

RECJOON

#백준 #문제추천 #라이벌추천 #알고리즘친구

RECSYS-14
(RECognizer)

boostcamp ai tech

CONTENTS

01 Intro

팀 소개
프로젝트 소개
아키텍처

02 모듈별 소개

데이터 수집
문제 추천
라이벌 추천
라이벌 문제 추천
Product serving

03 Result

결과 및 고찰
appendix

01 Intro

팀 소개

프로젝트 소개

아키텍처



01 Intro / 팀 소개



김은선



T3047



박정규



T3092



이서희



T3150



이선호



T3151



진완혁



T3214

라이벌 추천 모델링

데이터 EDA

라이벌 추천 모델링

Back-end 개발

데이터 수집, EDA

라이벌 문제 추천 모델링

Front-end 개발

라이벌 문제추천 모델링

Front-end 디자인

문제 추천 모델링

태스크 자동화

온오프라인 지표 개발

CI & CD 자동화

문제 추천 모델 전처리

01 Intro / 프로젝트 소개

코딩 테스트 연습해야하는데...
오, BOJ라는 좋은 사이트가
있네??



문제도 많고.. 좋긴 한데..
도대체 뭐부터 풀어야하지..?
내가 풀 만한 문제를
추천해줄 순 없을까?



혼자 하니까 지루하네..
나랑 비슷한 사람들이랑
같이하면 동기부여될 것 같은데..

