2022-Fall BKMS1 Final Project Report

백개의 리뷰: 전자제품 리뷰 총정리 서비스

Team 09
김정국 (데이터사이언스학과)
신성구 (데이터사이언스학과)
이준규 (데이터사이언스학과)
조유진 (데이터사이언스학과)
허상우 (데이터사이언스학과)

Motivation / Background

전자기기는 결코 저렴하지 않다. 더구나 한번 사면 수개월에서 수년을 함께해야 한다. 그래서 노트북, 스마트폰 등 전자제품을 사겠다고 마음 먹으면, 우리는 다른 어떤 제품을 소비할 때보다 신중하게 된다. 무게나 색상, 디자인은 적당한지, 다른 제품과 비교하면 어떤 강점이 있는지 꼼꼼히 따져보게 된다. 그런 점에서 먼저 제품을 구매해 실사용 후기를 남긴 타인의 리뷰들은 제품 구매를 고민 중인 소비자에게 가치가 매우 큰 데이터베이스다.

하지만 리뷰 데이터는 유통 플랫폼별로 뿔뿔이 흩어져 있다. 우리는 온라인 구매 최저가를 비교하기 위해 네이버 쇼핑에서 제품을 검색해보고, 비교적 양질의 리뷰 데이터가 많이 축적돼 있는 쿠팡에서 또 다시 제품을 검색해 상품후기나 리뷰를 확인해야 한다. 이 번거로움이 '백개의 리뷰' 서비스가 파고 들고자 하는 페인 포인트다.

Scenario

- Requirements / Assumptions

'백개의 리뷰'는 여러 쇼핑몰, 리뷰 사이트에서 동일 제품에 대한 리뷰들을 수집해 유저가 한 화면에서 다양한 방식으로 리뷰를 비교 조회해볼 수 있도록 한다. 현실적으로 존재하는 모든 쇼핑몰과 리뷰 사이트의 데이터를 수집하기는 어려워 양질의 리뷰가 대량으로 축적돼 있고 데이터 크롤링이 비교적 수월한 네이버쇼핑몰, 쿠팡, 꿀리뷰 세 채널로 한정하고자 한다.

유저가 리뷰를 통해 보고 싶은 정보는 여러 가지가 있겠지만 본 서비스에서는 유저가 특정 제품을 검색하면 그에 대한 긍정/부정 리뷰, 크기/성능/디자인 등 키워드별 리뷰, 리뷰 유튜브 영상을 함께 제공하고자 한다.

이러한 서비스를 위해서는 1) 파이썬을 활용해 리뷰 데이터를 크롤링하고 2) 이들을 동일한 컬럼, 인덱스 양식으로 맞춰 여러 데이터 소스의 리뷰를 하나의 데이터베이스로 통합해 MySQL에 연동시킨 뒤 3) 이러한 통합 DB로부터 필요한 리뷰 데이터만을 추출해 워드클라우드, 긍정/부정 감정 예측, 키워드 조회 등에 활용하는 절차가 요구된다.

- Technical challenges / Merit

네이버, 쿠팡 등 플랫폼들은 과도한 데이터 크롤링으로 자사에 축적된 데이터를 이용료 없이 사용하거나 서비스 이용에 지장이 생기는 것을 방지하기 위해 다양한 방식으로 크롤링 구현 난이도를 높여놓고 있다. 이에 단순한 코드만으로 웹 페이지 소스 코드에서 리뷰 데이터베이스를 구현하기는 어려워 깃허브 등 커뮤니티에 공개된 크롤링 모델들을 차용하여 작업했다.

또 하나의 어려움은 리뷰 본문들이 온라인 상의 자의적 표기법에 따라 작성됨에 따라 어떤 문법 규칙을 적용해 표준적인 전처리를 하기 어렵다는 점이다. 가령 키워드 추출 시에 해당 키워드가 포함된 리뷰 문장만을 뽑으려 해도 통상적인 문장 종결 표기법 온점(.)을 사용하지 않은 경우가 많아 문장 단위 데이터 추출이 어려웠다.

- Data required / Data format / Data characteristics / Source of data

네이버 / 쿠팡을 중심으로 꿀리뷰까지 추가해 1) 노트북은 삼성, LG, 애플 3개 제조사의 7개 제품, 2) 스마트폰은 삼성, 애플 2개 제조사의 5개 제품, 3) 음향기기는 삼성, 애플 2개 제조사의 4개 제품에 대한 리뷰 데이터를 수집하고자 한다. 노트북의 경우 제조사별 대표 제품이면서 리뷰 적재가 비교적 많은 1) 삼성 갤럭시북 플렉스, 갤럭시북 이온, 갤럭시북2 프로, LG 그램과 울트라PC, 애플 맥북 프로와 에어가 대상이 됐다. 2) 스마트폰은 삼성 갤럭시 Z플립4, Z폴드4, 갤럭시 S22와 애플 아이폰 14, 14 Plus를, 3) 음향기기는 삼성 갤럭시 버즈2, 버즈2 프로와 애플 에어팟 프로 2세대와 3세대를 대상으로 했다.

크롤링한 리뷰 데이터는 기본적으로 스프레드시트 형태로 1차 수집돼 통합 csv 파일로 편집, 이후 MySQL에 연동될 수 있도록 구성했다.

