데이트 마이닝

빅데이터 기반 서울 데이트 코스 추천

K H U P I D

경희대학교 산업경영공학과 김동혁 / 관광학과 류연주 / 산업경영공학과 유정수

KHUPID

INDEX

데 이 터 분 석 캡 스 톤 디 자 인

Table of Contents

데모 영상

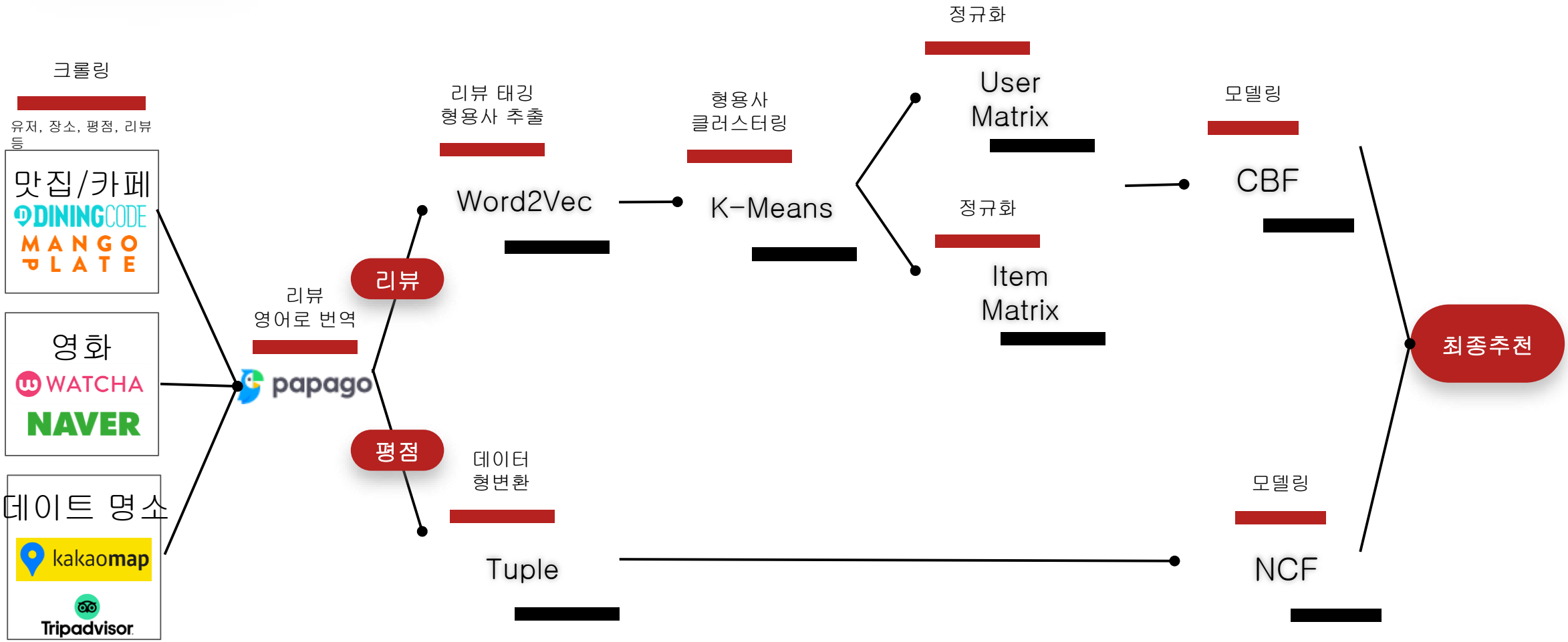
2020-1 데이터분석캡스톤 디자인

데이트 마이닝 : 빅데이터 기반 서울 데이트 코스 추천 알고리즘

데모영상

Team. KHUPID

플로우차트


크롤링
유저, 장소, 평점, 리뷰 등
맛집/카페
DINING CODE
MANGO PLATE
영화
WATCHA
NAVER
데이트 명소
kakaomap
Tripadvisor
리뷰 영어로 번역
papago
리뷰
평점
리뷰 태깅 형용사 추출
Word2Vec
데이터 형변환
Tuple
형용사 클러스터링
K-Means
정규화
User Matrix
정규화
Item Matrix
모델링
CBF
모델링
NCF
최종추천

