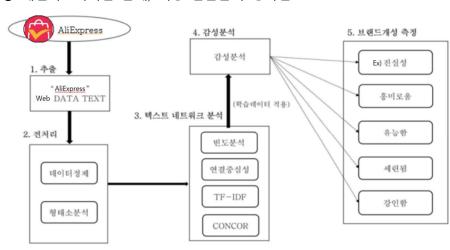
주간프로젝트 기획안

기획안 작성일자 : 2024년 5월 24일

조 명	1 조:TETO						
조 원	조장: 임태수 조원: 예동완, 남수형, 남학균, 김 원						
	주제: 알리익스프레스 내에 영향력있는 제품군의 가격과 평점, 리뷰들을 분석하여 이를 바탕으로 국내 사업자들에게 경쟁력 있는 가격대, 경쟁력 있는 제품의 방향설정에 도움을 줄 수 있는 데이터 분석 개요:						
프로젝트 주제 및 개요	 배경: 중국 온라인 직구의 한국 시장 진출 후 오픈마켓 플랫폼의 지각변동, 한국의 중소기업들이 우후죽순 밀려나가는 상황 목적: 알리의 잘팔리는 제품군의 가격대와 리뷰를 조사하여 해당 시장(가성비 선호)의 적정 가격대, 긍정 부정적 소비자 리뷰를 파악하여 국내 브랜드의 보완 점을 마련 타겟: 국내 오픈마켓 및 해외 오픈마켓을 유통채널로 쓰고 있는 국내 셀러및 제조업자 기대효과 알리익스프레스의 가격 경쟁력을 이길 수는 없으나 본 프로젝트의 데이터분석자료를 통해 적정 가격을 산출할 수 있게 된다. 파악된 알리익스프레스의 고질적인 문제(리뷰를 통해 분석)들을 본 프로젝트의데이터분석자료를 통해 파악하여 전략의 선택 집중을 할 수 있게 된다. 						
	● 해결하고자하는 문제, 최종 산출물의 청사진						
	AliFynness A zhalika E HallE-dad 3-28						

프로젝트 수행 방향 및 내용 (수행 방법/도구)



1. 데이터 수집:

알리익스프레스의 베스트셀러 각 파트의 Top 50 제품 웹 크롤링 df = pd.DataFrame(("상품명": [], '가격': [], '평점': [], '댓글': [], '별점': [], '링크': []))

2. 전처리:

데이터 정제 / 형태소분석, 데이터 편집 과정을 거쳐 데이터 전처리를 수행

키워드 필터링과 중복제거, (형태소 분석에서는 분석 언어를 한국어로 설정) 분석 품사로는 단순 품사 중 명사와 형용사를 선택 Positive
(e.g. adorable, great, awesome no regret, no complaint)
Negative

(e.g. junk,bad not good, no worth)

[불용 단어 제거 / 긍정/부정 단어 추출 사전 선정]