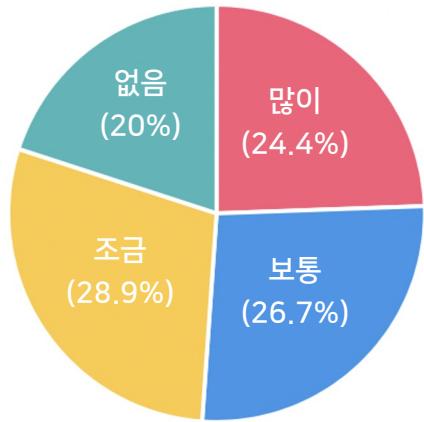


Baekjoon Online Judge 개인별 문제 및 라이벌 추천

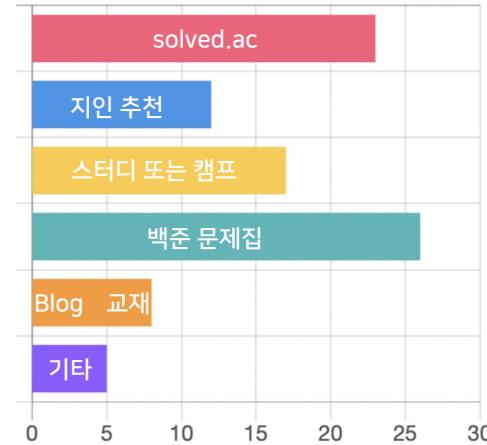
Objective: 알고리즘 학습 동기 부여

01 Intro / 프로젝트 소개

Q. 알고리즘 문제를 선택하는 데 있어서
얼마나 어려움을 느끼나요?

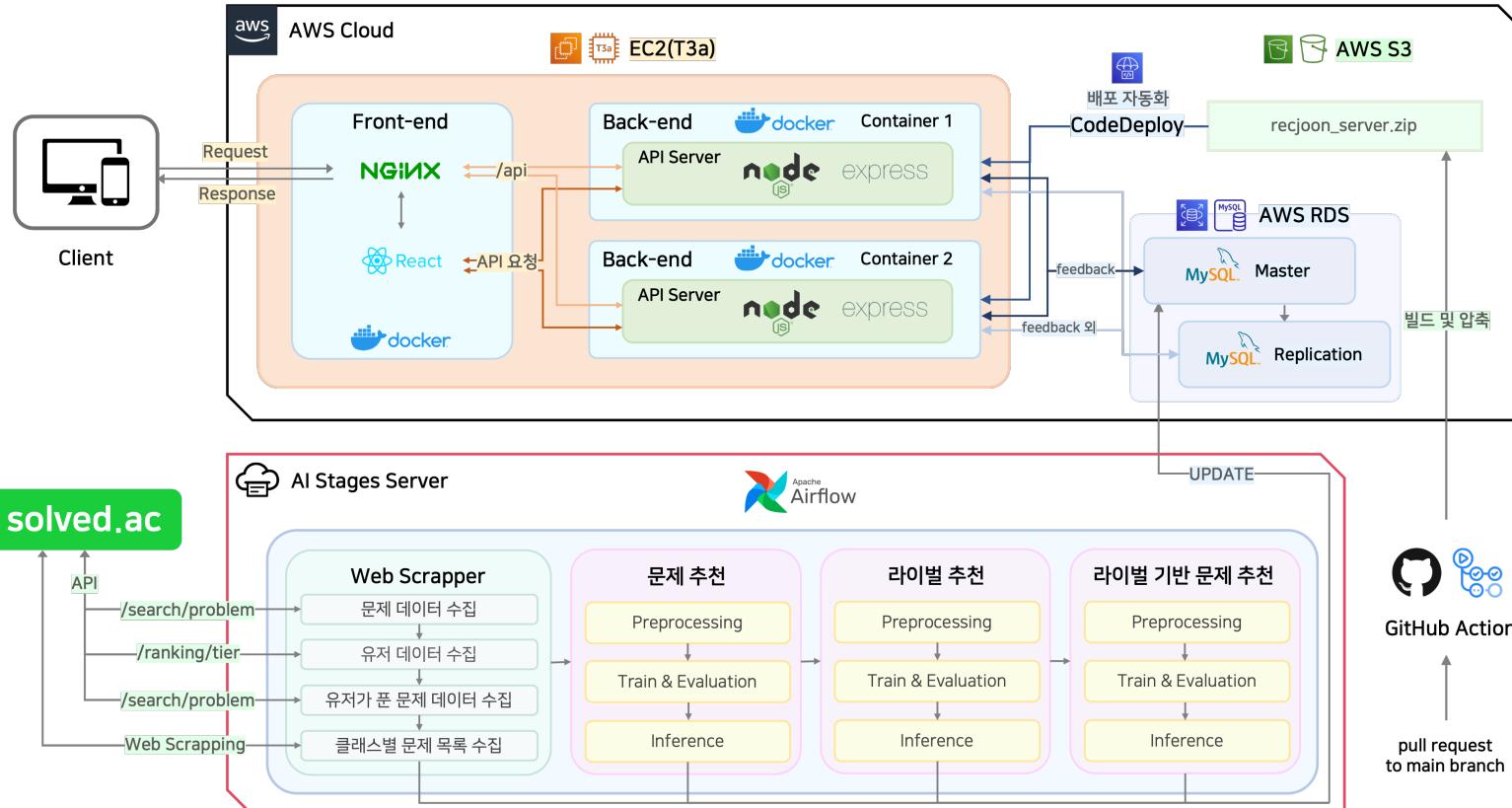


Q. 알고리즘 문제 선택 경로 또는
방법은 무엇인가요?

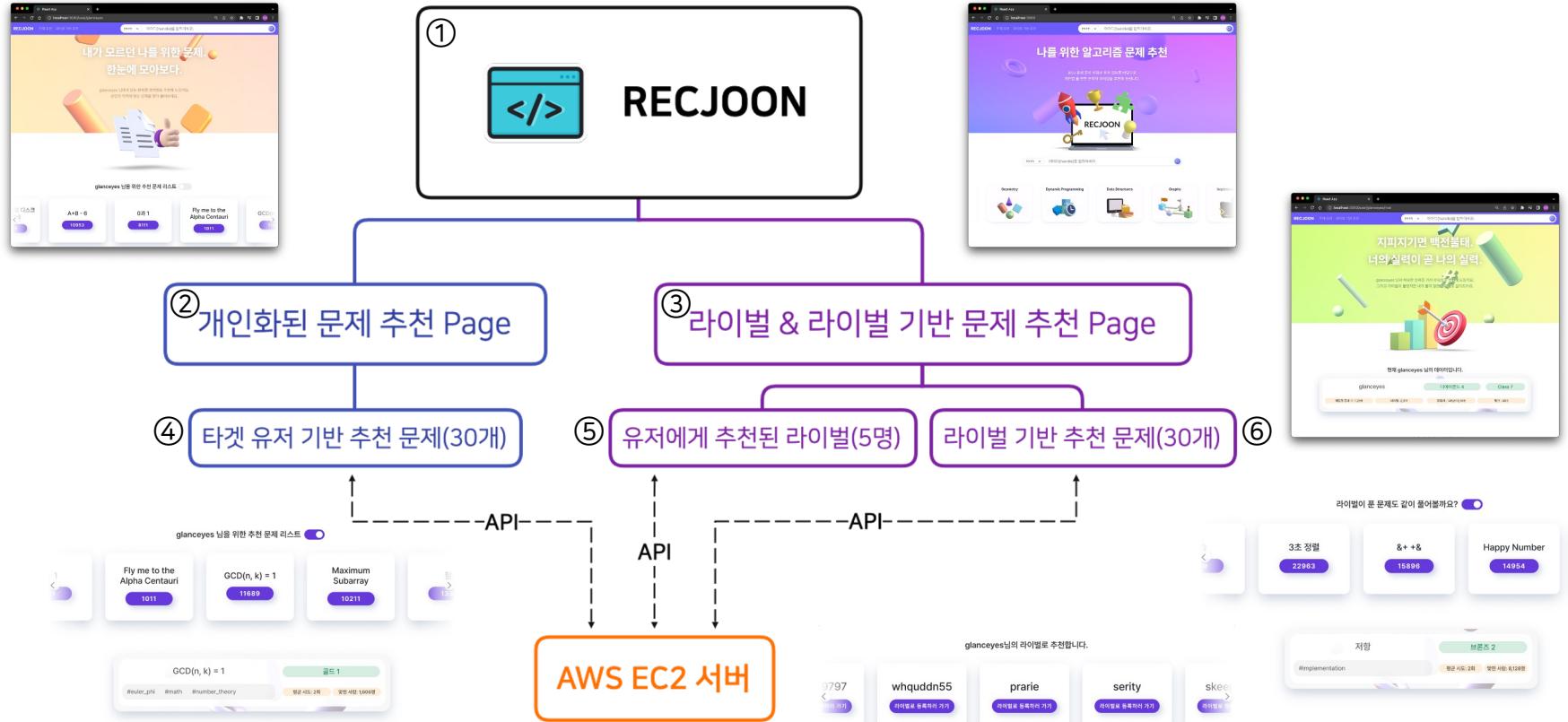


많은 사람들이 경험하는 알고리즘 문제 선택 공감대

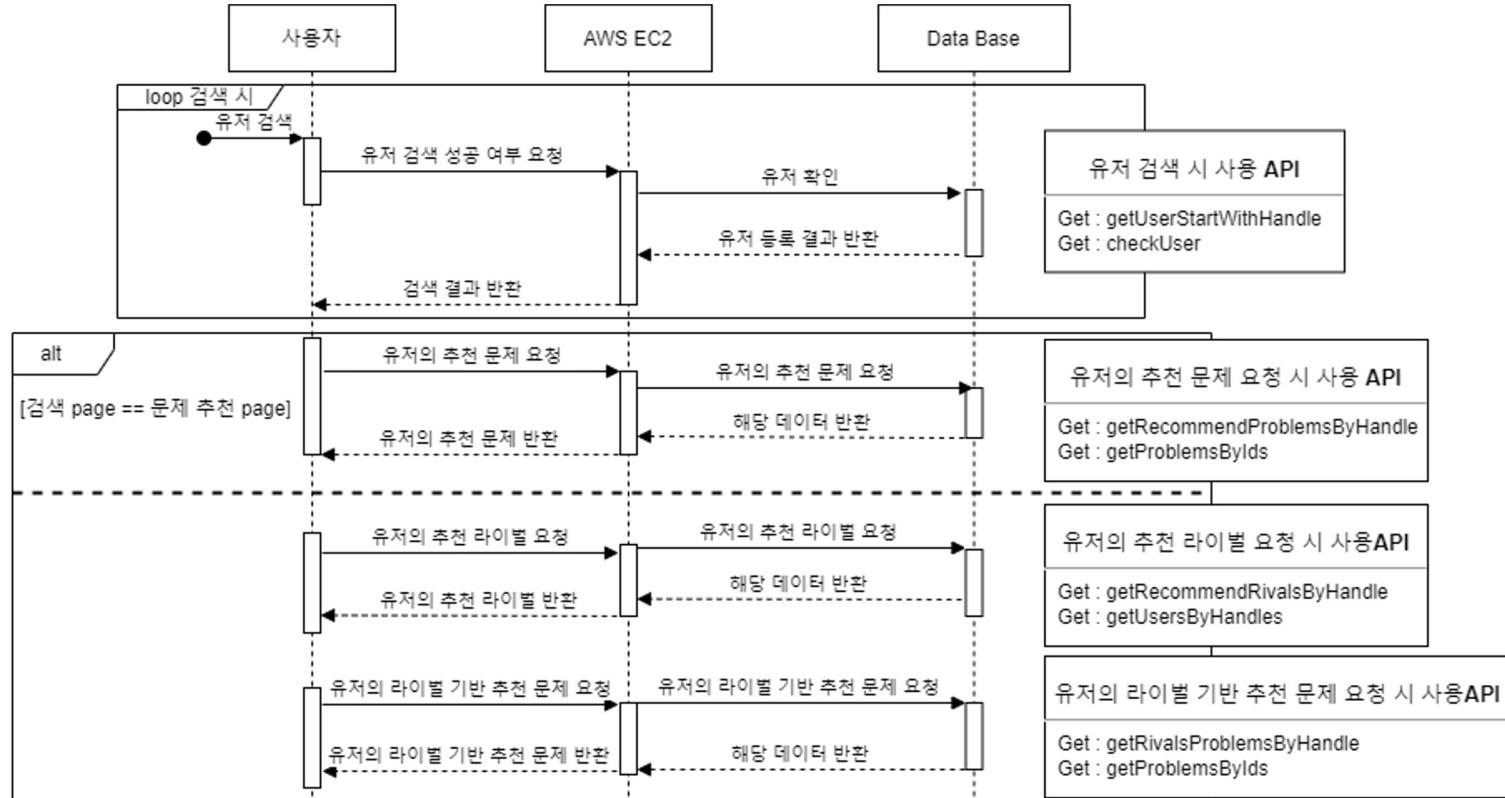
01 Intro / 아키텍처



01 Intro / 사이트맵



01 Intro / UML 시퀀스 다이어그램



01 Intro / 데이터 수집 방법 탐색

방법 1. Baekjoon Online Judge 웹 스크레이핑을 통한 수집? → X

BOJ에서 '채점 현황' 페이지는 웹 스크레이핑에 관하여 Disallow로 명시되어 있는데,
해당 페이지에서 스크레이핑을 하여 데이터를 얻을 수 있나요?

14501	틀렸습니다
9267	컴파일 에러
1000	채점 중 (64%)
9372	컴파일 에러
1003	시간 초과
2206	맞았습니다!!
25083	맞았습니다!!
2577	시간 초과

</>

BOJ는 모든 종류의 웹 스크레이핑을 **원칙적으로 금지하고 있습니다.**

BOJ 운영자에게 문의한 결과



BOJ에서 유저별 문제 풀이 순차 이력인 '채점 현황' 수집 불가

01 Intro / 데이터 수집 방법 탐색

방법 2. solved.ac API를 사용하여 수집 → 0

※ solved.ac 사용 유저에 한해 solved.ac의 데이터는 BOJ 데이터와 동기화됨

solved.ac API를 사용해서 문제와 라이벌 추천을 위한 데이터 수집이 가능할까요?



시간당 API 수행 Limit 안에서 데이터 수집이 가능합니다.

solved.ac 운영자에게 문의한 결과



solved.ac API를 통해서 BOJ 유저와 문제 데이터 수집

02 모듈별 소개

데이터 수집

문제 추천

라이벌 추천

라이벌 기반 문제 추천

Product Serving



02 모듈별 소개 / 데이터 수집

AC

solved.ac API¹⁾ 및 스크래핑을 통해 데이터를 수집

- 유저의 정보(14개의 특성)

사용자명	푼 문제 수	티어	레이팅	랭킹	...	경험치
------	--------	----	-----	----	-----	-----

- 문제의 정보(7개의 특성)

문제 ID	문제 제목	채점 가능 여부	맞은 사람 수	문제 레벨	평균 시도 횟수	태그
-------	-------	----------	---------	-------	----------	----