NCF 모델 최종 학습

분야별 데이터 수(유저 수, 아이템 수, 평가 수)와 학습 방법

	EAT	GO	WATCH
Total users	22173	20246	14795
Total Items	5397	1260	800
Total Ratings	101595	41315	40232

모델 공통 파라미터

MLP layers = [64, 32, 16, 8]

Activation function = tanh

Optimizer = adam

Call back = 5 patience

Random search hyper parameter

Hyper parameter tuning을 위한 random search 시행 및 최적 파라미터 결과 성능(분야별)

```
# RandomSearch
rs_neumf = RandomSearch(
    model = neumf,
    space=[
        Discrete("num_epochs", [50, 100, 150, 200]),
        Discrete("num_factors", [4, 8]),
        Discrete("batch_size", [128, 256, 512]),
        Continuous("lr", 0.001, 0.01)
    ],
    metric = fm,
    eval_method = ratio_split
)
```

EAT

```
print('Random search: ', rs_neumf.best_params)
```

```
Random search:  {'batch_size': 512, 'lr': 0.0057864482838717955, 'num_epochs': 50, 'num_factors': 8}
```

Go

```
print('Random search: ', rs_neumf.best_params)
```

```
Random search:  {'batch_size': 512, 'lr': 0.002259556863673461, 'num_epochs': 50, 'num_factors': 4}
```

WATCH

```
print('Random search: ', rs_neumf.best_params)
```

```
Random search:  {'batch_size': 128, 'lr': 0.007501990443131995, 'num_epochs': 100, 'num_factors': 8}
```

NCF(Neural Collaborative Filtering) model

각 분야 별 결과

EAT 결과

	F1@10	NDCG@10	Precision@10	Recall@10	Train (s)	Test (s)
RandomSearch_NeuMF	0.0149	0.0313	0.0097	0.0540	14024.7585	11.2788

Go 결과

	F1@10	NDCG@10	Precision@10	Recall@10	Train (s)	Test (s)
RandomSearch_NeuMF	0.0970	0.2482	0.0556	0.4532	3795.3764	2.3491

WATCH 결과

	F1@10	NDCG@10	Precision@10	Recall@10	Train (s)	Test (s)
RandomSearch_NeuMF	0.0150	0.0283	0.0098	0.0535	4831.9023	1.0951

	u_id	p_id	ncf_score
0	lily	529.0	5.586952e-03
1	lily	522.0	6.249547e-05
2	lily	758.0	0.000000e+00
3	lily	438.0	2.086163e-07
4	lily	279.0	1.043427e-02
5	lily	788.0	0.000000e+00
6	lily	266.0	5.364418e-07
7	lily	593.0	0.000000e+00
8	lily	578.0	6.079674e-06
9	lily	735.0	4.678965e-06
10	lily	391.0	0.000000e+00
11	lily	632.0	1.490116e-07
12	lily	490.0	1.873225e-03
13	lily	530.0	0.000000e+00
14	lily	100.0	0.000000e+00
15	lily	552.0	1.192093e-07

