

머글끼니

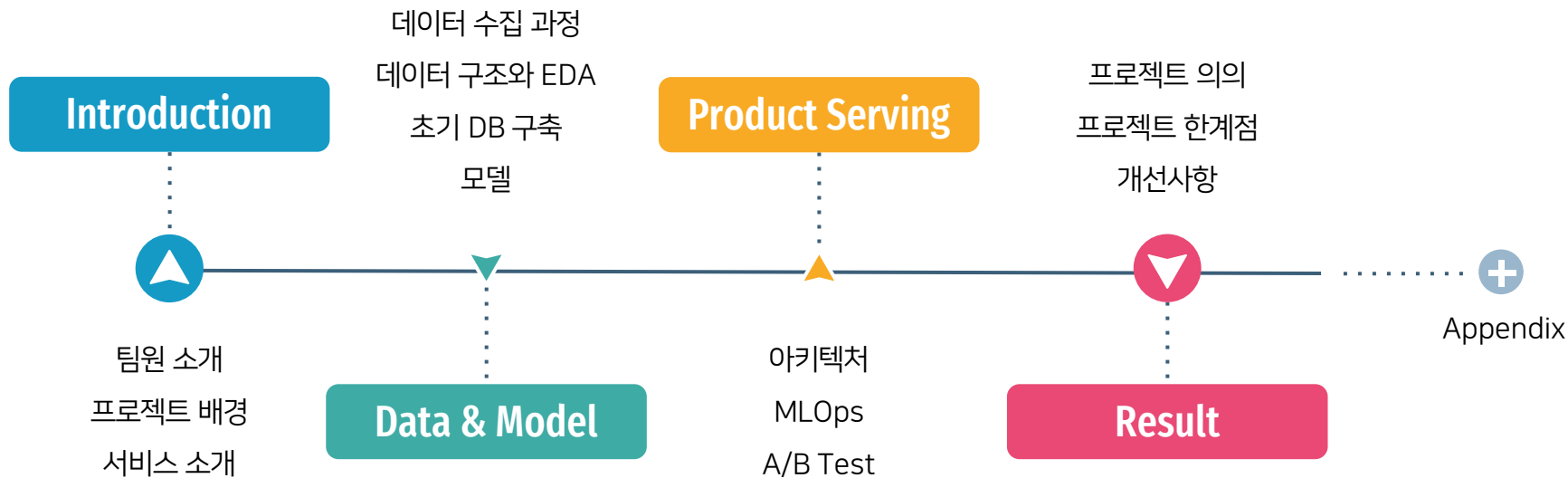
RECCAR조

RecSys - 04

RECCAR | 김성연 배성재 양승훈 조수연 홍재형 황선태



CONTENTS



1

Introduction

Team RECCAR



김성연



배성재



양승훈



조수연

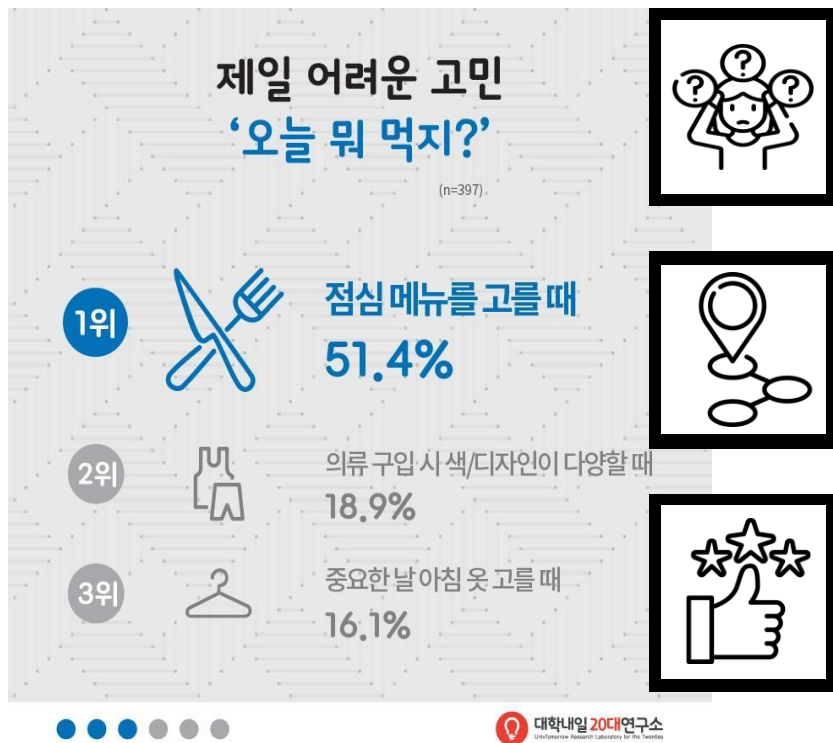


홍재형



황선태

문제 의식



20대의 가장 큰 고민, "오늘 뭐 먹지?"



새로운 지역에서 모였을 때



강력하게 추천해 줄 도구가 필요할 때

(출처: <https://www.vingle.net/posts/997221>)

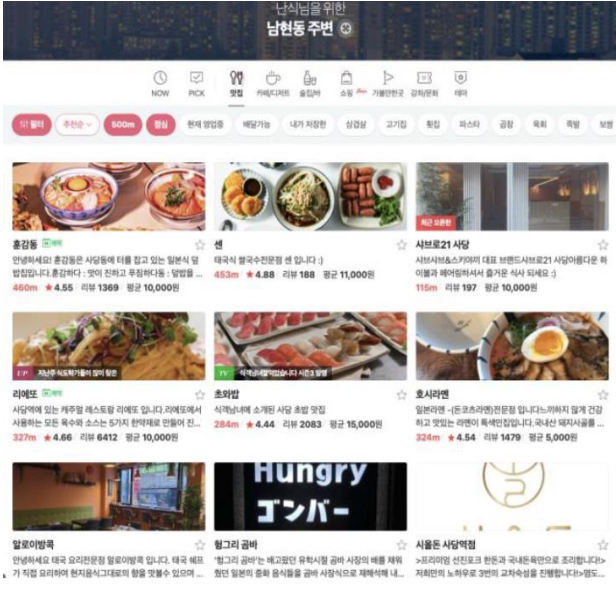
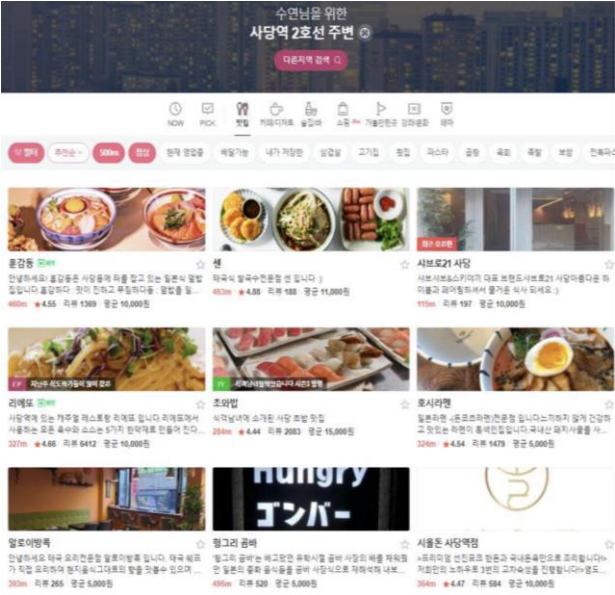
문제 의식

