Da, GO! Mall

온라인 쇼핑몰 및 스마트 폰에서의 추천 시스템
- 반응형 웹 전자기기 쇼핑몰

목차

1. Introduction
1. Data Preprocessing
1. DB Schema & Data Info
1. Recommendation System
1. 쇼핑몰 시연

1 Introduction

팀원 소개

이현준
대학생
한양대학교 컴퓨터소프트웨어학부

정지훈
대학생
한양대학교 컴퓨터소프트웨어학부

Monthly Planner

	이현준	정지훈
March	Recommendation Algorithm	Data Analysis
April	요구사항 분석	개발 tool 선정
May	ER모델 설계	데이터베이스 설계
June	쇼핑몰 layout 만들기	
July	Front-end와 Back-end 결합	
August	Collaborative filtering	Content-Based RecSys

Tool

Language	Python3
Front-end	Bootstrap
Back-end	Django, Rest Framework
Server	EC2
Database	Amazon RDS, MariaDB
Recommendation System	Pandas scikit-learn

2

Data Preprocessing

Amazon Electronics Dataset

May 1996 – Oct 2018

20,994,353

Electronics Review 11GB

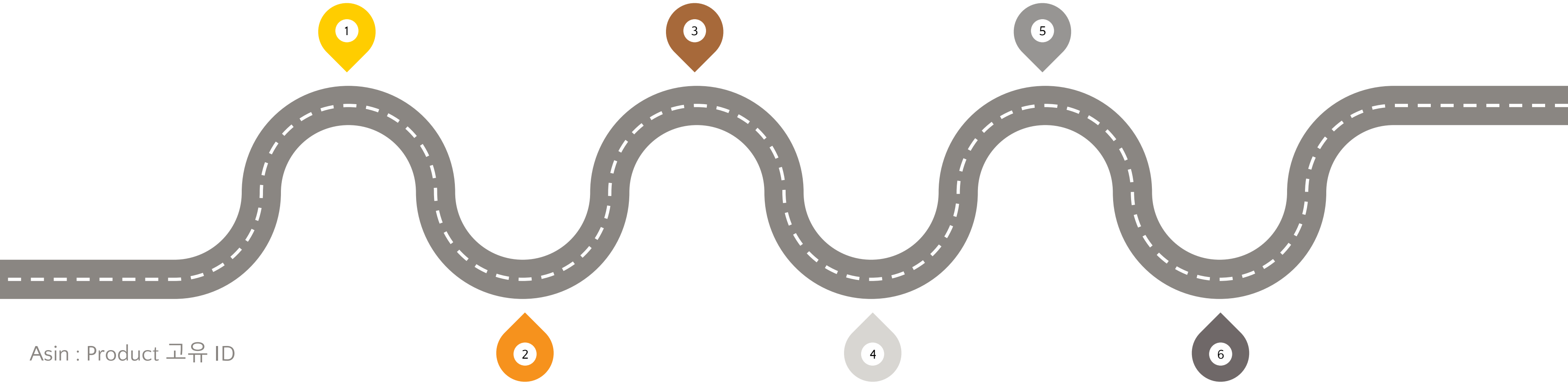
786,445

Electronics Meta Product 10GB

Too Big!

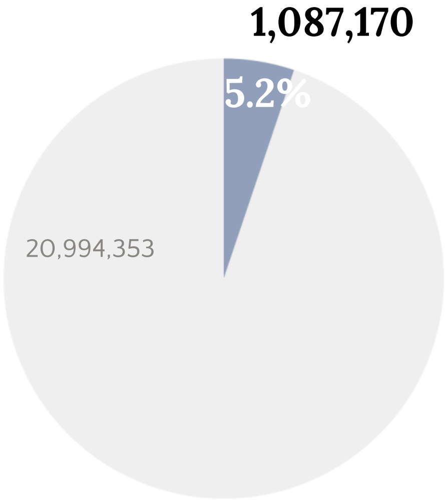
Data Preprocessing

1. 최신의 쇼핑몰 구현을 위해 2018년도 Review 추출
2. 실 구매한 리뷰를 바탕으로 Meta Product 추출
3. Asin, Title이 중복되고 추천에 필요한 Feature가 없는 Product Drop
4. 정제된 Product에 맞는 Review 재추출
5. Meta Dataset에서 Brand, Category, Also View Dataset 구축
6. 쇼핑몰 구현에 필요 없는 Meta, Review 각 Column Drop

