

카페추천시스템

[목표]

카페를 방문하는 목적이 다양해졌으므로 목적에 맞는 카페를 추천해주는 시스템이 필요

[역할]

데이터 수집 자동화, 전처리, 평가, 시각화

[데이터]

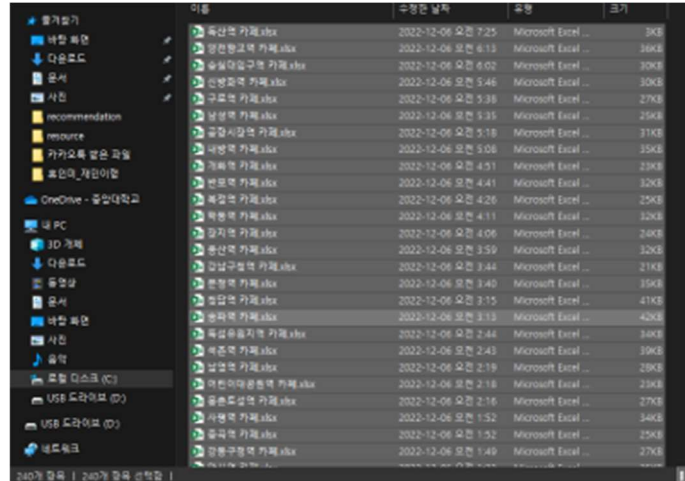


종	메뉴	리뷰	사진
어떤 점이 좋았어요? 📌			
✓ 55회 (47명 참여) 나도 참여			
☕	"커피가 맛있어요"	26	
📖	"집중하기 좋아요"	18	
👑	"태강이 청원해요"	15	
💖	"친절해요"	14	
🍹	"음료가 맛있어요"	10	
🏠	"인테리어가 멋져요"	10	
🍰	"디저트가 맛있어요"	9	
💰	"가성비가 좋아요"	7	
💬	"대화하기 좋아요"	7	

원하는 데이터가 없어 카페 데이터를 직접 크롤링을 통해 수집하기로 했다. 여러 웹 지도들을 살펴본 결과 '네이버 지도'의 리뷰 탭에서 이미 labeling된 지표들이 있어 활용하기로 했고, 영수증 인증을 통해 실제로 구매한 사람들만 리뷰를 남길 수 있다는 점에서 신뢰성도 확보했다.

```
14 browser = webdriver.Chrome("./chromedriver.exe")
15 browser.get("https://map.naver.com/v5/")
16 browser.implicitly_wait(10)
17 browser.maximize_window()
18
19 search = browser.find_element_by_css_selector("input.input_search")
20 search.click()
21 time.sleep(1)
22 search.send_keys("강남역 카페")
23 time.sleep(1)
24 search.send_keys(Keys.ENTER)
25 time.sleep(2)
```

위 스크린샷은 크롤링 코드의 일부분으로 실제 크롤링은 서울시에 위치한 역사명을 받아, 자동화된 크롤링을 수행하게 된다



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	0	히프스베이	이 맞!	681	전철역	249	특별한 메!	245	가성비가!	121	매장이 청!	74	음료가 맞!	25	커피가 맞!	15	인테리어?	11	대화하기!	6	좌석이 편!	4	사진이 잘
2	0	아미랑스	커피가 맞!	470	인테리어?	284	디저트가!	271	음료가 맞!	199	매장이 청!	190	전철역	156	사진이 잘	120	대화하기!	104	화장실이!	88	부가 좋아!	86	좌석이 편!
3	0	솔로랜드	커피가 맞!	117	전철역	91	음료가 맞!	54	인테리어?	54	매장이 청!	37	가성비가!	35	화장실이!	30	디저트가!	29	대화하기!	28	사진이 잘!	19	아늑해요!
4	0	다정도	병! 음료가 맞!	135	디저트가!	102	특별한 메!	92	전철역	79	인테리어?	63	매장이 청!	54	대화하기!	54	좌석이 편!	44	커피가!	39	부가 좋아!	3	화장실이!
5	0	이즈메이	가성비가!	37	전철역	35	디저트가!	32	커피가!	31	특별한 메!	18	음료가!	17	매장이 청!	14	인테리어?	4	대화하기!	3	사진이 잘!	2	부가 좋아!
6	0	블랙타운	커피가 맞!	503	전철역	302	음료가!	290	특별한 메!	178	매장이 청!	98	가성비가!	68	디저트가!	39	인테리어?	39	대화하기!	38	접종하기!	31	좌석이 편!
7	0	일구구일	병! 맞!	87	전철역	59	특별한 메!	43	매장이 청!	33	가성비가!	19	인테리어?	15	음료가!	4	커피가!	3	사진이 잘!	2	접종하기!	2	주차하기!

