가전제품 리뷰 기반
텍스트 분석 및 스코어링 및 추천 서비스

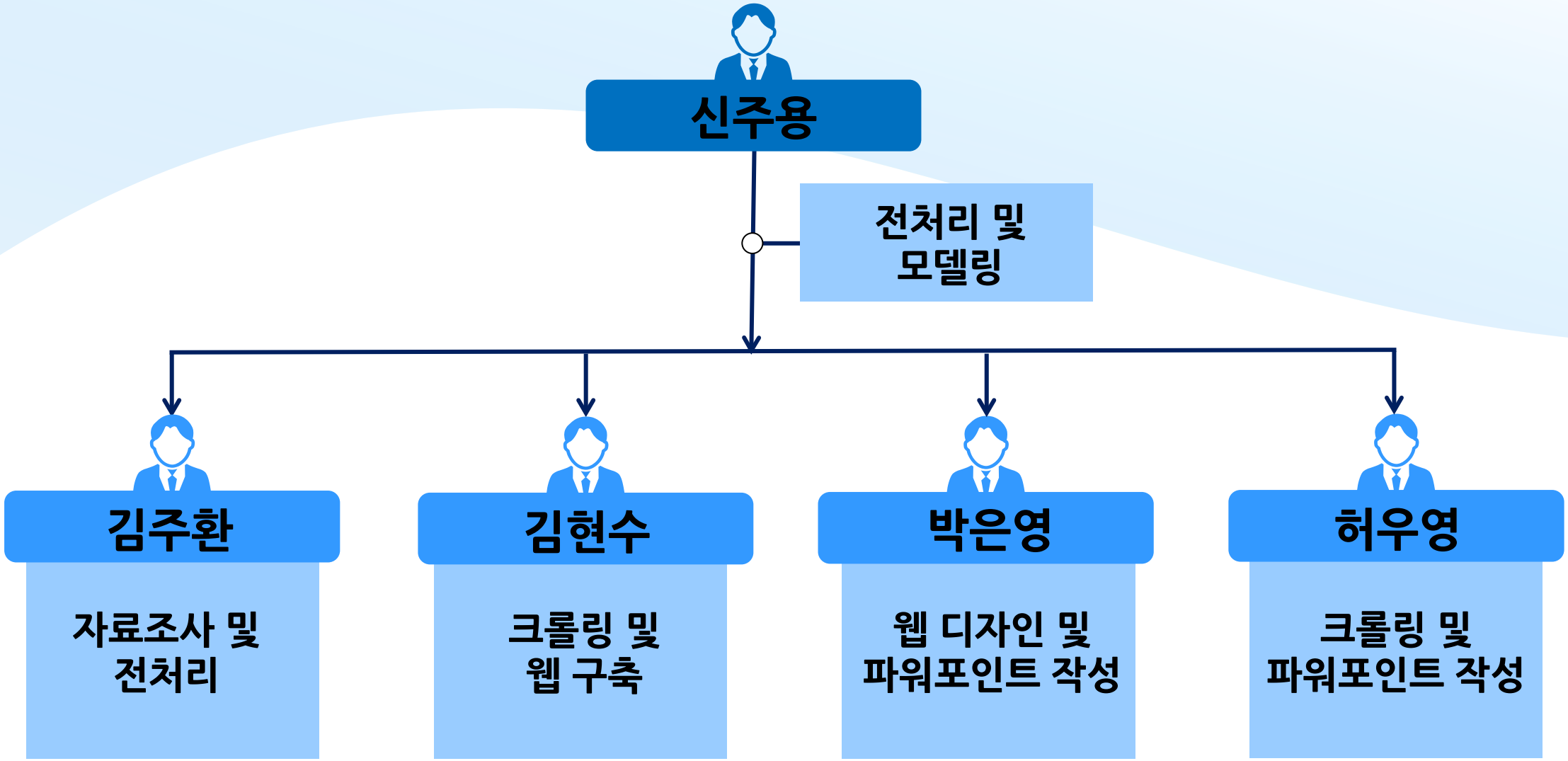
팀장 : 신주용

팀원 : 김주환, 김현수, 박은영, 허우영

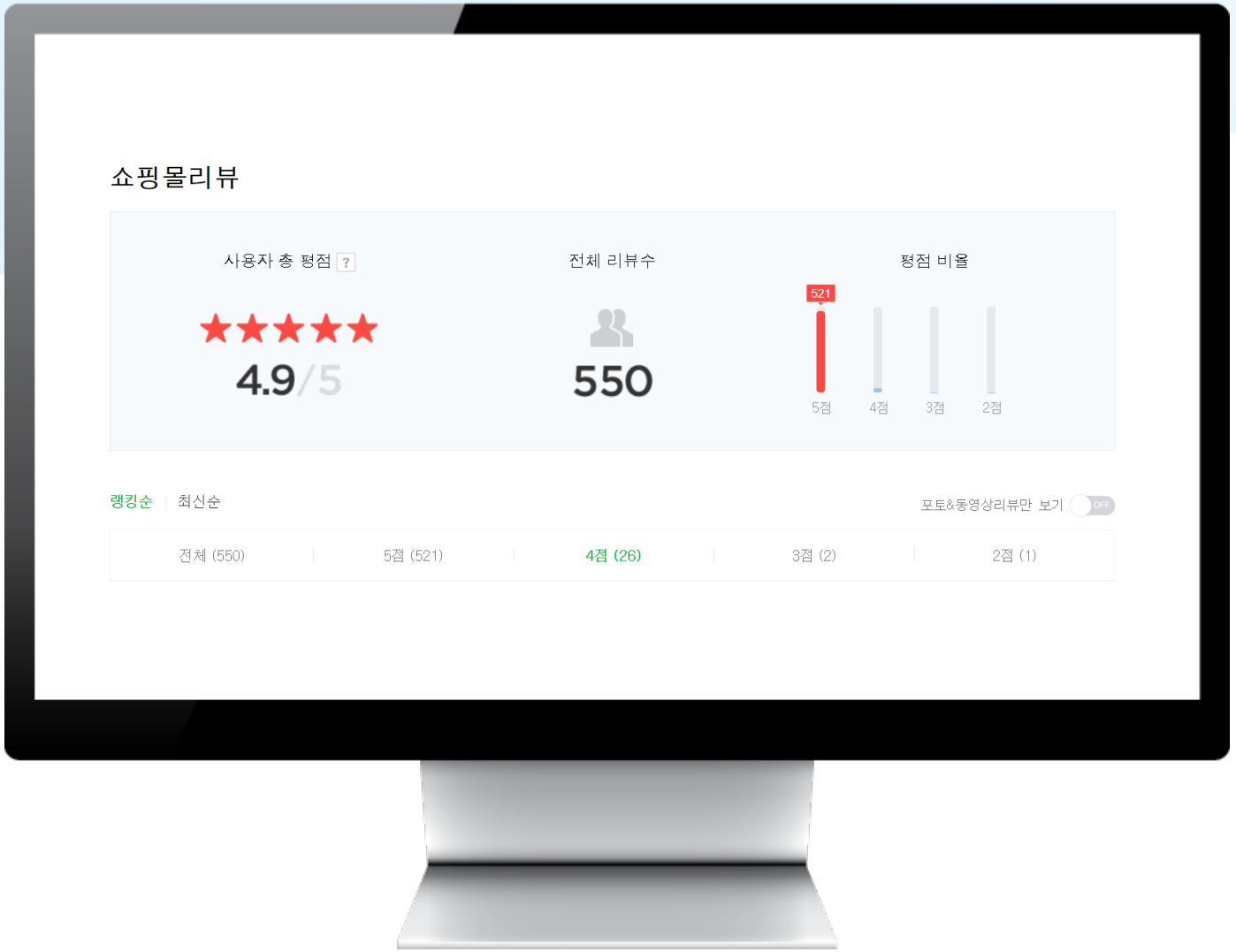
목차

01. 역할 분담 및 프로젝트 주제 소개

01. 역할 분담


신주용
전처리 및
모델링
김주환
자료조사 및
전처리
김현수
크롤링 및
웹 구축
박은영
웹 디자인 및
파워포인트 작성
허우영
크롤링 및
파워포인트 작성

