RAG using the Nebula graph and llamaindex

1 Knowledge Graph RAG Query Engine

1.1 Graph RAG

Graph RAG is an Knowledge-enabled RAG approach to retrieve information from Knowledge Graph on given task. Typically, this is to build context based on entities' SubGraph related to the task.

2 Download the Required Dependencies

[ ]:
```
%pip install llama-index-llms-azure-openai
%pip install llama-index-graph-stores-nebula
%pip install llama-index-llms-openai
%pip install llama-index-embeddings-azure-openai
%pip install llama-index
%pip install llama-index-readers-wikipedia
```

[ ]:
```
import logging
import sys

logging.basicConfig(
    stream=sys.stdout, level=logging.INFO
)  # logging.DEBUG for more verbose output
```

3 Configure openai Model

[ ]:
```
# For OpenAI

import os

os.environ["OPENAI_API_KEY"] = "Enter your Api Key"

# define LLM
from llama_index.llms.openai import OpenAI
from llama_index.core import Settings
```

```
Settings.llm = OpenAI(temperature=0, model="gpt-3.5-turbo")
Settings.chunk_size = 512
```

4 Configure Azure openai model

[ ]:
```
from llama_index.llms.azure_openai import AzureOpenAI
from llama_index.embeddings.azure_openai import AzureOpenAIEmbedding

# For Azure OpenAI
api_key = "Enter your api key"
azure_endpoint = "https://<your-resource-name>.openai.azure.com/"
api_version = "2023-07-01-preview"

llm = AzureOpenAI(
    model="gpt-35-turbo-16k",
    deployment_name="my-custom-llm",
    api_key=api_key,
    azure_endpoint=azure_endpoint,
    api_version=api_version,
)

# You need to deploy your own embedding model as well as your own chat␣
↪completion model
embed_model = AzureOpenAIEmbedding(
    model="text-embedding-ada-002",
    deployment_name="my-custom-embedding",
    api_key=api_key,
    azure_endpoint=azure_endpoint,
    api_version=api_version,
)
```

[ ]:
```
from llama_index.core import Settings

Settings.llm = llm
Settings.embed_model = embed_model
Settings.chunk_size = 512
```

5 load data from Wikipedia for Guardians of the Galaxy Vol. 3

[ ]:
```
from llama_index.core import download_loader

from llama_index.readers.wikipedia import WikipediaReader

loader = WikipediaReader()
```

```
documents = loader.load_data(
    pages=["Guardians of the Galaxy Vol. 3"], auto_suggest=False
)
```

6 Configure Nebula Graph for store

[ ]:
```
import os
os.environ["NEBULA_USER"] = "root"
os.environ["NEBULA_PASSWORD"] = "nebula"
os.environ[
    "NEBULA_ADDRESS"
] = "127.0.0.1:9669"  # default port for NebulaGraph

space_name = "llamaindex"
edge_types, rel_prop_names = ["relationship"], [
    "relationship"
]  # default, could be omit if create from an empty kg
tags = ["entity"]  # default, could be omit if create from an empty kg
```

7 Store the data

[ ]:
```
from llama_index.core import StorageContext
from llama_index.graph_stores.nebula import NebulaGraphStore

graph_store = NebulaGraphStore(
    space_name=space_name,
    edge_types=edge_types,
    rel_prop_names=rel_prop_names,
    tags=tags,
)
storage_context = StorageContext.from_defaults(graph_store=graph_store)
```

7.1 Perform Graph RAG Query

Finally, let's demo how to do Graph RAG towards an existing Knowledge Graph.

All we need to do is to use RetrieverQueryEngine and configure the retriver of it to be KnowledgeGraphRAGRetriever.

The KnowledgeGraphRAGRetriever performs the following steps:

- Search related Entities of the quesion/task
- Get SubGraph of those Entities (default 2-depth) from the KG
- Build Context based on the SubGraph

Please note, the way to Search related Entities could be either Keyword extraction based or Embedding based, which is controlled by argument retriever_mode of the

KnowledgeGraphRAGRetriever, and supported options are: - “keyword” - “embedding”(not yet implemented) - “keyword_embedding”(not yet implemented)

Here is the example on how to use RetrieverQueryEngine and KnowledgeGraphRAGRetriever:

[ ]:
```
from llama_index.core.query_engine import RetrieverQueryEngine
from llama_index.core.retrievers import KnowledgeGraphRAGRetriever

graph_rag_retriever = KnowledgeGraphRAGRetriever(
    storage_context=storage_context,
    verbose=True,
)

query_engine = RetrieverQueryEngine.from_args(
    graph_rag_retriever,
)
```

