



포스트 코로나 시대의 지역별 문화·여가생활을 위한 정책 제언

대한민국 17개 시도

DATA 53팀

목차

001 도입

- 현황 및 문제제기
- 주제선정
- 주제선정 및 과거 데이터 활용 방안

002 데이터

- 데이터 선정
- 데이터 전처리

003 분석 및 결과

- 국민여가 활동조사
 - 트리기반(Tree-based) 머신러닝
 - 주요 변수 추출
- 신한카드 데이터 토픽모델링
- 트립 어드바이저 크롤링

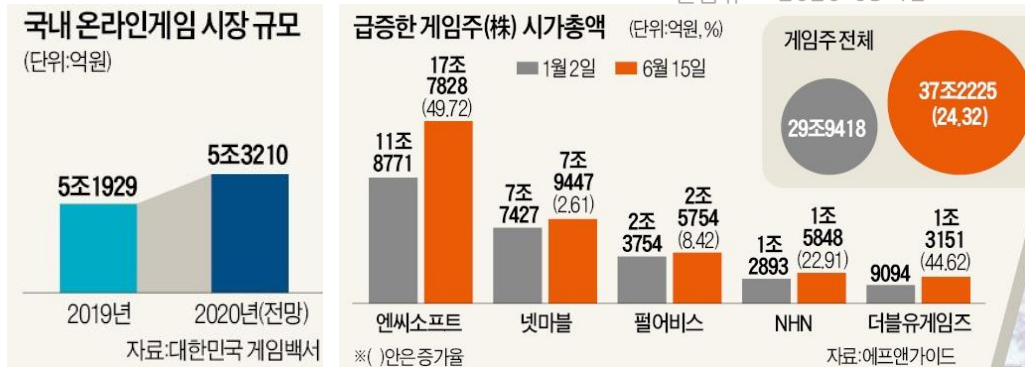
004 지역별 문화·여가 생활 위한 정책 제언

005 분석 의의 및 활용 방안 제언

1. 도입 코로나시대의 문화·여가생활 현황 및 문제제기

● 온라인 콘텐츠 소비 증가

- “코로나 '집콕'에 넷플릭스 '월월'..4월 439억 역대 최대 매출”
YTN뉴스 2020-05-26
- “코로나19로 한국 영화산업 매출 70%↓, 2만명 고용불안 전망”
연합뉴스 2020-05-12

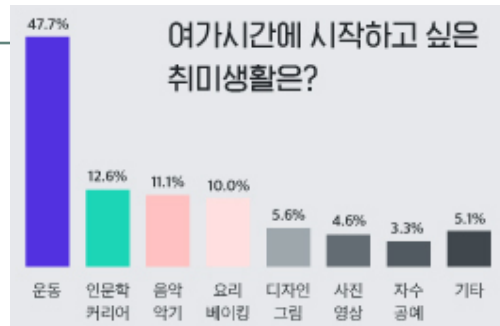


● 여가 시간의 증가, 취미생활의 수요 증가

잡코리아 · 알바몬이 성인 676명을 대상으로 '코로나19 발생 전후 여가 트렌드'에 대해 조사한 결과, 개인 여가시간이 코로나19 발생 이후 늘었다고 답한 사람이 67.5%에 달했다.

재택근무로 증가된 여가시간에 취미생활을 시작할 의향은?

있다 89.7% 없다 10.3%



출처: 동아일보 직장인 58% 재택근무 경험...90% “증가된 여가시간에 새 취미생활 하고파” 2020-03-31

● 증가한 여가 수요에 대한 공급 부족

여가시간을 보내는 방식도 달라졌다. 코로나 전후 여가시간을 어떻게 보내고 있는지 조사한 결과, 코로나19 발생 이전에는 여가시간이 생기면 ‘지인들과 만남을 가졌다’는 응답자가 가장 많았던 반면 **코로나19 발생 이후에는 여가시간이 생기면 ‘집에서 TV나 영화를 시청한다’고 답한 응답자가 가장 많았다.**

● 문화예술산업, 스포츠 산업, 관광산업의 위기에 따른 여가 생활 향유의 기회 감소



● 1인 활동의 증가 “코로나19에 혼자 여행 간다...캠핑·낚시용품 판매 증가” 연합뉴스 2020-05-01

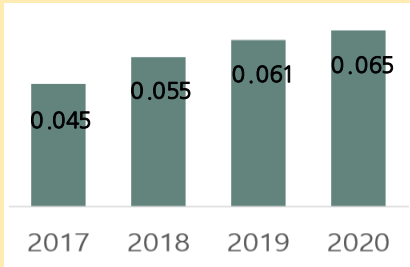
● 문제제기 “코로나로 인해 약해진 여가활동 인프라, 포스트 코로나의 급증하는 여가 수요 감당할 수 있을까?”

1. 도입 주제선정 및 과거 데이터 활용 방안

주제선정: 포스트 코로나 시대에 급증할 여가 활동을 위한 선제적 대응, 인프라 구축 필요

Q. 코로나로 인한 새로운 상황, 어디까지 미래의 대안에 반영해야 하는가?

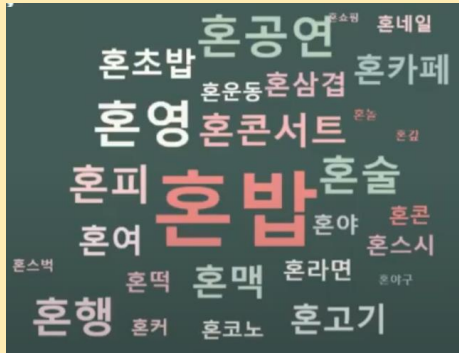
- 신한카드 데이터 전체 소비 빈도 중 인터넷 게임의 소비 빈도 비율



과거 데이터의 타당성

“코로나는 완전히 새로운 상황을 만들어 낸 것이 아니라, 기존에 존재하던 트렌드를 가속화시키고 우리의 눈 앞에 가져다 놓은 것이다. 즉 과거 데이터는 유의미하다”

- 2013년 시작한 혼밥, 2019년 혼○은 39개



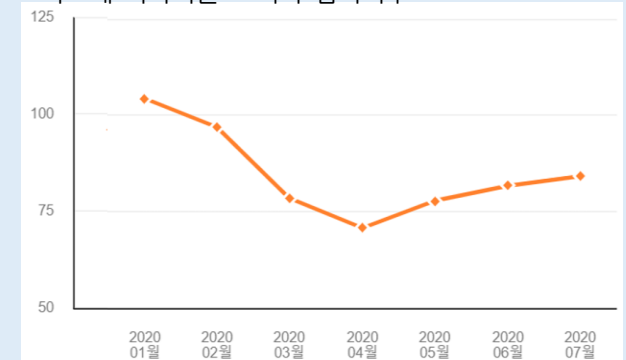
과거부터 코로나 시대까지 연속적으로 존재하는 경향은 미래에 적극적으로 반영

- 온라인 여가 생활의 증가와 혼카페, 혼여행 등 여가 생활을 혼자 즐기는 경향은 코로나 이전부터 존재하던 경향이다. 따라서 포스트 코로나 시대에 적극적으로 반영

과거에 존재하지 않다가 코로나 시대에 발생한 경향은 회복을 기대

- 현재 발생하는 문화/관광산업의 위기의 경우, 과거에 존재하지 않다가 코로나 시기에 발생한 경향이다. 따라서 포스트 코로나 시대에 회복될 것을 기대

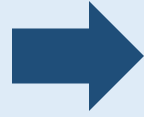
빠르게 회복되는 소비자 심리지수



e-나라지표 <http://www.index.go.kr/>

2. 데이터 데이터 선정

서울, 부산, 대구, 인천, 광주, 대전 울산, 세종,
경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주



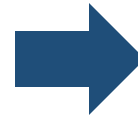
17개 시도 기준으로 문화여가생활의 경향 분석

국민여가 활동조사

- 코로나까지 이어진 문화·여가생활의 최신 경향 수집을 위해 2019년 raw 데이터 이용 (10060명이 381개의 질문에 응답)
- 17개의 거주지별로 문화여가 생활의 만족도에 가장 크게 기여하는 변수 추출
- 트리기반(tree-based) 머신러닝

신한카드 데이터

- 코로나까지 이어진 내국인의 문화 여가 생활의 최신 경향 수집을 위해 2019년 소비 데이터 KCTI_NATIVE_2019 이용
- 17개의 카드 소비 지역별 경향(나이, 성별, 여가 경향, 이용 시간대) 추출
- LDA 토픽 모델링 기법을 이용해 지역별 여가 소비 유형과 유형별 Term 분포 분석



트립 어드바이저 크롤링

거주지역별
여가생활
만족도 향상 기여
변수

소비지역별
여가 소비 유형

- 두 결과를 기반으로 17개 시도에 문화·여가 생활 인프라 구축을 위한 구체적 정책 제언



주요관광지점 입장객 통계

2. 데이터 국민여가 활동조사 변수선정

381개의 질문에 대한 응답으로부터 39개의 유의미한 독립변수 추출 / 종속변수: Q30 전반적 여가생활 만족도(binary data로 처리)

국민여가조사 항목 코드	기존 데이터	추출한 변수 설명	기존 형식	처리 방식 및 이후 형식
Q2_2_1, Q2_2_2, Q2_2_3, Q2_2_4, Q2_2_5	지난 1년간 가장 많이 참여한 1/2/3/4/5 순위 여가활동 동반자	Attr1: 여가를 혼자 즐겼는지의 여부 Attr2: 가족, 친척과 여가를 즐겼는지의 여부 Attr3: 그 외의 사람과 여가를 즐겼는지의 여부	1. 혼자서, 2. 가족(친척 포함)과 함께, 3. 친구(연인 포함)와 함께, 4. 직장 동료, 5. 동호회(종교단체 등 포함, 회원과 함께, 6. 기타	Attr1: 1번 응답은 1, 그 외 응답은 0으로 처리한 binary 데이터 Attr2: 2번 응답은 1, 그 외 응답은 0으로 처리한 binary 데이터 Attr3: 3, 4, 5, 6번 응답은 1, 그 외 응답은 0으로 처리한 binary 데이터
Q2_3_1, Q2_3_2, Q2_3_3, Q2_3_4, Q2_3_5	지난 1년간 가장 많이 참여한 1순위 여가활동 빈도	Attr4: 여가활동 1순위 참여 수치 빈도	1. 매일, 2. 일주일에 몇번 매일은 아님, 3. 한달에 2~3번, 4. 한달에 1번, 5. 몇 달에 1번	1번 응답은 5, 2번 응답은 4, 3번 응답은 3, 4번 응답은 2, 5번 응답은 1의 수치 빈도 데이터
Q2_4_1, Q2_4_2, Q2_4_3, Q2_4_4, Q2_4_5, Q2_4_1_N2, Q2_4_2_N2, Q2_4_3_N2, Q2_4_4_N2, Q2_4_5_N2	지난 1년간 가장 많이 참여한 1순위 여가활동 1회 여가 소요시간, 분	Attr5: 여가활동 1회 여가 소요시간 (분)	정수 수치 데이터	(Q2_4_1/2/3/4/5 * 60 + Q2_4_1/2/3/4/5_N2)/5의 수치 데이터
Q2_5_1, Q2_5_2, Q2_5_3, Q2_5_4, Q2_5_5	지난 1년간 가장 많이 참여한 1순위 여가활동 1회 기준 비용	Attr6: 여가활동 1회 기준 비용	정수 수치 데이터	(Q2_5_1/2/3/4/5의 합)/5의 수치 데이터
Q3	여가활동 주 목적	Attr7: 여가활동의 주 목적 개인인지의 여부	1. 건강을 위해, 2. 개인의 즐거움을 위해, 3. 마음의 안정과 휴식을 위해, 4. 대인 관계/교제를 위해, 5. 스트레스 해소를 위해, 6. 자기 계발을 위해, 7. 자기 만족을 위해, 8. 시간을 보내기 위해, 9. 가족과 시간을 함께하기 위해, 10. 기타	1, 2, 3, 5, 6, 7, 8, 10번 응답은 1, 4,9번은 0으로 처리한 binary 데이터
Q5	여가경력- 반복참여 경험	Attr8: 여가활동 반복 참여 경험 여부	1. 예, 2. 아니오	1번 응답은 1, 2번 응답은 0으로 처리한 binary 데이터
...				
Q15, Q15_M2, Q15_M3	연휴기간 중 여가활동 - 복수응답	Attr39: 연휴기간 중 바람직한 여가활동 여부	1.~88.의 여가활동 종류	세 응답이 모두 바람직하지 않은 여가활동(17, 19, 57, 58, 59, 74, 75, 62, 64, 71, 81, 82, 84번 응답)이면 0, 아니라면 1으로 처리한 binary 데이터 (한국민족문화대백과의 여가 정의 참고)