01. 프로젝트 주제 소개

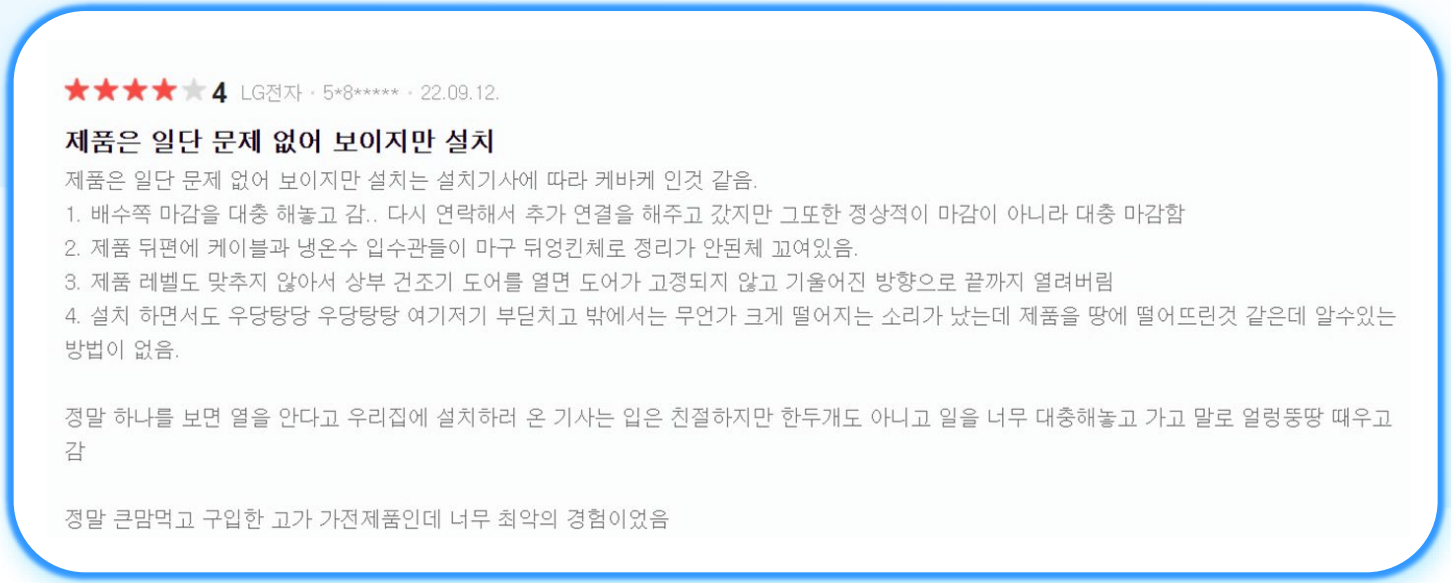


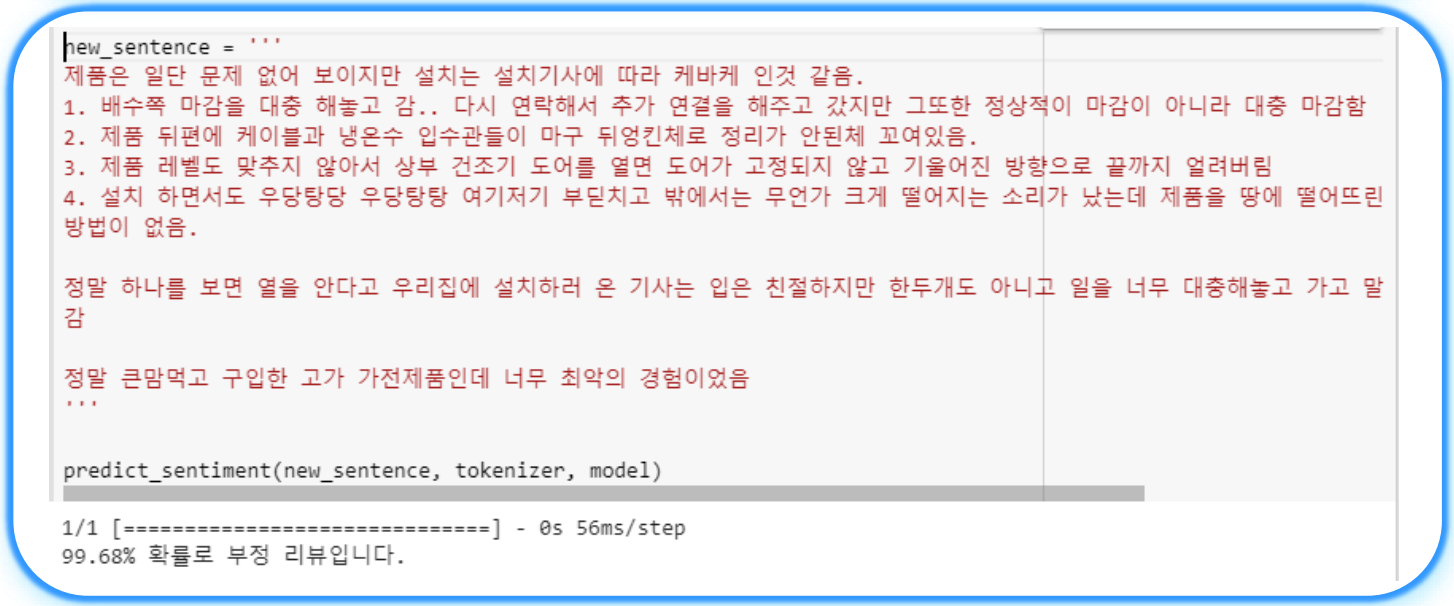
- 온라인에서 쇼핑을 할 때 리뷰를 보고 결정을 하는 경우가 많다.
- 대부분의 리뷰가 평점은 높은데 리뷰 자체는 신뢰성이 부족하다.
- 일일이 리뷰를 전부 읽기에는 양이 너무 많다.

01. 프로젝트 주제 소개

- 감성분석을 해보면 부정 99%가 넘어가는 리뷰인데 평점은 4점입니다.

- 이 4점은 작성된 리뷰와 상관없이 높은 점수라서 진짜 점수가 아니라고 보고, 글을 분석해서 나온 점수가 더 정확하다고 판단, 분석을 통해서 정확한 리뷰 점수를 알려주기 위해 시작하게 되었다.





02. 데이터 소개 및 탐색

02. 데이터 소개 및 탐색

NAVER Developers Products Documents Application NAVER D2 Support Forum API 상태 Search Here

Products > API 이용 안내 > API 소개

운영 정책
FAQ
BI 가이드
이용약관
상표사용 가이드

네이버 오픈 API 목록

네이버 오픈API 목록 및 안내입니다.

API명	설명	호출제한
검색	네이버 블로그, 이미지, 웹, 뉴스, 백과사전, 책, 카페, 지식iN 등 검색	25,000회/일
네이버 로그인	외부 사이트에서 네이버 로그인 기능 구현	없음
네이버 회원 프로필 조회	네이버 회원 이름, 이메일 주소, 휴대전화번호, 별명, 성별, 생일, 연령대, 출생연도, 프로필 조회	없음
Papago 번역	Papago 번역 인공신경망 기반 기계 번역	10,000글자/일
CLOVA Face Recognition	입력된 사진을 입력받아 얼굴윤곽/부위/표정/유명인 닮음도를 리턴	1,000건/일
데이터랩(검색어트렌드)	통합검색어 트렌드 조회	1,000회/일

