# UCLA Clubs Explorer

STAT 418 – Final Presentation

Hengyuan (David) Liu

Features

- Interactive Visualizations:
  - Select multiple Club Categories
  - Bar chart of club distribution
  - Most frequent words by category
  - Word cloud of common themes

- KNN Similarity Club Finder (API):
  - Input keywords + # of results
  - Output: Club Name, Category, Clickable Link, and Similarity Score

- Club Chatbot (API):
  - Powered by RAG + Gemini API
  - Accepts open-ended student questions
  - Returns grounded, concise, and linkable club recommendations

UCLA   Department of Statistics and Data Science
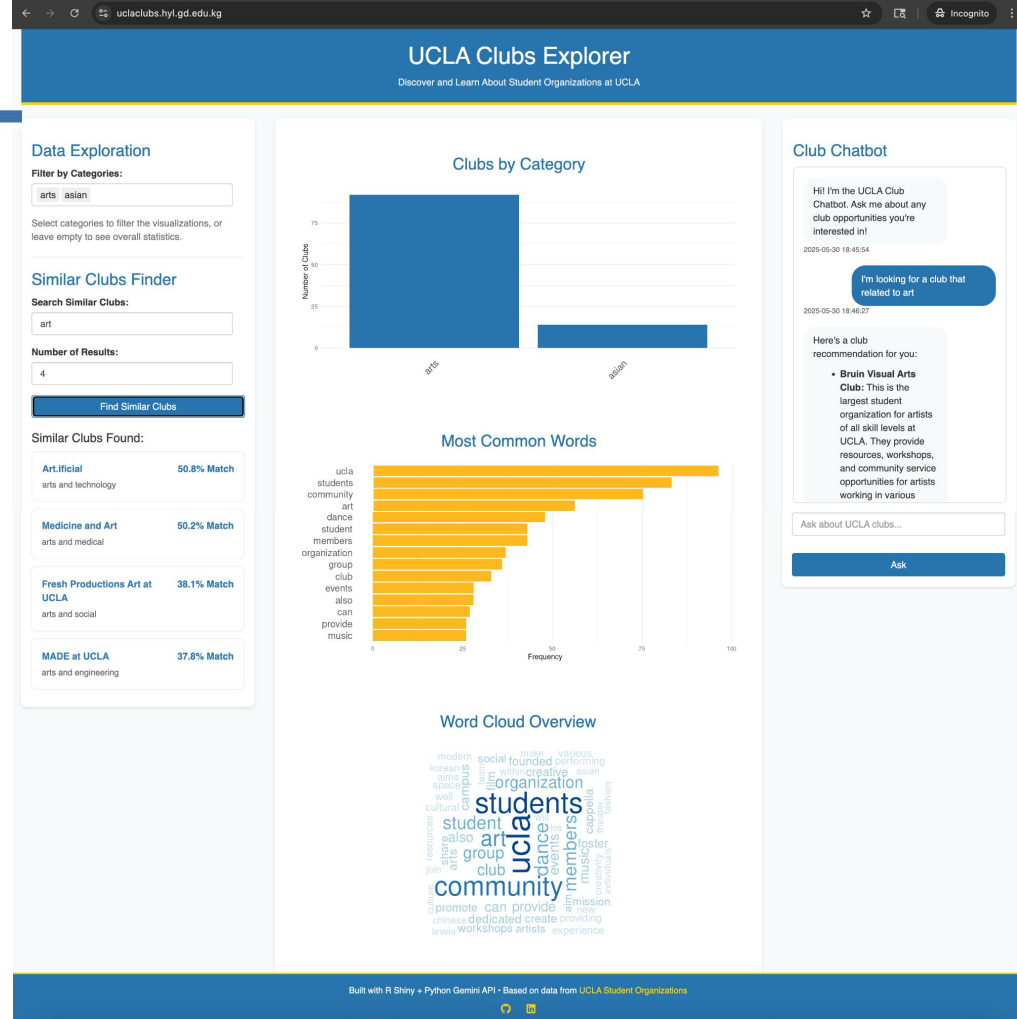
KNN and RAG+LLM Pipelines