2. 데이터 국민여가 활동조사 전처리

● 변수 처리 방법

Data를 두 그룹으로 묶어 binary data로 처리

- Attr10: 1: 매우 부족했다, 4: 보통, 7: 매우 충분했다
중 1~3을 아니오(0), 4~7를 예(1)로 처리

수치형 데이터로 처리

- 1: 매일, 2: 일주일에 몇 번, 매일은 아님, 3: 한 달에 2~3번, 4: 한 달에 1번, 5: 몇 달에 1번
- Attr4: 지난 1년간 가장 많이 참여한 여가활동 빈도 1, 2, 3, 4, 5를 5, 4, 3, 2, 1로 처리해 숫자가 빈도의 상대적 크기를 나타내도록 처리

수치 데이터 그대로 사용

- Attr9: 월평균 여가활동 지출액

Categorical data로 변환

- Attr35: 1 남자 응답을 1로, 2 여자 응답을 0으로 처리

Attr1	Attr2	Attr3	Attr4
1	4	0	12
1	4	0	14
4	1	0	24
3	2	0	22
2	1	2	17
2	2	1	15
4	1	0	20
2	0	3	17
4	1	0	15
2	1	2	18
0	1	4	21
2	3	0	21
2	1	2	12
2	2	1	20
0	3	2	18
3	0	2	18
3	2	0	20
2	1	2	17

...

Attr37	Attr38	Attr39	Satisf
1	1	1	1
0	1	1	1
0	1	1	1
1	1	1	1
1	1	1	0
1	1	1	0
0	0	1	0
0	1	1	1
0	0	1	1
0	0	0	0
1	1	1	0
1	1	1	1
0	0	1	1
0	0	1	1
0	0	1	1
1	0	1	1

Attr 31, 37, 38, 39: Domain Knowledge 이용 - 한국민족문화대백과, 한국학중앙연구원 <여가>

여가는 직장 업무에서 벗어나 개인이 자유롭게 자신이 재미와 흥미를 느끼는 일에 몰입할 수 있는 활동으로, 정신적이고 육체적인 균형을 위해 반드시 필요한 활동이다. ... 그러나 다른 한편에서는 불건전한 여가 활동도 많은 편이다. **경마와 경륜, 전자 도박, 도박 등 사행성 여가 활동이 적지 않은 편이고 점차 증가하고 있다. 그 외에도 화투나 폭음 등 불건전한 여가 활동이 있다.** ... 그러나 여가 활동의 불건전한 측면도 무시할 수 없다. 한국의 여가 시간을 가장 많이 잠식하고 있는 것은 단연 **TV이고 음주이다.** 하루 평균 3시간이 넘게 TV를 시청하고 있고, 음주 횟수가 일주일에 평균 2회 정도이다. 집에서 가장 많이 하는 것은 TV 시청이고, 사람을 만나서 가장 많이 하는 활동이 주로 음주이다.

17 스포츠 경기 간접관람 TV, DMB를 통한 관람

19 온라인게임 경기관람

57 홈페이지/블로그 관리

58 인터넷 검색/1인 미디어 제작/SNS

59 게임(온라인/모바일/콘솔게임 등)

74 TV시청(IPTV 포함)

75 모바일 콘텐츠, 동영상, VOD 시청

62 겜블(경마, 경륜, 카지노, 카드놀이, 고스톱, 마작 등)/복권

64 음주

71 낮잠

82 클럽/나이트/디스코/카바레 가기

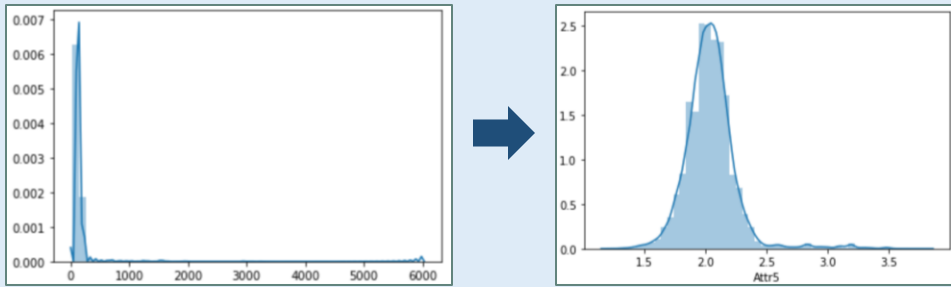
- 이에 따라 다음 항목들을 불건전한 여가로 설정, 인터넷 게임 등의 경우 일정 횟수 이상 응답 시 불건전한 여가로 분류
- 이 외에도 여가를 안 하는 경우(99)로 모든 항목이 이루어진 경우, 68 어학/기술/자격증 취득 공부, 학원 등 이용은 여가에 포함하지 않는 것으로 처리

2. 데이터 국민여가 활동조사 전처리

Skewness 조정

데이터가 한쪽으로 지나치게 치우친 경우 분석에 방해가 되므로 log와 2-power transformation을 통해 Attr5, 6, 9, 11의 skewness 조정

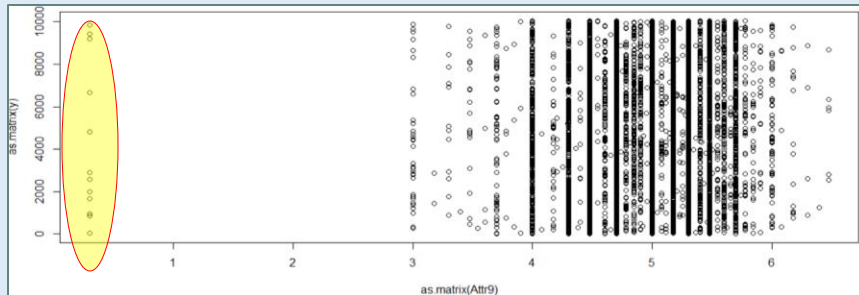
Example: Attr5의 log transformation



Outlier 확인

전체적인 경향에서 벗어나는 outlier 확인 후 중요하지 않다면 제거를 계획.
Attr9와 Attr11 모두 outlier가 유의미했기 때문에 제거하지 않고 그대로 진행

Example: Attr9의 outlier는 여가활동 지출액 0원에 해당, 여가비용에 대한 생각을 나타낼 것이라고 판단



결측치 채우기

<Attr20>

- Q20: 생활권 내 공공 문화여가시설 만족도
- Q19에서 이용여부 2아니오 응답한 사람은 Q20을 0으로 표기

<Attr28>

- Q30_1: 여가 생활 불만족 이유
- Q30에서 만족 응답한 사람은 Q30_1을 0으로 표기

<Attr29>

- 여가 생활 불만족 이유: 여가를 함께 즐길 사람이 없어서 (binary data)
- Q30에서 만족 응답한 사람&5번 이외 항목 응답한 사람

Binary data의 결측치 처리가 가능한 R 함수rpart, knn 이용

rpart: surrogate 변수를 사용하여 na를 다른 값을 대체

```
library(dlookr)
data2019 <- read.csv("data2019_rev.csv")
x<-impute_na(data2019,Attr20,Satisf,method="rpart")
write.csv(rpart_data, "data2019_rpart_log.csv")
```

knn: 유클리드 거리로 선정한 주변의 값들을 이용해 na를 대체

```
x<-impute_na(data2019,Attr20,Satisf,method="knn")
write.csv(rpart_data, "data2019_knn_log.csv")
```

2. 데이터 국민여가 활동조사 전처리

Standard Scaling

결측치 처리로 Attr20, 28, 29만 다른데 세 열이 모두 binary data이므로 결국 RPart와 KNN의 Scaling 결과는 동일하다.

```
# 잘 drop 됐는지 확인  
knnscale.head()
```

	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr9	Attr11
0	1	4	0	144	3.490801	5.810233	5.698970	5.301030
1	1	4	0	196	2.959995	4.806180	5.301030	5.301030
2	4	1	0	576	1.903090	2.301030	5.477121	5.903090
3	3	2	0	484	2.056905	4.580697	5.477121	5.477121
4	2	1	2	289	1.819544	3.869232	5.301030	5.477121

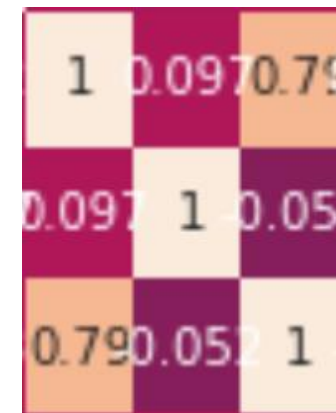
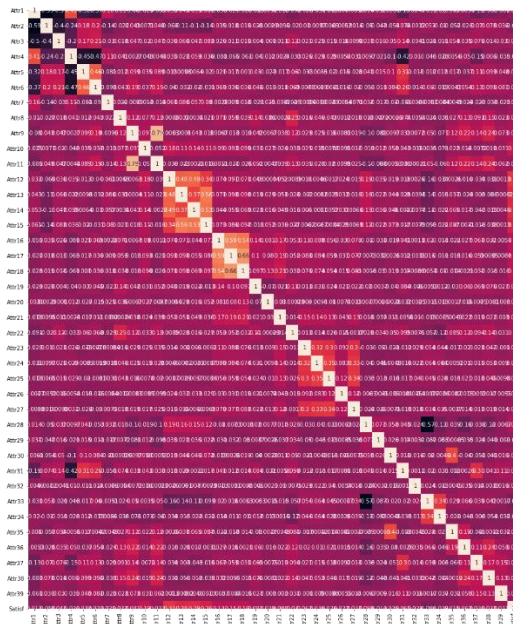
```
knnscale = StandardScaler().fit_transform(knnscale)  
knnscale = pd.DataFrame(knnscale)
```

knnscale

	0	1	2	3	4	5	6	7
0	-1.000099	2.232022	-1.251384	-1.919573	6.783539	2.083603	1.478228	0.389672
1	-1.000099	2.232022	-1.251384	-1.471681	4.297095	1.076820	0.681496	0.389672
2	1.378200	-0.282346	-1.251384	1.801374	-0.653742	-1.435140	1.034056	1.719509
3	0.585434	0.555777	-1.251384	1.008950	0.066769	0.850724	1.034056	0.778625
4	-0.207333	-0.282346	0.540889	-0.670644	-1.045095	0.137325	0.681496	0.778625
...	...	...	...	...	...	...	...	...
10055	-1.000099	-0.282346	1.437025	-0.670644	0.542025	0.549415	1.034056	1.270372
10056	-1.792866	-1.120468	3.229299	-0.369178	-0.043223	-0.209100	-0.834049	1.268647
10057	-1.000099	-1.120468	2.333162	-0.670644	0.542025	0.549333	0.875524	1.054591
10058	-0.207333	-1.120468	1.437025	0.285433	0.171118	0.349294	0.681496	0.778625
10059	0.585434	-0.282346	-0.355248	-1.704240	0.855623	0.524459	0.078791	0.778625

10060 rows × 8 columns

Correlation이 높은 변수 제거



Attr9 평균 여가활동 지출액과 Attr11 적절하다고 생각하는 여가비용(월평균)의 correlation이 0.79로 높게 나왔다. 실제로 소비한 정도가 더욱 중요하다고 보아 Attr11를 제거하였다.