‘서울시 지하철역 + 카페’ 라는 쿼리로 크롤링을 진행하여 총 2만여개의 카페 데이터와 103개의 feature를 수집 완료했다. 그 후 데이터들을 하나의 파일로 합치고 전처리를 진행했다.

1. 중복제거 : 역간 거리가 짧은 경우 중복되는 카페들이 있다.
2. 이상치 제거 : 방탈출카페, 팝업스토어 등 목적 데이터와 다른 데이터들을 제거했다.
3. Relabeling : 성능을 높이기 위해 지표를 통합/폐기하여 relabeling을 진행했다.

열1	friendly	dessert	interior	clean	coffee	special	bread	photo	non_coffee	tea	fresh	seat	talk
히프스베이	249	0	11	74	15	245	681	4	25	0	0	4	6
아미랑스	156	271	284	190	470	45	0	120	199	0	0	85	104
솔로랜드	91	29	54	37	117	9	0	19	54	5	0	4	28
다정도 병인 양	79	102	63	54	39	92	0	12	135	0	0	44	54
스즈메이글 맥대점	35	32	4	14	31	18	0	2	17	0	0	1	3

4. 스케일링 : 추후 모델링을 위해서 모든 데이터를 최대값으로 나눠서 0~1 사이 값을 갖도록 했다. (정규화)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	name	friendly	dessert	interior	clean	coffee	special	bread	photo	non_coffee	tea	fresh	seat	talk	concentrate	view	sofa	parking	glenty	food	big	cost	concept
2	히프스베이	0.365639	0	0.016153	0.130664	0.022026	0.359765	0	0.005874	0.036711	0	0	0.005874	0.008811	0.001468	0.005874	0.004405	0.001468	0	0	0	0.17768	0
3	아미랑스	0.331915	0.576596	0.040255	0.404255	1	0.095745	0	0.255319	0.423404	0	0	0.100851	0.221277	0.031915	0.182979	0.187354	0.178723	0	0	0	0.06385	0
4	솔로랜드	0.777778	0.247863	0.461538	0.336239	1	0.076923	0	0.362393	0.461538	0.042735	0	0.034188	0.239316	0.545299	0.111111	0.25641	0.025641	0	0	0	0.299145	0
5	다정도 병	0.585185	0.755556	0.466667	0.4	0.280889	0.681481	0	0.068889	1	0	0	0.325926	0.4	0.074074	0.17037	0.140741	0.148155	0	0	0	0.058259	0
6	이즈메이	0.545946	0.864865	0.308106	0.378378	0.837838	0.486486	0	0.054054	0.459459	0	0	0.027027	0.081081	0.027027	0.054054	0	0	0	0	0	1	0
7	블랙타운	0.600398	0.077335	0.077335	0.194831	1	0.353877	0	0.011828	0.576941	0	0	0.053678	0.075547	0.06163	0.017893	0.00994	0.00994	0	0	0	0.135189	0
8	일구구일	0.678161	0	0.172414	0.37931	0.054483	0.494253	1	0.022989	0.043977	0	0	0	0	0.022989	0	0	0.011494	0	0	0	0.218391	0
9	봉보루	0.415761	0.63587	0.475543	0	1	0	0	0	0	0	0	0	0	0	0.807065	0	0	0	0	0	0	0
10	분류해빙	0.392523	0	0.096075	0.411215	0	0.093458	0	0	0	0	0.280374	0	0	0.018692	0.009546	0.009546	0.17757	1	0.252336	0.064112	0	
11	고원 다지	0.726457	1	0.130045	0.475336	0.264574	0.363229	0	0.040359	0.192825	0	0	0.004484	0.017937	0.008969	0.004484	0.008969	0.013453	0	0	0	0.403587	0
12	카페 라지	0.461538	0.461538	0.230769	0.410256	1	0.025641	0.076923	0.025641	0.530462	0	0	0.128205	0.051282	0.025641	0	0.153846	0.051282	0	0	0	0	0
13	도너스콜	0.406842	1	0.039474	0.210526	0.197360	0.542105	0	0.078947	0.078947	0	0	0.026316	0	0	0.013158	0	0	0	0	0	0.157895	0
14	문래리	0.744681	0.574468	0.06383	0.425332	1	0.148936	0	0.306383	0.404255	0	0	0.212766	0.212766	0.042553	0	0.12766	0.021277	0	0	0	0.255319	0
15	비로계	0.429907	0.457944	0.336449	0.345394	1	0.074766	0	0.540187	0.364486	0	0	0.11215	0.545794	0.093458	0.35514	0.17757	0.065421	0	0	0	0.046729	0
16	메가MOC	0.444053	0.205882	0.073529	0.316176	1	0.079044	0	0.025735	0.543956	0	0	0.220588	0.172794	0.082721	0.036785	0.064338	0.007353	0	0	0	0.854779	0
17	고양명비	0.473604	1	0.210526	0.447368	0.421053	0.5	0	0.052632	0.184211	0	0	0.026316	0.026316	0	0.026316	0.026316	0.026316	0	0	0	0.236842	0
18	모글 베스	0.690476	1	0.095238	0.452381	0.230895	0.285714	0	0.02381	0.190476	0	0	0.02381	0.047619	0.02381	0	0	0	0	0	0	0.214206	0
19	커피매점	0.386364	0.25	0.254545	0.068182	1	0.386364	0	0.045455	0.568182	0	0	0.045455	0.090909	0	0	0.022727	0	0	0	0	0.294545	0
20	노리	0.209677	1	0.258065	0.185484	0.516129	0.314839	0	0.040123	0.17742	0	0	0.046129	0.193548	0.056432	0.016129	0.016129	0.016129	0	0	0	0.040123	0
21	아디에 A1	0.44136	0.232558	0.069767	0.232558	1	0.116279	0	0.023256	0.534854	0	0	0.116279	0.232558	0.116279	0.069767	0.093023	0.069767	0	0	0	0.395349	0
22	21stLab	0.407678	0	0.038718	0.197678	0.069696	0.064553	1	0.038718	0.068117	0	0	0.038718	0.038718	0.038718	0.038718	0.069696	0.007188	0	0	0	0.068117	0

