SK네트웍스 **Family AI** 과정 **13**기

데이터 전처리 인공지능 데이터 전처리 결과서

산출물 단계	데이터 전처리
평가 산출물	인공지능 데이터 전처리 결과서
제출 일자	2025.08.08
깃허브 경로	https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-3TEAM.git
작성 팀원	강지윤, 기원준, 우민규, 전진혁, 최호연

1. 문서 개요

- 프로젝트명: JJACKLETTE - 자동차 디자이너를 위한 프로토타입 생성 플랫폼
- 전처리 목적: 1) sLLM 모델 파인튜닝용 데이터 정제
  2) VectorDB, RDB 용 데이터 정제
- 문제 정의: 1) 차량 디자인 관련 지식과 기업의 목적 의식을 지닌 sLLM 모델을 만들기 위한 파인튜닝용 데이터셋 구축
  2) 세부 차량별 리뷰, 관련 뉴스 등을 검색하기 위한 RAG 용 데이터 전처리
  3) 차량별 분석 리포트 작성을 위한 RDB 용 데이터 구축

2. 데이터셋 개요

- 데이터 출처 및 수집 방법: 자체 수집
- 데이터 구성:
  - sLLM 파인튜닝용 데이터 예시
    - Hyundai_journal_articles.txt

Q. 디 올 뉴 넥쏘의 디자인 콘셉트와 핵심 키워드는 무엇인가?

설지훈 책임연구원 | 새로운 넥쏘는 단순히 스타일을 넘어 현대차의 수소전기차 기술에 담긴 '올곧은 신념'을 표현한 결과물이다. 아직 익숙지 않은 수소전기차 기술을 고객에게 친근하게 전달하기 위해 견고하고 신뢰감 있는 강성의 조형과 단단한 구조를 바탕으로 하나의 아이콘이 되는 스타일을 구현했다.

- 현대 모터스튜디오_디자인 관련 문서.pdf

- VectorDB용 데이터 예시
  - hyundai_car_history.json


```
[
  {
    "car_name": "포니(110) (1975~1985)",
    "year": "(1975~1985)",
    "explain": "포니는 대한민국 최초의 고유 모델 자동차이며, 이를 통하여 대한민국은 아시아에서는 일본에 이어 2번째로, 세계에서는 16번째로 고유 모델 자동차를 만든 국가가 되었다. 포니라는 차명은 당시 대한민국에서 공모를 통하여 지었는데, 수출 지향적인 부분도 반영이 되었다. 1974년 11월에 이탈리아에서 개최된 토리노 모터쇼에서 포니와 컨셉트 카인 포니 쿠페가 공개되었다. 포니 쿠페는 실제로 시판되지 못하여 컨셉트 카 단계에서 마무리되었다. 엔진은 당시 현대자동차의 기술 제휴 회사였던 미쓰비시 자동차의 1,238cc 새턴 엔진과 4단 수동변속기를 적용하였으며, 후륜구동 방식인 미쓰비시 랜서의 플랫폼을 기반으로 제작되었다. 1976년 1월 26일부터 판매가 시작되었으며, 가격은 2,289,200원 선에서 판매되었다.\n출처 : 위키피디아"
  },
  {
    "car_name": "포니2(110) (1982~1988)",
    "year": "(1982~1988)",
    "explain": "1982년 2월 19일에 페이스 리프트를 거친 포니 II가 출시되었다. 3도어 해치백과 5도어 스테이션 왜건은 판매가 부진하여 없어졌고, 5도어 해치백과 2도어 픽업 트럭만 생산되었다. 기존의 포니는 캐빈 룸과 트렁크 룸이 분리된 4도어 패스트백이였으나, 포니 II는 캐빈 룸과 트렁크 룸이 연결된 5도어 해치백으로 탈바꿈하였다. 현대자동차는 포니 II를 통하여 캐나다에 수출을 개시하였으며, 시속 5마일의 속도로 충돌해도 차체 손상을 막는 에너지 흡수형 범퍼가 달린 것이 특징이다. 캐나다 수출용은 1984년 5월 1일부터 대한민국에서 CX라는 트림으로 판매되었다. 1985년에 전륜구동 방식의 후속 차종인 포니 엑셀이 출시되었으나, 병행 판매되었다. 자가용은 1988년 4월에, 영업용은 1990년 1월에 단종되었다.\n출처 : 위키피디아"
  },
```