02. 데이터 소개 및 탐색

Feature names	Type	Description
Rss	-	RSS 컨테이너. RSS 리더기를 사용해 검색 결과를 확인할 수 있습니다.
rss/channel	-	검색 결과를 포함하는 컨테이너. channel 요소의 하위 요소인 title, link, description은 RSS에서 사용하는 정보이며, 검색 결과와는 상관이 없습니다.
rss/channel/lastBuildDate	dateTime	검색 결과를 생성한 시간
rss/channel/total	Integer	총 검색 결과 개수
rss/channel/start	Integer	검색 시작 위치
rss/channel/display	Integer	한 번에 표시할 검색 결과 개수
rss/channel/item	-	개별 검색 결과. JSON 형식의 결괏값에서는 items 속성의 JSON 배열로 개별 검색 결과를 반환합니다.
rss/channel/item/title	String	상품 이름. 이름에서 검색어와 일치하는 부분은 <b> 태그로 감싸져 있습니다.
rss/channel/item/link	String	상품 정보 URL
rss/channel/item/image	String	섬네일 이미지의 URL

Feature names	Type	Description
rss/channel/item/lprice	Integer	최저가. 최저가 정보가 없으면 0을 반환합니다. 가격 비교 데이터가 없으면 상품 가격을 의미합니다.
rss/channel/item/hprice	Integer	최고가. 최고가 정보가 없거나 가격 비교 데이터가 없으면 0을 반환합니다.
rss/channel/item/mallName	String	상품을 판매하는 쇼핑몰. 쇼핑몰 정보가 없으면 네이버를 반환합니다.
rss/channel/item/productId	Integer	네이버 쇼핑의 상품 ID
rss/channel/item/productType	Integer	상품군과 상품 종류에 따른 상품 타입. 상품군과 상품 종류에 따른 - 상품군: 일반상품, 중고상품, 단종상품, 판매예정상품 - 상품 종류: 가격비교 상품, 가격비교 비매칭 일반상품, 가격비교 매칭 일반상품
rss/channel/item/maker	String	제조사
rss/channel/item/brand	String	브랜드
rss/channel/item/category1	String	상품의 카테고리(대분류)
rss/channel/item/category2	String	상품의 카테고리(중분류)
rss/channel/item/category3	String	상품의 카테고리(소분류)

02. 네이버 쇼핑 크롤링

★★★★★5 애플 공식 브랜드스토어 · minc**** · 22.11.29. · 모델 선택: 에어팟 프로 2세대 MQD83KH/A

번개 같은 빠름으로? 사전예약 후 지난 10월 21일 수령해서 지금까지 한달 넘

번개 같은 빠름으로? 사전예약 후 지난 10월 21일 수령해서 지금까지 한달 넘게 사용중인데 정말 맘에

아마 많은 분들이 큰 기대를 갖고 제품을 구입하셨을 거예요. 한두푼 하는것도 아니고 이어폰 따위가 3
투자해서 스트레스 없이 편하게 약 2~3년간 사용할 수 있다면 뭐.. 나쁘지 않다고 생각을 했어요.(한달
지금부터 사용기를 정리해 보겠습니다.

```
<div class="reviewItems_review_text__dq0kE"> == $0
  <em class="reviewItems_title__AwHcz">
    "번개 같은 빠름으로? 사전예약 후 지난 10월 21일 수령해서 지금까지 한달 넘게 사용중인데 정말 맘에 듭니다.아마 많은 분들이 큰 기대를 갖고 제품을 구입하셨을 거예요. 한두푼 "
  </em>
  <p class="reviewItems_text__XrSSf">…</p>
</div>
<div class="reviewItems_review_thumb__3WU2G">…</div>
</div>
```

★★★★★5 삼성공식파트너 보보 · kynp**** · 21.10.17.

이거 출시된지도 모르고중소기업꺼 사려고 찾아보고 있었는데그거 샀으면 땅을 칠 뻔 했어요이건 후기도 …

이거 출시된지도 모르고
중소기업꺼 사려고 찾아보고 있었는데
그거 샀으면 땅을 칠 뻔 했어요
이건 후기도 안 보고
묻지도 따지지도 않고 바로 결제 했습니다

```
<div class="reviewItems_review_text__dq0kE"> == $0
  <em class="reviewItems_title__AwHcz">
    "이거 출시된지도 모르고중소기업꺼 사려고 찾아보고 있었는데그거 샀으면 땅을 칠 뻔 했어요이건 후기도 안 보고묻지도 따지지도 않고 바로 결제 했습니다배송은 엄청 빠르긴 한데역시 원래 "
  </em>
  <p class="reviewItems_text__XrSSf">…</p>
</div>
<div class="reviewItems_review_thumb__3WU2G">…</div>
</div>
```

★★★★★5 하이마트쇼핑몰 · 1*0***** · 22.07.14.

오랫동안 고민했던 문제들이 한번

오랫동안 고민했던 문제들이 한번에 해결되네요♡♡♡
가장 놀랐던것은 소음입니다 굉장히 조용하고 강력합니다
먼저 청소기는 돌리면서도 먼지 냄새가 나고 청소기 돌릴때는 전화기소리도 잘 못들었는데 삼성 비스포크
라 언제든지 가볍게 돌릴수 있어서 참 좋네요 무게는 무거운편이라는 후기를 보고 걱정했는데 제가 젊어선
생각보다 가벼워서 잘쓰고있습니다 강아지도 있고 탈모증상도 있어서 머리카락도 잘빠지는데 거치대에서

```
<div class="reviewItems_review__DqLYb"> == $0
  <div class="reviewItems_review_text__dq0kE">
    <em class="reviewItems_title__AwHcz">오랫동안 고민했던 문제들이 한번</em>
    <p class="reviewItems_text__XrSSf">…</p>
  </div>
```