추천을 위해 알고리즘을 선택해야 했다. 유저 피드백이 없는 Cold start 상태이기에 content based filtering을 바탕으로 하기로 했고 LightFM이라는 모델을 사용하여 cold start 문제를 그나마 해결하려고 했다.

[평가]

```

cafe_name = '서촌금상고로케'

result = recommend_cafe_list(data, cafe=cafe_name)

# data['name'] = data['열1']
# data.set_index('열1', inplace=True)
data.rename(columns={'열1': 'name'}, inplace=True)
result.rename(columns={'열1': 'name'}, inplace=True)
index = data.index[(data['name'] == cafe_name)]

user = data.iloc[index]

print(result['name'])

```

✓ 28.8s

14997 카페 홍대점
8497 부트브래드
14900 꿀넉쿠키 연남점
1037 일팔공일오
11737 뽕미제빵소
Name: name, dtype: object

유저가 긍정적으로 평가한 카페 명을 입력값으로 받으면, 해당 카페와 비슷한 5개의 카페들을 추천해주도록 설계했다.

왼쪽은 예시로 '서촌금상고로케'와 비슷한 feature들을 가진 카페들을 추천해주었고 아래는 해당 카페들의 feature로 비슷한 값을 갖고 있음을 알 수 있다.

열1	friend	dessert	interio	clean	coffee	special	bread	photo	non_co	tea	fresh	seat	talk
일팔공일오	174	0	26	97	36	126	326	9	10	0	0	3	2
서촌금상고로케	2397	0	115	1499	191	2250	2904	224	277	0	0	115	214
부트브래드	130	0	13	61	40	61	202	4	20	0	0	2	3
뽕미제빵소	131	0	3	98	30	89	231	0	17	0	0	0	1
꿀넉쿠키 연남점	162	0	46	98	38	176	209	30	28	0	0	9	32
카페 홍대점	266	0	111	152	63	363	478	96	63	0	0	34	48

```

# Lightgbm을 구현하여 shop value를 예측할 것
# lighgbm 구현

# library
import lightgbm as lgb # 없을 경우 cmd/anaconda prompt에서 install (LightGBM: Light Gradient-Boosting Machine)
from math import sqrt
from sklearn.metrics import mean_squared_error

# Lightgbm model
lgb_dtrain = lgb.Dataset(data = train_x, label = train_y) # LightGBM 모델에 맞게 변환
