역사 텍스트 디지털화 입문

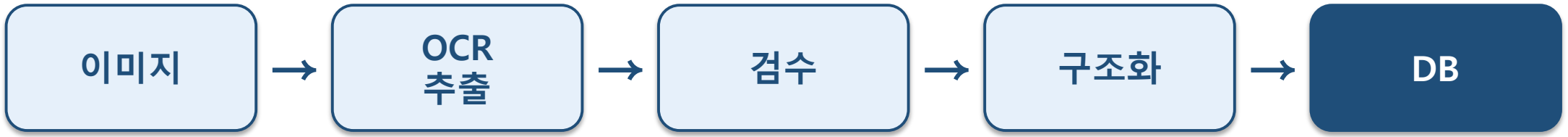
OCR에서 데이터베이스까지

이미지 → OCR 추출 → 검수 → 구조화 → 데이터베이스

2026년 3월 3일 | 줌 비대면 특강

도구: Google AI Studio (Gemini) | 준비물: Google 계정

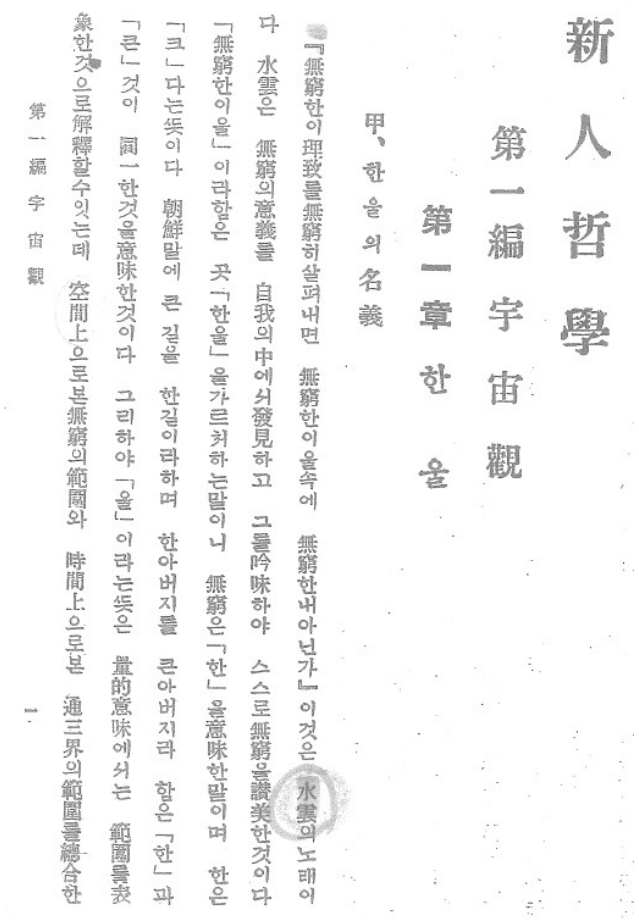
오늘의 목표



강의가 끝나면: **내 손으로 만든 검색 가능한 문장 DB 1개**

도구: Google AI Studio (aistudio.google.com)
준비물: Google 계정 (AI Studio + Google Sheets 공용)
방식: 강사 화면을 보며 동시에 따라하기

라이브 시연: AI가 읽습니다



新人哲學

第一編 宇宙觀

第一章 한 울

甲、한울의 名義

『無窮한이理致를無窮히살펴내면 無窮한이울속에 無窮한내아닌가』이것은 水雲의노래이다 水雲은 無窮의意義를 自我의中에서發見하고 그를吟味하야 스스로無窮을讚美한것이다 「無窮한이울」이라함은 곳「한울」을가르침하는말이니 無窮은「한」을意味한말이며 한은「크」다는뜻이다 朝鮮말에 큰 길을 한길이라하며 한아버지를 큰아버지라 함은「한」과「큰」것이 同一한것을意味한것이다 그리하야「울」이라는뜻은 量的意味에서는 範圍를表象한것으로解釋할수잇는데 空間上으로본無窮의範圍와 時間上으로본 通三界의範圍를總合한