ncf score data

KHUPID

리뷰 Word Clustering - word2vec, english

	comfortable	only	favorite	little	weak	first	want	female	overall	other	delicious
0	-0.000428	0.001228	-0.000180	0.000972	-0.000185	0.000038	0.000038	0.000261	0.000047	-0.000214	-0.000271
1	0.000183	0.000173	0.000307	0.001191	0.000233	0.000596	0.000923	0.000445	-0.000409	0.000150	-0.000466
2	0.000478	-0.000791	0.000173	-0.000499	0.001446	0.001369	0.000924	-0.000424	-0.000552	-0.000090	-0.000167
3	-0.000368	-0.001272	0.000377	0.000133	0.001009	0.000273	-0.000876	0.000082	-0.000340	0.000455	-0.000392
4	-0.000772	-0.000109	0.000543	-0.000014	0.000653	-0.000085	-0.000363	0.000610	-0.000167	-0.001110	-0.000231
...	...	...	...	...	...	...	...	...	...	...	...
295	0.000125	0.000061	-0.000115	0.000619	-0.001195	-0.001046	0.000390	0.000690	0.000301	0.000077	0.000119
296	-0.000071	0.000744	-0.000208	-0.000682	0.000181	0.000426	-0.000330	0.000058	-0.000014	0.000821	-0.000109
297	-0.000125	-0.000324	-0.000019	-0.000676	0.000222	0.000235	-0.000293	0.000132	-0.000472	-0.000355	0.000125
298	-0.000319	0.000006	-0.000451	0.000197	-0.000604	-0.000112	-0.000614	0.000107	-0.000610	0.000303	-0.000287
299	0.000050	-0.000063	-0.000577	0.000161	-0.000171	0.000441	0.000525	0.000271	0.000296	0.000548	0.000539

번역된 리뷰에서 tag 추출, word2vec으로 임베딩

0	1	2	3	4	5	6	7	8	9
cave	available	korea	kindness	hateful	sour	traditional	hidden	reasonable	atmosphere
ancient	uninvited	chinese	respectful	inflammatory	bitter	orthodox	invincible	sincere	busy
medieval	designated	indian	respect	NaN	acrimonious	conventional	silent	satisfactory	noisy
prehistoric	anonymous	asian	appreciate	NaN	NaN	unconventional	invisible	proper	calm
olden	unmanaged	african	gratitude	NaN	NaN	customary	forgotten	transparent	loud
paleolithic	mandatory	asia	praise	NaN	NaN	nonmainstream	unseen	doable	messy
primitive	specific	japan	NaN	NaN	NaN	mainstream	exist	fair	hectic
neolithic	optional	india	NaN	NaN	NaN	nontraditional	isolated	sufficient	chaotic
dinosaurs	preferred	mexican	NaN	NaN	NaN	NaN	unrecognized	appropriate	orderly
dinosaur	minimum	america	NaN	NaN	NaN	NaN	unaltered	rational	tangled
herbivorous	forbidden	africa	NaN	NaN	NaN	NaN	extinct	realistic	congested
NaN	onsite	australia	NaN	NaN	NaN	NaN	bankrupt	honest	disorderly

K-means clustering으로 word 그룹화

최종 클러스터 수 **EAT : 120**
GO : 100
WATCH : 100

user/item 클러스터별 점수 부여 및 정규화

u_id	0	1	2	3	4	5	6	7	8	9	...
5351	0.002153	0.068355	0.000000	0.007595	0.006359	0.136711	0.000000	0.034178	0.0	0.000000	...
3024	0.004500	0.000000	0.000000	0.003528	0.003692	0.000000	0.000000	0.031749	0.0	0.000676	...
5941	0.003436	0.009286	0.000164	0.002063	0.003455	0.000000	0.000432	0.013929	0.0	0.000395	...
13580	0.003460	0.000000	0.000000	0.000000	0.000000	0.000000	0.020440	0.000000	0.0	0.000000	...
6953	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...
...	...	...	...	...	...	...	...	...	...	...	...
u1s2&Koot	0.000000	0.000000	0.003412	0.000000	0.026897	0.000000	0.000000	0.000000	0.0	0.000000	...
u1s2&2019100861	0.000000	0.000000	0.010665	0.000000	0.046711	0.000000	0.000000	0.000000	0.0	0.000000	...
u1s2&Uynibus	0.000000	0.000000	0.010816	0.000000	0.018949	0.000000	0.000000	0.000000	0.0	0.000000	...
u1s2&dpwls258	0.000000	0.000000	0.010068	0.000000	0.026458	0.000000	0.000000	0.000000	0.0	0.000000	...
u1s2&쥬	0.000000	0.000000	0.003287	0.000000	0.034548	0.000000	0.000000	0.000000	0.0	0.000000	...