네이버 SmartAround 문제점

24세 남성

20세 여성

52세 여성



머글끼니 RECCAR조

RecSys - 04

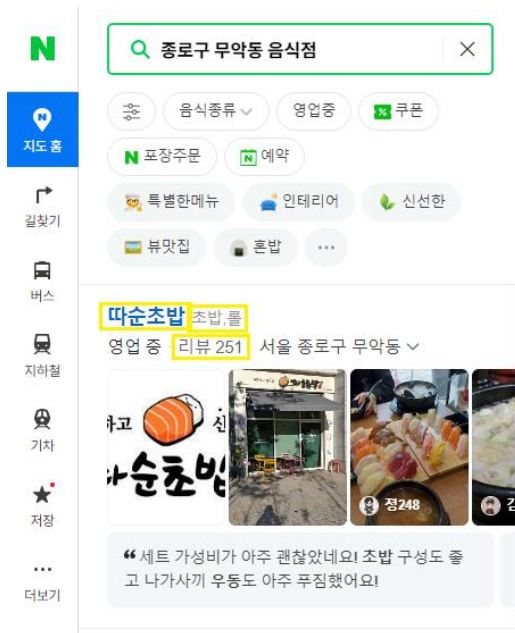
RECCAR | 김성연 배성재 양승훈 조수면 홍재형 황선태

boostcamp ai tech

2

Data & Model

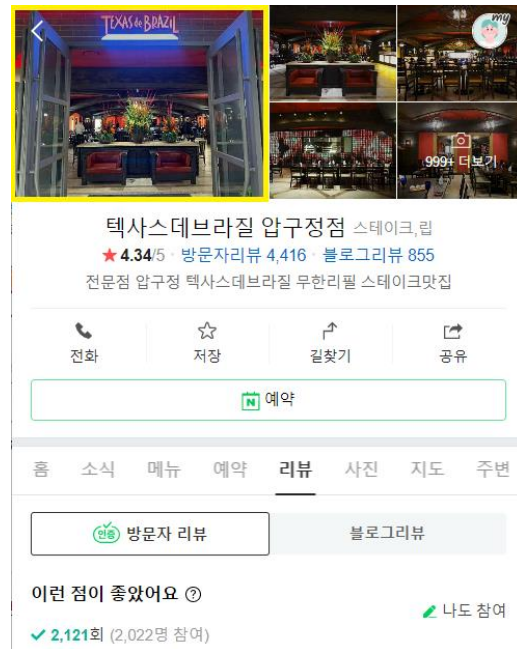
DATA - Data Crawling



음식점 크롤링

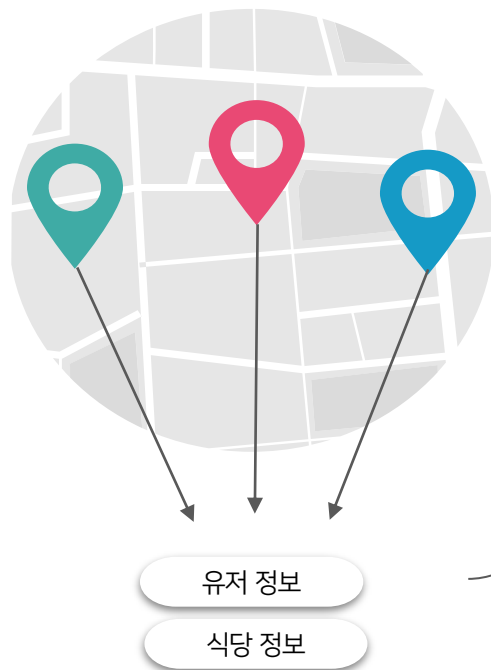


리뷰 크롤링

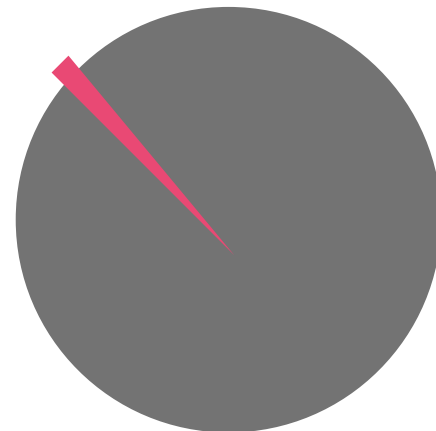


대표사진 크롤링

DATA



리뷰 5개 이하 유저 삭제



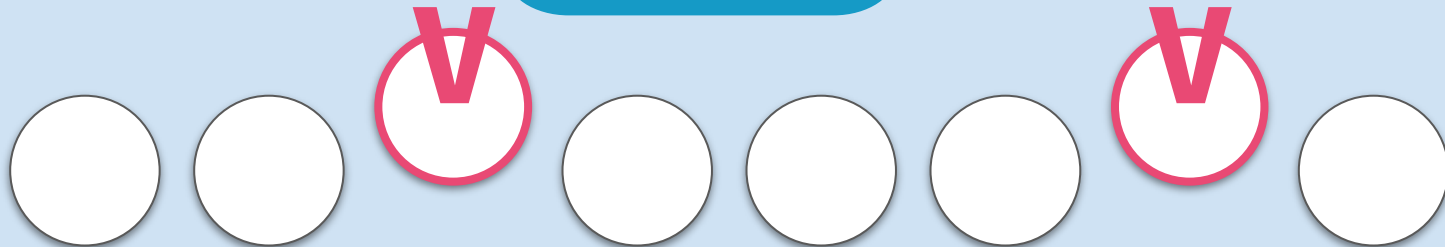
● : 0.26%
● : 99.74%

Sparse matrix

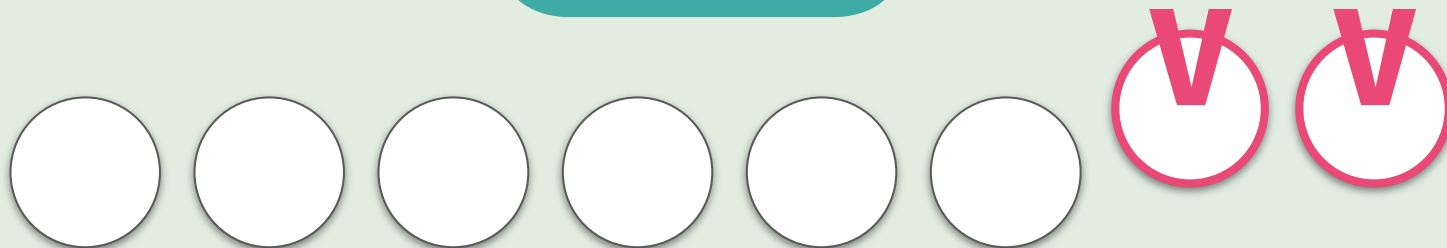
