Principais contribuições do artigo “RAGAS: Automated Evaluation of Retrieval Augmented Generation”, de Shahul Es et al

Aluno: Leandro Carísio Fernandes

- A ideia do artigo é propor um modelo automático para avaliação de pipelines RAG sem que seja necessário ter um conjunto de dados anotado por humanos.
- A proposta dos autores é usar o próprio LLM para avaliar o resultado.
- A nomenclatura utilizada é a seguinte: dada uma pergunta q, o sistema inicialmente irá retornar um contexto $c(q)$ e gerará uma resposta $a_S(q)$.
- Na avaliação, é verificado três itens:
    - Faithfulness: A fidelidade checa se a $a_S(q)$ é fiel à $c(q)$. Ou seja, se a resposta realmente é suportada pelos dados fornecidos pelo contexto. A avaliação é feita em dois passos por um modelo de linguagem:
        - Passo 1: O LLM separa $a_S(q)$ em sentenças.
        - Passo 2: O LLM responde se a sentença pode ser inferida do contexto ou não.
        - O resultado da métrica é a % de sentenças que pode ser inferida pelo contexto.
    - Answer relevance: A relevância da resposta indica se a resposta realmente responde à pergunta. Isso é feito em dois passos:
        - Passo 1: O LLM gera n possíveis perguntas para a resposta avaliada.
        - Passo 2: Extrai os embeddings de todas as possíveis perguntas geradas.
        - O resultado da métrica é a média da similaridade de cosseno entre a pergunta realmente feita e as n perguntas geradas.
    - Context relevance: A relevância do contexto índice se o contexto fornecido tem apenas a informação suficiente para responder à pergunta ou se possui mais informações desnecessárias:
        - O LLM extrai do contexto $c(q)$ todas as sentenças que são consideradas relevantes para responder q.
        - O resultado da métrica é a % de sentenças relevantes do contexto.