SK네트웍스 **Family AI** 과정 **12**기

데이터 전처리 인공지능 데이터 전처리 결과서

산출물 단계	데이터 전처리
평가 산출물	인공지능 데이터 전처리 결과서
제출 일자	2025.07.20
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-6TEAM
작성 팀원	조성지, 남의헌, 이정민

1. 문서 개요

- 프로젝트명: 마이무디(My Moody, AI 그림 심리 분석 및 페르소나 챗봇)
- 전처리 목적: 사용자의 그림 검사 결과에 나타난 감정을 분석하고, 이를 챗봇 페르소나의 5가지 유형(추진형, 내면형, 관계형, 안정형, 쾌락형)으로 분류하기 위함
- 문제 정의: 감정 분류와 관련된 논문 데이터에서 에니어그램 기반 감정 표현 키워드를 추출 및 정제하고, 이를 5가지 성격 유형으로 라벨링한 뒤, 키워드와 라벨에 맞는 심리 분석 결과 텍스트를 생성하여, 이러한 텍스트를 기반하여 분류 모델을 학습시키기 위한 데이터셋 구축

2. 데이터셋 개요

- 데이터 출처 및 수집 방법:
    - AI-HUB '감성 대화 말뭉치'
        - 데이터셋 설명 : 일반인 1,500명을 대상으로 하여 음성 15,700문장 및 코퍼스 27만 문장 구축 및 세대별 감성 대화 텍스트
        - 수집 방법 : JSON 파일의 형태로 직접 다운로드
    - DBPIA 및 RISS 논문
        - 데이터셋 설명 : 학술정보 포털 사이트에서 감정 분류와 관련된 논문 원문
        - 수집 방법 : 웹 사이트에서 원문 PDF 직접 다운로드 후, PyPDF2와 re 라이브러리를 활용한 Python 스크립트로 텍스트 추출
- 데이터 구성:

항목명	설명	예시
text	심리 분석 결과와 비슷한 텍스트	"당신은 일을 추진하는 데 있어서 명확한 목표가 필요하며 이를 위해 계획을 철저히 세우는 경향이 보입니다."

label	5가지 성격 유형	“추진형”
keyword	텍스트에서 추출한 핵심 키워드	“명확하다”