총 41460개의 식당,
382939명의 유저

DATA - Data Preprocessing

Random 분할



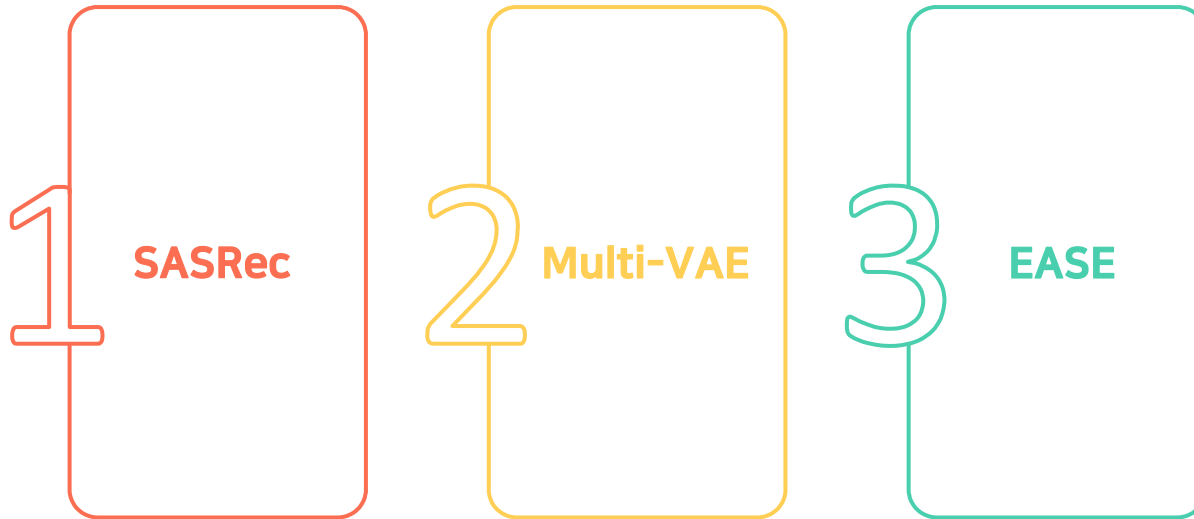
Time 분할



DATABASE - SQLite



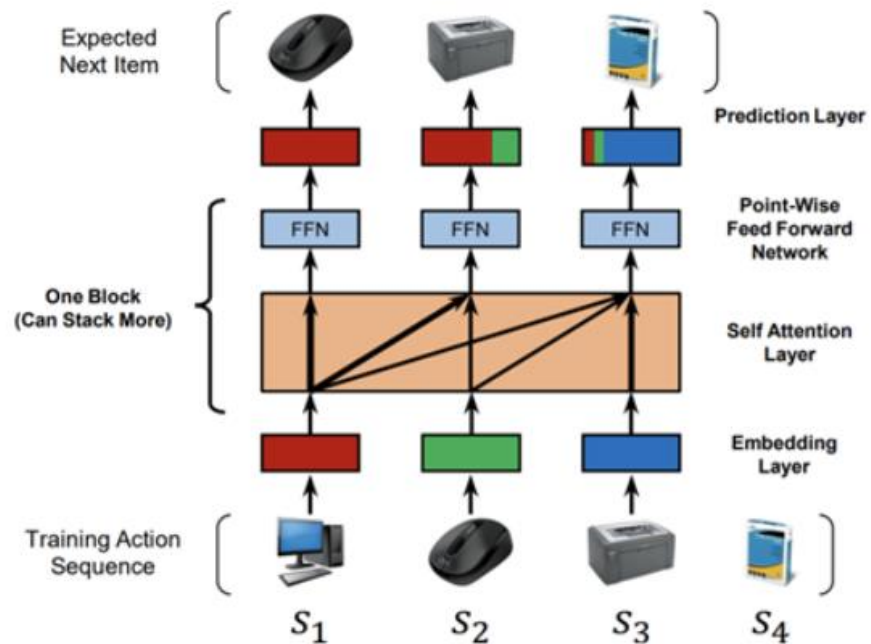
Model



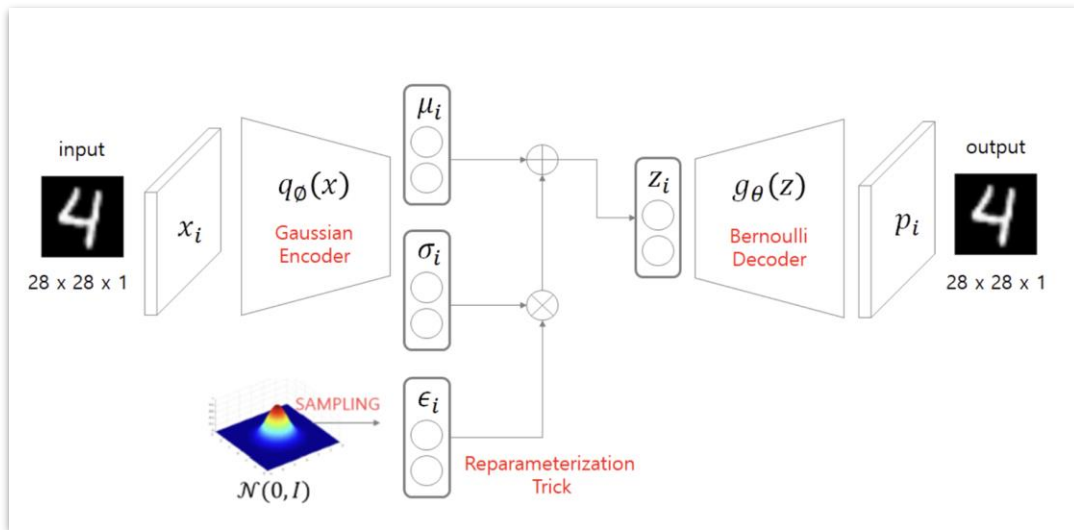
Model - SASRec

SASRec :

NLP에서 주로 쓰이는 Transformer 구조를
sequential recommendation에
적용한 모델



Model - MultiVAE

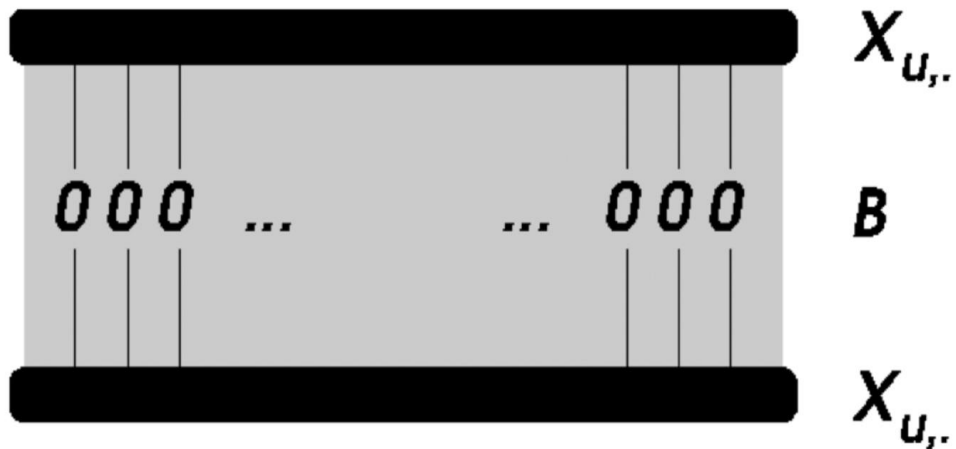


MultiVAE:

VAE 구조에 2개 이상의 가능성 중 하나를
선택하는 모델

multinomial likelihood를 사용하여
implicit feedback data를
더 잘 설명할 수 있음

Model - EASE



EASE :