3. 분석 및 결과 국민여가 활동조사 분석 LightGBM 머신러닝

모델 학습

복수 하이퍼 파라미터 튜닝 Grid Search

- 클래스 객체에 fit 메서드를 호출하면 grid search를 사용해 자동으로 복수 개의 내부 모델을 생성하고 이를 모두 실행시켜 최적의 파라미터를 찾아준다
- 실행 전에 범주형 자료를 .astype('category')를 통해 타입을 바꿔준다

모델 Test: k겹 교차검증(Cross-Validation)

	A	B	C	D	E
Cross Validation Iteration 1	Test	Train	Train	Train	Train
Cross Validation Iteration 2	Train	Test	Train	Train	Train
Cross Validation Iteration 3	Train	Train	Test	Train	Train
Cross Validation Iteration 4	Train	Train	Train	Test	Train
Cross Validation Iteration 5	Train	Train	Train	Train	Test

- 데이터셋의 크기가 작은 경우 테스트셋에 대한 성능 평가의 신뢰성이 떨어지게 된다
- 이를 방지하기 위해 모든 데이터가 최소 한 번은 테스트셋으로 쓰이도록 하는 것이 k겹 교차검증이다.

교차 검증 StratifiedKfold

- 데이터 수가 적은 경우 학습 데이터셋과 검증 데이터셋 둘 다 부족하다는 문제가 있다. 이를 해결하기 위해 사용하는 검증 방법이다

```
d_train = lgb.Dataset(x_train, label = y_train)
lgbm = LGBMClassifier(n_estimators=200, random_state=0)
```

```
params = {'max_depth': [10, 15, 20, 25, 30],
          'min_child_samples': [20, 40, 60, 80, 100],
          'subsample': [0.8, 1]}
```

```
grid = GridSearchCV(lgbm, param_grid=params)
grid.fit(x_train, y_train, early_stopping_rounds=100, eval_metric='F1',
        eval_set=[(x_train, y_train), (x_test, y_test)])
```

```
[98] valid_0's binary_logloss: 0.0211369    valid_1's binary_logloss: 0.0211369
[99] valid_0's binary_logloss: 0.0206676    valid_1's binary_logloss: 0.0206676
[100] valid_0's binary_logloss: 0.0202191    valid_1's binary_logloss: 0.0202191
[101] valid_0's binary_logloss: 0.0197172    valid_1's binary_logloss: 0.0197172
[102] valid_0's binary_logloss: 0.019127     valid_1's binary_logloss: 0.019127
[103] valid_0's binary_logloss: 0.0185037    valid_1's binary_logloss: 0.0185037
[104] valid_0's binary_logloss: 0.0179831    valid_1's binary_logloss: 0.0179831
[105] valid_0's binary_logloss: 0.0174758    valid_1's binary_logloss: 0.0174758
```

```
print("최적 파라미터: ", grid.best_params_)
lgbm_f1_score = f1_score(y_test, grid.predict(x_test), average='micro')
print("F1 score : {0:.4f}".format(lgbm_f1_score))
```

```
최적 파라미터: {'max_depth': 15, 'min_child_samples': 20, 'subsample': 0.8}
F1 score : 0.9022
```

```
lgbm = LGBMClassifier(n_estimators=1000, num_leaves=50, subsample=0.8,
                    min_child_samples=20, max_depth=10)
```

```
evals = [(x_test, y_test)]
lgbm.fit(x_train, y_train, early_stopping_rounds=100, eval_metric='F1',
        eval_set=evals, verbose=True)
```

```
score = f1_score(y_test, lgbm.predict(x_test), average='micro')
print("F1 score : {0:.4f}".format(score))
```

```
[9] valid_0's binary_logloss: 0.312369
[9] valid_0's binary_logloss: 0.310491
[10] valid_0's binary_logloss: 0.306133
[11] valid_0's binary_logloss: 0.304551
[12] valid_0's binary_logloss: 0.301747
[13] valid_0's binary_logloss: 0.30291
```

```
re_stf = RepeatedStratifiedKFold(n_splits = 10, n_repeats = 3, random_state = 1)
scores = cross_val_score(lgbm, X, y, scoring = "f1_micro", cv = re_stf)
scores.mean()
```

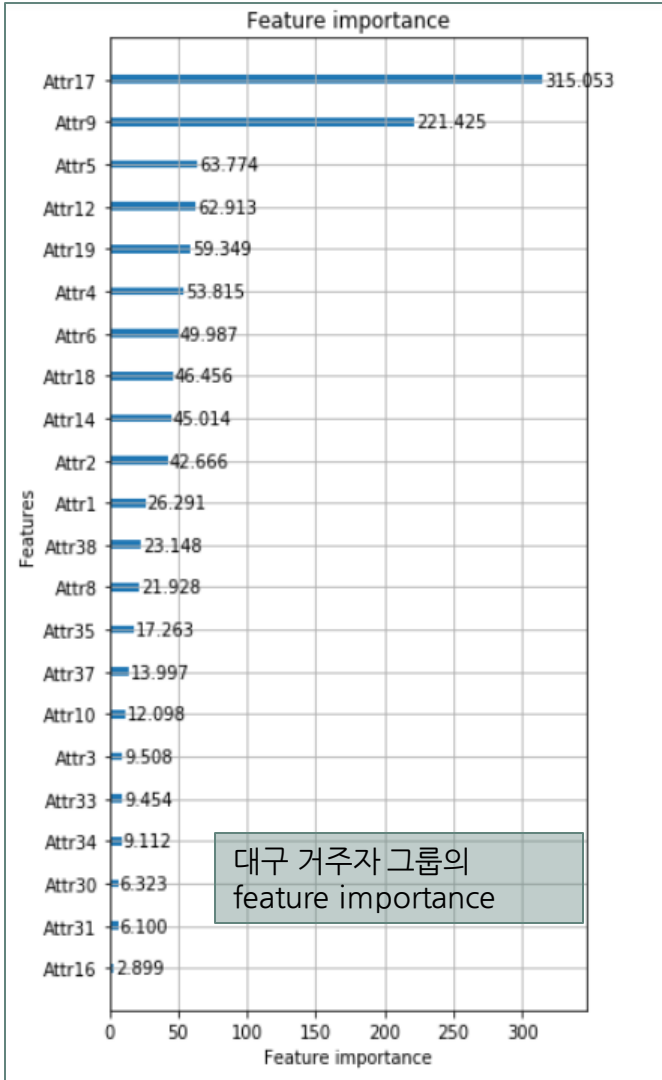
```
0.8754621848739498
```

```
scores = cross_val_score(lgbm, X, y, scoring = "roc_auc", cv = re_stf)
scores.mean()
```

```
0.8246912128643575
```

➡ 두 score 모두 높은 값을 보이므로, 머신러닝에 쓰인 변수들이 유의미하다고 할 수 있다

3. 분석 및 결과 국민여가 활동조사 분석 결과



● 결과

- 전체적으로 Attr12: 평일 여가시간 충분도(불충분), Attr13: 평일 여가시간 자유도, Attr5 여가활동 1회 여가 소요시간(적을수록) 등의 변수가 여가 활동 불만족의 주요 변인으로 나타났다.
- 즉, 코로나 이전 국민들은 여가를 즐길 시간과 여유가 없었다.
- 대체로 Attr9 월평균 여가활동 지출액 변수가 높을수록 여가 활동에 만족했다. 연령대를 나누어 다시 주요변인을 구해보니 10, 20대를 제외한 나머지 연령대는 모두 여가 활동 지출액이 큰 것이 여가 활동 만족에 주요 변인이 되었다.

그룹	정보성 높은 주요 변인 (만족도와의 관계성)	기타 주요 변인
10,20대	Attr6 1순위 여가활동 1회 기준 비용(낮을수록 만족도 높음), Attr38 연휴기간 중 바람직한 여가활동	Attr12, Attr5, Attr15 지난 1년간 여가시간 자유도(휴일)
30대	Attr6 (높을수록 만족도 높음), Attr17 생활권 내 공공문화 및 여가시설 프로그램 인지도	Attr14 휴일 여가시간 충분도, Attr12, Attr13, Attr5
40대	Attr6 (높을수록 만족도 높음), Attr10 여가비 충분한가 (만족하는 사람 중 여가비 충분하다고 답한 사람이 더 많음), , Attr4 여가활동 빈도	Attr12, Attr13
50+대	Attr18 생활권 내 공공문화 및 여가시설 프로그램 충분도, Attr6 (높을수록 만족도 높음), Attr9 월평균 여가활동 지출액(높을수록 만족도 높음)	Attr12, Attr5
서울	Attr6 (높을수록 만족도 높음), Attr36 관광/문화/여가 생활의 다양성, Attr1 여가 혼자서(혼자서 즐길수록 만족도 높음)	Attr12, Attr13, Attr5, Attr4 지난 1년간 가장 많이 참여한 여가활동 빈도
대구	Attr17, Attr9, Attr19 생활권 내 공공문화여가시설 이용여부(이용한 사람일수록 만족도 높음)	Attr5, Attr12

... *마찬가지로 연령대, 17개 시도 지역별 분석 결과 추출

3. 분석 및 결과 신한카드 데이터 분석 ① Topic Modeling LDA

LDA

주제를 표현하는 방법

- 단어 사용, Keyword Frequency, 단어들의 분포 사용
- 구체적으로 두 개의 서로 다른 확률 분포를 통해서 특정 문서에서 특정 단어가 사용되었다라고 가정

분포1: 각 문서에 대한 주제의 확률 분포

- 하나의 문서는 여러 개의 topic을 가지고 있고, 그 topic들에 대한 특정한 확률 분포 (Dirichlet 분포)가 있다고 가정

분포2: 각 Topic에 대한 단어의 확률 분포

- 각 topic에 대한 단어들이 갖고 있는 특정한 확률 분포 (Dirichlet 분포)가 있다고 가정

클러스터링 기법과 차이점

- 두 개의 확률 분포를 실제 문서/단어 정보를 가지고 역으로 추정
- Clustering 기법들과 달리 LDA는 주제에 대한 단어들의 확률 정보까지 제안

