**프로젝트 기획안**

기획안 작성일자 : 2025.05.27 / 기획안 수정일자 : 2025.07.15

| 조 명 | 3조 : 데이터 맛집 |
| --- | --- |
| 조 원 | 조장 : 황진웅 / 조원 : 김철기, 박건하, 성원선, 홍석우(퇴소) |
| 프로젝트 주제 및 개요 | 이커머스 판매 데이터 분석을 통한 수요 예측 및 추천 시스템 구축  - 마켓컬리 리뷰 분석 기반 자동 평점 및 맞춤형 상품 추천 시스템 |
| 프로젝트 수행 방향 | ◈ 프로젝트 수행 계획  1. 문제 정의  1.1. 마켓컬리 내 정량적 평점 시스템 부재로 인한 수요 예측의 제약  → 리뷰 텍스트 기반의 **자동 평점 산출 시스템** 도입  (예상화면)    2.1. 고객 맞춤형 정보 제공이 제한적이어 상품 선택에 대한 피로가 증가  → 리뷰 감정분석 및 외부데이터를 활용한 **맞춤형 상품 추천 시스템** 구축  (예상화면)    2. 데이터 수집  2.1. 데이터 목록(공통)  2.1.1. 웹 크롤링  · 마켓컬리 리뷰 데이터  · 11번가, 옥션 등 외부 사이트의 평점 데이터, 리뷰 데이터 이용  · 네이버 지도 리뷰와 해시 태그  2.1.2. API 및 다운로드  · 카카오맵 부정 리뷰 데이터  · 통계청 - 간편식 데이터  · 롯데카드 배달음식 빅데이터(판매량, 재구매율)  · GPT API - 질문 문장 데이터셋 생성  · Kaggle Dataset(Food.com Recipes and Interactions) - 해외 데이터  · 라이브커머스 판매량 데이터  · 데이터랩 검색량 데이터  · 기상청 API - 날씨 데이터  2.2. 데이터 활용  · **자동 평점 산출 시스템**  - 리뷰 데이터  - 카카오맵 부정 리뷰 데이터  · **키워드 자동생성 시스템**  - 리뷰 데이터  - 네이버 지도 리뷰와 해시 태그  · **맞춤형 상품 추천 시스템**  - 통계청의 간편식 데이터  - 롯데카드 배달음식 빅데이터  - 생성된 질문 문장 데이터셋  - Kaggle Dataset  - 라이브커머스 판매량 데이터  - 데이터랩 검색량 데이터  - 날씨 데이터  3. 전처리  3.1. **자동 평점 산출 시스템**  · 텍스트 정제 – 감정 분석 정확도 향상(이모지, 특수문자 삭제)  · 데이터 추가 수집 - 카카오맵 부정 리뷰 데이터(긍·부정 리뷰 비율)  3.2. **키워드 자동생성 시스템**  · 사전 생성  · keyword map 생성  3.3. **맞춤형 상품 추천 시스템**  · 텍스트 정제 – 의도 분류, 특징 추출 정확도 향상  · 문장 생성 - 통계청 간편식 데이터와 롯데카드 배달음식 빅데이터를  분석한 결과 기반으로 질문과 답변 문장 생성  4. 분석 및 시각화  4.1. **자동 평점 산출 시스템**  · TF-IDF를 이용해 외부 사이트 평점 데이터의 평점별 빈출어 추출  · 수집한 외부 사이트의 평점 데이터에서 상품별 평점 분포를 시각화  · 수집한 11번가, 옥션의 데이터와 카카오맵의 7만개 부정데이터를 합침    → 편향되어있던 분포가 어느정도 비율을 맞춤  4.2. **키워드 자동생성 시스템**  · 형태소별 빈도수 집계  · 리뷰에서 자주 함께 등장하는 단어 조합과 이들의 연관 규칙 분석  · 상위 주요 키워드 확인  4.3. **맞춤형 상품 추천 시스템**  · 키워드변화 탐색-계절/날씨별 자주 등장한 키워드 비교(워드클라우드)  · 계절/날씨에 따라 자주 등장하는 키워드 변화 추이  · 계절/날씨별 긍정 리뷰 수 비교 – 상품별 계절성 수요 분석  · 유사 상품 간 감정 비교(예: 상품 A vs B) – 상품 경쟁력 비교  · 연령대, 지역, 성별, 가구원수의 배달음식과 간편식의 판매량과 판매금액  그리고 재구매율  5. 1차 모델링  : (Model Prototyping /Baseline Model 구현)  5.1. **자동 평점 산출 시스템**  · BERT/KoBERT – 리뷰 감정 분석 → 긍정/부정 감정 점수 산출 및 분류  · 회귀 모델(XGBoost, LGBM) → 감정 점수와 외부 사이트 평점 기반 학습  5.2. **키워드 자동생성 시스템**  · KoBERT/KcBERT - 키워드 추출  · SBERT - 키워드 추출  5.3. **맞춤형 상품 추천 시스템**  · BERT/KoBERT – 문장 분석 → 특징 추출  · 시계열 모델(LSTM, GRU) → 날씨/시간 관련 데이터  6. 