگردانده شد جواب خوبی را برای سوال های پرسیده شده نداشتند و از لحاظ مرتبط بودن، وقتی که بازیاب FAISS وزن برابر ۱ داشت، اسناد بهتر و معنادار تری برگردانده شد.

ت) در این مرحله وزن های 0.3 به بازیاب BM25 و 0.7 به بازیاب FAISS داده شد. همچنین وزن های 0.2 به بازیاب BM25 و 0.8 به بازیاب FAISS داده شد که جواب ها فرقی نداشتند و همین انتخاب شد. سوالات زیر به بازیاب داده شد:

What is self-attention?

What is binary search tree?

Who is the president of Bolivia?

با بررسی جملات، مشاهده شد که برای سوال ها به خصوص برای سوال اول، جواب خوب و مرتبطی تولید شده است.

پاسخ بخش چهارم - پیاده سازی Router chain

الف) در این قسمت Router chain پیاده سازی شد. به همین منظور ابتدا به سایت معرفی شده مراجعه کردم و بعد از دریافت API Key آن را در محیط os ذخیره کردم.

ب) مدل خواسته شده با temperature=0 تعریف شد.

پ) در این مرحله، پرامپت زیر را نوشتم و از ChatPromptTemplate برای وصل کردن به chain و دادن پرامپت به مدل، استفاده کردم:

You are an intelligent assistant with the ability to classify user queries. Your task is to determine whether a given query is related to Natural Language Processing (NLP), Computer Science (CS), or neither of them. Use the following guidelines to make your classification:

Natural Language Processing (NLP): Queries related to understanding, generating, or processing human language using computational methods. This includes topics like sentiment analysis, language translation, text generation, speech recognition, and other linguistics-based computational tasks.

Computer Science (CS): Queries related to general computer science topics excluding NLP. This includes areas such as algorithms, data structures, programming languages, software engineering, databases, networking, artificial intelligence (excluding NLP-specific topics), computer architecture, and cybersecurity.

If the query is related to NLP, just return the string VectorStore.

If the query is related to CS, just return the string SearchEngine

If the query is related to neither NLP nor CS, just return the string None.

{output_instructions}

query: {query}

ت) در اینجا parser خواسته شده پیاده سازی شد تا خروجی مدل یکی از مقادیر None, VectorStore, SearchEngine باشد. در نهایت هم chain ساخته شد. کد این قسمت:

```
class ChosenTool(BaseModel):
    tool_name: Literal['None', "VectorStore", "SearchEngine"] =
Field(description="the tool that was chosen by LLM in question routing
stage")

router_chain_parser = PydanticOutputParser(pydantic_object=ChosenTool)
router_chain = prompt | llm | router_chain_parser
```

ث) در اینجا مدل باید یک جواب قطعی و در واقع یک fact ای را بیان کند و نیاز به خلاقیت ندارد. یعنی اگر prompt ای مربوط به VectorStore باشد، همیشه خروجی باید همین باشد نه چیز دیگری. برای همین temperature را 0 قرار می دهیم.

پاسخ بخش پنجم – پیاده سازی Search Engine Chain

الف) انجام شد.