第一編 宇宙觀 一

1931년 국한문혼용 세로쓰기 철학서

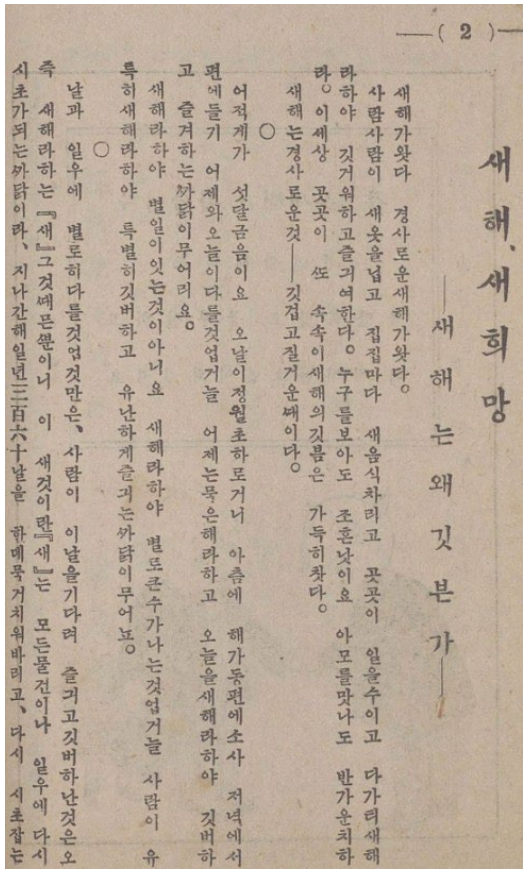
이돈화, 《新人哲學》 (천도교중앙종리원, 1931)

→ AI가 이것을 읽어냅니다

AI는 "자연어로 지시하는 연구 조수"
코딩을 배우는 것이 아닙니다
→ 무엇을 원하는지 말할 줄 알면 됩니다

⚠ 단, AI는 그럴듯하게 틀릴 수 있음
→ 최종 판단은 반드시 연구자 본인

실습 ① OCR: 어린이「새해 새희망」(1924)



—(2)—

새해、새희망

—새해는 왜 깃분가—

새해가왓다 경사로운새해가왓다。

사람사람이 새옷을닙고 집집마다 새음식차리고 곳곳이 일을수이고 다가리새해라하야 깃거워하고즐기여한다。누구를보아도 조흔낫이요 아모를만나도 반가운치하라。이세상 곳곳이 또 속속이새해의깃븜은 가득히찻다。

새해는경사로운것——깃겁고즐거운째이다。

○

어적게가 섯달금음이요 오날이정월초하로거니 아츰에 해가동편에소사 저녁에서편에들기 어제와오늘이다를것업거늘 어제는묵은해라하고 오늘을새해라하야 깃버하고 즐겨하는까닭이무어리요。

새해라하야 별일이잇는것이아니요 새해라하야 별로큰수가나는것업거늘 사람이 특별히새해라하야 특별히깃버하고 유난하게즐겨하는까닭이무어뇨。

○

날과 일우에 별로히다를것업것만은、사람이 이날을기다려 즐겨하고깃버하난것은오즉 새해라하는『새』그것때문이니 이 새것이란『새』는 모든물건이나 일우에 다시시초가되는까닭이라、지나간해일년三百六十날을 한묵거치워바리고、다시 시초잡는

이 이미지는 1924년에 발행된 어린이 잡지입니다.
세로쓰기 텍스트를 오른쪽→왼쪽 순서로,
원문 그대로 추출해 주세요.

AI Studio 세팅 (첫 실습 전 한 번만)

① 모델: Gemini 3.1 Pro (화면 상단)
② Temperature: 0.0 (오른쪽 Run settings)
③ Safety Settings: Off (오른쪽 하단)

※ 이미지 업로드: 채팅창 ＋ 버튼 또는 드래그 앤 드롭

OCR 결과 검수: 원문 대조는 필수

옛 표기 ≠ OCR 오류

☑ 정확히 읽은 것 (옛 표기 그대로)
"깃븐가" → 기쁜가 | "왓다" → 왔다
"닙고" → 입고 | "조흔" → 좋은
1920년대 표기법이 지금과 다를 뿐입니다

