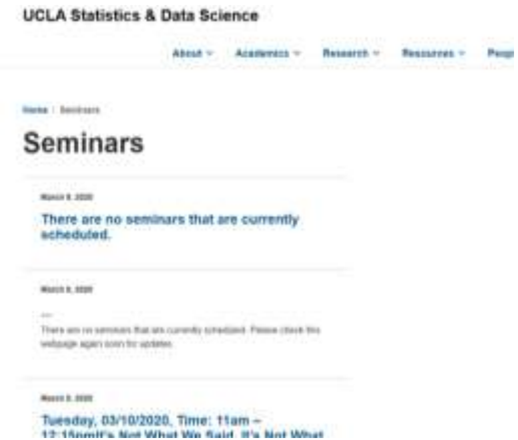
a QA System for UCLA Statistics Seminars

Fanchao Lin
STATS 418 Final Project

Data Scraping and Processing



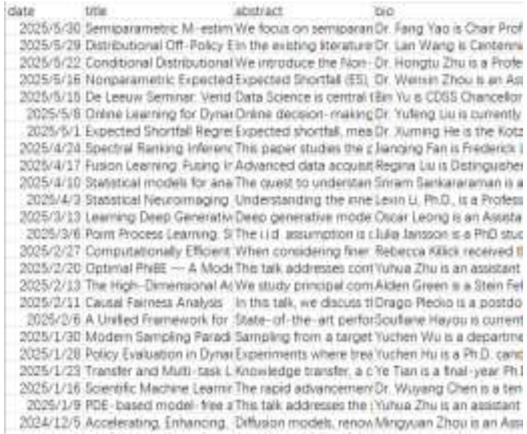
UCLA Statistics & Data Science

Seminars

There are no seminars that are currently scheduled.

Tuesday, 03/10/2020, Time: 11am –

2024 – 2025 Acad. Year



date	title	abstract	bio
2025/5/30	Semiparametric M-estim	We focus on semiparan	Dr. Fang Yao is Chair Prof
2025/5/29	Distributional Off-Policy E	In the existing literature	Dr. Lan Wang is Centenni
2025/5/22	Conditional Distributional	We introduce the Non-	Dr. Hongtu Zhu is a Profe
2025/5/16	Nonparametric Expected	Expected Shortfall (ES),	Dr. Weixin Zhou is an As
2025/5/15	De Leeuw Seminar: Verid	Data Science is central t	Bin Yu is CDSS Chancellor
2025/5/8	Online Learning for Dynai	Online decision-making	Dr. Yufeng Liu is currently
2025/5/1	Expected Shortfall Regre	Expected shortfall, mea	Dr. Xuming He is the Kotz
2025/4/24	Spectral Ranking Inferenc	This paper studies the c	Jianqing Fan is Frederick
2025/4/17	Fusion Learning: Fusing Ir	Advanced data acquisit	Regina Liu is Distinguishe
2025/4/10	Statistical models for ana	The quest to understan	Sriram Sankararaman is a
2025/4/3	Statistical Neuroimaging	Understanding the inne	Lexin Li, Ph.D., is a Profess
2025/3/13	Learning Deep Generativ	Deep generative mode	Oscar Leong is an Assista
2025/3/6	Point Process Learning: S	The i.i.d. assumption is c	Julia Jansson is a PhD stuc
2025/2/27	Computationally Efficient	When considering finer	Rebecca Killick received t
2025/2/20	Optimal PhiBE — A Mode	This talk addresses cont	Yuhua Zhu is an assistant
2025/2/13	The High-Dimensional A	We study principal com	Alden Green is a Stein Fel
2025/2/11	Causal Fairness Analysis	In this talk, we discuss t	Drago Plecko is a postdo
2025/2/6	A Unified Framework for	State-of-the-art perfor	Soufiane Hayou is current
2025/1/30	Modern Sampling Paradi	Sampling from a target	Yuchen Wu is a departme
2025/1/28	Policy Evaluation in Dynai	Experiments where trea	Yuchen Hu is a Ph.D. canc
2025/1/23	Transfer and Multi-task L	Knowledge transfer, a c	Ye Tian is a final-year Ph.I
2025/1/16	Scientific Machine Learnir	The rapid advancemen	Dr. Wuyang Chen is a ten
2025/1/9	PDE-based model-free a	This talk addresses the	Yuhua Zhu is an assistant
2024/12/5	Accelerating, Enhancing,	Diffusion models, renow	Mingyuan Zhou is an Ass

Seminar formats vary across years, making pattern matching difficult.

