**Project Outline: Student Clustering via Course Keywords and LDA**

**GitHub Repository:** [CBS-Github-DND](https://github.com/dyuthiii/CBS-Github-DND.git)

**1. Project Goal**

This project explores how students’ elective course choices at Columbia Business School can be analyzed and grouped into **“interest clusters”**. By extracting meaningful keywords from course titles and descriptions and then applying **Latent Dirichlet Allocation (LDA)** (a topic modeling technique), we can reveal hidden patterns in how students pick courses.

In simpler terms: the idea is to see if students naturally fall into groups like *“finance-heavy electives”*, *“entrepreneurship & innovation”*, or *“data & analytics”*—without us pre-defining those categories.

**2. Workflow Overview**

The project has three main stages, each tied to the documents you uploaded:

**A. Initial EDA (Exploratory Data Analysis)**

* Looked at basic descriptive statistics:
  + Which elective courses are most popular?
  + Which instructors appear most often?
  + How course popularity shifts over time or across student subgroups.
* Produced some **correlation checks** and thought about visualizations like a heatmap to see which courses “travel together” (i.e., are often taken by the same students).
* Identified and cleaned **data quirks** (e.g., date formats showing up as course names, missing student IDs).

**B. Keyword Extraction & NLP Prep**

* Created **clean course names** (combining SIS vs. warehouse names).
* Broke down course names and descriptions into **unigrams, bigrams, trigrams** (single words, two-word combos, three-word combos).
* Applied **TF-IDF** to highlight which phrases stand out in courses.
* Experimented with **RAKE** and **YAKE** keyword extractors:
  + YAKE worked well for short texts like course titles.
  + Cleaned outputs manually (e.g., removing terms like “Half Term” or “NYC Summer Immersion Seminar” that don’t add meaning).
* Built lookup tables to decide which keywords to keep and which to drop.

**C. Student Clustering with LDA**

* **Pipeline steps:**
  1. Normalize and lemmatize keywords (so “investing” → “invest”).
  2. Build a **student × keyword matrix** (rows = students, columns = keywords, values = counts).
  3. Run **LDA topic modeling** to discover latent themes (“topics”).
  4. Assign each student a distribution of topics (e.g., 60% Finance, 40% Entrepreneurship).
* Implemented a **fallback system** to ensure no course is left untagged:
  1. If title keywords fail → use description keywords.
  2. If those fail → pull from YAKE’s top candidates.
* Exported CSVs to make the results auditable and reusable:
  1. lda\_student\_topic\_distribution.csv → what cluster each student belongs to.
  2. lda\_topic\_term\_weights.csv → top keywords per cluster.
  3. fallback\_na\_resolution\_log.csv → record of how missing values were filled.

**3. Why It Matters**

* **For administrators**: Helps understand trends in student interests over time.
* **For instructors**: Identifies how their courses “fit” within broader themes.
* **For students**: Could eventually support personalized course recommendations.

**4. Next Steps**

* Improve keyword coverage by refining the **keep map** (the curated keyword list).
* Experiment with **guided LDA** (seeding known themes like “Finance” or “Entrepreneurship”).
* Explore “pathways” created by the LDA to look at Future Job Function and Future Job.
* Build dashboards for **visual exploration** (e.g., topic prevalence across years).

✅ In short: This project takes messy course text, cleans it into usable keywords, and then uses LDA to uncover hidden clusters of student interests. The GitHub repo holds the code and notebooks that make this process reproducible.