Embarrassingly Shallow Autoencoders

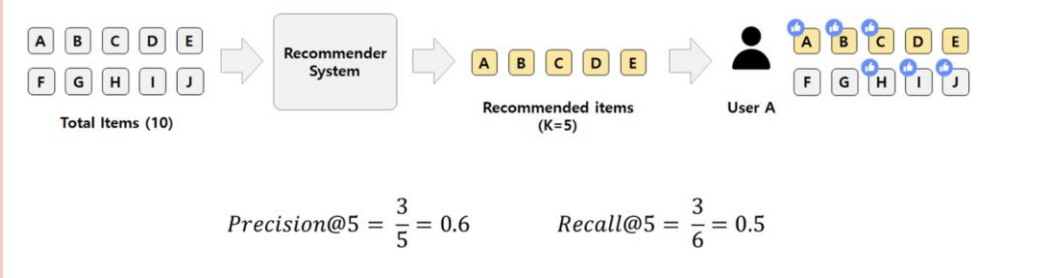
for Sparse Data

neighborhood-based 접근법과 유사한

단순한 구조의 모델

Model - Offline Test Metric

recall @ 20



Personalization

	A	C	B	D	X	Z
0	1	1	1	1	0	0
1	1	1	1	0	1	0
2	1	1	1	0	0	1



Model - Offline Test

offline test

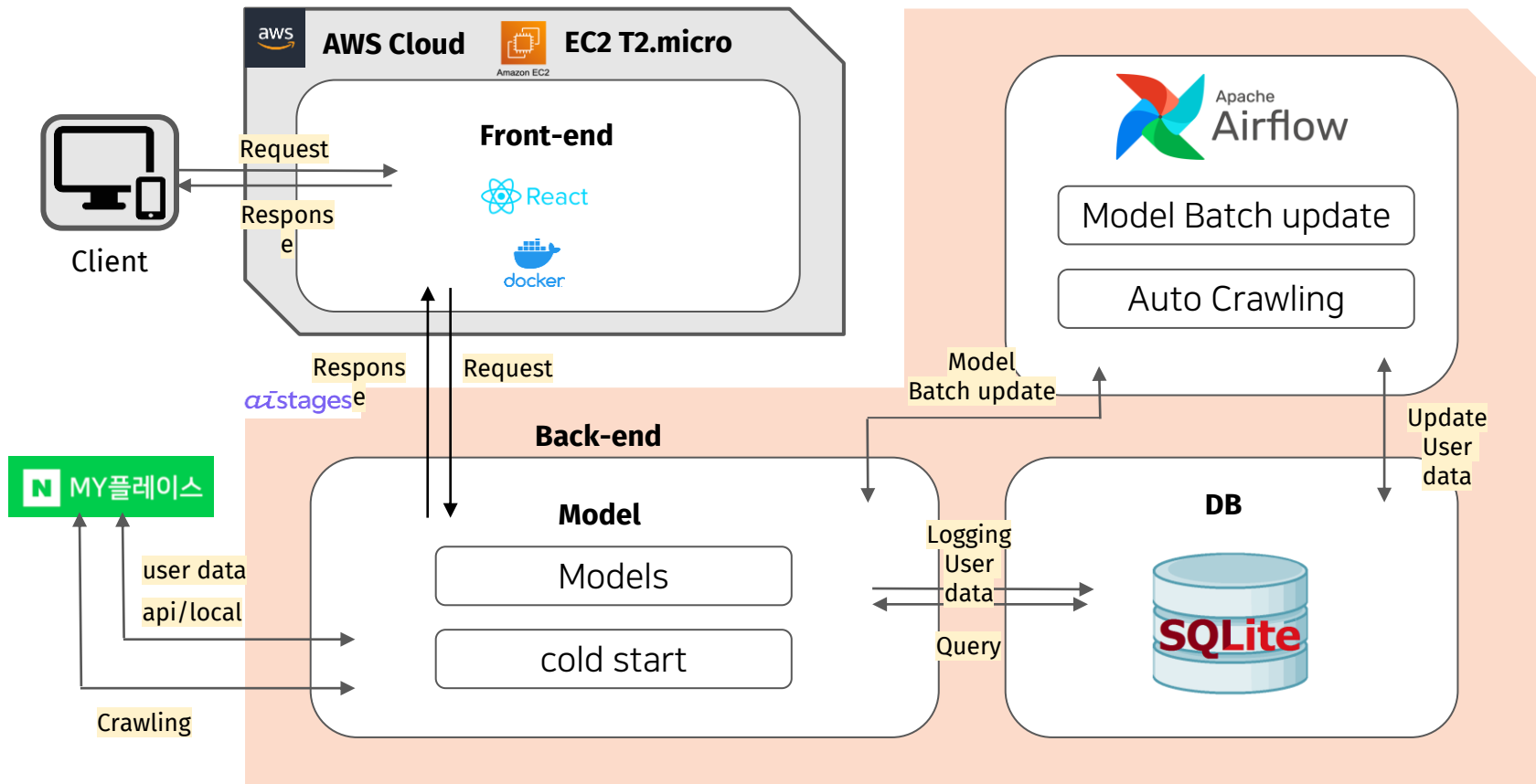


recall@20			Personalization
model	rand	time	
SASRec	5.65%	5.96%	0.00669
MultiVAE	11.23%	10.02%	0.00253
EASE	29.10%	24.29%	0.00334
⋮			
단순 random	0.01%	0.01%	
단순 인기도	0.03%	0.03%	

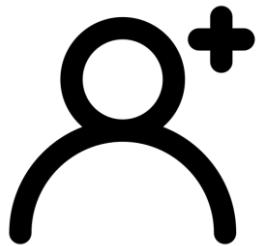
3

Product Serving

Product Serving - Architecture



Product Serving - User Scenario



새로운 유저



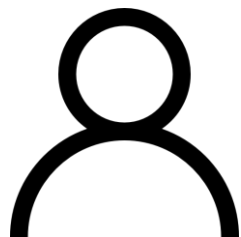
음식 호불호 조사



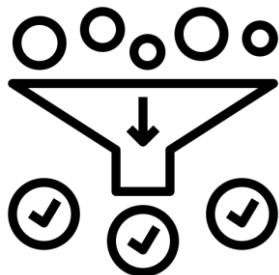
인기도 기반 모델 사용



추천



기존 유저

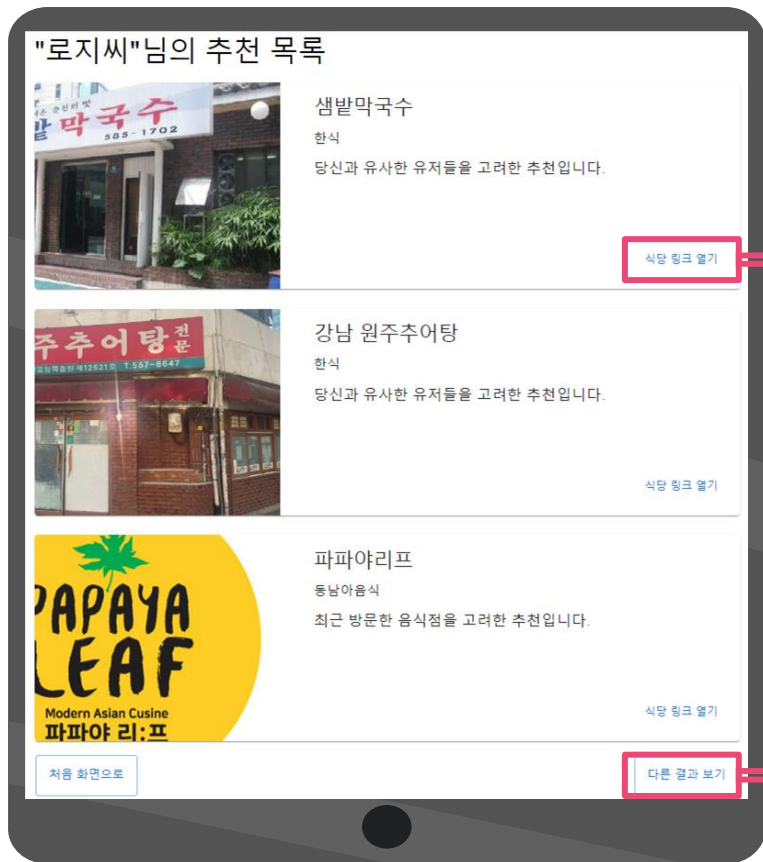


모델



추천

Product Serving - A/B test

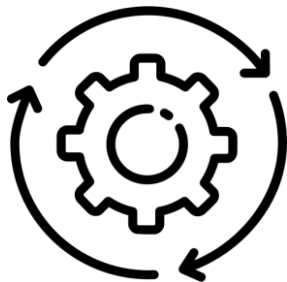


긍정 feedback



부정 feedback

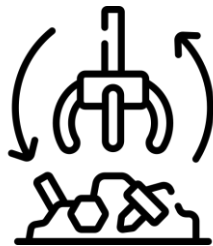
Product Serving - Airflow



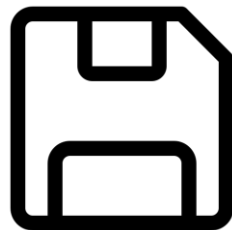
크롤링 자동화



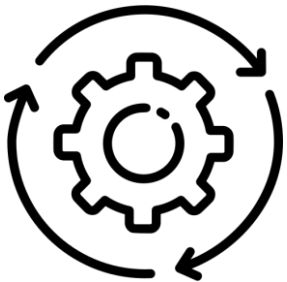
일일 업데이트



유저 정보 크롤링



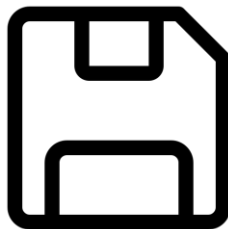
유저 정보 저장



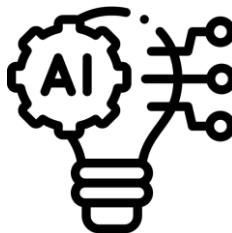
모델 학습 자동화



사용자 사용 정보

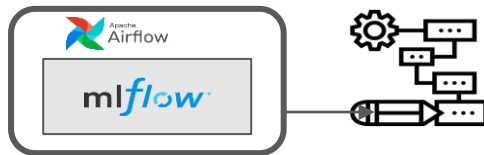


사용 정보 저장



모델 재학습

Product Serving - MLflow



파라미터 및 평가지표 기록

학습 시점 기록

학습 결과 모델 파일 (*.pt / *.pkl 등) 기록 및 저장

<input type="checkbox"/>	Run Name	Created	Duration	Source	Models
<input checked="" type="checkbox"/>	sassy-squid-596	22 hours ago	2.6min	batch_modeling.py	pytorch
<input type="checkbox"/>	adventurous-ant-699	1 day ago	18.6min	batch_modeling.py	-
<input type="checkbox"/>	capable-auk-759	1 day ago	9.8min	batch_modeling.py	sasrec/3
<input type="checkbox"/>	capable-auk-759	1 day ago	9.8min	batch_modeling.py	sasrec/2
<input type="checkbox"/>	capable-auk-759	1 day ago	13.4min	batch_modeling.py	sasrec/1

Parameters (6)	
Name	Value
multivae_batch_size	500
multivae_data_type	time
multivae_dropout_rate	0.5
multivae_num_epochs	1
multivae_p_dims	[100, 400]
multivae_top_K	20
Metrics (1)	
Name	Value
multivae_recall	0.08

4

Result

사용자 위치 기반 식당 추천 서비스 in Seoul

설명 가능한 추천

어떤 이유에서 추천된
음식점인지 사용자에게 설명

모델 개선 자동화

Airflow를 통해
주기적인 크롤링과
모델 학습하는 환경 구축

개인화 추천

회원가입 없이 네이버 플레이스 리뷰 내
역만으로 개인 맞춤형 추천

개선 사항 (발전 계획, 방향)










온라인 테스트 결과를 이용한 오프라인 테스트 평가와 모델 고도화



서울 이외의 지역에 확대 적용

Appendix

목차

	팀원 역할
	협업 방식
	System Spec
	Data Crawling
	MLOps
	Model
	설문조사 결과

팀원 역할



김성연

모델링

데이터베이스(SQLite)

