Relatório Técnico — Projeto Chatbot com Hugging Face (RAG)

Willian Garcia de Araujo

Contexto do Projeto

Este projeto faz parte da atividade da **Aula 3 de PLN (Processamento de Linguagem Natural)**, cujo objetivo era implementar um **chatbot baseado em RAG (Retrieval-Augmented Generation)** utilizando:

- **FastAPI** para o backend
- **SentenceTransformers + FAISS** para indexação semântica dos textos fornecidos
- **Inference API da Hugging Face** para geração de respostas usando um LLM (modelo de linguagem)
- Frontend simples para interação em formato de chat

O fluxo previsto:

1. Usuário envia textos pelo endpoint /ingest → armazenados como embeddings no índice FAISS
2. Usuário faz perguntas pelo endpoint /ask
3. O sistema recupera os textos relevantes e envia como contexto para um modelo hospedado no Hugging Face
4. O modelo retorna a resposta baseada no contexto

Tecnologias e Componentes Implementados

- FastAPI + uvicorn
- sentence-transformers/all-MiniLM-L6-v2 para embeddings
- faiss-cpu para busca vetorial
- Integração com https://api-inference.huggingface.co/models/{HF_MODEL}
- Arquivo .env contendo:
- HUGGINGFACEHUB_API_TOKEN=hf_...
- HF_MODEL=Qwen/Qwen2.5-0.5B-Instruct
- Frontend web em HTML/CSS/JS estilo chat, usando fetch no /ingest e /ask

---

Resultados Parciais

- O backend sobe corretamente em http://localhost:8000
- O Swagger UI em /docs mostra e executa o /ingest com sucesso
  Exemplo:

{
"status": "added",
"text": "A FastAPI é um framework moderno e rápido para construir APIs em Python.",
"total_docs": 1
}

- Os embeddings são criados e armazenados corretamente

Problema Encontrado

Ao tentar executar o endpoint /ask para fazer perguntas, ocorre o erro:

{

"detail": {

"non_json_body": "Not Found"

}

}

- Código HTTP: **404 Not Found**
- Isso indica que a **Inference API do Hugging Face está rejeitando a requisição**
- Foi confirmado que:
  - A variável HF_MODEL está correta (Qwen/Qwen2.5-0.5B-Instruct ou TinyLlama/TinyLlama-1.1B-Chat-v1.0)
  - A variável HUGGINGFACEHUB_API_TOKEN está presente e visível no /health
  - O payload enviado segue o formato esperado pela API:

    {

    "inputs": "ping",

    "options": {"wait_for_model": true}

    }

- Porém, ao testar diretamente no PowerShell com:

  ```
  Invoke-RestMethod `
  -Uri "https://api-inference.huggingface.co/models/Qwen/Qwen2.5-0.5B-Instruct" `
  -Headers @{Authorization="Bearer $Env:HUGGINGFACEHUB_API_TOKEN"; "Content-Type"="application/json"} `
  -Method Post `
  -Body '{ "inputs":"ping", "options":{"wait_for_model":true} }'
  ```

- a resposta foi:

  {"error":"Invalid credentials in Authorization header"}

Ou seja: a Hugging Face **não está aceitando os tokens gerados para as contas dos alunos**, devolvendo erro de **credenciais inválidas**, impedindo qualquer resposta do modelo.

Impacto

- Sem acesso à Inference API, não é possível completar a etapa de **G (Generation)** do RAG
- O chatbot consegue ingerir e recuperar documentos localmente, mas **não consegue gerar respostas**

- Este problema está ocorrendo com **outros colegas da turma**, indicando falha geral no ambiente de desenvolvimento ou limitação imposta pelo Hugging Face

---

Conclusão

- O backend, embeddings e FAISS estão funcionando corretamente
- O problema está exclusivamente na **validação de token da Hugging Face**
- Sem um token aceito, não é possível avançar no desenvolvimento.