❌ **AI가 틀릴 수 있는 부분**

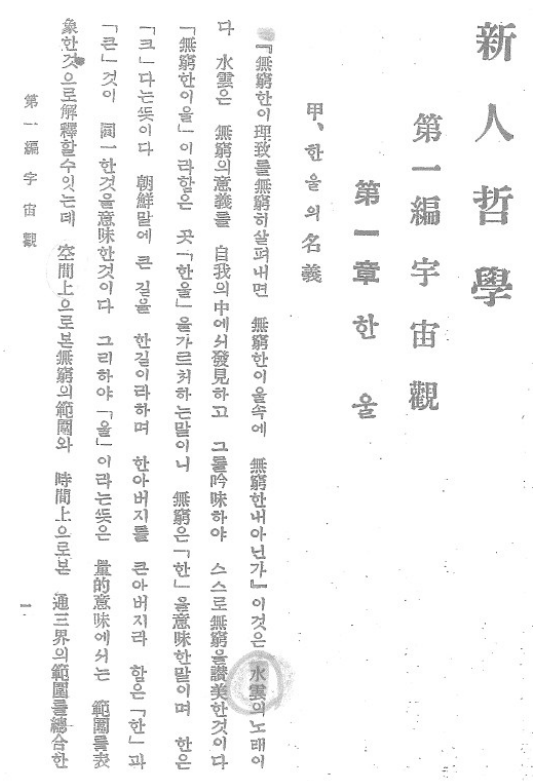
- 비슷한 한자 혼동: 大↔太, 己↔己, 日↔日
- 글자 누락 또는 없는 글자 생성

🔍 **검수 원칙**

1. 결과를 메모장에 복사
2. 원문 이미지와 나란히 놓고 대조
3. 글자 단위로 확인

"반드시 원문과 대조 검수가 필요합니다"

실습 ② OCR: 신인철학 C001·C002 (1931)



新人哲學

第一編宇宙觀

第一章한 울

甲、한 울의 名義

『無窮한이理致를無窮히살펴내면 無窮한이울속에 無窮한내아닌가』이것은 水雲의노래이다 水雲은 無窮의意義를 自我의中에서發見하고 그를吟味하야 스스로無窮을讚美한것이다 「無窮한이울」이라함은 곳「한울」을가르쳐하는말이니 無窮은「한」울意味한말이며 한은「크」다는뜻이다 朝鮮말에 큰길을 한길이라하며 한아버지를 큰아버지라 함은「한」과「큰」것이 同一한것을意味한것이다 그리하야「울」이라는뜻은 量的意味에서는 範圍를表象한것으로解釋할수잇는데 空間上으로본無窮의範圍와 時間上으로본 通三界의範圍를總合한