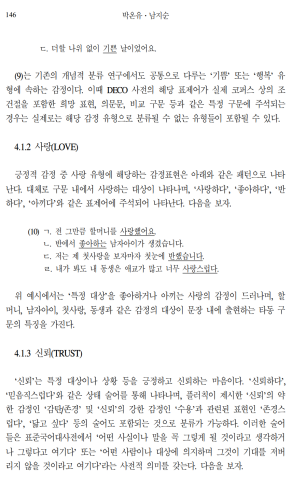
- 원본 데이터 샘플(5~10건 첨부):


```
"content": {"HS01": "직장에서 막내라는 이유로 나에게만 온갖 심부름을 시켜. 일도 많은 데 정말 분하고 섭섭해.", "SS01": "관련 없는 심부름을 모두 하게 되어서
"content": {"HS01": "얼마 전 입사한 신입사원이 나를 무시하는 것 같아서 너무 화가 나.", "SS01": "무시하는 것 같은 태도에 화가 나셨군요. 상대방의 어떤 행동이
"content": {"HS01": "직장에 다니고 있지만 시간만 버리는 거 같아. 진지하게 진로에 대한 고민이 생겨.", "SS01": "진로에 대해서 고민하고 계시는군요. 어떤 점이
"content": {"HS01": "성인인데도 진로를 아직도 못 정했다고 부모님이 노여워하셔. 나도 섭섭해.", "SS01": "부모님의 노여움에 섭섭하시군요. 이런 상황을 어떻게 하
"content": {"HS01": "퇴사한 지 얼마 안 됐지만 천천히 직장을 구해보려고.", "SS01": "천천히라도 직장을 구해 보려고 하시는군요. 특별한 이유가 있으신가요?", "
"content": {"HS01": "졸업반이라서 취업을 생각해야 하는데 지금 너무 느긋해서 이래도 되나 싶어.", "SS01": "취업에 대해 걱정이 되는군요.", "HS02": "응. 느긋
"content": {"HS01": "요즘 직장생활이 너무 편하고 좋은 것 같아!", "SS01": "직장생활이 편하고 좋으시다니 좋아 보여요. 다니고 계신 회사만의 장점이 있나요?",
"content": {"HS01": "취업해야 할 나이인데 취업하고 싶지가 않아.", "SS01": "취업하고 싶지 않으시군요. 혹시 어떤 이유가 있을까요?", "HS02": "아직 조금 더
"content": {"HS01": "면접에서 부모님 직업에 대한 질문이 들어왔어.", "SS01": "그때 어떤 생각이 들었나요?", "HS02": "무척 당혹스러웠어. 부모님 직업이 면접이
"content": {"HS01": "큰일이야. 부장님께 결재받아야 하는 서류가 사라졌어. 한 시간 뒤에 제출해야 하는데 어디로 갔지?", "SS01": "중요한 서류가 사라져서 당혹스
"content": {"HS01": "나 얼마 전에 면접 본 회사에서 면접 합격했다고 연락받았었는데 오늘 다시 입사 취소 통보받아서 당혹스러워.", "SS01": "회사에서 합격 통보
"content": {"HS01": "길을 가다가 우연히 마주친 동네 아주머니께서 취업했냐고 물어보셔서 당황했어.", "SS01": "취업에 대한 질문이 당혹스러우셨군요.", "HS02":
"content": {"HS01": "어제 합격 통보를 받은 회사에서 문자를 잘못 발송했다고 연락이 왔어. 너무 당혹스럽고 속상해.", "SS01": "잘못 발송된 문자로 당황하셨군요.
"content": {"HS01": "나 오늘 첫 출근 했는데 너무 당황스러웠어!", "SS01": "무슨 일 있었나요?", "HS02": "버스 타고 카드를 찍으려고 하는데 지갑을 바꿔서 가
"content": {"HS01": "이번에 직장을 이직했는데 글쎄 만나고 싶지 않은 사람을 만나서 아주 당황스럽더라고.", "SS01": "이직하신 직장에서 만나고 싶지 않으신 분을
"content": {"HS01": "코로나 때문에 뭘 할 수가 없어. 취직 준비를 해야 하는데 시험이 줄줄이 취소되니 말이야. 마비라는 말이 이럴 때 참 어울리는 것 같아.", "SS
"content": {"HS01": "오늘 회사에서 큰 실수를 한 것 같아.", "SS01": "무슨 실수를 하셨나요?", "HS02": "중요한 발표가 있었는데 많은 사람 앞에서 말을 해야 하
"content": {"HS01": "요즘 취업 관련해서 떠올리기만 해도 온몸이 마비될 것 같아.", "SS01": "온몸에 마비될 것 같은 기분이 드신다니 현재 어떤 상황인지 여쭤봐도