  - Hyundai_car_reviews.json


```
  {
    "data_id": "5167",
    "car_name": "더 뉴 투싼 Hybrid",
    "review": "구매를 하기전 타사 차량과 고민을 많이 했지만  선루프와 오디오  옵션을 제외한 나머지 옵션이 다 들어간  인스피레이션 트림으로 결정 하여 계약을 진행 하였고 친절하고 정확한 카마스터님의  케어로 차량이 우리 가족에게 오기까지 정말 즐겁고 신나는 날들 이었습니다.\n\n17년간 탔던  우리 가족 차를  보내고  리니를  만났던 순간 마치 전 부터 우리 가족인듯 하나도 어색 하지 않았어요. 딸아이는  발을 동동 구르며  너무 좋아 했고  몇일간 고민해서  우리 차가 된 투싼이 에게 리니라는 이름을 만들어 주었습니다.\n\n리니를 타는 순간 그동안 걱정했던 '가격을 너무 비싸게 준게 아닌가?' , '조금 더 큰 차량으로 해야 하는건 아닌가?', '너무 빨리 결정한 건 아닐까?'  등 이런 생각들은 깔끔하게 없어졌습니다. \n가벼운 터치로 전원이 들어오며 부드러운 스타트 정숙한 고속 주행에 \n자율 주행 모드  모든게  새롭고 편안했습니다. 와이프와 딸아이는 리니의 디자인에 반해 버렸고  저는 이 아이에 모든것에 반해 버렸습니다.\n\n저는 차에 대해 잘 몰라서 주행이 어쩌니 기능이 어쩌니 엔진이 어쩌니 잘 모르겠습니다. 하지만 20년간 무사고 운전한  노련한 운전자의 느낌으로  우리 가족에게는 리니가 아주 오랫동안 건강하게  잘 지낼것 같습니다. \n\n성함은 밝히기는 어렵지만 카마스터님께  감사 드립니다. 처음부터 끝까지 하나 하나 챙겨주시고 신경써주셔서  마지막에는 죄송한 마음까지 들었어요^^",
    "tags": {
      "성능": "편의기능이 다양해요",
      "공간": "많은 짐도 거뜬해요",
      "디자인": "매력적이에요",
      "승차감": "하차감이 즐거워요"
    }
  },
  {
```


  - other_issue_articles.txt

```
[1] [컨슈머인사이트] 기아 PV5, '레고' 닮은 혁신성에도 소비자 반응 미지근
o 기아의 다목적 전기차 'PV5'에 대한 소비자 반응이 뜨뜻미지근하다. 승용, 레저, 화물밴, 특장차
o 기아의 다목적 전기차 'PV5'에 대한 소비자 반응이 뜨뜻미지근하다. 승용, 레저, 화물밴, 특장차 등 확장성과 가격 이점을 두루 갖췄다는 평가에도 소비자 인지도, 관심도, 구입의향 모두 신차 평균 이하였다. '레고'처럼 조립 가능한 신개념 전기차라는 획기적인 콘셉트가 소비자에게 제대로 전달되지 않은 듯하다.
□ 자동차 리서치 전문기관 컨슈머인사이트는 2021년 11월 시작한 신차 소비자 초기 반응(AIMM : Auto Initial Market Monitoring) 조사에서 앞으로 2년 내 신차 구입의향이 있는 소비자(매주 500명)에게 출시 전후 1년 이내(출시 전, 출시 후 각각 6개월)의 국산·수입 신차 모델(페이스 리프트 제외)에 대한 인지도, 관심도, 구입의향 등을 묻고 있다. 구입의향은 '그 모델을 구입할 가능성이 얼마나 있습니까?'라는 4점 척도 문항에 '구입할 가능성 조금(3점)+많이(4점) 있다' 응답 비율이다.
■ 출시 전후 구입의향 상승 폭 크지 않아
o 7월 3주(14일 시작 주) 조사에서 'PV5' 구입의향은 7%였다[그림]. 2년 내 신차 구입 예정자 100명 중 7명 정도만 실제 구입을 고려한다는 뜻이다. 조사 대상 신차 13개 모델 중 9위로, 전체 평균 수준(7%)에 턱걸이했다. 관심도(8%)와 인지도(16%)는 각각 평균치(10%, 22%)에도 못 미쳤다. 특히 인지도의 평균 대비 열세가 점점 커지고 있는 것이 문제다.

o 출시 전후 구입의향 변동도 별로 없었다. 출시 주(6월 1주) 4%에서 조금씩 상승해 최근 3주 연속 7%를 기록했으나 출시 전에도 4~6% 사이에서 오르내렸던 것을 고려하면 박스권에 머물고 있다고 봐도 무방하다. 출시 전후 수주간 구입의향이 급상승했다 차츰 내려오는 일반적인 패턴에 비해 '조용한' 모습이다.
■ 기아 PBV 라인업의 첫 모델
o PV5는 기아가 처음 선보인 '목적 기반 모빌리티(PBV, Purpose Built Vehicle) 라인업의 첫 모델이다. 하나의 플랫폼에서 승합, 화물, 특장, 캠핑 등 용도에 따라 레고 조립하듯 변형 가능한 구조다. 경쟁 모델로 현대차 '스타리아'가 거론되긴 하지만 연료 계통부터 사용 목적, 소비자층까지 다른 새로운 개념의 차라는 평가가 타당성이 있어 보인다. 정부와 지자체 보조금을 받으면 2000만원대 중후반(카고 모델), 또는 3000만원대 중후반(패신저)에 구입 가능한 점도 매력이다.
```