03. 데이터 전처리 과정 및 시각화

03. 데이터 전처리 과정 및 시각화

- 워드 클라우드

텍스트 데이터 자료형으로

Tokenize (형태소 분석)

불용어 제거

TF-IDF Vectorizer

04. 적용한 분석 기법 및 모델 소개

04. Kobert-tokenizer

- SKTBrain's

한국어 위키피디아에서 얻은 500만 개 이상의 문장을 대상으로 pretrain된 KoBERT tokenizer를 활용

1) Load the pre-trained tokenizer

```
[8] !git clone https://github.com/monologg/KoBERT-Transformers.git
    !mv KoBERT-Transformers/kobert_transformers/tokenization_kobert.py /content

    clear_output() # clear the output
```

```
from tokenization_kobert import KoBertTokenizer

tokenizer = KoBertTokenizer.from_pretrained('monologg/kobert') # sentencepiece 라이브러리가 먼저 import 되어있어야 합니다.
# tokenizer = BertTokenizer.from_pretrained('bert-base-multilingual-cased', cache_dir='bert_ckpt', do_lower_case=False)
```

```
Downloading (...)zer_78b3253a26.model: 100% 371k/371k [00:00<00:00, 2.61MB/s]
Downloading (...)solve/main/vocab.txt: 100% 77.8k/77.8k [00:00<00:00, 2.40MB/s]
Downloading (...)okenizer_config.json: 100% 51.0/51.0 [00:00<00:00, 775B/s]
Downloading (...)lve/main/config.json: 100% 426/426 [00:00<00:00, 5.35kB/s]
```

2) browse the usage of BertTokenizer

```
[11] print(tokenizer.tokenize("사전 학습된 토크나이저에 대한 테스트입니다.")) # Only tokenize the sentence

['_사전', '_학습', '된', '_', '토크', '나', '이', '저', '에', '_대한', '_테스트', '입니다', '.']

[12] print(tokenizer.encode("사전 학습된 토크나이저에 대한 테스트입니다.")) # Tokenizing + Tokens to sequence numbers

[2, 2625, 4954, 5899, 517, 7630, 5655, 7096, 7199, 6896, 1682, 4736, 7139, 54, 3]

[13] print(tokenizer.encode("사전 학습된 토크나이저에 대한 테스트입니다.", max_length=30, padding='max_length')) # Tokenizing + Tokens to sequence numbers + Padding

[2, 2625, 4954, 5899, 517, 7630, 5655, 7096, 7199, 6896, 1682, 4736, 7139, 54, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

print(tokenizer.convert_ids_to_tokens(tokenizer.encode("사전 학습된 토크나이저에 대한 테스트입니다."))) # tokenizer.convert_ids_to_tokens(SEQUENCE OF TOKEN NUMBERS)

# 실제로는 tokenize() 함수 적용 후 length(17)보다 encode() 함수 적용 결과인 sequence의 length(19)가 2 만큼 더 큽니다.
# 이는 아래와 같이 특수 토큰인 '[CLS]' & '[SEP]' 토큰이 자동으로 추가되었기 때문입니다.

# '[CLS]' : Special Classification token(CLS), 모든 문장의 가장 첫 번째(문장의 시작) 토큰으로 삽입됩니다.
# '[SEP]' : Special Separator token(SEP), 문자열이 2개의 문장으로 구성되었을 때 첫 번째 문장과 두 번째 문장을 구별하기 위해 삽입됩니다. (여기서는 문장이 하나이므로 큰 의미가 없습니다.)

['[CLS]', '_사전', '_학습', '된', '_', '토크', '나', '이', '저', '에', '_대한', '_테스트', '입니다', '.', '[SEP]']
```

04. 적용한 분석 기법 및 모델 소개

- 전통적인 머신러닝 XGBoost 성능

```
Training accuracy : 0.9165
Test accuracy : 0.8859
```

- Bidirectional LSTM Network 성능

```
1875/1875 [==============================] - 4s 2ms/step
0.885315058442966
```

- Bert 성능

```
1875/1875 [==============================] - 52s 25ms/step
0.91506177780001
```

- KoBert 성능

```
0.9338368932686375
```

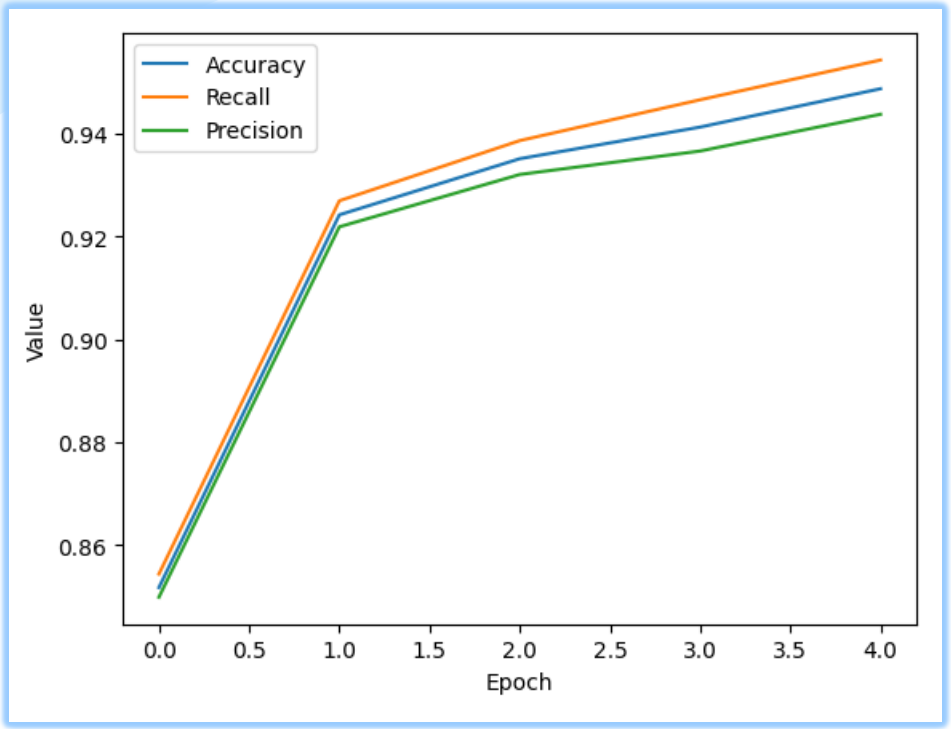
- KoBERT 모델 선정한 이유
  다른 모델보다 정확도가 높고, 성능이 좋다.