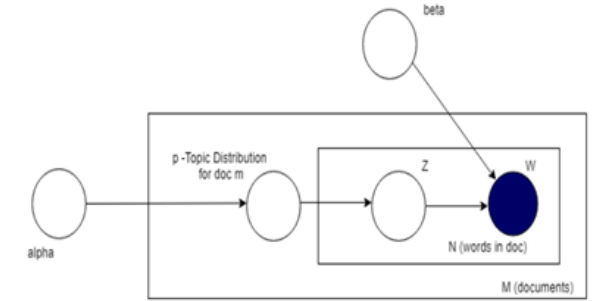
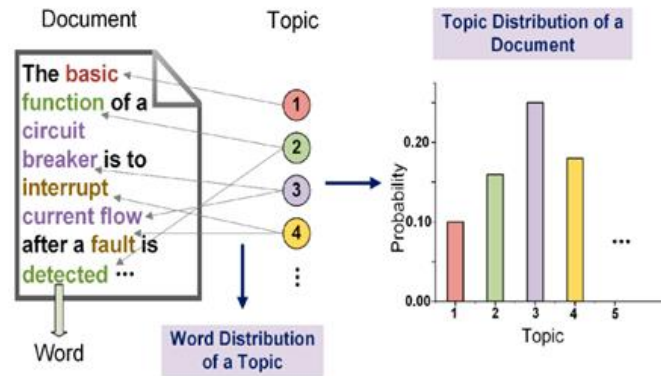


Plate notation of LDA model

- 실제로 사용된 단어들과 문서 정보를 이용해서 역으로 확률 분포 추정

- Bayesian 통계를 사용한 확률적 생성 모형

➡ 신한카드 데이터에 적용함으로써 카드 사용 유형화 시도

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

posterior distribution
(이 분포를 찾아야 한다.
즉, 이 값을 최대화하는
값들을 찾아야 한다)

Joint Distribution

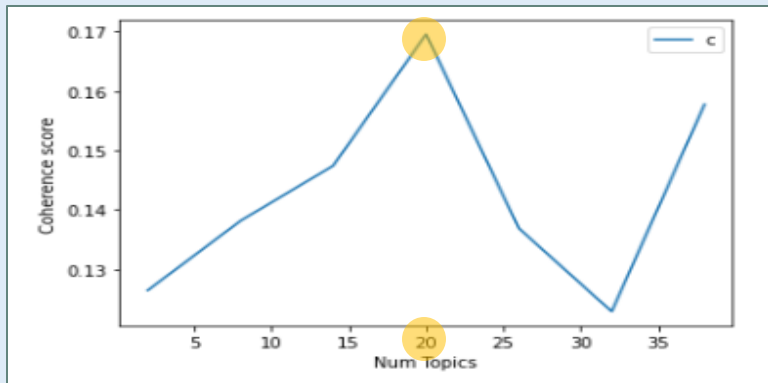
Evidence

3. 분석 및 결과 신한카드 데이터 분석 ① Topic Modeling LDA

LDA의 핵심: 최적의 토픽 개수 정하기

- 즉, 지역별 문화 여가 생활 유형을 가장 잘 드러내는 결과를 가져다 주는 ‘한 유형 내의 소비 특성 개수’ 정하기
- Coherence와 Perplexity, 두 가지 기준 이용

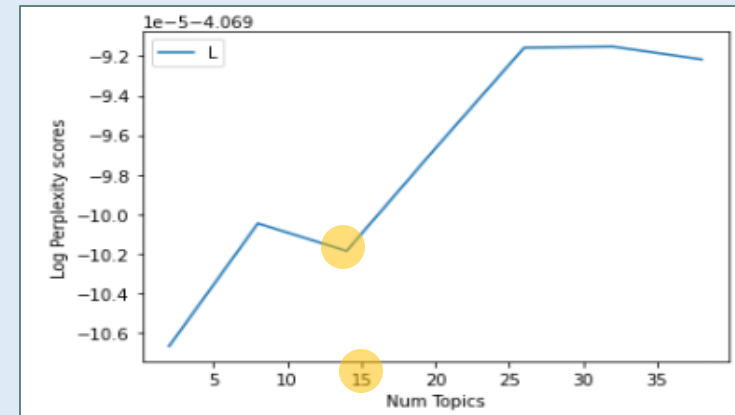
Coherence



- 각 주제와 관련이 높은 top n 개의 단어들 간의 유사도를 계산
- PMI (Pointwise Mutual Information) 이용
- 최고 Coherence 값의 토픽 수로 결정

Coherence 계산 결과 최적의 토픽 개수
(=소비 유형 개수): 20개

Perplexity



- 최저 Perplexity 값의 토픽 수로 결정

$$Perplexity(w) = \exp \left[-\frac{\log \{p(w)\}}{\sum_{d=1}^D \sum_{j=1}^V n^{jd}} \right]$$

$p(w)$: corpus의 발생확률

Perplexity 계산 결과 최적의 토픽 개수
(=소비 유형 개수): 15개

3. 분석 및 결과 신한카드 데이터 분석 결과 ① Topic Modeling LDA

● LDA 토픽 모델링 결과 (20개, Coherence 값 참조)

토픽	주요 변수	분석결과	
		소비 지역	해당 지역 소비 유형
1	경기_사용장소, 취미오락, 외식, WHITE, F, 활동, M, RED, 휴식, 패션쇼핑, 20, 40, 50	경기	낮~초저녁 시간대 외식과 의류 쇼핑.
2	전남_사용장소, 광주_거주지, 종합쇼핑, 취미오락, M, WHITE/RED, 활동, 휴식, 30, 50, 20	전남	낮~저녁 시간대 종합 쇼핑. 광주 지역 거주자.
3	서울_사용장소, 경기_사용장소, 전남_사용장소, 경기_거주지, 인천_거주지, 여행, 숙박, M/F, WHITE, 활동, RED, 휴식, 30, 20, 50	경기, 서울, 전남	낮~초저녁 시간대 숙박 시설 결제. 경인 지역 거주자.
4	경기_사용장소, 전남_사용장소, 광주_사용장소, 취미오락, 취침, M, WHITE/RED, 목욕, 유흥, 휴식, 20, 50, 30	경기, 광주, 전남	저녁~새벽 시간대 목욕 시설과 유흥.
5	경기_사용장소, 세종_거주지, 인천_거주지, 취미오락, WHITE, 활동, M, 외식, RED, F, 휴식	경기	낮~밤 시간대 외식. 인천/세종 지역 거주자.
6	경기_사용장소, 경기_거주지, 취미오락, 활동, WHITE, F/M, RED, 독서, 휴식, 애완동물돌보기, 20, 30, 50	경기	낮~초저녁 시간대 독서, 반려동물 물품 구매. 경기 지역 거주자.
7	경기_사용장소, 서울_사용장소, 경기_거주지, 스포츠활동, M, WHITE, 레저 스포츠, 활동, RED, 휴식, F, 골프, 헬스, 취침	경기, 서울	낮~새벽 시간대 레저 스포츠 활동. 경기 지역 거주자.
8	전북_사용장소, 서울_사용장소, 경기_사용장소, 전북_거주지, 문화예술활동, 활동, WHITE, M, 공연관람, RED, 휴식, 사진촬영	서울, 전북	밤~새벽 시간대 공연관람과 사진 촬영 등 문화예술 활동. 전북 지역 거주자.
9	대구_사용장소, 대구_거주지, 인터넷게임, 취미오락, M, WHITE, 휴식, 활동, F, 30, 50	대구	밤~새벽 시간대 인터넷 게임. 대구 거주자.
10	서울_사용장소, 경기_사용장소, 미용, 취미오락, 활동, WHITE, 휴식	경기, 서울	평일 저녁 시간대 미용 관련 구매.

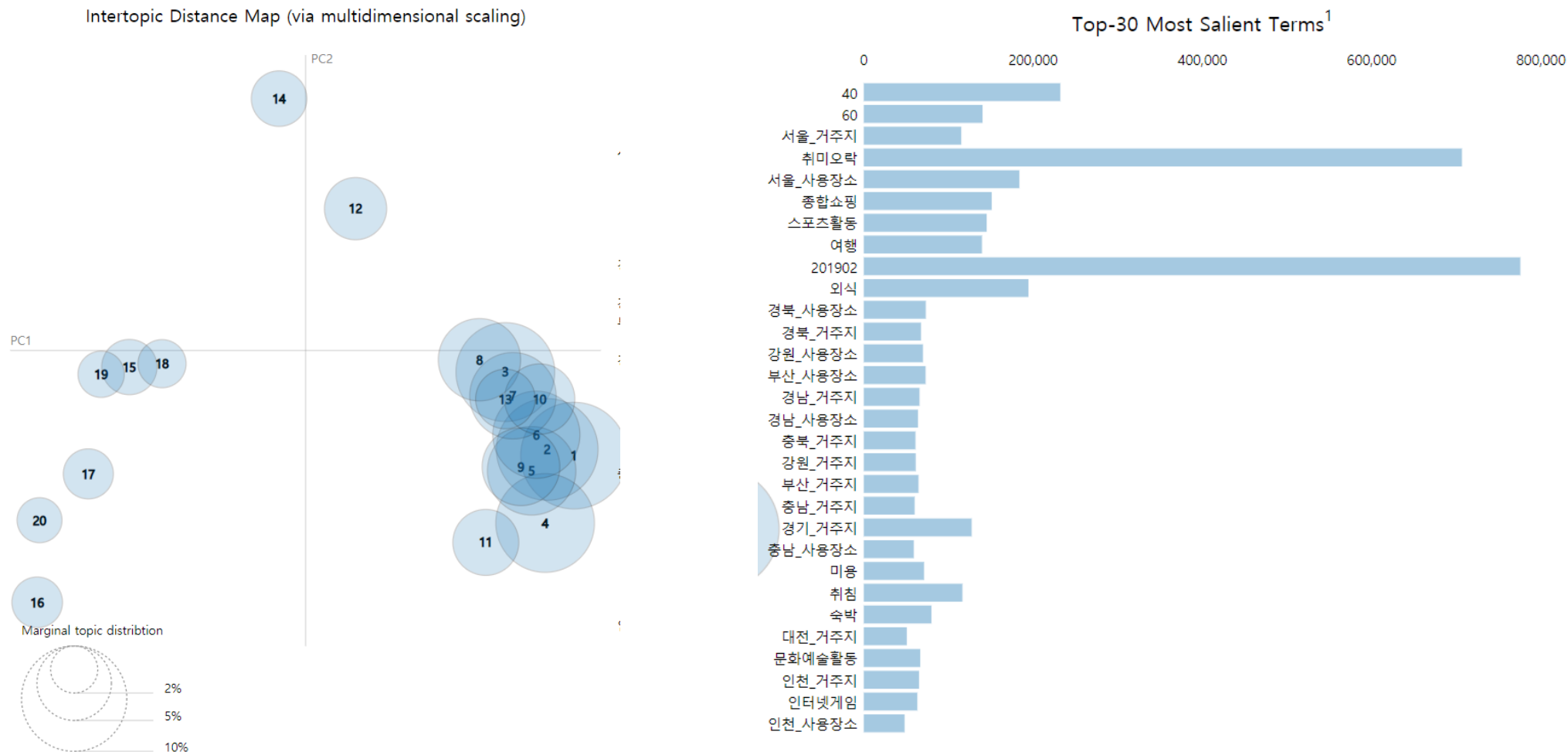
3. 분석 및 결과 신한카드 데이터 분석 결과 ① Topic Modeling LDA