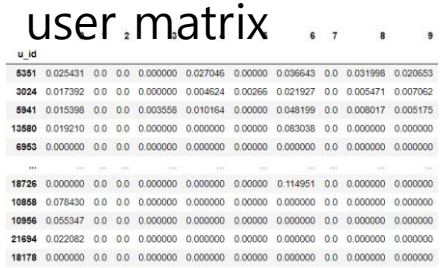
설문을 통해 얻은 새로운 user data 추가

KHUPID

CBF 점수 산출

유저와 장소간의 코사인 유사도를 통해 CBF 점수 산출

user matrix



u_id										
5351	0.025431	0.0	0.0	0.000000	0.027046	0.00000	0.036643	0.0	0.031998	0.020653
3024	0.017392	0.0	0.0	0.000000	0.004624	0.00265	0.021927	0.0	0.005471	0.007062
5941	0.015398	0.0	0.0	0.003558	0.010164	0.00000	0.048199	0.0	0.008017	0.005175
13580	0.019210	0.0	0.0	0.000000	0.000000	0.00000	0.083038	0.0	0.000000	0.000000
6953	0.000000	0.0	0.0	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...
18726	0.000000	0.0	0.0	0.000000	0.000000	0.00000	0.114951	0.0	0.000000	0.000000
10858	0.078430	0.0	0.0	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000
10956	0.055347	0.0	0.0	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000
21694	0.022082	0.0	0.0	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000
18178	0.000000	0.0	0.0	0.000000	0.000000	0.00000	0.000000	0.0	0.000000	0.000000

u1s2&쥬	0.000000	0.000000	0.003287	0.000000	0.034548	0.000000	0.000000	0.000000	0.0	0.000000

p_id	0	1	2	3	4	5	6	7	8	9
721	0.000000	0.000000	0.000000	0.00000	0.0	0.000000	0.000000	0.0	0.000000	0.000000
1832	0.053308	0.000000	0.000000	0.00000	0.0	0.000000	0.046086	0.0	0.000000	0.000000
741	0.034111	0.000000	0.000000	0.00000	0.0	0.000000	0.000000	0.0	0.000000	0.055405
218	0.042261	0.000000	0.000000	0.00000	0.0	0.077576	0.182678	0.0	0.079761	0.000000
3276	0.000000	0.000000	0.000000	0.00000	0.0	0.000000	0.000000	0.0	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...
3793	0.003882	0.000000	0.000000	0.00000	0.0	0.000000	0.000000	0.0	0.000000	0.018916
2809	0.016743	0.000000	0.000000	0.00000	0.0	0.006147	0.043423	0.0	0.018959	0.024475
3297	0.016136	0.002938	0.001383	0.00235	0.0	0.006439	0.021229	0.0	0.007945	0.010256
445	0.000000	0.000000	0.000000	0.00000	0.0	0.000000	0.039573	0.0	0.000000	0.000000
3592	0.000000	0.000000	0.000000	0.00000	0.0	0.000000	0.060478	0.0	0.000000	0.000000

두 벡터간의 cosine similarity 측정

	preference
0	0.004246
1	0.000000
2	0.001651
3	0.000020
4	0.000324
...	...
1404	0.000004
1405	0.000000
1406	0.000082
1407	0.000016
1408	0.000262

Place matrix

user에 특성에 맞는 item을 탐색 후 ranking

모델 결과 결합

CBF 모델 결과

p_id	cbf_score
643	0.936923
962	0.935923
2481	0.928743
1210	0.926539
3554	0.926400
...	...

