From Nodes to Knowledge

GraphRAG for Everyone

Sethu Pavan Venkata Reddy Pastula

About Me

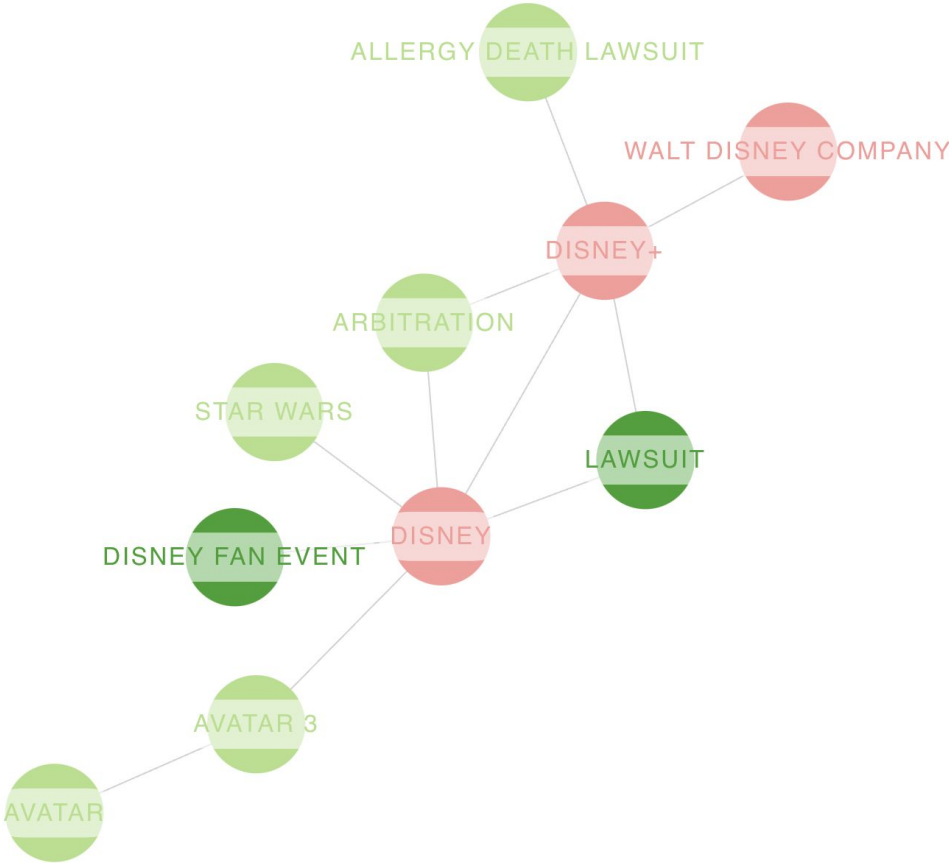
- Software Engineer at Microsoft
  - Production bugs to unblock companies from deploying Microsoft software to millions.
- Thesis on AI for Fashion ~~Deep Generative Networks for Virtual Try-On (VTON)~~ .
  - Trained a GAN model to generate given clothing item on a target person's body realistically accounting for occlusions.
  - Fine Tuned SD2 on Fashion dataset to generate text-guided VTON images.
- Software Engineer at STFC, UKRI
  - Built a bridge that allows you to run MATLAB functions in a C++ software and vice-versa.

Agenda

- Provide an intuition of GraphRAG
  - What is GraphRAG?
  - Why is GraphRAG better than naive RAG?
  - How does GraphRAG work?
  - Limitations of GraphRAG
  - Consideration for GraphRAG
  - Demo
    - Graph
    - Inference

