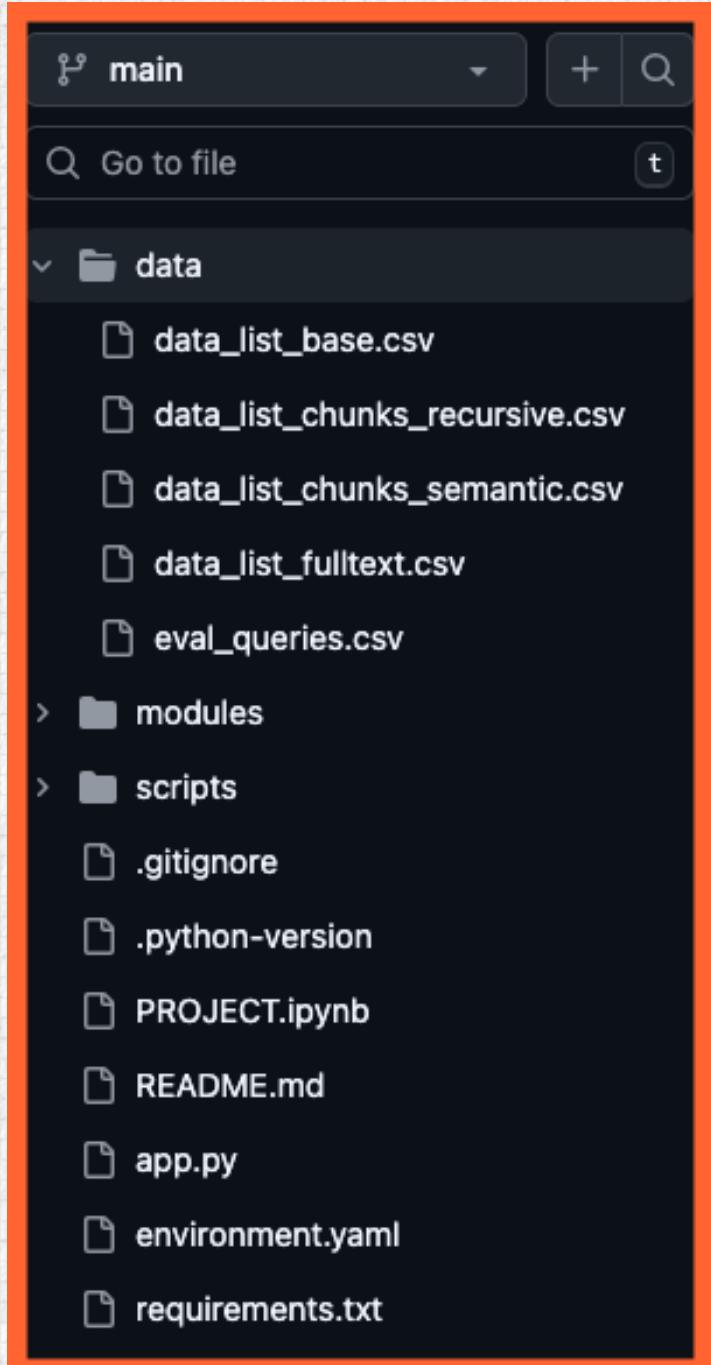


# 메타데이터 CSV, 왜 중요할까?



base.csv

✓ "어디에 어떤 PDF가 있는지 설계도"

chunks.csv

✓ "실제 RAG 검색에 쓰는 텍스트 조각들"

fulltext.csv

✓ "PDF에서 글자만 뽑아온 원문 저장소"

보다  
효율적인  
학습 가능

# 쉽게 비유를 들자면,

**base.csv**

넷플릭스  
작품리스트 DB

작품ID, 제목, 장르, 개봉일, 러닝타임, 링크 등등  
"넷플릭스에 어떤 컨텐츠들이 올라와 있는가?"  
→ 넷플릭스 작품 목록이 base.csv