X

NCF 모델 결과

p_id	ncf_scores
2210	0.013282
1944	0.000021
420	0.015204
2900	0.029513
4571	0.161402
...	...

=

p_id	name	address	result
4142	중앙해장	서울특별시 강남구 영동대로86길 17 육인빌딩	0.532276
1068	동경산책	서울특별시 성북구 보문로34길 45	0.505191
1712	몽탄	서울특별시 용산구 백범로99길 50	0.503543
3468	은달 왕 돈까스	서울시 성북구 동선동1가 2-9	0.491465
1658	모리돈부리	서울특별시 관악구 관악로12길 6	0.458265
2704	쉐이크쉑 두타점	서울특별시 중구 을지로6가 18-12	0.454874
4412	카레시오	서울특별시 관악구 관악로14길 28	0.453978
1587	메이크샐러드	서울특별시 동대문구 휘경로2길 16 1F	0.448346
747	뉴질랜드스토리	서울특별시 송파구 송파동 32-1	0.434375
4772	토끼정	서울특별시 강서구 하늘길 38 롯데몰 김포공항점 MF	0.433607

지역구 필터링

EAT

p_id	name	address	district	result
972	더훈	서울특별시 용산구 독서당로 87	용산구	0.117684
2778	스시올로지	서울특별시 마포구 동교로 266-11	마포구	0.058028
5186	할리스커피	서울특별시 금천구 가산동 459-11	금천구	0.010796
5281	호 파스타	서울시 광진구 화양동 12-52	광진구	0.007456
2892	식샤	서울특별시 성북구 성북동1가 36-1	성북구	0.004602
...	...	...	...	...
5415	후니도니	서울특별시 종로구 종로 19 르메이에르종로타운 B1	종로구	NaN
1519	맛짱	서울특별시 중랑구 중랑역로 116	중랑구	NaN
3843	이탈리아노501	서울특별시 도봉구 해등로32길 76	도봉구	NaN
2480	서초장어타운	서울특별시 서초구 반포대로28길 77 큰대문집	서초구	NaN
5425	후토이	서울특별시 강서구 강서로7길 42	강서구	NaN

GO

p_id	p_name	address	district	result
g84	도깨비코인노래연습장	서울 광진구 자양로18길 19 해민빌딩 3층 (우)05043	광진구	0.045244
f231	신촌반지클럽	서울 서대문구 연세로 25-1 5층 (우)03788	서대문구	0.018419
3	국립민속박물관&국립민속박물관 어린이박물관	종로구	종로구	0.014978
c113	해브어코믹스데이	서울 마포구 도화4길 23 2층 (우)04169	마포구	0.004622
a59	스핀볼링센터 가든5점	서울 송파구 중민로 52 지하1층 57호 (우)05839	송파구	0.004398
...	...	...	...	...
387	장충동 족발 골목	중구	중구	0.000000
e38	엑스케이프 홍대점	서울 마포구 홍익로3길 44 호곡빌딩 7층 (우)04039	마포구	0.000000
e63	셜록홈즈 영등포점	서울 영등포구 영중로4길 25-1 4층 (우)07304	영등포구	0.000000
e67	키이스케이프 강남더오름	서울 강남구 테헤란로6길 30 지하1층 (우)06240	강남구	0.000000
g92	세븐스타코인노래연습장 장안동점	서울 동대문구 답십리로72길 28 2층 (우)02637	동대문구	0.000000