"content": {"HS01": "어제도 야근 오늘도 야근이야. 너무 힘들어.", "SS01": "야근이 많아서 힘드신 것 같아요. 어떤 상황인지 자세히 말씀해 주시겠어요? ", "HS0
"content": {"HS01": "우리 회사는 정말 사내 분위기가 좋아.", "SS01": "사내 분위기가 좋아서 즐거우시군요.", "HS02": "즐거운 일이 매일 생길 거 같은 기분이야
"content": {"HS01": "오늘 내가 다니는 회사가 참 좋은 직장이라는 생각이 들었어.", "SS01": "직장이 좋으시군요. 자세히 알려주실 수 있나요?", "HS02": "회사에
"content": {"HS01": "회사에서 전공시험을 봤어. 오늘 시험 결과가 나왔어.", "SS01": "시험 결과가 어떠셨나요?", "HS02": "열심히 준비한 만큼 원하던 점수가 나
"content": {"HS01": "면접관에게 완전히 속았어. 면접일에 알려준 연봉과 실수령액이 꽤 차이가 나네.", "SS01": "연봉과 실수령액이 차이가 나시는군요.", "HS02":
"content": {"HS01": "지인이 취업시켜준다면서 사기를 쳤어. 배신감이 들어.", "SS01": "지인에게 당한 사기라 더 배신감이 크시군요. 무슨 일이 있었는지 알려주시
"content": {"HS01": "우리 부모님은 나를 버린 거나 다름없어. 내가 진로에 대해 걱정을 할 때 나를 걱정 해주거나 돌봐 주지를 않았어.", "SS01": "부모님으로부터
"content": {"HS01": "계속 취업이 안 되니까 버려진 기분이야.", "SS01": "계속 취업이 안 돼서 버려진 기분이군요.", "HS02": "내가 능력이 없어서 취업이 안 되
"content": {"HS01": "이번에 중소기업에 취업하게 되었어. 친구들에게 말하기가 조금 부끄러워.", "SS01": "중소기업에 취업해 부끄러운 마음이 들었군요.", "HS02"
"content": {"HS01": "저번 주에 친구와 같은 회사에서 같이 면접을 봤는데 나만 떨어져서 너무 창피해.", "SS01": "친구분과 같이 면접을 봤는데 혼자만 떨어지셔서
"content": {"HS01": "요즘 청년 실업률이 너무 심각한 거 같아. 취업해야 하는데 기업들이 채용을 많이 하지를 않아.", "SS01": "청년 실업률의 심각성에 대해서 고
"content": {"HS01": "직장에서 부당한 일을 겪어서 너무 화가 나.", "SS01": "부당한 일을 당해서 속상하시군요. 주변 사람에게 말해보면 감정이 조금은 나아질까요?
"content": {"HS01": "오늘 엄마랑 진로 문제로 크게 싸웠어. 그래서 너무 화가 나.", "SS01": "어머니랑 진로 문제로 싸우셨군요. 자세하게 알려주실 수 있나요?",
"content": {"HS01": "오늘 면접은 너무 망했어!", "SS01": "면접 때 무슨 일이 있으셨나요?", "HS02": "낙하산이 있었던 것 같아. 모두 그 사람한테만 말을 걸고
"content": {"HS01": "부모님이 자꾸 취업하라고 잔소리해. 나도 아는데 자꾸 잔소리하니까 화가 나.", "SS01": "취업하라는 잔소리 때문에 화가 나는군요. 어떻게 하
"content": {"HS01": "이제 주변 친구들은 하나둘씩 다 취업에 성공하는데 아직 나만 못해서 불안해. 이러다 나만 못하는 것은 아닐까 걱정돼.", "SS01": "친구는 이
"content": {"HS01": "나는 내 꿈이 무엇인지 몰라서 불안해.", "SS01": "꿈이 무엇인지 모르셔서 불안하시군요.", "HS02": "꿈은 모르겠지만 컴퓨터로 무언가를 작
"content": {"HS01": "코로나 때문에 출근하는 것이 불안해.", "SS01": "출근으로 인한 코로나 감염이 불안하신가 봐요. 불안한 상황을 극복할 방법이 있을까요? ",
"content": {"HS01": "내년이 졸업인데 취업 문제 때문에 걱정이야.", "SS01": "취업 때문에 걱정이 많으시군요. 어떤 부분이 가장 걱정되시나요? ", "HS02": "학점
```