- 유저별 푼 문제 리스트

문제의 개수 : 약 23,000개
사용자의 수 : 약 65,000명
Interaction 수 : 약 840만개

02 모듈별 소개 / 문제 추천

사용할 추천 모델 종류 선택

Sequential model

유저가 일정 기간동안 순차적으로 어떠한 문제를 풀고 맞거나 틀렸는지를 고려



Temporal feature를 고려할 수 있지만
순차 데이터인 '채점 현황' 데이터 수집 불가

Non-sequential model

유저가 그동안 어떠한 문제를 풀었는지를 고려

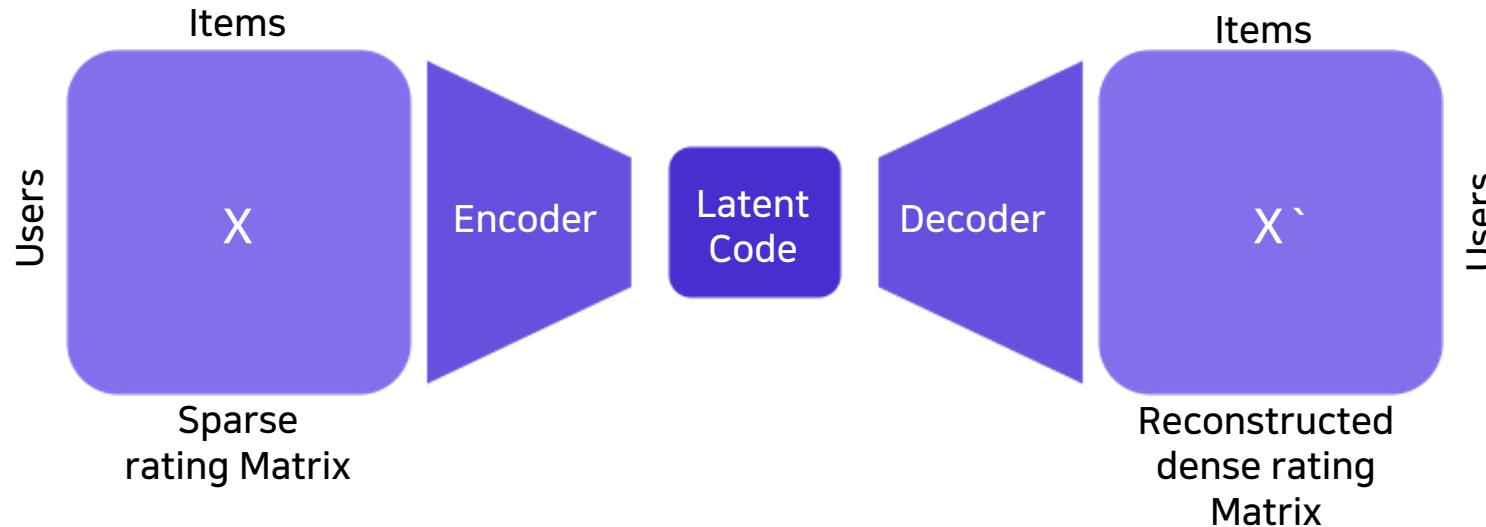


Non-sequential model 선택 및 탐색

02 모듈별 소개 / 문제 추천

주 사용 모델 - RecVAE¹⁾

VAE²⁾를 Interaction Data의 Reconstruction에 활용하여 Matrix Completion으로 접근한 Multi-VAE³⁾에서 더 나은 성능을 구현하기 위해 몇가지 구조를 변형한 모델.



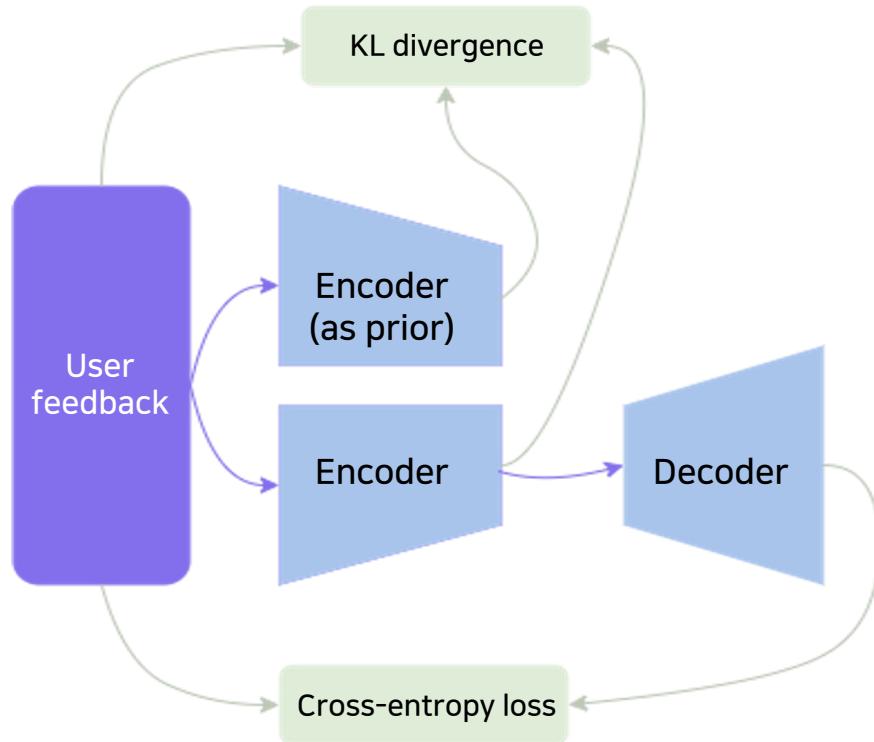
1) 'RecVAE: A New Variational Autoencoder for Top-N Recommendations with Implicit Feedback', 2019

2) 'Auto-Encoding Variational Bayes', Kingma and Welling, 2014

3) 'Variational autoencoders for collaborative filtering', Liang et al., 2018

02 모듈별 소개 / 문제 추천

Training



KL divergence

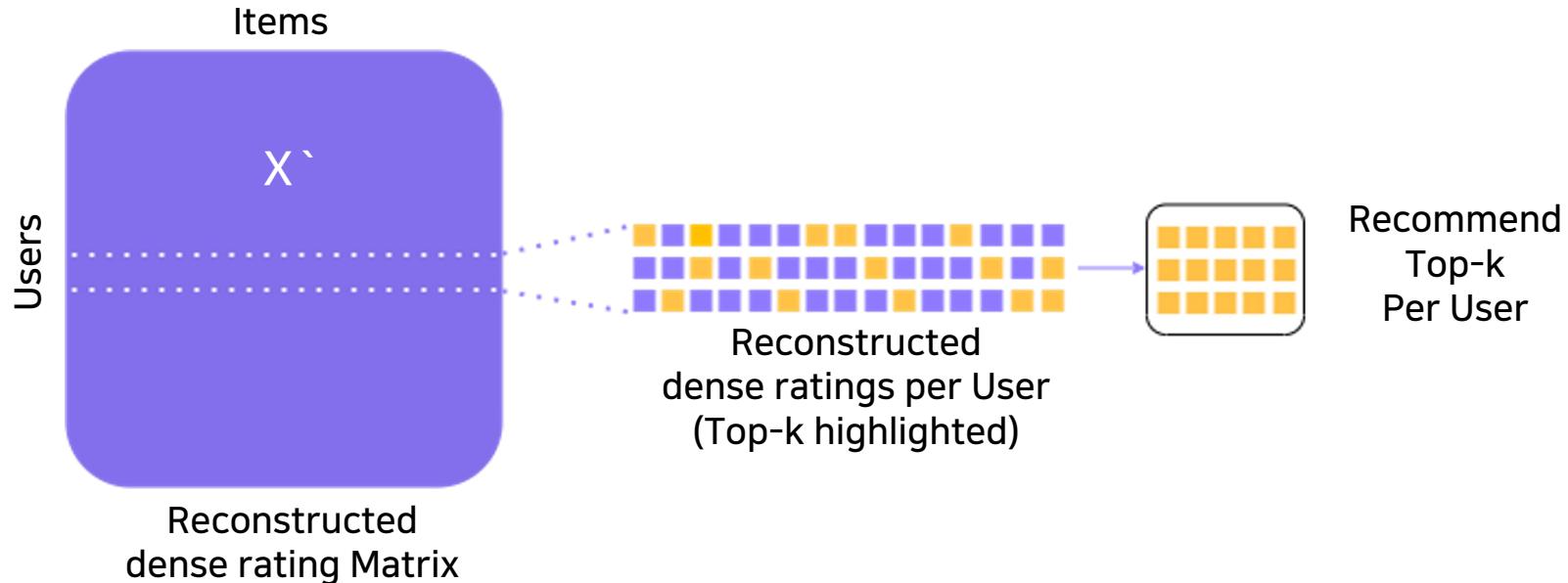
- 새로운 인코더 아키텍처를 구성
- Prior distribution에서 $p(z)$ 대신 $p(z|\phi_{old}, x)$ 사용
- beta - VAE⁴⁾ 논문에서 활용한 beta hyperparameter 적용

Encoder와 Decoder를 번갈아 가면서 학습하는 Alternating Update 적용

(Cross-entropy loss) - (KL divergence)가 최소가 되도록 Training

02 모듈별 소개 / 문제 추천

Inference



02 모듈별 소개 / 문제 추천

Input

5번 이상 문제를 푼 유저와 10번 이상 풀린 문제들만 적용.

	problem_1	problem_2	problem_3	...	problem_m
user_1	1	0	1	...	0
user_2	0	0	0	...	1

사용자의 문제별 rating
1 : 풀었음
0 : 안 풀었음



Output

	problem_1	problem_2	problem_3	...	problem_m
user_1	-inf	3.2243	-inf	...	2.1493
user_2	4.1922	2.8749	-1.2934	...	-inf