user가 희망하는 지역구 중
result의 합을 가장 크게만드는 지역구로 추천

최종 결과 출력

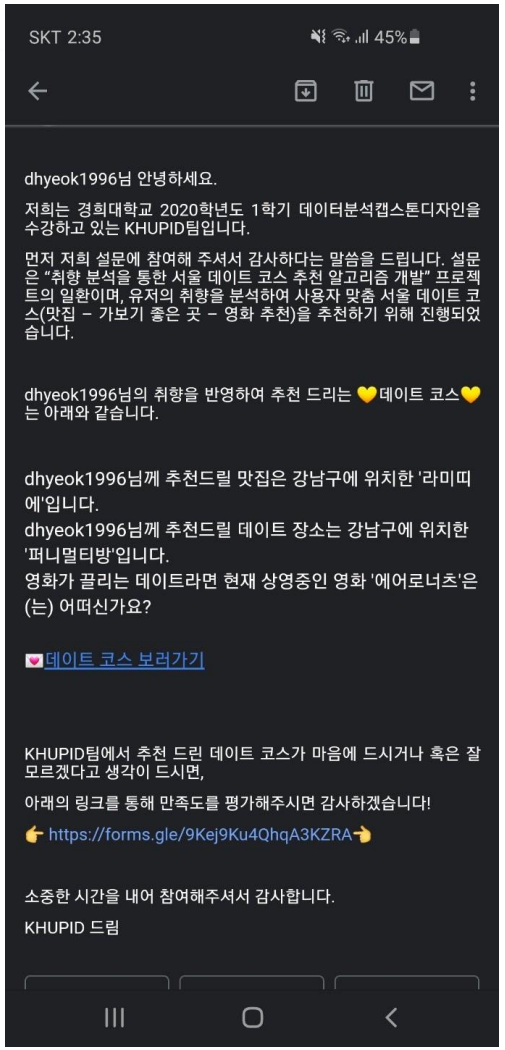
지역구를 한 개만 선택했을 경우

```
dhyeok1996님께 추천드릴 맛집은 강남구에 위치한 '라미띠에'입니다.
dhyeok1996님께 추천드릴 데이트 장소는 강남구에 위치한 '퍼니멀티방'입니다.
영화가 끌리는 데이트라면 현재 상영중인 영화 '에어로너츠'은(는) 어떠신가요?
```

지역구를 두 개 이상 선택했을 경우

```
쥬님! 용산구에서의 데이트를 계획하고 계신가요?
쥬님께 추천드릴 맛집은 용산구에 위치한 '더훈'입니다.
쥬님께 추천드릴 데이트 장소는 용산구에 위치한 '남영볼링센터'입니다.


혹시나! 광진구에서의 데이트를 계획하고 계신다면?
맛집으로는 광진구에 위치한 '호 파스타'와
데이트 장소로는 광진구에 위치한 '도깨비코인노래연습장'을(를) 추천드려요.
영화가 끌리는 데이트라면 현재 상영중인 영화 '로스트 인 파리'은(는) 어떠신가요?
```



User 별 추천 결과를 지도 링크와 함께 전송

흥미로웠던 부분

어느 가족	3
에어로너츠	1
도쿄 타워	1
결백	3
시간을 달리는 소녀	1
마담 프루스트의 비밀정원	1
1917	1
너의 췌장을 먹고싶어	3
프랑스여자	2
나는보리	1
콜 미 바이 유어 네임	1
안녕, 미누	1
레미: 집 없는 아이	1
프리즌 이스케이프	2
인생	1
배짱이들	1
국도극장	1
바닷마을 다이어리	1
전망 좋은 방	1
검은 여름	1
아홉 스님	1

영화: 한쪽에 치우치지 않고 user별로 골고루 추천함

lseinuu님! 영등포구에서의 데이트를 계획하고 계신가요?
lseinuu님께 추천드릴 맛집은 영등포구에 위치한 '양키스그릴'입니다.
lseinuu님께 추천드릴 데이트 장소는 영등포구에 위치한 '한강시민공원 여의도지구(여의도한강공원)'입니다.
💌 첫번째 데이트 코스 보러가기

혹시나! 서초구에서의 데이트를 계획하고 계신다면?
맛집으로는 서초구에 위치한 '라모라'와
데이트 장소로는 서초구에 위치한 '뒷벌공원'을(를) 추천드려요.

