

# SK네트웍스 Family AI과정 6기

## 모델링 및 평가 수집된 데이터 및 전처리 문서

### □ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : 수집된 데이터 및 전처리 문서
- 제출 일자 : 2025.02.28
- 깃허브 경로 : <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-2Team>
- 작성 팀원 : 이세화, 박미현, 안형진, 전하연, 정유진

### 1. 웹툰\_네이버

|                |   |
|----------------|---|
| 개요             | <p><b>데이터 설명:</b><br/>총 3,297개의 정보로 구성되어 있으며, 22개의 컬럼을 포함.</p> <p><b>컬럼들:</b><br/><b>rating:</b> 웹툰의 평균 평점 (예: 9.92)<br/><b>thumbnail:</b> 웹툰 썸네일 이미지 URL<br/><b>id:</b> 웹툰 고유 ID<br/><b>type:</b> 웹툰 유형 (웹툰)<br/><b>platform:</b> 플랫폼 (네이버 웹툰)<br/><b>title:</b> 웹툰 제목<br/><b>status:</b> 연재 상태 (완결/연재/휴재)<br/><b>update_days:</b> 연재 요일<br/><b>genre:</b> 장르 (로맨스, 스릴러, 개그 등)<br/><b>views:</b> 조회수 (현재는 '-'로 표시된 값 존재)<br/><b>synopsis:</b> 웹툰 줄거리<br/><b>author:</b> 작가명<br/><b>illustrator:</b> 일러스트레이터<br/><b>original:</b> 원작 여부<br/><b>age_rating:</b> 연령 등급 (전체 이용가, 12세 이용가 등)<br/><b>price:</b> 무료/유료/기다리면 무료 여부<br/><b>likes:</b> 좋아요 수<br/><b>url:</b> 웹툰 링크<br/><b>episode:</b> 에피소드 수<br/><b>first_episode:</b> 웹툰의 최초 공개일<br/><b>keywords:</b> 웹툰 관련 키워드(회귀물 등)-&gt;LLM이 적절한 웹툰을 찾는 데 주요한 역할<br/>키워드와 장르, 줄거리는 LLM이 사용자가 요구하는 웹툰을 추천하기 위한 척도로 사용.</p> <p><b>데이터 수집목적:</b><br/>사용자 맞춤형 웹툰 추천 서비스를 제공하기 위해,<br/>좋아요 수, 평점 등의 데이터를 활용하여 인기 웹툰을 분석하여 취향에 맞는<br/>확률이 높은 웹툰을 추천하기 위함.</p> |
| 데이터 저장<br>및 관리 | <p><b>데이터 저장 방식:</b><br/>JSON파일 형식으로 크롤링하여 MySQL에 저장.</p>   |

|                   |  |
|-------------------|--|
| <p>데이터 전처리 과정</p> | <p><b>전처리 단계 및 방법 설명:</b></p> <p><b>1. 크롤링과 동시에 저장 과정에서의 전처리</b></p> <ul style="list-style-type: none"> <li>• 데이터가 없으면 모두 “-”로 저장하여 MySQL에 저장할 땐 null값 처리</li> <li>• <b>synopsis</b>: “줄거리”의 줄바꿈 제거</li> <li>• <b>first_episode</b>: 첫 화의 날짜에서 가져오는 형식은 YY.MM.DD이지만 다른 데이터들과의 통일성을 위해 YYYY.MM.DD형식으로 변환 ex) 22.05.25 -&gt; 2022.05.25</li> <li>• <b>price</b>: “dailyPass”값이 True이면 “기다리면 무료”, selenium으로 ‘유료’ 표시를 찾아 데이터가 존재하면 유료, 아니면 무료로 기입</li> <li>• <b>keywords</b>: LLM의 원활한 키워드 인식을 위해 리스트 형식의 데이터를 “, “.join을 사용하여 리스트의 요소들을 문자열로 결합하였다.<br/>ex) [“사이다”, “느와르”, “먼치킨”] -&gt; “사이다, 느와르, 먼치킨”</li> <li>• <b>update_days</b>: API엔 keywords와 마찬가지로 리스트 형식으로 요일이 표시되어 있었으나 “, “.join으로 하나의 문자열로 결합함과 동시에 모든 문자열 뒤에 “요일”글자가 붙도록 하였다 - 역시 LLM의 인식을 돕기 위해<br/>ex) [“수”, “금”] -&gt; “수요일, 금요일”<br/>매일 연재되는 작품은 “월요일, 화요일, …, 일요일”의 형식을 따랐다.</li> </ul> <p><b>2. 데이터를 분류하고 수정하는 단계에서의 전처리</b></p> <ul style="list-style-type: none"> <li>• 모델별로 수집한 데이터 나누기 <ul style="list-style-type: none"> <li>◦ 기본 모델: 모든 장르 데이터</li> <li>◦ 로맨스 모델(남, 여): “로맨스”</li> <li>◦ 액션 모델: “액션”, “무협”, “액션/무협”, “판타지”, “현대판타지”, “먼치킨”</li> <li>◦ 로판 모델: “로맨스 판타지”</li> </ul> </li> </ul> <p><b>3. 데이터를 저장하고 난 후의 추가 전처리</b></p> <ul style="list-style-type: none"> <li>• likes(즐거찾기 수), rating(평점), episode(회차수), days_since_release(첫 웹툰 공개일로부터의 경과일수) 를 정규화하여 가중치를 장르별로 다르게 뒤 웹툰의 최종 score를 계산. LLM은 이 score를 참고하여 추천.</li> </ul> |
| <p>데이터 전처리 결과</p> | <p><b>결과</b><br/>2025.2.25일 기준 3,297개 웹툰의 데이터 크롤링 완료<br/>(추후 웹툰이 늘어나면 데이터 수는 변동가능)</p> <p><b>향후 데이터 사용계획</b><br/>데이터는 벡터 임베딩하여 LLM이 유사한 웹툰 콘텐츠를 효율적으로 검색할 수 있게 할 예정입니다. 또한, 챗봇의 성능을 지속적으로 모니터링하면서 추가적인 데이터를 수집하고, 벡터 데이터베이스에서 불필요한 키 값을 정리하여 최적화할 계획입니다.</p>   |

## 2. 웹툰\_카카오

|              |  |
|--------------|--|
| 개요           | <p><b>데이터 설명</b></p> <p>총 2,672개의 웹툰 정보로 구성</p> <ul style="list-style-type: none"> <li>- <b>thumbnail</b>: 웹툰 썸네일 이미지 URL</li> <li>- <b>id</b>: 웹툰 고유 ID</li> <li>- <b>type</b>: 웹툰 유형 (웹툰)</li> <li>- <b>platform</b>: 플랫폼 (네이버 웹툰)</li> <li>- <b>title</b>: 웹툰 제목</li> <li>- <b>status</b>: 연재 상태 (완결/연재 중/휴재)</li> <li>- <b>update_days</b>: 연재 요일</li> <li>- <b>genre</b>: 장르 (로맨스, 스릴러, 개그 등)</li> <li>- <b>views</b>: 조회수</li> <li>- <b>synopsis</b>: 웹툰 줄거리</li> <li>- <b>author</b>: 작가명</li> <li>- <b>illustrator</b>: 일러스트레이터</li> <li>- <b>original</b>: 원작 여부</li> <li>- <b>age_rating</b>: 연령 등급 (전체 이용가, 15세 이용가 등)</li> <li>- <b>price</b>: 무료/유료/기다리면 무료 여부</li> <li>- <b>likes</b>: 좋아요 수</li> <li>- <b>url</b>: 웹툰 링크</li> <li>- <b>episode</b>: 에피소드 수</li> <li>- <b>first_episode</b>: 웹툰의 최초 공개일</li> <li>- <b>keywords</b>: 웹툰 관련 키워드(회귀물 등)-&gt;LLM이 적절한 웹툰을 찾는 데 주요한 역할</li> <li>- 키워드와 장르, 줄거리는 LLM이 사용자가 요구하는 웹툰을 추천하기 위한 척도로 사용됨.</li> </ul> <p><b>데이터 수집목적</b></p> <ul style="list-style-type: none"> <li>- LLM + RAG 기반 추천 시스템에서 작품 정보를 메타데이터로 활용하여 사용자의 선호 작품과 유사한 작품 추천을 위함.</li> </ul> |
| 데이터 자동화 및 검증 | <p><b>데이터 수집 자동화 프로세스</b></p> <ul style="list-style-type: none"> <li>- Selenium + Requests 기반 크롤링을 통해 자동 수집             <ol style="list-style-type: none"> <li>1. <b>로그인 처리</b>: 카카오웹툰 플랫폼은 로그인 후 일부 데이터 접근 가능 → .env 파일을 활용한 로그인 자동화</li> <li>2. <b>작품 목록 URL 크롤링</b>: 전체 작품 URL 수집 후 pickle로 저장</li> <li>3. <b>상세 정보 크롤링</b>: 작품 ID를 활용하여 API 요청 및 HTML 크롤링을 병행, Selenium을 통해 추가 정보(장르, 조회수, 좋아요 수) 확보</li> <li>4. <b>신작웹툰 추가</b>: 신작 탭의 API 요청하여 신작만 따로 크롤링 후 작품 ID를 비교하여 중복된 데이터는 제거</li> <li>5. <b>데이터 저장 및 검증</b>: 기존 데이터와 비교하여 새로운 작품 추가 or 기존 데이터 업데이트</li> </ol> </li> </ul>  |
| 데이터 저장 및 관리  | <p><b>데이터 저장 방식</b></p> <ul style="list-style-type: none"> <li>- 로컬 JSON 저장 (.json)</li> <li>- MYSQL 저장 후 관리 예정</li> </ul>   |