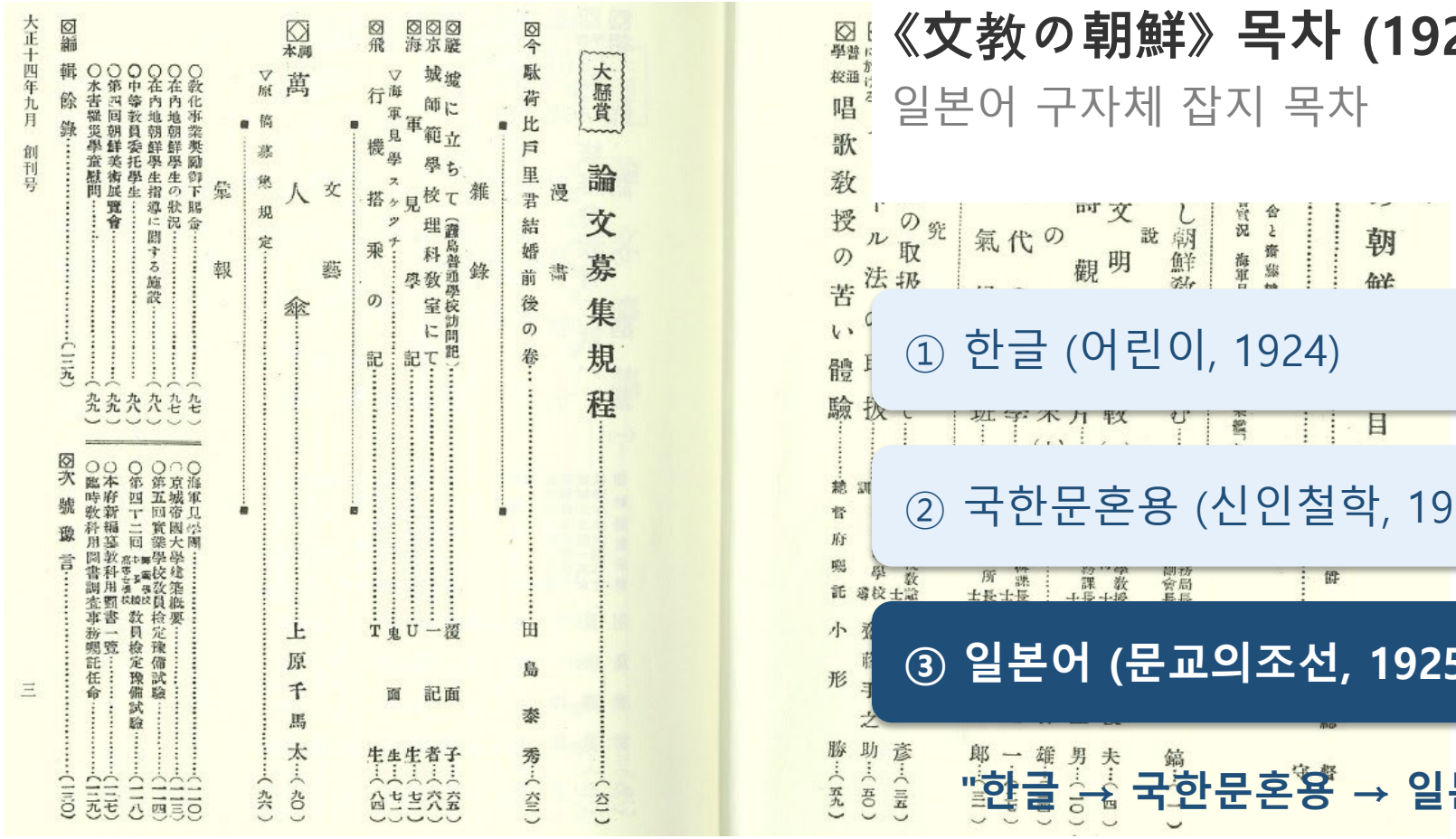
新人哲學 二

宇宙全體를 가르켜「울」이라말한것이니 이意味에서울이라는것은 宇宙의全的全量을가르켜하는말이며 質的意味에서 이「울」이라함은「우리」라는뜻이니 우리집 우리民族 우리人類라는것과같은「우리」이나 우리는곳나와同類를包含하야말하는것임으로 한울은곳「큰나」라는뜻으로解釋할수잇다 이를漢字로말하면 한울은大我라는뜻으로 個體되는小我에對하야 小我와 關係되는「우리」를大我라 名稱한것이다 朝鮮에서 神的對象을普通으로「한우님」「하나님」이라이름하나 이것은朝鮮말의語源에 어울어진것이며 設使그러한말이朝鮮固有語라할지라도 이것은우리가말하는「한울」이라는對象과는 槪念에서甚遠한差異點이잇다 「하나님」이라는名辭는 基督敎信者들이 붙으는이름으로 이것은 純全히一神敎的人格的上帝를가르키불으는이름이오 「한울」이라함은人格的神을 가르쳐하는말이아니라 部分에對한全的이라는意味이며 小我에對한大我라는이름이오 一層나아가 宗敎的名稱으로볼지라도 「한울」은汎神觀的이며 萬有神觀으로解釋할수잇고 또는이것을哲學的又는科學的으로본다면 小我에對한大我는 때를따르고 곳을따라 方便的으로使用할수잇으니 한個人이自己團體를생각할때에 이를大我라볼수잇으며 한百姓이 自己國家를 생각할때에이를大我라할수잇고 한사람이社會를생각

이 이미지는 1931년에 출판된 국한문혼용 세로쓰기 철학서입니다.
세로쓰기 텍스트를 오른쪽→왼쪽 순서로, 원문 그대로 추출해 주세요.
한자는 원문 그대로 표기하세요.
문장이 끝나지 않았으면 줄바꿈 없이 이어서 타이핑하세요.

