동반자 유무에 따른 관광객의 온라인 리뷰 차이분석:
토픽모델링 기법을 중심으로

경기대학교 최지나
경기대학교 가정혜

목차

1.서론

2.이론적 배경

3.연구 설계

4.연구 결과

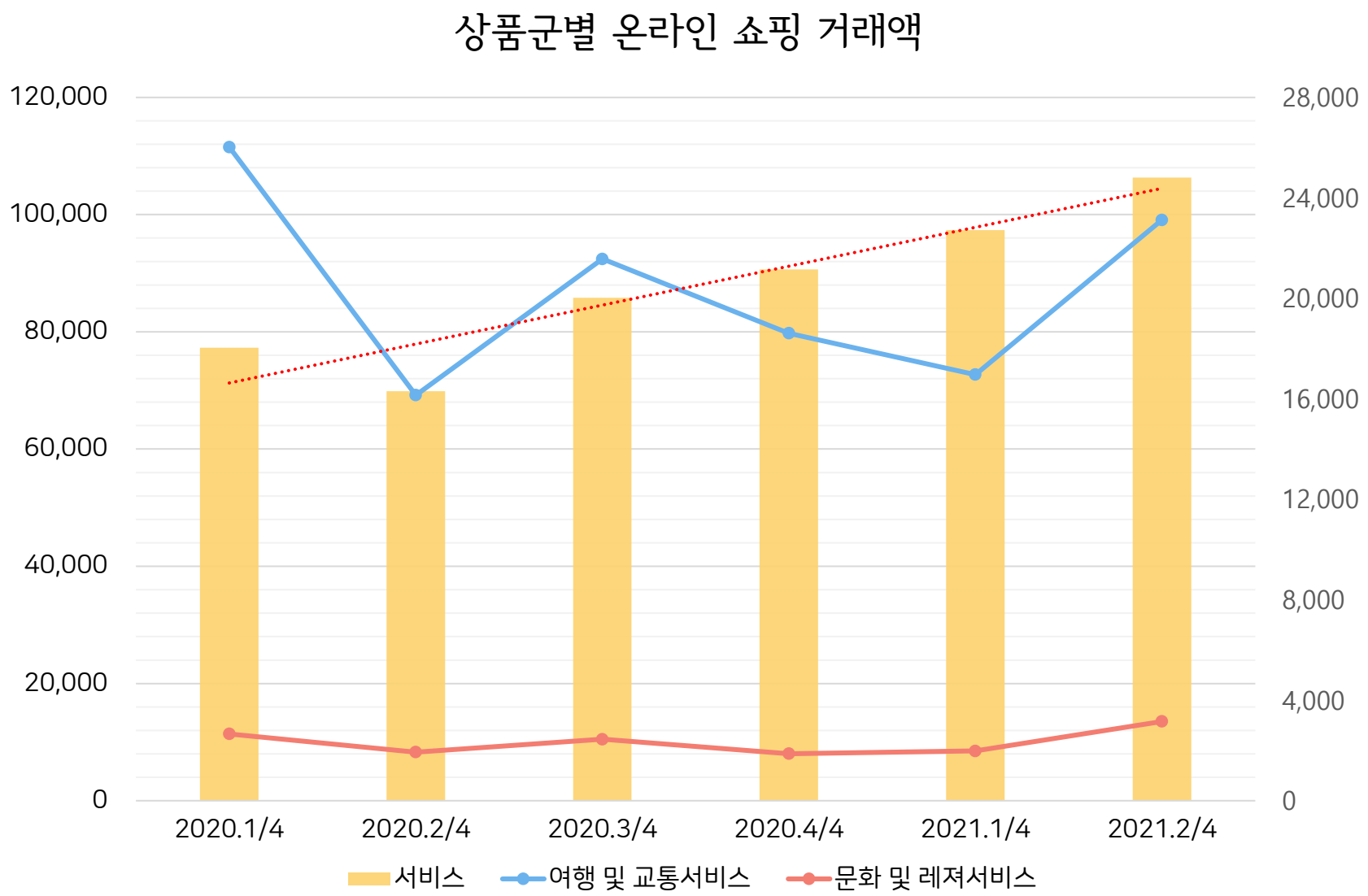
5.결론

온라인 시장의 확대

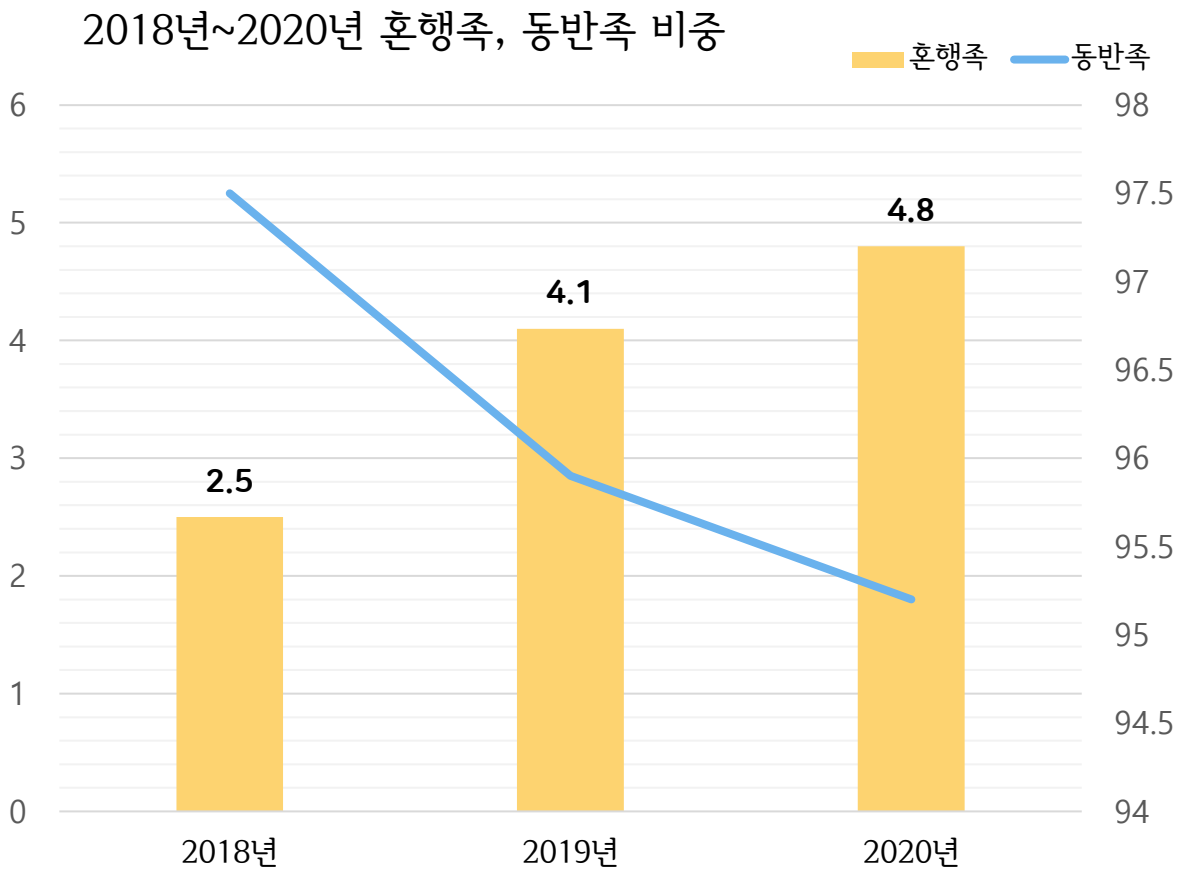
온라인 쇼핑을 통한 '서비스' 부문 거래액의 지속적 증가

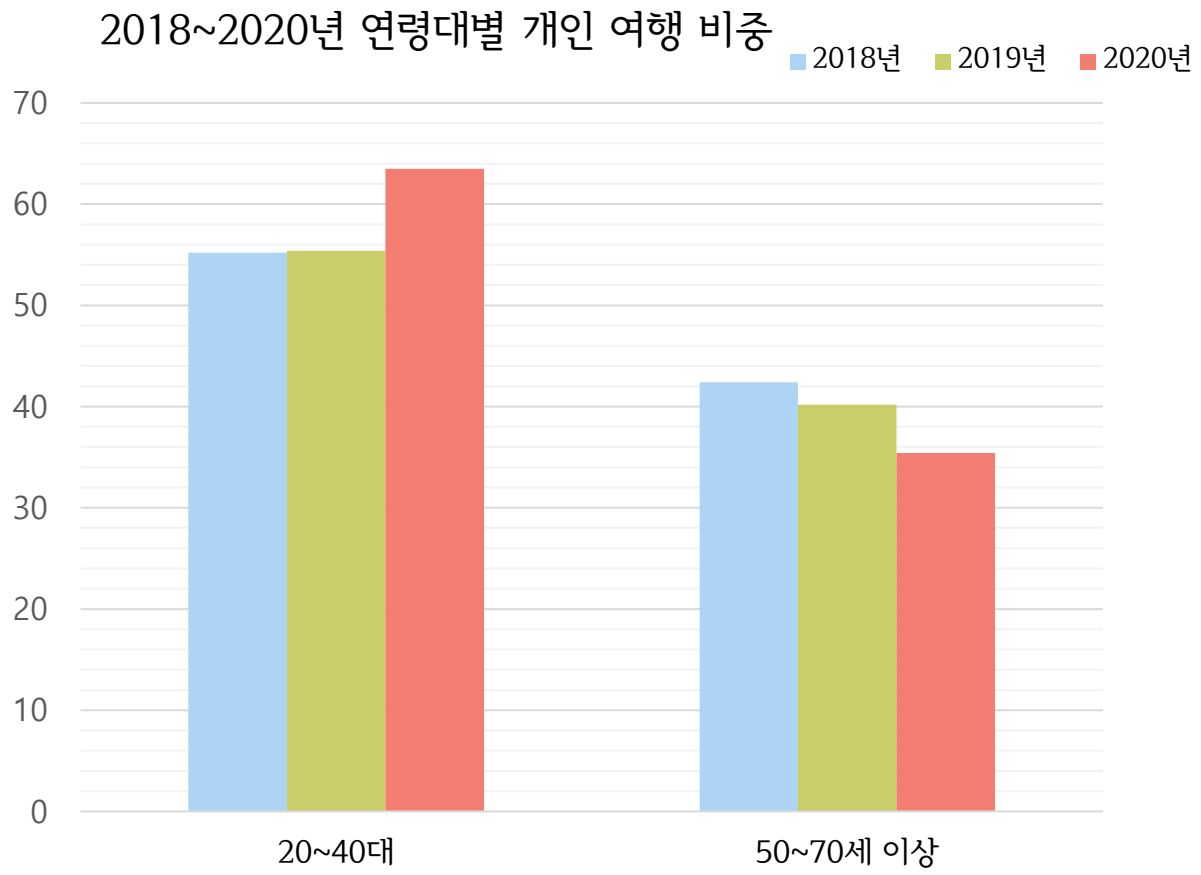
온라인 쇼핑에서 매출로 이을 때의 리뷰 및 평점의 역할 ↑
(최자영, 김현아, 김용범(2020)

평점<리뷰 제목/내용의 감성 정도
(최자영, 김현아, 김용범(2020)



개인 여행의 증가





- ✓ 동반족의 비중이 압도적이지만, 혼행족의 증가 추세를 확인할 수 있음.
- ✓ 20~40대의 혼행족의 비중이 점차 증가 → 지속적으로 개인 여행의 비중이 증가함을 추측 가능.
- ✓ 박영욱, 정규엽(2021)은 향후 과제로서 여행행태별 감성의 차이를 분석하는 것을 언급.

감성분석

순서

데이터 수집 → 자연어 처리 —(감성사전 구축)→ 학습 → 분석 → 결과 확인

오타 또는 맞춤법 수정
불용어 제거
형태소 분리

(사전 사용)

