Test #3: extractor.py (V1 vs V2)

Text, OCR/Slide Segmentation, and Vision Fallback

August 14, 2025

1 Goal

Compare two concept-extraction pipelines on lecture PDFs and document practical trade-offs:

- **V1 (baseline)**: sentence → embeddings → agglomerative clustering → cluster representative → LLM to name a concept and give a 1–5 difficulty.
- **V2 (improved precision)**: V1 backbone + (i) admin/noise filtering, (ii) anchor-similarity gate to keep only ML/DL/VLM concepts, (iii) de-dup, (iv) difficulty backstops.

2 Colab & Data (summary)

Secrets: `OPENAI_API_KEY` via Colab "Secrets". Files: Google Drive folder with `DS301_Lecture-5_NYU.pdf` (text-friendly) and `DS-UA-301_Lecture-12-NYU.pdf` (image-heavy). Models: `text-embedding-3-small` for vectors; `gpt-4o-mini` for concepts.

3 Method

3.1 Shared steps (V1/V2)

Extract text (PyMuPDF), segment, embed, cluster, choose a representative per cluster, ask the chat model for *concept + one-sentence explanation + difficulty*. V2 adds: regex noise guard (e.g., "new documentation", "meeting notes"), domain anchor gate (cosine similarity vs. curated anchors), and de-duplication.

3.2 Slide/OCR adaptations

Slide decks often lack punctuation and contain short headers. We therefore:

- Use a **slide/bullet segmenter**: split on newlines and bullet markers; keep short headers (≥8–10 chars).
- Use **OCR fallback** (Tesseract) when native text is sparse; merge OCR text with native text.
- **Adapt** cluster count and similarity threshold for short inputs (fewer segments ⇒ fewer clusters; slightly lower gate, e.g. $\tau = 0.62$).

3.3 Vision fallback (optional)

If after OCR/segmentation text is still too weak, rasterize a few pages and query a vision LLM to extract 3–10 canonical, 1–4 word concept bullets, then run the same V2 filter on the synthesized text. (Not required in the latest run, but kept as a robust option.)

4 Results

doc_id	V1 Count	V2 Count	V1 Spurious	V2 Spurious	V1 Time (s)	V
DS301_Lecture-5_NYU.pdf	12	6	6	0	28.27	

Table 1: Text-friendly deck: V2 removes spurious items with a modest latency increase.

PDF A: `DS301_Lecture-5_NYU.pdf` **(text-friendly).**

PDF B: `DS-UA-301_Lecture-12-NYU.pdf` **(image-heavy slides).** Latest run (with slide segmentation + OCR; no vision needed):

doc_id	V1 Count	V2 Count	V1 Spurious	V2 Spurious	V1 Time (s
DS-UA-301_Lecture-12-NYU.pdf	4	0	1	0	8.3

Table 2: Image-heavy deck: V1 finds a few headings (one flagged spurious); V2 is conservative and returns zero, preserving precision.

5 Analysis & Thinking

Why V2 shines on PDF A. The text is continuous; headings and stray admin strings appear but are filtered by V2. The anchor gate passes true ML/DL items (e.g., RAG, Dimensionality Reduction) and rejects generic ones, hence Spurious=0.

Why V2 can be empty on PDF B. Slides are mostly figures and terse headers. Even after OCR, segments are short and context-poor. The anchor gate favors precision and rejects borderline labels; V1, which lacks the gate, surfaces a few labels but includes one spurious item.

What we tried. (1) Bullet/slide segmentation to keep short titles; (2) OCR at 300 dpi to recover text from images; (3) adaptive cluster count and threshold for short inputs; (4) vision fallback (kept as an option). The latest configuration yields acceptable precision on the image-heavy deck, albeit with lower recall (Table 2).

6 Trade-offs & Guidance

- **Precision vs. recall**: V2 is intentionally conservative. On text-friendly PDFs, this is ideal (Table 1). On image-heavy decks, it may return zero concepts while V1 still emits a few labels (Table 2).
- **Latency/cost**: V2 adds post-filtering; OCR adds rasterization time; vision adds API cost and the largest latency. Use them conditionally.
- **Practical policy**:
  1. Run V2 on extracted text (with slide segmentation).
  2. If V2 returns 0 and the deck is clearly slide/figure-heavy, either (a) lower the similarity gate slightly (e.g., $\tau = 0.62$) *or* (b) trigger the vision fallback on a small subset of pages.
  3. If course/topic is known, **extend anchors** with course-specific terms (e.g., VLM: Flamingo, Perceiver Resampler, gated cross-attention) to improve recall without sacrificing precision.

7 Conclusion

V2 should be the default: it consistently removes unrelated/admin topics and yields clean concept sets on text-friendly material. For image-heavy slides, the best results come from the slide/OCR pathway with an adaptive gate; when recall is still insufficient, enable the vision fallback on a few representative pages to recover core concepts.

Appendix: Key Settings

- Embeddings: `text-embedding-3-small`. Chat: `gpt-4o-mini`.
- Clustering: Agglomerative; clusters auto-scaled by segment count for slides.
- Noise filter (examples): `new documentation`, `commit message`, `meeting notes`.
- Anchors (subset): ML/DL basics (transformers, attention, embeddings, RAG, vector DB, dimensionality reduction, fine-tuning) + VLM terms (Flamingo, Perceiver Resampler, gated cross-attention, CLIP/ALIGN).
- Thresholds: $\tau = 0.68$ (text-rich), $\tau \approx 0.62$ (short slide segments).
- OCR: Tesseract at 300 dpi; merged with native text. Vision fallback kept as optional last resort.