lgb_param = {'max_depth': 20, # original: 10
             'learning_rate': 0.01, # Step Size
             'n_estimators': 1000, # Number of trees
             'objective': 'regression'} # 목적 함수 (L2 Loss)
lgb_model = lgb.train(params = lgb_param, train_set = lgb_dtrain) # 학습 진행
lgb_model_predict = lgb_model.predict(test_x) # test data 예측
print("RMSE: {}".format(sqrt(mean_squared_error(lgb_model_predict, test_y)))) # RMSE

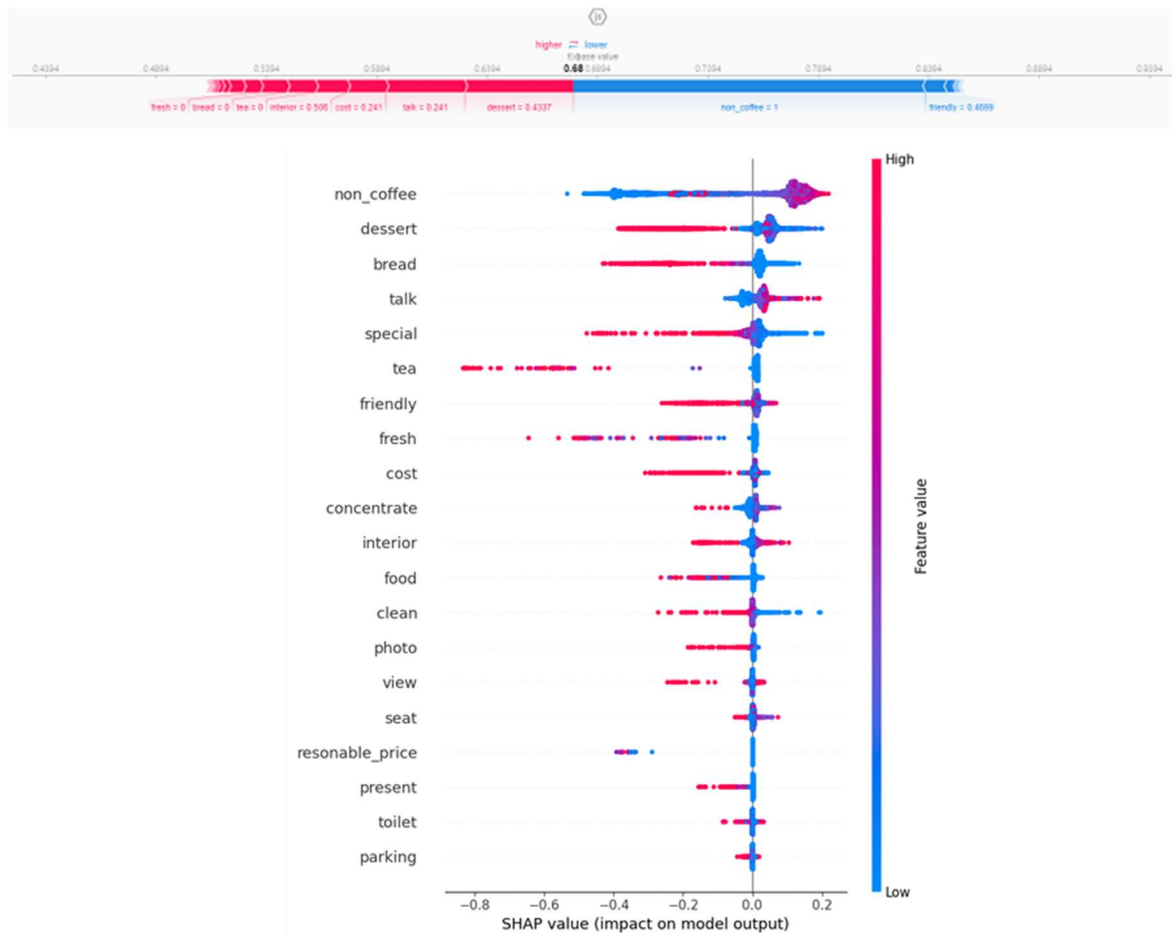
```

✓ 1.1s

[LightGBM] [Warning] Accuracy may be bad since you didn't explicitly set num_leaves OR 2*max_depth > num_leaves. (num_leaves=31).
[LightGBM] [Warning] Accuracy may be bad since you didn't explicitly set num_leaves OR 2*max_depth > num_leaves. (num_leaves=31).
[LightGBM] [Warning] Auto-choosing col-wise multi-threading, the overhead of testing was 0.001783 seconds.
You can set 'force_col_wise=true' to remove the overhead.
[LightGBM] [Info] Total Bins 4493
[LightGBM] [Info] Number of data points in the train set: 10682, number of used features: 41
[LightGBM] [Info] Start training from score 0.689399
RMSE: 0.1278267900246556

LightFM을 통해 정확도 평가를 진행했을 때 RMSE값이 0.1278로 높은 정확도를 갖고 있음을 알 수 있다.

LightFM 모델을 사용해 예측한 결과에 대해서 해석/평가하기 위해 SHAP를 사용했다. SHAP는 우리가 relabeling한 특성들이 모델의 예측에 얼마나 기여하는지를 나타내고 특성들간의 상관관계를 나타낼 수 있었다.



해당 그래프들은 '커피가 맛있다'라는 지표로 기준으로 양/음의 상관관계를 가진 칼럼들을 보여준 것이다. 예를 들면 '커피가 맛있다'고 평가한 고객들은 '디저트가 맛있다'라는 평가도 같이 남기지만 '논커피가 맛있다'라는 평가는 남기지 않는 경향이 있다는 것을 알 수 있었다.

