Friday, Feb 7 - with Remi (mentor)

- All members ask questions about objective of project

- #1 beauty in the world
- Many different sub brands within l'oreal
- Focus on e commerce and advertisement
- Work heavily with amazon and own e commerce website
- Optimize promotion detect when products drop
- Have some genAI use cases for CMI (marketing team)
-

Last year compared graph look up rather than vector look up

- Results were similar results graph is more complicated but both accomplished similar results
- graphRAG should be better than vectorRAG if graph is set up the right way
- Graph DB
- How to evaluate how good a graph is?
- Optimize way to construct graph for q and a with specific data and better than vectorRAG
    - Not sure if its solved yet
    - Maybe start with replicating microsoft or other companies
    - Try manually
- When we encode graph right way do we see gap in performance when querying?
- Huggingface dataset for ecommerce
    - Amazon 1 Million
        - Description price metadata
- Experimental just to see how encoding data would work
- Last year used columbia resources

Next week:

- **Read up on Past Report and look through past semester's**
- **Read paper on how to encode paper**
- **Remind remi to send report to us so we can read it**
- **Find data**
- **Confirm approach multiple datasets**

Friday, Feb 14 - with Remi (mentor) and Professor Sining (instructor)

Questions:

- What are we exactly doing? We understand the project proposal was to construct a graph for graphRAG to increase performance. However, we noticed that last semester's group already accomplished this task using almost an identical tech stack. What do you want us to do differently?
- Use case for this project?
    - Bc if we increase retrieval efficiency then the vector will be the best
    - Usage more detailed chatbot?

Notes:

- Schema used for knowledge graph?
- Can do manually?
- Look into how Neo4j and microsoft construct the knowledge graph
- Directed vs Undirected graphs
- Graph used because if want to look for blue color clothes graph should be able to work but vector shouldn't
- Last year's repo:
- https://github.com/engie4800/dsi-capstone-fall-2024-loreal-ragvsgraphrag/blob/main/S2_L'Ore%CC%81al_RagVsGraphRag_final_report.pdf
- https://github.com/engie4800/dsi-capstone-fall-2024-loreal-rag-capstone
- OpenAI Resources/tokens
- Langchain

TODO for Tuesday (2/18):

- Figure out a timeline for our project
- Look into Neo4j logistics
- Afterwards find out how to store amazon dataset
- Get OpenAI Tokens

Friday, Feb 21 - with Remi (mentor)

Questions:

- GitHub issue: tutorial code usage
    - Remi: Use it given no L'Oreal affiliated dataset
- Choice of knowledge graph form:
    - Remi: try different options and optimize the graph structure on specific application scenarios/datasets with evaluation/benchmark

Notes: **Weekly Progress Presentation (Team Report)**

- Jiayi: Introduce capstone timeline
- Pei:
    - Setup Python env and OpenAI API key usage
    - Reproduce the evaluation of VectorRAG with evaluation library `ragas`
- Jiaheng: import Amazon 2023 dataset, use LLaMA to generate synthetic questions and response, modify the pipeline of last team and try to reproduce their result
- Brandon:
    - Data and Neo4j setup: create account and instance,
    - Data engineering: convert tabular data (csv) to unstructured text (md)
    - Graph Construction: extract entity and relationship by GPT, generate Cypher query, extract information from knowledge graph
- Jiayi: Different knowledge graph forms for GraphRAG
    - Entity-Relationship
    - Hierarchical
    - Event-based
    - Attribute
    - Causal

TODO:

- Dataset source similar to realistic application scenarios: customer reviews maybe helpful
- Find out use case related to realistic scenarios: input & output, evaluation
- Final Goal: find out way to construct best graph

Friday, Feb 28 - with Remi (mentor)

Questions:

- GitHub issue: tutorial code usage
    - Remi: Use it given no L'Oreal affiliated dataset
- Choice of knowledge graph form:
    - Remi: try different options and optimize the graph structure on specific application scenarios/datasets with evaluation/benchmark

Notes: **Weekly Progress Presentation (Team Report)**

- Jiayi: Introduce capstone timeline
- Pei:
    - Setup Python env and OpenAI API key usage
    - Reproduce the evaluation of VectorRAG with evaluation library `ragas`
- Jiaheng: import Amazon 2023 dataset, use LLaMA to generate synthetic questions and response, modify the pipeline of last team and try to reproduce their result
- Brandon:
    - Data and Neo4j setup: create account and instance,
    - Data engineering: convert tabular data (csv) to unstructured text (md)
    - Graph Construction: extract entity and relationship by GPT, generate Cypher query, extract information from knowledge graph
- Jiayi: Different knowledge graph forms for GraphRAG
    - Entity-Relationship
    - Hierarchical
    - Event-based
    - Attribute
    - Causal

TODO:

- Dataset source similar to realistic application scenarios: customer reviews maybe helpful
- Find out use case related to realistic scenarios: input & output, evaluation
- Final Goal: find out way to construct best graph

Friday, Mar 7 - with Nicole Brye (mentor assistant filling in for Remi) and Professor Sining (instructor)

Questions:

- How to overcome KG construction max length issue?
    - Reduce to smaller chunks/batches
    - Could try use the classical way to construct KG (without LLM)
- Can we see code of past semesters constructing KG?
    - Past semester GitHub has code for querying from KG but not constructing

Notes: **Weekly Progress Presentation (Team Report)**

- Last week had some issues with OpenAI API versioning
    - Fixed this week
- Some issues with OpenAI to construct Cypher Queries for constructing graph
    - Fixed:
        - Markdown files too large (tokens per minute) TPM limit reached
        - String too long
    - Not Fixed: Maximum context length limit exceeded
        - Limit: 8192 tokens while messages resulted in 270668
        - Unsure how to approach this
- One approach could be to remove descriptions from markdown file as currently not used in prompting
- Vector Search / Semantic Search for GraphRAG
- Embedding the demo KG with GenAI API of Neo4j
- Found LLMGraphTransformer which is another route

TODO:

- Prep for Midterm report
- Prep for Mini Conference

Friday, Mar 14 - with Remi (mentor) and Nicole Brye (mentor assistant)

Questions:

- Can we use subset of dataset (beauty and health)?
    - reduce scope so that issues with OpenAI token costs mitigated

Notes: **Weekly Progress Presentation (Team Report)**

- Shared information from mini-conference.
- Received feedback from mini-conference to work with small subset of data to test functionality

TODO:

- Midterm Report