● LDA 토픽 모델링 결과 (20개, Coherence 값 참조)

토픽	주요 변수	분석결과	
		소비 지역	해당 지역 소비 유형
11	60, M, 취미오락, 활동, 평일, 휴일, 외식, 휴식	없음	낮~저녁 시간대에 취미오락과 외식을 즐기는 60대 남성.
12	서울_사용장소, 인천_거주지, 제주_거주지, 세종_거주지, F, 활동, 주말, 평일, 취미오락, 30, 50, 20	서울, 세종, 인천, 제주	취미오락을 주로 낮~초저녁 시간대에 취미오락.
13	경기_사용장소, 전남_사용장소, 경기_거주지, 인천_거주지, 스포츠활동, 스포츠용품구, M, 평일, 주말, 30, 50, 20, 스키	경기, 전남	스포츠활동, 스포츠용품 구매를 즐기는 남성.
14	서울_거주지, 서울_사용장소, 제주_사용장소, F, 활동, 관광쇼핑, 미술공예참여, 여행사, 주말, 평일, 30	서울, 제주	낮~초저녁 시간대에 관광쇼핑과 미술공예 활동을 즐기는 여성.
15	부산_사용장소, 울산_사용장소, 세종_사용장소, 부산_거주지, 울산_거주지, 활동, F	부산, 울산	낮~초저녁 시간대 소비활동을 즐기는 여성.
16	경남_거주지, 경남_사용장소, 인천_사용장소, 음악감상	경남, 인천	음악감상
17	충남_거주지, 대전_거주지, 충남_사용장소, 자전거 활동, F, M, 휴일, 평일	대전, 충남	자전거 활동을 즐기는 여성과 남성.
18	강원_사용장소, 강원_거주지, 교육훈련, 활동, F, 평일, 30	강원	학원비 및 교육비를 결제하는 강원도 지역 30대 여성
19	충북_거주지 충북_사용장소, 대전_사용장소, 종교활동, 활동, F, 휴일, 평일	대전, 충북	낮~초저녁 시간대에 종교 활동을 즐기는 여성.
20	경북_사용장소, 경북_거주지, 활동, F, 휴일, 평일, 50	경북	휴일과 평일 초저녁 시간에 카드 소비를 한 50대 여성

3. 분석 및 결과 신한카드 데이터 분석 결과 ① Topic Modeling LDA

LDA 토픽 모델링 토픽 및 단어분포 시각화 (20개, Coherence 값 참조)



3. 분석 및 결과 신한카드 데이터 분석 결과 ① Topic Modeling LDA

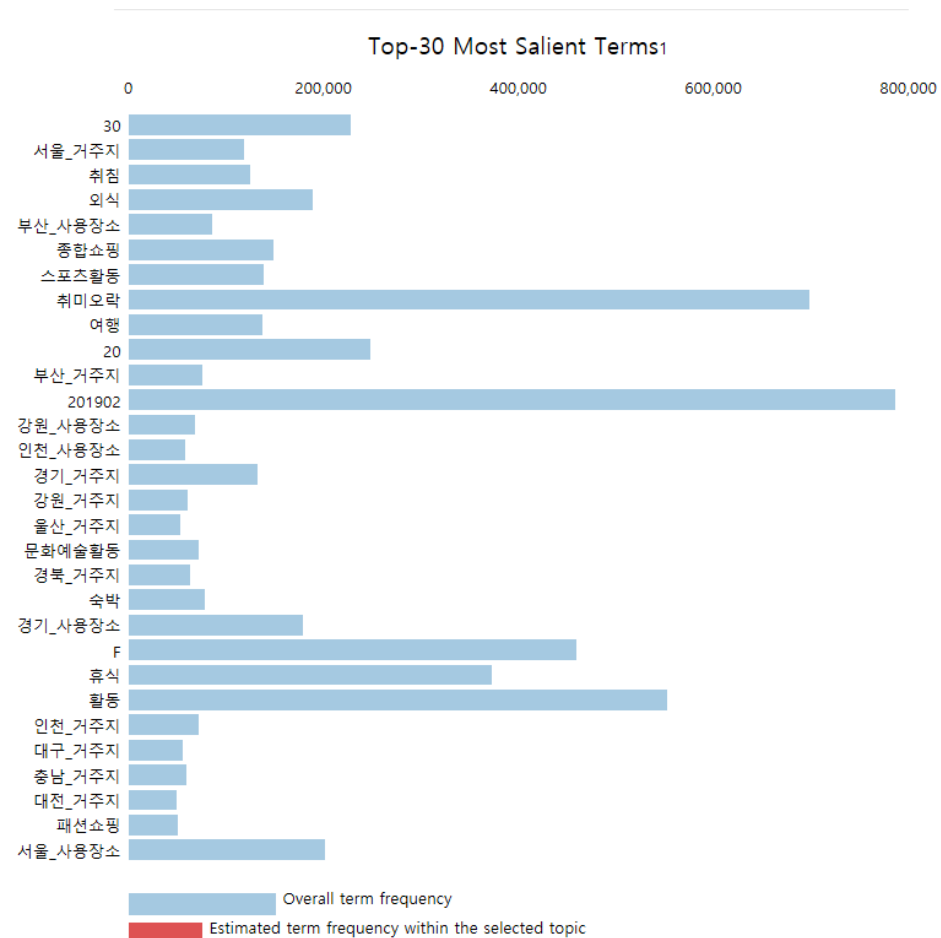
LDA 토픽 모델링 결과 (10개, Perplexity 값 참조)

*15개 중 term 들의 확률 분포값이 지나치게 작거나 term 개수가 적은 Topic 5개를 제외함

토픽	주요 변수	분석결과	
		소비 지역	해당 지역 소비 유형
1	여행, 숙박, M, 평일, 활동, 휴일, 40, 경기_사용장소, 서울_사용장소, 경기_거주지, 인천_거주지	경기, 서울	여행, 숙박에 돈을 많이 쓰는 40대 남성, 경기/인천 거주
2	스포츠활동, M, 활동, 평일, 휴일, 레저스포츠, 휴식, 스포츠용품구매, 경기_사용장소, 서울_사용장소	경기, 서울	퇴근 후 또는 평일에 스포츠 활동 (레저스포츠, 스포츠용품구매, 골프)를 즐기는 남성
3	취미오락, 경기_사용장소, 경기_거주지, 활동, F, 미용, M, 40, 50, 60, 교육훈련, 목욕	경기	취미오락, 교육훈련, 목욕, 미용을 즐기는 40~60대
4	20, 취미오락, M, 인터넷게임, 평일, 휴식, 충북_거주지, 휴일, 활동, 충북_사용장소, 서울_사용장소, 미용	서울, 충북	20대 남성, 취미오락과 인터넷 게임을 휴식시간에 즐김, 미용에도 관심이 있음
5	취미오락, 휴식, M, 충남_거주지, 평일, 충남_사용장소, 30, 목욕, 휴일, 유흥	충남	30대 취미오락, 목욕, 유흥을 즐기는 30대 남성, 미용에도 관심이 있음
6	문화예술활동, 대구_거주지, 활동, 공연관람, 평일, M, 휴일, F, 휴식, 서울_사용장소	서울	대구에 주로 거주하며, 문화예술활동과 공연관람, 사진촬영을 즐기는 20대
7	취침, 취미오락, M, 평일, 휴일, 외식, 20, 서울_사용장소, F, 유흥, 40, 50	서울	취미오락, 외식, 유흥을 취침시간에 즐기는 20, 40~50대
8	서울_거주지, F, 서울_사용장소, 애완동물돌보기, 활동, 휴일, 음악감상, 취미오락	서울	서울에 주로 거주하며 휴일 활동시간에 애완동물을 위한 소비를 즐긴다, 음악감상과 취미오락 즐기는 여성
9	경북_거주자, 패션쇼핑, 취미오락, F, 활동, 대구_사용장소, 평일, 대구_사용장소, 평일, 경북_사용장소, 휴일	대구, 경북	경북에 주로 거주하며 패션쇼핑, 취미오락에 관심있는 여성
10	인천_거주지, 대전_거주지, 30, 활동, F, 대전_사용장소, 독서, 세종_거주지, 취미오락, 휴일, 평일, M	대전	인천, 대전에서 주로 거주하며 활동시간에 독서 및 취미오락을 즐기는 30대

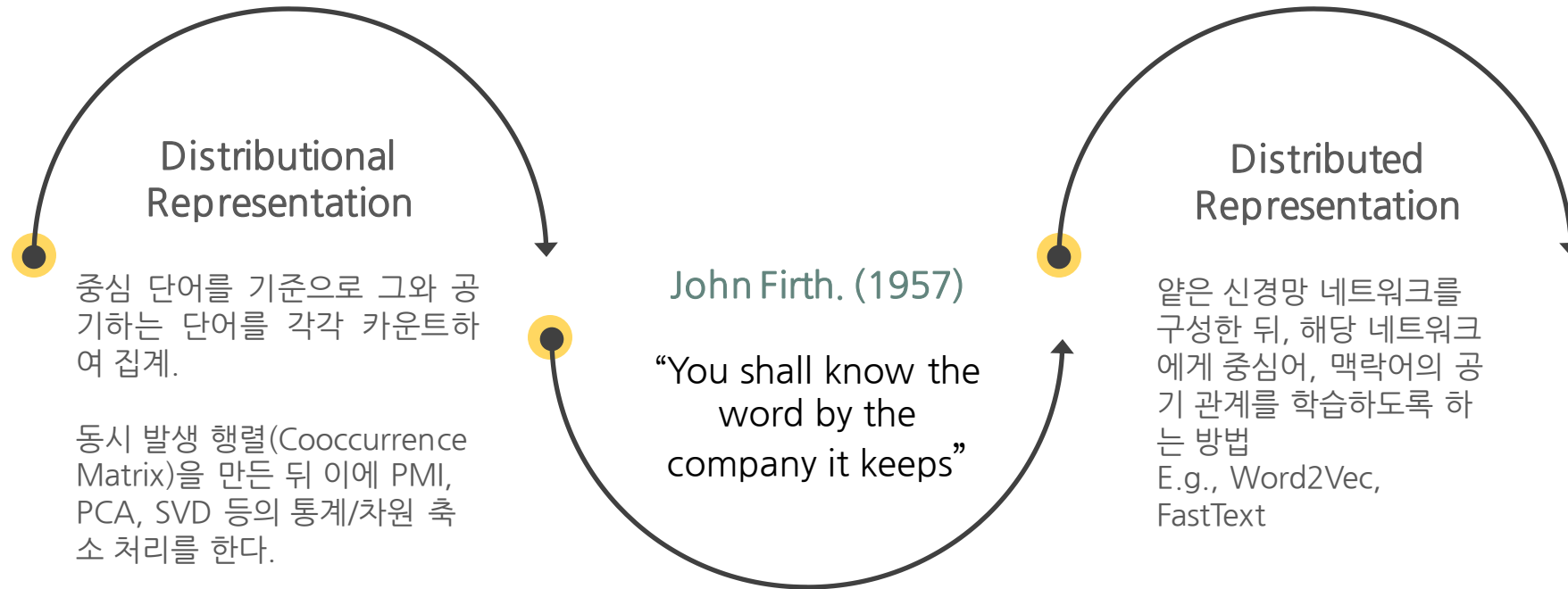
3. 분석 및 결과 신한카드 데이터 분석 결과 ① Topic Modeling LDA

LDA 토픽 모델링 토픽 및 단어분포 시각화 (10개, Perplexity 값 참조)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))]] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

3. 분석 및 결과 신한카드 데이터 분석 ② Vector Space Model



동시 등장 관계를 통계적으로 계산하거나, 뉴럴 네트워크에 학습시킴으로써 얻은 벡터들이

단어들 간의 중요한 관계성, 의미 정보를 담고 있을 것이라는 접근 방법.

