SK네트웍스 **Family AI**과정 **6**기

데이터 전처리 인공지능 데이터 전처리 결과서

☐ 개요

- 산출물 단계 : 데이터 전처리
- 평가 산출물 : 인공지능 데이터 전처리 결과서
- 제출 일자 : 2025.02.27
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-2Team
- 작성 팀원 : 이세화

데이터 전처리 개요	**데이터 수집일자** - 2025.02.12 (기존 크롤링 데이터를 벡터화 진행) - 2025.02.27 (추가 크롤링 진행 중, 이 데이터를 벡터화하여 최종적으로 사용 예정) **데이터 양** **[학습 data]** - 웹툰: 23,585 개 웹소설: 156,973 개 - 테스트를 마무리하고 최종 테스트 단계부터 사용될 예정. 배포 시에는 이 데이터들 사용. **[테스트 data]** - 프롬프트 엔지니어링을 효율적으로 진행하기 위해 sample 데이터를 사용 플랫폼(네이버, 네이버시리즈, 카카오, 카카오페이지)별로 각 장르에서 20개씩 가져온 데이터로 테스트 작업 진행. - 장르와 플랫폼들의 다양성을 잘 인식하는지 실험해보고자 데이터 양을 기준으로 나누기보단 각각에서 일부를 가져옴. - 벡터화 하여 chroma db에 저장.

전처리 과정

전처리 도구

```python
# JSON 데이터를 벡터 스토어에 저장하는 함수
def json_to_vector_store(json_filename, store_name):
    with open(json_filename, "r", encoding="utf-8") as f:
        total = json.load(f)

    documents = []
    for idx in range(len(total)):
        content = total[idx]
        merged_text = f"{content['title']}({content['platform']})-{content['description']}, 장르:{content['genre']}, 키워드: {content['keywords']}"
        if isinstance(content['keywords'], str):
            keywords = content['keywords']
        else:
            keywords = ", ".join(content['keywords'])
        documents.append(
            Document(
                page_content=merged_text,
                metadata={
                    "title": content["title"],
                    "platform": content["platform"],
                    "update_days": content["update_days"],
                    "genre": content["genre"],
                    "price": content["price_price"],
                    "episode": content["episode"],
                },
            )
        )
```

데이터 추출 방식

[모델]

- "text-embedding-ada-002" 모델로 벡터화 진행.
- "text-embedding-3-small","text-embedding-ada-002","Bespin SRoBERTa" 3개의 모델로 각각 진행함.
  그 중 "text-embedding-ada-002" 모델로 벡터화 했을 때 사용자 쿼리를 가장 잘 인식 해 이 모델로 진행.
  1) "text-embedding-3-small"
     query: "여주가 짝사랑하는 로맨스 웹툰 추천해줘."
     -> 주체('여주'), 장르('로맨스') 인식, 종류('웹툰') 을 잘 인식하지 못하고 짝사랑하는 내용이 담긴 일상물 혹은 웹소설을 추천함.

  2) "Bespin SRoBERTa"
     query: "여주가 짝사랑하는 로맨스 웹툰 추천해줘."
     -> 장르('로맨스')를 가장 잘 인식하게 함, 그러나 로맨스 웹툰 중 아무 웹툰을 추천함.

불필요한 데이터 제거 기준

- 크게 2가지 기준에 해당하면 metadata에 넣지 않음

[1]

- llm 모델에 사용될 정보는 아니고, MySQL에서 정보로 담고 있어야 할 키 값들
  1) id
  2) type -> (각각의 벡터스토어가 존재해 사용 안 됨)
  3) thumbnail
  4) age_rating -> (회원정보의 나이로 나이제한에 맞는 작품 추천으로 사용)
  5) url

[2]

- 사용자가 자주 질문하지 않아 넣지 않아도 된다고 판단한 부분(네이버지식인글을 토대로 검사함. 3.모델링 및 평가_수집된 데이터 및 전처리 문서_0차 6번을 참조)
    1) author
    2) illustrator
    3) status

