TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

RAG Evaluation

TIM 175 WEEK 8 LAB

This week, we will work on evaluating the RAG implementation from last week to analyze our outputs and improve our implementation. We will also add to the spreadsheet from last week to create a completed annotated dataset. This will include an **individual deliverable (due Saturday at 11:59pm)** and a **team deliverable (due Monday 11:59pm)**

Readings:
See Week 8 Prelab or course spreadsheet.

Submission Link
Week 8 TIM 175 Submission Form (Spring 2025)

Brief Task Overview

- Setup your copies of the Evaluation Spreadsheet and Colab.
- Manual Evaluation of Retrieved Context
- Manual Evaluation of User response
- RAGAS Evaluation of User response
- Iteration and Refinement of RAG Pipeline

You can use ChatGPT or other GenAI tools to inform any part of the assignment but: (1) you need to first form your own independent thoughts, (2) every word included in the submission needs to be something you've read, thought about, and decided to include, and (3) you should strive towards submitting the highest quality work you can rather than mediocre work that meets the requirements.

Setup:

Google Spreadsheet: Create a copy of this Evaluation template and paste your RAG results from last week which has your 5 user queries, ground truths, query objects, retrieved context and user responses into this sheet. We will be completing additional columns for evaluations this week.

Note: for the ground truths, please put the actual podcast episode name rather than just the transcript number. Also, please add the actual Q&A pair text that you know is relevant for the user query.

Evaluation Google Colab: Make a copy of this Google Colab to edit. In the upcoming tasks (Task 2), this Colab will have you go through evaluation metrics on our RAG system using the RAGAS library. The Google Colab will:

- Load your dataset from last week into a dataframe
- Convert it into an EvaluationDataset object for it to be interpretable for RAGAS.
- Loop through each point in our dataset to evaluate the LLM's response

RAG Pipeline Google Colab: Make a copy of your Google Colab from last week to refine your prompts in Task 4. This can be your individual submission or your team submission.

Change runtime type to T4 GPU. On your copy of the Google Colab, look at the top right corner and you will see your connection status and a dropdown menu for "Additional connection options". From there change your runtime type to T4 GPU and save. GPU connection is preferred for AI tasks over CPUs primarily due to their parallel processing capabilities. This will make your notebook run more efficiently.

Task 1: Manual Evaluation of Retrieved Context

You will start by doing a manual evaluation of the retrieved contexts. In the tab called "Manual RAG Evaluation", fill out the following additional columns:

1. **Retrieval Reflection:** In the 'Retrieval Reflection' column, reflect on the retrieved contexts for each user query from last week's exercise and consider the quality of the retrieval:

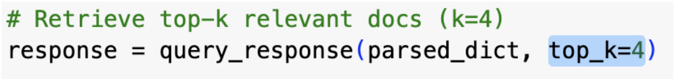
    a. *E.g were the retrieved passages relevant? Were there passages that were relevant, but were not returned? Are the retrieved passages returned in the ideal rank order (most relevant to least relevant?)*

    b. Add a short reflection evaluating the retrieved context in your copy of the Evaluation Spreadsheet.

2. **Reference Contexts:** Now create a ground truth for the contexts that should have been retrieved.

    a. **Note**: to actually determine the true reference contexts, you would in principle have to read through ALL of the transcripts in our dataset. We do not expect you to do this - you only have to follow the below instructions.

    b. *Modify the code from Week 7 Lab to retrieve 10 pieces of context rather than 4 by editing the part where it says "top_k=4" to change it to "top_k=10"*

    ```
    # Retrieve top-k relevant docs (k=4)
    response = query_response(parsed_dict, top_k=4)
    ```

    c. *From the list of top 10 retrieved passages, remove any irrelevant passages,*

    d. *If the passages could be reordered for relevancy, re-order them (most relevant to least),*

    e. *For the user queries in which you already knew of a relevant passage, if you did not see that passage included, and if the relevant passage you knew of is better than the top 4 in your reordered list, then add your user query into the list where it belongs,*

    f. *Take the top 4 out of the resulting list as your ground truth,* Fill in the manual evaluation columns for the RAG retrieval outputs in the “manual evaluations” tab of your copy of the Google Spreadsheet for all 5 queries. You will submit your copy of the spreadsheet in the Google Form.