〈표 8〉 성격유형별 형용사 키워드

성격 유형	형용사 키워드
1번	현명하다 공정하다 심각하다 불안하다 옳다 착하다 올바르다 강하다 정확하다 충분하다 철저하다 급하다 심하다 진지하다 날카롭다 집착하다 확실하다 완벽하다 초조하다 치열하다 엄격하다 지혜롭다 성실하다 불편하다 자유롭다 필요하다 당연하다 당당하다 재미있다 평범하다 다양하다 성숙하다 새롭다 성장하다 끊임없다 짐요하다 탁월하다 멋지다 사소하다 다르다
2번	다정하다 다감하다 따뜻하다 친절하다 친하다 깊다 너그럽다 예민하다 지나치다 강하다 필요하다 민감하다 아프다 불편하다 사랑스럽다 행복하다 즐겁다 아름답다 사소하다 친숙하다 교만하다 거만하다 단순하다 낯설다 불안하다 괴롭다 자유롭다 세심하다 충실하다 진지하다 변하다 뭉클하다 성장하다 성숙하다 섬세하다 멋지다 복잡하다 솔직하다 원하다 좋아하다
3번	부지런하다 원하다 좋다 행복하다 과하다 다재다능하다 친근하다 친절하다 편리하다 유능하다 똑똑하다 꾸준하다 끊임없다 여전하다 명확하다 독특하다 특별하다 뛰어나다 진지하다 유익하다 필요하다 거짓되다 무시하다 교묘하다 무리하다 힘들다 다르다 적응하다 기대하다 성장하다 따뜻하다 철저하다 충실하다 유용하다 확실하다 성공하다 훌륭하다 새롭다 정확하다 완벽하다
4번	따뜻하다 우울하다 외롭다 강하다 고집스럽다 심하다 특별하다 뛰어나다 아름답다 우아하다 다르다 아프다 민감하다 훌륭하다 고상하다 중요하다 세련되다 멋지다 차분하다 진솔하다 소중하다 섬세하다 생소하다 초연하다 까다롭다 격렬하다 들쑥날쑥하다 예민하다 독특하다 풍부하다 슬프다 무섭다 그립다 독특하다 소중하다 친절하다 성장하다 심각하다 괴롭다 두렵다
5번	민감하다 현명하다 세다 높다 중요하다 흥미롭다 새롭다 필요하다 기대하다 교만하다 인색하다 신중하다 진지하다 재미있다 영리하다 예리하다 날카롭다, 자세하다 초연하다 무관심하다 무신경하다 차갑다 완고하다 탐욕스럽다 맑이 없다 추리하다 꼼꼼하다 바쁘다 놀랍다 신비롭다 어렵다 유명하다 단순하다 명쾌하다 냉정하다 꾸준하다 성장하다 정확하다 딱딱하다 복잡하다

성격 유형	형용사 키워드
6번	불가능하다 충직하다 따뜻하다 착하다 조심스럽다 충실하다 공손하다 상냥하다 친절하다 충성스럽다 정직하다 올바르다 성공하다 신중하다 진지하다 심각하다 무겁다 결연하다 완고하다 유용하다 익숙하다 초조하다 소심하다 순수하다 우유부단하다 불안하다 걱정스럽다 우아하다 단순하다 안전하다 섬세하다 위험하다 험난하다 두렵다 낯설다 만족하다 용감하다 중요하다 힘겹다 평범하다
7번	제미있다 즐겁다 흥미롭다 신비롭다 새롭다 놀랍다 신기하다 엉뚱하다 황당하다 기발하다 풍부하다 열렬하다 열광하다 치열하다 민첩하다 신속하다 산만하다 강하다 행복하다 멋지다 즐겁다 특별하다 기쁘다 멍랑하다 활기차다 발랄하다 유쾌하다 경쾌하다 힘차다 순수하다 가볍다 무르다 무책임하다 다채롭다 다양하다 대단하다 남다르다 똑똑하다 훌분하다 독특하다
8번	진지하다 솔직하다 치열하다 안전하다 뛰어나다 위험하다 엄격하다 충직하다 둔감하다 단순하다 힘세다 강하다 자신만만하다 대담하다 넓다 조화롭다 정의롭다 정당하다 유능하다 통명스럽다 무정하다 둔하다 거칠다 과격하다 급격하다 신랄하다 용감하다 열광하다 위대하다 오만하다 치열하다 썩썩하다 약하다 위험하다 엄숙하다 적합하다 올바르다 창피하다 새롭다 어렵다
9번	고집스럽다 무신경하다 상냥하다 평화롭다 너그럽다 여유롭다 겸손하다 편안하다 침착하다 평온하다 조화롭다 정의롭다 느긋하다 지루하다 우유부단하다 장황하다 안일하다 나태하다 태만하다 초연하다 완고하다 충실하다 순수하다 단순하다 간단하다 평범하다 따뜻하다 안전하다 급격하다 적절하다 연결하다 간결하다 진솔하다 깔끔하다 다양하다 필요하다 여전하다 적절하다 남다르다 어렵다



146 박은음 · 남지순

ㄷ. 더할 나위 없이 기쁜 날이었어요.

(9)는 기존의 개념적 분류 연구에서도 공통으로 다루는 '기쁨' 또는 '행복' 유형에 속하는 감정이다. 이때 DECO 사전의 해당 표제어가 실제 코퍼스 상의 조건절을 포함한 희망 표현, 의문문, 비교 구문 등과 같은 특정 구문에 주석되는 경우는 실제로는 해당 감정 유형으로 분류될 수 없는 유형들이 포함될 수 있다.

