

## 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 : 25.02.17.
- 깃허브 경로 : [git@github.com:SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-4Team.git](https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-4Team.git)
- 작성 팀원 : 고성주 , 김동명(수정)

개요	<ul style="list-style-type: none"><li>• 데이터 설명 : 네이버 영화 리뷰, TMDB 영화 키워드, 왓차피디아 영화 리뷰</li><li>• 데이터 수집목적 : 리뷰를 감정 분류하여, 영화 평점을 계산 모델 생성. 줄거리 및 키워드를 이용하여 영화 추천 모델 생성</li></ul>
데이터 자동화 및 검증	<ul style="list-style-type: none"><li>• 데이터 수집 자동화 프로세스 : beautifulsoup과 selenium을 활용하여, 각 사이트별 크롤링 자동화 진행</li><li>• 세부 내용은 제목, 장르, 명대사, 리뷰, 감독, 배우 등등</li><li>• 검증 : 초기 1회에 한하여 이미지에 대한 검증 수작업 진행 (크롤링 및 해상도는 자동으로 진행되나, 제목과 이미지의 매칭 여부 확인 필요)</li></ul>
데이터 저장 및 관리	<ul style="list-style-type: none"><li>• 데이터 저장 방식(영화 기본정보, 리뷰)<ul style="list-style-type: none"><li>◦ 엑셀 파일로 관리</li><li>◦ 텍스트의 경우 엑셀에 입력되지 않는 글자(특수문자) 제외하고 수집</li><li>◦ 리뷰 수집 시 셀렉터가 바뀐 영화 30여개 추가 크롤링</li></ul></li><li>• 이미지 : 별도 DB없이 드라이브에 저장. (영화제목을 파일명으로 사용하여, 영화 이미지 검색시 파일명을 검색하여 사용)</li><li>• Primary key<ul style="list-style-type: none"><li>◦ 시리즈의 경우 1,2,... 와 같이 숫자로 구분하며 동일 제목의 영화는 인지도가 높은 영화로 선정(manual)</li></ul></li></ul>
데이터 전처리 과정	<ul style="list-style-type: none"><li>• 전처리 단계 및 방법 설명<ul style="list-style-type: none"><li>◦ 영화 기본정보는 따로 수정 X</li><li>◦ 영화 이미지명 호출에 편리하도록 조정.</li><li>◦ 이모티콘은 감정분류가 가능하며, 특수 이미지는 크롤링 단계에서 필터링됨</li></ul></li></ul>
데이터 전처리 결과	<ul style="list-style-type: none"><li>• 결과<ul style="list-style-type: none"><li>- 7005개의 영화에 대한 데이터 확보</li><li>- 이 중 인지도에 따라 약 3천개의 영화 선정</li><li>- 리뷰 개수 30개 미만의 영화 제거(28개) 리뷰 갯수 평균 128.5, 중간값 147, 1분위수 100, 3분위수 150</li></ul></li><li>• 향후 데이터 사용계획<ul style="list-style-type: none"><li>- 향후 새로운 영화 혹은 시리즈 있을 시 데이터 추가.</li></ul></li></ul>