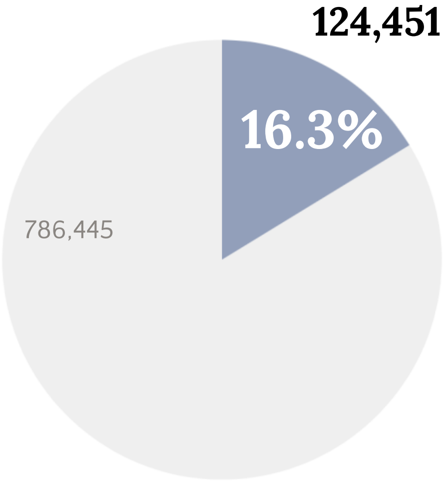


Asin : Product 고유 ID

Data Preprocessing



Review



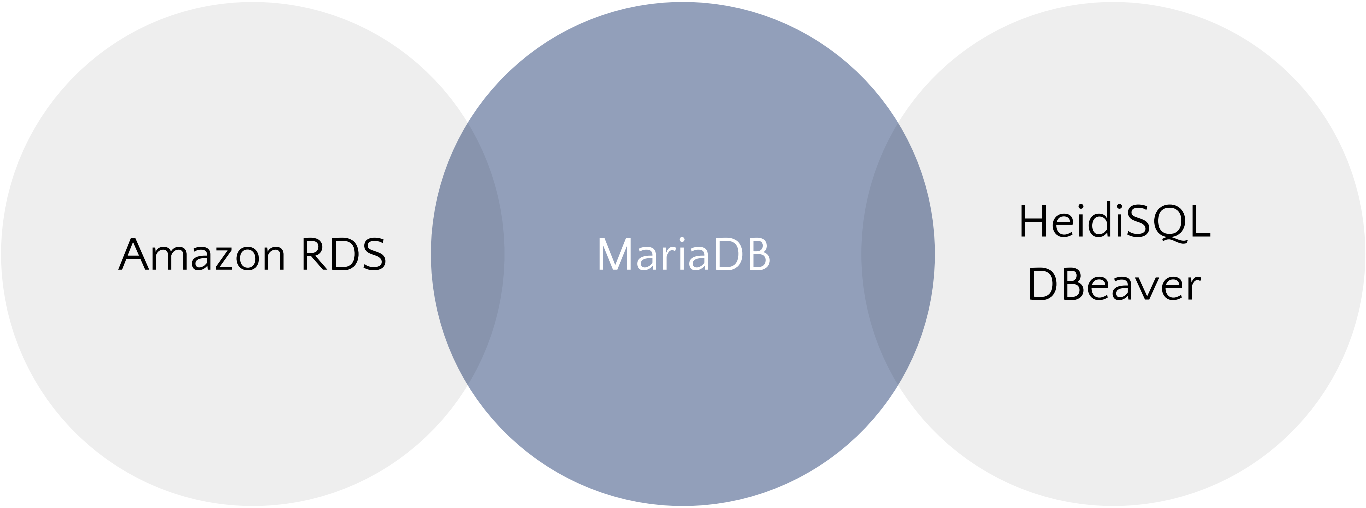
Meta

3

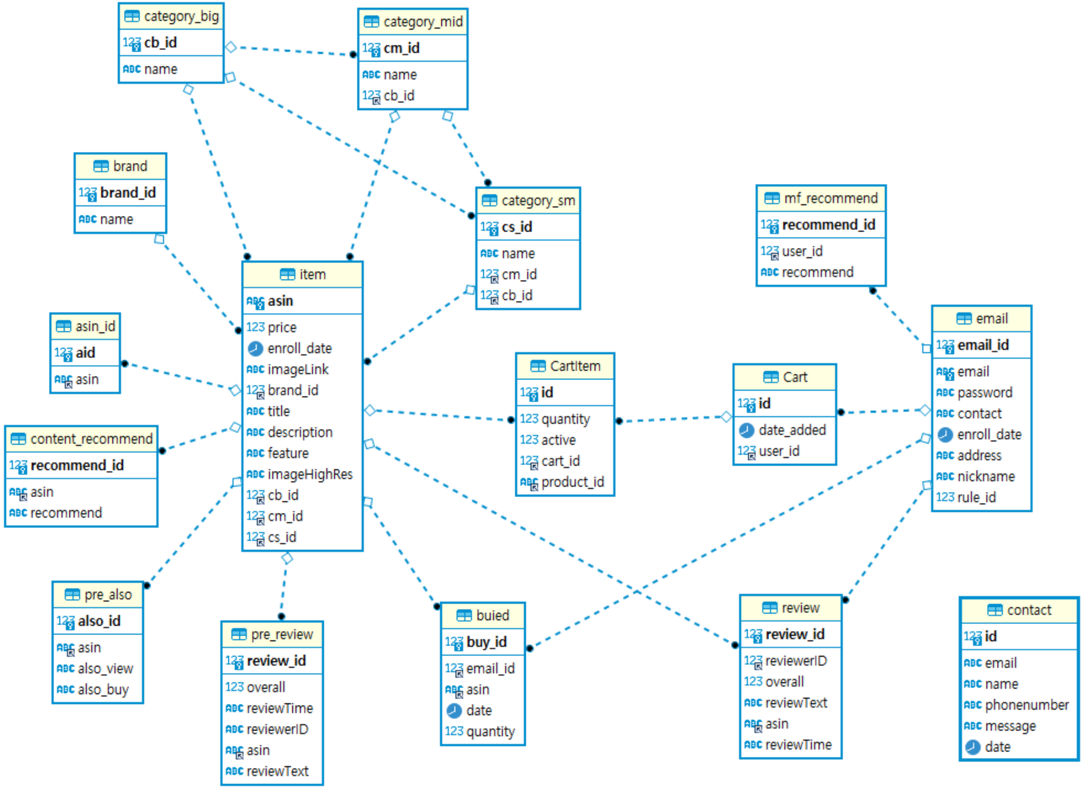
DB Schema & Data Info

Amazon RDS MariaDB

Tool



DB schema



category_big
cb_id
name
category_mid
cm_id
name
cb_id
brand
brand_id
name
category_sm
cs_id
name
cm_id
cb_id
mf_recommend
recommend_id
user_id
recommend
item
asin
price
enroll_date
imageLink
brand_id
title
description
feature
imageHighRes
cb_id
cm_id
cs_id
asin_id
aid
asin
content_recommend
recommend_id
asin
recommend
CartItem
id
quantity
active
cart_id
product_id
Cart
id
date_added
user_id
email
email_id
email
password
contact
enroll_date
address
nickname
rule_id
pre_also
also_id
asin
also_view
also_buy
pre_review
review_id
overall
reviewTime
reviewerID
asin
reviewText
buied
buy_id
email_id
asin
date
quantity
review
review_id
reviewerID
overall
reviewText
asin
reviewTime
contact
id
email
name
phonenumber
message
date

Category