모델을 통해 재구성된
rating에서 기존에 풀었던
문제를 제외한 상위 K개의
문제를 추천

02 모듈별 소개 / 문제 추천

AutoEncoder 기반 모델 비교

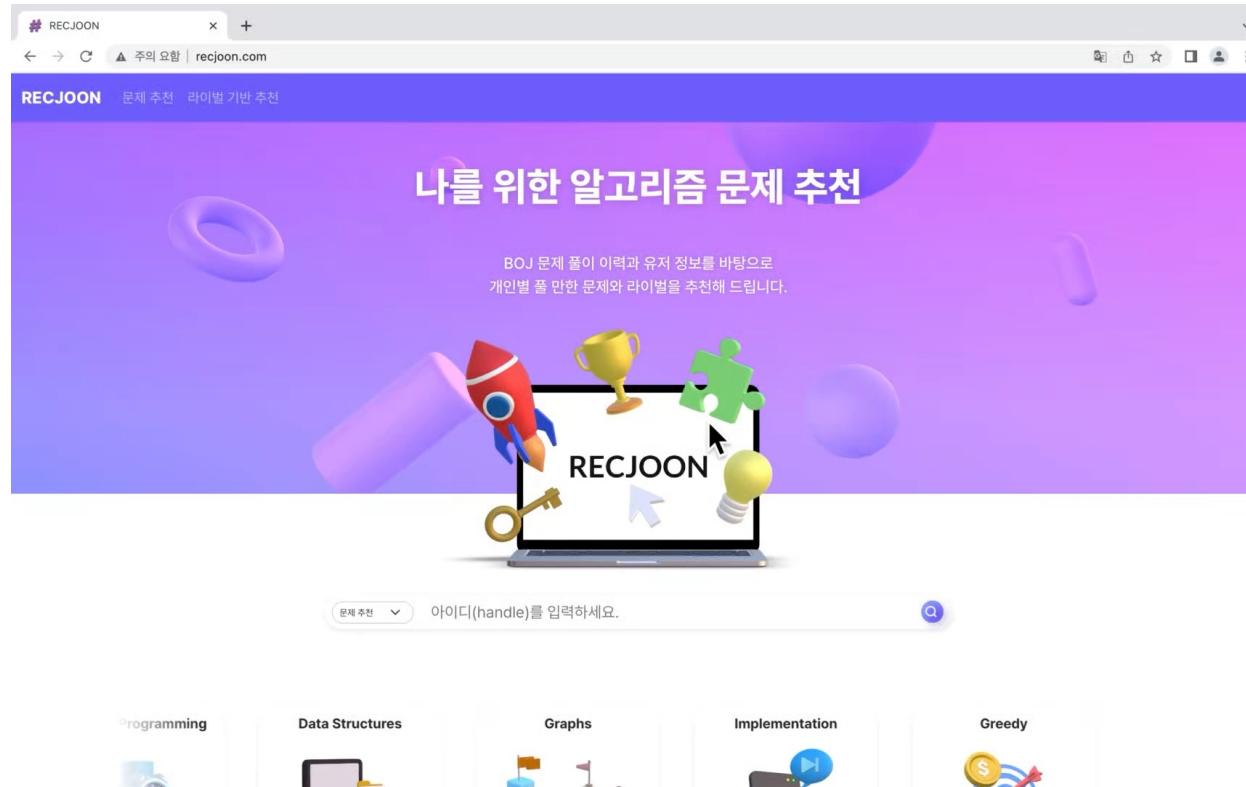
	Multi-DAE	Multi-VAE	RecVAE
NDCG @ 100	0.7270	<u>0.7326</u>	0.7302
Recall @ 10	0.7213	0.7238	<u>0.7283</u>
Recall @ 20	0.7086	0.7108	<u>0.7132</u>

- NDCG @ K : Top-K 추천 리스트의 순서에 가중치를 두어 성능을 평가하는 지표.
- Recall @ K : 유저의 관심 아이템 중 추천된 K개의 비율을 평가하는 지표

추천된 아이템의 순서에 의미를 두지 않기 때문에, Recall의 값이 가장 높은 RecVAE를 주요 사용 모델로 선정.

데이터 업데이트마다 Recall 값에 따라 유동적으로 모델 적용.

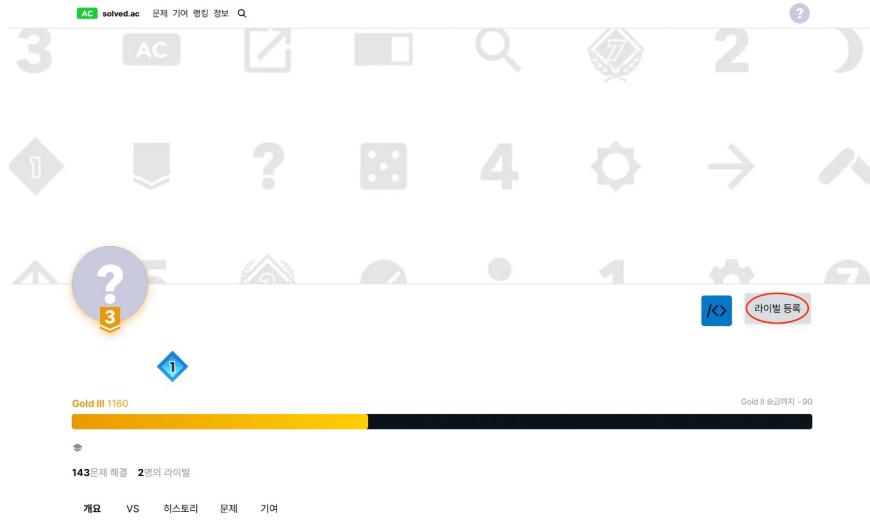
02 모듈별 소개 / 문제 추천 시연 영상



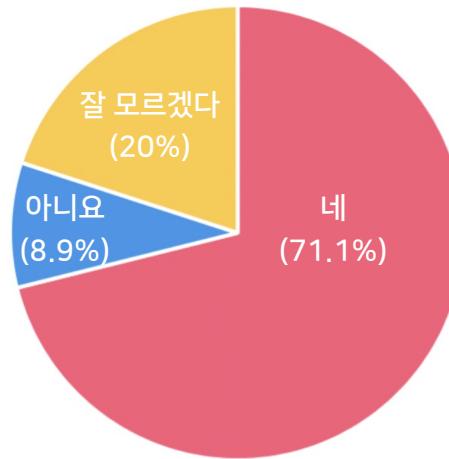
02 모듈별 소개 / 라이벌 추천

solved.ac

의 라이벌 기능



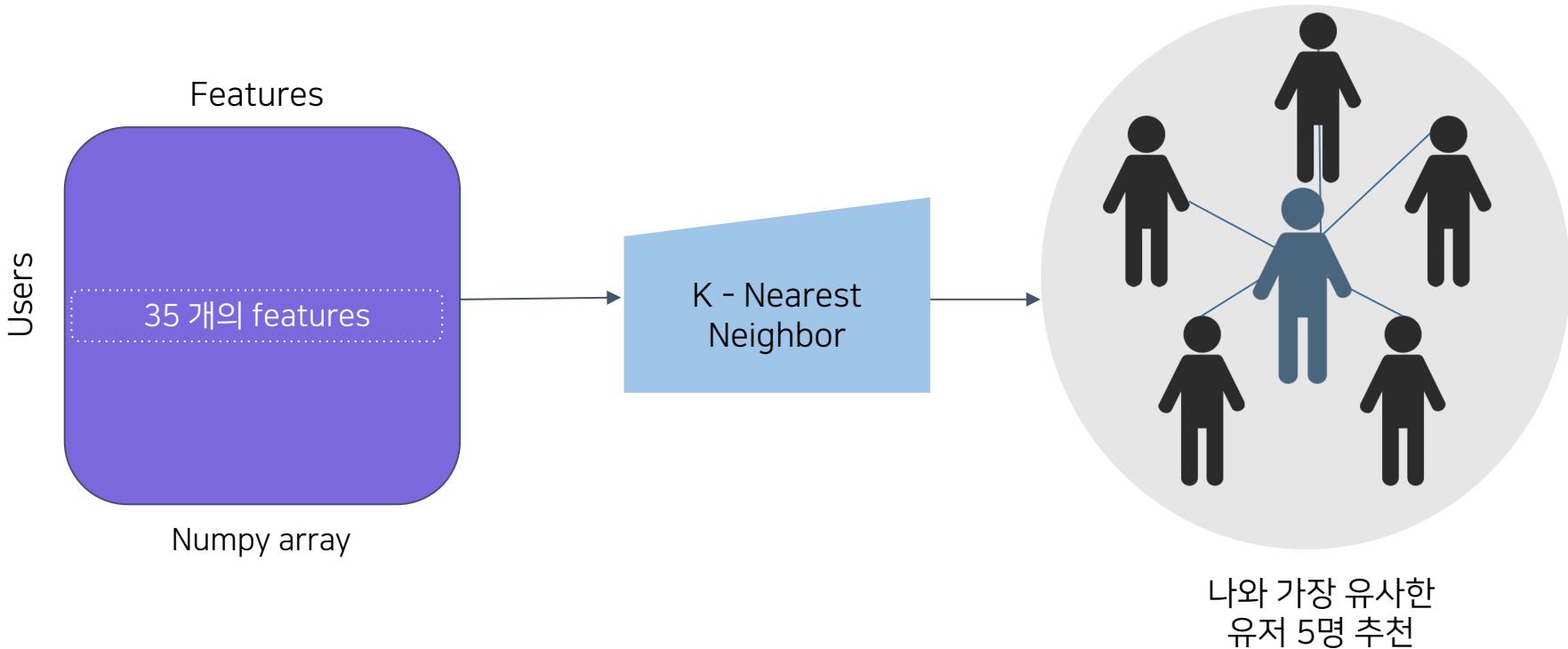
Q. 라이벌 추천을 통해 자신의 실력 향상과 문제 풀이의 동기부여가 될 것 같다고 생각하나요?