|                       |  |
|-----------------------|--|
| <p>데이터 전처리<br/>과정</p> | <p>전처리 단계 및 방법 설명</p> <ul style="list-style-type: none"> <li>• <b>id</b><br/>URL에서 정규 표현식(<code>re.search(r'/content/.+/(Wd+)', url)</code>)을 사용하여 ID 값만 추출하고 int로 변환하여 저장</li> <li>• <b>first_episode</b><br/>API에서 2020-09-28T00:00:00Z 같은 ISO 형식으로 제공 날짜 문자열에서 YYYY-MM-DD 부분만 추출 2020.09.28 형식으로 변환하여 저장</li> <li>• <b>status</b><br/>연재/완결/휴재 중 하나로 반환<br/>COMPLETED, END_OF_SEASON, SEASON_COMPLETED, EPISODES_PUBLISHING 등 여러 상태 값이 존재 <ul style="list-style-type: none"> <li>- COMPLETED → 완결</li> <li>- END_OF_SEASON, SEASON_COMPLETED, EPISODES_PUBLISHING → 연재,</li> <li>- EPISODES_NOT_PUBLISHING → 휴재 등으로 매핑</li> </ul> </li> <li>• <b>update_days</b><br/>MON, TUE와 같은 영어 약자로 제공<br/>MON → 월요일, TUE → 화요일 등으로 변환<br/>여러 요일이 있을 경우 월요일, 목요일처럼 쉼표로 구분</li> <li>• <b>price</b><br/>무료/유료/기다리면 무료 여부를 보다 명확하게 정리<br/>"waitForFree" 값이 존재하면 "기다리면 무료"<br/>"freePublishing"이 False이면서 무료 제공 에피소드 개수(freeEpisodeCount)가 전체 에피소드와 같다면 "무료"<br/>"freePublishing"이 False이면 "유료"</li> <li>• <b>views&amp;likes</b> <ul style="list-style-type: none"> <li>- 정수로 반환</li> <li>- 천 단위의 ',' 가 있다면 제거</li> <li>- '억' -&gt; 100,000,000 , '만' -&gt; 10,000 을 곱한 것으로 반환</li> </ul> </li> </ul> |
| <p>데이터 전처리<br/>결과</p> | <ul style="list-style-type: none"> <li>• <b>결과</b><br/>크롤링을 통해 2025.02.24 기준 카카오 웹툰 총 2,672개의 웹툰 데이터를 확보</li> <li>• <b>향후 데이터 사용계획</b><br/>MySQL과 ChromaDB를 활용한 벡터 기반 검색 기능 추가<br/>유사한 키워드, 장르를 기반으로 유저 맞춤형 웹툰 추천</li> </ul>   |

### 3. 웹툰\_카카오페이지

|                     |   |
|---------------------|---|
| <p>개요</p>           | <p><b>데이터 설명</b></p> <ul style="list-style-type: none"> <li>- 카카오페이지에서 웹툰에 대한 정보를 크롤링 함.</li> <li>- 각 장르별로 크롤링 함.(판타지, 드라마, 로맨스, 로판, 무협, 액션, BL)</li> </ul> <p>총 데이터 : 17,616개<br/>         [판타지]: 1,790개<br/>         [드라마]: 3,194개<br/>         [로맨스]: 5,019개<br/>         [로판]: 2,298개<br/>         [무협]: 1,678개<br/>         [액션]: 1,410개<br/>         [BL]: 2,227개</p> <p><b>데이터 수집목적</b></p> <ul style="list-style-type: none"> <li>- RAG기반 챗봇에서 웹툰 추천 서비스를 제공하기 위한 핵심 자료로 활용됨. 챗봇은 이 데이터들을 데이터베이스에서 검색하여 사용자의 질의에 대해 적절한 추천을 수행함.</li> <li>- 회원의 데이터 베이스에 작품의 고유 아이디를 mapping해서 회원의 취향을 분석하고 다양한 추천시스템에 적용해볼 수도 있음.</li> </ul>   |
| <p>데이터 자동화 및 검증</p> | <p><b>데이터 수집 자동화 프로세스</b></p> <pre># 현재 날짜 확인 today = datetime.now()  # 다음 실행 날짜 설정 (매월 26일 오후 2:33) next_run = datetime(today.year, today.month, 26, 14, 33)  # 만약 오늘이 26일 이후라면 다음 달로 설정 if today.day &gt; 26 or (today.day == 26 and today.hour &gt;= 14 and today.minute &gt;= 33):     if today.month == 12: # 12월이면 다음 실행 날짜를 다음 해 1월로 변경         next_run = datetime(today.year + 1, 1, 26, 14, 33)     else:         next_run = datetime(today.year, today.month + 1, 26, 14, 33)  # 실행 대기 시간 계산 wait_time = (next_run - today).total_seconds()</pre> <ul style="list-style-type: none"> <li>- 1달에 1번 매월 정해진 시각에 자동적으로 크롤링 되는 코드로 구현함.</li> <li>- 다만 19세 이용가 작품의 경우 로그인을 해야 크롤링 할 수 있는데, 핸드폰으로 2차 인증을 거쳐야 해서 이 부분은 다시 크롤링 할 때마다 수동적으로 인증해줘야 함.</li> <li>- 정보 업데이트가 너무 늦다는 한계가 있음. 현재는 한 번 크롤링을 하는데 약 50시간이 소요되어 이와 같은 방식으로 진행 하였으나 실제 상업적으로 이용하고자 한다면 API를 쓰는 방식으로 진행하면 됨.</li> </ul> <p><b>검증</b></p> <pre>2025-02-26 14:32:25,800 - INFO - 다음 크롤링 실행 예정 시간: 2025-02-26 14:33:00 2025-02-26 14:33:00,003 - INFO - 웹툰 크롤링을 시작합니다... 2025-02-26 14:33:00,007 - INFO - ===== WebDriver manager ===== 2025-02-26 14:33:01,091 - INFO - Get LATEST chromedriver version for google-chrome 2025-02-26 14:33:01,134 - INFO - Get LATEST chromedriver version for google-chrome 2025-02-26 14:33:01,195 - INFO - Driver [C:\Users\Playdata\.wdm\drivers\chromedriver\win64\133.0.6943.14 ache DevTools listening on ws://127.0.0.1:59251/devtools/browser/8edb3f98-f53e-431b-a60e-ad6e8fa12c39 2025-02-26 14:33:07,688 - INFO - [1/3] 크롤링 완료: 아웃 2025-02-26 14:33:11,438 - INFO - [2/3] 크롤링 완료: 시그리드 Created TensorFlow Lite XNNPACK delegate for CPU. 2025-02-26 14:33:15,162 - INFO - [3/3] 크롤링 완료: 열 명의 신재들이 모두 내 제자 2025-02-26 14:33:15,164 - INFO - 총 3개의 웹툰 정보를 저장하였습니다: webtoon_crawling_results.json</pre> <ul style="list-style-type: none"> <li>- 매월 26일 14:33에 크롤링 되는 것으로 코드가 작동하는 지에 대한 검증을 완료함.(데이터 양이 방대하여 3개의 작품으로만 검증)</li> </ul> |

