

Postgrado en inteligencia artificial.

Seminario del Estado Actual de la Tecnología

Capstone project
| Charmie RAG |

José Javier Barrios Velásquez - 21000478

Charmie

- Una nueva forma de descubrir tu carrera -





índice general

Resumen	3
Objetivo del Proyecto	3
Alcance	3
Arquitectura del Sistema.....	4
Diagrama de Arquitectura General	4
Componentes Técnicos	5
Modelos de IA	5
Modelo LLM.....	5
Base de Datos Vectorial	6
FAISS	6
Ventajas de FAISS	6
Referencias	7



NOTA

Lo que se necesita instalar para ejecutar los colabs y que apis se necesitan está en README.MD

Resumen

Objetivo del Proyecto

Desarrollar un sistema de Recuperación Aumentada por Generación (RAG) para proporcionar información precisa y contextualizada sobre pensums universitarios de la Universidad Galileo, asistiendo en el proceso de orientación vocacional de estudiantes.

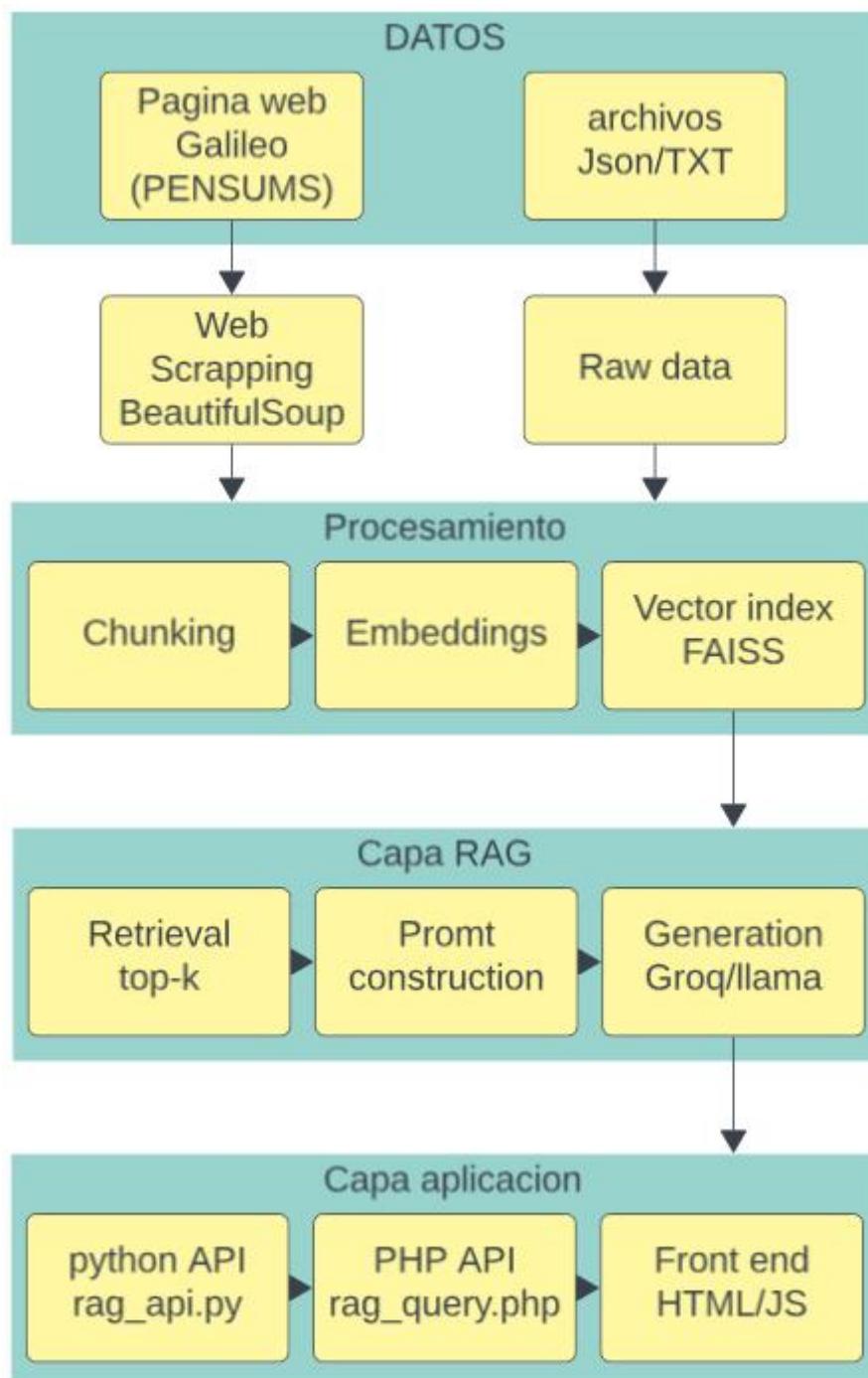
Alcance

- Extracción automatizada: Web scraping de pensums universitarios
- Procesamiento inteligente: Chunking semántico por ciclos académicos
- Búsqueda vectorial: Sistema de embeddings multilingüe con FAISS
- Generación de respuestas: LLM Llama 3.3 70B vía Groq API
- Interfaz de usuario: Chat web integrado con PHP
- Evaluación de calidad: Métricas automatizadas con DeepEval



Arquitectura del Sistema

Diagrama de Arquitectura General





Componentes Técnicos

Modelos de IA

modelo: "paraphrase-multilingual-MiniLM-L12-v2"

dimensión: 384

idioma: Multilingüe (optimizado para español)

tamaño: 420 MB

velocidad: 1000 oraciones/seg

max_seq_length: 128 tokens

Justificación de elección:

- Soporte nativo para español
- Balance entre velocidad y calidad
- Tamaño manejable para deploy local

Modelo LLM

modelo: "llama-3.3-70b-versatile"

proveedor: Groq

parámetros: 70 mil millones

contexto: 8,192 tokens

temperatura: 0.2 (más determinístico)

max_tokens: 2,000

Justificación de elección:



- API gratuito de alta velocidad
- Excelente comprensión del español
- Capacidad de seguir instrucciones complejas
- Generación coherente y precisa

Base de Datos Vectorial

FAISS

Facebook AI Similarity Search (FAISS) es una biblioteca de código abierto para búsqueda eficiente de similitud y clustering de vectores densos.

Características Implementadas

Tipo de índice: IndexFlatL2

Métrica de distancia: L2 (Euclídea)

Dimensión de vectores: 384

Total de vectores: 72 chunks

Tamaño en disco: 2.5 MB

Memoria RAM requerida: 50 MB

Ventajas de FAISS

Ventaja	
Velocidad	Búsqueda en < 10ms para 72 vectores
Precisión	Búsqueda exacta con IndexFlatL2
Escalabilidad	Soporta millones de vectores
Costo	100% gratuito, sin límites



Referencias

FAISS: <https://github.com/facebookresearch/faiss>

Sentence-Transformers: <https://www.sbert.net/>

Groq API: <https://console.groq.com/docs>

DeepEval: <https://docs.confident-ai.com/>

BeautifulSoup: <https://www.crummy.com/software/BeautifulSoup/>