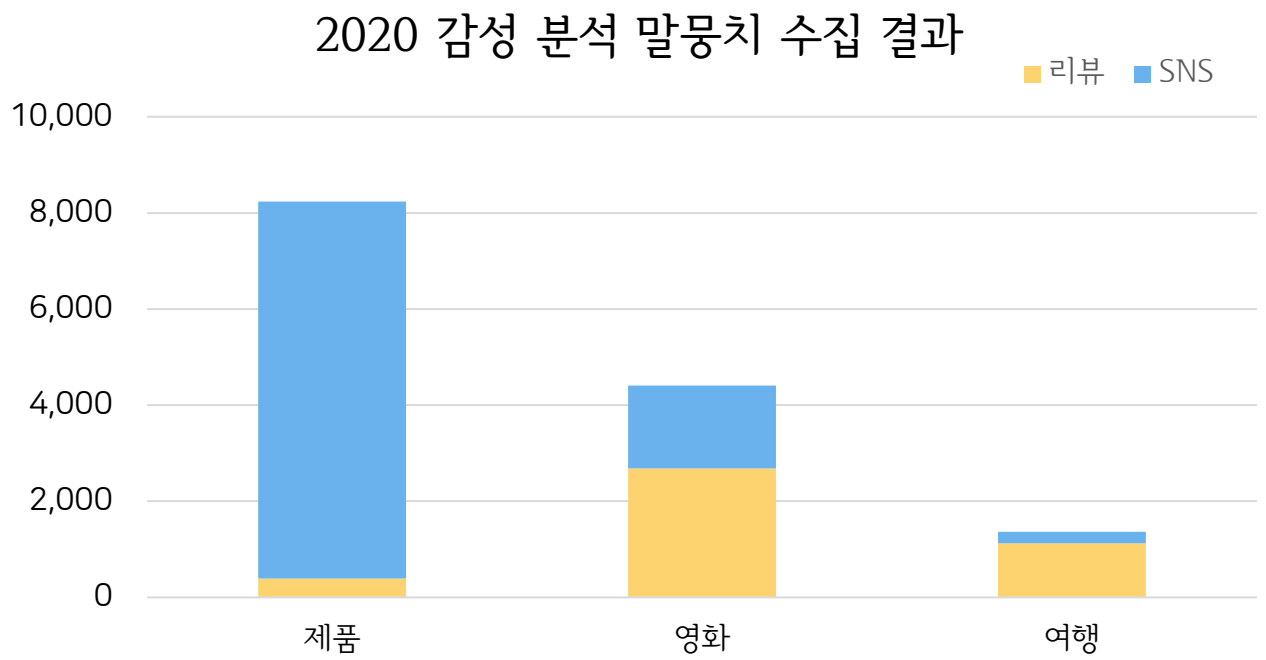
문제점

사전 구축을 위한 정제된 말뭉치 필요

→ 미국(2,000억 단어),

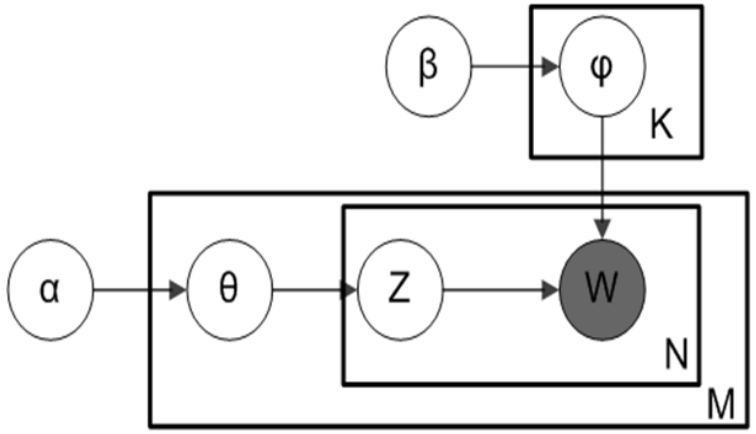
일본, 중국, 유럽(100억 단어 이상)

구축된 한국어 감성 사전의 빈약함.



LDA 토픽모델링

LDA : 각 단어나 문서의 숨겨진 주제를 찾아내어 문서와 키워드별로 주제끼리 묶어주는 비지도 학습 알고리즘



M : 문서의 개수 N: 문서에 속한 단어의 개수
Θ : 문서의 토픽 디리클레(Dirichlet) 분포
Φ : 주제의 단어
Z : 해당 단어가 속한 토픽의 번호
W : 실제 관측 가능한 값
K : 토픽 개수
α : 문서들의 토픽 분포의 밀집도
β : 문서 내 단어들의 토픽 분포의 밀집도

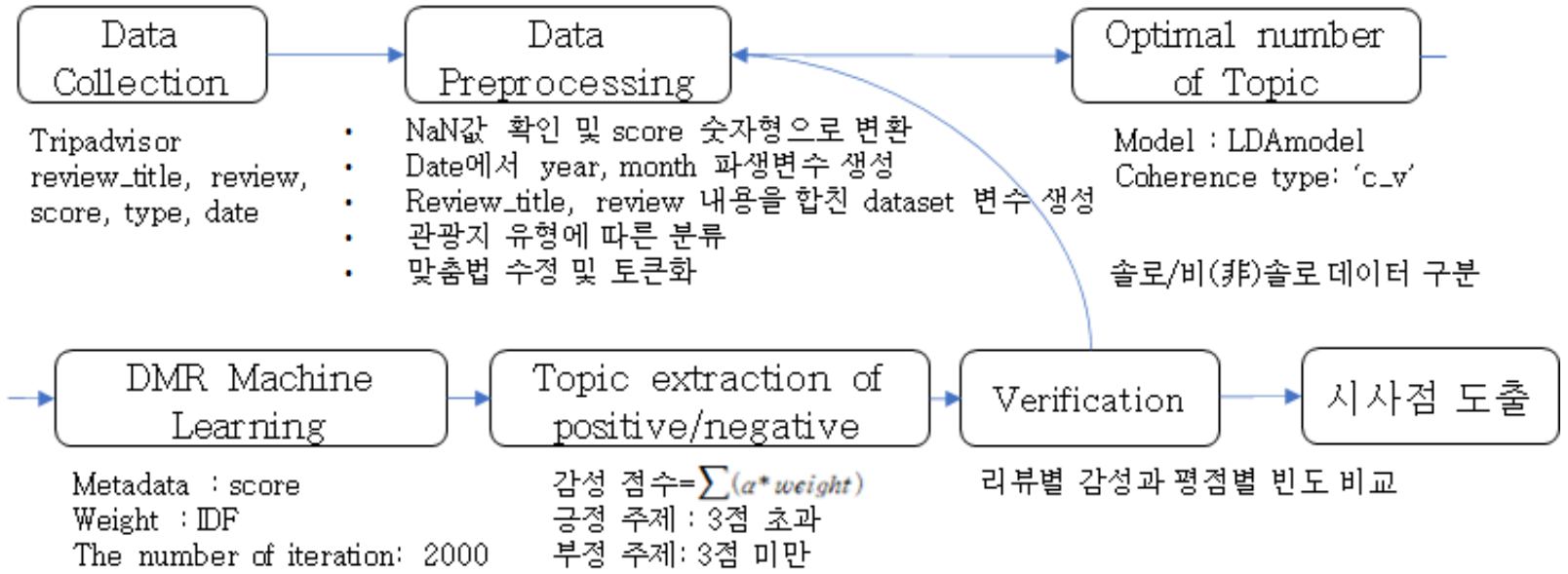
	모든 단어의 동일한 가중치	평점 데이터 미활용
문제	출현 빈도 多 → 높은 빈도의 단어 위주로 주제 도출	LDA는 리뷰 내용만을 이용→ 토픽의 감성분석 필요
해결책	IDF(Inverse Document Frequency) 적용	DMR(Dirichlet Multinrmial Regression) 이용

$\log(N/n) + 1$ N: 전체 문서의 개수, n: 특정 단어가 포함된 문서의 개수

특정 단어가 모든 문서에 출현한다면 IDF=1
해당 단어를 포함한 문서의 수가 적을수록 값↑

연구 단계 및 과제

연구 단계



연구 과제

RQ1 : DMR을 통해 고객들이 남긴 평가점수(1~5점) 그룹 별 주제 비중을 파악하여 긍정과 부정 주제로 분류

RQ2 : 긍·부정으로 분류된 주제에 대해 각 주제와 관련된 리뷰를 추적 후,

해당 리뷰가 실제 긍정(4~5점) 또는 부정(1~2점) 그룹에서 얼마나 자주 출현했는지 파악

RQ3 : DMR을 통한 긍·부정 주제와 점수 별 출현 빈도를 비교하여 본 연구 방법의 타당성 검토

RQ4 : 여행 타입 별 데이터로 분류하여 RQ1, 2, 3을 재진행

RQ5 : 검증이 완료된 주제에 대해 여행 타입별로 영향을 주는 관광지 유형에 따른 속성을 파악하여 시사점 도출

데이터 수집 및 정제

수집 및 정제

<원본 데이터>

review_title	review	score	date
대한민국의 역사	대한민국의 역사가 잠들어 있는 곳. 서울을 방문했다면 꼭 방문	풍선 5개 중 5.0	2020년 10월
국민이 공감하는	경복궁은 국민들이 자주 찾는곳으로 작성자는 주말에 자주 가 하고 있음.	풍선 5개 중 5.0	2020년 7월
산책하기 좋은 경	날씨 좋은 날 종종 산책하러 경복궁에 가는데 마음이 편온해지	풍선 5개 중 5.0	2020년 3월
Good	Goooooood 다 좋습니다 다음에 또 오고 싶네요 근처 관광지도	풍선 5개 중 5.0	2020년 4월
가족단위로 방문	요새 더더욱 코로나로 인해 사람 방문이 적음. 두자녀 동반시	풍선 5개 중 5.0	2020년 4월 • 가족
하늘이 내린 큰	조선 개국 4년째인 1395년에 처음으로 세운 으뜸 궁궐이다.	풍선 5개 중 5.0	2020년 3월