**fulltext.csv**

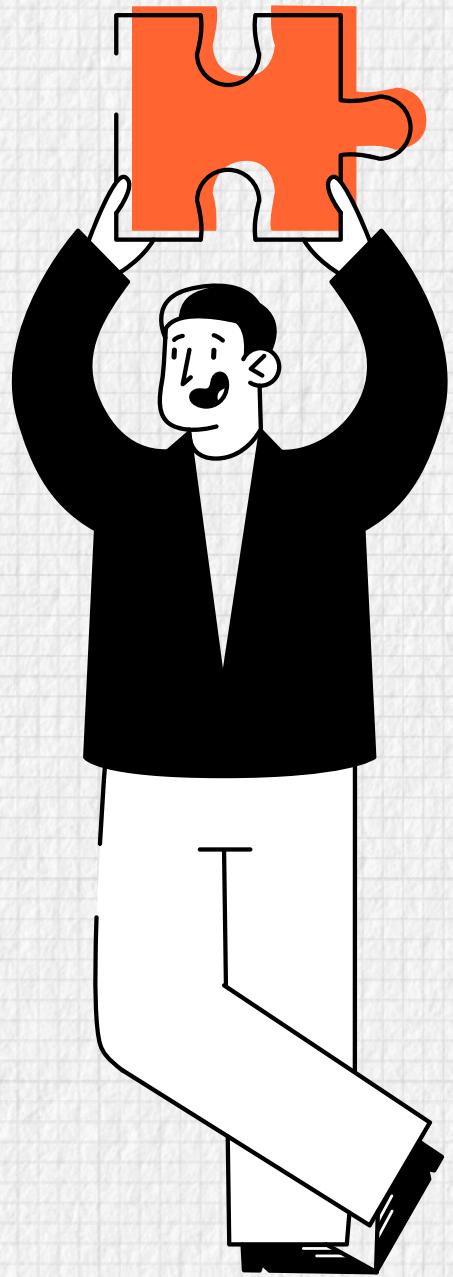
영화/드라마  
풀 영상 파일

영화 한편 전체 mp4  
드라마 한 시즌 전체 영상  
→ RFP 문서 1건 전체 텍스트 = "풀 타임 영상"

**chunks.csv**

하이라이트 클립들  
+ 메타데이터

명장면만 2분씩 잘라서 클립으로 만들어 둔 것들  
\* 이때 각 클립에 작품ID, 시즌, 에피소드, 타임스탬프 정보를 붙임  
→ 클립 단위로 검색해서 top-k만 골라서 LLM에 보여줌



# 만약 하나라도 없으면?

base.csv

- RFP 목록관리
- 파일 경로 관리
- API 결과 매핑용 테이블

fulltext.csv

- PDF 파싱 결과를 한번 만 해두고 계속 재활용하기 위한 캐시

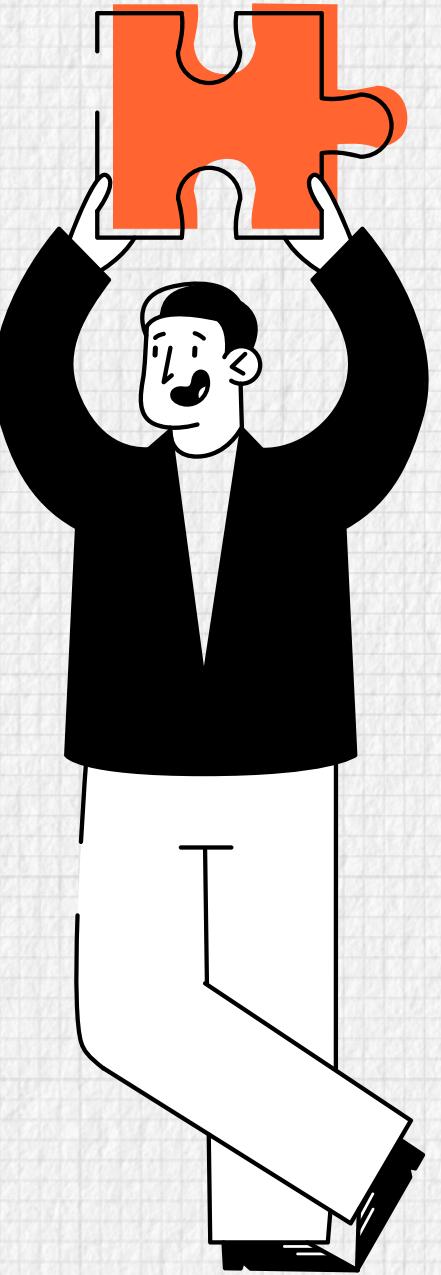
chunks.csv

- RAG 인덱스 테이블

사용자가 "이 정보 어디서 가져왔어?" 라고 물을때 원본 PDF 링크를 줄수가 없음.  
→ **족보 없는 답변**이 됨.