데이터 전처리

metric 정의

Airflow



배성재

데이터 크롤링

프론트엔드(React)

서비스 배포

식당 좌표 수집

MLflow



양승훈

모델링

백엔드(FastAPI)

서비스 배포

MLflow

Airflow



조수연

모델링

백엔드(FastAPI)

PPT



홍재형

데이터 크롤링

데이터베이스(SQLite)

프론트엔드(React)

백엔드(FastAPI)

Airflow



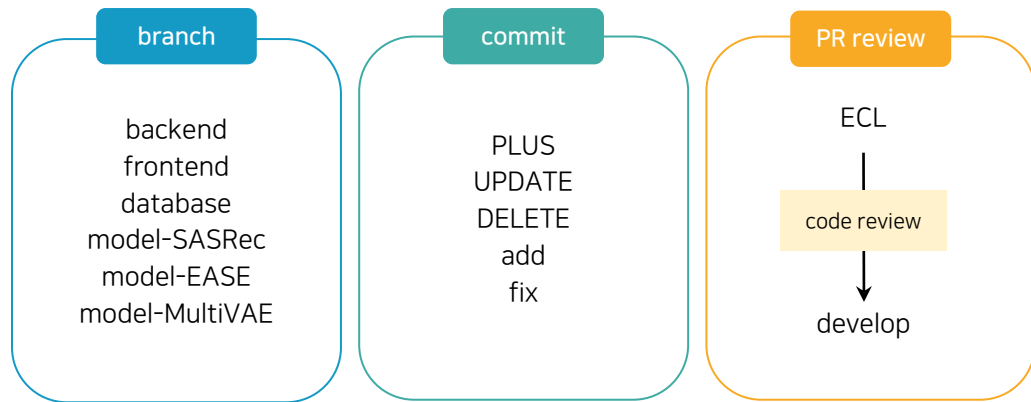
항선태

모델링

프론트엔드(React)

발표

협업 방식 - git, slack, notion



진행 사항, 할 일, 수정 사항 등
팀원 간 공유



프로젝트 회의록, 아이디어, 회고,
모델 성능 등 기록

협업방식 - Zep, CI/CD

회의를 2D 캐릭터 기반의 메타버스 플랫폼인 Zep에서 진행했습니다.



Black 23.1.0
documentation



GitHub Actions



코드 포매팅 도구인 Black 을 Github Action을 이용하여 일관성 있는 코드를 작성하였습니다.

협업-Poetry

pyproject.toml

```
[tool.poetry.dependencies]
python = ">=3.8,<3.12"
uvicorn = "^0.20.0"
fastapi = "^0.89.1"
pillow = "^9.4.0"
torch = "^1.13.1"
pandas = "^1.5.3"
scipy = "^1.10.0"
tqdm = "^4.64.1"
scikit-learn = "^1.2.1"
beautifulsoup4 = "^4.11.2"
requests = "^2.28.2"
pytorch-lightning = "^1.9.0"
```

❌ conda 가상환경에 beautifulsoup4 설치하면서
모듈 간 충돌 에러 발생

💡 이를 해결하기 위해, 의존성을 관리해주는 Poetry 활용




`poetry add` 를 통해 새로운 모듈을 설치하면,
자동으로 기존 모듈과의 의존성 check

`poetry export` 를 통해 [requirements.txt] 파일로 추출할
수 있어,
poetry를 사용하지 않는 팀원들과 협업 가능

requirements.txt

```
beautifulsoup4==4.11.2 ; python_version >= "3.8" and python_version < "3.12"
certifi==2022.12.7 ; python_version >= "3.8" and python_version < "3.12"
charset-normalizer==3.0.1 ; python_version >= "3.8" and python_version < "3.12"
click==8.1.3 ; python_version >= "3.8" and python_version < "3.12"
colorama==0.4.6 ; python_version >= "3.8" and python_version < "3.12" and platform_system == "Windows"
fastapi==0.89.1 ; python_version >= "3.8" and python_version < "3.12"
h11==0.14.0 ; python_version >= "3.8" and python_version < "3.12"
idna==3.4 ; python_version >= "3.8" and python_version < "3.12"
jinja2==3.1.2 ; python_version >= "3.8" and python_version < "3.12"
```

System Spec

 Database	<ul style="list-style-type: none">- SQLite	Backend 서버 내부에 [*.db] 파일로 존재
 Backend	<ul style="list-style-type: none">- Server : Nvidia Tesla v100- Web Framework : FastAPI	DB <-> Frontend 간 통신 & Airflow / MLflow 기능 구현
 Frontend	<ul style="list-style-type: none">- Server : Amazon EC2 (t2.micro)- Docker Container- Frontend Library : React.js	EC2 서버 내부에 Docker 설치 node.js 이미지 기반 Container 실행

Data Crawling

네이버 MY플레이스 리뷰 데이터 선정이유

- 영수증을 인증한 실제 고객들의 데이터라서 타 사이트 리뷰보다 신뢰도가 높다고 판단
- 국내 리뷰 데이터 중 제일 많은 데이터를 가지고 있다.

뉴스홈 | 최신기사

네이버 마이플레이스 "올해 840만명이 리뷰 2억건 작성"

송고시간 | 2022-12-21 16:27

×

**영수증, 처음부터 끝까지
잘 나오게 찍어보세요**
상세 내역까지 기록해 드립니다.

소소식당 영수증

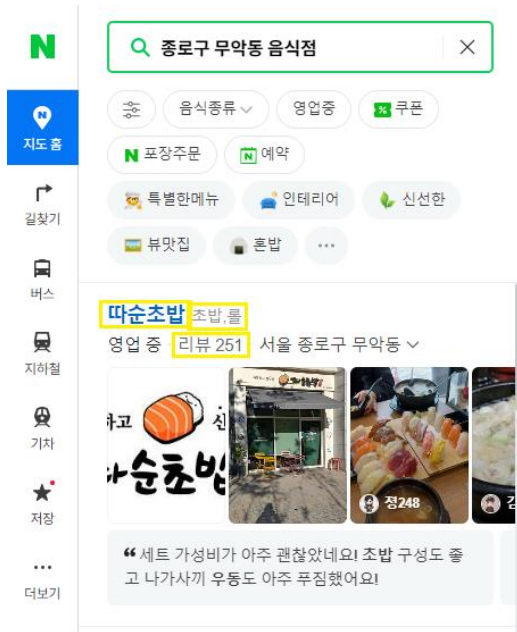
매장명	소소식당	
전화번호	010-XXX-XXXX	
사업자번호	000-00-0000	
주소	성남시 분당구 불정로 6	
매출일	2019.08.30	

상품명	수량	금액
매콤 카레밥	2	22,000
오리엔탈 새우 샐러드	1	13,500
치즈 마쉬룸 오믈렛	3	35,000

합계금액 **70,500**

확인

Data Crawling



1.Kcrawling_rest_server

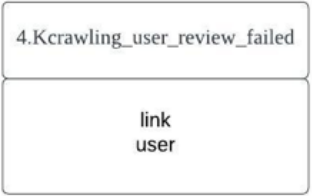
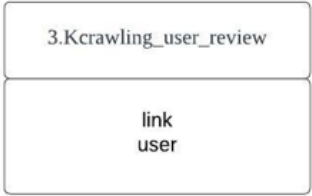
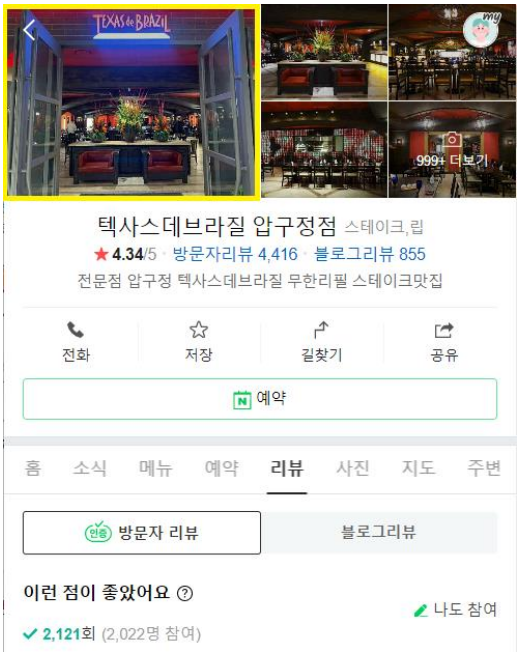
restaurant
tag
url
review

2.Kcrawling_concat

restaurant
tag
url
review
len
rating
count

네이버 지도에서 음식점을 검색한 화면에서
음식점이름, 음식점 종류, 리뷰개수, 마이플레이스 링크를 가져옵니다.
페이지를 넘기면서 정보를 가져와야 했기에 셀레니움을 이용해서 크롤링을
진행했습니다.