파이썬(Python)의 Selenium 패키지 이용하여 크롤링(Crawling)

- 리뷰 제목, 리뷰 내용, 평점, 날짜 및 타입

→ 평점과 날짜 및 타입의 필요한 형태로의 변형

- 토픽 모델링을 위한 리뷰 제목 및 리뷰 내용을 합친 변수 생성('dataset')

→ 네이버 맞춤법 검사기 라이브러리 'HanSpell' 이용하여 맞춤법 검토

- 'dataset' 열의 결측치 제거
- 관광지 유형에 따른 분류 기준대로 파일 생성

<관광지 분류>

활동 장소	여가 장소	2차적 요소	부가적 요소
국립중앙박물관, 한국 전쟁 기념관, 트릭아이뮤지엄 서울, 명동난타극장, 러브뮤지엄, 삼성미술관 리움	경복궁, 북한산 국립공원, 창덕궁, 한강공원, N 서울 타워, 남산공원, 청계천, 조계사, 봉은사, 하늘공원, 이화여자대학교, DDP, 서울숲, 여의도 한강공원, 서울스카이	명동 쇼핑 거리, 인사동, 롯데월드타워& 몰, 북촌 한옥마을, 홍대앞 거리, 별마당 도서관, 광장시장	KTX, 서울메트로

DMR 학습 준비 및 학습

Tokenize

파이썬(Python)의 Selenium 패키지 이용하여 크롤링(Crawling)

- 한글 형태소 라이브러리 'KoNLPy'의 'Okt' 클래스 이용
- 불용어 리스트 : 우리, 함께, 있습니다

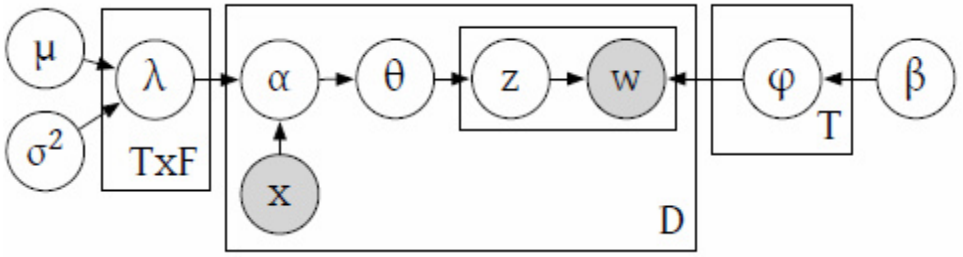
최적 토픽 개수 판단

응집도 : 각 토픽에서 상위 비중을 차지하는 단어들이 의미적으로 얼마나 유사한지를 나타내는 척도

Cf. 혼란도 : 주제가 적정 개수일 때는 확률적을 일관되게 단어가 주제에 할당되는 정도

- 응집도가 높거나 혼란도가 낮다고 해서 무조건적으로 적절한 토픽 개수는 아니기 때문에 결과를 보고 연구자가 판단해야 한다.
- 'tomotopy' 라이브러리를 통해 1개~20개 사이의 DMR 실시 후 응집도 계산 수행

DMR



	메타 데이터		
$a = \exp(\lambda)$	A	B	C
주제1	0.341	2.565	1.216
주제2	2.377	0.326	1.216

가정: 문헌의 주제 분포를 관장하는 하이퍼 파라미터 α 가 문헌의 메타데이터에 따라 다를 것.
x : 문헌의 메타데이터
λ : 평균이 μ이고 표준편차가 σ인 정규분포를 따르는 메타데이터 별 하이퍼 파라미터 결정 값

원래 LDA에서는 α를 정해야 하지만, 학습과정에서 최적해로 수렴.

메타데이터를 가지는 어떤 문헌이 각 주제를 얼만큼 포함하는지 확인 가능
→ 메타데이터 A인 문헌의 주제분포는 D(0.341, 2.377)
→ 메타데이터 A를 가지는 어떤 문헌이 주제2를 평균적으로 얼만큼 포함하는지 :
2.377/(0.341+2.377) = 0.874

감성점수 계산 및 타당성 검증

감성점수 계산

1. 주제 단어 확률분포 P(word | topic), 리뷰 주제 확률분포 P(topic | review), 점수 그룹 별 주제의 비중 α
2. P(word | topic) 통해 해당 주제의 의미와 내용 파악
3. 각 주제에서 1점부터 5점까지의 α값을 그 주제의 α값의 합으로 나누어 비중 계산

p주제 감성점수= $\sum_{q=1}^{5}(a_{p,q}/\sum^{q} a_{p,q} \times q)$

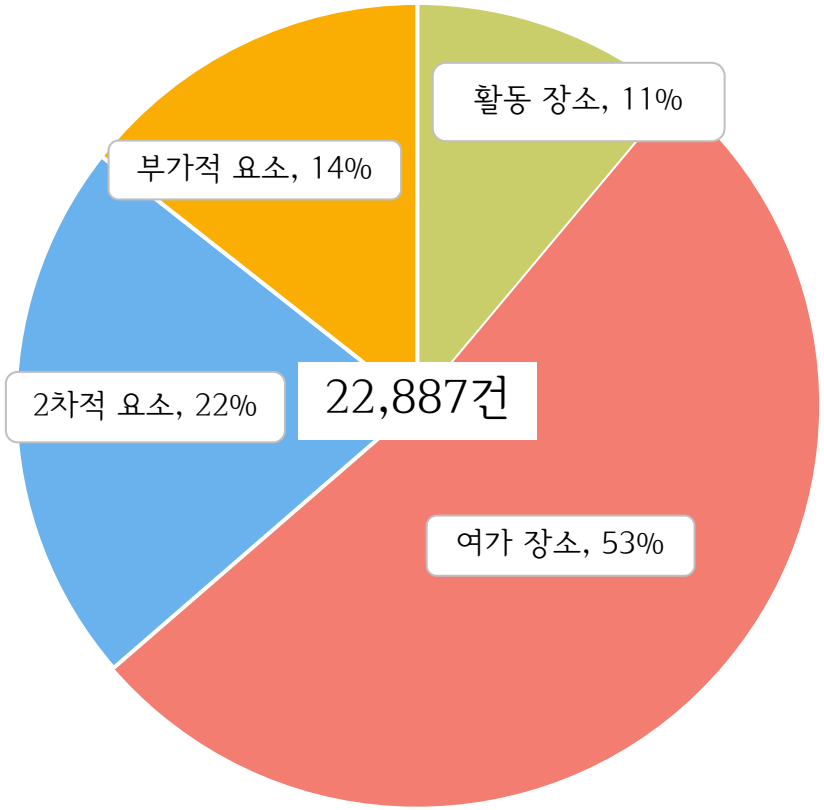
p: 주제 번호, q : p 주제의 q점수 그룹의 α값에 대한 비중

타당성 검증

1. 리뷰 주제 확률분포 P(topic | review)가 50%이상의 확률을 보이는 주제에 대해 그 주제의 리뷰라고 판단
2. 주제, 점수별 실제 빈도와 기대빈도를 구한 후, 그 둘의 차이를 기대빈도로 나누어 빈도비율 계산
3. 빈도 비율을 통한 긍/부정과 감성 점수를 통한 긍/부정의 일치 하는지 비교