	Big	Medium	Small		Big	Medium	Small
Camera & Photo	1	15	88	Security & Surveillance	9	10	20
Accessories & Supplies	2	10	74	Television & Video	10	16	25
GPS, Finders & Accessories	3	6	18	Home Audio	11	7	28
Computers & Accessories	4	14	88	Video Projectors	12	1	–
eBook Readers & Accessories	5	11	12	Wearable Technology	13	7	9
Headphones	6	6	12	Service Plans	14	1	–
Car & Vehicle Electronics	7	5	29	Electronics Warranties	15	1	–
Portable Audio & Video	8	16	33	계	15	126	436

Brand

Total Brand : 17,170 Total Product : 124,415

Rank	Id	Name	Count	Rank	Id	Name	Count
1	76	Sony	1,817	9	3,719	Fintie	647
2	18	Dell	1,505	10	1,033	Asus	623
3	55	Samsung	1,358	11	1,162	Monoprice	599
4	237	Generic	1,165	12	195	Apple	597
5	155	HP	1,120	13	16	SanDisk	586
6	97	Canon	840	14	57	Garmin	563
7	93	Nikon	712	15	2,033	SquareTrade	558
8	2,506	Neewer	704	계	15	-	13,394

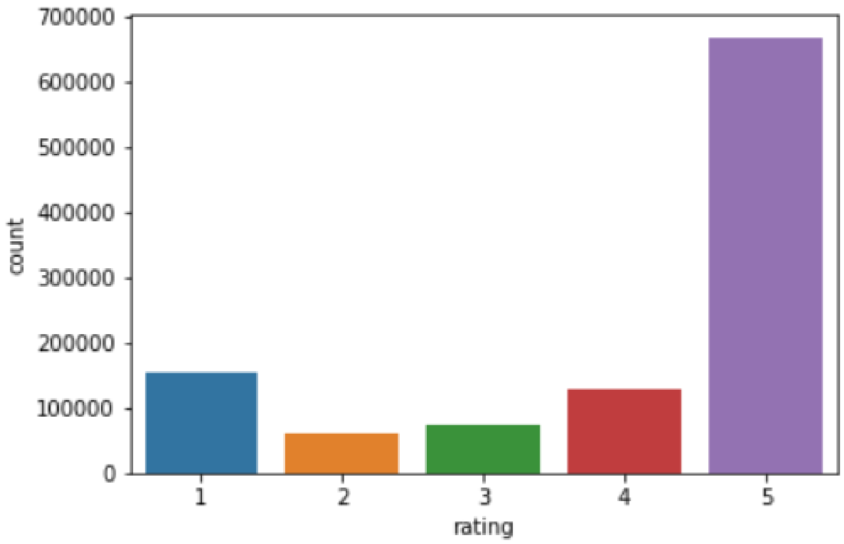
Review

	count	mean	std	min	25%	50%	75%	max
rating	1087170.00	4.01	1.48	1.00	3.00	5.00	5.00	5.00

