# DSCI 510 – Fall 2025 – University of Southern California

Instructor: Dr. Alexey Tregubov

## Final Project Progress Report

## Title: Which Topics Are Trending Across Online Learning Platforms?

### Project Scope Update

I approached this project by examining how topic trends have shifted across Udemy and Coursera over time and comparing them to global public interest using Google Trends data through the pytrends API. I built a full pipeline that loads and cleans the raw Kaggle datasets, normalizes and aggregates topics by year, and retrieves yearly search interest for "machine learning." The pipeline now runs end-to-end and saves all analytics and visualizations directly into the `results/` folder. To keep everything organized and reproducible, I separated the code into modular components: data processing in `fp_topics.py`, API handling in `fp_trends.py`, and visualization in `main.py`.

### Data Source Table

| **Source** | **Udemy Online Education Courses (Kaggle)** | **Coursera Courses and Skills Dataset 2025 (Kaggle)** | **Google Trends (pytends API)** |
|---|---|---|---|
| **Description** | ~3,678 courses with metadata (title, subject, publish date, etc) | ~3,404 courses with subject, institution, and skills data | Public web search interest data from Google, retrieved through the pytrends library |
| **Type** | File | File | API |

## API Integration

I used the Pytrends library, an unofficial Google Trends API to pull search interest data for the keyword “machine learning.” I did this by creating a `TrendReq` object, building the payload, and calling `interest_over_time()` to retrieve weekly search interest. After collecting the raw weekly data, I aggregated it into yearly averages so it would match Udemy’s course publication years. The final outputs include both a CSV file (`trends_machine_learning_yearly.csv`) and a quick preview file (`trends_machine_learning_preview.json`), which help me verify that the data was loaded and aggregated correctly.

## Findings / Results So Far

I found that Udemy’s course distribution has been steadily shifting toward Web Development, Graphic Design, and Business Finance, while Coursera remains centered on Data Science, Computer Science, and Business. When I compared these platform trends to global Google search interest, I saw that “machine learning” has shown consistent growth since 2010. After running a regression between Udemy’s year-to-year topic share changes and prior-year Google Trends values, I got an $R^2$ of about 0.036, which tells me the correlation is weak but still detectable. All of the processed data and visualizations from this analysis are saved in `results/outputs/` and `results/figs/`.

## Issues / Difficulties

I noticed that the Coursera dataset includes only a small range of publication years, which made it harder to line up both platforms for a clean, long-term comparison. While pulling data from the Google Trends API, I also ran into rate limits (HTTP 429), so I added retry delays in `fp_trends.py` to keep the pipeline stable. On top of that, I had to manually align topic labels across Udemy and Coursera since their category systems don’t match directly, and that extra normalization step took some careful work.

## How to Run

When I run:

```
cd src
Python -m src.main
```

It triggers the entire pipeline from end to end. The script loads the datasets, cleans and standardizes the topic labels, pulls the Google Trends data, and then generates all of the CSV files and plots. By the time it finishes, everything is neatly saved in the `results/` folder, so I can review the outputs without running any extra steps.

Next step, limitation

The Coursera dataset represents a 2025 snapshot rather than a multi-year catalog, limiting direct trend comparison across time. However, by contrasting Udemy's historical topic evolution (2011–2017) with Coursera's 2025 distribution, I highlight both long-term and current shifts in online education focus. Future work could extend this analysis using Coursera's historical data if available.