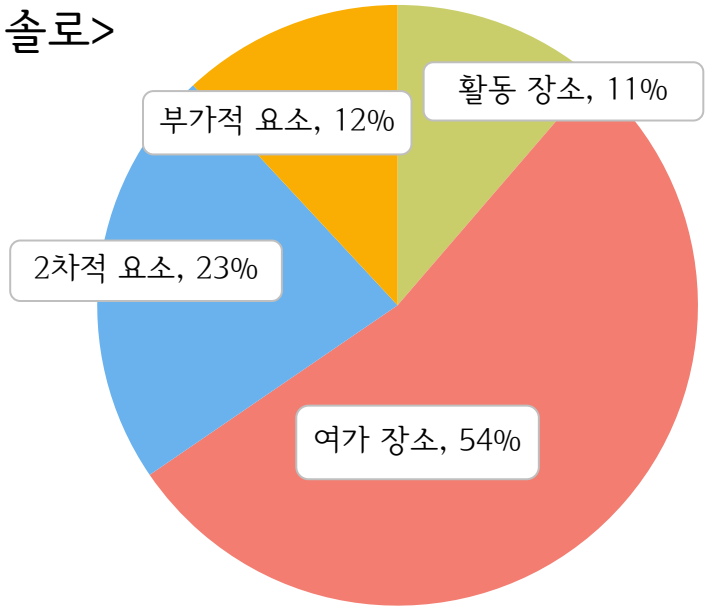
표본의 특성

<전체 데이터>



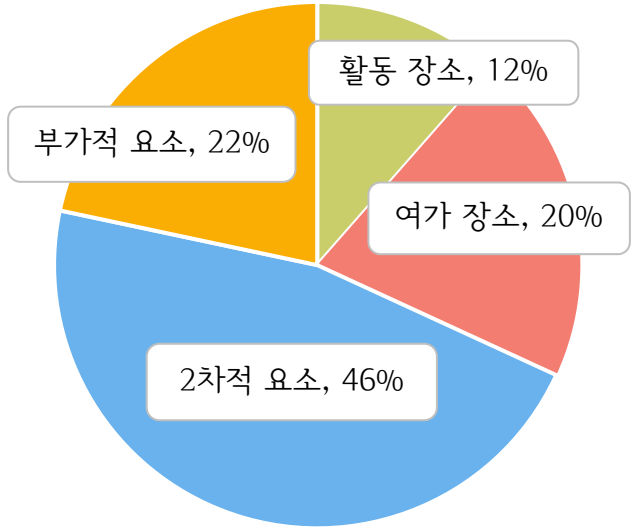
15,399건

<비(非)솔로>



3,617건

<솔로>



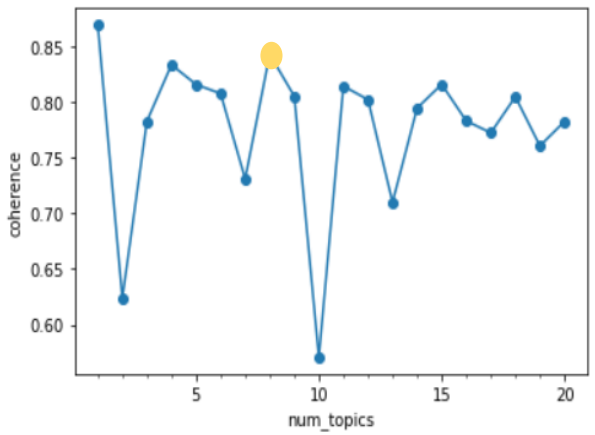
DMR 결과 - 응집도 점수

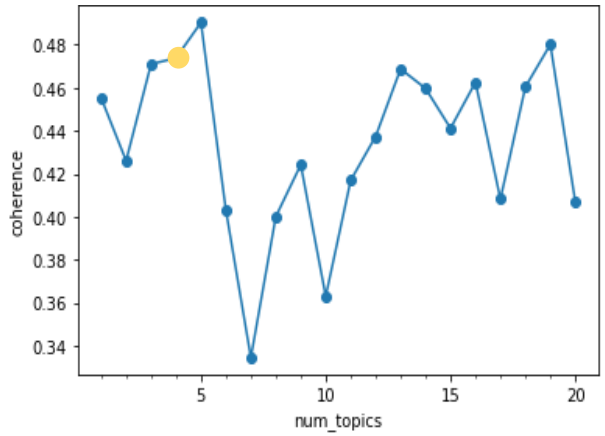
<전체 데이터>

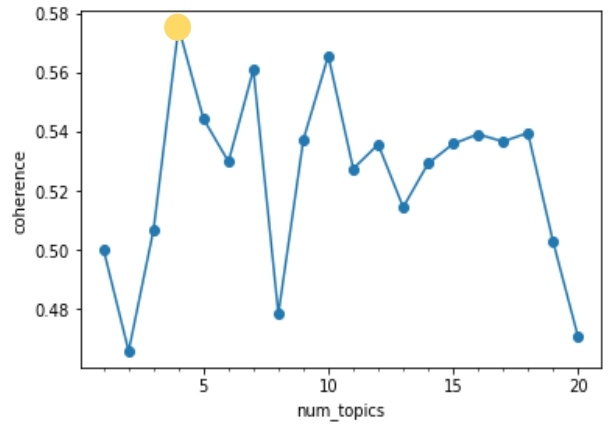
<비(非) 솔로>

<솔로>

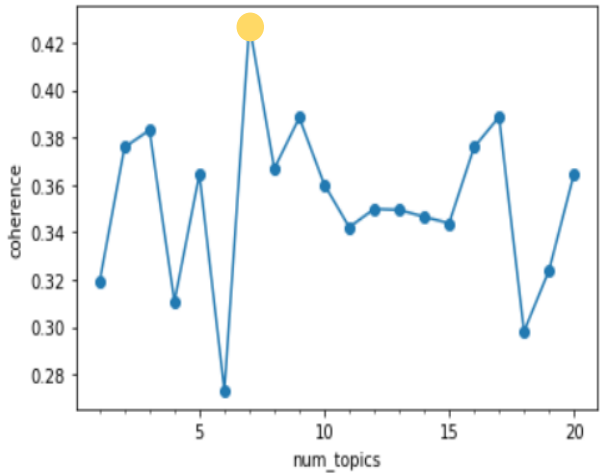
<활동 장소>

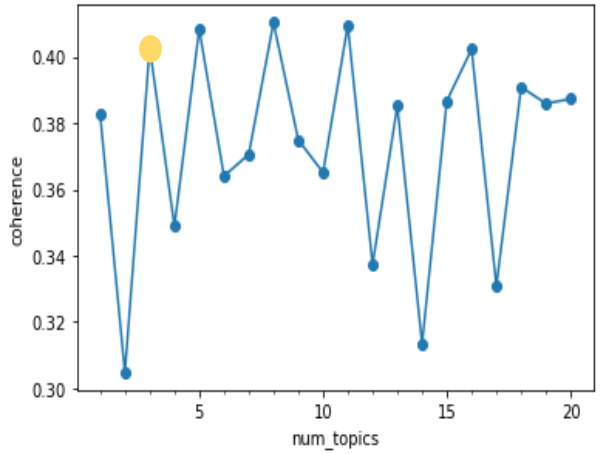


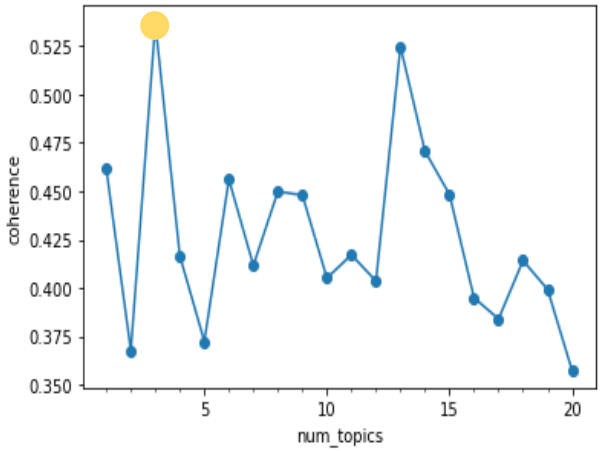




<여가 장소>

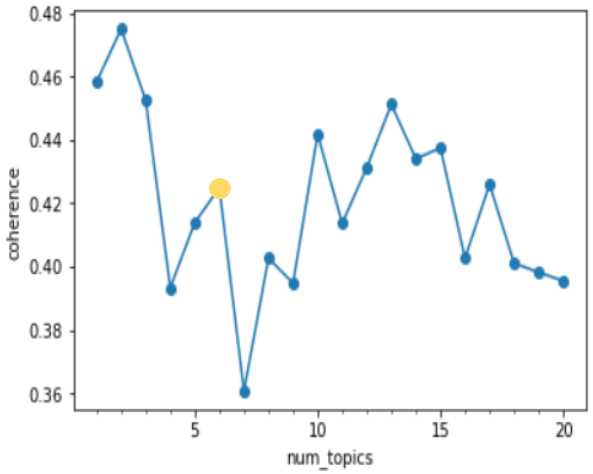




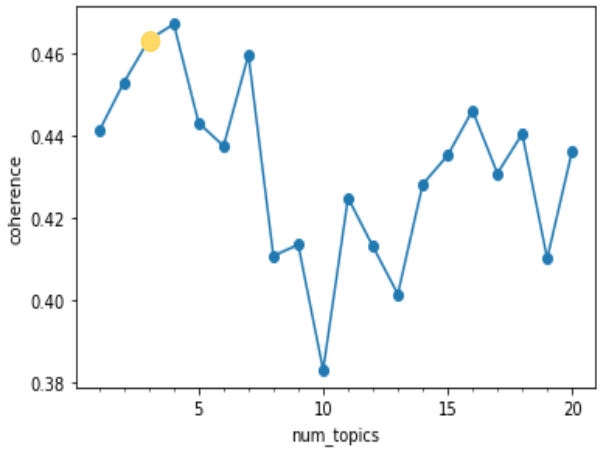


DMR 결과 - 응집도 점수

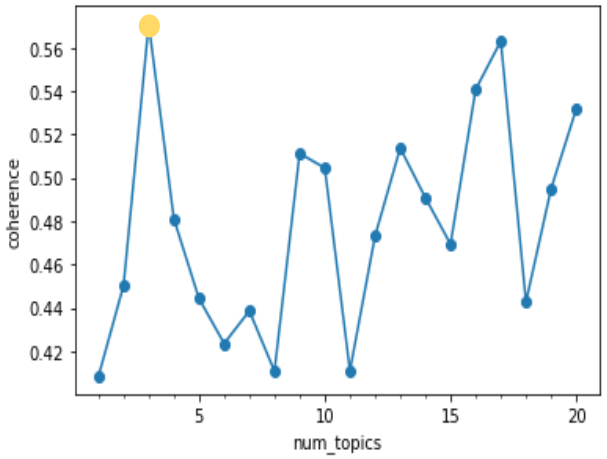
<전체 데이터>



<비(非) 솔로>

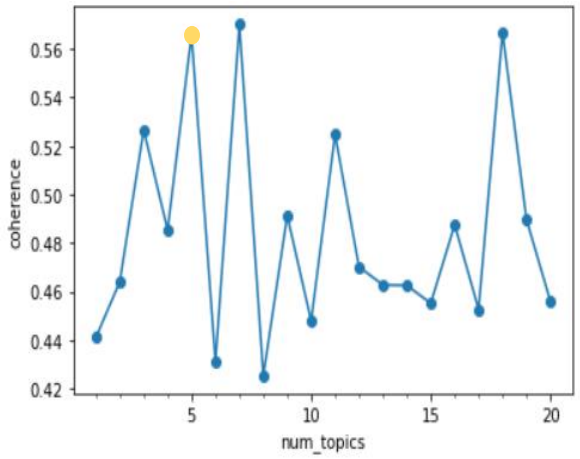


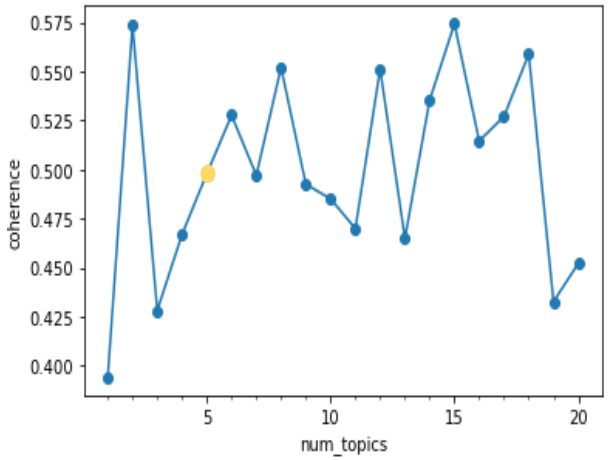
<솔로>

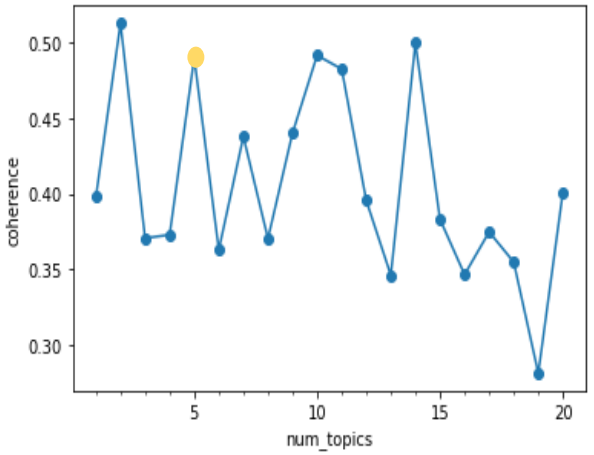


<2차적 요소>

<부가적 요소>







활동 장소 - 전체 데이터

Topic	Sentiment score	Topic Title	Representative top 10 words				
1	2.98 (Neg)	미술관	미술관	삼성	리움	이태원	해설
			운영	게임	혼자	월요일	보면
2	3.17 (Pos)	자식 교육	아이	어린이	교육	있어	되어
			비행기	있어서	탱크	많고	전쟁기념관
3	2.36 (Neg)	재미	재미있는	그것	사진	공연	재미
			극장	당신	난타	재미있고	친구
4	2.87 (Neg)	사진	사진	얼음	가격	아이	트릭
			친구	재미있는	시간	재미	티켓
5	1.99 (Neg)	미술관 전시	전시	작품	미술관	너무	기획
			입장료	리움	부족한	있어	모네
6	3.89 (Pos)	역사 박물관	좋아요	관람	중앙	외국인	전시
			생각	있어서	우리나라	방문	역사
7	3.15 (Pos)	가이드	예술	가이드	시간	건물	아주
			여기	하지	합니다	매우	방문
8	2.96 (Neg)	전쟁사	전쟁	역사	많은	무료	표시
			기념관	모든	정보	시간	매우