- KoBERT는 기존 BERT의 한국어 성능 한계를 극복하기 위해 개발되었다. 위키피디아나 뉴스 등에서 수집한 수백만 개의 한국어 문장으로 이루어진 대규모 말뭉치(corpus)를 학습하였으며, 한국어의 불규칙한 언어 변화의 특성을 반영하기 위해 데이터 기반 토큰화(Tokenization) 기법을 적용하여 기존 대비 27%의 토큰만으로 2.6% 이상의 성능 향상을 이끌어 냈다.

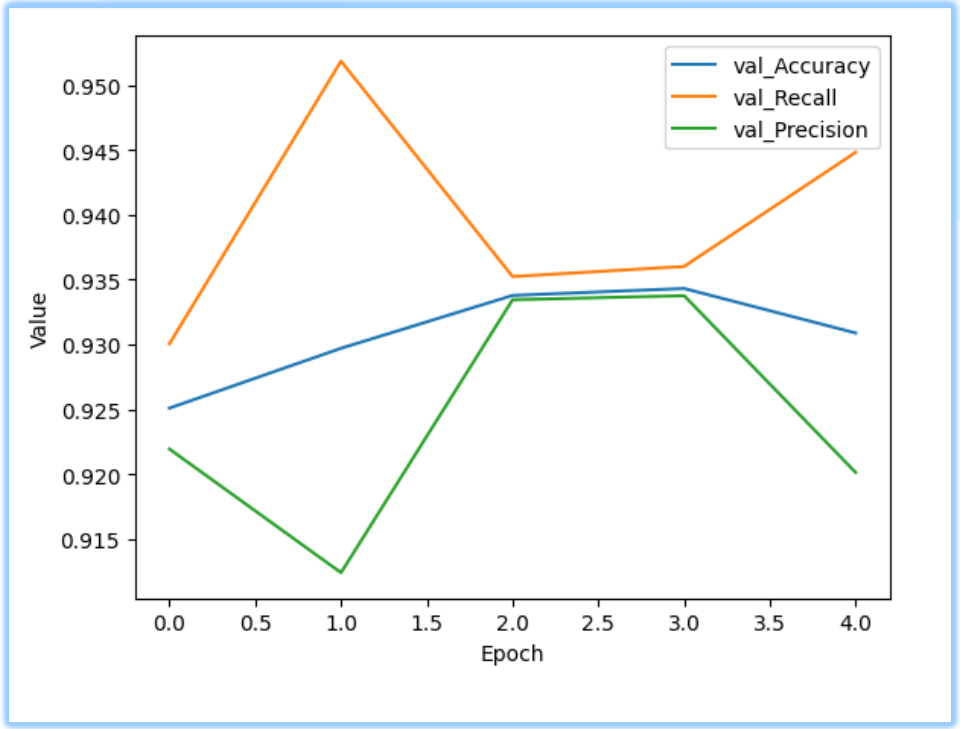
05. 모델링 평가 지표

05. 모델링 평가 지표

Epoch for Train Accuracy Recall Precision



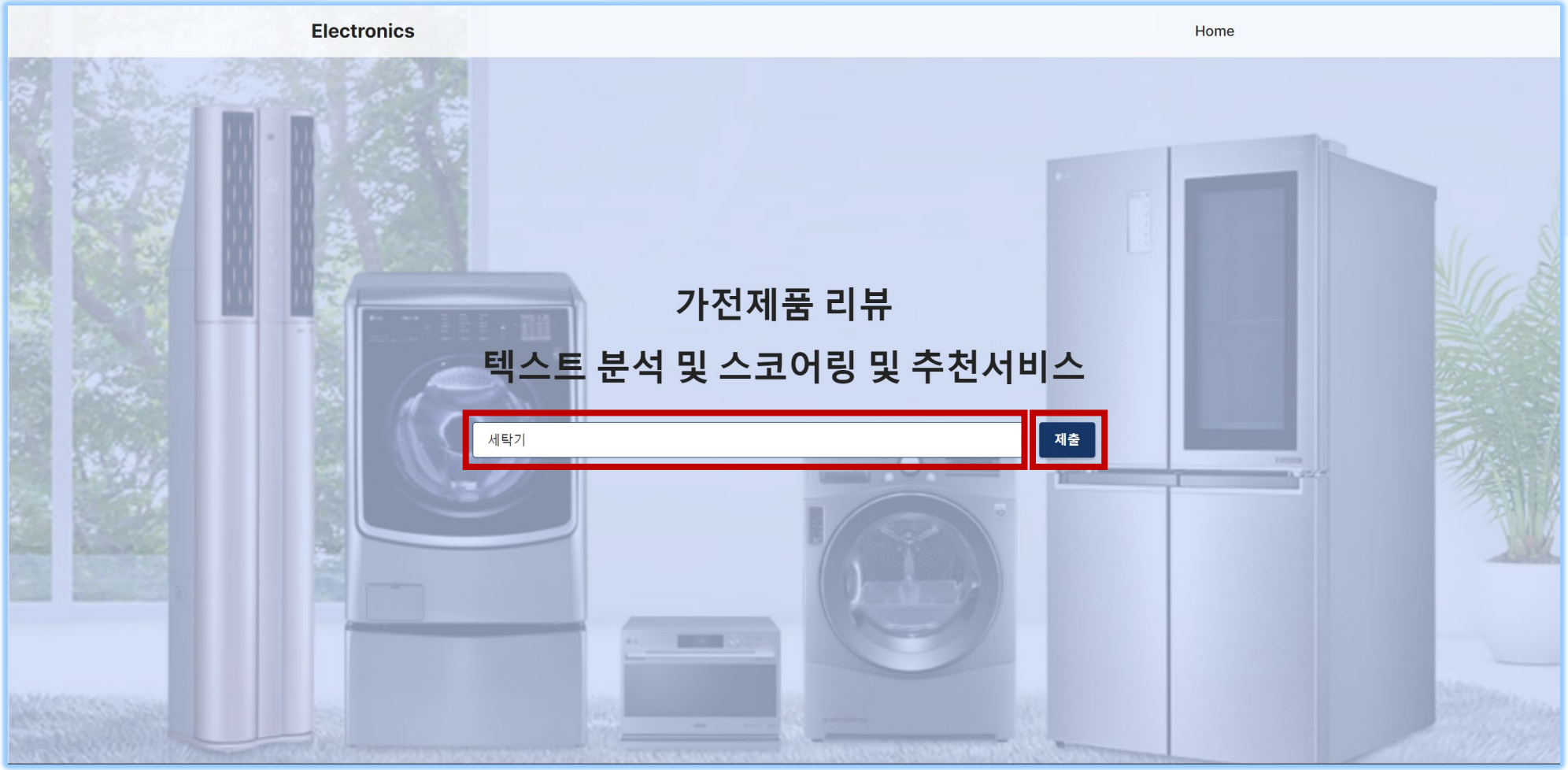
Epoch for Validation Accuracy Recall Precision