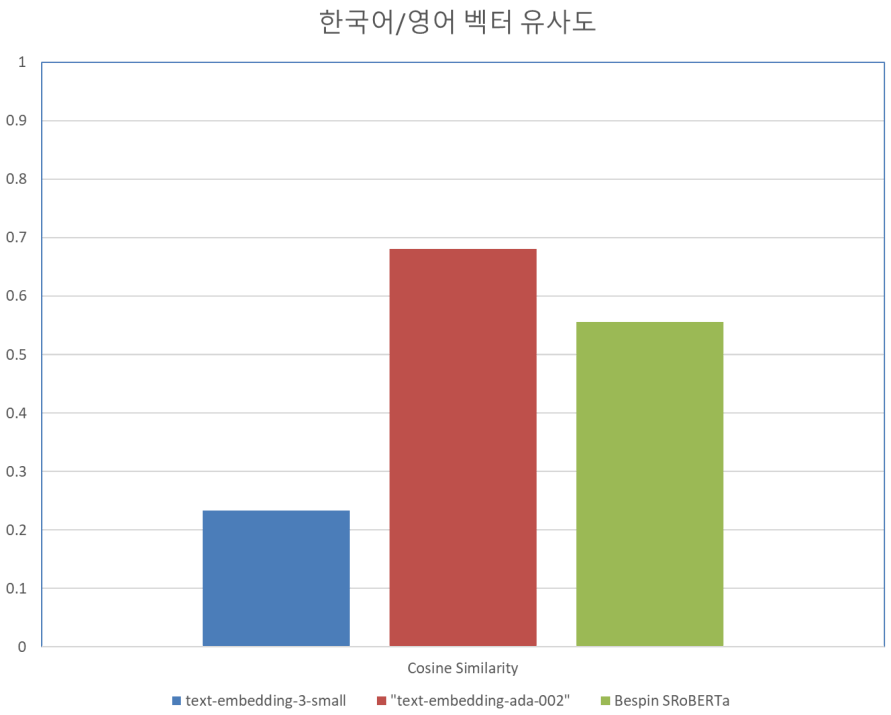
[3]

- 따로 자체 웹툰 평가 지표 score를 만들어 추출할 예정. score의 feature로 사용되는 데이터라 metadata에는 필요 없음.
    1) views
    2) rating
    3) likes
    4) first_espisode
    5) comments_count
    6) recent_comments_count

정제 방법

[한글-영어 벡터 유사도]

- 성능이 안 나오는 이유 중 하나가 사용자 쿼리에 쓰이는 단어들을 벡터 데이터베이스에서 잘 못 찾아서라고 가정함.
- 사용하고자 했던 모델의 한글-영어 벡터 유사도를 확인하고자 함. 자주 쓰이는 '장르'라는 단어와 'genre'의 벡터간 유사도(Cosine Similarity)를 살펴봄.



- 유사도가 가장 높은 모델도 0.7 미만의 값을 보임.
- page_context 에 영어로 작성했던 부분을 모두 한국어로 바꿈.

1) genre -> ‘장르’, keywords -> ‘키워드’
2) 이 정제 방법을 쓰고 훨씬 나은 답변을 내보임. 특히 장르를 잘 구분함.
3) 그러나 이렇게 바꿈으로서 이번 프로젝트에서는 한국어 사용자로 제한됨. 더 많은 시간을 투자해 다양한 언어 모델로 각각 사용이 가능하게끔 발전시키고 싶음.

[이름 변경]

- llm모델을 테스트하는 과정에서 줄거리를 더 잘 인식해야한다는 결론이 나와 기존 줄거리를 의미하던 ‘description’ -> ‘synopsis’로 변경