|                        |   |
|------------------------|---|
| <b>데이터 저장<br/>및 관리</b> | <b>데이터 저장 방식</b> <ul style="list-style-type: none"> <li>- 현재는 local에 json형식으로 저장함.</li> <li>- MySQL</li> </ul>  |
| <b>데이터<br/>전처리 과정</b>  | <b>전처리 단계 및 방법 설명</b> <p><b>[형식 변환]</b></p> <ul style="list-style-type: none"> <li>- <b>조회수</b> : 정수로 반환 <ul style="list-style-type: none"> <li>1) 천 단위의 ‘,’ 가 있다면 제거</li> <li>2) ‘억’ -&gt; 100,000,000 , ‘만’ -&gt; 10,000 을 곱한 것으로 반환</li> <li>ex) 2,280만 -&gt; 22800000</li> </ul> </li> <li>- <b>연재 상태</b> : 연재/완결/휴재 중 하나로 반환 <ul style="list-style-type: none"> <li>1) ‘완결’ 혹은 ‘휴재’ 가 아니면 모두 연재로 반환</li> <li>ex) ‘완결’ -&gt; 완결, ‘월,수 연재’ -&gt; 연재</li> </ul> </li> <li>- <b>연재 요일</b> : 연재 중인 작품은 연재 요일을 형식에 맞게 반환 <ul style="list-style-type: none"> <li>1) 연재 상태가 ‘연재’ 인 경우 ‘월 연재’ 중 앞에 요일만 가져옴</li> <li>2) ‘,’ 가 있다면 그 단위로 끊음</li> <li>3) 뒤에 요일을 붙임. 여러 요일인 경우 , 로 연결해 하나의 문자열로 반환</li> <li>ex) ‘월 연재’ -&gt; ‘월요일’, ‘화,목 연재’ -&gt; “화요일,목요일”</li> </ul> </li> <li>- <b>이용가</b> : 띄어쓰기 <ul style="list-style-type: none"> <li>1) 나이 뒤에 띄어쓰기</li> <li>ex) ‘19세이용가’ -&gt; ‘19세 이용가’</li> </ul> </li> <li>- ‘조회수’는 숫자로 변환해야 점수 책정에 용이하여 전처리를 진행하였고, ‘연재 상태’, ‘연재 요일’, ‘이용가’는 작품들이 서로 다른 플랫폼에서 온다는 점 때문에 같은 의미인데도 통일되지 않아 팀원들과 상의해 위와 같은 규칙으로 정함.</li> </ul> <p><b>[중복 데이터 제거]</b></p> <ul style="list-style-type: none"> <li>- 웹툰끼리 ‘title’과 ‘author’이 모두 같은 경우 같은 작품으로 인식하고 고유 아이디를 동일하게 변환.</li> <li>- ‘title’이 같은 것으로 진행했었으나 제목이 같고 다른 작품인 경우가 있어서 위와 같이 다시 진행.</li> </ul> |

## 데이터 전처리 결과

### 결과

```
{
  "comments_count": 0,
  "recent_comments_count": 0
},
{
  "id": "61472803",
  "type": "webtoon",
  "platform": "카카오페이지",
  "title": "미행",
  "status": "완결",
  "update_days": "-",
  "thumbnail": "https://page-images.kakaoentcdn.com/download/resource?kid=3uuob/hzVqM4TaA1/VYQoIeiZpSes5Nr2kQtry0&filename=61472803_thumbnail.jpg",
  "genre": "드라마",
  "views": 26000,
  "rating": 8.3,
  "synopsis": "비가 내리는 날이면\n잡은 여자들이 사라진다...\n\n인터넷에 떠도는 괴담으로 불안에 떠는 도시.\n하지만 언론에서는 실종 사건을\n\nkeywords": "스릴러, 호러, 연쇄살인",
  "author": "복화",
  "illustrator": "그린비",
  "original": "-",
  "age_rating": "15세 이용가",
  "price": "기다리면 무료",
  "url": "https://page.kakao.com/content/61472803",
  "episode": 40,
  "comments_count": 35,
  "recent_comments_count": 0
},
{
  "id": "65517649",
  "type": "webtoon",
  "platform": "카카오페이지",
  "title": "국립자유경제고등학교 세실고 외전(울컬러)",
  "status": "연재",
  "update_days": "목요일",
  "thumbnail": "https://page-images.kakaoentcdn.com/download/resource?kid=H8AM5/hAFPKadJdK/F1Kx1aV1UHKMSF81XubWS0&filename=65517649_thumbnail.jpg",
  "genre": "드라마",
  "views": 1668,
  "rating": 10.0,
  "synopsis": "[매주 목요일 업데이트]\n\n세실고 학생회의 사무실 이전 작업 중\n\n학년이자 학생회 부회장이 된 송예랑이\n\nkeywords": "학원물",
  "author": "양혜석",
  "illustrator": "이현지",
  "original": "-",
  "age_rating": "전체 이용가",
  "price": "유료",
  "url": "https://page.kakao.com/content/65517649",
  "episode": 15,
  "comments_count": 16,
  "recent_comments_count": 9
},
{
  "id": "62169485",
  "type": "webtoon",
  "platform": "카카오페이지",
  "title": "미소짓는 남자",
  "status": "완결",
  "update_days": "-",
  "thumbnail": "https://page-images.kakaoentcdn.com/download/resource?kid=H8AM5/hAFPKadJdK/F1Kx1aV1UHKMSF81XubWS0&filename=62169485_thumbnail.jpg",
  "genre": "드라마",
  "views": 1668,
  "rating": 10.0,
  "synopsis": "[매주 목요일 업데이트]\n\n세실고 학생회의 사무실 이전 작업 중\n\n학년이자 학생회 부회장이 된 송예랑이\n\nkeywords": "학원물",
  "author": "양혜석",
  "illustrator": "이현지",
  "original": "-",
  "age_rating": "전체 이용가",
  "price": "유료",
  "url": "https://page.kakao.com/content/62169485",
  "episode": 15,
  "comments_count": 16,
  "recent_comments_count": 9
}
```

### 향후 데이터 사용계획

- 벡터 임베딩 적용하여 사용자의 쿼리와 유사한 콘텐츠 검색으로 사용될 예정.
- 챗봇의 성능을 확인해 가며 추가로 더 많은 정보를 가져오거나, 벡터 데이터베이스에는 필요없는 키 값을 삭제하여 사용할 예정

#### 4. 웹소설\_네이버시리즈

### 개요

#### 1. 데이터 설명

- 1) 데이터 출처 : 네이버 시리즈 (NAVER SERIES) 웹소설  
<https://series.naver.com/novel/categoryProductList.series?categoryTypeCode=all>
- 2) 데이터 유형: 웹소설 정보(JSON 형식)
- 3) 수집일자 : 2025.02.27
- 4) 데이터 양 ( 101,648 )
  - 장르별 데이터
    - 로맨스 - 46,147
    - 로판 - 15,683
    - 판타지 - 12,415
    - 현판 - 10,123
    - 무협 - 5,461
    - 미스터리 - 690
    - 라이트노벨 - 967
    - BL - 10,162
  - 필수 수집 항목
    - Type (웹소설)
    - Platform (네이버 시리즈)
    - ID (작품 고유 번호)
    - Title (작품의 제목)
    - Status (연재 상태) : 완결 또는 연재중
    - Thumbnail (작품 대표 이미지)
    - Genre (작품의 장르)
    - Views (작품 열람 수)
    - Rating (별점)
    - Like (좋아요 수)
    - Synopsis (줄거리)
    - Keywords (해시태그, 키워드)
    - Author (작가명)
    - illustrator (작화, 삽화가)
    - Age rating (이용 연령)
    - First episode (첫 에피소드 공개일)
    - Comments (댓글 수)
    - Price (가격 상세)
    - Episode (총 작품 수)
    - URL (작품 링크)