💌 두번째 데이트 코스 보러가기

🎬 영화가 끌리는 데이트라면 현재 상영중인 영화 '너의 췌장을 먹고 싶어'은(는) 어떠신가요?

데이트 장소: 지역구에 상관없이 둘 다 공원을 추천함

흥미로웠던 부분2

쥬님! 용산구에서의 데이트를 계획하고 계신가요?
쥬님께 추천드릴 맛집은 용산구에 위치한 '더훈'입니다.
쥬님께 추천드릴 데이트 장소는 용산구에 위치한 '남영볼링센터'입니다.

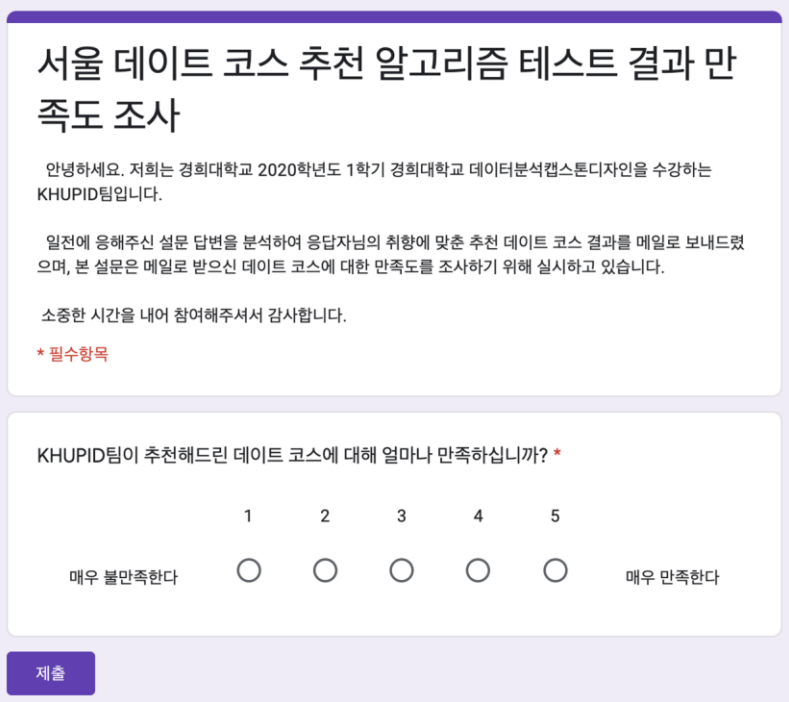
혹시나! 광진구에서의 데이트를 계획하고 계신다면?
맛집으로는 광진구에 위치한 '호 파스타'와
데이트 장소로는 광진구에 위치한 '도깨비코인노래연습장'을(를) 추천드려요.
영화가 끌리는 데이트라면 현재 상영중인 영화 '로스트 인 파리'은(는) 어떠신가요?

*연*의남자님! 노원구에서의 데이트를 계획하고 계신가요?
*연*의남자님께 추천드릴 맛집은 노원구에 위치한 '로지다이닝'입니다.
*연*의남자님께 추천드릴 데이트 장소는 노원구에 위치한 '마이툰'입니다.

혹시나! 송파구에서의 데이트를 계획하고 계신다면?
맛집으로는 송파구에 위치한 '맥주유소'와
데이트 장소로는 송파구에 위치한 '세븐스타코인노래연습장 송파구청점'을(를) 추천드려요.
영화가 끌리는 데이트라면 현재 상영중인 영화 '어느 가족'은(는) 어떠신가요?

데이트 장소: 실제 커플인 사용자에게 둘 다 '코인노래방'을 추천함

최종 만족도 조사



서울 데이트 코스 추천 알고리즘 테스트 결과 만족도 조사

안녕하세요. 저희는 경희대학교 2020학년도 1학기 경희대학교 데이터분석캡스톤디자인을 수강하는 KHUPID팀입니다.