Reviewer 평균평점	**Count**
1	127,736
5	461,625

Product 평균평점	**Count**
1	10,192
5	47,200

평점	**Count**
1	154,921
2	61,452
3	72,037
4	129,819
5	668,941
계	1,087,170



Top 5 Review

Total Reviewer : 825,733

Top 5 Reviewer	**Review ID**	**Count**
1	A10SE0U42ABS9S	95
2	A2O5RXNP1U7UXC	50
3	A3RW697Y2EZWUL	47
4	ALDWXFCCVIGGR	47
5	A2ZDJWO9G0CTL	45
계	5	284

Total Product : 124,415

Top 5 Product	**Asin**	**Count**
1	B017YEA6QW	3860
2	B01F9RGWF0	3673
3	B014I8T4MO	3628
4	B015IEW1AY	3354
5	B00F0DD0I6	3227
계	5	17,742

4-1

Content-Based

TF-IDF Vectorizer & Cosine Similarity

Content-Based RecSys

TF-IDF Vectorizer

TF : 단어가 문서 내에서 얼마나 자주 등장하는지

IDF : 단어가 문서 전체에서 얼마나 공통적으로 등장하는지

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$FV \in \mathbb{R}^{124,415 \times 4,660,813}$$

Cosine Similarity

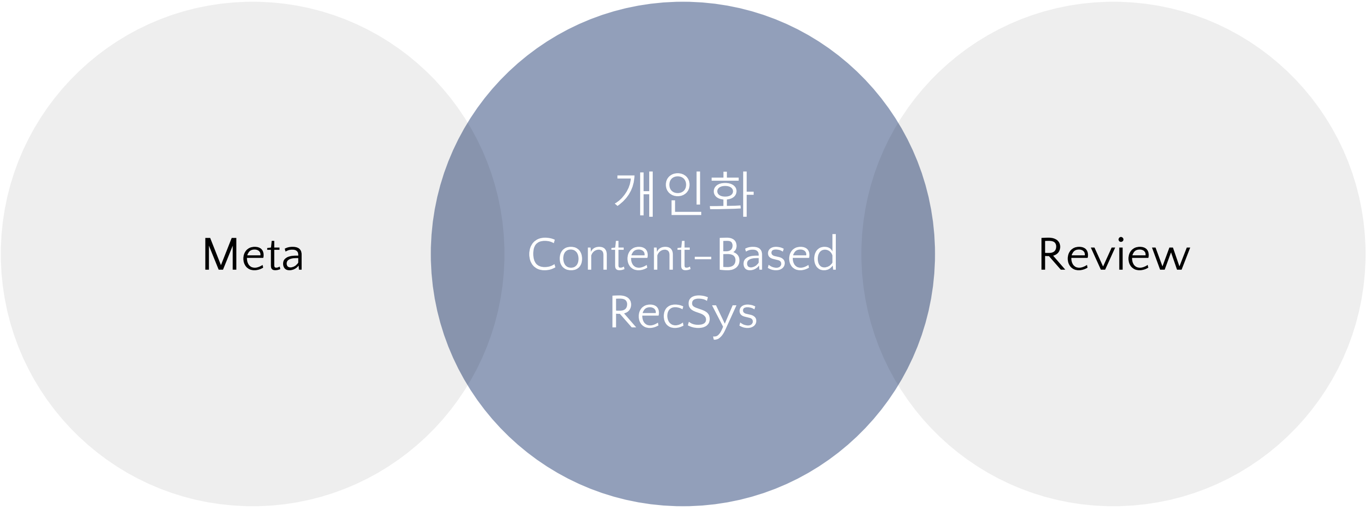
Feature Vector 유사도를 통해 비슷한 Meta 계산

Result

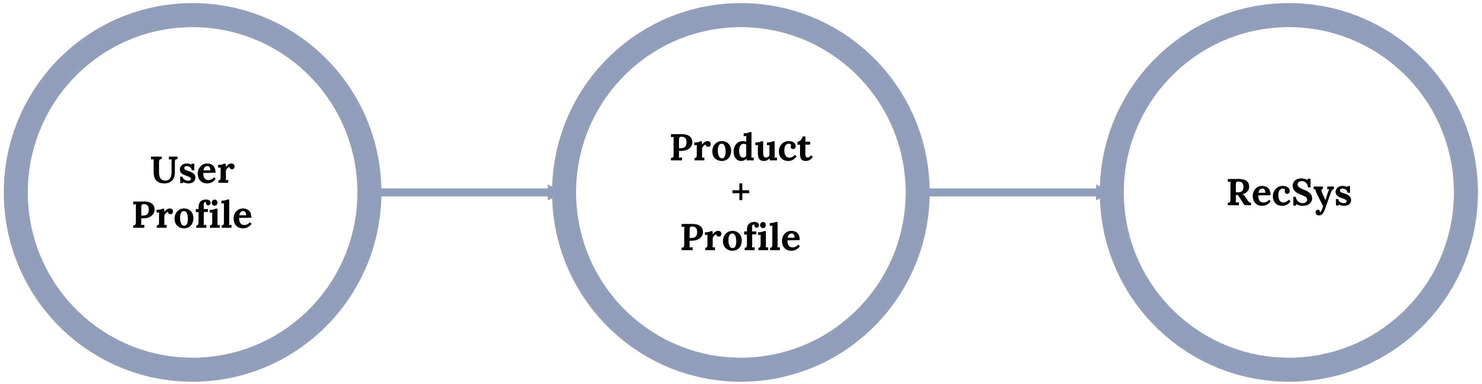
예시	Asin	유사도
LG 21:9 QHD IPS 34" Monitor	B01GLYBTP8	0.78
	B01GLRK0RI	0.73
	B01GLCY9UM	0.67
	B01GLXVY7M	0.59