#### 2. 데이터 수집목적

- 1) 웹소설 추천 시스템 구축을 위한 기초 데이터 확보
- 2) 사용자의 관심, 이용 패턴을 기반한 콘텐츠 추천의 최적화
- 3) 장르 별 인기 작품 및 성향 분석
- 4) 구독자 성향과 패턴 연구를 통한 개인 맞춤 추천



### 3. 데이터 수집 자동화 프로세스

1) 수집 방법: 네이버 시리즈 크롤링 및 BeautifulSoup, Selenium 활용

○ 주요 절차 :

- 네이버 시리즈 내 웹소설 목록 접근 및 카테고리 내 장르별 작품 상세 페이지 크롤링
- JSON 형식으로 데이터 저장
- 10만건이 넘는 거대 데이터로 짧은 기간의 정기적인 데이터 업데이트 어려움.
- 신작 웹소설 업데이트 주기 (검색일로부터 2주전까지)
- 특별관 '신작 웹소설' 에서 제공하는 데이터, 최신순 'NEW' 배너 작품
- 해당 서비스에서 신작 업데이트 주기가 2주 인 것으로 확인되어, 2주마다 월요일 오전 10시 작품 상세 페이지 크롤링한다.

**NAVER 시리즈**

로그인

웹소설 만화 e북 MY | 장바구니 0 작품명 · 작가명 · 출판사명 검색

웹소설 홈 카테고리 TOP 100 프리패스 매일10시무료 타임딜 추천장르 로먼스 로맨스 판타지

**신작 웹소설**  
총 335개 작품

완결작 보기 최신순

| 19세이상 이용가능  | 19세이상 이용가능  | 19세이상 이용가능   | 19세이상 이용가능   | 19세이상 이용가능   |
|---|---|--|--|--|
| <b>NEW</b> 19세이상 이용가능<br><b>연재 웹소설</b><br>백한송이<br>★★★★★ 0.0<br>3화 무료          | <b>NEW</b> 19세이상 이용가능<br><b>죽느니 뭐사</b><br>아겠습니다, 이...<br>보리수염<br>★★★★★ 10.0<br>3화 무료    | <b>NEW</b> 19세이상 이용가능<br><b>해어질 때</b><br>알아야 할 것들...<br>정은현<br>★★★★★ 0.0                       | <b>NEW</b> 19세이상 이용가능<br><b>서로를 부</b><br>를 수 있었던 ...<br>지연<br>★★★★★ 0.0                  | <b>NEW</b> 19세이상 이용가능<br><b>아한 겨울</b><br>[BL][단행본]<br>아라미스, 하오츠...<br>★★★★★ 0.0    |
| <b>NEW</b> 19세이상 이용가능<br><b>검을 찢은</b><br>꽃 [BL]<br>달영재<br>★★★★★ 10.0<br>3화 무료 | <b>NEW</b> 19세이상 이용가능<br><b>새별집 사생</b><br>마도 사는 법<br>옥뚜무<br>★★★★★ 9.1<br>50화 무료 · 2일 남음 | <b>NEW</b> 19세이상 이용가능<br><b>피해 소설 속</b><br>박막으로부터 ...<br>Hideka<br>★★★★★ 9.0<br>30화 무료 · 2일 남음 | <b>NEW</b> 19세이상 이용가능<br><b>결대적 이글</b><br>남 [극점]<br>조연진<br>★★★★★ 10.0<br>5화 무료 · 139일 남음 | <b>NEW</b> 19세이상 이용가능<br><b>소설이 끝</b><br>났는데 접착제 ...<br>이조원<br>★★★★★ 10.0<br>5화 무료 |

## 4. 데이터 검증

### 1) 검사

- 제목, 장르 등의 필수 수집 항목 누락 여부 확인
- URL 크롤링 이후, 판매 중지된 작품의 경우 다음과 같이 누락된 정보를 가진채로 수집이 되어 결측 데이터로 삭제가 필요함.
- 전처리 과정에서 삭제

### 판매 중지된 상품 페이지입니다.

해당 상품은 판매중지 되었습니다.  
다른 상품을 이용해주세요.

이전 페이지

시리즈 홈

```
{
  "type": "웹소설",
  "platform": "네이버 시리즈",
  "id": 2760846,
  "title": "-",
  "status": "-",
  "thumbnail": "-",
  "genre": "-",
  "views": "-",
  "rating": "-",
  "like": "-",
  "synopsis": "-",
  "keywords": "-",
  "author": "-",
  "illustrator": "-",
  "age_rating": "-",
  "first_episode": "-",
  "comments": "-",
  "price": "-",
  "episode": "-",
  "url": "https://series.naver.com/novel/detail.series?productNo=2760846"
},
```

### 2) 데이터 중복 처리

- 동일 작품이 반복 크롤링되지 않도록 필터링
- 동일 작품이어도 구매단위가 다른 [단행본] 삭제하지 않기로 협의

### 3) 데이터 정제

- 타입 확인
  - rating – float
  - id, Like, age\_rating, comments, episode – int
- 형식
  - "first\_episode": "(2025.02.19.)" 날짜 "YYYY.mm.dd." 형식으로 맞추고, "(" 괄호 제거
  - like, comments : ", " 심표 제거

### 4) 유효성 검사

- URL 응답 코드 확인, 이미지 링크 정상 불러오기 여부 체크

```
import json
import requests
from urllib.parse import urlparse

def load_data():
    """데이터 로드 및 기본 검증"""
    try:
        with open("MUNES_ROMANCE_1-46147.json", "r", encoding="utf-8") as f:
            main_data = json.load(f)
        with open("url_ROMANCE_01.json", "r", encoding="utf-8") as f:
            url_list = json.load(f)
        return main_data, url_list
    except FileNotFoundError as e:
        print(f"파일을 찾을 수 없습니다: {e.filename}")
        exit(1)

def extract_product_id(url):
    """URL에서 상품 ID 추출"""
    parsed = urlparse(url)
    return parsed.path.split('/')[1].split('?')[0]

def check_url_status(url):
    """URL 응답 상태 확인"""
    try:
        response = requests.head(url, timeout=10, allow_redirects=True)
        return response.status_code
    except requests.RequestException as e:
        return f"Error: {str(e)}"

def check_image(url):
    """이미지 링크 유효성 검증"""
    try:
        response = requests.get(url, stream=True, timeout=10)
        if 'image/' in response.headers['Content-Type']:
            return "Valid Image"
        return "Invalid Content"
    except Exception as e:
        return f"Image Error: {str(e)}"

def main():
    main_data, url_list = load_data()
```

```
# ID 불일치 검증
mismatch_log = []
for idx, (item, url) in enumerate(zip(main_data, url_list)):
    id_item = extract_product_id(item['url'])
    id_url = extract_product_id(url)
    if id_item != id_url:
        mismatch_log.append((idx, item['url'], url))

# URL 상태 & 이미지 검사
validation_report = []
for idx, item in enumerate(main_data):
    webtoon_url_status = check_url_status(item['url'])
    image_status = check_image(item['thumbnail']) if 'thumbnail' in item else None

    validation_report.append({
        "index": idx,
        "url_status": webtoon_url_status,
        "image_status": image_status
    })

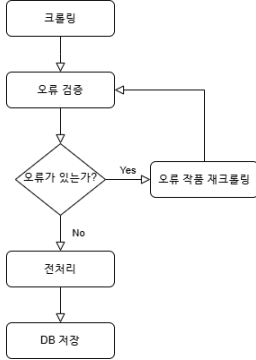
# 결과 출력
print("\n=== ID 불일치 항목 ===")
for log in mismatch_log:
    print(f"[{log[0]}] 메인URL: {log[1]}\n      비교URL: {log[2]}\n")

print("\n=== 유효성 검사 결과 ===")
for report in validation_report[:5]: # 샘플 5개 출력
    status_icon = "✅" if report['url_status'] == 200 else "❌"
    print(f"[{report['index']}] {status_icon} URL: {report['url_status']} |
```

```
if __name__ == "__main__":
    main()
```