06. 모델을 활용한 앱 서비스에 대한 소개

06. 모델을 활용한 웹 서비스에 대한 소개

- Index.html



06. 모델을 활용한 웹 서비스에 대한 소개

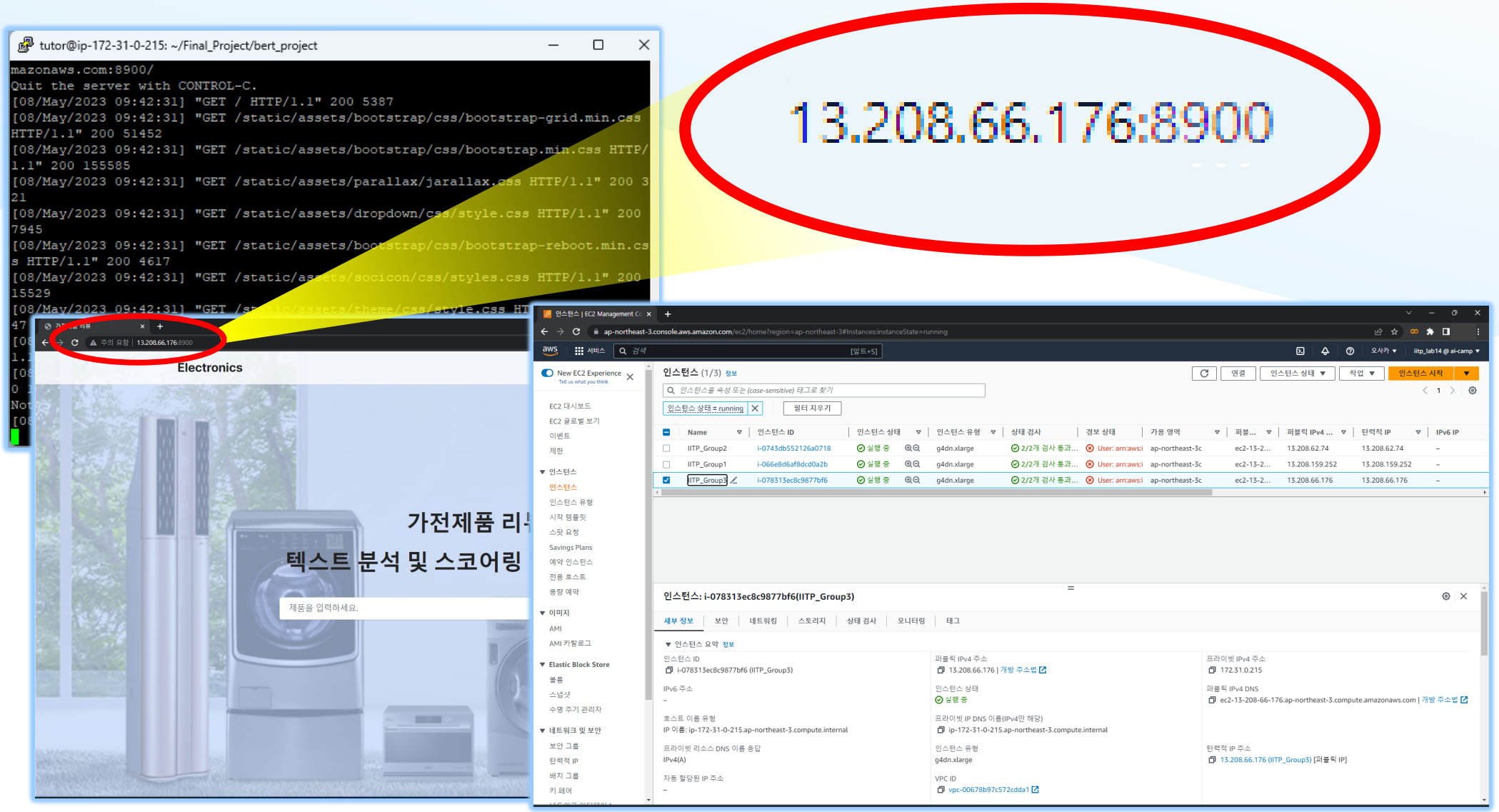
- detail.html

06. 모델을 활용한 웹 서비스에 대한 소개

- result.html

07. AWS 배포 과정

07. AWS 배포 과정


tutor@ip-172-31-0-215: ~/Final_Project/bert_project
mazonaws.com:8900/
Quit the server with CONTROL-C.
[08/May/2023 09:42:31] "GET / HTTP/1.1" 200 5387
[08/May/2023 09:42:31] "GET /static/assets/bootstrap/css/bootstrap-grid.min.css HTTP/1.1" 200 51452
[08/May/2023 09:42:31] "GET /static/assets/bootstrap/css/bootstrap.min.css HTTP/1.1" 200 155585
[08/May/2023 09:42:31] "GET /static/assets/parallax/jarallax.css HTTP/1.1" 200 321
[08/May/2023 09:42:31] "GET /static/assets/dropdown/css/style.css HTTP/1.1" 200 7945
[08/May/2023 09:42:31] "GET /static/assets/bootstrap/css/bootstrap-reboot.min.css HTTP/1.1" 200 4617
[08/May/2023 09:42:31] "GET /static/assets/socicon/css/styles.css HTTP/1.1" 200 15529
13.208.66.176:8900
주의 요함 | 13.208.66.176:8900
Electronics
가전제품 리뷰
텍스트 분석 및 스코어링
제품을 입력하세요.
인스턴스 | EC2 Management Co
ap-northeast-3.console.aws.amazon.com/ec2/home?region=ap-northeast-3#Instances:instanceState=running
인스턴스 (1/3) 정보