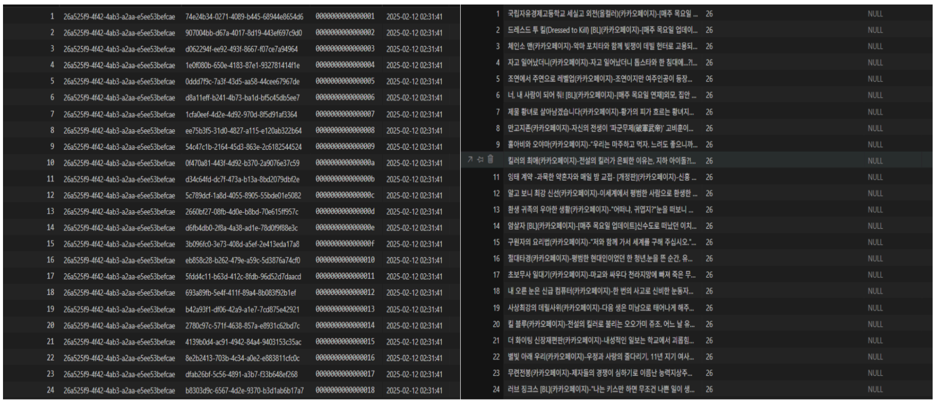
- ‘월 연재’ -> ‘월요일’
  1) query: “수요일에 볼 만한 웹소설 추천해줘.” -> 이런 경우 모든 모델에서 ‘수요일’을 전혀 인식하지 못하고 ‘월요일’, ’토요일’에 연재되는 작품을 추천함.
  2) 월 연재의 벡터값과 ‘월요일’의 벡터값 간의 유사도가 낮음이 원인으로 보고, 크롤링 데이터에서 연재 유무와 요일을 분리하고 ‘월요일’같은 형식으로 통일함.

[추후 예정]

- llm 모델 성능이 잘 나오냐? 안 나오냐? 성능에 따라 아래 전처리를 다양하게 시도 할 예정
  1) ‘?’,’!’ 등의 특수 문자 제거
  2) ‘synopsis’를 문장 단위로 / 다 합쳐진 page_context를 token수 단위로 chunking