일전에 응해주신 설문 답변을 분석하여 응답자님의 취향에 맞춘 추천 데이트 코스 결과를 메일로 보내드렸으며, 본 설문은 메일로 받으신 데이트 코스에 대한 만족도를 조사하기 위해 실시하고 있습니다.

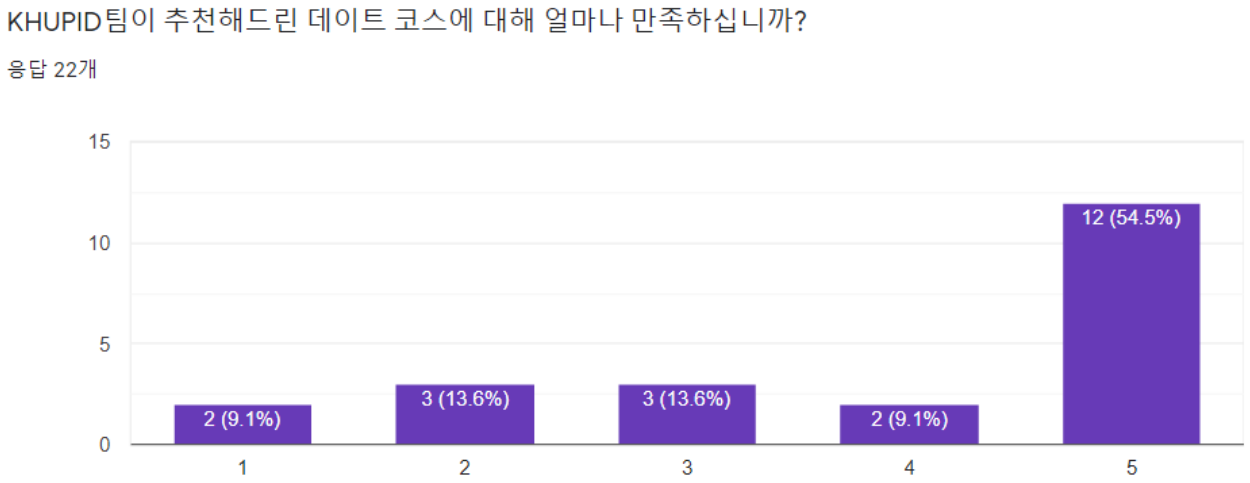
소중한 시간을 내어 참여해주셔서 감사합니다.

* 필수항목

KHUPID팀이 추천해드린 데이트 코스에 대해 얼마나 만족하십니까? *

	1	2	3	4	5	
매우 불만족한다	○	○	○	○	○	매우 만족한다

제출



평균 **3.863636364**

시사점

1. NCF의 cold start 문제를 해결하기위해 contents based filtering을 도입
-> 최근 추천 트렌드에서 아직까지 좋은 성능을 발휘하는 hybrid 방식으로 추천

2. 기존의 hybrid 방식에서의 CBF는 item의 특성(영화로 따지면 배우, 감독, 장르, 줄거리 등)의 유사도를 비교했지만,

우리 프로젝트에선 해당 리뷰데이터를 통해 user와 item에 대한 특성을 재정의하고 취향을 반영함

아쉬웠던 점, 한계점

시간 제약 관계로 앱 서비스로 제작을 못한 것,
그로 인해 2차 만족도 설문 및 신규 사용자에 대한 모델 강화 부분을 하지 못한 것

동혁

- 리뷰들을 현재 운영되고 있는 맛집이나 영화, 지도 사이트 등에서 크롤링 할 수밖에 없다는 것. 실제 서비스를 운영하며 데이터를 축적하면 좀 더 원활하게 구현할 수 있을 듯하다.

연주

- 각각 다른 플랫폼들에서 데이터를 가져오게 되어서, 아이템과 리뷰를 매칭시키는데 한계가 있었다.

정수

- 리뷰에서 장소를 잘 나타내는 형용사를 추출하는 것이 중요한데 그 부분이 부족했다고 생각

감사합니다.