- RDB용 데이터 예시
  - Hyundai_car_reviews.json


```
},
{
  "data_id": "5114",
  "car_name": "디 올 뉴 팰리세이드",
  "review": "기존 투싼과의 9년 동행을 마무리하며 새로운 차량을 고민하던 중에 신형 팰리세이드가 출시된다는 정보를 처음에 접하고 \n많은 기대를 했었지만
  가격이 부담스러워 포기하려던 순간에 최고급사양 및 옵션만 쳐다보다가 관점을 바꾸어 익스클루시브 트림을 \n자세히 살펴보니 사양이 개인적으로 너무 만족스러워
  옵션만 일부 추가, 구매를 하게 되었고 지금은 운행하면서 만족하는 중입니다.  \n\n기본 사양이지만 예전에 운행하던 차량에서는 볼 수 없었던 편리한 기능,
  안전한 부분들이 매우 많습니다.\n기본 안전사양과 왠만한 편의 이 은 다 들어가 있고, 1, 2열 2중 접합유리, 1열 통풍시트, 터널에 진입시 창문이 자동으로
  닫히고, 내기 순환모드로 전환, 탑승 전 미리 휴대폰을 통해 환기 가능, 시동제어 가능 등등 기존에는 경험해보지 못했던 것들을 새로운 차량을 통해 경험해보면서
  \n운행하고 싶은 마음이 저절로 생기고 있습니다. \n\n새로운 기술이 적용된 하이브리드 파워트레인을 선택하지 않은 부분이 아쉽게 느껴지지만(가격 상승을
  무시할 수 없었습니다)\n현재 약 1300km를 운행하면서 가솔린 파워트레인도 만족하는 중입니다. \n정숙성에 있어서도 하브에 비교해 뒤쳐진다고 느껴지지
  않습니다. 차량은 매우 조용하고 운전하기 편안하며 엔진소리도 개인적으로는 전혀\n귀에 거슬리지 않습니다. 또한 연비도 나쁘지 않습니다. 개인차가 있겠지만
  저같은 경우, 고속도로와 시내도로가 혼합된 출퇴근길을 운행하는데 12~14km/L를 기록하고 있고 얼마전 홍천으로 가족여행을 다녀왔을 때는 15.4km를 기록하기도
  했습니다. \n넓은 실내공간 덕분에 가족들이 더 좋아합니다. 이제 캠핑갈 때 스트레스 받지 않습니다. 짐도 가득 넣을 수 있고 탑승 공간도 넉넉합니다.
  \n\n웅장한 외관, 고급스러운 실내 공간과 운행의 편안함 등 앞으로도 만족스럽게 오랜기간 운행해보려 합니다. \n트림과 옵션을 고민 하시는 분들에게
  익스클루시브 트림이 매우 합리적인 선택이 될 수 있을 것이라고 생각합니다.",
  "tags": {
    "성능": "편의기능이 다양해요",
    "공간": "많은 짐도 거뜬해요",
    "디자인": "마음에 쏙 하차감",
    "승차감": "운전이 즐거워요"
  }
},
```


  - 그랜저.csv


```
항목,프리미엄 (A/T),익스클루시브 (A/T),아너스 (A/T),캘리그래피 (A/T),캘리그래피 (블랙익스테리어) (A/T),캘리그래피 (블랙잉크) (A/T)
전장,"5,035 mm","5,035 mm","5,035 mm","5,035 mm","5,035 mm","5,035 mm"
전폭,"1,880 mm","1,880 mm","1,880 mm","1,880 mm","1,880 mm","1,880 mm"
전고,"1,460 mm","1,460 mm","1,460 mm","1,460 mm","1,460 mm","1,460 mm"
축거,"2,895 mm","2,895 mm","2,895 mm","2,895 mm","2,895 mm","2,895 mm"
윤거 (전),"1,628 mm","1,628 mm","1,624 mm","1,624 mm","1,624 mm","1,624 mm"
윤거 (후),"1,635 mm","1,635 mm","1,631 mm","1,631 mm","1,631 mm","1,631 mm"
승차정원,5,5,5,5,5,5
공차중량,"1,700 kg","1,700 kg","1,715 kg","1,715 kg","1,715 kg","1,735 kg"
연료탱크,50 ℓ,50 ℓ,50 ℓ,50 ℓ,50 ℓ,50 ℓ
트렁크 (후) 용량,480 ℓ,480 ℓ,480 ℓ,480 ℓ,480 ℓ,480 ℓ