|  |   |
|--|---|
|  | <div data-bbox="370 224 651 264"><b>5. 데이터 저장 방식</b></div> <div data-bbox="427 300 1088 371"><ul style="list-style-type: none"><li>• 형식: JSON 파일</li><li>• 저장소: 로컬 MySQL DB 및 ChromaDB 벡터 저장소</li></ul></div> <div data-bbox="370 407 691 448"><b>6. 데이터 백업 및 관리</b></div> <div data-bbox="427 483 1096 555"><ul style="list-style-type: none"><li>• 주기적인 백업을 통해 데이터 손실 방지</li><li>• 데이터 변경 감지를 위한 버전 관리 시스템 적용 협의</li></ul></div>  |
| <div data-bbox="178 1146 343 1218"><b>데이터<br/>전처리 결과</b></div> | <div data-bbox="370 721 600 761"><b>7. 데이터 전처리</b></div> <div data-bbox="427 801 1394 1142"><div data-bbox="427 801 786 842">1) 과정(전처리 단계 및 방법)</div><div data-bbox="427 842 1394 1142"><ul style="list-style-type: none"><li>• 결측값 처리: 조회수, 좋아요 수 등 "-"로 표기된 데이터 제거 또는 대체<ul style="list-style-type: none"><li>◦ Views (작품 열람 수) - 제공되지 않는 데이터 → 결측치 삭제</li><li>◦ Keywords (해시태그, 키워드) - 제공되지 않는 데이터로 전체DB에서 네이버 시리즈만 검색되는 정보 → 결측치 → 작품의 장르와 같게 작성</li><li>◦ illustrator (작화, 삽화가) - 제공되지 않는 데이터 → 결측치 삭제</li></ul></li><li>• 텍스트 정제: 특수 문자, HTML 태그 제거 후 자연어 처리(NLP) 적용</li><li>• 중복 제거: 동일한 작품이 중복 저장되지 않도록 ID 기준 중복 필터링</li><li>• 토큰화 및 임베딩: 줄거리 및 키워드를 자연어 처리하여 벡터화</li></ul></div></div> <div data-bbox="370 1182 681 1223"><b>8. 데이터 전처리 결과</b></div> <div data-bbox="416 1263 1414 1644"><div data-bbox="416 1263 507 1303">1) 결과</div><div data-bbox="427 1303 1414 1415"><ul style="list-style-type: none"><li>• 데이터 품질 개선을 위해 결측 DB와 불필요한 정보 제거 및 중복 데이터 최소화</li><li>• 검색 및 추천 효율성을 위해 작품 정보를 구조화하여 효율적인 검색 가능</li><li>• 장르별 벡터 임베딩 데이터 생성하여 유사도 검색을 위한 벡터 데이터 활용 가능</li></ul></div><div data-bbox="416 1415 724 1456">2) 향후 데이터 사용 계획</div><div data-bbox="427 1456 1370 1644"><ul style="list-style-type: none"><li>• LLM 기반 웹소설 추천 시스템에 활용</li><li>• RAG(Retrieval-Augmented Generation) 기술과 결합하여 AI 챗봇 추천 기능 개선</li><li>• 사용자 선호도를 분석하여 개별 맞춤형 콘텐츠 추천</li><li>• 플랫폼별 인기 작품 비교 분석을 통해 추천 서비스 최적화</li></ul></div></div> |

## 5. 웹소설\_카카오페이지

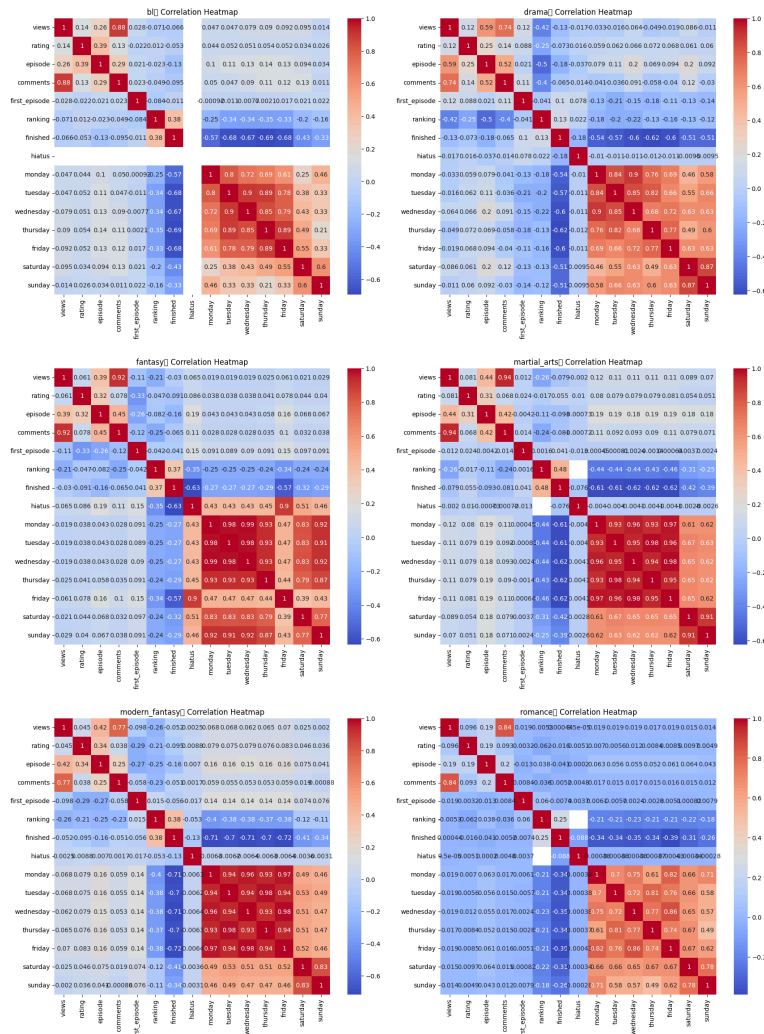
|                     |   |
|---------------------|---|
| <p>개요</p>           | <p><b>데이터 설명</b></p> <ul style="list-style-type: none"> <li>• 플랫폼: 카카오페이지</li> <li>• 작품 타입: 웹소설</li> <li>• 2025년 2월 24일 기준 서비스</li> <li>• 각 장르별 데이터 양 <ul style="list-style-type: none"> <li>◦ 판타지: 10,553</li> <li>◦ 현판: 7,692</li> <li>◦ 로맨스: 20,537</li> <li>◦ 로판: 8,989</li> <li>◦ 무협: 4,013</li> <li>◦ BL: 3,239</li> <li>◦ 드라마: 302</li> </ul> </li> </ul> <p><b>데이터 수집목적</b></p> <ul style="list-style-type: none"> <li>• 카카오페이지의 웹소설 정보를 저장</li> <li>• 챗봇 화면에서 썸네일 / 기본 정보 제공</li> <li>• 벡터스토어 생성용</li> </ul>   |
| <p>데이터 자동화 및 검증</p> | <p><b>1. 데이터 자동화</b></p> <ul style="list-style-type: none"> <li>◦ 크롤링 및 오류 검증, 오류 처리 후 전처리하는 mySQL 저장 코드 생성 <ol style="list-style-type: none"> <li>1. 크롤링</li> <li>2. 오류 검증 <ul style="list-style-type: none"> <li>■ 오류가 없을 시<br/>→ 바로 전처리로 이동</li> <li>■ 오류가 있을 시<br/>→ 오류가 있는 작품 재크롤링 후<br/>→ 2번 다시 시도</li> </ul> </li> <li>3. 전처리</li> <li>4. mySQL 저장</li> </ol> </li> </ul>  <pre> graph TD     A[크롤링] --&gt; B[오류 검증]     B --&gt; C{오류가 있는가?}     C -- Yes --&gt; D[오류 작품 재크롤링]     D --&gt; B     C -- No --&gt; E[전처리]     E --&gt; F[DB 저장]     </pre> <ul style="list-style-type: none"> <li>◦ 신작 데이터 업데이트 주기 <ol style="list-style-type: none"> <li>1. 매주 월요일 오전 10시 신작 작품 전체 크롤링 <ul style="list-style-type: none"> <li>■ 표지에 있는 <b>신작</b> 배너로 식별</li> <li>■ 신작의 기준: 일주일로 예상 (25.02.26 오후 5시 24분 기준 25.02.20에 등록된 작품 전체, 25.02.19일 일부 두 작품이 신작)</li> <li>■ 중복을 방지하기 위해 중복 제거 처리</li> </ul> </li> </ol> </li> <li>◦ 전체 데이터 업데이트 주기 <ol style="list-style-type: none"> <li>1. 매달 1일 <ul style="list-style-type: none"> <li>■ 판매 중지 작품</li> <li>■ 랭킹 데이터: 월간 TOP 300 데이터</li> </ul> </li> <li>2. 에피소드, 연재 상태, 별점, 조회수, 댓글 수</li> </ol> </li> <li>◦ 신작 / 데이터 업데이트 하는 파이썬 파일 생성 후 매 주기마다 스케줄러로 구동(예정)</li> </ul> |