Task 2: Manual Evaluation of User Response

You also need to do manual evaluations for the actual user response generated from your prompt workflow from last week. This is the response that will be returned to the user. In the same tab called “Manual RAG Evaluation”, conduct a manual evaluation for the user response. This has been broken down into multiple columns to help you assess the following:

3. **Reference:** First, create a ground truth response based on the ground truth reference contexts from the previous task (the resulting top 4 relevant passages when starting with top 10 passages). The ground truth response is the ideal response you'd like to return to the user that is both helpful to the student and incorporates rich stories from professionals.

4. **Evaluation Metrics:** For each of these, you should provide a manual 0/1 label and a rationale

    a. *Response relevancy - whether the response answered the user query in a way that will likely be helpful to the student,*

    b. *Format and tone - whether the response was formatted or written in a way that feels natural and supportive,*

    c. *Faithfulness - whether the response is faithful to the retrieved contexts,*

    d. *Community - whether the response connects students with the community, e.g. whether it introduces community members by name, organization, and occupation, and provides a soundcloud link for students to continue listening,*

    e. *Rich stories - whether the response is grounded in rich stories from the career journey of professionals in the community, to the extent possible given the retrieved context,*

Fill in the manual evaluation columns for the user response in the “manual evaluations” tab in the Google Spreadsheet for all 5 queries. You will submit your copy of the spreadsheet in the Google Form.

Task 3: RAGAS Evaluation of User Response:

In this task, you will be using the RAGAS library to evaluate the dataset through the code provided to you in the Google Colab. The google Colab will guide you through the following steps:

- Loading your RAG results and evaluation spreadsheets from last week into a dataframe. This should include:
    - ○ Queries - your user queries

    - Context_list - the list of contexts the rag system retrieved per given query
    - User response - the final LLM output from your RAG pipeline
- Converting this evaluation dataset into an EvaluationDataset object for it to be interpretable for RAGAS.
- Using this EvaluationDataset to loop through each query in our dataset to evaluate the Retrieved Context and LLMs response

Your task is to run the colab with your dataset and then in the "RAGAS Evaluation" tab of your spreadsheet fill in the "RAGAS Evaluation of User Response" columns on faithfulness and response_relevancy for each of your queries.

Then write a short reflection of the RAGAS results for each query in the "RAGAS Evaluation" column to add what you think they mean and what they help you understand about your results and whether they align with your manual evaluations.

Note: we are not going to do RAGAS evaluation of retrieved context ("Context Recall" and "Context Precision") for this assignment. However, if someone would like to try doing this, we'll provide some extra credit :)

Task 4: Iteration and Refinement of RAG Pipeline:

Now that you have a solid grasp of implementing and evaluating Retrieval-Augmented Generation (RAG), focus on refining your approach to achieve better results. Use a copy of your RAG Pipeline Google Collab from last week to refine your RAG pipeline prompts.

1) **Optimize your RAG Pipeline - Prompt 1**: Based on your evaluation, go back and refine your first prompt from last week to improve the query objects generated for retrieval results. If your prompt was not adding the industry filter before, you can do so now. Make sure your prompt workflow is handling cases when no industry filter is required as well as when an industry filter should be used.

2) **Optimize your RAG Pipeline - Prompt 2**: Based on your evaluation, go back and refine your second prompt from last week to improve the user response. After doing the detailed manual evaluations and the RAGAS evaluations above, you should have gotten an idea of what you can consider improving in your user responses.

3) **Add Another round of queries and manual evaluations:** After refining your process, Add 3 new queries to your dataset and complete manual evaluations again (Task1 and Task 2), and compare the updated results with your previous queries to check for improvements. You do not need to repeat RAGAS evaluations.

Submission Instructions

Reflect and submit. Finally, reflect on the individual activity. Submit your responses to the following questions directly on the Google Form along with your latest version of the RAG Colab Pipeline link and your results and evaluation spreadsheet. **IMPORTANT: MAKE SURE TO GIVE US PERMISSIONS TO ACCESS DOCUMENTS**:

1. What did you learn about your results from doing the manual evaluations ? What did you feel was good/strong in your RAG results from last week? What did you feel were gaps/weaknesses in your results?