나중에 Chunk 크기를 바꾸고 싶을때, 다시 처음부터 PDF 수백장을 OCR 해야함  
→ 시간과 비용이 엄청나게 소모

질문을 할때마다 전체 텍스트를 AI에게 통째로 읽혀야 함  
→ AI가 너무 길어서 못읽는다고 리젝하거나 비용이 수십배로 나옴.



**즉, 실무에서는 청크 크기를 다시 바꾸거나, 메타데이터 칼럼을 늘리거나, 로그를 분석해야 하기 때문에 3단 구조로 나눔**

# Contents

# 공고번호 누락 이수 & 해결

#	Column	Non-Null Count
0	공고 번호	82 non-null
1	사업명	100 non-null
2	사업 금액	99 non-null
3	발주 기관	100 non-null
4	공개 일자	100 non-null
5	입찰 참여 시작일	74 non-null
6	입찰 참여 마감일	92 non-null
7	사업 요약	100 non-null
8	파일형식	100 non-null
9	파일명	100 non-null



	Column	Non Null Count
0	공고 번호	98 non-null
1	공고 차수	98 non-null
2	사업명	98 non-null
3	project_id	98 non-null
4	사업 금액	98 non-null
5	발주 기관	98 non-null
6	공개 일자	98 non-null
7	입찰 참여 시작일	98 non-null
8	입찰 참여 마감일	98 non-null
9	사업 요약	98 non-null
10	doc_index	98 non-null
11	공개 월	98 non-null

파라미터	의미
serviceKey	개인 인증키
pageNo	페이지 번호
numOfRows	한 페이지 결과 수
inqryDiv	조회 구분
inqryBgnDt	조회 시작일시
inqryEndDt	조회 종료일시
bidNtceNo	입찰 공고 번호
type	응답 포맷

총 1건

---

공공행정 조달청

활용신청 [승인] 조달청\_나라장터 입찰공고정보서비스

계정 개발 신청일 2025-12-12 만료예정일 2027-12-12

나라장터 API 보강 대상 행 개수 (전체): 29	
-	공고번호 기반(ID) 조회 대상: 17
-	사업명/기관명 기반 조회 대상: 12
공고 번호	사업명
0 20241001798	한영대학교 특성화 맞춤형 교육환경 구축 - 트랙운영 학사정보시스템 고도화
12	NaN [사전공개] 학업성취도 다차원 종단분석 통합시스템 1차 고도화 용역
13	NaN [입찰공고] 산학협력단 정보시스템 운영 용역업체 선정
14	NaN 건설통합시스템(CMS) 고도화
16	NaN 예약발매시스템 개량 ISMP 용역

- 제공된 CSV를 보면 "공고 번호"가 비어있거나 이상한 값인 행이 몇개 있었음.
  - 옆의 표는 나라장터 API 파라미터인데, "공고 번호"가 비어있으면 해당 RFP는 API로 검증/보강이 불가하고, 나중에 RAG 결과를 원본 공고와 연결하기도 어려움.
  - 그래서, 누락된 공고번호는 나라장터에서 수동 검색하여 수기로 "공고 번호"만 찾아내서 기존 CSV에 덮어씌웠고, 원문 PDF와 매핑하였습니다.
  - 처음에는, 사업명은 누락된것이 없었기 때문에 "공고번호가 없는 행은 사업명으로 필터링해서 정보를 가져와" 하는 조건을 했었지만, API를 호출하는 코드를 작성중에, 사업명이나 특정이름에 대한 파라미터가 없는지를 모르고 시행착오를 겪었음.

# UI 프로토타입

KAG 네고 (더미)

- 현재는 로컬 실행 확인용 더미(generator/judge)를 씁니다.
- 검색(Qdrant) 결과 Top-k를 보여주고, 더미 답변을 출력합니다.