예시 링크

개인화 Content-Based RecSys



개인화 Content-Based RecSys



$$UP(u) = \frac{\sum_{i=1}^{n}(\mathcal{S}_i * FV_i)}{\sum_{i=1}^{n}\mathcal{S}_i}$$

$n = |R_u|$ (User의 Reviews)

$\mathcal{S}_i$: i번째 Review의 Score

FV_i : i번째 Review 제품의 FV(Feature Vector)

$$PFV = \beta * FV_p + (1 - \beta) * UF(u)$$

PFV : Personal FV

FV_p : 선택된 Product FV

$$CS \in \mathbb{R}^{1 \times 124{,}415}$$

Persona

하이피플

개인화 Content-Based RecSys

Content-Based

Recommend

PIN YUAN Lightnin...

Lighting to HDMI ...

ebasy HDMI Adapte...

ASAITEKE Compatib...

개인화 RecSys

Recommend

1080P Audio Conne...

PIN YUAN Lightnin...

Lighting to HDMI ...

Lighting to HDMI ...

개인화 Content-Based RecSys

Result

예시	Asin	유사도
B00F4QM67A	B00ESL71RW	0.33
	B008E0U51W	0.32
	B00M6OEV12	0.25
	B00M7NEB1C	0.17

예시	Asin	유사도
B00F4QM67A	B00GXJMWUI	0.161
	B00ESL71RW	0.160
	B008E0U51W	0.158
	B00OZ4QRMW	0.149
	⋮	
6등	B00M6OEV12	0.122
10등	B00M7NEB1C	0.084

4-2

Recommendation System

Collaborative filtering (Matrix Factorization)

Preprocessing

Still big data

- Transaction : 1,087,170
- User : 825,733
- Product : 124,415

	userId	productId	rating
0	A2ZCK9686DTWU0	106171327X	5
1	A3G5NNV6T6JA8J	106171327X	5
2	AFML7PYI3LERI	106171327X	5
3	A1G0HYMR02WM2W	106171327X	4
4	A1T8B3I8KRS3W0	106171327X	5
...	...	...	...
1087165	A34QQHRRG65E08	B01HJF9W84	5
1087166	A2ASY291VN2H6F	B01HJGOOMW	2
1087167	A3TJ82VBTRW3O6	B01HJGOOMW	1
1087168	A10GEXRHH85UDI	B01HJF704M	1
1087169	AZVJAG9XGX1R	B01HJF704M	5

1087170 rows × 3 columns

Preprocessing

Too sparse

- Most people do not give a rating
- Users who have rated more than 50 products : Only 2

	UserId	Ratings
0	A10SE0U42ABS9S	95
1	A2O5RXNP1U7UXC	50
2	A3RW697Y2EZWUL	47
3	ALDWXFCCVIGGR	47
4	AM2IJT2H3SG	45
5	A2ZDJWO9G0CTLL	45
6	A37XP21TPO5ZED	43
7	A2NF3AZ6GSIBEY	40
8	A22BQ2P4R61T3E	40
9	A2XP8CV9ES33AM	35

Preprocessing

Solution

- Drop users who have rated less equal than 5 products
  - Transaction : 1,087,170 => 66,352
  - User : 825,733 => 7,934
  - Product : 124,415 => 27,646

```
Unique USERS who have rated 6 or more products : 7934
Unique USERS dropped      : 817799
Unique ITEMS remaining    : 27646
Unique ITEMS dropped      : 96769

Final length of the dataset : 66352
```