본 분석에서는 신한 카드 데이터 변수 중 사용장소-세부 업종(변수명: GB2, GB3)의 관계성을 모델링하기 위해 Vector Space Model을 이용. 카운트 기반해 동시 등장 행렬을 구축한 뒤 (P)PMI/LSA의 기법을 이용. 변수 인스턴스 간의 거리는 cosine metric을 이용해서 계량.

3. 분석 및 결과 신한카드 데이터 분석 ② Vector Space Model

1. 공기행렬 구축 (Co-Occurrence Matrix)

- 1개의 Row 내에 어떤 변수들이 공기하였는 지 카운트한 뒤 이를 동시 발생 행렬 (Co-occurrence Matrix) 형태로 구축

v1	v2	v3	gb3	gb2	sex_ccd	cln_age_r	ta_ym	daw_ccd	apv_ts_dl
세종	세종	.	취미오락	종합쇼핑	M	40	201901	WHITE	활동
대전	세종	.	취미오락	외식	M	20	201901	WHITE	휴식
서울	세종	.	취미오락	외식	F	40	201901	RED	휴식
경남	세종	.	취미오락	종합쇼핑	M	50	201901	RED	휴식
충북	세종	.	취미오락	종합쇼핑	F	40	201901	WHITE	활동
충북	세종	.	취미오락	목욕	M	30	201901	RED	활동
부산	세종	.	취미오락	외식	M	50	201901	WHITE	휴식
대전	세종	.	스포츠활동	레저스포츠	M	30	201901	WHITE	활동
세종	세종	.	취미오락	종합쇼핑	M	20	201901	WHITE	휴식

전체의 관계성이 포착될 수 있도록 공기의 기준이 되는 Window 크기를 10으로 설정

2. PMI 또는 SVD 진행

- 기존의 동시발생행렬은 0인 원소가 지나치게 많은 희소행렬 (Sparse Matrix)이므로 이를 통계적으로 개선하기 위해
 - ① 점별 상호정보량(PMI)를 계산
 - ② 차원축소 (SVD)를 진행함.

$$\text{pmi}(X, i, j) = \log \left(\frac{X_{ij}}{\text{expected}(X, i, j)} \right) = \log \left(\frac{P(X_{ij})}{P(X_{i*}) \cdot P(X_{*j})} \right)$$

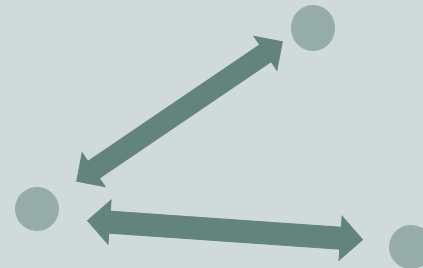
$$A_{m \times n} = T_{m \times m} S_{m \times m} D_{n \times m}^T$$

$$\begin{pmatrix} \vdots & \vdots & \vdots & \vdots \end{pmatrix}_{A_{3 \times 4}} = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots \end{pmatrix}_{T_{3 \times 3}} \begin{pmatrix} \cdot & \cdot & \cdot \end{pmatrix}_{S_{3 \times 3}} \begin{pmatrix} \vdots & \vdots & \vdots \end{pmatrix}_{D_{4 \times 3}^T}^T$$

3. 소비 지역, 거주지별 코사인 거리를 계량

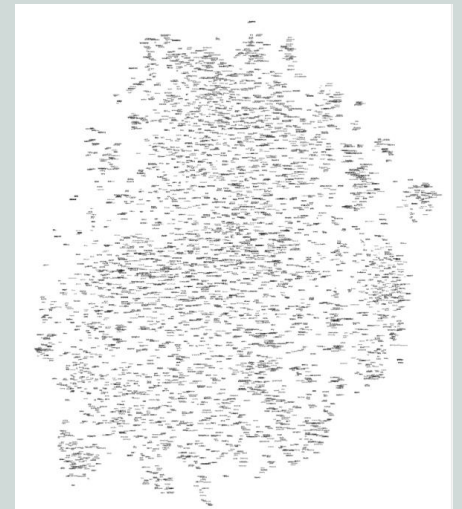
- 벡터 공간상 가까이 있는 이웃 (neighbor)들의 목록을 뽑아냄

$$\text{cosine}(u, v) = 1 - \frac{\sum_{i=1}^n u_i \times v_i}{\|u\|_2 \times \|v\|_2}$$



4. 시각화

- Tsne로 2차원으로 차원을 축소한 뒤 시각화 진행.



3. 분석 및 결과 신한카드 데이터 분석 결과 ② Vector Space Model

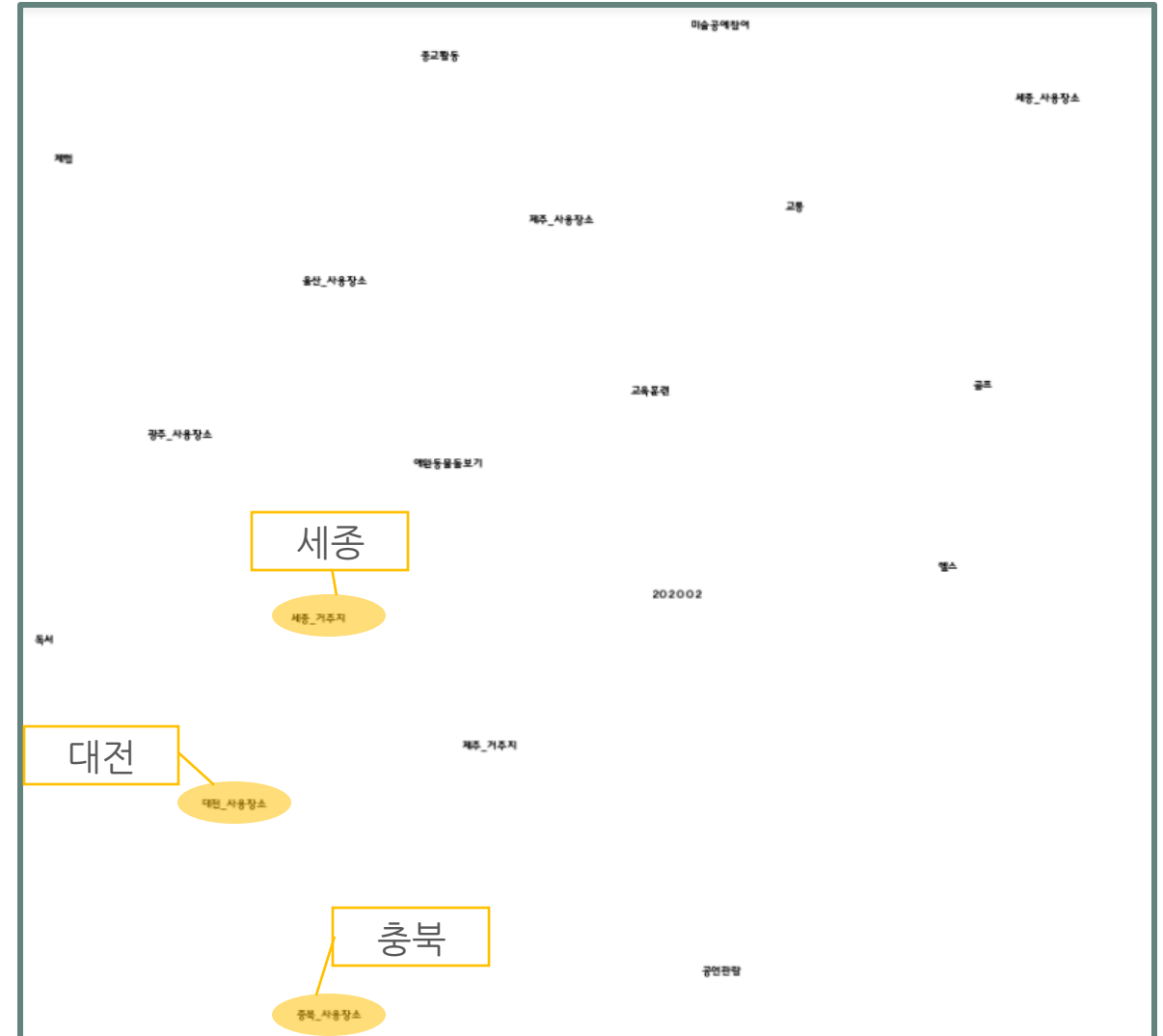
서울 거주지 SVD: 코사인 거리가 적게 나온 15개 neighbor

	0
서울_거주지	0
경기_거주지	3.02E-06
강원_사용장소	0.000148
인천_거주지	0.00017
40	0.000171
부산_사용장소	0.00023
30	0.000247
대전_사용장소	0.00025
서울_사용장소	0.000267
부산_거주지	0.000295
50	0.000316
경남_사용장소	0.000341
전북_사용장소	0.000343
경남_거주지	0.000348

서울 사용장소 PMI: 코사인 거리가 적게 나온 15개 neighbor

	0
서울_사용장소	0
WHITE	0.290675
활동	0.371438
서울_거주지	0.378897
제주_사용장소	0.410141
F	0.418133
201901	0.430068
경기_거주지	0.437421
201902	0.485764
문화예술활동	0.497967
50	0.505475
20	0.510928
대전_사용장소	0.511774
40	0.512347

(우측) 2020년 신한카드 데이터 벡터 시각화 결과. 세종_거주지, 대전_사용장소, 충북_사용장소 등이 가까이 나온 것이 인상적이다.