유니크 포인트 추가  6.1. **자동 평점 산출 시스템**의 확장  · 해시태그-키워드 자동생성 기능  → 사용자 후기 탐색 피로도 감소 기능 추가  → Logistic Regression, KcBERT 이용  6.2. **맞춤형 상품 추천 시스템**의 확장  · 다양한 조건을 고려한 정밀 추천 기능  → 시간, 날씨, 감정 점수, 상품 속성을 고려하여 현재 가장 적합한 상품 추천(LightFM, XGBoost Ranker)  · 사용자 취향에 맞춘 맞춤 추천 기능  → 리뷰를 기반으로 구매자의 취향을 유형화하여 유사한 반응을 보인 고객에게 맞춤형 상품을 추천(K-Means, GMM 등)  · 챗봇 기능(BERT, KcBERT, Transformer)  → 고객의 현재 상황(감정 등)에 맞는 음식 추천  → Kaggle의 emotion dataset으로 학습  → 대화 로직  → 문장 생성에 대한 특징 설정(질문과 답변)    7. 2차 모델링  7.1. **자동 평점 산출 시스템**  7.1.1 감정 분석 기반 평점 예측 모델  · 감정 분석 프로세스(KoBERT fine-tuning)  → 리뷰 텍스트로부터 감정 점수(0 또는 1 범위의 연속값) 산출하여 회귀모델 입력으로 사용   | review\_id | 리뷰 텍스트 | 감정 점수 (KoBERT) | 외부평점 (1~5) | 리뷰 길이 | 주요 키워드 | | --- | --- | --- | --- | --- | --- | | R001 | 배송 빠르고 신선해요 | 0.91 | 5.0 | 20자 | 배송, 신선 | | R002 | 기대보다 별로였어요 | 0.28 | 2.0 | 15자 | 실망, 맛 | | R003 | 괜찮았지만 재구매는 안 할 듯 | 0.55 | 3.0 | 22자 | 무난 |   7.1.1.1. 감정점수 산출  · KoBERT를 활용해 리뷰 텍스트를 감정 점수(0~1 사이 연속값)로 변환  · 불용어 제거 및 정제된 문장 처리로 길이가 짧은 리뷰 제외  예시) 배송이 빨라서 좋았어요 → 0.89 (긍정)  맛이 기대 이하입니다 → 0.32 (부정)  7.1.2 회귀 기반 평점 예측 모델  · 감정 점수(KoBERT), 리뷰 길이, TF-IDF/CountVectorizer 결과, 외부 평점 등을 입력하여 1~5 사이의 실수형 평점 예측   | item\_id | 감정 점수 | TF-IDF 유사도 | 리뷰 길이 | 외부 평균 평점 | 예측 평점 (y) | | --- | --- | --- | --- | --- | --- | | 101 | 0.92 | 0.35 | 85 | 4.7 | 4.8 | | 102 | 0.30 | 0.12 | 42 | 3.1 | 2.9 | | 103 | 0.73 | 0.29 | 63 | 4.0 | 4.2 |   7.1.2.1. 학습 및 평가 전략  · Train/Valid/Test 비율 = 8:1:1  · MAE, RMSE를 평가 지표 성능 분석  · 예측 오차가 큰 리뷰는 별도 수집하여 정성 분석 수행  7.1.2.2. 예측 결과 활용 방안  · 상품 상세 페이지에 자동 평점 표시하여 리뷰 수가 부족한 신상품의 초기 평점 보완.  · 마켓컬리 운영팀의 개선/검토용 지표로도 활용 가능  7.1.2. 분류와 회귀 비교  · 모델  - LSTM  - BiLSTM + CNN  - Roberta + BiLSTM  7.2. **키워드 자동생성 시스템**  7.2.1. 해시태그-키워드 자동생성 기능  · 리뷰 감정 분석(KcBERT/KoBERT)  → 키워드 : ‘가성비’, ‘가성비가 좋아요’, ‘배송이 빨라요’ 등  · 키워드별 리뷰 수 계산(XGBoost, LightGBM, CatBoost)  7.2.2. 키워드/해시태그 추출 모델 적용  · KeyBERT 및 SentenceTransformer 기반 핵심 키워드 추출  · 리뷰별 핵심 키워드 1~3개 자동 추출  7.2.3. 불용어 사전 적용  · 한국어 불용어 직접 수집 및 커스텀 리스트 적용  7.2.4. 자동 해시태그 생성 로직 개발  · 사전 기반 자동 태깅, 키워드 매칭 방식  7.3. **맞춤형 상품 추천 시스템**  7.3.1. 다양한 조건을 조합한 상품 추천 및 우선순위 표시 기능  7.3.1.1. 