Preprocessing

Result

	userId	productId	rating
3	A1G0HYMR02WM2W	106171327X	4
16	A2T7D9P4TM7L83	106171327X	5
17	AUGBCGUG05AH7	106171327X	5
59	A2WBQ18YJ2CYYF	B000001ON6	3
61	A288U5R1WTBGEH	B000001OM5	4
...	...	...	...
1087017	A7JA5G6OMWYSR	B01HIS365C	4
1087053	A2LJ2R6YU6VG6M	B01HITUTV0	1
1087058	A7CDO4MIXAONA	B01HIUOLZ4	5
1087100	A1QGNRP0RTECUA	B01HIZK1B2	5
1087107	A2L12USPGEMCTM	B01HIZEW1C	5

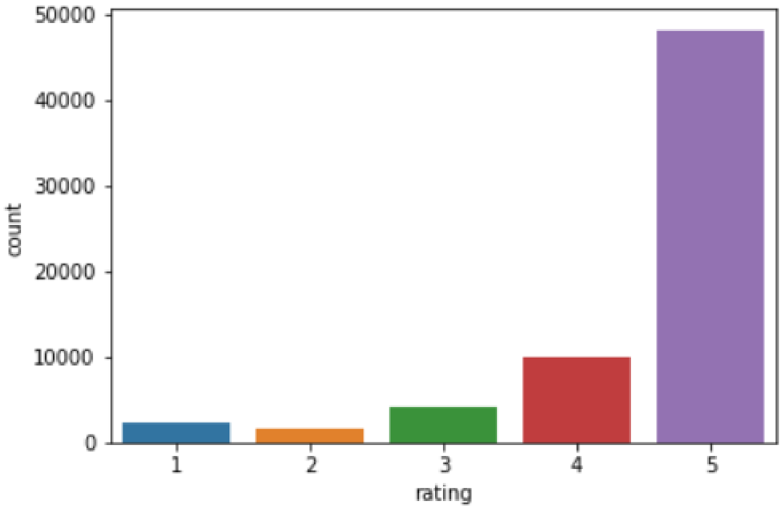
66352 rows × 3 columns

Preprocessing

Result

평점	Count
1	2,359
2	1,653
3	4,152
4	10,027
5	48,161
계	66,352

	count	mean	std	min	25%	50%	75%	max
rating	66352.00	4.51	0.98	1.00	4.00	5.00	5.00	5.00



Preprocessing

Label Encoding

	userId	productId	rating
3	942	1	4
16	3702	1	5
17	7603	1	5
59	3888	22	3
61	2511	20	4
...	...	...	...
1087017	6220	27640	4
1087053	3264	27641	1
1087058	6209	27642	5
1087100	1531	27645	5
1087107	3233	27644	5

66352 rows × 3 columns

Train Test Split

- test_size=0.05

	userId	productId	rating
1074638	1654	26991	5
613000	2204	25468	4
785216	1449	9383	1
217160	2580	12999	5
72169	7149	5434	5
...	...	...	...
943888	2301	20006	5
61096	7295	4759	5
291839	4401	15955	5
259866	332	14825	5
643433	1224	26357	4

63034 rows × 3 columns

	userId	productId	rating
749817	1307	6500	5
25103	1771	2253	5
478285	1310	21764	5
93533	2437	6618	5
456038	7523	21231	5
...	...	...	...
899498	4508	17325	5
54879	1597	4397	5
53186	2209	4364	5
685897	86	256	5
110253	7146	7632	5

3318 rows × 3 columns

Preprocessing

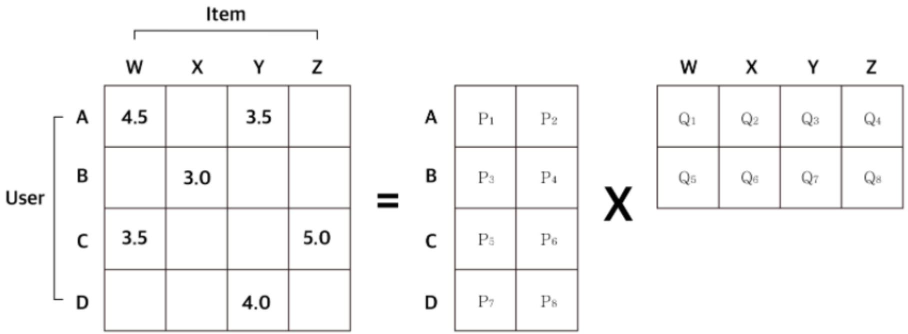
Rating Matrix

- Make rating matrix with train data set

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7929	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7930	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7931	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7932	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7933	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

7934 rows × 27646 columns

Matrix Factorization
(Gradient Descent)