질문

고려대학교 관련문서좀 줘

RAG Top-k(답변에 사용할 문서 수)

5 3

질문하기

1 10

답변(더미)

※ [DUMMY ANSWER] (로컬 실행 확인용)  
질문: 고려대학교 관련문서좀 줘

참고한 문서 Top-k 요약:

- [1] doc\_id=20240541684, chunk\_id=123, source=proj\_063 :: 「소프트웨어 진흥법」제50조제3항에 따라 위와 같이 소프트웨어사업 과업 내용변경을 요청합니다. 년 월 일 신청인 (서명 또는 인) 한국농어촌공사의 장 귀하 <table><thead></thead><tbody><tr><td>처 리 절 차</td></tr><tr><td>과업심의위원회 심의결과 및 조치계획 신청서 작성 → ...
- [2] doc\_id=B5202401500, chunk\_id=169, source=proj\_023 :: 수지점을 설정하고 개발 방안을 수립하여야 한다. # (1) 취수지점 조사 - ① 취수지점의 선정은 안정적인 취수확보, 민원발생가능성, 친환경적인 설계, 취수 - 원 이전 가능성, 운영관리의 용이성, 경제성 등을 종합적으로 검토하여 최적의 위치를 선정하여야 한다. - ② 취수지점은 수리권 확보가 가능하고 장래의 유로변경...
- [3] doc\_id=20240633082, chunk\_id=9, source=proj\_074 :: # [FILE: 경기도 안양시 호계체육관 배드민턴장 및 탁구장 예약 시스템 구축 용역.pdf] # 제 안 요 청 서 | 사업 명 | 호계체육관 배드민턴장 및 탁구장 예약 시스템 구축 ||...| - || 주관부서 | 체육과 || 사업 담당자 | 시설?급 박종일 | 2024. 6. # 목 차 | · 사업의 개요...

최종 답변(더미): 위 문서들을 근거로 요구사항/조건/유지보수/네트워크 등 항목이 포함됩니다.

참고 문서 Top-k(요약)

[1] doc\_id=20240541684 | chunk\_id=123 | source=proj\_063 | mode=recursive  
「소프트웨어 진흥법」제50조제3항에 따라 위와 같이 소프트웨어사업 과업 내용변경을 요청합니다. 년 월 일 신청인 (서명 또는 인) 한국농어촌공사의 장 귀하 <table><thead></thead><tbody><tr><td>처 리 절 차</td></tr><tr><td>과업심의위원회 심의결과 및 조치계획 신청서 작성 → 접수 → 심의·의결 개최 통보 신청인 처리기 관: 한국 농 어 촌 공 사</td></tr></tbody></table> |

[2] doc\_id=B5202401500 | chunk\_id=169 | source=proj\_023 | mode=recursive  
수지점을 설정하고 개발 방안을 수립하여야 한다. # (1) 취수지점 조사 - ① 취수지점의 선정은 안정적인 취수확보, 민원발생가능성, 친환경적인 설계, 취수 - 원 이전 가능성, 운영관리의 용이성, 경제성 등을 종합적으로 검토하여 최적의 위치를 선정하여야 한다. - ② 취수지점은 수리권 확보가 가능하고 장래의 유로변경, 부유물 및 퇴사 등에 의 - 한 취수 장애가 없는 곳이라야 하며, 공사 및 유지관리 측면을 고려하여 선정하 - 여야 한다. - ③ 취수지점의 수질은 공업용수 수질기준에 적합하게 정수처리를 할 수 있는지를 - 판단하는데 중요한 자료를 제공할 수 있어야 하며, 각종 취수시설을 위한 충분 - 34 -

그라디오

Gradio ~~★~~

- 파이썬 함수만 있으면 바로 웹 UI 생성(코랩에서도 가능)

- 링크 1개로 팀원 및 멘토와 공유가 가능

- 최대 문서수, 문서분할방법, 청크크기, 청크오버랩, 임베딩 모델 등을 선택가능하게 하여 실시간 성능평가가 가능하였음.

- 성능 평가 또한 답변 박스에 1-10 점수를 주게 만들어 스스로 평가하게 하였고, 질문 csv 답안지를 만들어서 @ 점수를 매기고 평가를 할수 있었음.