3. 분석 : 빈도 분석, TF-IDF, 연결중심성 네트워크 분석, LDA 토픽 모델링

$$TF - IDF = TF \times \frac{1}{DF}$$

TF = 문서 내 특정단어의 빈도수

DF = 여러문서내의 특정단어 빈도수

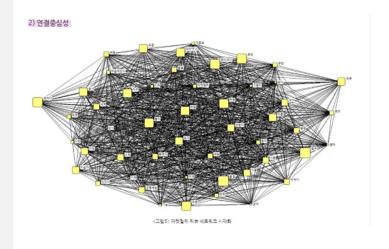
IDF = DF의 역수

4. 시각화

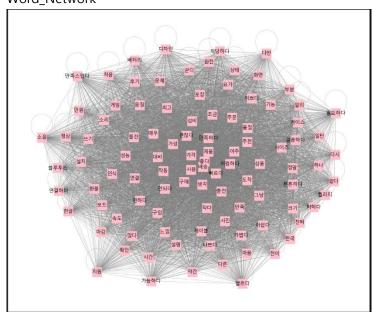


단어	빈도	연결중심성	TF-IDF	단어	빈도	연결중심성	TF-IDF
좋다	1680	0.115	1373.846	장바구니	131	0.021	442.732
식품	851	0.072	1243.516	자주	125	0.015	409.409
배송	520	0.047	957.165	기능	109	0.012	383.56
편리하다	491	0.04	911.175	불편	108	0.017	361.808
사용	485	0.047	957.165	감사	107	0.013	359.502
마켓컬리	433	0.056	936.041	문앞	104	0.015	354.65
주문	302	0.041	757.519	결제	97	0.017	357.032
안된다	301	0.036	727.354	편하다	91	0.006	317.134
없다	290	0.044	710.91	생각	91	0.044	320.219
빠르다	289	0.019	677.097	아침	86	0.012	307.715
많다	271	0.033	664.333	맛있다	85	0.012	307.305
같다	246	0.034	645.384	검색	85	0.015	321.309
만족	243	0.024	619.913	음식	77	0.012	286.604
어플	223	0.034	602.457	새벽	76	0.009	282.881
정말	221	0.024	600.445	마트	74	0.012	281.971
샛별배송	211	0.02	571.111	요즘	71	0.009	276.223
품질	200	0.016	553.974	그렇다	71	0.015	271.64
다양	193	0.02	528.391	마음에듬	70	0.009	263.571
구매	184	0.029	535.372	아쉽다	70	0.01	268.917
포장	167	0.02	494.474	오뜌	66	0.012	263.724
물품	164	0.02	486.674	배달	65	0.009	251.805
쇼핑	150	0.018	452.249	문의	59	0.012	243.945
가격	144	0.021	442.389	비싸다	59	0.009	235.753
쿠폰	142	0.019	452.84	된요	58	0.01	228.622
리뷰	136	0.016	446,668	다음날	57	0.009	224.68

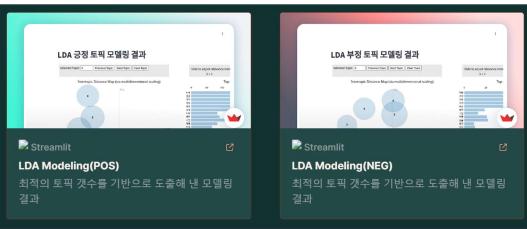
〈표4〉마켓컬리 리뷰 텍스트 분석 결과 (상위50위)



$Word_Network$



LDA



5. 대시보드 화면

- Gamma Ai
- Streamlit
- LOOM (웹크롤링 화면 녹화)

● 데이터소개

- AliExpress Web_Crawling (제품:290 개, 댓글: 29,000 개)
- KOSIS 국가통계포탈 '해외직접구매액 조사'

● 수행도구

- Pandas, Numpy, Matplotlib, Sklearn, Konlpy_Okt, Korean_Sentiment Model from Hugging Face, Bert Model from Google, Java Script, BS4, Selenium, Github, Streamlit, Gamma Ai, Powerpoint, Sklearn - LogisticRegression, Sklearn_GridSearchCV

프로젝트

조직 (구성원

및 역할)

● 역할분담

팀 장: 임태수 - 데이터 수집 및 전처리, 모델분석, 시각화 부 팀장: 예동완 - 데이터 수집 및 전처리, 모델분석, 시각화

조원: 김원 - 데이터 수집 및 전처리, 모델분석, 시각화

조원: 남수형 - 데이터 수집 및 전처리, 모델분석, 시각화 조원: 남학균 - 데이터 수집 및 전처리, 모델분석, 시각화

● 일정

5/11 ~18: 대 주제 선정주제 선정 및 일정 수립

- 과제 주제 후보 별 귀무가설, 대립가설 설정

- 사전 학습 (웹크롤링, 텍스트마이닝 이론)

프로젝트

추진 일정 5/18 ~ 5/24: Raw Dataset 구축

- 알리익스프레스 카테고리별 베스트 50 상품 웹 크롤링
- 알리익스프레스 베스트 50 상품의 댓글, 별점별 숫자 크롤링

5/22 ~ 5/27: 데이터 전처리

5/27 ~ 6/05: EDA, 시각화

- 리뷰에 대한 긍정/부정 별 지수 산출
- 가격, 별점, 리뷰 시각화
- 빈도 분석, TF-IDF, CNN 분류 모델로 긍정/부정 분류