

Data Science Professional Practicum (DSCI 560)
Final Project Description and Suggested Plan
Instructor: Young Cho, Ph.D.

This semester's Final Project goal is to build a complete working solution for electronic/virtual Teaching Assistants for a university course. You will likely rely on all the skills you've gained while completing your laboratory assignments. You will need to also learn additional skills for the domain-specific functions necessary to make a complete system.

There are weekly demonstration video reports that you must submit to show your incremental progress. By April 8th (at the end of the second week), you must have a demo video of a basic working platform that takes user questions and answers them to the best of its ability. This platform can be a custom Piazza client or a web-based user interface that you built with Piazza-like functions.

Every week, you must scrape and incorporate more data with increasing complexities and integrate embedding algorithms with incremental improvements. The demonstration videos must describe these new data and the algorithms that improved the system. All team members must speak during the demonstration videos.

1) Minimum Requirements

- a) A working web-based user interaction with electronic/virtual TA**
 - i) Piazza interface (automated login, real-time data collection, and answers)**
 - ii) Web forum interface (automated login, real-time data collection, and answers)**
 - iii) Custom client/server (user database/login, real-time data collection, and answers)**
- b) Answers with images and videos**
 - i) Images associated with the answers in topic documents**
 - ii) Images from associated videos (snapshots with timestamps and links)**
 - iii) Transcripts of the videos for answering questions**

2) Suggested Project Schedule/Milestones

- a) March 25: Identify specific courses with available data types.**
 - i) Tests/Images (books, PDFs, websites, Wikipedia, lecture notes, forums, blackboard, etc.)**
 - ii) Videos (lecture videos, YouTube videos, tutorials, etc.)**
 - iii) Build programs to scrape and process an initial set of data.**
 - iv) Build a database of data chunks and embeddings from the set of data.**
- b) April 1: Functional user interface**
 - i) Identify and collect new question postings to answer.**
 - ii) Automated context matching using embedding and database retrieval functions (OpenAI, huggingface, etc.)**

- iii) Integrated LLM-based answer result (GPT, LLAMA, etc.)
- iv) Build and test the Question and Answer interface (you may use whatever is available like <https://github.com/hfaran/piazza-api> or build your own)
- c) April 8: Images and Videos
 - i) Derive a way to embed or associate images and videos with the text answers.
 - ii) Display relevant images for the answers (or not display any, if there are none.)
 - iii) Display relevant video URL with time location (or not output any, if there are none.)
- d) April 15: Real-time data collection/embedding.
 - i) Identify new answers and evaluate for database viability.
 - ii) Distinguishing relatively static content (core course topics) from dynamically changing content (course logistics)
 - iii) Automatic evaluation algorithms for updating the embedding database.
- e) April 22: Improvements to the algorithms
- f) April 29: Final Touches/Demo Video
 - i) Record and post a professional quality project demonstration video
- 3) In-person Final Progress Presentations
 - a) Mandatory attendance: April 16, 18, 23, 25
 - b) Up to 7 teams per date (randomly assigned)
 - c) Expectations will be different for assigned dates
- 4) Submissions
 - a) Weekly progress demo videos due dates: April 1, 8, 15, 22, 29
 - b) Final Demo Video: May 8