Data Crawling



마이플레이스 링크에는 사용자가 남긴 리뷰와 음식점 정보가 있습니다.

사용자이름, 리뷰내용, 방문일자, 방문횟수, 음식점 대표 이미지를 수집합니다.

스크롤을 내려야 리뷰를 볼 수 있어서 셀레니움을 이용했습니다.

Data Crawling

NAVER Developers

Products

Documents

Application

NAVER D2

Support

Forum

API 상태

Search Here

로그인

블로그

뉴스

책

성인 검색어 판별

백과사전

영화

카페글

지식iN

지역

오타변환

웹문서

이미지

쇼핑

전문자료

개요

검색 API와 블로그 검색 개요

검색 API는 네이버 검색 결과를 뉴스, 백과사전, 블로그, 쇼핑, 영화, 웹 문서, 전문정보, 지식iN, 책, 카페글 등 분야별로 볼 수 있는 API입니다. 그 외에 지역 검색 결과와 성인 검색어 판별 기능, 오타 변환 기능을 제공합니다.

블로그 검색은 검색 API를 사용해 네이버 검색의 블로그 검색 결과를 반환하는 RESTful API입니다. 블로그 검색 결과를 XML 형식 또는 JSON 형식으로 반환합니다. API를 호출할 때는 검색어와 검색 조건을 쿼리 스트링(Query String) 형식의 데이터로 전달합니다.

블로그 검색은 검색 API를 사용하며, 검색 API의 하루 호출 한도는 25,000회입니다.

검색 API 특징

검색 API는 비로그인 방식 오픈 API입니다.

비로그인 방식 오픈 API는 네이버 오픈API를 호출할 때 HTTP 요청 헤더에 클라이언트 아이디와 클라이언트 시크릿 값만 전송해 사용할 수 있는 오픈 API입니다. 클라이언트 아이디와 클라이언트 시크릿은 네이버 오픈API에서 인증된 사용자인지 확인하는 수단입니다. [네이버 개발자 센터](#)에서 애플리케이션을 등록하면 클라이언트 아이디와 클라이언트 시크릿이 발급됩니다.

참고

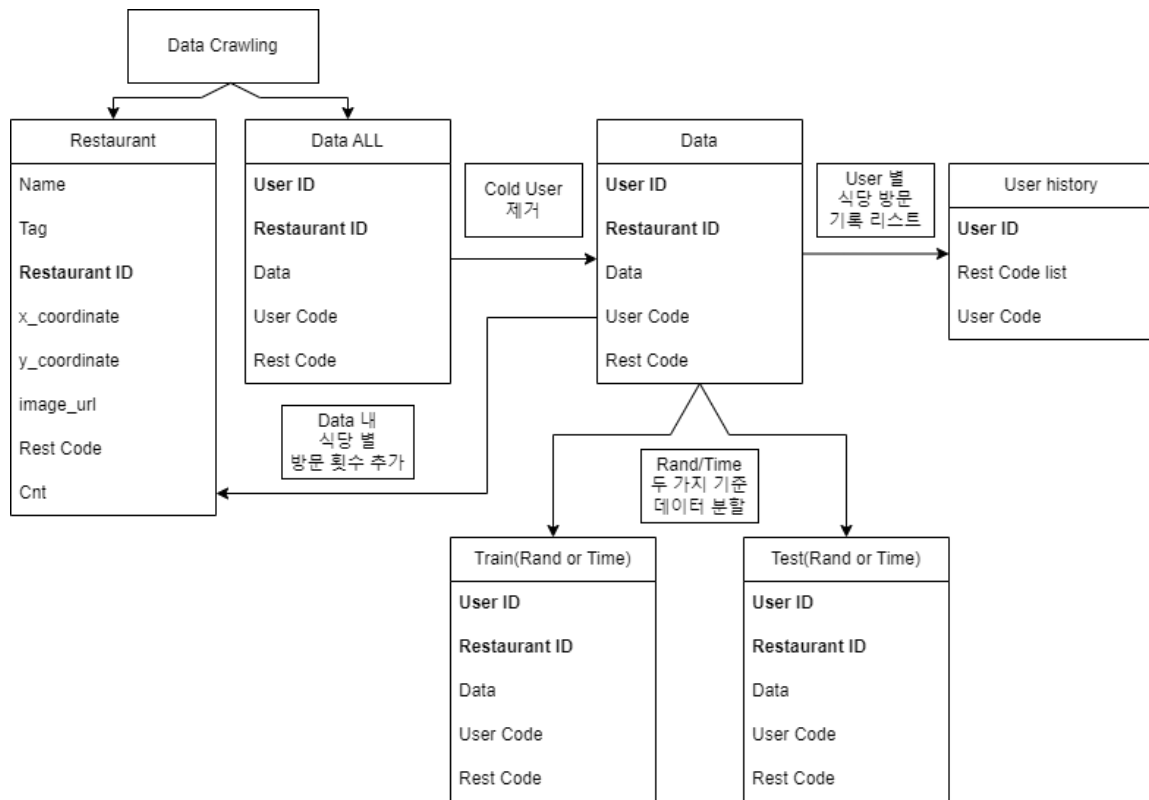
네이버 오픈API의 종류와 클라이언트 아이디, 클라이언트 시크릿에 관한 자세한 내용은 "[API 공통 가이드](#)"를 참고하십시오.

5.Kcrawling_get_rest_info

x
y
image

음식점의 x,y 좌표를 얻기위해 NAVER Developers에서 지원해주는 api를 활용했습니다.

Data Crawling

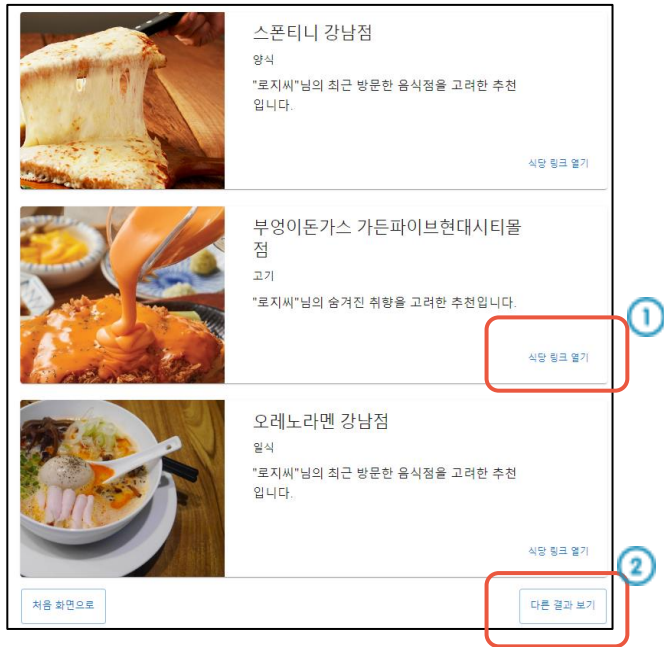


유저와 음식점 단위로 크롤링을 진행한 뒤 Cold 유저를 제거한 Data 테이블을 제작합니다.

이를 통해 User history 테이블과 모델 학습에 필요한 Train, Test 테이블이 제작됩니다.

백엔드에서 모델 서빙하는 과정에서는 Restaurant 테이블과 User history 테이블만 사용하게 됩니다.