난이도 상승: 국한문혼용 세로쓰기
- C001 → C002 순서로 각각 OCR
- 결과를 메모장에 이어 붙이기
- 이 텍스트가 후반부 구조화 입력 자료

시연: 일본어 문헌도 읽힙니다


大正十四年九月　創刊号

大懸賞　論文募集規程……（六二）

漫畫

今駄荷比戶里君結婚前後の卷……田島泰秀（六三）

雜錄

巖邊に立ちて（鹽島普通學校訪問記）……覆面子（六五）
京城師範學校理科教室にて……一記者（六八）
軍見學スケッチ……U生（七二）
飛行機搭乘の記……鬼面生（八四）

文藝

萬人傘……上原千馬太（九〇）

彙報

教化事業奬勵御下賜金……（九七）
在內地朝鮮學生の狀況……（九七）
在內地朝鮮學生指導に關する施設……（九八）
中等教員委托學生……（九八）
第四回朝鮮美術展覽會……（九九）
本府新編纂教科用圖書一覽……（一二七）
水害罹災學童慰問……（九九）
臨時教科用圖書調査事務嘱託命……（一二九）

編輯餘錄……（一二九）

次號豫言……（一三〇）

普通學校唱歌教授の苦い體驗……總督府嘱託小形勝（五九）

《文教の朝鮮》목차 (1925.09)

일본어 구자체 잡지 목차

① 한글 (어린이, 1924)

② 국한문혼용 (신인철학, 1931)

③ 일본어 (문교의조선, 1925)

"한글 → 국한문혼용 → 일본어까지, AI가 읽어냈습니다"

전반부 정리

① **어린이 (1924)** 한글 세로쓰기 [쉬움]

② **신인철학 (1931)** 국한문혼용 세로쓰기 [중간]

③ **문교의조선 (1925)** 일본어 구자체 [시연]

OCR은 완벽하지 않습니다 — 반드시 원문과 대조 검수가 필요합니다

▶ 후반부: 신인철학 텍스트를 검색 가능한 문장 단위 DB로 만듭니다

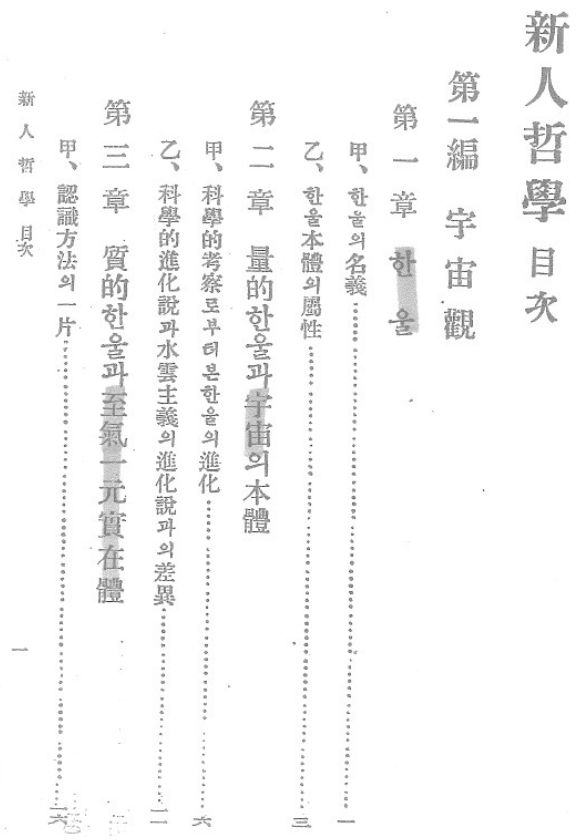
10분 휴식

Google AI Studio에 로그인이 아직 안 되신 분은 이 시간에 시도해 주세요
aistudio.google.com

Google Sheets도 열어두시면 좋습니다
sheets.google.com

전반부에서 신인철학 C001+C002 OCR 결과를
메모장에 저장해 두셨는지 확인해 주세요

구조화란 무엇인가


新人哲學 目次

第一編 宇宙觀

第一章 한울

甲、한울의名義 …… 一

乙、한울本體의屬性 …… 三

第二章 量的한울과宇宙의本體

甲、科學的考察로부터본한울의進化 …… 六

乙、科學的進化說과水雲主義의進化說과의差異 …… 二

第三章 質的한울과至氣一元實在體

甲、認識方法의一片 …… 二六

新人哲學 目次 一

텍스트 뭉치 → 문장 DB = 검색·분석 가능

서류 더미 → 엑셀 표로 정리하는 것과 같습니다

《新人哲學》 계층 구조 (A01 목차 기준)