4.1.2 사랑(LOVE)

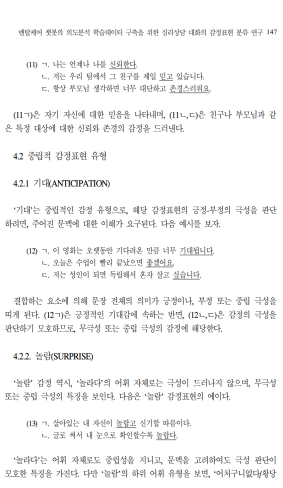
긍정적 감정 중 사랑 유형에 해당하는 감정표현은 아래와 같은 패턴으로 나타난다. 대체로 구문 내에서 사랑하는 대상이 나타나며, '사랑하다', '좋아하다', '반하다', '아끼다'와 같은 표제어에 주석되어 나타난다. 다음을 보자.

(10) ㄱ. 전 그만큼 할머니를 사랑했어요.
ㄴ. 반에서 좋아하는 남자아이가 생겼습니다.
ㄷ. 저는 제 첫사랑을 보자마자 첫눈에 반했습니다.
ㄹ. 내가 봐도 내 동생은 예쁘고 착한 너무 사랑스럽다.

위 예시에서는 '특정 대상'을 좋아하거나 아끼는 사랑의 감정이 드러나며, 할머니, 남자아이, 첫사랑, 동생과 같은 감정의 대상이 문장 내에 출현하는 타동 구문의 특징을 가진다.

4.1.3 신뢰(TRUST)

'신뢰'는 특정 대상이나 상황 등을 긍정하고 신뢰하는 마음이다. '신뢰하다', '믿음직스럽다'와 같은 상태 술어를 통해 나타나며, 플러칙이 제시한 '신뢰'의 약한 감정인 '감탄/존경' 및 '신뢰'의 강한 감정인 '수용'과 관련된 표현인 '존경스럽다', '닮고 싶다' 등의 술어도 포함되는 것으로 분류가 가능하다. 이러한 술어들은 표준국어대사전에서 '어떤 사실이나 말을 꼭 그렇게 될 것이라고 생각하거나 그렇다고 여기다' 또는 '어떤 사람이나 대상에 의지하여 그것이 기대를 저버리지 않을 것이라고 여기다'라는 사전적 의미를 갖는다. 다음을 보자.



텐탈에서 찾보의 의도분석 학습데이터 구축을 위한 심리상담 대화의 감정표현 분류 연구 147

(11) ㄱ. 나는 언제나 나를 신뢰한다.
ㄴ. 저는 우리 팀에서 그 친구를 제일 믿고 있습니다.
ㄷ. 항상 부모님 생각하면 너무 대단하고 존경스러워요.

(11ㄱ)은 자기 자신에 대한 믿음을 나타내며, (11ㄴ,ㄷ)은 친구나 부모님과 같은 특정 대상에 대한 신뢰와 존경의 감정을 드러낸다.

4.2 중립적 감정표현 유형

4.2.1 기대(ANTICIPATION)

'기대'는 중립적인 감정 유형으로, 해당 감정표현의 긍정-부정의 극성을 판단하려면, 주어진 문맥에 대한 이해가 요구된다. 다음 예시를 보자.

(12) ㄱ. 이 영화는 오랫동안 기다려온 만큼 너무 기대됩니다.
ㄴ. 오늘은 수업이 빨리 끝났으면 좋겠어요.
ㄷ. 저는 성인이 되면 독립해서 혼자 살고 싶습니다.

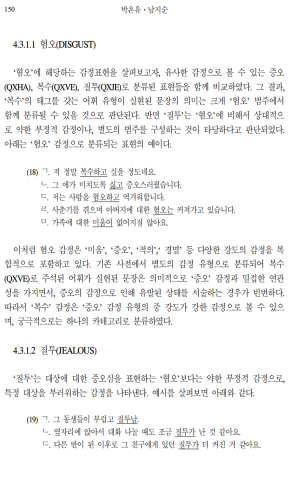
결합하는 요소에 의해 문장 전체의 의미가 긍정이나, 부정 또는 중립 극성을 띠게 된다. (12ㄱ)은 긍정적인 기대감에 속하는 반면, (12ㄴ,ㄷ)은 감정의 극성을 판단하기 모호하므로, 부극성 또는 중립 극성의 감정에 해당한다.