여가 장소 - 전체 데이터

Topic	Sentiment score	Topic Title	Representative top 10 words				
1	3.39 (Pos)	고궁 풍경	투어	궁전	가이드	가든	정원
			시크릿	팰리스	역사	사원	아름다운
2	1.82 (Neg)	다양한 컨셉	건물	디자인	쇼핑	마을	전통
			지역	사진	거리	대학	북촌
3	3.38 (Pos)	도심 공원	도시	산책	아름다운	많은	스트림
			따라	시간	공원	방문	아주
4	2.36 (Neg)	한복	입어	아름다운	방문	궁전	사진
			시간	전통	한국	팰리스	사람
5	1.93 (Neg)	가이드 투어	궁전	방문	역사	한국	시간
			팰리스	건물	하는	투어	아름다운
6	3.26 (Pos)	전망	타워	전망	케이블카	타고	버스
			멋진	남산	도시	사랑	풍경
7	2.23 (Neg)	한강 공원	한강	좋아요	공원	너무	야경
			산책	도심	입니다	좋습니다	자전거

2차적 요소 - 전체 데이터

Topic	Sentiment score	Topic Title	Representative top 10 words				
1	2.75 (Neg)	외국인 쇼핑	외국인	명동	관광객	너무	중국인
			사람	롯데	중국	많고	타워
2	2.91 (Neg)	시설 유형	명동	재래시장	서울	방문	가격
			롯데	하지만	백화점	입구	거리
3	3.74 (Pos)	쇼핑 지역	많은	좋은	상점	지역	거리
			레스토랑	장소	사람	모든	서울
4	3.47 (Pos)	전통 기념품	기념품	좋은	전통	인사동	한국
			가게	많은	상점	지역	예술
5	2.85 (Neg)	야외 먹거리	먹거리	빈대떡	광장시장	홍대	거리
			김밥	외국인	인사동	구경	육회
6	3.44 (Pos)	화장품	많은	모든	제품	화장품	좋은
			길거리	가게	경우	상점	시장

부가적 요소 – 전체 데이터

Topic	Sentiment score	Topic Title	Representative top 10 words				
1	3.26 (Pos)	효율성	이용	환승	호선	수단	최고
			출퇴근	나라	교통	시간	세계
2	2.23 (Neg)	출퇴근	거리	홍콩	있어	하게	다른
			서로	아침	위치	매일	경험
3	3.16 (Pos)	결제 시스템	카드	사용	티켓	영어	여행
			시스템	쉽게	방법	모든	좋은
4	1.43 (Neg)	무선 인터넷	인터넷	와이파이	무선	세계	손쉬운
			통한	무료	된다	게다가	시간
5	3.77 (Pos)	서비스	기차	매우	깨끗하고	쉽게	시간
			아주	사용	여행	사람	좋은

1차적 요소 - 집단 구분

Topic	Sentiment score	Topic Title	Representative top 5 words				