ب) TavilySearchResults از کتابخانه langchain_community.tools به عنوان موتور جستجو تعریف شد و کوئری "what is AI?" برای امتحان و دیدن خروجی آن داده شد.

```
[41] from langchain_community.tools.tavily_search import TavilySearchResults

search = TavilySearchResults(max_results=5)
await search.ainvoke("what is AI?")

[{'url': 'https://www.mediapost.com/publications/article/392343/cis-has-plenty-of-smart-ases-this-year-it-will-t.html',
'content': 'Incubated by a team of AI experts at Stagwell's U.K.-based locaria agency, which they developed the generative AI platform capable of automatically analyzing, categorizing and managing creative advertising assets based on how well they perform -- or don't. advertisement advertisement "AI is a quantum leap for creativity\and\productivity that will drive a new age of value creation for modern business. Username Password Forgot? Become a free MediaPost member now to read this article log in if you are already a member Username Password Forgot? Username Password Remember me Forgot your password? Subscribe today to gain access to every Research Intelligence article we publish as well as the exclusive daily newsletter, full access to The MediaPost Cases, first-look research and daily insights from Joe Mandese, Editor in Chief. "Whether you're a global player or a nascent challenger,\and\developing an approach to the three E's of AI\and\is the best way to stay ahead in this year of competition, and there's no marketing network boasting the breadth of full-service digital transformation and self-service AI products like SmartAssets to help clients adapt. Dubbed "SmartAssets," the startup was the winner early last year of an internal "Shark tank"-like Stagwell competition, as well as the beneficiary of a grant of up to $1 million to get things rolling. '),
'url': 'https://www.wlox.com/2024/06/24/mccc-partners-with-main-workforce-ai-courses/',
'content': '"But if you use Amazon if you use any sort of application on your phone AI has been in the background now we're just taking it out of the background and bringing it into the hands of the end user the individuals day-to-day. Dr. Napier also says that while many people may worry about the use of AI, it has actually been around since the 1950s and only popularized in the last decade or so. (WLOX) - Mississippi Gulf Coast Community College and the Mississippi Artificial Intelligence Network (MAIN) are partnering to meet the growing demand for AI skills in Mississippi's workforce. You know, the bias and everything that's involved in how we can use it to enhance our skills, but not replace our judgment or our creativity." Because it will be part of their lives now and into the future," said Dr. Kollin Napier, an MCCC employee and Director of MAIN. '),
'url': 'https://21stcenturychronicle.com/ai-friend-or-foe/',
'content': '(Newsfeature) Related Posts Hesa mission lines up to 'touch the Sun' NCC to unveil strategic plan for telecom sector in January F6 launches 5-year roadmap on data protection, expects over N125billion revenue Coy begins campaign to curb human trafficking Philanthropist launches hub to train 1,700 youths on ICT annually CO28 president hits back at climate denial claims Recent Posts Archives Categories Meta @ "Advancement in AI technology has come to stay and it is becoming an essential tool for problem-solving and decision-making." AI, as of today, is at the heart of many technologies, including smart devices and voice assistants such as Siri on Apple devices", Tinuola Popoola, a System Analyst and Cyber Safety Advocate, told the News Agency of Nigeria (NAN). By Jessica Dogo Artificial Intelligence (AI) has been around since the 1950s, but its recent boom has caused exciting wave of interest, as it becomes more accessible to the public with increasing roles in man's daily activities. Created at a Dartmouth Conference of 1956, AI started as a field study by scientists. Mr Jide Awe, a Science, Technology and Innovation (STI) Policy Advisor and Founder, Jidaw.com, said that AI had the potential to be a powerful tool for Nigeria to tackle insecurity and other development challenges. "AI's capability to enhance the analysis of network traffic patterns enables it to quickly identify unusual activities, which in turn assists in promptly detecting and responding to cyber threats. '),
'url': 'https://muckatingjournal.com/news/state-and-regional/illinois/tech-companies-want-to-build-artificial-general-intelligence-but-who-decides-when-agi-is-attained/article_cfd0e206-34c0-5bfc-a52d-a1fcf877828.html',
'content': '"The surging interest in women's college basketball prompted the NCAA to double down on its investment last summer by backing the inaugural Wom. Klaus Mäkelä, just 28,
```

شکل ۱ خروجی TavilySearchResults

همان طور که میبینید یک لیستی از دیکشنری هاست که هر کدام شامل url و content است.

ت) برای این که حداکثر ۵ سند مرتبط را بازبایی کند، max_results=5 قرار دادم. برای این که فرمت خروجی به فرمت خواسته شده باشد، یک تابع نوشتیم که خروجی موتور جستجو را میگیرد و به ازای هر یک از اعضای لیست، آن را تبدیل به یک Document می کند. این Document ها را به یک لیست جدید اضافه کردم و برگرداندم که خروجی این chain استاندارد باشد.