|                 |   |
|-----------------|---|
| 데이터 자동화<br>및 검증 | <h2>2. 데이터 검증</h2> <ul style="list-style-type: none"> <li>○ 개별 데이터 검증             <ol style="list-style-type: none"> <li>1. 요소 누락</li> <li>2. 요소 데이터 타입 확인                     <ul style="list-style-type: none"> <li>■ id, views, comments, episode: int</li> <li>■ rating: float</li> <li>■ 나머지: str</li> </ul> </li> </ol> </li> <li>○ 전체 데이터 검증             <ol style="list-style-type: none"> <li>1. 중복작 제거                     <ul style="list-style-type: none"> <li>■ 중복작의 정의                             <ol style="list-style-type: none"> <li>1. 같은 플랫폼, 같은 내용, 다른 제목                                     <ul style="list-style-type: none"> <li>■ (이용권)이 붙은 작품은 같은 작품으로 간주 → (이용권)이 붙은 작품을 삭제</li> <li>■ 단행본과 아닌 작품 → 가격, 작품 ID 등이 다르기 때문에 다른 작품으로 간주</li> </ul> </li> <li>2. 다른 플랫폼, 같은 작품</li> </ol> </li> </ul> </li> <li>2. 판매 중지 작품 제거                     <ul style="list-style-type: none"> <li>■ url 크롤링 기준 날짜와 상세 크롤링 날짜 사이에 판매 중지 작품 발생</li> <li>■ 해당 url에 들어가서 표지 배너에 <b>판매 중지</b> 가 있다면 해당 작품 삭제</li> </ul> </li> </ol> </li> </ul> |
| 데이터 저장<br>및 관리  | <ul style="list-style-type: none"> <li>● json 형태로 장르별 저장</li> <li>● mySQL에 contents 테이블로 저장</li> </ul>  |
| 데이터 전처리<br>과정   | <ol style="list-style-type: none"> <li>1. 타 플랫폼과 통일             <ul style="list-style-type: none"> <li>○ 키 값 통일(해당하는 키만 사용)</li> </ul> </li> </ol> <pre> "id": 고유번호숫자(매핑 용도), "type": 웹툰/웹소설, "platform":네이버 웹툰/네이버 시리즈/카카오웹툰/카카오페이지, "title": 작품 이름, "status": 연재/완결/휴재, "update_days": 월요일, 수요일/ 월요일/ 월요일, 화요일, 수요일...일요일(매일) "thumbnail": 표지 이미지 url, "genre": 장르, "views": 열람자수, "rating": 별점, - 카카오페이지웹툰,카카오페이지웹소설, 네이버시리즈웹소설 "likes": 좋아요,관심, - 카카오웹툰,네이버시리즈소설,네이버웹툰 "synopsis": 줄거리, "keywords": 키워드, "author":글 작가, "illustrator": 그림 작가, "original": 원작자, - 웹툰 하는 사람들만 "age_rating": 전체 이용가/12세 이용가/15세 이용가/19세 이용가 "price": 무료/유료/기다리면 무료 "url": base_url, "episode": 총 회차수, "comments": 총 댓글 수, "first_episode": 최초 연재일(YYYY.MM.DD.) </pre>  |

## 데이터 전처리 과정

### 2. 통합 인기 점수

- 장르별로 각 요소별 상관관계를 그려본 결과 차이를 발견



- 이 차이를 토대로 조회수, 별점, 댓글 수, 첫 연재일, 회차 수, 완결 여부 등에 가중치를 부여한 통합 인기 척도 생성
  - 월간 랭킹 순위와의 상관성을 부여하기 위해 장르별 TOP 300의 키워드와 시놉시스로 키워드 분석(예정)
  - 키워드 점수 + 독자 반응 점수(조회수, 별점, 댓글 수) + 최신성(첫 연재일, 회차 수, 완결 여부)를 합하여 최종 인기 척도 생성(예정)
  - 현재 드라마, BL, 무협, 현대 판타지 작업 완료
  - 로맨스, 로맨스 판타지, 판타지 작업 진행 중

## 데이터 전처리 결과

### 3. 인기 점수의 예시(가장 데이터 양이 적은(302개) 드라마 장르)

- 랭킹 기반 분석

- 공식:

$$I_{\text{랭킹}} = 1 - \frac{\text{ranking} - 1}{299}$$

$$I_{\text{rad}} = w_1 \cdot I_{\text{랭킹}} + w_2 \cdot N_{\text{조회}} + w_3 \cdot N_{\text{별점}} + w_4 \cdot N_{\text{댓글}} + w_5 \cdot N_{\text{키워드}}$$

- features

| feature  | 가중치  | 이유                          |
|----------|------|-----------------------------|
| rankng   | 0.5  | 장르 내 상대적인 인기를 직접적으로 반영      |
| views    | 0.2  | 사용자들의 실제 관심도(클릭 수, 유입량)를 반영 |
| rating   | 0.15 | 작품에 대한 독자의 만족도를 반영          |
| comments | 0.1  | 사용자들의 반응과 커뮤니티 활성도를 반영      |
| keywords | 0.05 | 인기 있는 작품의 공통적인 특징을 반영       |

- 1) 랭킹

- 월간 랭킹을 기반으로 작품의 상대적 위치를 확인.
- 하지만 랭킹만으로는 작품의 절대적인 인기를 판단하기 어려움.

- 2) 조회수 (Views)

- 사람들이 많이 찾는 작품일수록 대중적인 관심이 높다고 볼 수 있음.
- 특정 플랫폼에서 조회수 편향이 있을 수 있으므로, 로그 정규화를 활용하여 변동성을 줄임.

- 3) 평점 (Rating)

- 일부 작품은 평점이 낮은 경우에도 높은 조회수를 기록하는 경우가 있어 가중치를 0.3으로 설정하여 보조적 지표로 활용.

- 4) 댓글 수 (Comments)

- 커뮤니티 활동이 활발할수록 작품의 파급력이 크다는 것을 의미.
- 단, 댓글이 무조건 긍정적인 반응은 아닐 수 있음
- 따라서, 가중치를 0.2로 설정하여 반영.

- 5) 키워드 가중치 (Keyword Frequency)

- 인기작에서 자주 등장하는 키워드가 유사한 작품에도 영향을 미칠 가능성이 높음.
- 예를 들어, 법정 드라마에서 ‘변호사’, ‘판사’ 키워드가 반복 등장하면 해당 키워드를 가진 작품이 인기가 높을 확률이 큼.
- 하지만 키워드는 다른 요소들보다 영향력이 낮기 때문에 가중치를 0.1로 설정.

- 정규화 방법

$$N_{\text{조회수}} = \frac{\text{조회수} - \min(\text{조회수})}{\max(\text{조회수}) - \min(\text{조회수})}$$

$$N_{\text{별점}} = \frac{\text{별점} - \min(\text{별점})}{\max(\text{별점}) - \min(\text{별점})}$$

$$N_{\text{댓글수}} = \frac{\text{댓글 수} - \min(\text{댓글 수})}{\max(\text{댓글 수}) - \min(\text{댓글 수})}$$

$$N_{\text{키워드}} = \frac{\text{키워드 출현 횟수} - \min(\text{키워드 출현 횟수})}{\max(\text{키워드 출현 횟수}) - \min(\text{키워드 출현 횟수})}$$

- 평가(실제 랭킹과 비교)

- 복정기 (5위) → 랭킹 기반 = 0.7347 → 실제 랭킹이 높지만, 조회수 및 평점이 상대적으로 낮아 예상보다 낮은 척도.
- 드래곤 라자 (7위) → 랭킹 기반 = 0.7617 → 비교적 랭킹 기반 척도와 일치함.