- Database workloads

OLTP

1. 신규 리뷰 추가 시 DB 업데이트

INSERT INTO user_table, review_table

(ex)

INSERT INTO teamdb9.user_table VALUES (35, 'Sangwoo Heo', ‘2022-12-12’, 11074, ‘맥북 에어’);

INSERT INTO teamdb9.review_table VALUES (7, 'Apple 2022 맥북 에어, 실버, M2 8코어, GPU 8코어, 256GB, 8GB, 30W, 한글', '맥북 에어', '쿠팡답게 칼배송입이다 미드나잇과 스그중 무엇을 살까 고민하다 미드나잇을 케이스 씌우고 쓰기로했습니다 지문때문에 케이스는 필수이며 실물은 무지무지 이쁩니다 미드나잇 쓰실분은 꼭 케이스도 쓰시길..', 5);

OLAP

1. 검색어로 제품명 입력 시 워드클라우드, 긍정/부정 분류, 키워드 분류, 유튜브 리뷰 영상 검색 작업 위해 해당 제품 리뷰 데이터 끌어오기

f'SELECT r.review FROM review_table r WHERE r.condensed_name = {search}'

2. 최근 등록된 리뷰

SELECT r.condensed_name, r.review FROM review_table r

JOIN user_table u ON u.review_id = r.review_id

ORDER BY u.datetime DESC LIMIT 7

3. 지금 HOT한 리뷰

SELECT r.condensed_name, p.product_image_link, COUNT(r.review_id) AS rcnt FROM

review_table r

JOIN user_table u ON u.review_id = r.review_id

JOIN product_table p ON p.product_id = r.product_id

GROUP BY r.condensed_name ORDER BY rcnt DESC LIMIT 3

WHERE u.datetime > (Last day of Last month)

((((

1) 최근 24시간 내 검색 로그 데이터를 기준으로 검색량이 가장 많은 제품 M개를 추려 [메인 화면]-[지금 HOT한 제품]에 제품명과 제품 이미지를 보여준다. (이미지 데이터는 어디서 끌어온다고 하지????)
2) 리뷰 데이터베이스에 적재된 시간을 기준으로 최신 N개 제품을 뽑아 제품명과 리뷰 본문을 서비스 [메인 화면]-[최근 등록된 리뷰] 부분에 보여준다.
3) 통합된 리뷰 데이터베이스로부터 유저가 입력한 검색 대상 제품에 관한 리뷰만을 끌어온다.
4) 3) 데이터를 기반으로 긍정/부정 sentiment 예측 확률이 높은 순서로 각각 상위 5개 리뷰들을 각각 [검색 화면]에 보여준다.
5) 3) 데이터를 기반으로 불용어, 제품명 등 의미가 적은 단어들을 제외한 워드클라우드를 구현하여 사과 모양의 이미지로 [검색 화면]에 보여준다.
6) 3) 데이터를 기반으로 불용어를 제외하고 크기 / 색상 등 키워드별 리뷰 문장 5개씩을 [검색 화면]에 보여준다.

))))

How to model data

Keyword Classification

- explore the data to select only those reviews that mention color, size, and price within the sentence after a sentence-by-sentence tokenization of the Natural Language Toolkit (NLTK) package

Sentiment analysis

-data for modeling : 200,000 Korean review with star rating from Naver shopping
-형태소 분석기 Mecab을 이용해 형태소 토큰화(morpheme tokenization)을 하고, tensorflow karas 라이브러리의 GRU를 이용해 모델 학습.

Word cloud

- Word Cloud visualizes and shows the size of frequently appearing words in relative sizes, providing an intuitive glimpse of the influence of commonly used words.
- KoNLPy (package for natural Korean language processing) was used for Korean text analysis.
- Wordcloud package was used for word visualization.

1. Bring up the crawled comment file and save it in the list variable
2. Create Okt Analysis Module from Konlpy Package
3. Use repetitive sentences to classify shapes and match parts

Architectural design (end to end) / why you select particular database systems for required functionalities

######유진 메모

1. Data source : Naver shopping, coupang,...
2. (Python+human) Data crawling & preprocessing : review. csv
3. (MySQL) Database : Query
    a. User Input : new review, date_time, user ID
    b. Product Info : image, manufacturer,

4. (Python) Processing
    a. NLP- Sentiment analysis, word cloud visualization, keyword classification
    b. UI/UX : TKINTER
5. Service to User

Results / demonstration

######유진 메모
-Explain Prototype 'Review everywhere everything all at once'(영어 제목 어떰? 번뜩이는 아이디어~~~~~~)

Limitation (other than time), what prevents from doing better

##유진메모

-We extracted only Korean reviews.

IT reviews are written in various languages around the world, and products sold first in other countries such as the United States may deliver faster information on English reviews.

Reflection on doing this project/lessons learned from this project

##유진메모

-Review everywhere everything all at once

In addition to integrated reviews of Coupang and Naver Shopping, a platform where consumers recently purchase a lot of IT products, you can also analyze high-quality reviews such as review sites such as honey reviews and reviews of small shopping malls such as student welfare stores.

Team member introduction, roles, and contribution

1) 신성구 (데이터사이언스학과)
- 개발 방향 총괄
- 유튜브 리뷰 영상 크롤링
- MySQL Database 구축
- 데모 ui 설계 및 개발

2) 이준규 (데이터사이언스학과)
- 네이버 리뷰 크롤링 ````
- 데이터베이스 통합
- MySQL Database 구축

3) 김정국 (데이터사이언스학과)
- 쿠팡, 학생복지스토어 리뷰 크롤링
- 리뷰 워드클라우드 코드 구현

4) 조유진 (데이터사이언스학과)
- 꿀리뷰, 네이버 리뷰 크롤링
- 긍/부정 리뷰 분류

5) 허상우 (데이터사이언스학과)
- 쿠팡 리뷰 크롤링
- 키워드별 리뷰 분류
- 보고서 초안 작성

Future work / possible extensions

1) 유저별 검색 로그 데이터를 별도 데이터베이스에 적재하여 이전에 검색했던 제품과의 리뷰 비교가 가능하도록 할 수 있다.