출처: 백준 온라인 저지 문제 추천 서비스 예상 선호도 조사
대상: 45명 (코딩테스트 준비 오픈 카카오톡 채팅방, S' 대학교 ICPC Team Slack 채널 등)

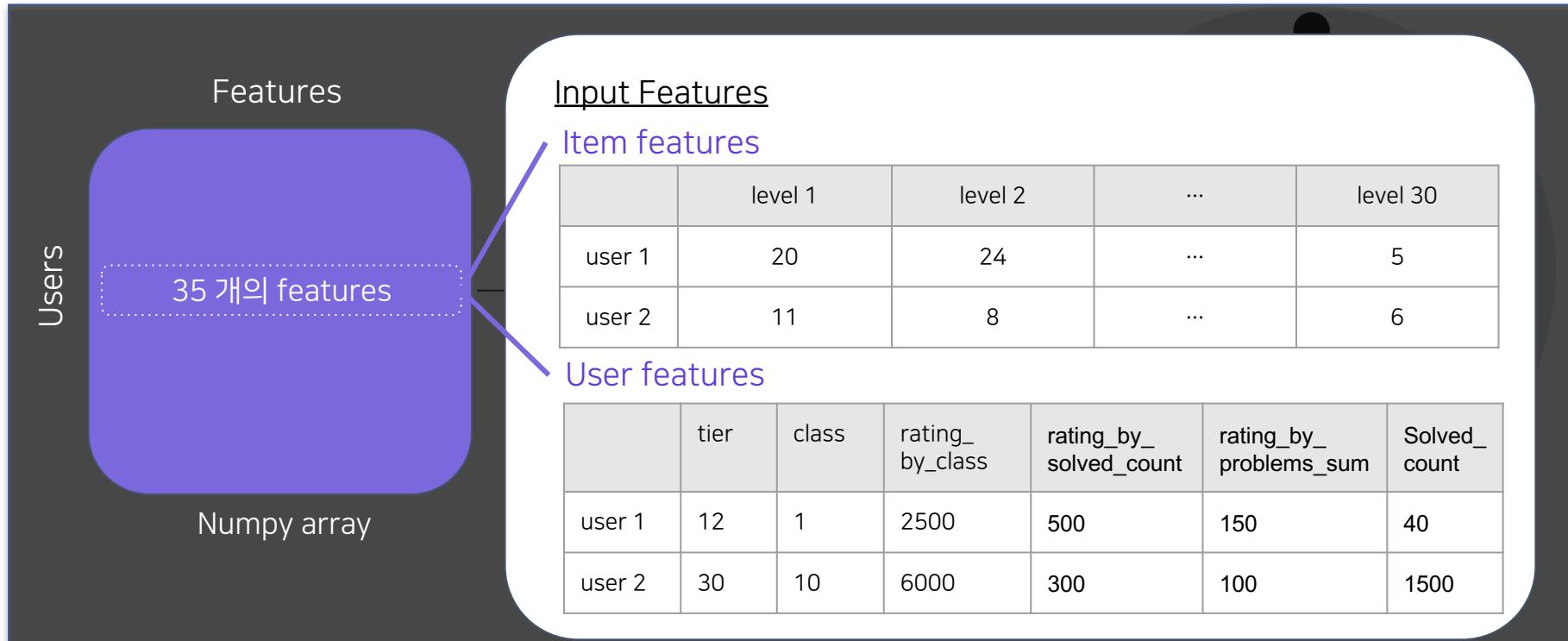
02 모듈별 소개 / 라이벌 추천

Process



02 모듈별 소개 / 라이벌 추천

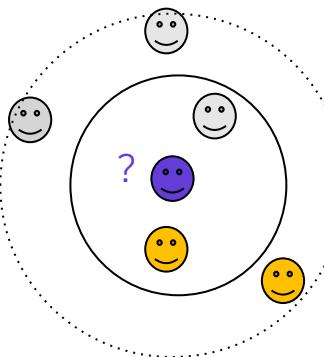
Input



02 모듈별 소개 / 라이벌 추천

Model

KNN



데이터의 특성을 기반으로 제일 근접한 k개의 요소의 라벨을 참조하여, 타겟 데이터가 어떤 라벨에 속하는지 분류하는 알고리즘

K - Nearest Neighbor

Experiments Result

Cosine	0.2164
Matrix Factorization	0.1565
Collaborative MF	0.1549
Light GCN	0.2160
DBSCAN + KNN	0.1599
Kmeans + KNN	0.3345
KNN	0.1285

02 모듈별 소개 / 라이벌 추천

Result

Evaluation

백준 Solved.AC Rating 산출법

유저 별 상위 100개 문제의
난이도 값의 합

CLASS에 따른 보너스

푼 문제 수에 따른 보너스

기여 수에 따른 보너스
(최대 25)

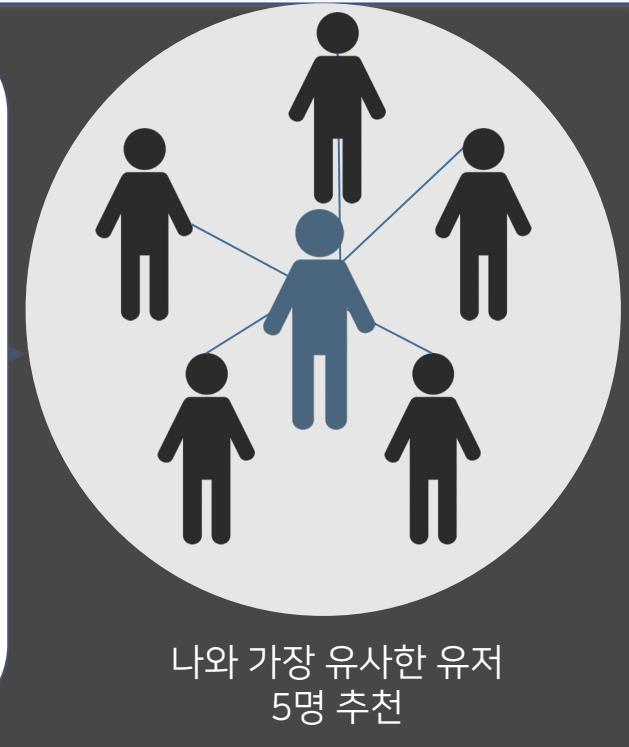
Mean

라이벌 추천 성능 지표

$$\text{sum} \left(\frac{|\text{나의 난이도 합} - \text{라이벌 난이도 합}|}{\text{최대 난이도 합}} \right) \text{라이벌 수}$$

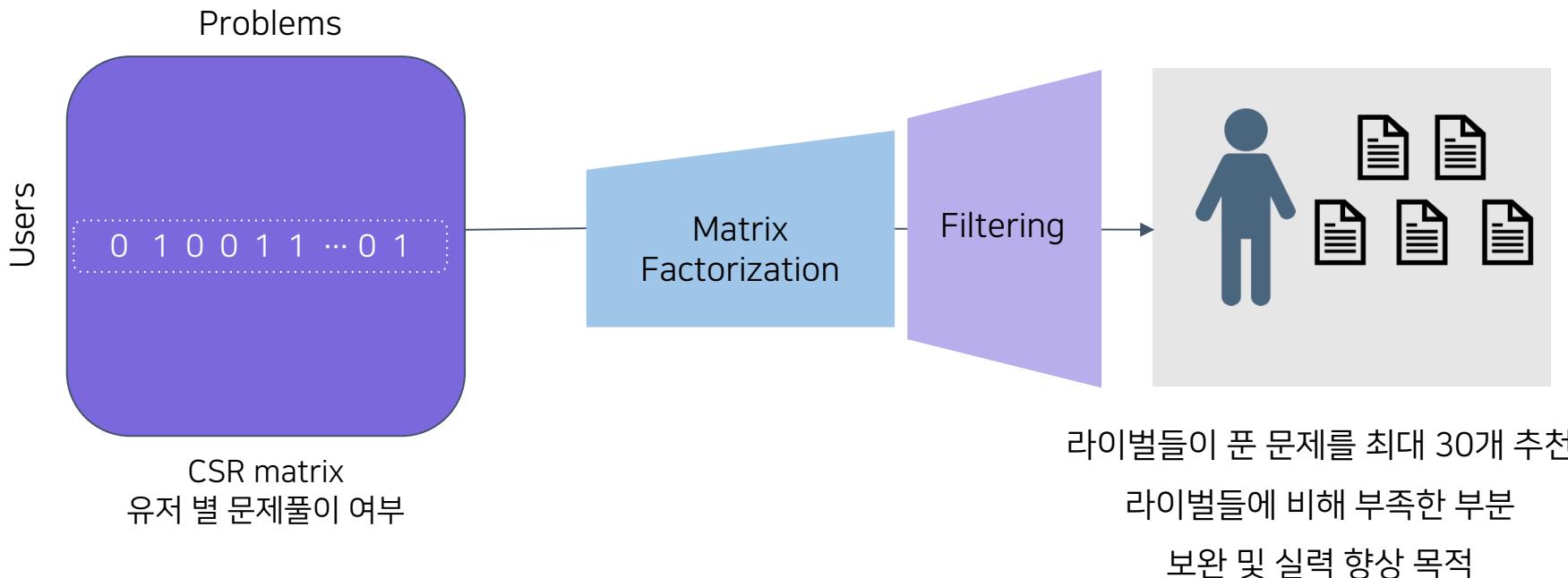
$$\text{sum} \left(\frac{|(\text{나의 클래스보너스 점수} - \text{라이벌 클래스보너스 점수})}{\text{최대 클래스보너스 점수}} \right) \text{라이벌 수}$$

