

# 일 일 보 고 서

프로젝트	무신사 온라인 소비자 리뷰 기반 데이터 분석		
작 성 자	team3	일 시	2025-11-04

금 일 실 시 사 항	비 고
<p>[곽동원]</p> <ul style="list-style-type: none"><li>- 데이터 처리 과정 논의</li><li>- 수집한 사전학습 데이터, 파인튜닝 데이터 선택 기준 정의</li><li>- 사전학습 데이터 EDA</li><li>- 데이터 정제 규칙 확립 및 전처리</li><li>- 사전학습, 파인튜닝 데이터 전처리</li></ul> <p>[강지연]</p> <ul style="list-style-type: none"><li>- 데이터 처리 과정 논의</li><li>- 수집한 사전학습 데이터, 파인튜닝 데이터 선택 기준 정의</li><li>- 사전학습 데이터 EDA 및 병합</li><li>- 데이터 정제 규칙 확립 및 전처리</li><li>- 사전학습, 파인튜닝 데이터 전처리</li></ul> <p>[안호용]</p> <ul style="list-style-type: none"><li>- 데이터 처리 과정 논의</li><li>- 수집한 사전학습 데이터, 파인튜닝 데이터 선택 기준 정의</li><li>- 사전학습 데이터 EDA 및 병합</li><li>- 데이터 정제 규칙 확립 및 전처리</li><li>- 사전학습, 파인튜닝 데이터 전처리</li></ul>	

---

[정수아]

- 데이터 처리 과정 논의
- 수집한 사전학습 데이터, 파인튜닝 데이터 선택 기준 정의
- 사전학습 데이터 EDA 및 병합
- 데이터 정제 규칙 확립 및 전처리
- 사전학습, 파인튜닝 데이터 전처리

## 상세 실시 사항

1. 수집한 사전학습 데이터, 파인튜닝 데이터 선택 기준 정의
    - 1.1. 라벨이 있는 데이터와 없는 데이터로 분리
      - 사전학습 : 라벨이 없는 리뷰 데이터
      - 파인튜닝 : 라벨링 되어 있는 리뷰 데이터
  2. 나눠있는 사전학습 데이터 EDA 및 병합
    - 각 데이터 파일 형식이 다름
      - 하나의 데이터 파일(.txt)로 통일
    - 결측치, 중복값(이상치) 확인 및 제거
  3. 데이터 정제 규칙 확립 및 전처리
    - 3.1 Kc-Bert 텍스트 정제 규칙 참고
    - 3.2 학습 텍스트 정제 규칙 재정립
      - 한글(자/모음) 및 영어, 특수문자, 그리고 유니코드 이모지, 숫자 를 제외한 글자를 제거
      - soynlp를 사용해 댓글 내 중복 문자열 축약
      - 글자 단위 10글자 이하 제거
      - 중복 문장 제거
-

<p>4. 데이터 정제 규칙 기반, 사전학습, 파인튜닝 데이터 전처리</p> <p>5. 목적 파인튜닝 데이터 EDA</p> <p>5.1 kc-bert모델을 이용하여 목적 파인튜닝 데이터 분류</p> <ul style="list-style-type: none"> <li>- 다른 리뷰데이터로 kc-bert모델 파인튜닝하여 목적 파인튜닝 데이터 분류 특화 시도</li> <li>- kc-bert 모델 학습</li> </ul>	
명 일 예 정 사 항	비 고
<ul style="list-style-type: none"> <li>● kc-bert 모델 학습</li> <li>● kc-bert 모델 이용 목적 파인튜닝 데이터 분류 특화</li> <li>● DAPT(도메인 적응 사전 학습) 학습</li> <li>● 크롤링 데이터 재정의</li> </ul>	

---

이 슈 사 항

---

---