خروجی بعد از اعمال این parser این شد:

```
[ ] Document(metadata={'url': 'https://www.ibm.com/topics/artificial-intelligence'}, page_content='But since the advent of electronic computing (and relative to some of the topics discussed in this article) important events and milestones in the evolution of artificial intelligence include the following:\nPut AI to work in your business with IBM's industry-leading AI expertise and portfolio of solutions at your side.\nIn it, they delve into four potential goals or definitions of AI, which differentiates computer systems on the basis of rationality and thinking vs. acting;\nHuman approach:\nIdeal approach:\nAlan Turing's definition would have fallen under the category of "systems that act like humans."\nIBM acquires Manta to complement data and AI governance capabilities;\nIBM Watsonx Orchestrate;\nIBM Watsonx Assistant;\nExplore Gen AI learning for developers;\nWeak AI-also called Narrow AI or Artificial Narrow Intelligence (ANI)-is AI trained and focused to perform specific tasks. While a number of definitions of artificial intelligence (AI) have surfaced over the last few decades, John McCarthy offers the following definition in this 2004\paper\and\resides outside ibm.com), " It is the science and engineering of making intelligent machines, especially intelligent computer programs. Artificial general intelligence (AGI), or general AI, is a theoretical form of AI where a machine would have an intelligence equaled to humans; it would have a self-aware consciousness that has the ability to solve problems, learn, and plan for the future.').
Document(metadata={'url': 'https://www.coursera.org/articles/what-is-artificial-intelligence'}, page_content='Learners are advised to conduct additional research to ensure that courses and other credentials pursued meet their personal, professional, and financial goals.\n1 unlocks unlimited opportunities;\nCoursera Footer;\nPopular AI Content;\nPopular Programs;\nPopular Skills;\nPopular Career Resources;\nCoursera;\nCommunity;\nMore Yet, despite the many philosophical disagreements over whether "true" intelligent machines actually exist, when most people use the term AI today, they're referring to a suite of machine learning-powered technologies, such as Chat GPT or computer vision, that enable machines to perform tasks that previously only humans can do like generating written content, steering a car, or analyzing data.\nFor Everyone course, you'll learn what AI can realistically do and not do, how to spot opportunities to apply AI to problems in your own organization, and what it feels like to build machine learning and data science projects.\nRegardless of how far we are from achieving AGI, you can assume that when someone uses the term artificial general intelligence, they're referring to the kind of sentient computer programs and machines that are commonly found in popular science fiction.\nSome of the most common examples of AI in use today include:\nchatGPT: Uses large language models (LLMs) to generate text in response to questions or comments posed to it.\n').
Document(metadata={'url': 'https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence'}, page_content='These include the launch of Apple's Siri and Amazon's Alexa voice assistants; IBM Watson's victories on Jeopardy; self-driving cars; the development of the first generative adversarial network; the launch of TensorFlow, Google's open source deep learning framework; the founding of research lab OpenAI, developers of the GPT-3 language model and Dall-E image generator; the defeat of world go champion Lee Sedol by Google DeepMind's AlphaGo; and the implementation of AI-based systems that detect cancers with a high degree of accuracy.\n2020s. Throughout the centuries, thinkers from Aristotle to the 13th century Spanish theologian Ramon Llull to René Descartes and Thomas Bayes used the tools and logic of their times to describe human thought processes as symbols, laying the foundation for AI concepts such as general knowledge representation.\nAs the airline giant moves more of its data workloads to the cloud, tools from Intel's Granulate are making platforms such as ...;\nThe vendor's new platform, now in beta testing, combines its existing lakehouse with AI to better enable users to manage and ...;\nIn summary, AI's ethical challenges include the following:\nAI governance and regulations;\nDespite potential risks, there are currently few regulations governing the use of AI tools, and where laws do exist, they typically pertain to AI indirectly. The rapid evolution of AI technologies is another obstacle to forming meaningful regulation of AI, as are the challenges presented by AI's lack of transparency that make it difficult to see how the algorithms reach their results.').
Document(metadata={'url': 'https://meng.wisc.edu/news-stories/ai-artificial-intelligence-what-is-the-definition-of-ai-and-how-does-ai-work/'), page_content='Dec 21, 2023 · Artificial Intelligence (AI) works by simulating human intelligence through the use of algorithms, data, and computational power. The goal is to\and...').
Document(metadata={'url': 'https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-ai'}, page_content='Sven Blumberg is a senior partner in McKinsey's Düsseldorf office; Michael Chui is a partner at the McKinsey Global Institute and is based in the Bay Area office, where Iareina Yee is a senior partner; Kia Javanmardian is a senior partner in the Chicago office, where Alex Singla, the global leader of QuantumBlack, AI by McKinsey, is also a senior partner; Kate Smaje and Alex Sukharevsky are senior partners in the London office.\nWhat's more, the specific areas in which companies see value from AI have evolved, from manufacturing and risk to these;\nand one set of companies continues to pull ahead of
```

شکل ۲ خروجی TavilySearchResults بعد از اعمال parser

پاسخ بخش ششم – پیاده سازی Relevancy Check Chain

الف) پرامپت زیر را برای این قسمت نوشتم:

You are expert in checking relevancy. Given a query and a document and you have the task to check whether or not the document is relevant to the query.

If the document and query was related to each other just return relevant otherwise just return irrelevant.

{output_instructions}

document: {document}

query: {query}

ب) از همان مدل تعریف شده در قسمت ۴ استفاده شد.

