TIM 175: BUSINESS STRATEGY AND INFORMATION SYSTEMS
SPRING 2025

RAG Metrics and Evaluation

TIM 175 WEEK 8 PRELAB

Last week we did an introduction to RAG, and how it can be used to ground LLM results in data sources. Just like normal LLM outputs, it's important to have metrics to evaluate RAG systems. In this exercise, we'll cover various metrics for RAG, as well as a helpful library that will make implementing them much easier. **This individual prelab is due Tuesday 11:59pm.**

Readings: We mark up to two readings with a ⭐ that we suggest you read

- Building A Generative AI Platform ⭐
- List of available metrics - Ragas ⭐

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

Brief Task Overview:

- Brainstorming Metrics for RAG
- Understanding existing Metrics for RAG
- Manually Evaluating with RAG metrics
- Using RAGAS for Evaluating with RAG metrics

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:

1. Make a copy of this Google Document and make sure to update sharing permissions.
2. Complete Activity 1, Activity 2 and Activity 3 on your copy of this Google Document.
3. Make a copy of this Google Colab for Activity 4 and follow the instructions to complete the activity. Add your reflections for Activity 4 on your copy of this document.
4. Submissions:
    a. Submit your copy of this document
    b. Submit your copy of the Google Collab

RAG Metrics and Evaluation

Activity 1: Brainstorming Metrics for RAG

As a reminder, RAG consists of the following aspects:

1. **Retrieval**: retrieving the most semantically relevant documents from a database,
2. **Augmentation**: using those retrieved documents to augment an LLM prompt,
3. **Generation**: generating a useful output from the LLM prompt,

It's important to remember that all of these aspects are important for good results. For example, you may have a perfect prompt, but if the retrieved documents aren't relevant you won't get good results (and vice versa).

With this in mind, take a minute to think about these three steps and how we might design metrics to evaluate them. For the time being, don't worry about exactly how you might obtain these metrics, just think about the high level of what's important to quantify about a specific RAG system. As a reminder, you can look at some of the week 2 metrics for inspiration.

In the box below, list at least 3 potential metrics we might consider for a RAG system. For each metric, write a short description of its purpose (no more than 1 sentence)

1. Metric name - short description
2.
3.

Activity 2: Understanding Standard Metrics for RAG

Now that you've come up with some of your own metrics, let's take a look at some of the standard ones we will be using. Go through the list of 6 RAG metrics in the RAGAS library documentation HERE - you can skip the multimodal ones. Try and understand each one carefully, as a few are quite similar!

Summarize each metric down to a single line like you did with your metrics above and write a short reflection about what you understand.

1. Context Precision - short description
2. Context Recall -
3. Context Entities Recall -
4. Noise Sensitivity -
5. Response Relevancy -
6. Faithfulness -

Write a short reflection of your understanding of the 6 metrics. Not just description of the metric but what do you think these are useful for, which cases are each of these most relevant for, which cases would they fail?

Activity 3: Manual Evaluation with RAG metrics

Once you fully understand the above metrics, let's test them out in a hypothetical scenario. Fill in the scores for the 3 queries below. In this scenario, the "Database of Facts" has been embedded and stored in a vector database. The retrieved documents for each query are listed by their index in the database (1-9). Assume the generated output has access to only the set of retrieved documents.

You only need to put a yes (Y) or no (N) for each metric, no need to calculate anything for now!

Database of Facts:

1. Bananas are berries, but strawberries aren't
2. Bananas are naturally radioactive
3. While oranges are known for Vitamin C, strawberries actually have more
4. A day on Venus is longer than a year on Venus
5. Octopuses have three hearts
6. Starfish have no brains or blood
7. Sharks existed before trees
8. The shortest war in history lasted only 38 to 45 minutes
9. A cloud can weigh over a million pounds

Query	Retrieved Indices	Output	**Scores (Y/N)**	**Reasoning**
How different are sea creatures to us really?	5	The vast diversity of ocean life can often feel like an alien planet! Octopuses have 3 whole hearts, and that's just the tip of the iceberg when it comes to the ocean.	Context Precision - Context Recall - Noise Sensitivity - Response Relevancy - Faithfulness -	

What fruit should I be eating to stay healthy?	1,2,3,4, 9	Different fruits each have their strengths and weaknesses. If you're looking for antioxidants, then blueberries are your go-to. If you want vitamin C, nothing beats an orange!	Context Precision - Context Recall - Noise Sensitivity - Response Relevancy - Faithfulness -	
What's the oldest thing on earth?	4,7	There are a ton of really old things on earth. While not the oldest species, some sharks are older than trees themselves! Although, if those sharks were on venus, they would've experienced less full days than years! The question remains whether that makes them older, or younger?	Context Precision - Context Recall - Noise Sensitivity - Response Relevancy - Faithfulness -	

Activity 4: Using RAGAS for Evaluating with RAG metrics

The metric definitions above come from a library called RAGAS. They can be used to automate evaluation for RAG. Now that you've gotten the chance to familiarize yourself with them, it's time to actually try implementing them in code.

1. Make a copy of the following document: TIM175 Week 8 PreLab - RAG Evaluation
2. Go through parts 1-3 to set up a simple RAG setup using your document
   a. This should all look very similar to last week!
3. Follow the instructions for the activity and test 5 queries on 3 different metrics. Place your results below and reflect on the experience.

Link to your collab:

Query:	Metric Tested	Score
Query 1		
Query 2		
Query 3		

Query 4		
Query 5		

Please write a short paragraph reflection on this activity. Were there any metrics that stood out to you? Were there any surprising results? If not, summarize your experience implementing and evaluating the RAG system.

Submission Instructions:

After completing all the activities, fill out the submission form including a link to your copy of **this google document** on it.