$$\text{sum} \left(\frac{|\text{나의 문제풀이 보너스 점수} - \text{라이벌 문제풀이 보너스 점수}|}{\text{최대 문제풀이 보너스 점수}} \right) \text{라이벌 수}$$



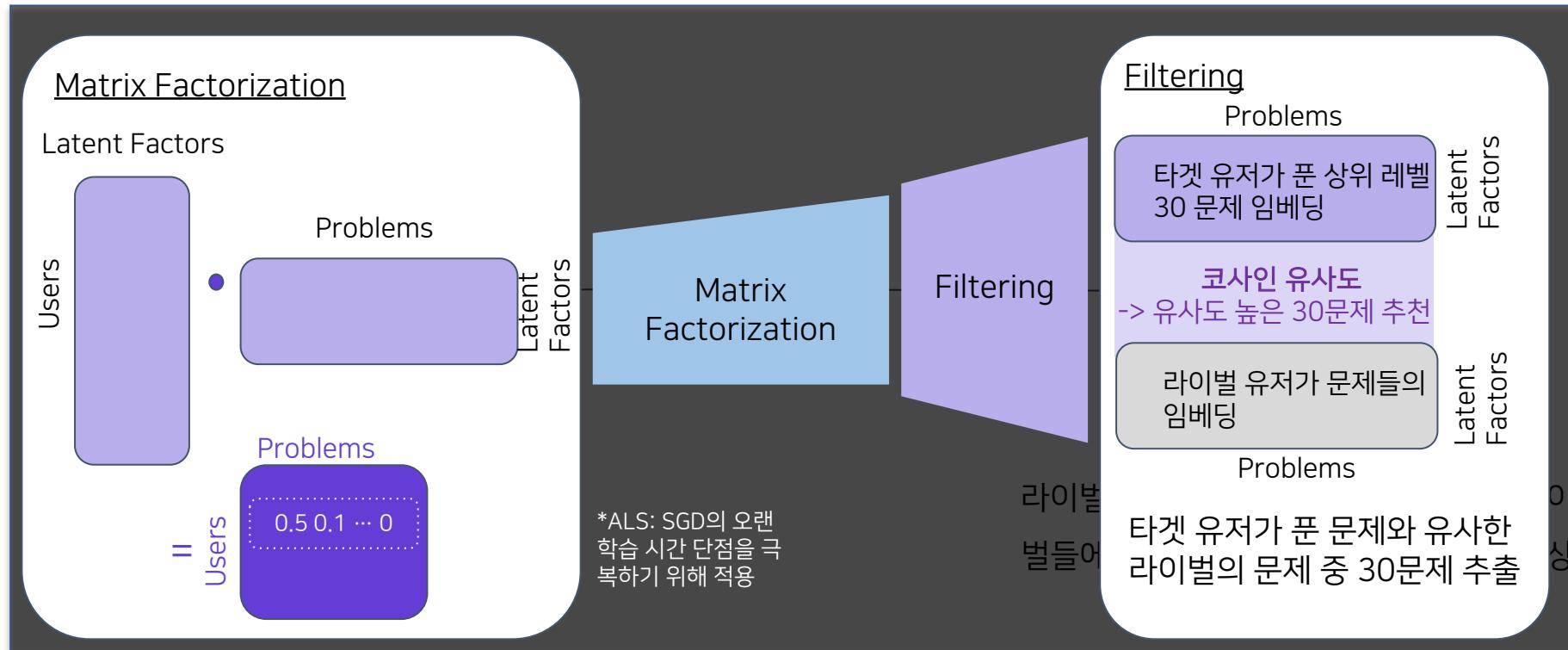
02 모듈별 소개 / 라이벌 기반 문제 추천

Process



02 모듈별 소개 / 라이벌 기반 문제 추천

Model 유저의 문제풀이 패턴을 고려한 문제를 추천해주고자 ML 기법 도입.



02 모듈별 소개 / 라이벌 기반 문제 추천

Result

Evaluation

라이벌 기반 문제 추천 성능 지표

$$\frac{\sum \left(\frac{\text{내가푼문제의 난이도합}}{\text{내가푼문제수}} - \frac{\text{추천된문제의 난이도합}}{\text{추천된문제수}} \right)}{\text{전체유저수}}$$

* 내가 푼 문제는 상위 레벨 100문제의 난이도 합으로 설정

Experiments Result

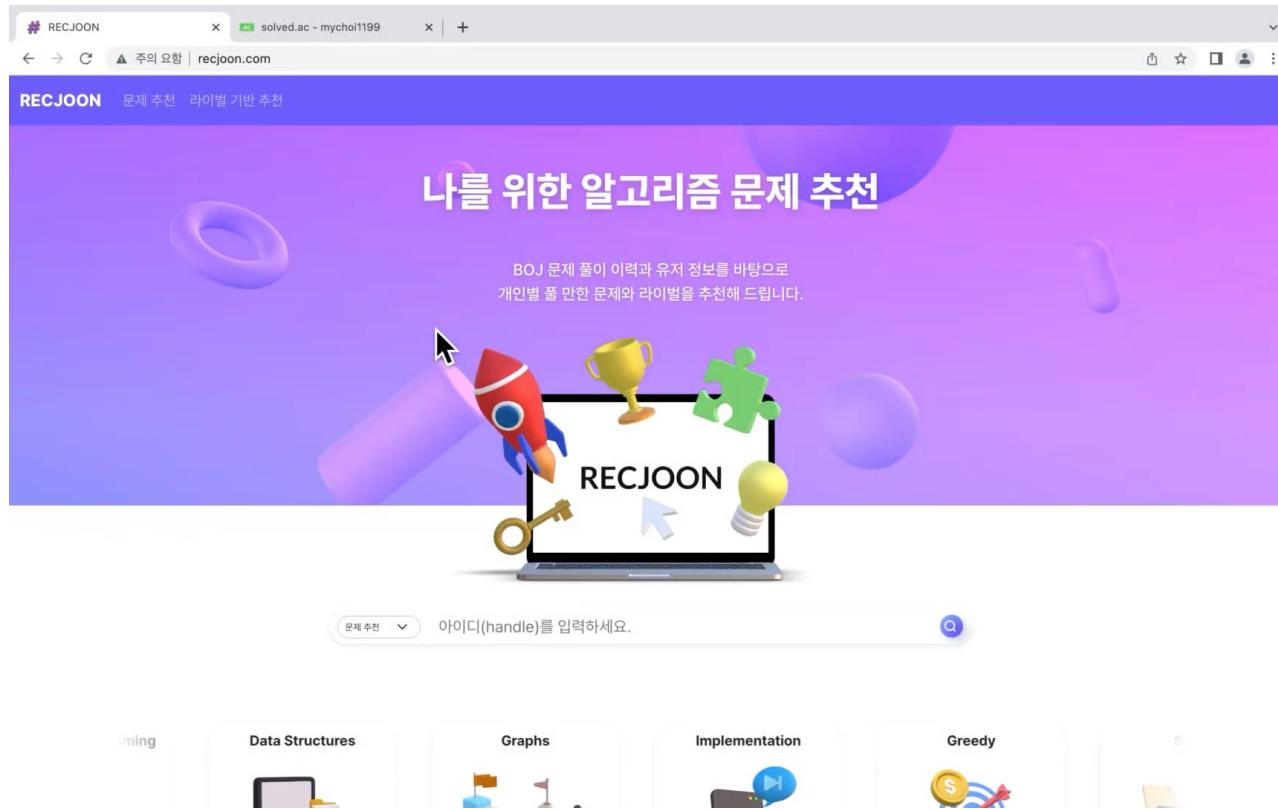
실험 모델	지표
Item based Collaborative Filtering	3.036
Matrix Factorization - ALS	1.258
Bayesian Personalized Ranking	1.419