2. How did the RAGAS library evaluations help you understand your results better? How did your manual evaluations compare with the RAGAS evaluations for the user responses? Was there any discrepancy and what do you think is the reason for any difference between them?

3. How did you optimize your RAG pipeline in the last task? What changes did you make to your prompt 1 for creating user queries and what changes did you make in your prompt 2 for creating user responses?

4. How did your results change in the second iteration? What is better in your results after the second iteration of the RAG pipeline? What is still not performing well or what metrics are still giving poor results? How could you improve that?

Team Assignment Instructions

In-section peer review of evaluation. Among the team members who are present at the section, conduct a peer review of each other's work. We don't have a strong preference for how you assign peer reviewers, but one way to do it is to use a cycle A → B → C → D → A. In other words, if you have 4 team members present (A, B, C, D), have A peer review B, B peer review C, C peer review D, and D peer review A.

Examining your team member's evaluation spreadsheet and latest prompts, taking note of specific ways they could have improved them. Then:

- Write a two sentence reflection on what your team member did well in. For example, how did they write the manual evaluations and RAGAS evaluations? How did they update their prompts to improve results? What was good about the new set of results they generated?
- Write a two sentence reflection on a concrete, specific way in which your partner could have improved their prompts and results and evaluations. You should focus on what is most important / would make the biggest impact or improvement.

Note: The course staff will share some initial observations and pointers, but this is only based on a quick review of the individual submissions. You should take what they say into account, but make sure to be detail-oriented in thinking about the submission you are reviewing and how it should improve.

Compile your reviews. Create another copy of the RAG results and evaluation spreadsheet for your team submission and compile the reviews from all of the participating team members in the 'Peer Reviews' tab. Specify for each review, who was the reviewer, who was being reviewed, the two sentence reflection on what your team member did well in, and the two sentence reflection on the most important way they could have improved.

We are looking at the quality of your reviews, so the way to maximize your points is to write the most helpful critique that points out the biggest way each team member can improve.

Iterate as a team. Work together to create a single final team version of the RAG pipeline and results, building on the work you have already done in the individual submission and considering what you might want to draw from each submission (e.g. refined prompts for creating query objects, refined prompts for creating user response). Note: you do not have to use something from every individual submission. You should be trying to create a final RAG Pipeline workflow that produces the highest quality outcomes you can achieve.

Document your final pipeline for 15 diverse user queries. Run your 15 queries from last week through your new version of the RAG pipeline and document the results in your team version of the RAG Results and Evaluation Spreadsheet.

Complete Manual evaluations for the 15 user queries. You can divide the queries amongst yourselves and complete the manual evaluation of retrieved outputs (2 columns) and manual evaluation of user responses (4 columns) for each query.

Reflect and submit. Submit your responses to the below questions directly on the Google Form along with your final prompt workflow and final spreadsheet. **IMPORTANT: MAKE SURE TO GIVE US PERMISSIONS TO ACCESS DOCUMENTS**

1. Write 3-4 sentences on your team's thought process. How did you create your final RAG Pipeline? What improvements did you make? What feedback did you incorporate (from peer review or your tutor)?
2. Write 3-4 sentences on your team's results and evaluation. How did the results perform in manual evaluations? What were the good/strong aspects of your results? Where were your results still lacking?
3. Write a sentence or two describing the team dynamics. Were there any challenges you faced working in your team and how did you overcome them?
4. Please list each member of your team, whether they attended and engaged in section discussions, and their specific contributions.

Evaluation Rubrics

Individual submissions will be graded on a Check+, Check, Check-, Minus+, Minus scale according to the below rubric. The purpose of individual submissions is primarily to ensure that all members are contributing to their team, so graders will not be providing feedback on these.

Check+ Outstanding, one of the best in the class (102%),
Check High quality, though not one of the best in minor ways (95%),
Check- Completed the work, but needs significant improvement (80%),
Minus+ Low quality or missing significant portions (40%)
Minus Did not do the work or barely did any work (0%)

Team submissions will be graded according to detailed rubrics to be posted on the rubric spreadsheet