비(非)솔로, 활동 장소							
1	4.37 (Pos)	입장료	전쟁	역사	많은	무료	시간
2	2.28 (Neg)	가족	전시	관람	아이	좋아요	있어서
3	4.15 (Pos)	추억/사진	사진	재밌는	경험	친구	재미
4	3.01 (Not Defined)	한국	전쟁	나라	너무	그녀	않고
솔로, 활동 장소							
1	4.35 (Pos)	시점/시간	전시	시간	방문	모든	입니다
2	3.86 (Pos)	관람	사진	미술관	문화재	작품	국립
3	4.02 (Pos)	정보	정보	투어	최고	가이드	하루
4	3.91 (Pos)	설명	전쟁	거대한	대해	가지	영어
비(非)솔로, 여가 장소							
1	3.57 (Pos)	고궁 투어	궁전	아름다운	방문	투어	한국
2	2.50 (Neg)	공원	한강	좋아요	너무	공원	산책
3	2.53 (Neg)	전망	타워	전망	케이블카	타고	버스
솔로, 여가 장소							
1	3.60 (Pos)	전망	타워	전망	케이블카	남산	타고
2	3.65 (Pos)	시점/시간	궁전	방문	아름다운	시간	투어
3	3.50 (Pos)	산책	산책	하기	경복궁	한강	따라

2차적 및 부가적 요소 - 집단 구분

Topic	Sentiment score	Topic Title	Representative top 5 words				
비(非)솔로, 2차적 요소							
1	3.38 (Pos)	인기 관광지	외국인	사람	너무	구경	먹거리
2	3.59 (Pos)	기념품	좋은	기념품	인사동	가게	많은
3	3.93 (Pos)	다양한 가게	많은	모든	상점	좋은	제품
솔로, 2차적 요소							
1	2.71 (Neg)	외국인 관광객	외국인	명동	관광객	사람	너무
2	3.53 (Pos)	쇼핑 번화가	상점	많은	음식	지역	가게
3	3.65 (Pos)	인기 관광지	한국	인사동	홍대	많고	사람
비(非)솔로, 부가적 요소							
1	2.72 (Neg)	시설	계단	엘리베이터	에스컬레이터	부산	ktx
2	3.08 (Not Defined)	편리성	이용	환승	호선	너무	좋아요
3	2.52 (Neg)	정시성	메트로	시간	이용	되어	세계
4	3.34 (Pos)	외국어 지원	중국어	기차	연결	영어	아주
5	4.39 (Pos)	결제 시스템	사용	카드	쉽게	시스템	여행
솔로, 부가적 요소							
1	3.20 (Pos)	결제 시스템	영어	사용	여행	카드	방법
2	3.64 (Pos)	이용성	이용	시간	수단	교통	호선
3	3.20 (Pos)	편리성	하지	하는	출구	지하철역	매우
4	2.28 (Neg)	국가 사업	KTX	국가	생각	상당히	특히
5	2.85 (Neg)	애국심	최고	우리나라	세계	나라	한국