Filtering



추천된 문제와 타겟유저가 푼 문제의 난이도 차이가 적을수록 잘 추천된 문제라고 판단

02 모듈별 소개 / 라이벌 추천 시연 영상

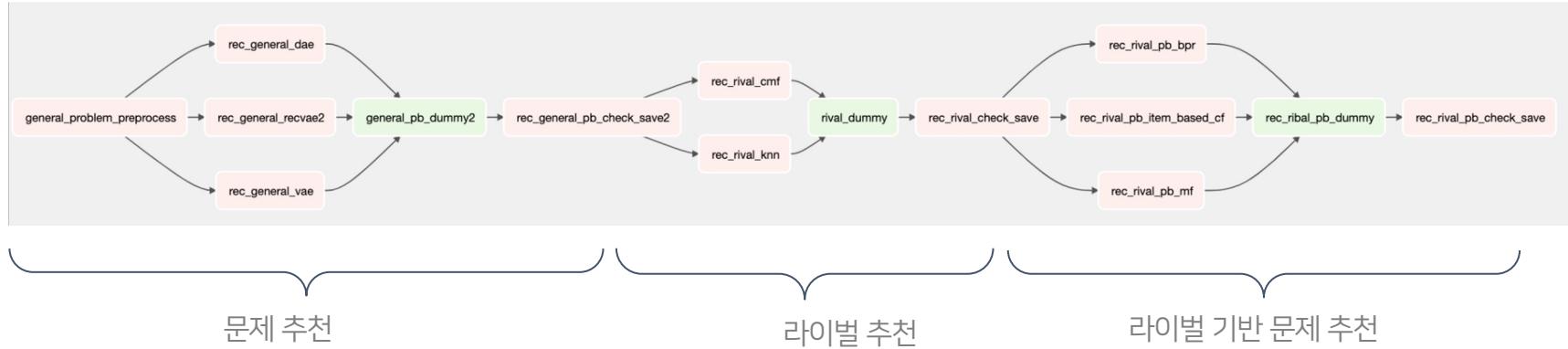


02 모듈별 소개 / Product Serving



Airflow DAGs를 통한 배치 스케줄 관리

- 주기: 1주일 1번

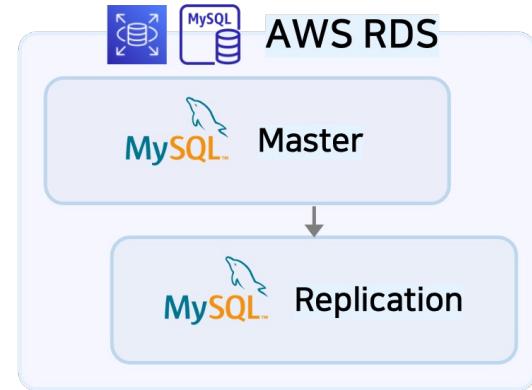
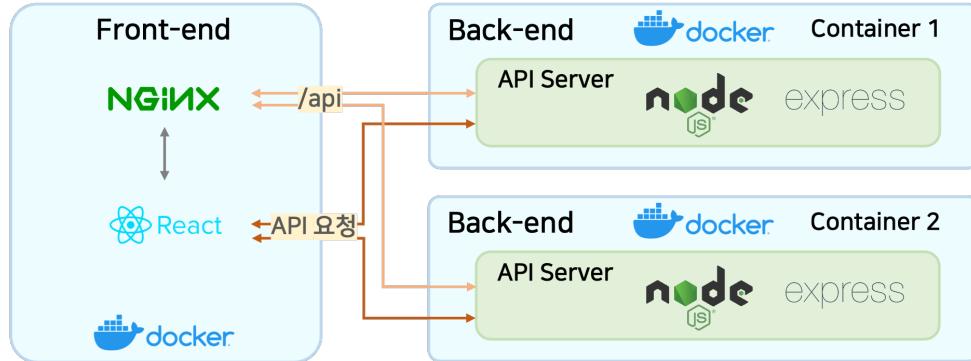


Airflow를 통한
다양한 모델 학습 및 추론 병렬화

사전에 정의한 지표 결과 값 비교 시 가장 잘 나온 모델의 추론 결과 사용

02 모듈별 소개 / Product Serving

NGINX를 사용한 무중단 배포



용도에 맞게 나눈 데이터베이스

웹 스크래핑 또는 모델 Inference 과정에서 데이터 업로드 오류가 발생할 경우 실제 서비스에서의 장애 발생 예방
쓸 수 있는 데이터베이스(Master)와 이를 복제한 읽기 전용의 데이터베이스(Replication)

03 Result

결과 및 고찰

Appendix



03 Result / 결과 및 고찰

프로젝트 의의

<서비스 측면>

- 알고리즘 학습 시 문제 선정의 고충 제거
- 나와 실력이 유사한 유저들을 추천받음으로써 학습 시 동기 부여

<개발 측면>

- 자동화 배포를 통해 유저의 최신 정보를 반영한 라이벌과 문제 추천

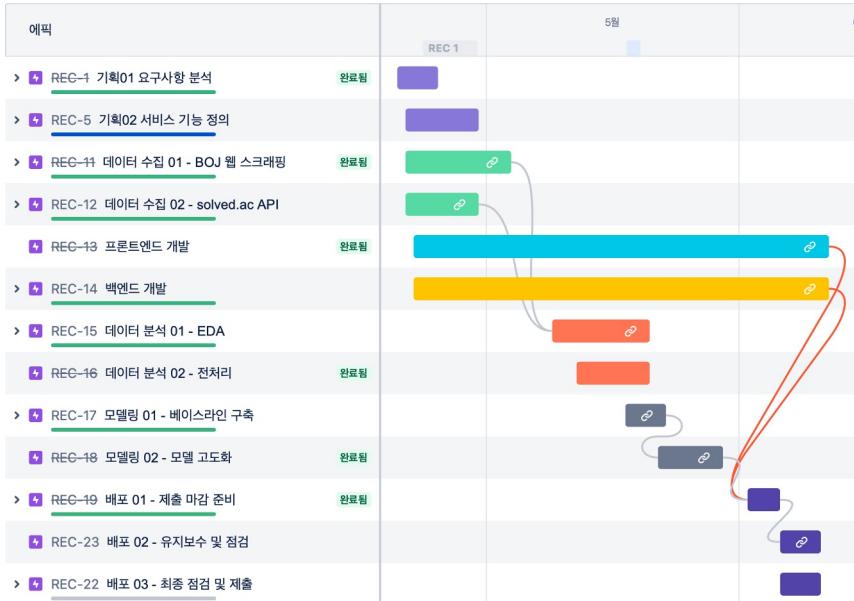
03 Result / 결과 및 고찰

더 시도해 볼만한 사항

- AWS EC2 인스턴스 오토스케일링
 - 실제 서비스 개시 후 유동적인 트래픽 대비와 클라이언트의 요청 처리 분산
- 추천되는 문제의 레벨 범위 고도화
 - 본인의 레벨에 맞는 문제들이 필터링 될 수 있도록 문제 레벨의 범위를 구하는 공식 적용

03 Result / Appendix

Jira를 통한 협업 및 로드맵 구축



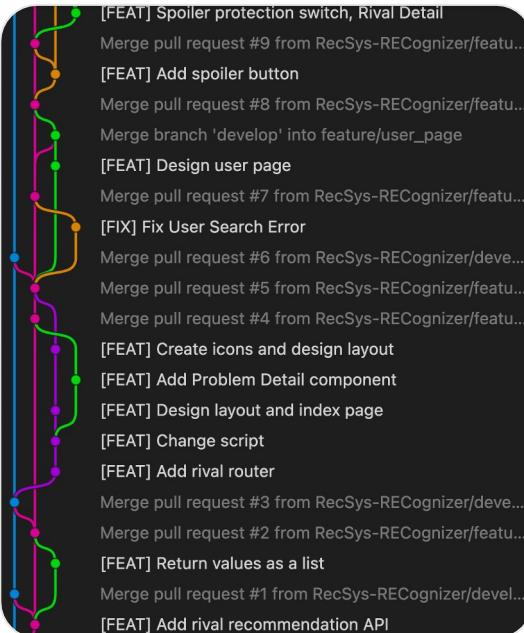
Notion을 활용한 태스크 및 실험 관리

A Notion table titled 'Default view' showing tasks and experiments. The table includes columns for task type, name, start date, end date, and assignees.