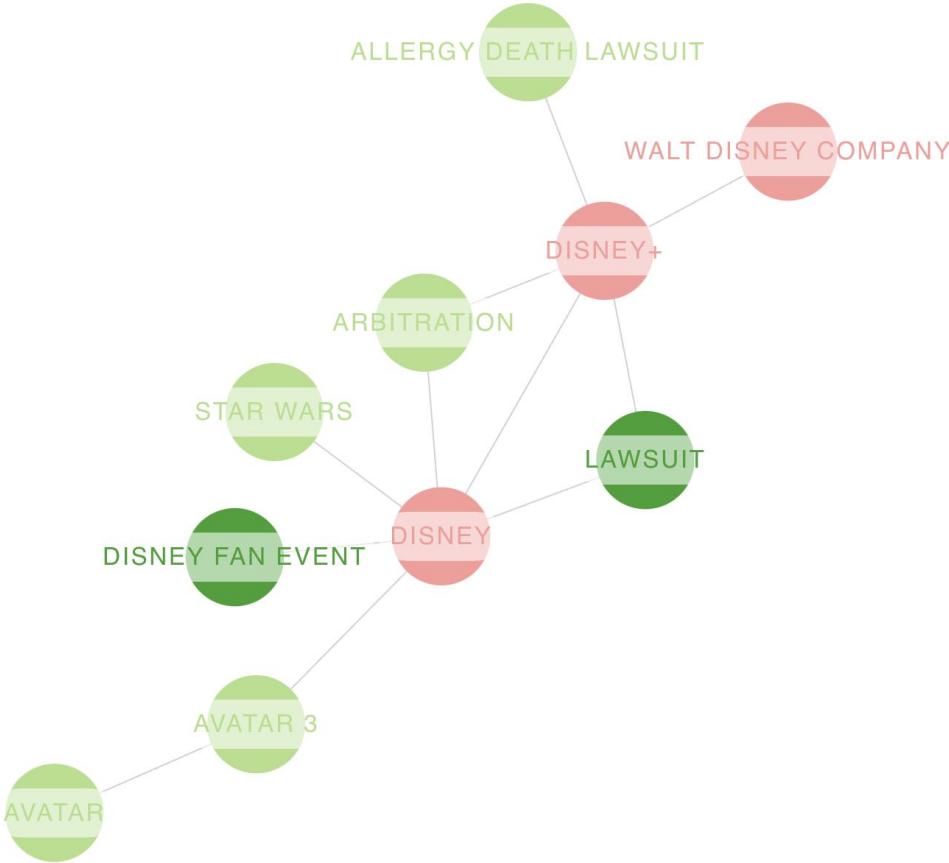
What is GraphRAG?

- Structured, hierarchical RAG
  - Using Knowledge Graphs
  - As opposed to naive semantic-search on plain text.
- Knowledge Graphs
  - Nodes
  - Relationships



What is GraphRAG?

- Structured, hierarchical RAG
  - Using Knowledge Graphs
  - As opposed to naive semantic-search on plain text.
- Knowledge Graphs
  - Nodes
  - Relationships
- Good for
  - Finding recurring themes
  - Different perspectives
  - Bigger picture



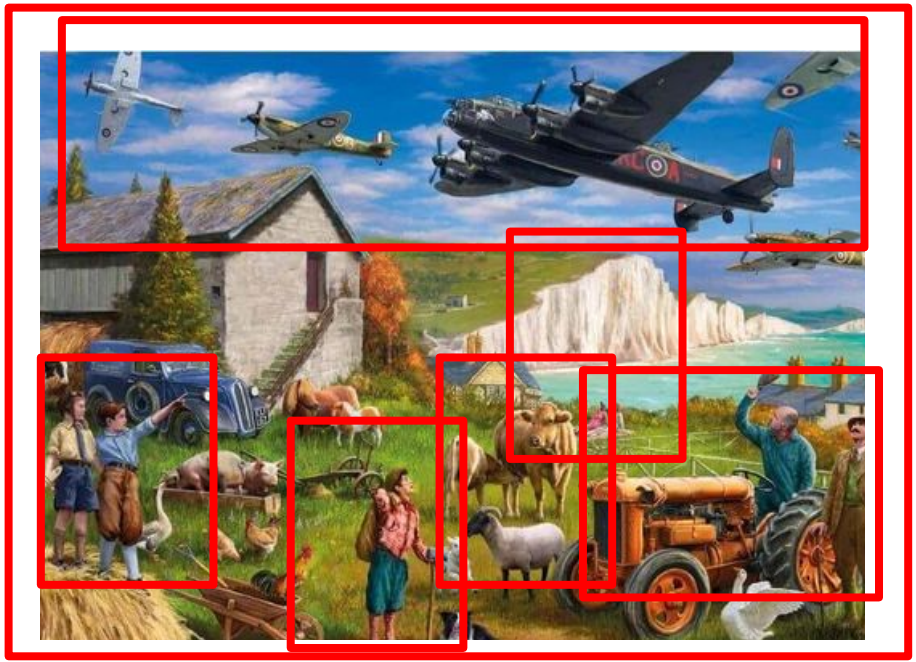
Why GraphRAG is better : A Visualisation

Question: What are people in village doing?

Naive RAG

The men, a kid are looking at the sky, another kid is pointing at the sky and the calf is drinking milk.

GraphRAG



Everyone in the coastal village has stopped to wonder at the flying fleet of war planes.