Basic Concept

$$\mathbf{R} \approx \mathbf{P} \times \mathbf{Q}^T = \hat{\mathbf{R}}$$

$$\hat{r}_{ij} = p_i^T q_j = \sum_{k=1}^{k} p_{ik} q_{kj}$$

Formula

$$e_{ij}^2 = (r_{ij} - \textstyle\sum_{k=1}^{K} p_{ik} q_{kj})^2 + \frac{\beta}{2} \sum_{k=1}^{K} (||P||^2 + ||Q||^2)$$

$$p'_{ik} = p_{ik} + \alpha \frac{\partial}{\partial p_{ik}} e_{ij}^2 = p_{ik} + \alpha(2 e_{ij} q_{kj} - \beta p_{ik})$$
$$q'_{kj} = q_{kj} + \alpha \frac{\partial}{\partial q_{kj}} e_{ij}^2 = q_{kj} + \alpha(2 e_{ij} p_{ik} - \beta q_{kj})$$

Reference

https://minkithub.github.io/2020/06/15/matrix_factorization/

https://yamalab.tistory.com/92

Matrix Factorization
(Gradient Descent)

Training

- Latent parameter : 8
- Alpha on weight update : 0.1
- Beta on weight update : 0.1
- Training epochs : 200

Matrix Factorization
(Gradient Descent)

Result

```
-   Iteration: 10 ; cost = 0.2115
    Iteration: 20 ; cost = 0.1743
    Iteration: 30 ; cost = 0.1630
    ...
    Iteration: 180 ; cost = 0.1407
    Iteration: 190 ; cost = 0.1404
    Iteration: 200 ; cost = 0.1401
```

	0	1	2	3	4	5	6	7	8	9	10
0	2.91	4.67	4.79	4.39	4.72	4.74	4.78	5.01	5.28	4.68	4.83
1	2.28	4.65	4.86	4.16	4.24	4.56	4.46	5.87	5.33	4.39	4.63
2	2.46	4.14	4.54	3.94	4.28	4.34	4.39	3.98	5.33	4.30	4.42
3	1.47	4.19	3.72	3.36	3.48	3.52	3.86	5.05	3.94	3.64	3.81
4	3.07	5.02	5.07	4.55	4.99	5.02	4.96	3.48	5.49	4.84	5.19
...	...	...	...	...	...	...	...	...	...	...	...
7929	3.11	4.93	5.02	4.62	4.94	4.91	4.96	4.76	5.50	4.87	5.06
7930	3.18	5.04	5.17	4.72	5.01	5.03	5.08	4.95	5.72	4.98	5.14
7931	3.11	5.01	5.04	4.64	4.95	4.96	5.02	4.81	5.47	4.89	5.08
7932	3.12	4.92	5.00	4.61	4.93	4.91	4.97	4.71	5.47	4.86	5.03
7933	3.18	4.85	5.02	4.63	4.94	4.89	4.96	4.80	5.56	4.87	5.01

7934 rows × 27646 columns

Matrix Factorization
(Gradient Descent)

Result

- Inverse transform label

	1059950073	106171327X	140053271X	1935009311	1936170671	9565727689	9573212919	9573213893	9800359796	9803751263	9806010728	9806010914	9966569863
A0455374D4OQJ3XCOGGT	2.91	4.67	4.79	4.39	4.72	4.74	4.78	5.01	5.28	4.68	4.83	4.33	4.97
A100ULP2P8U71Z	2.28	4.65	4.86	4.16	4.24	4.56	4.46	5.87	5.33	4.39	4.63	4.09	4.75
A1012HOQZDKO80	2.46	4.14	4.54	3.94	4.28	4.34	4.39	3.98	5.33	4.30	4.42	4.35	4.52
A1013Q9SD2KIE1	1.47	4.19	3.72	3.36	3.48	3.52	3.86	5.05	3.94	3.64	3.81	3.40	4.02
A1028XZRNI0NRP	3.07	5.02	5.07	4.55	4.99	5.02	4.96	3.48	5.49	4.84	5.19	4.37	5.09
...	...	...	...	...	...	...	...	...	...	...	...	...	...
AZXQVHQVUB6W1	3.11	4.93	5.02	4.62	4.94	4.91	4.96	4.76	5.50	4.87	5.06	4.53	5.13
AZYZVPNT9YR2L	3.18	5.04	5.17	4.72	5.01	5.03	5.08	4.95	5.72	4.98	5.14	4.72	5.23
AZZEPOCDCD1PN	3.11	5.01	5.04	4.64	4.95	4.96	5.02	4.81	5.47	4.89	5.08	4.52	5.16
AZZW780H8VJ8N	3.12	4.92	5.00	4.61	4.93	4.91	4.97	4.71	5.47	4.86	5.03	4.52	5.12
AZZYJH0XNZ896	3.18	4.85	5.02	4.63	4.94	4.89	4.96	4.80	5.56	4.87	5.01	4.57	5.11

7934 rows × 27646 columns

Matrix Factorization
(Gradient Descent)