상호작용 기반  · LightFM 구성 예시   | 리뷰 ID | 상품명 | 감정 분석 결과 | 상호작용 값 | | --- | --- | --- | --- | | User\_1 | 상품A | 긍정 | 1 | | User\_2 | 상품B | 부정 | 0 | | User\_3 | 상품A | 긍정 | 1 |   · 리뷰 감정 분석 설계   | 문제 | 해결 방법 | | --- | --- | | 사용자 ID 없음 | 리뷰 단위로 가짜 유저 ID 생성  (user\_1, user\_2...) | | 클릭/구매 없음 | 리뷰 감정 분석 결과로 상호작용 구성  (1=긍정, 0=부정) | | 추천 방식 | 감정 기반 콘텐츠 추천 or 유사 아이템 추천으로 활용 |   → 각 리뷰 하나 = 하나의 사용자로 간주  → 상호작용은 감정 분석 결과로 대체 (긍정/부정 → 1/0)  7.3.1.2. 시간대, 날씨, 계절 등 추가 조건  · 상품 feature에 따른 정보를 메타데이터로 추가  (벡터화해서 LightFM 모델의 인자로 학습)  예시) 상품 카테고리 → 과일, 야식  날씨 태그 → 더움, 추움  계절 태그 → 여름, 겨울  감정 키워드 → 시원한, 따뜻한  시간대 → 새벽, 점심  · XGBoost Ranker 구성 예시   | item\_id | 감성 점수 | 회원등급 점수 | 리뷰 길이 | 긍/부정 라벨 (y) | | --- | --- | --- | --- | --- | | 101 | 0.92 | 2 | 85 | 1 | | 102 | 0.23 | 0 | 40 | 0 | | 103 | 0.75 | 1 | 63 | 1 | | 104 | 0.33 | 0 | 38 | 0 |   예측값 기반으로 추천 우선순위 정렬 가능  → 추천 순서: 101 > 103 > 102 > 104  7.3.1.3. 챗봇  · 문장 생성 - GPT API 이용  → 전체 6만개의 문장  → 특징 조합에 따른 문장 생성  → 생성된 문장과 문장 생성에 사용된 특징 그리고 타겟값 저장  → 타겟값으로 사용되는 열에 결측치가 있는지 확인  · 코드 연결  → 질문 데이터와 타겟값을 학습(KcBERT/BERT)  → 특징 추출(BERT 모델 또는 커스텀 함수)  → 답변 데이터와 연결  8. 결과 발표  · 웹 대시보드 이용  · 시각자료 준비(시연 영상, streamlit 등)  · ppt 발표 자료  ◈ 학습 모델  · 크롤링 : Selenium, BeautifulSoap  · 챗봇 : KoBERT, Transformer, KcBERT, CNN  · 시계열 : RNN, LSTM, GRU, CNN  · 키워드 : CountVectorizer, Word Frequency, SBERT  · 감정 분석 : XGBoost, Light GBM, KoBERT, CatBoost, KcBERT  · 다양한 조건을 조합해서 상품을 추천 : LightFM, CountVectorizer  · 상품의 우선순위를 학습 : XGBoost Ranker |
| 프로젝트 조직  (구성원 및 역할) | ◈ 공통 진행 업무  · 마켓컬리, 외부 사이트 리뷰 데이터 크롤링  · 데이터 전처리  · 모델 학습 및 설계 실험  · 진행 상황에 대한 피드백  ◈ 담당 프로젝트  · 자동 평점 시스템 : 김철기  · 키워드 자동생성 시스템 : 박건하  · 맞춤 상품 추천 시스템 : 성원선, 황진웅, 홍석우(퇴소)  ◈ 역할  조장) 황진웅 : 프로젝트 전반 관리 / 추천 모델 설계 및 문장 생성  조원1 김철기 : 데이터 관리 담당 / 자동 평점 부여 및 챗봇 구현  조원2 박건하 : 프로젝트 진행 관리 담당 / 키워드 자동 생성  조원3 성원선 : 회의 내용 기록 담당 / 챗봇 구현 및 대시 보드 설계  (퇴소) 홍석우 : 수행 일지 및 문서 담당 / 추천 모델 설계 |
| 프로젝트 추진 일정 | ◈ 일정  5/14 ~ 5/23 : 주제 선정  5/26 ~ 5/30 : 주제 선정 및 데이터 수집  6/02 ~ 6/06 : 데이터 수집 및 전처리 작업, 분석 및 시각화, 1차 모델링  6/09 ~ 6/13 : 모델 개선 및 파라미터 튜닝과 분석, 데이터 수집  6/16 ~ 6/20 : 2차 모델링  6/23 ~ 6/27 : 모델 개선 및 파라미터 튜닝  6/30 ~ 7/11 : 최종 모델링  7/14 ~ 7/15 : 챗봇 구현 및 발표 자료 작성 |