GraphRAG : Deep Dive

Indexing

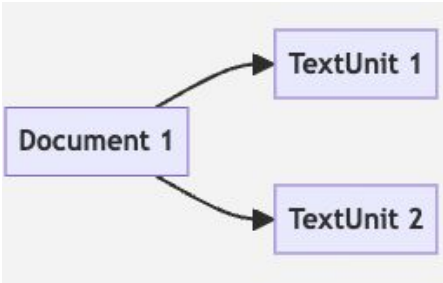
- 4 Step process to generate a Knowledge Graph using LLM

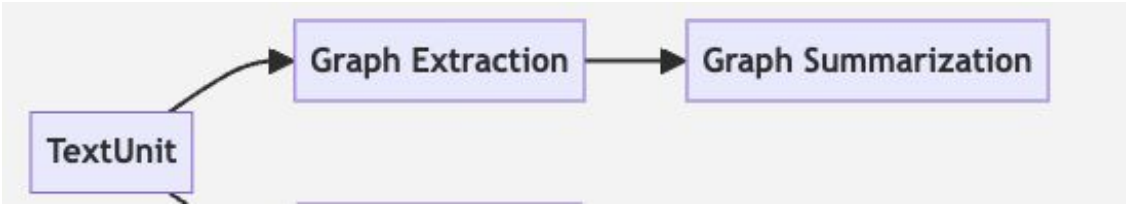
Querying

- Answering user's question using Knowledge Graph and LLM
- Global Search
  - Best suited for answering questions that require comprehensive understanding of the dataset
  - Eg: *What are the 5 most important events to review in this dataset?*
- Local Search
  - Suited for questions on a specific entities that have less number of related entities
  - Eg: *What is Disney's lawsuit about?*

GraphRAG: Indexing

- Converts your private data into a Graph
- 4 Step Process:
  a. **Chunking** makes Chunks.
  b. **Extraction** makes Entities and Relationships.
  c. **Community Detection** generates a hierarchy of communities using Leiden technique.
     - Applies recursive community clustering until we reach a community size threshold.
  d. **Community Summaries** makes Summaries for detected communities at multiple levels of granularity.





GraphRAG: Querying - Local Search

- Best suited for answering questions related to specific entities that have a small number of related entities
- Similarity search with LanceDB embeddings store
- **Extract semantically similar entities**
  - Get “semantically” similar entities
  - For each entity E, find neighbours with highest similarity score to embeddings of E.
  - Given entities, retrieve all records of relationships between these entities.
  - Use this data to construct the context and send it to an LLM to get an answer

GraphRAG: Global Search

- Best suited for answering questions that comprehensive understanding of the dataset
  - Eg:*What are the 5 most important events to review in this dataset?*
- Step 1: Gather and fit community reports into context window
- Step 2: Generate, rank and remove “internal” answers
  - Understand what community level we are exploring.
  - Generate answer for the user question using each community report as an answer
  - Rerank all answers based on quality score
  - Remove all answers below threshold for score
- Step 3: Prepare final answer
  - Combine all “internal” answers above threshold and let LLM prepare final answer

GraphRAG: Limitations

- Latency
  - Local Search: Latency similar to naive RAG.
  - Global Search: High latency and varies based on community level, size of dataset, LLM
  - Fast version of GraphRAG is coming soon
- Cost
  - ~ $3 to build a graph for 4930 lines of text with GPT-4o-mini
- Data format
  - Only .txt and .csv supported.
  - No multi-modality yet.
- Data ingestion
  - No easy way to incorporate new data into existing graph

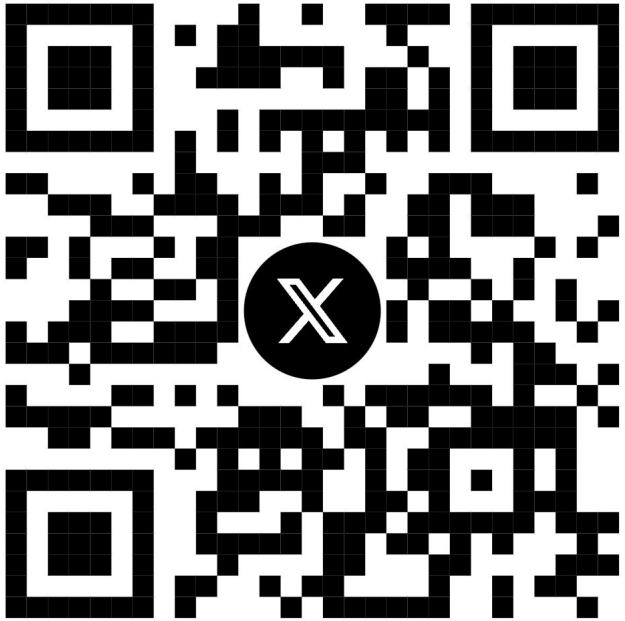
GraphRAG: Considerations

- Trade latency for accuracy
  - How feasible is it to wait for a really good answer?
- Cost
  - Indexing and freq
  - Global Search
  - Local models, SLMs
- Highly customisable
  - Prompt Tuning to any arbitrary domain. Eg: Space, Cancer Research, Politics
- How often do you find your RAG to be incomplete in its answers?
  - How often do you think your users might need Global Search?
- Data exploration tool
  - If you are building something, Graph is a great and fun way to understand the data.

GraphRAG : Demo!

- Gephi
- Prompt tuning + Parquet files
- GraphRAG Visualiser

Thank you!



Feel free to reach out!