第一編 宇宙觀 → E01-C00-S00-P00-S00

第一章 한울 → E01-C01-S00-P00-S00

甲, 한울의名義 → E01-C01-S01-P00-S00

乙, 한울本體의屬性 → E01-C01-S02-P00-S00

문장 한 줄 한 줄이 표의 한 행이 됩니다

문장 DB: 4열 구성

열 이름	설명	예시
local_id	계층적 위치 (E=편, C=장, S=절, P=문단, S=문장)	E01-C01-S01-P01-S01
page_info	원본 페이지 태그	C001
line_class	행 유형	STRUCT / TEXT
kr_text	텍스트 내용	『無窮한이理致를…』

이 4열이면 검색·필터·정렬이 모두 가능합니다

(완성 DB는 10열이지만, 핵심은 이 4열 — 나머지는 필요할 때 추가)

완성 DB 시연: 인내천요의 2,254행

같은 저자 이돈화의 다른 책《人乃天要義》(1924)

10열 2,254행의 완성된 문장 DB

Google Sheets에서 시연할 내용:

① 필터: line_class로 STRUCT만 보기 → 목차 구조가 한눈에
② 검색: Ctrl+F → "한울" → 해당 문장들이 바로 찾아짐
③ 정렬: local_id로 정렬하면 원문 순서 복원

▶ "오늘 신인철학 첫 두 페이지로 이것을 시작합니다"

실습 ③ 구조화: 신인철학 → 문장 단위 TSV

다음은《新人哲學》(이돈화, 1931) 본문의 OCR 텍스트입니다.
이 텍스트를 문장 단위로 분절하여 탭으로 구분된 TSV 형식으로 출력해 주세요.

열 구성:
- local_id: 계층적 위치 (예: E01-C01-S01-P01-S01)
  E##=편, C##=장, S##=절, P##=문단, S##=문장
- page_info: 원본 페이지 태그 (예: C001)
- line_class: STRUCT(제목/소제목) 또는 TEXT(본문)
- kr_text: 텍스트 내용

규칙:
- 편명, 장명, 절명은 line_class를 STRUCT로 분류
- 본문은 의미가 완결되는 문장 단위로 분절하여 각각 한 행으로
- 첫 줄은 헤더(열 이름)로 출력 / 구분자는 탭(Tab)

C001 → C002 순서로 처리 (헤더 행은 한 번만)

⚠ **"문장 경계가 여러분과 다를 수 있습니다"**
→ 이것이 연구자 판단이 필요한 지점

DB 완성: Google Sheets

① 붙여넣기

- Google Sheets 새 스프레드시트 열기
- AI가 출력한 TSV 전체 복사 → A1 셀에 붙여넣기
- 탭 구분자로 열이 자동 분리되는 것 확인

② 필터 + 검색

- 첫 행 선택 → 데이터 → 필터 만들기
- line_class 필터: STRUCT만 → 목차 구조 추출
- Ctrl+F → "한울" 검색 → 문장 바로 찾기

③ 성과 확인

- 이 스프레드시트 = 내가 만든 신인철학 문장 DB
- 같은 방법 → 책 전체 → 2,000행 이상

🎉 **"내 자료가 검색 가능한 데이터베이스가 되었다"**

정리 + 다음 단계

① 이미지 업로드 → OCR → ② 원문과 대조 검수 → ③ 문장 단위 구조화 (TSV) → ④ Sheets → 필터/검색 DB

다음 단계 안내

단계	내용	배포 자료
다음 단계 1	대량 정밀 OCR + 번역	일본 사상서 판독 및 번역 가이드
다음 단계 2	텍스트 정제·일괄 편집	VS Code 멀티커서 가이드
다음 단계 3	논문/자료 코퍼스 관리	VS Code 논문 검색 가이드

모든 자료(핸드아웃, 이미지, 결과물, 다음 단계 가이드)는 사후 배포됩니다