SK네트웍스 **Family AI**과정 **6**기

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
- 작성 팀원 : 고성주

개요	• 데이터 설명 : 네이버 영화 리뷰, TMDB 영화 키워드 • 데이터 수집목적 : 리뷰를 감정 분류하여, 영화 평점을 계산 모델 생성. 키워드를 이용하여 영화 추천 모델 생성
데이터 자동화 및 검증	• 데이터 수집 자동화 프로세스 : beautifulsoup과 selenium을 활용하여, 각 사이트별 크롤링 자동화 진행 • 검증 : 초기 1회에 한하여 이미지에 대한 검증 수작업 진행 (크롤링 및 해상도는 자동으로 진행되나, 제목과 이미지의 매칭 여부 확인 필요)
데이터 저장 및 관리	• 데이터 저장 방식(영화 기본정보, 리뷰) ○ sql로 관리: • 이미지 : 별도 DB없이 드라이브에 저장. (영화제목을 파일명으로 사용하여, 영화 이미지 검색시 파일명을 검색하여 사용)

데이터 전처리 과정	• 전처리 단계 및 방법 설명
	  ○ 영화 기본정보는 따로 수정 X
	  ○ 영화 이미지명 호출에 편리하도록 조정.

데이터 전처리 결과	• 결과
	  - 7005개의 영화에 대한 데이터 확보
	• 향후 데이터 사용계획
	  - 향후 새로운 영화 혹은 시리즈 있을 시 데이터 추가.