Then we can query it like:

[ ]:
```
from IPython.display import display, Markdown

response = query_engine.query(
    "Tell me about Peter Quill?",
)
display(Markdown(f"<b>{response}</b>"))
```

Entities processed: ['Star', 'Lord', 'Marvel', 'Quill', 'Galaxy', 'Guardians', 'Guardians of the Galaxy', 'Star-Lord', 'Peter Quill', 'Peter']

Entities processed: ['Star', 'Lord', 'Marvel', 'Quill', 'Galaxy', 'Guardians', 'Guardians of the Galaxy', 'Star-Lord', 'Peter Quill', 'Peter']

Graph RAG context:

The following are knowledge sequence in max depth 2 in the form of `subject predicate, object, predicate_next_hop, object_next_hop ...` extracted based on key entities as subject:

Guardians, is member of, Guardians, was experimented on, by the High Evolutionary

Guardians, is member of, Guardians, considered to tell, origins

Guardians, is member of, Guardians, origins, team-up movie

Guardians, is member of, Guardians, befriended, his fellow Batch 89 test subjects

Guardians, is member of, Guardians, sought to enhance and anthropomorphize animal lifeforms, to create an ideal society

Guardians, is member of, Guardians, is creator of, Rocket

Guardians, is member of, Guardians, is, Mantis

Guardians, is member of, Guardians, is half-sister of, Mantis

Guardians, is member of, Guardians, is, Kraglin

Guardians, is member of, Guardians, developed psionic abilities, after being abandoned in outer space

Guardians, is member of, Guardians, would portray, Cosmo

Guardians, is member of, Guardians, recalls, his past

Guardians, is member of, Guardians

Guardians, is member of, Guardians, focus on, third Guardians-centric film

Guardians, is member of, Guardians, is, Rocket

Guardians, is member of, Guardians, backstory, flashbacks

Guardians, is member of, Guardians, is former second-in-command of, Ravagers

Quill, is half-sister of, Mantis, is member of, Guardians

Quill, is half-sister of, Mantis, is, Mantis

Quill, is in a state of depression, following the appearance of a variant of his dead lover Gamora

Quill, is half-sister of, Mantis

Peter Quill, is leader of, Guardians of the Galaxy, is sequel to, Guardians of

Peter Quill is the leader of the Guardians of the Galaxy and the main protagonist of the Guardians of the Galaxy films. He was raised by a group of alien thieves and smugglers, and was abducted from Earth as a child. He is half-human, half-Celestial, and has the ability to wield an energy weapon called the Infinity Stone. He is set to return to the MCU in May 2021.

[ ]:
```
response = await query_engine.aquery(
    "Tell me about Peter Quill?",
)
display(Markdown(f"<b>{response}</b>"))
```

INFO:openai:message='OpenAI API response'
path=https://api.openai.com/v1/completions processing_ms=611
request_id=1c07a89e18f19ac7bbc508507c2902d9 response_code=200
Entities processed: ['Star', 'Lord', 'Marvel', 'Quill', 'Galaxy',

'Guardians', 'Guardians of the Galaxy', 'Star-Lord', 'Peter Quill', 'Peter']

INFO:openai:message='OpenAI API response'
path=https://api.openai.com/v1/completions processing_ms=992
request_id=6517cb63da3364acd33e816a9b3ee242 response_code=200

Entities processed: ['Star', 'Lord', 'Marvel', 'Quill', 'Galaxy', 'Guardians', 'Guardians of the Galaxy', 'Star-Lord', 'Peter Quill', 'Peter']

Graph RAG context:

The following are knowledge sequence in max depth 2 in the form of `subject predicate, object, predicate_next_hop, object_next_hop ...` extracted based on key entities as subject:

Guardians, is member of, Guardians, was experimented on, by the High Evolutionary

Guardians, is member of, Guardians, considered to tell, origins

Guardians, is member of, Guardians, origins, team-up movie

Guardians, is member of, Guardians, befriended, his fellow Batch 89 test subjects

Guardians, is member of, Guardians, sought to enhance and anthropomorphize animal lifeforms, to create an ideal society

Guardians, is member of, Guardians, is creator of, Rocket

Guardians, is member of, Guardians, is, Mantis

Guardians, is member of, Guardians, is half-sister of, Mantis

Guardians, is member of, Guardians, is, Kraglin

Guardians, is member of, Guardians, developed psionic abilities, after being abandoned in outer space