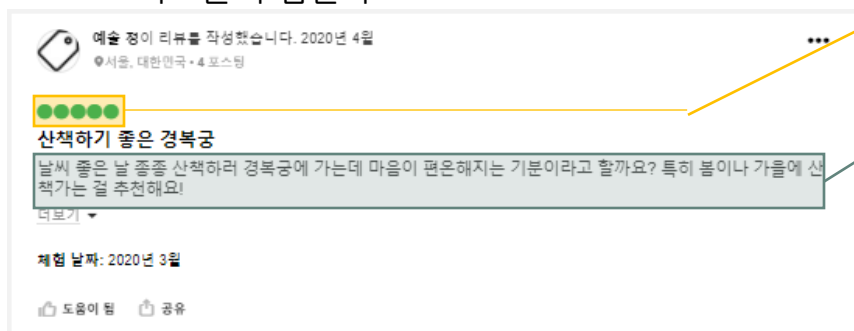


3. 분석 및 결과

● 트립어드바이저 크롤링

여행에 특화된 어플리케이션의 관광지 별점 및 줄글 후기 크롤링

- 관광지 선정 기준: 주요관광지점 입장객통계, 국민여가 활동 조사 분석 결과의 17개 시도 특징
- 크롤링 대상은 별점 후기와 줄글 후기로 선정하였으며 동적 웹페이지의 HTML, CSS 소스코드를 분석하여 수집 대상 객체를 정의한다
- 셀레니움 모듈 설치, 크롬 웹드라이버 설치, 크롤링 데이터를 모을 XML 및 TXT 파일 위치 설정 후 파이썬 코드로 자동적으로 자료를 수집한다



● 파이썬 크롤링 코드 (URL 파싱, 수집 대상 설정 등)

```
for j in range(0,page_reviews,1):

    #print(i)
    reviews = driver.find_elements_by_css_selector(".Dq9MAugU.LnVzGwUB") # May change for
#nt page design
    rating_code = reviews[j].find_element_by_css_selector(".ui_bubble_rating") # May chan
#different page design
    cls_attr = rating_code.get_attribute("class")
    cls_attr = str(cls_attr).split("ui_bubble_rating bubble-")
    url_parts = url.split('-Reviews')
    url_parts[0] += "-Reviews-or"
    url_front = url_parts[0]
    url_end = url_parts[1]
    return url_front, url_end
```

● 크롤링 결과 17개 시도 주요 관광지에 대한 후기 수집

	A	B	C	D	E	F	G	H	I	J	K	L
1	지역	loc	score	review								
2	서울	경복궁	50	날씨 좋은 날 종종 산책하러 경복궁에 가는데 마음이 편해지는 기분이라고 할까요? 특히 봄이나								
3	서울	경복궁	50	Gooooooooo 다 좋습니다 다음에 또 오고 싶네요 근처 관광지도많고 먹을거리도많네요 서울올때마다								
4	서울	경복궁	50	요새 더더욱 코로나로 인해 사람 방문이 적음. 두자녀 동반시 성인 입장무료. 지금 시기에 나들이 하								
5	서울	경복궁	50	조선 개국 4년째인 1395년에 처음으로 세운 으뜸 궁궐이다. '하늘이 내린 큰 복'이라는 뜻으로 경복								
6	서울	경복궁	50	한국 건축물의 아름다움을 느낄 수 있는 곳이고, 한복을 입었을 때 입장료는 무료. 주변에 한복 대여								
7	서울	경복궁	50	조선의 성곽 안에는 박물관도 있고 경회소 등 볼것이 많다. 외국인들 및 연인들도 주변에서 한복 입								
8	서울	경복궁	40	대부분의 사람들이 여행을 간다면 그 나라의 유서 깊은 궁전을 찾게 된다. 한국에서는 그 곳이 바로								
9	서울	경복궁	50	몇번을 가도 감탄하게 되며 산책하기 최고인 경복궁 매번 갈때마다 정교함과 운치에 또 한번 반하게								
10	서울	경복궁	50	구정전날 방문했는데 관광객이 참 많았습니다 무료입장이었고 한복입은 관광객이 많더군요 때맞추								
11	서울	경복궁	50	넓고 아름다운 우리나라 궁궐 건축물도 보면서 역사공부도 할 수 있는 곳이에요. 외국인들이 한국인								
12	서울	경복궁	50	친구랑 다녀왔는데 엄청 넓어요. 공사중인 곳 있어서 조금 아쉬워서 날씨 따뜻해지면 또 가려구요~								
13	서울	경복궁	50	도심 한 가운데서 만날 수 있는 조선왕조의 궁궐로 소란스러운 가운데 조용함을 느낄 수 있는 공간								
14	서울	경복궁	50	조선 최고의 궁궐입니다. 너무 좋아요 야경은 두 말 할 것도 없이 완전히 좋습니다. 즐기 전에 한 번								

● 크롤링 데이터 분석:워드클라우드, 감성분석

- 후기 명사 추출 후 워드클라우드를 생성해 관광지 특색, 개선점 분석
- 일부 관광지에 대한 추가적 인사이트를 위해 감성분석 통해 영향 끼친 요인 분석



```
show_most_informative_features(clf_1)
```

neg coef	word	pos coef	word
-7.0121	관광명소	-3.1001	경복궁
-7.0121	다음	-3.3232	한복
-7.0121	방문객	-3.6919	방문
-7.0121	식당	-3.7540	야간
-7.0121	유모차	-4.0503	개장
-7.0121	카페	-4.0854	한국
-7.0121	학생	-4.1034	사람
-6.7244	가까이	-4.1217	외국인
-6.7244	가능	-4.1987	정화루
-6.7244	가시	-4.2606	무료

4. 지역별 문화·여가 생활 위한 정책 제언 인천, 전남

타겟 지역 (17개 시도 중)

○ 인천

“음악 감상”

해당 지역의
토픽모델링
소비유형 키워드

송도달빛축제공원

타겟 문화·여가생활
장소
(주요관광지점 입장객
통계 기반 선정)



신한카드데이터 Topic 16

음악 감상이 주 여가활동. 인천 거주지 소비 8210건 중 '활동시간 중 인천사용 장소' 소비는 799건으로 10% 미만. 거주지 내 여가활동이 부족함을 알 수 있음.

국민여가데이터

짧은 여가시간을 고려하여 간단하게 즐길 수 있는 다양한 장르의 음악 행사 장려.

현재 인천시는 다수의 음악 페스티벌을 주최하지만 타겟 연령대가 다양하지 못하다는 단점이 존재.

종합 분석 내용

- 인천 내 공연, 버스킹 등 짧고 간단하게 즐길 수 있는 음악 관련 행사 장려
- 코로나 이후 사회 인식에 대비하여 노래방 부스 등 타인과의 접촉 없이 혼자 즐길 수 있는 여가시설 마련.

구체적인 정책 제언
내용

○ 전남

“숙박” “스포츠”

담양 죽녹원



출처: 죽녹원 홈페이지

신한카드데이터 Topic 13

스포츠에 관심이 많은 남성이 다수였음. 숙박 키워드가 높게 나온 것으로 보아 가족, 연인 단위로 여행을 오는 경우가 많을 것으로 파악.



- 지자체 측에서 숙박시설 관리 강화
- 담양 죽녹원에서 산책 등 스포츠 활동, 스포츠 용품 판매

4. 지역별 문화·여가 생활 위한 정책 제언 강원, 경기

○ 강원 “ 교육훈련 ” “ 여성 ” “ 30대 ”

강릉시 주민자치센터

2020년 하반기 교2동 주민자치센터 수강생 모집 공고

2020년도 교2동주민자치센터 문화강좌(프로그램)
하반기 수강생 모집을 아래와 같이 공고합니다.

○ 아 래 ○

- 공 고 일 : 2020년 07월 08일(수)
- 접수일시 : 2020년 07월 20일(월)~07월24일(금)까지 선착순 정원마감
07월 20일(월) 09:10분 (교2동주민자치센터 신분증지참)
07월 23일(목)부터 교2동주민자치센터 및 타동접수
- 접수장소 : 교2동주민자치센터 사무실(3층)
- 접수방법 : 방문접수 원칙 (야간반 전화접수 가능)
- 프로그램현황 (08월~12월+5개월)

프로그램명	정원(명)	회수(요일)	시 간	강의실	담당강사	수강료
요가수련A	25	주2회(토,목)	09:50~10:50	제1강의실	윤가람	육만원 (08월~12월)
요가수련B	25	주2회(토,목)	11:00~12:00	제1강의실	윤가람	육만원 (08월~12월)
요가수련A	25	주2회(토,수)	18:50~19:50	제1강의실	이유심	육만원 (08월~12월)
요가수련B	25	주2회(토,수)	20:00~21:00	제1강의실	이유심	육만원 (08월~12월)
라틴댄스	26	주2회(토,수)	10:00~11:30	제1강의실	김민수	월요일만 (08월~12월)
다이어트댄스	25	주2회(토,목)	19:00~20:30	제1강의실	박도유	월요일만 (08월~12월)
노래교실	60	주2회(토)	13:30~15:00	제1강의실	김선영	육만원 (08월~12월)
프로그래밍	20	주1회(월)	14:00~16:00	제2강의실	최경숙	월요일만 (08월~12월)
프로그래밍	20	주1회(수)	14:00~16:00	제2강의실	최경숙	월요일만 (08월~12월)
아로마테라피	20	주1회(목)	14:00~16:00	제1강의실	김정화	월요일만 (08월~12월)
서예명예교과목	20	주1회(월)	09:30~12:30	제3강의실	한영호	월요일만 (08월~12월)

수강료 입금계좌 : 9002-1359-3981-4 (새마을)예금주 :
교2동주민자치센터

- *기타 자세한 사항은 033-642-4612로 문의하여 주시기 바랍니다.
- *프로그램별 수강신청 인원이 70% 미만일 경우 해당 프로그램 폐강.

신한카드데이터 Topic 18

강원도 주 소비자층은 교육비에 주로 지출하는 30대 여성. 여가활동과 무관한 소비 내역이 나온 것으로 보아 강원도의 여가시설이 부족한 것으로 보임.

국민여가데이터

강원도민은 여가비용이 낮을수록 여가의 만족도가 높았으며, 공공시설의 인지도가 낮았음.

강릉시의 문화교육프로그램을 살펴본 결과, 수업료가 6만원으로 다소 비용적 부담이 있을 것으로 짐작됨.

- 공공여가시설 홍보 증진
- 주민자치센터 문화프로그램 수강료 개선

○ 경기 “ 스포츠 ” “ 숙박 ” “ 서울 거주 ”

양평 두물머리



출처: 경기도청 홈페이지

신한카드데이터 Topic 3

숙박 지출이 많은 점, 서울 거주자가 많은 점으로 보아 수도권인 경기도로 가족 단위의 여행을 오는 40대 남성일 것으로 판단.

신한카드데이터 Topic 7

여가 활동으로 미용에 관심이 있고 스포츠를 즐기는 모습을 보임.



트립어드바이저
크롤링 데이터
기반
워드클라우드

- 서울 거주자가 접근하기 좋은 근교 관광지로 두물머리 선정. 연꽃 족욕, 팩, 마사지 등 이색적인 미용 상품 출시. 스포츠를 즐기는 점, 워드클라우드에서 ‘한강’이 높게 나온 점을 감안하여 한강을 롤모델 삼아 산책로 조성 제안. ‘주차’가 눈에 띄는 점으로 보아 주차 문제와 대중교통 확충이 해결될 필요가 보임.
- 가족 단위로 스포츠를 즐기기 위해 과천 경마공원 제안. 가족 단위로 승마 체험 및 인근 캠핑장에서 글램핑 가능. 여행과 스포츠 여가활동 모두 향유 가능.

4. 지역별 문화·여가 생활 위한 정책 제언 전북, 제주

○ 전북 “공연 관람” “사진 촬영” “혼자”



군산 선유도

국민여가데이터

여가 비용 및 시간이 높을수록, 여가를 혼자 보낼수록 여가 만족도가 높음을 확인할 수 있었음.

신한카드데이터 Topic 8

전북의 카드 사용장소와 이용회원 거주지의 일치도가 매우 높는데, 이를 토대로 전북도민은 거주지에서 주로 소비함을 파악.

토픽 모델링 결과 공연 관람과 사진 촬영을 키워드로 선정.