데이터 전처리 결과

- 결과



1	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	74e24b34-0271-4089-b445-68944e8654d6	0000000000000001	2025-02-12 02:31:41	1	국립자유경제고등학교 세실고 외전(올컬러)(카카오페이지)-[매주 목요일 ...	26	NULL
2	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	907004bb-d67a-4017-8d19-443ef697c9d0	0000000000000002	2025-02-12 02:31:41	2	드레스드 투 킬(Dressed to Kill) [BL](카카오페이지)-[매주 목요일 업데이...	26	NULL
3	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	d062294f-ee92-493f-8667-f07ce7a94964	0000000000000003	2025-02-12 02:31:41	3	체인소 맨(카카오페이지)-악마 포치타와 함께 빚쟁이 데빌 헌터로 고용되...	26	NULL
4	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	1e0f080b-650e-4183-87e1-932781414f1e	0000000000000004	2025-02-12 02:31:41	4	자고 일어났더니(카카오페이지)-자고 일어났더니 톱스타와 한 침대에...?!...	26	NULL
5	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	0ddd7f9c-7a3f-43d5-aa58-44cee67967de	0000000000000005	2025-02-12 02:31:41	5	조연에서 주연으로 레벨업(카카오페이지)-조연이지만 여주인공이 등장...	26	NULL
6	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	d8a11eff-b241-4b73-ba1d-bf5c45db5ee7	0000000000000006	2025-02-12 02:31:41	6	너, 내 사람이 되어 줘! [BL](카카오페이지)-[매주 목요일 연재]외모, 집안 ...	26	NULL
7	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	1cfa0eef-4d2e-4d92-970d-8f5d91af3364	0000000000000007	2025-02-12 02:31:41	7	제물 황녀로 살아남겠습니다(카카오페이지)-황가의 피가 흐르는 황녀지...	26	NULL
8	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	ee75b3f5-31d0-4827-a115-e120ab322b64	0000000000000008	2025-02-12 02:31:41	8	만고지존(카카오페이지)-자신의 전생이 '파군무제(破軍武帝)' 고비훈이...	26	NULL
9	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	54c47c1b-2164-45d3-863e-2c6182544524	0000000000000009	2025-02-12 02:31:41	9	홀아비와 오야마(카카오페이지)-"우리는 마주하고 먹자. 느려도 좋으니까...	26	NULL
10	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	0f470a81-443f-4d92-b370-2a9076e37c59	000000000000000a	2025-02-12 02:31:41		킬러의 최애(카카오페이지)-전설의 킬러가 은퇴한 이유는, 지하 아이돌?!...	26	NULL
11	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	d34c64fd-dc7f-473a-b13a-8bd2079dbf2e	000000000000000b	2025-02-12 02:31:41	11	잉태 계약 -과묵한 약혼자와 매일 밤 교접- [개정판](카카오페이지)-신흥 ...	26	NULL
12	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	5c789dcf-1a8d-4055-8905-55bde01e5082	000000000000000c	2025-02-12 02:31:41	12	알고 보니 최강 신선(카카오페이지)-이세계에서 평범한 사람으로 환생한 ...	26	NULL
13	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	2660bf27-08fb-4d0e-b8bd-70e615ff957c	000000000000000d	2025-02-12 02:31:41	13	환생 귀족의 우아한 생활(카카오페이지)-"어떠냐, 귀엽지?"눈을 떠보니 ...	26	NULL
14	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	d6fb4db0-2f8a-4a38-ad1e-78d0f9f88e3c	000000000000000e	2025-02-12 02:31:41	14	암살자 [BL](카카오페이지)-[매주 목요일 업데이트]신수도로 떠났던 이치...	26	NULL
15	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	3b096fc0-3e73-408d-a5ef-2e413eda17a8	000000000000000f	2025-02-12 02:31:41	15	구원자의 요리법(카카오페이지)-"저와 함께 가서 세계를 구해 주십시오."...	26	NULL
16	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	eb858c28-b262-479e-a59c-5d3876a74cf0	0000000000000010	2025-02-12 02:31:41	16	절대타경(카카오페이지)-평범한 현대인이었던 한 청년.눈을 뜬 순간, 유...	26	NULL
17	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	5fdd4c11-b63d-412c-8fdb-96d52d7daacd	0000000000000011	2025-02-12 02:31:41	17	초보무사 일대기(카카오페이지)-마교와 싸우다 천라지망에 빠져 죽은 무...	26	NULL
18	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	693a89fb-5e4f-411f-89a4-8b083f92b1ef	0000000000000012	2025-02-12 02:31:41	18	내 오른 눈은 신급 컴퓨터(카카오페이지)-한 번의 사고로 신비한 눈동자...	26	NULL
19	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	b42a93f1-df06-42a9-a1e7-7cd875e42921	0000000000000013	2025-02-12 02:31:41	19	사상최강의 데릴사위(카카오페이지)-다음 생은 미남으로 태어나게 해주...	26	NULL
20	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	2780c97c-571f-4638-857a-e8931c62bd7c	0000000000000014	2025-02-12 02:31:41	20	킬 블루(카카오페이지)-전설의 킬러로 불리는 오오가미 쥬조. 어느 날 유...	26	NULL
21	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	4139b0d4-ac91-4942-84a4-9403153c35ac	0000000000000015	2025-02-12 02:31:41	21	더 화이팅 신장재편판(카카오페이지)-내성적인 일보는 학교에서 괴롭힘...	26	NULL
22	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	8e2b2413-703b-4c34-a0e2-e883811cfc0c	0000000000000016	2025-02-12 02:31:41	22	별빛 아래 우리(카카오페이지)-우정과 사랑의 줄다리기, 11년 지기 여사...	26	NULL
23	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	dfab26bf-5c56-4891-a3b7-f33b648ef268	0000000000000017	2025-02-12 02:31:41	23	무련전봉(카카오페이지)-제자들의 경쟁이 심하기로 이름난 능력지상주...	26	NULL
24	26a525f9-4f42-4ab3-a2aa-e5ee53befcae	b8303d9c-6567-4d2e-9370-b3d1ab6b17a7	0000000000000018	2025-02-12 02:31:41	24	러브 징크스 [BL](카카오페이지)-"나는 키스만 하면 무조건 나쁜 일이 생...	26	NULL

- 향후 사용계획
- 현재 위 데이터로 llm 모델 테스트 작업 진행 중.
- 앞으로도 목표 성능이 나올 때까지 데이터 전처리&프롬프트 엔지니어링을 진행 예정.
- 모델 작업이 완료되면 total_data 동일하게 벡터화하여 최종 배포에 사용.