## 데이터 전처리 결과

- 전체 기반 분석

- 공식 :  $I_{total} = w_1 \cdot N_{조회} + w_2 \cdot N_{별점} + w_3 \cdot N_{댓글}$
- features

| feature  | 가중치 | 이유                     |
|----------|-----|------------------------|
| views    | 0.5 | 작품의 절대적 인기를 나타내는 주요 지표 |
| rating   | 0.3 | 사용자 만족도를 측정하는 보조 지표    |
| comments | 0.2 | 커뮤니티 반응과 사용자 몰입도를 측정   |

### 1) 조회수 (Views)

- 작품이 얼마나 많은 사람들에게 노출되었는지를 의미하는 주요 지표.
- 절대적 인기를 판단하는 핵심 요소로, 가장 높은 가중치(0.5)를 부여.
- 조회수 편차가 클 경우 로그 정규화를 적용하여 변동성을 줄임.

### 2) 평점 (Rating)

- 작품을 본 독자들이 남긴 평가 점수.
- 특정 작품은 조회수는 높지만 평점이 낮을 수 있음.
- 따라서, 보조적인 품질 평가 지표로 0.3의 가중치를 부여.

### 3) 댓글 수 (Comments)

- 작품이 사람들에게 얼마나 화제가 되었는지를 측정.
- 작품이 인기가 많더라도 댓글이 많지 않을 수 있음(예: 조용한 명작).
- 가중치 0.2를 부여하여 반영.

- 정규화 방법

$$N_{조회수} = \frac{조회수 - \min(조회수)}{\max(조회수) - \min(조회수)}$$

$$N_{별점} = \frac{별점 - \min(별점)}{\max(별점) - \min(별점)}$$

$$N_{댓글수} = \frac{댓글수 - \min(댓글수)}{\max(댓글수) - \min(댓글수)}$$

- 극단적 분포 처리

- 정규화: 위와 같음
- 로그 변환
- 결측값 처리
  - 평점이 없는 경우: 전체 데이터의 평균 평점으로 대체
- 최소값 보정
  - 조회수, 댓글 수가 0인 경우, 최소값을 1로 강제 설정

- 평가(실제 랭킹과 비교)

- 신비소설 무(巫) (12위) →  $I_{total} = 0.9148$  → 랭킹(12위)에 비해 조회수와 평점이 압도적으로 높아 전체 인기에서는 1위 수준.
- 만행무승 (9위) →  $I_{total} = 0.7517$  → 전체 인기도에서도 높은 점수를 기록.
- 조선야설 (10위) →  $I_{total} = 0.7208$  → 전체 인기도에서도 비교적 높은 순위 유지.



|                   |  |
|-------------------|--|
| <p>데이터 전처리 결과</p> | <ul style="list-style-type: none"> <li>• 데이터 신뢰도 <ul style="list-style-type: none"> <li>◦ 평점은 조작될 가능성이 높으므로, 조회수와 댓글 수 같은 실제 사용자 반응 데이터를 주요 feature로 활용.</li> <li>◦ 조회수는 데이터 수집 방식(네이버, 카카오 등 플랫폼 차이)에서 편차가 있을 수 있어 정규화 과정을 거쳐 비교 가능하도록 조정.</li> </ul> </li> <li>• 트렌드 반영 <ul style="list-style-type: none"> <li>◦ 인기 작품들이 어떤 요소를 포함하는지 분석하기 위해 키워드 빈도를 활용.</li> <li>◦ 단순한 조회수 기반 인기작보다, 현재 트렌드를 반영한 작품이 더 오래 사랑받는 경향이 있어, 이를 보정하는 feature 추가.</li> </ul> </li> <li>• 극단적인 데이터 분포 문제 해결 <ul style="list-style-type: none"> <li>◦ 조회수가 극단적으로 높은 경우를 대비해 로그 정규화 적용.</li> <li>◦ 평점이 "-"인 경우, 평균 평점으로 대체하여 비교 가능하게 조정.</li> </ul> </li> </ul>  |
| <p>데이터 전처리 결과</p> | <ul style="list-style-type: none"> <li>• 결과<br/>이전)</li> </ul> <pre> {   "id": 66065996,   "type": "웹소설",   "platform": "카카오페이지",   "title": "구도만리",   "status": "매일 연재",   "thumbnail": "https://page-images.kakaoentcdn.com/download/resource?kid=cKTh0/hAJrQ5XZQI/1lz16S",   "genre": "드라마",   "views": "9.9만",   "rating": "10.0",   "synopsis": "소년행 작가님의 신작 &lt;구도만리&gt;가 새롭게 연재를 시작합니다!\n\n외판 산 중 암자에서",   "keywords": "시대물, 역사, 시대역사, 명작스테디셀러",   "author": "소년행",   "age_rating": "전체이용가",   "price": "1일 기다무 대여권",   "url": "https://page.kakao.com/content/66065996",   "episode": "전체 118",   "comments": "전체 189",   "first_episode": "25.02.13",   "ranking": 1 }, </pre> <p>이후)</p> <pre> {   "id":66065996,   "type":"웹소설",   "platform":"카카오페이지",   "title":"구도만리",   "status":"연재",   "update_days":"월요일, 화요일, 수요일, 목요일, 금요일, 토요일, 일요일",   "thumbnail":"https://page-images.kakaoentcdn.com/download/resource?kid=cKTh0/hAJrQ5XZQI/1lz16S",   "genre":"드라마",   "views":99000,   "rating":10.0,   "synopsis":"소년행 작가님의 신작 &lt;구도만리&gt;가 새롭게 연재를 시작합니다!\n\n외판 산 중 암자에서",   "keywords":"시대물, 역사, 시대역사, 명작스테디셀러",   "author":"소년행",   "age_rating":"전체 이용가",   "price":"기다리면 무료",   "url":"https://page.kakao.com/content/66065996",   "episode":118,   "comments":189,   "first_episode":"25.02.13",   "ranking":1.0,   "I_rank_adjusted":0.8399330535,   "I_total":0.7485944781 }, </pre> <ul style="list-style-type: none"> <li>• 향후 데이터 사용계획 <ul style="list-style-type: none"> <li>◦ RAG를 활용한 작품 검색 데이터</li> <li>◦ 웹 어플리케이션에서 사용자에게 보여줄 작품 정보보</li> </ul> </li> </ul> |

## 6. 네이버 지식인 글(웹툰)

|              |   |
|--------------|---|
| 개요           | <ul style="list-style-type: none"> <li>데이터 설명 <ul style="list-style-type: none"> <li>네이버 지식인에 ‘웹툰 추천’ 혹은 ‘웹툰 ㄸㄸ’이라고 검색했을 때 나오는 질문글과 답글을 크롤링 함.</li> </ul> </li> <li>데이터 수집목적 <ul style="list-style-type: none"> <li>이번 프로젝트의 챗봇의 목표가 위와 같이 웹 사이트에 웹툰을 추천해달라는 글을 올리는 대신 사용되는 것이기 때문에 사용자의 예상 질문과 알맞은 답을 확인하는 평가 데이터셋으로 사용 예정</li> </ul> </li> </ul>  |
| 데이터 자동화 및 검증 | <p>데이터 수집 자동화 프로세스</p> <ul style="list-style-type: none"> <li>api 호출로 1달에 한 번 크롤링 요청.</li> </ul>  |
| 데이터 저장 및 관리  | <ul style="list-style-type: none"> <li>데이터 저장 방식 <ul style="list-style-type: none"> <li>MySQL에 저장. 검색 도구로 사용 시 vector화 해서 벡터 데이터베이스에 저장 예정.</li> </ul> </li> </ul>  |
| 데이터 전처리 과정   | <p><b>전처리 단계 및 방법 설명</b></p> <p><b>[중복 데이터 제거]</b></p> <ul style="list-style-type: none"> <li>딕셔너리를 사용해 중복 데이터 저장 방지. ‘웹툰 추천’ 에서 검색된 글은 ‘웹툰 ㄸㄸ’을 포함하고 있어도 새로 저장되지 않음.</li> </ul> <p><b>[무의미한 데이터 삭제]</b></p> <ul style="list-style-type: none"> <li><b>답변이 없는 글</b><br/>답변의 개수가 0인 경우 데이터에서 제외</li> <li><b>삭제된 답변</b><br/>답글의 내용이 ‘삭제된 답변입니다.’ 인 경우 답글에서 제외</li> <li><b>ai 답변</b><br/>네이버 지식인 자체 ai 답변 기능이 있음. ai 답변의 불확실한 데이터는 학습하지 않는게 맞다고 판단하여 제거</li> </ul> |

## 데이터 전처리 과정

### 1개 답변

최적 ▼ 추천순 ▼



지식이

필대신

#AI답변 #지식이는실험중

좀비 아포칼립스 웹툰을 찾고 계시는군요! 좀비 아포칼립스 웹툰을 추천해 드릴게요.