- 전북도민을 대상으로 하는 사진 공모전
- 전북 관광지 (ex. 선유도) 곳곳에 포토존 확대

○ 제주 “관광쇼핑” “미술공예참여”

제주 세화해변 카페공작소



신한카드데이터 Topic 14

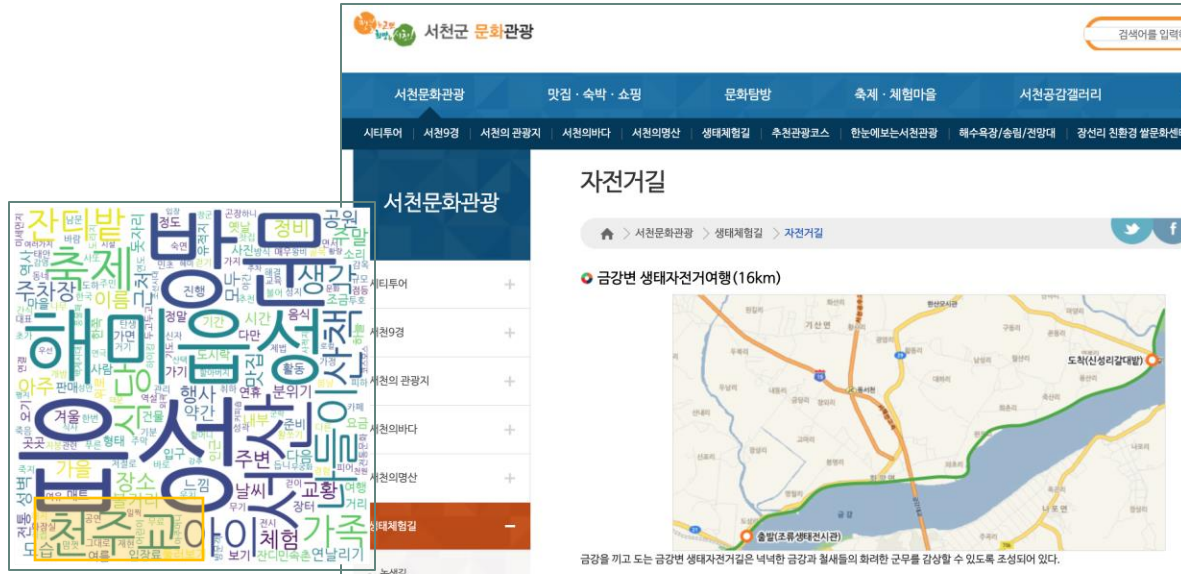
관광쇼핑, 미술공예활동을 주로 즐김.

이미 관광이 충분히 발달한 지역이기 때문에 미술공예 여가활동에 초점을 둘 것을 제안함.

- 관광쇼핑과 미술공예를 융합, 여행 기억을 담을 수 있는 공예품 제작
- 관광객뿐 아니라 거주자들도 즐길 수 있는 미술공예행사 기획. • 타인과의 접촉에 구애받지 않고 각자 즐길 수 있는 공예 장소 마련 또는 공예 키트 판매.

4. 지역별 문화·여가 생활 위한 정책 제언 충남·충북, 부산

○ 충남·충북 “종교활동” “자전거” “휴일”



출처: 서천군청 홈페이지

신한카드데이터 Topic 17, 19

충북, 충남에서는 휴일에 각각 종교활동과 자전거 활동이 주 여가활동임.

워드클라우드를 통해 해미읍성이 천주교와 밀접함을 확인.



- 서천 자전거 여행 코스 등 자전거 관련 활동 제안, 홍보 확대.
- 서산 해미읍성 등 종교 관련 유적지에서 종교 행사 강화.

○ 부산 “거주지 내”

부산_거주지	0
경남_거주지	0.000003732636587
부산_사용장소	0.000008248184448
경기_사용장소	0.00001106178742
경북_거주지	0.00001234775312

부산_사용장소	0
부산_거주지	0.000008248184448
인천_거주지	0.00001786347601
경남_거주지	0.00001827371533
대전_사용장소	0.00001827676915

국민여가데이터

여성의 여가 만족도가 더 낮음.

신한카드데이터 Topic 15

Vector Space Model 중 SVD 를 참고하여 부산 거주지와 부산 사용장소의 거리가 매우 가까워 부산 거주민은 부산 내에서 소비함을 알 수 있음.

 감천문화마을



출처: 감천문화마을 홈페이지

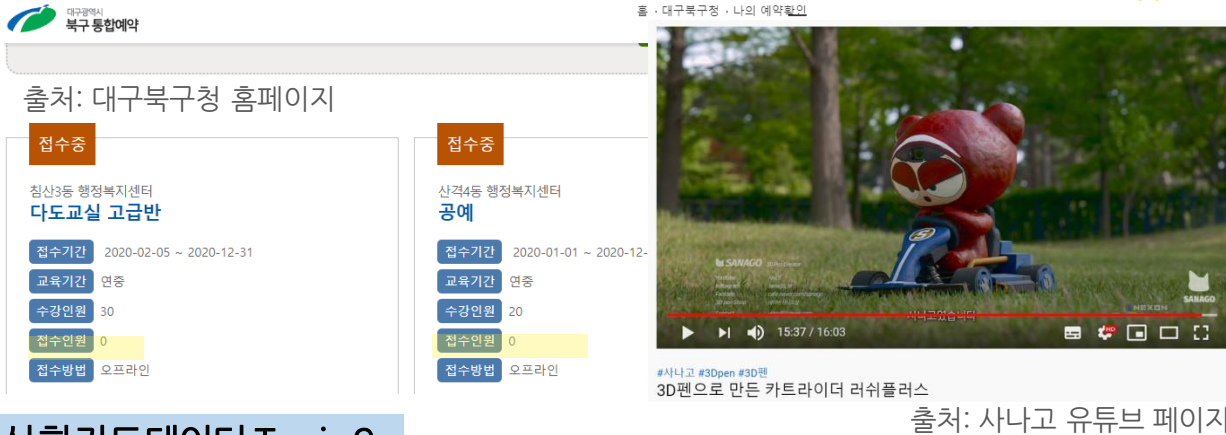
• 부산 내 여가시설 확충

- 감천문화마을 등 부산 내 유명 관광지 적극 활용, 벽화그리기 봉사활동, 유명인 벽화 등 이색 여가활동 고안.

4. 지역별 문화·여가 생활 위한 정책 제언

○ 대구·대전

“ 인터넷 게임 ” “ 휴식 (19시~02시) ”



신한카드데이터 Topic 9

인터넷 게임이 주 여가활동. 대전 사용장소 SVD도 대구 사용장소 SVD와 매우 가까운 거리에 위치. 인터넷 게임은 불건전한 여가활동으로 분류하였기 때문에 **건전한 여가활동의 필요성 강조.**

대구 문화센터 접수인원이 0인 것으로 보아 공공여가시설 이용도가 매우 낮음.

- 인터넷 게임을 테마로 하는 공예, 전시 등 관련 활동을 할 수 있는 인프라 마련.
- 대전 국립중앙과학관에서 게임 VR 체험, 게임 관련 과학 체험을 제공할 수 있을 것으로 보임.
- 문화센터 프로그램의 **홍보를 강화** 및 수강 기간을 1년 이내로 축소 제안.

○ 서울 “서울 거주” “관광쇼핑”



신한카드데이터 Topic 14

서울의 전통 유적지에 한국인의 방문이 적다는 의식 재고의 필요성. 서울에 거주하며 서울에서 여가를 보내는 사람이 많음을 토픽 모델링을 통해 확인.

경복궁 크롤링 결과 한국인 방문객이 의외로 많음을 발견.

국민여가데이터

관광, 문화, 여가 생활의 다양성 척도가 높게 나왔으며, 혼자 즐길수록 만족도가 높은 추세를 보임. 서울시민들은 다양한 여가활동을 즐기고 있으며 시설 구축이 비교적 준수하기 때문에 특정 여가활동을 제시하기보다 현재의 기조를 유지.

- 경북궁 야간개장의 확대 또는 개선 필요
- 혼자 즐길 수 있는 다양한 여가 활동 제공

4. 지역별 문화·여가 생활 위한 정책 제언 특징 없는 지역, 기타 제언

◦ 특징 없는 지역들: 울산, 세종, 경남, 경북, 광주

울산, 세종, 경남, 경북, 광주의 경우 표본 및 데이터 부족으로 소비자층 및 여가생활을 분석하기에 어려움이 있었으며, 분석의 한계점에 해당함. 따라서 지역적 특성을 살린 여가활동을 구체적으로 제시하기에는 한계가 존재.

국민여가 활동조사 데이터 분석 결과를 바탕으로 정책 제언

- 전국적인 추세를 보았을 때 생활권 내 공공문화여가시설을 이용하는 사람이 더 만족도가 높았으므로 여가생활의 만족도를 증진을 위해 공공문화여가 활동을 활성화할 필요가 있어 보임.
- 연령대별 특성을 반영하여 타겟팅한 문화·여가활동 구성.

◦ 기타 제언 - 홍보 방법 등

홍보 방법

- 사진촬영 키워드를 활용하여 사진을 SNS 에 업로드하거나 공모전을 제출하면 인센티브 제공. 거주지 내 여가활동을 장려하기 위해 자신이 거주하는 지역의 사진 행사에 참여하는 참여자에게 혜택.
- 공공여가시설 홍보에 개선이 필요해 보임. 홈페이지 내 배너 광고, 또는 거주자에게 문자메세지로 전달하는 등 현재보다 적극적인 홍보를 통해 거주자들의 여가활동 참여 증진.

많은 사람들과 접촉하는 대중교통과 달리 전동 킥보드나 자전거, 드라이브 코스는 포스트 코로나를 고려했을 때 타인과의 접촉이 적은 여가활동으로 매우 적합. 따라서 사람들이 안전하게 즐길 수 있는 여가생활임을 주지시킬 필요.

5. 분석 의의 및 활용 방안 제언

◦ 포스트 코로나 시대의 문화·여가 생활 준비의 방향성 제시

17개 시도에 대해 국민여가활동조사 데이터, 신한카드 데이터, 트립어드바이저 크롤링 데이터 분석 결과를 반영한 각 지역별 정책 마련의 방향성을 4장 지역별 지역별 문화·여가 생활 위한 정책 제언 내용을 바탕으로 구성하여 국민들이 포스트 코로나 시대의 문화·여가 생활을 누릴 수 있도록 대비할 수 있음.

◦ 미래의 문화·여가 트렌드 예측 가능성 제시

2장 데이터 내용에서 볼 수 있듯 머신러닝 등의 방법을 활용하여 미래의 문화·여가 트렌드를 예측할 수 있으며, 이를 바탕으로 만족도가 높은 문화·여가 산업을 유지 및 성장시킬 수 있음.

◦ 실시간으로 생성되는 빅데이터의 특성을 적극 활용

빅데이터는 실시간으로 생성되는 특성을 가지며, 이러한 데이터를 즉각적, 지속적으로 반영하며 실시간 분석까지 이룬다면 문화·여가 산업을 더욱 활성화시킬 수 있을 것으로 판단됨.

- 사진 키오스크 제안: 관광객들이 여가생활을 할 때 사진을 많이 찍으므로 이러한 사진 자체가 중요한 데이터가 될 수 있음. 키오스크에 여가, 관광 목적, 성별, 연령대 등 간단한 정보를 입력 시 고품질 사진 인화 및 전송 서비스를 제공하는 키오스크를 설치하여 관광객에게는 특별한 추억을 제공하는 한편, 빅데이터 수집 가능. 현재진행형으로 축적되는 데이터를 바탕으로 개선점에 발빠르게 대응할 수 있을 것으로 전망.