타입	제목	시작일	종료일	담당자
기획	요구사항 분석	2022/04/21	2022/04/24	Truth Glanceyes
데이터 수집	solved.ac API	2022/04/21	2022/04/29	Glanceyes
데이터 수집	BOJ 웹 스크래핑	2022/04/21	2022/05/03	Truth Glanceyes Juke
기획	서비스 기능 정의	2022/04/22	2022/04/29	All
개발	Back-end 설계 및 개발	2022/04/22	2022/05/20	Sunny Brill Glanceyes Juke
개발	Front-end 개발	2022/04/22	2022/06/01	Glanceyes Juke
데이터 분석	EDA	2022/05/04	2022/05/12	All
데이터 분석	전처리	2022/05/12	2022/05/15	Sunny Brill Truth
개발	모델 베이스라인 구축	2022/05/16	2022/05/20	Sunny Brill Truth
개발	모델 설계 및 개발	2022/05/19	2022/05/29	All
배포	제출 마감 준비	2022/06/02	2022/06/05	All
배포	테스트 배포	2022/06/05		All
배포	유지보수 및 점검	2022/06/06	2022/06/10	All
배포	최종 점검 및 제출	2022/06/06	2022/06/10	All
배포	최종 배포	2022/06/10		All

03 Result / Appendix

Git branch 관리

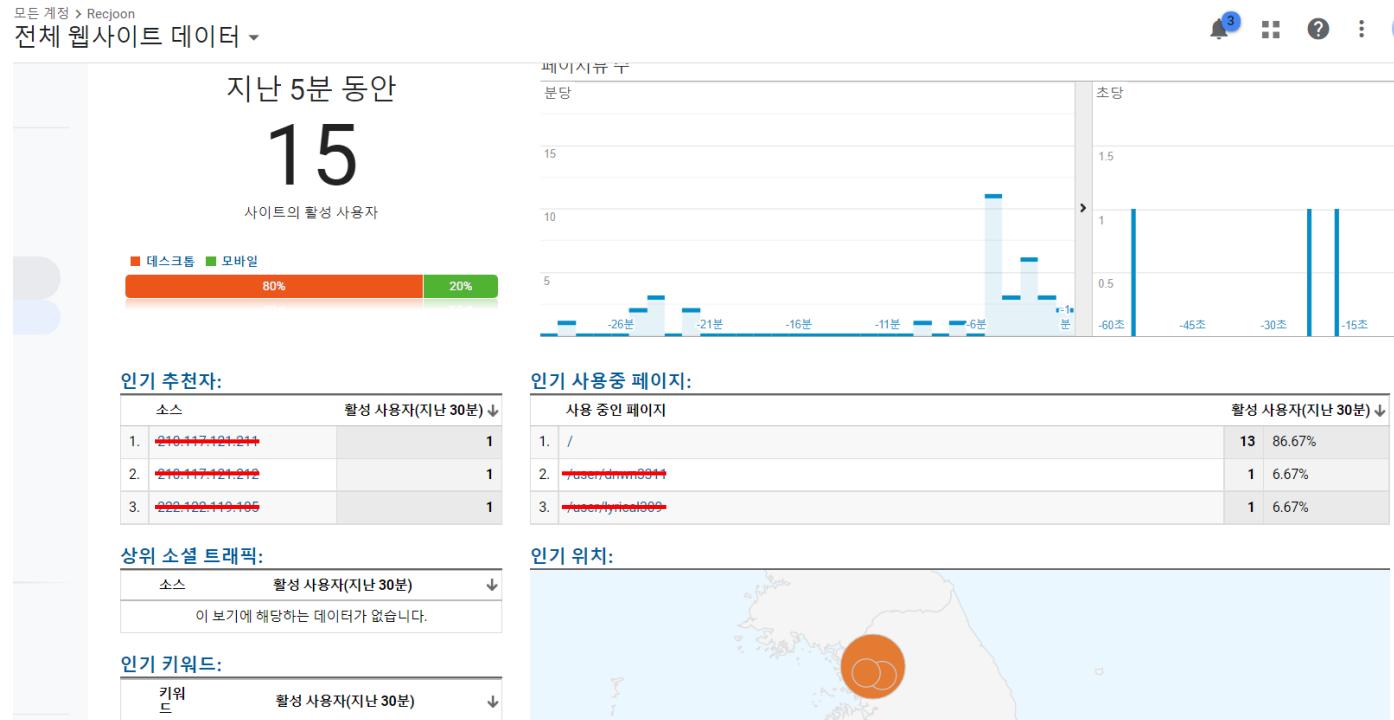


GitHub Action을 통한 Code Deployment 자동화

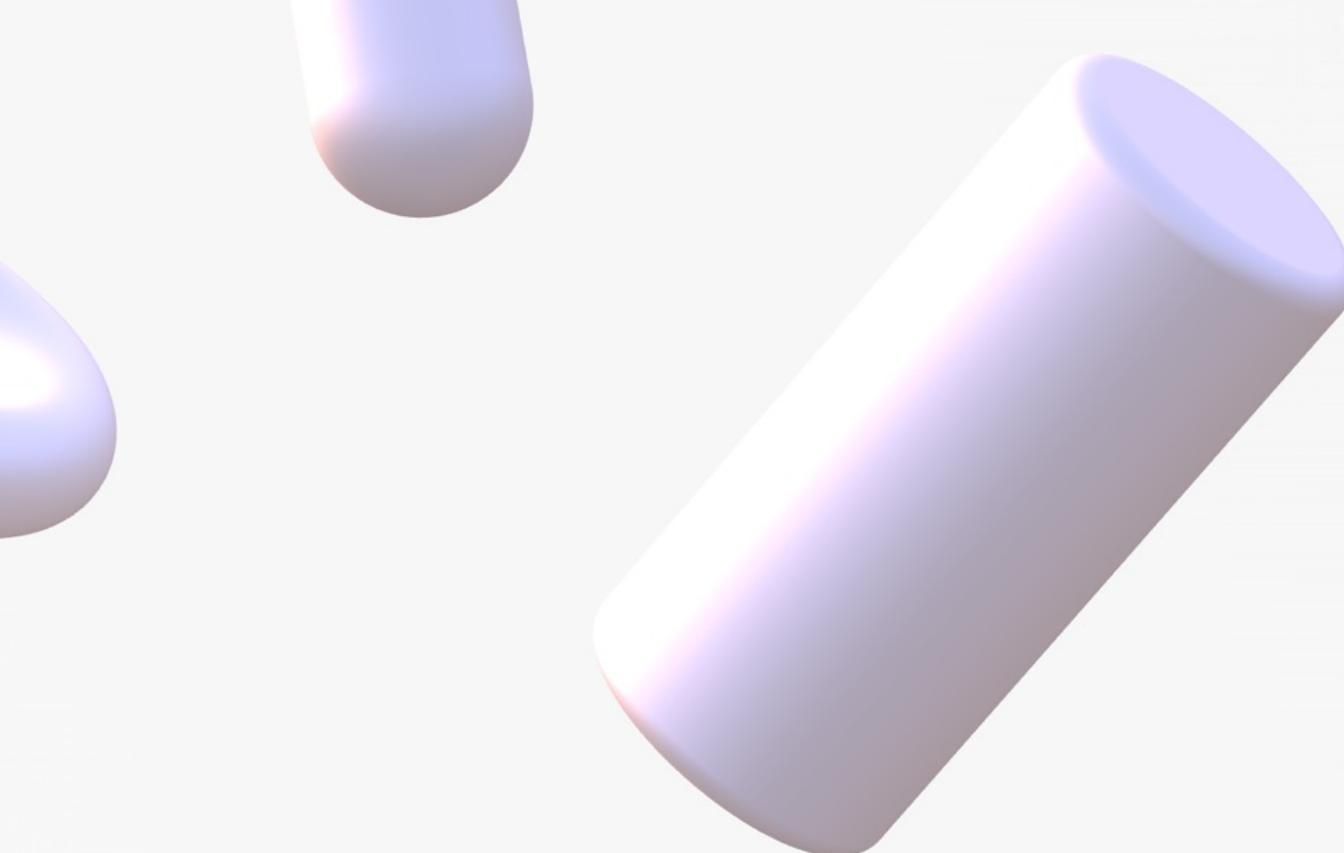
7 workflow runs	Event ▾	Status ▾	Branch ▾	Actor ▾
✓ [FEAT] Beta 1.0.1 for test deploy #7: Pull request #23 closed by Glanceyes	develop	12 hours ago 2m 35s	...	
✓ [FEAT] Beta 1.0 for test deploy #6: Pull request #17 closed by Glanceyes	develop	2 days ago 2m 34s	...	
✓ [FEAT] Beta index page deploy #5: Pull request #6 closed by Glanceyes	develop	6 days ago 2m 31s	...	
✓ [FEAT] Return API value as list deploy #4: Pull request #3 opened by Glanceyes	develop	9 days ago 2m 29s	...	
✓ Merge pull request #1 from RecSys-RERecognizer/de... deploy #3: Commit f237d32 pushed by Glanceyes	main	9 days ago 2m 18s	...	
✓ [FEAT] Add rival recommendation API deploy #2: Pull request #1 opened by Glanceyes	develop	9 days ago 2m 0s	...	
✓ [FEAT] First commit deploy #1: Commit 98946e3 pushed by Glanceyes	main	10 days ago 2m 0s	...	

03 Result / Appendix

구글 태그 매니저와 애널리틱스를 활용한 성과 모니터링



Q&A

A large, semi-transparent, light purple cylinder is positioned diagonally across the center of the frame. Behind it, a smaller, rounded, light purple shape is visible. The background is a plain, light gray.

감사합니다.