```

3. 전처리 프로세스 개요

- 전체 흐름도:

  ① 수집 → ② 텍스트 추출 → ③ 세부 내용별 Chunking → ④ JSON 파싱

- 전처리 파이프라인 요약:

단계	목적	수행 작업	사용 도구 / 라이브러리
텍스트 추출	Raw text 변환	텍스트 추출 및 불필요한 부분 제거	Pdfloader, textloader
Chunking	개별 데이터 분리	리뷰별 / 기사별 / 주제별 분리	OpenAI, TextSplitter
JSON 파싱	텍스트 전처리	목적에 따라 QA / Document 형식	OpenAI

4. 세부 전처리 단계

- 데이터별 전처리 방법

파일명	파일 유형	**Chunk** 기준	**JSON** 파싱 형식
현대 모터스튜디오_디자인 관련 문서.pdf	.pdf	단일 질문별	1️⃣
현대자동차 디자인 철학에 내재하는 미의식의 신경학적 해석.pdf	.pdf	단일 질문별	1️⃣
자동차 개발 단계에서의 인간공학의 역할.pdf	.pdf	세부 주제별	2️⃣
자동차 차체 형태 및 디자인이 공기역학 성능에 미치는 영향에 대한 연구.pdf	.pdf	단일 질문별	1️⃣
Hyundai Motor Company Identity Design Guide Book.pdf	.pdf	단일 질문별	1️⃣
현대 디자인 모토.txt	.txt	단일 질문별	1️⃣
차체 및 구조 설계의 모든 것.txt	.txt	단일 질문별	1️⃣
new_articles.txt	.txt	Article 별	1️⃣
preview_articles.txt	.txt	Article 별	1️⃣
total_articles.txt	.txt	Article 별	1️⃣
other_issue_articles.txt	.txt	Article 별	2️⃣
other_new_articles.txt	.txt	Article 별	2️⃣
other_preview_articles.txt	.txt	Article 별	2️⃣

other_total_articles.txt	.txt	Article 별	2
interview_articles.txt	.txt	Article 별	1
hyundai_journal_articles.txt	.txt	Article 별	1
hyundai_car_history.json	.json	개별 뉴스별	2
hyundai_car_reviews.json	.json	개별 리뷰별	2
Car_specs.zip (.csv)	.csv	-	3

- 데이터 용도별 JSON 파싱 형식
  - 1: sLLM 파인튜닝용 데이터
    - { {“Q”:”~~~~~~~”, “A”:”~~~~~~~”}, …. }
  - 2: RAG VectorDB용 데이터
    - { {“page_content”:”~~~~~~~”,
      “Metadata”: {“title”:”~~~~~~~”,
      “Car_name”:”~~~~~~~”,
      “Category”:”~~~~~~~”,
      “Tags”:[“~~~~~~~”, “~~~~~~~”, …],
      “Brand”:”~~~~~~~”
      }
      }, …. }
  - 3: RDB 리포트 분석용 데이터
    - 작성 필요

5. 전처리 후 데이터 요약

- sLLM 파인튜닝용 데이터 분리 기준 및 방법:
  - 기준: 무작위 분할
  - 비율: Train 80% / Valid 10% / Test 10%
- 분리 코드:

```
from sklearn.model_selection import train_test_split

# 전체 데이터 로드
df = load_all_qa_data(QA_CONTEXT_DIR)

# Train/Test 분할
train_valid, test = train_test_split(df, test_size=0.1, random_state=42)

# Train/Validation 분할
train, valid = train_test_split(train_valid, test_size=0.1111, random_state=42)  # 0.1111 ≈ 10% of 90%
```

- 분리 후 건수:

구분	데이터 수
학습 데이터	266건
검증 데이터	50건

테스트 데이터	40건

- 전처리 후 전체 건수:
  - sLLM 파인튜닝용 데이터 : 266 (train) + 40 (test) + 50 (validation) = 356건
  - RAG VectorDB용 데이터 : 602건
  - RDB 리포트 분석용 데이터 : 500(review) + 33(csv) = 533건