MLOps



① “식당 링크 열기” 버튼 Click

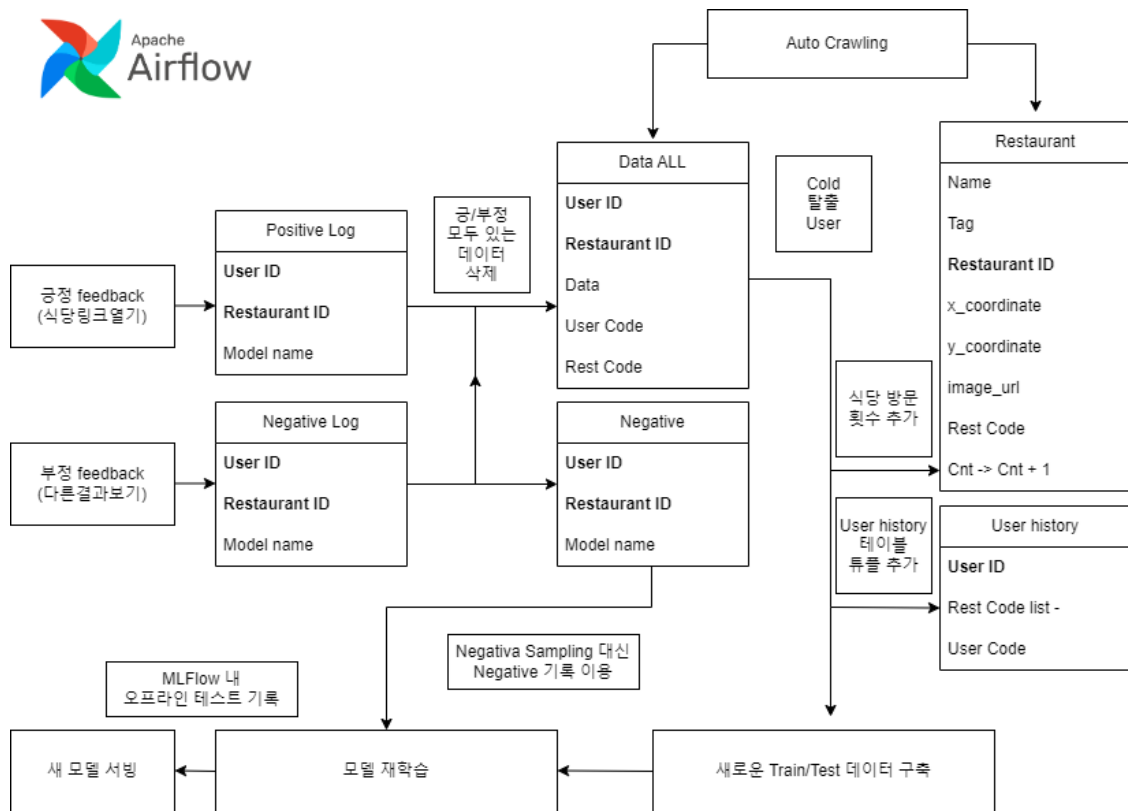
- 현재 유저의 해당 식당에 대한 Positive Feedback으로 설정
- DB의 “positive” 테이블에 (유저 / 식당 / 모델) 데이터 추가

16	6130db4973adbe125329a3e4	1,217,232,435	ease
17	5b62e55e10599e4793d54f8a	1,683,909,619	ease
18	5c9eb9a4ba69d2b1f5c55cbe	38,272,988	sasrec

② “다른 결과 보기” 버튼 Click

- 화면에 보이는 3개의 식당 모두 Negative Feedback으로 설정
- DB의 “negative” 테이블에 (유저 / 식당 / 모델) 데이터 추가

MLOps



영수증 리뷰와 Positive
반응이 5개 이상인 유저는
Cold Start를 탈출합니다.

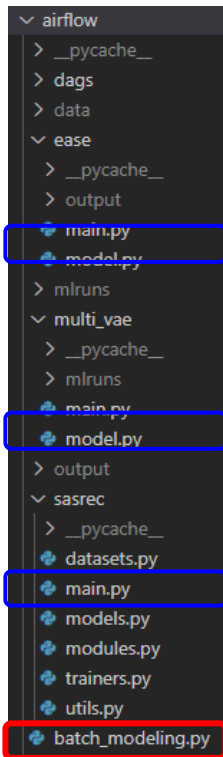
3일마다 데이터를 크롤링
해서 최신 데이터를 수집
합니다.

16시~17시, 03시~04시
점심 이후, 저녁 이후
시간대에 모델을
재학습합니다.

Positive, Negative Log 내
모델 별 비율을 통해
A/B Test를 진행합니다.

MLOps

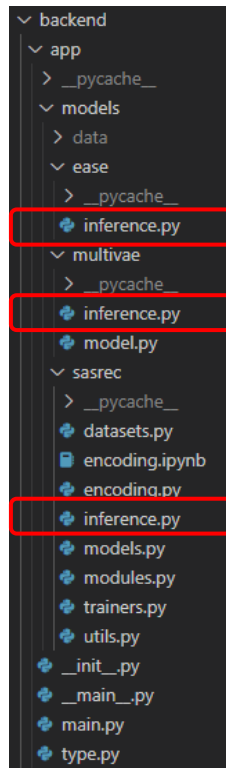
<Airflow>



- Airflow를 통한 주기적 batch 학습
- [airflow/batch_modeling.py] 실행 시, 각 모델의 [main.py]가 실행되면서 학습 진행
- SASRec / MultiVAE -> model.pt 파일 생성
- EASE -> model.pkl 파일 생성
- 생성된 .pt / .pkl 파일 [backend/app/models/data/] 경로로 저장

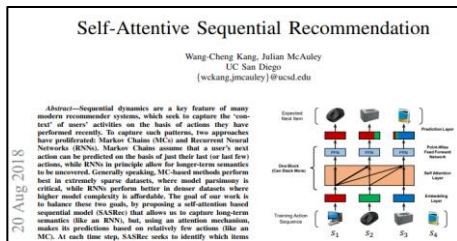


<Backend>



- 사용자의 추천 요청 발생
- 각 모델의 [inference.py] 실행
- 각 모델에 해당하는 .pt 또는 .pkl 파일 로드
- Top K개 음식점 추천 결과 도출 후 사용자에게 전달

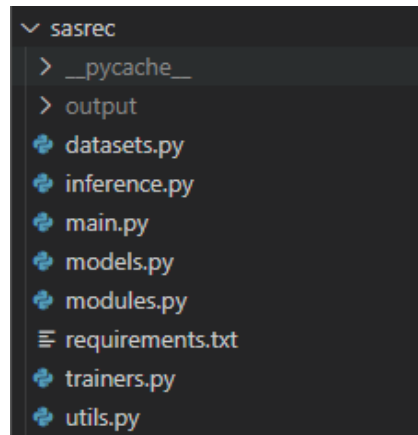
Model - SASRec



<https://arxiv.org/pdf/1808.09781.pdf>

<Hyperparameter>

- hidden_size = 500
- num_hidden_layers = 2
- num_attention_heads = 2
- attention_probs_dropout_prob = 0.2
- hidden_dropout_prob = 0.3
- max_seq_length = 150
- learning_rate = 0.001
- epochs = 5



(model/sasrec 폴더)

- datasets.py : 데이터 셋 가공후 데이터 로더로 만드는 부분
- inference.py : 학습 된 pt 파일을 이용해 output을 뽑아냄
- main.py : 메인 함수 (python main.py 로 실행 가능)
- models.py : SASRec 메인 모델이 들어가 있음.
- modules.py : SASRec 모델을 위한 보조 도구가 있는 부분.
- requirements.txt : 라이브러리(pip install -r requirements.txt)
- trainers.py : 학습 진행과 Recall 값을 뽑아주는 부분
- utils.py : 간단한 함수들의 모음

Model - EASE



<https://arxiv.org/pdf/1905.03375.pdf>

<Hyperparameter>

Lambda = 500

```
self.B = np.float16(B)
self.pred = X.dot(B)
```

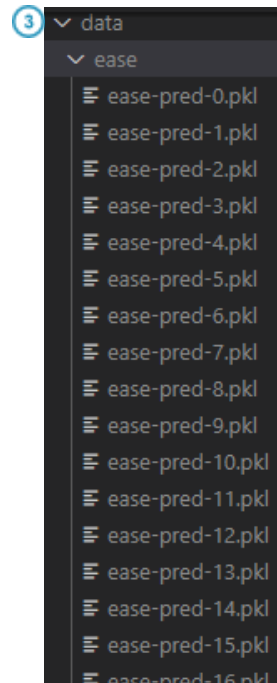
(382940, 41456) * (41456, 41456) 차원의 행렬 연산 과정에서 RAM 용량 부족으로 Memory Error 발생

```
1 X_cur = model.X[ start : end ]
pred_cur = X_cur.dot(model.B)
pred_cur = np.float16(pred_cur) 2 ## 용량 줄이기
if data_type == 'time':
    with open(pkl_path + f'ease/ease-pred-{i}.pkl', 'wb') as f:
        pickle.dump(pred_cur, f, pickle.HIGHEST_PROTOCOL) 3
```