4.2.2. 놀람(SURPRISE)

'놀람' 감정 역시, '놀라다'의 어휘 자체로는 극성이 드러나지 않으며, 부극성 또는 중립 극성의 특징을 보인다. 다음은 '놀람' 감정표현의 예이다.

(13) ㄱ. 잃어있는 내 자신이 놀랍고 신기할 따름이다.
ㄴ. 글로 써서 내 모으로 확인할수록 놀랐다.

'놀라다'는 어휘 자체로도 중립성을 지니고, 문맥을 고려하여도 극성 판단이 모호한 특징을 가진다. 다만 '놀람'의 하위 어휘 유형을 보면, '어처구니없다/황당



150 박은음 · 남지순

4.3.1.1 혐오(DISGUST)

'혐오'에 해당하는 감정표현을 살펴보고자, 유사한 감정으로 볼 수 있는 증오(QXHA), 복수(QXVE), 질투(QXJE)로 분류된 표현들을 함께 비교하였다. 그 결과, '복수'의 태그를 갖는 어휘 유형이 실현된 문장의 의미는 크게 '혐오' 범주에서 함께 분류될 수 있을 것으로 판단된다. 반면 '질투'는 '혐오'에 비해서 상대적으로 약한 부정적 감정이나, 별도의 범주를 구성하는 것이 타당하다고 판단되었다. 아래는 '혐오' 감정으로 분류되는 표현의 예이다.

(18) ㄱ. 저 정말 복수하고 싶을 정도네요.
ㄴ. 그 애가 미치도록 싫고 증오스러웠습니다.
ㄷ. 저는 사람을 혐오하고 역겨워합니다.
ㄹ. 사춘기를 겪으며 아버지에 대한 혐오는 커져가고 있습니다.
ㅁ. 가족에 대한 미움이 없어지질 않아요.

이처럼 혐오 감정은 '미움', '증오', '적의', '경멸' 등 다양한 강도의 감정을 복합적으로 포함하고 있다. 기존 사전에서 별도의 감정 유형으로 분류되어 복수(QXVE)로 주석된 어휘가 실현된 문장은 의미적으로 '증오' 감정과 밀접한 연관성을 가지면서, 증오의 감정으로 인해 유발된 상태를 서술하는 경우가 빈번하다. 따라서 '복수' 감정은 '증오' 감정 유형의 중 강도가 강한 감정으로 볼 수 있으며, 궁극적으로는 하나의 카테고리로 분류하였다.

4.3.1.2 질투(JEALOUS)

'질투'는 대상에 대한 중오심을 표현하는 '혐오'보다는 약한 부정적 감정으로, 특정 대상을 부러워하는 감정을 나타낸다. 예시를 살펴보면 아래와 같다.

(19) ㄱ. 그 동생들이 부럽고 질투남.
ㄴ. 엄처리에 앉아서 대화 나눌 때도 조금 질투가 난 것 같아요.
ㄷ. 다른 팬이 된 이후로 그 친구에게 있던 질투가 더 커진 거 같아요.

3. 전처리 프로세스 개요

- 전체 흐름도:

  ① 데이터 수집 → ② 이상치 처리 → ③ 정규화 → ④ 데이터 변환 → ⑤ 데이터 분리

- 전처리 파이프라인 요약:

단계	목적	수행 작업	사용 도구/라이브러리
이상치 처리	노이즈 제거	PDF 내 무관 정보, 특수기호, 분류 모호 데이터 제거	Python
정규화	텍스트 전처리	특수문자, 숫자, 중복 공백 제거	re
데이터 변환	키워드 추출 및 라벨링	'...다' 형태 키워드 추출, 5가지 성격 유형 매핑	Python
데이터 분리	학습/검증 분할	train:test = 7:3	Scikit-learn

4. 세부 전처리 단계

4.1 결측치 처리

- 결측치 존재 여부: 없음
- 원본 데이터를 바탕으로 label과 keyword가 모두 존재하는 데이터만 추출하여 데이터셋을 직접 구성했으므로, 별도의 결측치가 존재하지 않음