Guardians, is member of, Guardians, would portray, Cosmo

Guardians, is member of, Guardians, recalls, his past

Guardians, is member of, Guardians

Guardians, is member of, Guardians, focus on, third Guardians-centric film

Guardians, is member of, Guardians, is, Rocket

Guardians, is member of, Guardians, backstory, flashbacks

Guardians, is member of, Guardians, is former second-in-command of, Ravagers

Quill, is half-sister of, Mantis, is member of, Guardians

Quill, is half-sister of, Mantis, is, Mantis

Quill, is in a state of depression, following the appearance of a variant of his dead lover Gamora

Quill, is half-sister of, Mantis

Peter Quill, is leader of, Guardians of the Galaxy, is sequel to, Guardians of the Galaxy

Peter Quill, was raised by, a group of alien thieves and smugglers

Peter Quill, would return to the MCU, May 2021

path=https://api.openai.com/v1/completions processing_ms=2384
request_id=b5a7e601affa751fbc7f957f3359a238 response_code=200

Peter Quill is the leader of the Guardians of the Galaxy and the main protagonist of the Guardians of the Galaxy films. He was raised by a group of alien thieves and smugglers, and was abducted from Earth as a child. He is half-human, half-Celestial, and has the ability to wield an energy weapon called the Infinity Stone. He is set to return to the MCU in May 2021.

7.2 Include nl2graphquery as Context in Graph RAG

The nature of (Sub)Graph RAG and nl2graphquery are different. No one is better than the other but just when one fits more in certain type of questions. To understand more on how they differ from the other, see this demo comparing the two.

While in real world cases, we may not always know which approach works better, thus, one way to best leverage KG in RAG are fetching both retrieval results as context and letting LLM + Prompt generate answer with them all being involved.

So, optionally, we could choose to synthesise answer from two piece of retrieved context from KG: - Graph RAG, the default retrieval method, which extracts subgraph that's related to the key entities in the question. - NL2GraphQuery, generate Knowledge Graph Query based on query and the Schema of the Knowledge Graph, which is by default switched off.

We could set with_nl2graphquery=True to enable it like:

[ ]:
```
graph_rag_retriever_with_nl2graphquery = KnowledgeGraphRAGRetriever(
    storage_context=storage_context,
    verbose=True,
    with_nl2graphquery=True,
)

query_engine_with_nl2graphquery = RetrieverQueryEngine.from_args(
    graph_rag_retriever_with_nl2graphquery,
)
```

[ ]:
```
response = query_engine_with_nl2graphquery.query(
    "What do you know about Peter Quill?",
)
display(Markdown(f"<b>{response}</b>"))
```

Graph Store Query:

```
MATCH (p:`entity`)-[:`relationship`]->(m:`entity`) WHERE p.`entity`.`name` == 'Peter Quill'
RETURN m.`entity`.`name`;
```

Graph Store Response:

{'m.entity.name': ['May 2021', 'as a child', 'Guardians of the Galaxy', 'a group of alien thieves and smugglers', 'half-Celestial']}

Entities processed: ['Star', 'Lord', 'Marvel', 'Quill', 'Galaxy', 'Guardians', 'Guardians of the Galaxy', 'Star-Lord', 'Peter Quill', 'Peter']

Entities processed: ['Star', 'Lord', 'Marvel', 'Quill', 'Galaxy', 'Guardians', 'Guardians of the Galaxy', 'Star-Lord', 'Peter Quill', 'Peter']

Graph RAG context:

The following are knowledge sequence in max depth 2 in the form of `subject predicate, object, predicate_next_hop, object_next_hop ...` extracted based on key entities as subject:

Guardians, is member of, Guardians, was experimented on, by the High Evolutionary

Guardians, is member of, Guardians, considered to tell, origins

Guardians, is member of, Guardians, origins, team-up movie

Guardians, is member of, Guardians, befriended, his fellow Batch 89 test subjects

Guardians, is member of, Guardians, sought to enhance and anthropomorphize animal lifeforms, to create an ideal society

Guardians, is member of, Guardians, is creator of, Rocket

Guardians, is member of, Guardians, is, Mantis

Guardians, is member of, Guardians, is half-sister of, Mantis

Guardians, is member of, Guardians, is, Kraglin

Guardians, is member of, Guardians, developed psionic abilities, after being abandoned in outer space

Guardians, is member of, Guardians, would portray, Cosmo

Guardians, is member of, Guardians, recalls, his past

Guardians, is member of, Guardians

Peter Quill is the leader of the Guardians of the Galaxy and was abducted from Earth as a child. He is half-human and half-Celestial, and was raised by a group of alien thieves and smugglers. He would return to the MCU in May 2021.