RMSE

- Test data : 3,318
- Fault : 0
- <rmse result> : 0.993

```
testing = X_test.values.tolist()
total = 0
fault_cnt = 0
for user, item, rating in testing:
  try:
    if inv.at[labels[0][user],labels[1][item]] > 5:
      total += pow(5 - rating, 2)
    elif inv.at[labels[0][user],labels[1][item]] < 0:
      total += pow(0 - rating, 2)
    else:
      total += pow(inv.at[labels[0][user],labels[1][item]] - rating, 2)
  except:
    fault_cnt += 1
print(len(testing))
print("fault : " + str(fault_cnt))
print("<rmse result>")
print(np.sqrt(total / (len(testing)-fault_cnt)))
```

```
3318
fault : 0
<rmse result>
0.9933968939433877
```

Matrix Factorization

Make recommendation

- Choose 2 user personas
- Picked ten recommended products

A1NSAEZV6AQP2X	
B00MVKUVLU	5.746542106269716
B01F477FGU	5.730950875882733
B001P3O3O0	5.663567840838001
B00SGO8TNC	5.654597784090665
B008UGMLWQ	5.650658692444763
B00359EXZG	5.644564594008128
B012DTEIVM	5.632355036349783
B0007LZH7S	5.616694873190977
B01BTR988M	5.597965152430909
B011IH6MI2	5.593652301642797

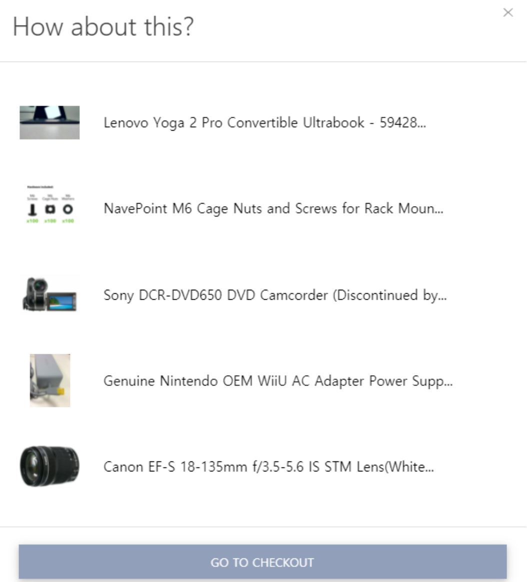
Personas 1

A1O09DT1PREH9Y	
B01F477FGU	5.969532289134793
B00MVKUVLU	5.947387374059887
B00SGO8TNC	5.828674618558825
B008UGMLWQ	5.8065453294618035
B0007LZH7S	5.782643237447696
B001P3O3O0	5.78054518865192
B00UVT5FC8	5.767132308793703
B00VICHYKS	5.754560076527632
B01BTR988M	5.754520466995791
B0075SUKIC	5.751409675844602

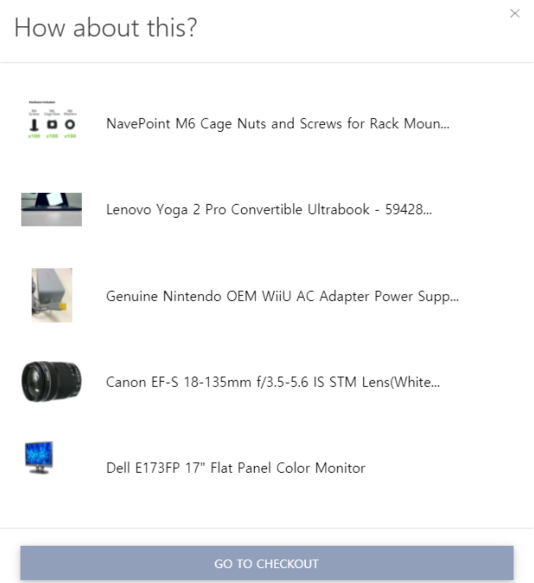
Personas 2

Matrix Factorization

Recommendation



Personas 1



Personas 2

Reason why MF(Gradient Descent)?

SVD in SURPRISE package

```
algo = SVD(n_factors=8, biased=False)
algo.fit(train)
```

```
<surprise.prediction_algorithms.matrix_factorization.SVD at 0x7f96f8af1cd0>
```

Result

```
predictions = algo.test(test)
print('RMSE : ', accuracy.rmse(predictions, verbose=False))
```

```
RMSE :  2.9554583706668343
```

Reason why MF(Gradient Descent)?

Matrix Factorization with ALS

- r_lambda = 40
- nf = 200
- alpha = 40

Result

- With Transaction : 66,352

사용 가능한 RAM을 모두 사용한 후 세션이 다운되었습니다. 런타임 로그 보기 ✕

- With Transaction : 8,920
  - rmse : 3.707

```
446
fault : 0
<rmse result>
3.707173355570345
```

Reference

Y. Hu et al, Collaborative Filtering for Implicit Feedback

5 쇼핑몰 시연

http://54.180.154.189:8000/

Thanks!

Any questions ?