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 : 02.17.
- 깃허브 경로 : git@github.com:SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-4Team.git
- 작성 팀원 : 조하늘, (수정) 정지원, 김동명

데이터 전처리 개요	• 문서유형 : TEXT - 영화 정보, 영화 리뷰, 학습 데이터

전처리 과정	● 영화 LIST ○ 불필요한 영화 목록 제거 (15분 영화, 비주류 영화 등) ○ 영화 추천 인플루언서 → 불필요한 영화 LIST 필요 없음 ● 영화 정보 ○ 전처리 할 데이터 없음 ○ 영화 특성 상 일부 고유명사가 중요한 요소인데 전처리 진행 시엔 해당 요소 손실 우려가 있음 ○ 필요 토큰과 불필요한 토큰의 구분 불가 ■ (줄거리) 변호사 직업을 가진 주인공이 펼치는 법정 드라마 → “변호사” 토큰 필요 ■ (줄거리) 변호사 직업을 가진 주인공이 길에서 우연히 첫사랑을 만난 로맨스 → “변호사” 토큰 불필요 ■ 키워드 생성 시에 줄거리 고려하여 필요한 키워드만 생성 → 전처리 불필요 ● 영화 리뷰, 학습 데이터 ○ 한국어 특성 상 조사, 어미에 따라 내용이 달라지기 때문에 전처리할 경우 의미 손상 가능성이 높음 ○ 특정 이모티콘 ‘ㅠㅠㅠㅠ’, ‘ㅜㅜ’ 이 감정을 나타내고 있기 때문에 불용어 처리 불가 ■ 예시와 같은 이모티콘들은 감정 클러스터 분석 모델에서 감정으로 분석 가능하여 불용어 처리할 필요 없음 ○ 학습 데이터의 경우 비속어 전처리가 되어 있음 ○ 엑셀에 입력되지 않는 글자(특수문자) 제외하고 수집 ○ 시리즈의 경우 1,2… 와 같이 숫자로 구분하며 동일 제목의 영화는 인지도가 높은 영화로 선정(manual)
데이터 전처리 결과	● 영화 LIST 수정 → 2934개 영화 DB 확보(제목 일치, 리뷰 30개 이상, 기본 정보 모두 포함) ● 리뷰 갯수 30개 미만의 영화 제거(28개) ● 리뷰 갯수 평균 128.5 중간값 147, 1/4분위수 100, 3/4분위수 150