인스턴스 상태 = running
필터 지우기
Name
인스턴스 ID
인스턴스 상태
인스턴스 유형
상태 검사
경보 상태
가용 영역
IITP_Group2 i-0743db552126a0718 실행 중 g4dn.xlarge 2/2개 검사 통과... User: arn:aws:i ap-northeast-3c
IITP_Group1 i-066e8d6af8dcd0a2b 실행 중 g4dn.xlarge 2/2개 검사 통과... User: arn:aws:i ap-northeast-3c
IITP_Group3 i-078313ec8c9877bf6 실행 중 g4dn.xlarge 2/2개 검사 통과... User: arn:aws:i ap-northeast-3c
인스턴스: i-078313ec8c9877bf6(IITP_Group3)
세부 정보
보안
네트워킹
스토리지
상태 검사
모니터링
태그
인스턴스 요약 정보
인스턴스 ID
i-078313ec8c9877bf6 (IITP_Group3)
퍼블릭 IPv4 주소
13.208.66.176 | 개방 주소법
프라이빗 IPv4 주소
172.31.0.215
IPv6 주소
인스턴스 상태
실행 중
퍼블릭 IPv4 DNS
ec2-13-208-66-176.ap-northeast-3.compute.amazonaws.com | 개방 주소법
호스트 이름 유형
IP 이름: ip-172-31-0-215.ap-northeast-3.compute.internal
프라이빗 IP DNS 이름(IPv4만 해당)
ip-172-31-0-215.ap-northeast-3.compute.internal
프라이빗 리소스 DNS 이름 응답
IPv4(A)
인스턴스 유형
g4dn.xlarge
탄력적 IP 주소
13.208.66.176 (IITP_Group3) [퍼블릭 IP]
자동 할당된 IP 주소
VPC ID
vpc-00678b97c572cdda1

08. 후속 과제

08. 후속 과제

- Model과 서비스단 분리하여 향후 유지보수를 위한 모듈화 작업
- 댓글, 비동기 통신 크롤링 사용하기
- 서비스 속도를 위해 Model에서 for문 사용하지 않기
- 특정 제품 입력 시, 데이터 가져오는 방법 연구
- 사용자에게 더 다양한 기능을 추가하여 사용자 편의성 제공
- 감정 분석된 평점에 따라 제품의 브랜드에 순위 변동
- 다른 모델을 사용하여 모델 간 성능 평가

Thank you.