하우스키퍼: 채용택 작가와 유현 작가가 만든 SF 아포칼립스 작품으로 네이버 웹툰과 네이버 시리즈에서 2019년 1월 22일부터 연재되었어요. AI와 좀비 아포칼립스가 결합된 독특한 설정과 화려한 액션 신이 특징이에요. 또한 한국 웹툰 최초로 애니메이션화가 예정되어 있으며, 슈퍼 스트링의 12번째 작품으로서 다른 웹툰과의 연계도 기대돼요.

폰투스: 2023년 5월 30일 기준으로 호평을 받고 있는 작품으로, 스토리와 작화가 모두 뛰어나다는 평가를 받아요. 이 웹툰을 보지 않았다면 인생 낭비라는 말이 있을 정도로 강력 추천되고 있어요.

극야: 2018년 12월 22일부터 연재를 시작했으며, 남극에 파견된 한중일 연구원들이 외계 바이러스로 인해 차례로 감염되면서 생존과 지구 보호를 위해 싸우는 이야기를 담고 있어요. 3국의 캐릭터가 등장하며, 영어를 공용어로 사용하기 때문에 웹툰에서는 4개 국어를 만나볼 수 있어요. 이 작품은 시즌 2가 2022년에 공개되었으며, 산악 구조 웹툰처럼 보이는 섬네일로 인해 초기에는 주목받지 못했지만, '본격 남극 좀비 물'이라는 문구로 인해 인기를 얻게 되었어요.

극야 외 추천 작품: '어느 날 갑자기 서울은', '살아남은 로맨스', '미시령', '그림자의 밤', '좀간', '스위트홈', '염총 소년', '지금 우리 학교는', '리턴 서바이벌', '블러디 발렌타인: 인류 종말', '드림사이드', '심해수' 등이 있어요.

위 추천 작품들은 모두 좀비 아포칼립스 장르를 다루고 있으며, 각기 다른 매력과 스토리를 가지고 있어요. 취향에 따라 선택하시면 좋을 것 같아요.

#AI답변

#웹툰추천

#좀비

#아포칼립스

#네이버웹툰

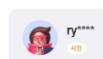
#연재중인웹툰

2025.02.21.

- 삭제된 답변과 ai 답변을 제거했더니 답변이 아무것도 남지 않은 상태라면 데이터에서 제외.

### [형식 변환]

#### - 들여쓰기



ny\*\*\*

본인 등록 정보 정보

그레이트 빠전- 시즌1 종료후 잠든게 단점

천재의사 아무진

레벨업 닥터 최기석

메디컬 환생

등이 카카오페이지에 있습니다

개인적으로 원작소설을 다 읽어봤고 개인적인 호불호는

소설원작상 닥터 최태수 > 그레이트 빠전 > 레벨업 닥터 > 메디컬환생 > 천재의사 아무진 순이었습니다.

☆ 안타깝지만 모든 작품들이 원작을 50% 도 완성하지 못했습니다.

2024.05.17.



```
<p class="se-text-paragraph se-text-paragraph-align- " style id="SE-d5cc33fa-bc5d-49b6-a8aa-1410b5b0c0e">
  <span style class="se-fs- se-ff- " id="SE-b13d2070-ea3d-4a7f-b325-cf4671e0073">그레이트 빠전- 시즌1 종료후 잠든게 단점 </span> == $0
</p>
<!-- SE-TEXT -->
<!-- SE-TEXT { -->
<p class="se-text-paragraph se-text-paragraph-align- " style id="SE-9ce0f9a6-2ecc-4815-808b-fdea07abd994">
  <span style class="se-fs- se-ff- " id="SE-cf9fb26c-a49d-46aa-9013-6092d7a713dd">
    &ZeroWidthSpace</span>
  </p>
<!-- SE-TEXT -->
<!-- SE-TEXT { -->
<p class="se-text-paragraph se-text-paragraph-align- " style id="SE-143df223-d4a5-4b8f-bb7d-b1641185de31">
  <span style class="se-fs- se-ff- " id="SE-92f4cf7-ca9d-4fde-86a4-664807533c05">천재의사 아무진 </span>
  </p>
<!-- SE-TEXT -->
<!-- SE-TEXT { -->
<p class="se-text-paragraph se-text-paragraph-align- " style id="SE-7758534a-91a1-46f0-bd3c-c79b4da71c25">
  <span style class="se-fs- se-ff- " id="SE-a5f3f712-6917-41cb-b708-d01dc60c6b4c">
    &ZeroWidthSpace</span>
  </p>
<!-- SE-TEXT -->
<!-- SE-TEXT { -->
<p class="se-text-paragraph se-text-paragraph-align- " style id="SE-c9a717fe-f143-4333-b964-523415e6b0de">
  </p>
<!-- SE-TEXT -->
<!-- SE-TEXT { -->
```

지식인의 답글 형식은 다음과 같이 한 줄씩 담겨 있음. 들여쓰기를 하는 줄로 내용이 담겨 있지 않고 ‘ZeroWidthSpace;’인 경우 답글의 내용에서 제외. 나머지 글들은 띄어쓰기로 이어 하나의 문자열로 합함.

## 데이터 전처리 결과

### ● 결과

```
주 버터를 위한 할 그림자의 참 됨은 연이 미망의 고백 드레논의 평생을 가져다 갔습니다. 다시, 아슬하게 흐르는 불완전 신데렐라를 전생연분 임
"그저 여영일분(카카페) 줄가미(네이버) 내가 죽기로 결심한 것은(네이버) 벡투더 진비(네이버) 이기적연애론(네이버) 토마토가 돼라!(네이버) 불
}
{
"title": "웹툰 추천",
"question": "안녕하세요!! 정략결혼 해서 남주가 여주한테 엄청 막 대하고 여주는 가족한테 상처받고 어려서 남주한테 상처받는 내용의 웹툰 추천해
"recommend": [
"네이버웹툰> 나를 미워하던 남편이 기억을 잃었다 혼한 빙의물인 줄 알았다 <카카오페이지> 남편을 내 편으로 만드는 방법 누군가 내 몸에 빙의
"나를 미워하는 남편이 기억을 잃었다 내가 죽기를 바랄때가있다 이거 두 개 꼭 보시길 바랍니다!!!"
]
},
{
"title": "월요일웹툰 추천해주세요!!",
"question": "월요일 웹툰 추천해주세요!! 로맨스 일절 없고, 사람 캐릭터고, 잔인한건 없는그런 재밌는 웹툰 있습니다ㅠㅠ",
"recommend": [
"뚝 달은 말 ☆☆드려요",
"다육이는 잘 자란다> 추천드려요! 다육이가 살아가는 이야기인데, 블랙코미디류예요"
]
},
{
"title": "웹툰 추천해주세요!",
"question": "‘/중간에서 만나’ 라는 웹툰이랑 비슷한 웹툰 추천해주세요! 로맨스 웹툰이요!",
"recommend": [
"남사친의 법칙이 가장 비슷할것 같은데..",
"현대를 세계 서열 0위님, 제발 그만해 주세요 어느날 짝남에게 공작님이 빙의했다. 솔로스를 암호는 002 전력고백(신) 용한소녀 아이스크림 소년
"비슷한지는 잘모르겠는데 불완전 신데렐라를 추천 드립니다"
```

### ● 향후 데이터 사용계획

- 이 파일을 벡터화해서 검색하는 방식도 고려 중.