As a workaround, I extracted the full titles and content blocks directly.

Then, I used the OpenAI API to extract seminar topics, speaker info, and summaries line by line.

The Model

VECTOR DATABASE:

- The processed **title, abstract, and speaker bio** of each seminar are combined into a single chunk (450 words on average, around 600 tokens).
- Each chunk is embedded using the **text-embedding-3-small** model.
- The resulting embeddings are stored in a vector database using **ChromaDB**.

RAG RETRIEVAL + Q&A:

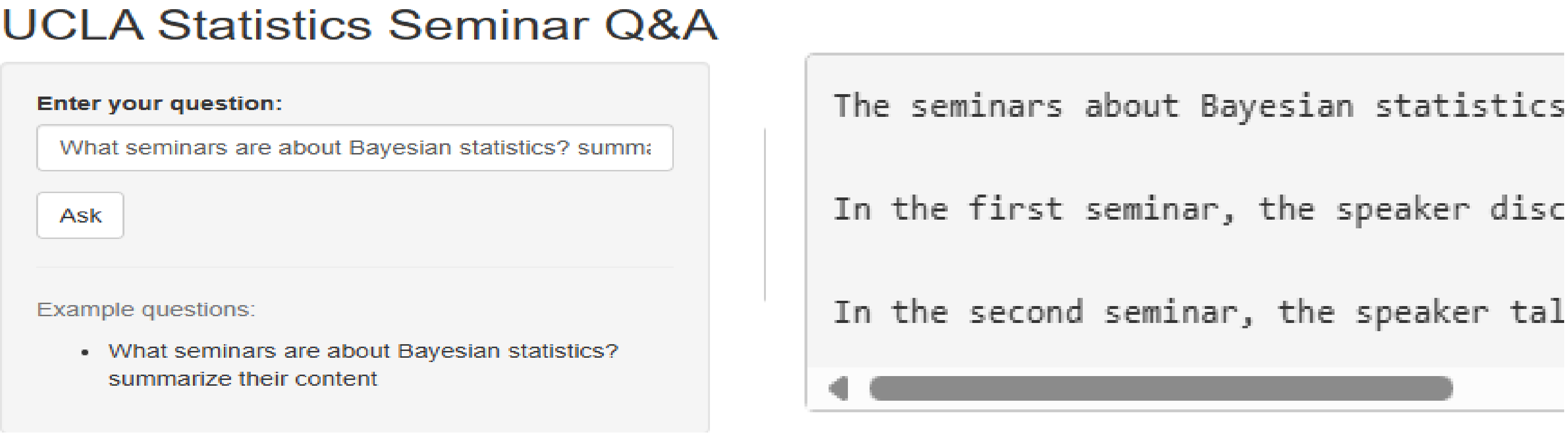
- The user's question is embedded using the same model.
- An approximate nearest neighbor (ANN) search is performed in ChromaDB based on cosine similarity, retrieving the **top 3** most relevant chunks.
- The retrieved chunks are combined with the user question and passed to **GPT-4 Turbo** to generate a final response.

Deployment

- **Frontend**: Deployed R Shiny application on shinyapps.io
- **Backend API**: Flask-based REST API containerized with Docker
- **Cloud Hosting**: Flask API deployed on Google Cloud Run
- **Container Registry**: Docker image pushed to Docker Hub for reuse
- **Integration**: Shiny frontend communicates with Flask API via HTTPS (JSON POST requests)

Demo

https://lfc787878.shinyapps.io/UCLA_seminar_QA/



UCLA Statistics Seminar Q&A

Enter your question:

What seminars are about Bayesian statistics? summa

Example questions:

- What seminars are about Bayesian statistics? summarize their content

```
The seminars about Bayesian statistics

In the first seminar, the speaker disc

In the second seminar, the speaker tal
```

Demo

Q: What seminars are about Bayesian statistics? summarize their content

A: The seminars about Bayesian statistics are "Bayesian Inference in Nonparanormal Graphical Models" and "Scaling and Generalizing Bayesian Inference."

In the first seminar, the speaker discusses Bayesian solutions for nonparanormal graphical models, which generalize Gaussian models for continuous variables. The talk focuses on two Bayesian solutions, one with priors on transformation functions and the other with a rank-based likelihood function, and studies their numerical performance through simulation and real data application.

In the second seminar, the speaker talks about recent research on addressing limitations in Bayesian statistics, including handling massive data sets and developing generic algorithms for approximating the posterior. The speaker also discusses how approximate posterior inference algorithms have revolutionized the field, making Bayesian statistics a usable and general-purpose language for data analysis.