پ) در این جا کلاس CheckQuery را نوشتم که خروجی مدل relevant یا irrelevant باشد. آن را به PydanticOutputParser دادم. در نهایت chain خواسته شده ساخته شد.

```
class CheckQuery(BaseModel):
    relevancy: Literal["relevant", "irrelevant"] = Field(description="This specifies whether the user's query is relevant or irrelevant")

relevancy_check_parser = PydanticOutputParser(pydantic_object=CheckQuery)
relevancy_check_chain = prompt | llm | relevancy_check_parser
```

ت) برای این که ممکن است هر یک از vector store یا SearchEngine ممکن است فایل هایی نامرتبط برگردانند یا این که همان اول مسیری اشتباه آمده (بین VectoStore و SearchEngine و Fallback) و در اینجا یک بار دیگر ارتباط سند با پرسجو چک می شود و در صورت مرتبط نبودن هر سند برگردانده شده آن را از اسناد ذخیره شده در state حذف می کند. اگر این chain نباشد، خطا ممکن است منتشر شود ولی با وجود آن، دقت بالاتر می رود و خطا کمتر خواهد شد.

پاسخ بخش هفتم – پیاده سازی Fallback Chain

الف) پرامپی که برای این قسمت ساختم:

You are a responsive chatbot in the field of Computer Science (CS) with special expertise in Natural Language Processing (NLP). You are very helpfully.

You only answer questions that are related to NLP or CS. Generate the answer considering your limitations.

If the question is not related to NLP or CS, apologize and explain that these questions are not in your field of expertise.

Finally, politely ask him to ask questions only about NLP or CS.

Current conversations:\n\n{chat_history}\n\n

human: {query}

(ب) در این جا از مدل با `temperature=0.1` استفاده کردم.

(پ) در این مرحله برای سابقه‌ی چت، آن را به فرمت مناسب، همانند آن چه که در کارگاه پیاده سازی شده بود، در آوردم و `chain` را به صورت زیر کامل کردم:

```
fallback_chain = (
    {
        "chat_history": lambda x: "\n".join(
            [
                (
                    f"human: {msg.content}"
                    if isinstance(msg, HumanMessage)
                    else f"AI: {msg.content}"
                )
                for msg in x["chat_history"]
            ]
        ),
        "query": itemgetter("query") ,
    }
    | fallback_prompt
    | llm
    | StrOutputParser()
)
```

پاسخ بخش هشتم – پیاده سازی `Generate With Context Chain`

(الف) پرامپت زیر را برای این بخش نوشتم:

You are a helpful chatbot. You are given a query and a context and you must generate the answer using the given context. Just generate answer based on the given context and not more. If your context was not enough just explain your knowledge is limit and can not answer the query.

context: {context}\n\n

query: {query}

ب) از همان مدل با `temperature=0.1` استفاده کردم.

پ) در نهایت زنجیر به صورت زیر ساخته شد:

```
generate_with_context_chain = generate_with_context_prompt | llm |  
StrOutputParser()
```

پاسخ بخش نهم – آماده سازی گراف با استفاده از Langgraph

در این بخش ابتدا state گراف به همان صورت داده شده، تعریف شد. گره ها را همان طور که خواسته شده بود، تعریف کردم. در اینجا روند کار گراف را توضیح خواهم داد.

ابتدا `router_node` به عنوان گره ی ورودی عمل خواهد کرد. کوئری کاربر را میگیرد و به `router_chain` می دهد. خروجی آن سه حالت ممکن دارد:

اگر `SearchEngine` باشد به گره ی `search_engine`، اگر `VectorStore` باشد به گره ی `vector_store` و در نهایت اگر `None` باشد به گره ی `fallback` خواهد رفت.