4.2 이상치 처리

- 이상치 정의 기준:
  - PDF에서 추출된 페이지 번호, 참고문헌, 표 제목 등 무관한 정보
  - 여러 성격 유형에 걸쳐 중복으로 나타나거나, 특정 유형으로 분류하기 모호한 키워드 또는 문장

- 처리 방식 및 영향:

항목	기준	처리 방식	제거 수
lines	정규식 remove_pattern(제목, 페이지 수)에 일치하는지 여부	해당 패턴과 일치하는 행 제거	8건
full_text	section_title과 next_section_title 사이의 텍스트	지정된 섹션의 텍스트만 추출	1건
keywords	label_keywords 리스트 내 중복된 키워드	해당 행 제거	6건

- 이상치 처리 영향 : 정규식, 섹션 추출, 중복 키워드 제거 등의 이상치 처리를 통해 노이즈가 줄고 데이터 품질이 향상되어 모델의 분류 정확도와 효율성 개선에 기여함

4.3 정규화 및 표준화

- 텍스트 정규화:
    - 키워드 추출(토큰화)
        - 특수문자 제거: “날카롭다,” → “날카롭다”
        - 단어 분리: ‘현명하다공정하다심각하다’ → ‘현명하다’, ‘공정하다’, ‘심각하다’
- 사용 라이브러리: PyPDF2, re, json, scikit-learn

4.4 데이터 변환 및 생성

- 레이블 인코딩: PDF 원문에서 추출한 숫자 형태의 유형(예: "1번")을, 미리 정의된 type_to_label 딕셔너리를 사용하여 "추진형", "관계형" 등 실제 모델이 학습할 텍스트 라벨로 변환

```
type_to_label = {
    "1": "추진형",
    "2": "관계형",
    "3": "추진형",
    "4": "내면형",
    "5": "내면형",
    "6": "안정형",
    "7": "쾌락형",
    "8": "추진형",
    "9": "안정형",
}
```

- 데이터셋 생성: 추출된 키워드와 매핑된 라벨을 {"label": "라벨명", "keyword": "키워드"} 형식의

JSON 객체로 변환함

```
results = []
seen = set()
for kw in all_kws:
    if kw not in seen:
        results.append({"label": label, "keyword": kw})
        seen.add(kw)
```

- 데이터 증강: 학습 데이터의 양이 부족하다고 판단하여 감정 유형별로 예문을 추가 생성함. 유형별 포함된 에니어그램 수의 차이로 인해 발생하는 데이터 불균형을 해소하고자 예문 수에 차등을 둠

    - 목적: 감정 유형 간 학습 편향을 줄이고, 모델의 분류 성능 향상을 도모함
    - 방식
        - 내면형(4,5), 안정형(6,9) → 예문 2개 생성
        - 관계형(2), 쾌락형(7) → 예문 3개 생성
        - 추진형(1,3,8) → 예문 1개 생성

5. 학습/검증 데이터 분리

- 분리 기준 및 방법:

    - 기준: 무작위 분할

    - 비율: Train 70% / Test 30%

- 분리 코드:

```
from sklearn.model_selection import train_test_split, StratifiedKFold

train_df, eval_df = train_test_split(df, test_size=0.3, stratify=df['label'], random_state=42)
```

- 분리 후 건수:

구분	데이터 수
학습 데이터	468건
테스트 데이터	201건

6. 전처리 결과 요약 및 평가

- 전처리 후 전체 건수: 360건 → 354건
- 증강 후 학습에 사용된 전체 건수: 354건 → 669건

- 품질 향상 지표:

    - 노이즈 제거: PDF 파싱 과정에서 발생한 무관 정보(페이지 번호, 제목 등) 제외

    - 중복값 제거: 6건 제외

    - 레이블 정리 및 불균형 개선

- 향후 활용 방안:

- 정규화 및 표준화 등의 전처리 후 얻은 데이터셋은 향후 텍스트 분류, 대화 응답 분기 조건 설계, LLM 학습데이터 구축 등에 사용

7. 변경 이력

변경일	변경자	변경 내용	비고
2025-07-18	조성지	최초 작성	-
2025-07-18	남의헌	학습/검증 데이터 내용 추가	데이터 증강 추가 의문 제기
2025-07-20	이정민	최종 수정	-