Stage	K-NN (Flask)	RAG + Gemini (Flask)
Prep	TF-IDF → vectorizer.joblib	CSV → LangChain docs
Store	In-RAM TF-IDF matrix	Embeddings in ChromaDB
Query	/predict_clubs → cosine K-NN	/chat → Chroma retrieve
Process	Rank top-k clubs	Prompt → Gemini Flash
Output	JSON {name, link, sim}	Text answer with links
Serve	Docker → Cloud Run	Docker → Cloud Run
Secrets	None Needed	GOOGLE_API_KEY (Secret Manager)
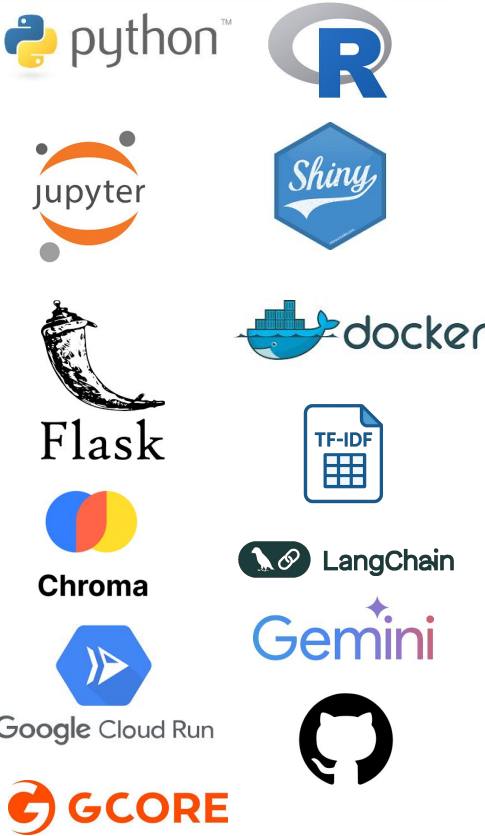
Architecture

☐	✔	ucla-clubs-app	Container
☐	✔	ucla-clubs-knn	Container
☐	✔	ucla-clubs-rag-api	Container

https://ucla-clubs-app-980752141572.us-west1.run.app

https://ucla-clubs-knn-980752141572.us-central1.run.app

https://ucla-clubs-rag-api-980752141572.us-central1.run.app

UCLA Department of Statistics and Data Science
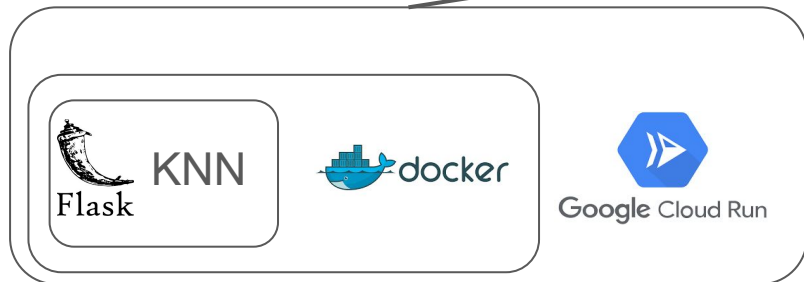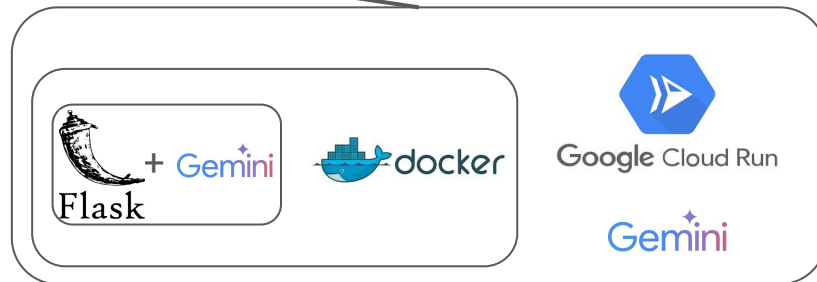
Prospects & Future Directions

- **Stronger NLP:** Data Preprocessing,Clean & normalize text (lemmatize, de-duplicate), Upgrade to contextual embeddings
- **RAG Refinement:** Prompt-engineering playbook, Fine-tune Gemini on club FAQs
- **Maintenance:** scrape community.ucla.edu → rebuild vectors/Chroma DB quarterly/yearly on GitHub
- **Ops & Monitoring:** Cloud Run alerts + dashboards, cybersecurity
- **User Feedback Loop:** " 👍/👎 helpful?" in Shiny UI, log queries & retrain on signals

Demo:

- https://uclaclubs.hyl.gd.edu.kg/
- https://hengyuanliu.shinyapps.io/uclaclubs/
- https://ucla-clubs-app-980752141572.us-west1.run.app