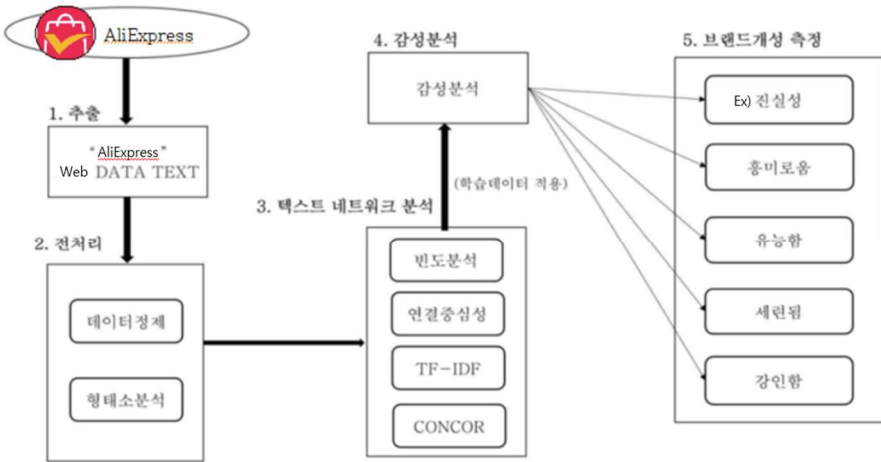


주간프로젝트 기획안

기획안 작성일자 : 2024년 5월 24일

조 명	1 조 : TETO
조 원	조장: 임태수 조원: 예동완, 남수형, 남학균, 김 원
프로젝트 주제 및 개요	<p>주제 :</p> <p>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</p> <p>개요 :</p> <ul style="list-style-type: none"> - 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 셀러 및 제조업자 - 기대효과 - 알리익스프레스의 가격 경쟁력을 이길 수는 없으나 본 프로젝트의 데이터 분석자료를 통해 적정 가격을 산출할 수 있게 된다. - 파악된 알리익스프레스의 고질적인 문제(리뷰를 통해 분석)들을 본 프로젝트의 데이터 분석자료를 통해 파악하여 전략의 선택 집중을 할 수 있게 된다.
프로젝트 수행 방향 및 내용 (수행 방법/도구)	<p>● 해결하고자하는 문제, 최종 산출물의 청사진</p> 

1. 데이터 수집:

알리익스프레스의 베스트셀러 각 파트의 Top 50 제품 웹 크롤링

```
df = pd.DataFrame({"상품명": [], '가격': [], '평점': [], '댓글': [], '별점': [], '링크': []})
```

2. 전처리:

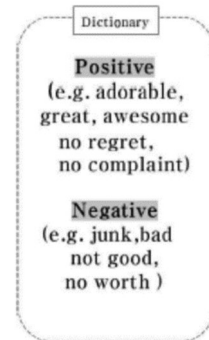
데이터 정제 / 형태소분석,

데이터 편집 과정을 거쳐 데이터 전처리를 수행

키워드 필터링과 중복제거,

(형태소 분석에서는 분석 언어를 한국어로 설정)

분석 품사로는 단순 품사 중 명사와 형용사를 선택



[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 = 문서내 특정단어의 빈도수

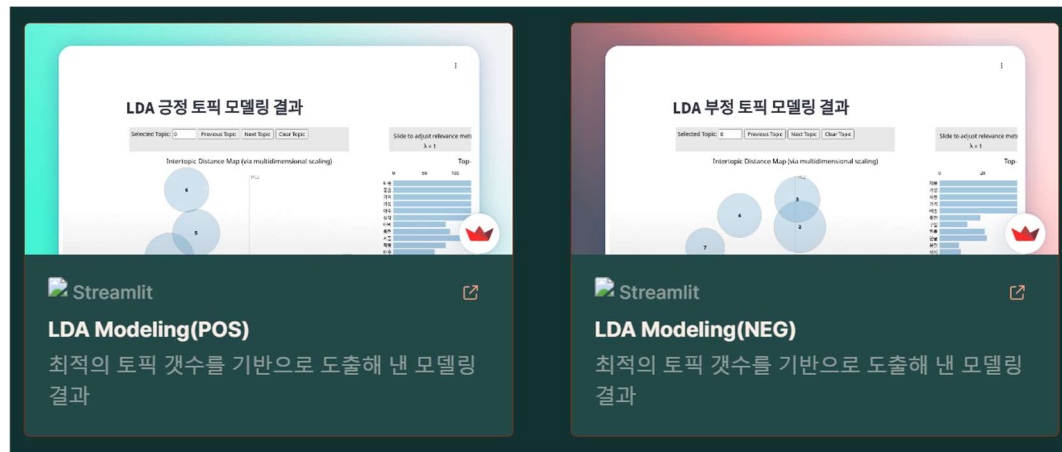
DF = 여러문서내의 특정단어 빈도수

IDF = DF의 역수

4. 시각화



LDA



5. 대시보드 화면

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

● 데이터소개

- AliExpress Web_Crawling (제품:290 개, 댓글: 29,000 개)
- KOSIS 국가통계포털 '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

프로젝트 추진 일정

● 일정

5/11 ~18: **대 주제 선정주제 선정 및 일정 수립**

- 과제 주제 후보 별 귀무가설, 대립가설 설정
- 사전 학습 (웹크롤링, 텍스트마이닝 이론)

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
-