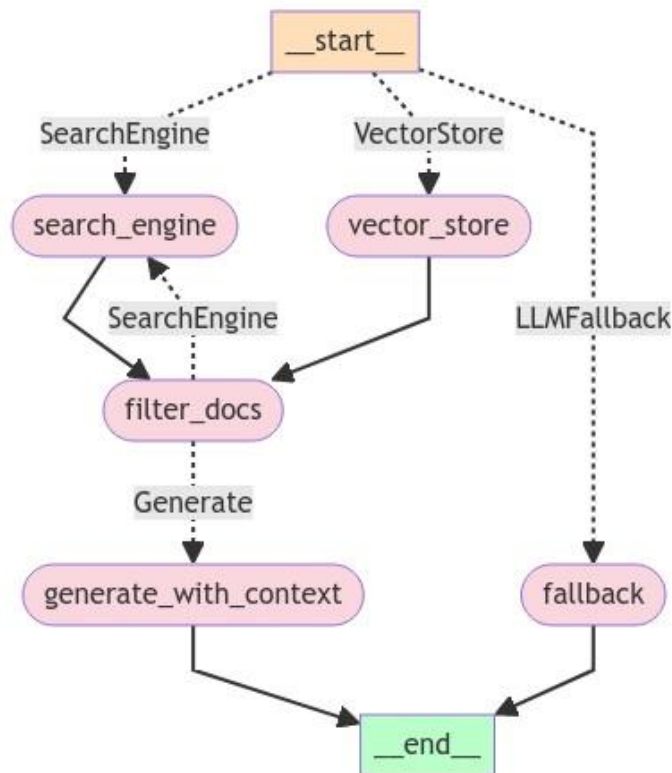
گره ی `fallback` کوئری کاربر و `chat_history` را می گیرد و به `fallback_chain` می دهد. پاسخ تولید شده را در `generation state` می گذارد و به `end` می رود.

گره ی `vector_store` کوئری کاربر را به `retriever_chain` می دهد و اسناد برگردانده شده را در `documents state` قرار می دهد. بعد از آن به گره ی `filter_docs` می رود. این گره، از `state` کوئری و `documents` را می گیرد و یک لیست جدید می سازد که شامل اسناد مرتبط با کوئری است که برای این کار، تک تک سند ها در `documents` را توسط `relevancy_check_chain` با کوئری چک می کند و اگر مرتبط بودند، آن ها را به این لیست جدید اضافه می کند. برای مقدار `State` هم این لیست جدید را قرار می دهد. در یال مربوط به آن، در تابع `check_documents` خالی بودن یا نبودن `documents` را چک می کند. اگر خالی بود، `Generate` و گرنه `SearchEngine` را برمی گرداند. اگر `Generate` برگرداند، به گره ی `generate_with_context` می رود و گرنه به `search_engine` می رود.

گره ی `search_engine` کوئری کاربر را می گیرد و در یک تابع داخلی `async` جواب `search_engine_chain` را برمی گرداند. این جواب را در `documents state` ذخیره می کند.

گره ی `generate_with_context` از `state` کوئری و `documents` را می گیرد و با استفاده از `generate_with_context_chain` جواب نهایی را تولید می کند، در `generation` می گذارد و به گره ی نهایی `end` می رود.

خروجی گراف به صورت زیر بدست آمد:



شکل ۳ گراف نهایی

در اینجا چند نمونه از سوال هایی که از چت بات کردم و جواب هایی که دریافت کردم را می گذارم:

Q: what are the binary search trees?

A: A Binary Search Tree is a data structure used in computer science for organizing and storing data in a sorted manner. Each node in a Binary Search Tree has at most two children, a left child and a right child, with the left child containing values less than the parent node and the right child containing values greater than the parent node.

Q: Hi I'm Zahra!

A: Hi Zahra! Nice to meet you! I'm a chatbot with expertise in Computer Science, particularly in Natural Language Processing (NLP). I'm here to help answer any questions you may have in these areas. What would you like to know or discuss?

Q: what are transformers?

A: According to the context, transformers are a neural architecture that can handle distant information. They are made up of stacks of transformer blocks, each of which is a multilayer network that maps sequences of input vectors to sequences of output vectors of the same length. These blocks are made by combining simple linear layers, feedforward networks, and self-attention layers, the key self-attention innovation of transformers. Self-attention allows a network to directly extract and use information from arbitrarily large contexts.

Q: what is chatbot?

A: According to the context, a chatbot is a conversational system designed to mimic the appearance of informal human conversation. It is a type of conversational agent that can carry on longer and more unstructured conversations, similar to human-human interaction.

Q: why sky is blue?

A: I'm happy to help! However, I must apologize that the question "why sky is blue?" is not within my field of expertise, which is specifically focused on Computer Science (CS) and Natural Language Processing (NLP).

As a chatbot, I'm designed to provide assistance and answer questions related to CS and NLP, but I'm not equipped to provide explanations on topics outside of these domains, such as physics or atmospheric science.

If you have any questions related to NLP or CS, I'd be more than happy to help! For example, you could ask me about topics like language models, machine learning, algorithms, or programming languages. Please feel free to ask, and I'll do my best to provide a helpful response.

Q: Tell me some programming languages.

A: Based on the provided context, some programming languages mentioned are:

1. Python

2.JavaScript

3.C++

4.Rust

5.HTML/CSS