- 1 X 행렬을 분해해서 B 행렬과 행렬곱 하는 방식으로 연산
- 2 자료형을 float32에서 float16으로 변환해서 용량 최소화
- 3 .pkl 파일을 여러 개로 쪼개서 저장 (0 ~ 191까지 총 192개의 .pkl 파일)

```
''' load pred matrix '''
ith_file, user_idx = divmod( user, args.thres )
4 with open(args.data_dir + f'ease/ease-pred-{ith_file}.pkl', 'rb') as f:
    pred = pickle.load(f)
    pred = pred[user_idx]
    .....
```

- 4 inference.py에서 추천 결과 계산할 때, 해당 유저 번호가 속하는 .pkl 파일만 불러와서 계산



Model - MultiVAE

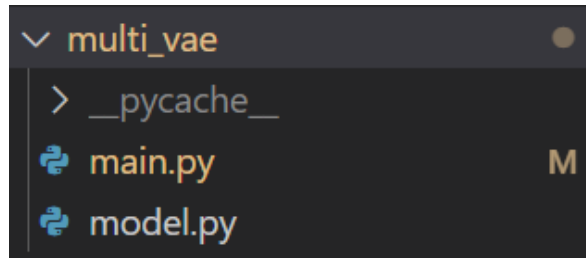


<https://arxiv.org/pdf/1802.05814.pdf>

<Hyperparameter>

- p_dims = [100, 400]
- dropout_rate = 0.5
- weight_decay = 0.01
- anneal_cap = 0.2
- total_anneal_steps = 200000
- learning_rate = 0.005
- batch_size = 500
- epochs = 100

(model/multi_vae 폴더)



main.py

데이터셋 불러오기
학습 진행
평가 진행

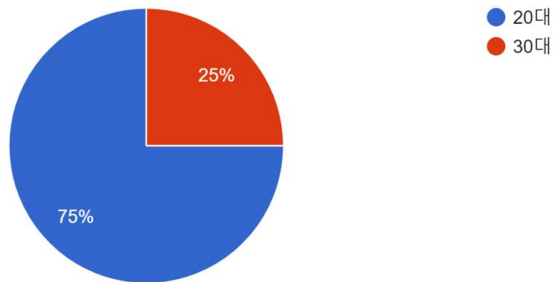
model.py

데이터셋 가공
모델 클래스 정의
손실함수 정의

설문조사 결과 - 개인정보

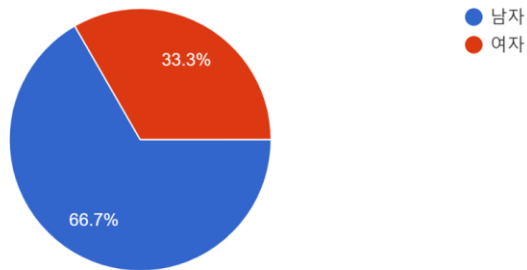
당신의 나이는?

응답 12개



당신의 성별은?

응답 12개



부스트캠프 내 서비스를 이용한 이용자 12명을

대상으로 한 설문조사입니다.

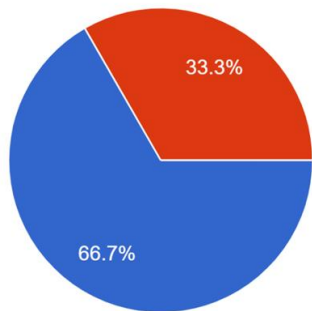
20대, 30대에 젊은 세대 시각을 담았으며

남성의 비율이 다소 높습니다.

설문조사 결과 - 메뉴 고민 관련

평소 식사 메뉴에 대해 고민 많이 하시나요?

응답 12개



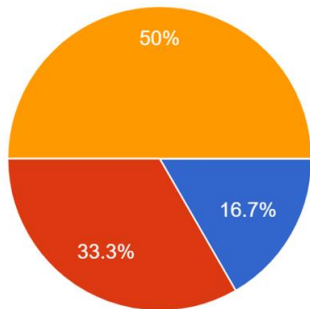
- 고민이 많습니다.
- 보통입니다.
- 메뉴 고민을 오래하지 않습니다.

식사 메뉴에 대해 대부분 고민을 많이하며 메뉴 고민을 오래하지 않는 사람은 없었습니다.

중립 장소에서 진행하는 모임이 있을 때 아무거나 괜찮다는 사람이 절반, 주도적으로 메뉴를 정하는 사람은 드물었습니다.

중립 장소에서 진행하는 모임이 있을 때 어떻게 메뉴를 정하시나요?

응답 12개



- 주도적으로 메뉴를 정하는 스타일입니다.
- 먹고 싶거나 먹기 싫은 메뉴에 대해 가볍게 의견제시를 합니다.
- 아무거나 괜찮다고 얘기합니다.

=> 2030 세대는 음식 메뉴를 정하는데 어려움이 있다고 볼 수 있습니다.

설문조사 결과 - 사용 만족도 조사

메뉴 추천 서비스 "먹을끼니" 만족하시나요?

응답 11개

넵

조금 아쉽습니다.

네

필요한 서비스이지만 개선이 필요합니다

네 만족합니다! 적당한 추천 개수와 네이버 사이트로 바로 연결이 되어서 정보 확인하기도 쉽고 좋은 것 같아요! ㅎㅎ

아쉬운점이 있지만 만족스럽습니다.

너무 훌륭합니다

4

개쩔탱이네요

설문조사 결과 - 개선사항

메뉴 추천 서비스 "먹을끼니"의 개선 사항을 적어주세요!

응답 11개

잘 모르겠습니다

사용법 사진이 도움말 속 말고 밖으로 꺼내져 있으면 좋을거 같아용 ! 부산도 서비스 부탁드립니다 ~!

마이플레이스에 리뷰를 남긴 적이 없어서 그런지 제가 좋아할만한 식당을 추천해준다는 느낌보다는 처음 싫어하는 카테고리를 제외하고 보여준 느낌이었습니다. '어떻게 사용하나요?'와 같은 친절한 설명은 좋았고, 아이디어 자체는 정말 필요한 주제인 것 같아요!

음..Link갖고오도록 하는 부분에서 사이트를 갖다가 복사하고 뒤로가기를 누르는 것보단.. 조금 사용자 사용측면에서 거치는 부분이 줄었으면 좋겠습니다. 안내페이지도 UI부분 더 이뻐지면 좋겠네용

1. 네이버 마이플레이스를 원래 사용하지 않는 입장에서, 서비스에 접근하고 이용하는 방식이 불편하고 어려웠습니다. 쉽게 로그인을 할 수 있었으면 좋겠습니다.
2. 싫어요 음식이 애매합니다. 양식/중식/한식 같은 대분류를 의미하는건지, 햄버거/자장면/한정식 같은 단품을 의미하는지 와닿지 않았습니다.
3. 싫어요 항목의 이미지들이 다들 너무 맛있게 생겼습니다. 싫어요를 누름에 어려움이 있습니다.
4. A/B테스트이거나, 아니면 마이플레이스 이용 내역이 있어야 정말 추천이 되는 것인지는 모르겠지만 서비스 체험 입장에서는 추천받은 가게들이 주로 유명 체인점 위주로 나오는 것 같아서 아쉬웠습니다.

설문조사 결과 - 개선사항

메뉴 추천 서비스 "먹을끼니"의 개선 사항을 적어주세요!

응답 11개

+ 개인적으로 제가 하고 싶었던 주제였는데 이렇게 보니 정말 반가웠습니다! 화이팅입니다!!

싫어하는 음식 선택할 때, 3개만 선택해야하는 줄 모르고 더 많이 골랐다가, 3개로 변경하고 확인 누르려니깐 안 눌러지네요! 뒤로 갔다가 다시 싫어하는 음식 선택하는 페이지로 이동해야 확인이 눌러집니다!

유저도 내 로그 정보들을 확인해보고 싶어요~!

경기도도 해주세요

첫 화면에서 제출을 눌렀는데도 추천결과가 안나와요.... 크롬환경입니다.
그리고 식사랑 카페 선택하고 싫어하는 음식을 고르는 부분이 똑같아서 다른 사진을 보여주는게 좋을거 같습니다.

추후에 발표하거나 어떻게 작동하는지 물어볼 여지가 있을 것 같습니다. UI UX적으로 어느정도 해결하긴 해야 하는데, 지금은 시간이 안되니 ㄸㄸ.. 주위 지인들에게 설명을 드리고 실제로 와닿는지 한번 시도해 보면 좋을 것 같습니다. 고생하셨습니다 ㅎㅎ

만족도는 주로 리커트 척도로 1~5점 레이팅 하는 게 편할 것 같습니다.

https://ko.wikipedia.org/wiki/%EB%A6%AC%EC%BB%A4%ED%8A%B8_%EC%B2%99%EB%8F%84