타당성 검증 - 전체 데이터

Type	Topic	1 score	2 score	3 score	4 score	5 score	Sentiment Score
활동 장소	1 (Pos)	-1.000	-1.000	-1.000	-0.563	0.384	2.98 (Neg)
	2 (Pos)	-1.000	-1.000	0.067	0.838	-0.367	3.17 (Pos)
	3 (Pos)	-1.000	-1.000	-0.786	-0.263	0.225	2.36 (Neg)
	4 (Neg)	6.938	1.977	1.646	0.058	-0.277	2.87 (Neg)
	5 (Pos)	-1.000	-1.000	-1.000	0.944	-0.297	1.99 (Neg)
	6 (Neg)	0.234	-0.383	-0.075	0.189	-0.074	3.89 (Pos)
	7 (Neg)	-0.479	0.825	0.014	-0.038	0.008	3.15 (Pos)
	8 (Pos)	-1.000	-0.549	-0.122	-0.123	0.082	2.96 (Neg)
여가 장소	1 (Pos)	-0.793	-0.777	-0.494	-0.123	0.246	3.39 (Pos)
	2 (Not Defined)	-0.695	0.206	0.242	0.125	-0.159	1.82 (Neg)
	3 (Pos)	-0.309	-0.318	-0.202	-0.115	0.153	3.38 (Pos)
	4 (Pos)	-1.000	-0.729	-0.609	0.040	0.144	2.36 (Neg)
	5 (Neg)	0.685	0.642	0.245	-0.011	-0.073	1.93 (Neg)
	6 (Neg)	0.741	0.375	0.297	0.098	-0.167	3.26 (Pos)
	7 (Pos)	-0.339	-0.075	-0.266	-0.097	0.148	2.23 (Neg)

타당성 검증 - 전체 데이터2

Type	Topic	1 score	2 score	3 score	4 score	5 score	Sentiment Score
2차적 요소	1 (Neg)	3.509	2.105	0.814	-0.080	-0.447	2.75 (Neg)
	2 (Not Defined)	0.411	-0.589	0.415	-0.076	-0.049	2.91 (Neg)
	3 (Pos)	-0.695	-0.378	-0.394	0.027	0.159	3.74 (Pos)
	4 (Pos)	-0.669	-0.566	-0.131	0.008	0.095	3.47 (Pos)
	5 (Not Defined)	-0.212	0.224	0.225	-0.012	-0.081	2.85 (Neg)
	6 (Pos)	-0.621	-0.559	-0.166	0.029	0.088	3.44 (Pos)
부가적 요소	1 (Not Defined)	0.393	-0.335	0.486	0.028	-0.066	3.26 (Pos)
	2 (Pos)	-1.000	-1.000	-1.000	0.470	-0.093	2.23 (Neg)
	3 (Neg)	-0.084	0.360	-0.370	-0.005	0.040	3.16 (Pos)
	4 (Pos)	-1.000	-1.000	-0.281	-0.160	0.119	1.43 (Neg)
	5 (Not Defined)	-1.000	3.543	0.369	0.143	-0.131	3.77 (Pos)

타당성 검증 - 비(非) 솔로

Type	Topic	1 score	2 score	3 score	4 score	5 score	Sentiment Score
활동 장소	1 (Pos)	-0.66	-0.69	-0.27	-0.05	0.07	4.37 (Pos)
	2 (Neg)	-0.27	0.35	0.23	0.23	-0.14	2.28 (Neg)
	3 (Neg)	1.65	0.83	0.23	-0.10	0.00	4.15 (Pos)
	4 (Neg)	-1.00	1.97	0.93	0.00	-0.13	3.01 (Not Defined)
여가 장소	1 (Pos)	-0.01	0.006	-0.03	0.009	0.00	3.57 (Pos)
	2 (Pos)	-0.30	-0.04	-0.18	-0.07	0.10	2.50 (Neg)
	3 (Neg)	0.43	0.04	0.30	0.08	-0.13	2.53 (Neg)
2차적 요소	1 (Neg)	1.08	0.97	0.58	-0.03	-0.26	3.38 (Pos)
	2 (Pos)	-0.39	-0.15	-0.16	0.03	0.04	3.59 (Pos)
	3 (Pos)	-0.69	-0.79	-0.41	0.00	0.20	3.93 (Pos)
부가적 요소	1 (Neg)	-1.00	8.07	2.70	-0.15	-0.29	2.72 (Neg)
	2 (Not Defined)	0.80	-1.00	1.09	0.13	-0.15	3.08 (Not Defined)
	3 (Not Defined)	-1.00	-0.19	0.77	-0.24	0.03	2.52 (Neg)
	4 (Not Defined)	-1.00	-1.00	0.85	-0.29	0.06	3.34 (Pos)
	5 (Neg)	0.05	0.08	-0.26	0.02	0.02	4.39 (Pos)

타당성 검증 - 솔로

Type	Topic	1 score	2 score	3 score	4 score	5 score	Sentiment Score
활동 장소	1 (Pos)	0.00	-0.33	-0.21	-0.14	0.10	4.37 (Pos)
	2 (Not Defined)	0.00	-1.00	0.67	0.36	-0.22	2.28 (Neg)
	3 (Neg)	0.00	3.12	0.37	0.04	-0.12	4.15 (Pos)
	4 (Pos)	0.00	-1.00	-0.38	0.03	0.05	3.01 (Not Defined)
여가 장소	1 (Pos)	-1.00	0.82	-0.12	0.29	-0.20	3.60 (Pos)
	2 (Pos)	-0.24	-0.29	0.01	-0.06	0.05	3.65 (Pos)
	3 (Neg)	1.08	0.09	0.05	-0.05	0.01	3.50 (Pos)
2차적 요소	1 (Neg)	2.48	1.98	0.89	-0.18	-0.45	2.71 (Neg)
	2 (Pos)	-0.15	-0.30	-0.07	-0.01	0.06	3.53 (Pos)
	3 (Pos)	-0.62	0.08	-0.18	0.11	0.00	3.65 (Pos)
부가적 요소	1 (Pos)	-1.00	-1.00	-0.74	-0.05	0.09	3.20 (Pos)
	2 (Neg)	1.11	0.05	0.59	0.19	-0.13	3.64 (Pos)
	3 (Neg)	-1.00	3.71	0.14	-0.37	0.12	3.20 (Pos)
	4 (Pos)	-1.00	-1.00	-1.00	-1.00	0.52	2.28 (Neg)
	5 (Pos)	-1.00	-1.00	-0.59	-0.47	0.26	2.85 (Neg)

결과 정리 - 전체 데이터

긍정 주제

자식 교육(활), 고궁 풍경(여),
도심 공원(여), 쇼핑 지역(2차),
전통 기념품(2차), 화장품(2차)

부정 주제

사진(활), 가이드 투어(여),
외국인 쇼핑(2차)

Topic number		Sentiment score		
		Positive (Over 3point)	Not Defined	Negative (Under 3point)
Frequency ratio	Positive 1, 2점(-) 4, 5점(+)	Group 1 활동 장소: 2 여가 장소: 1, 3 2차적 요소 : 3, 4, 6		Group 2 활동 장소: 1, 3, 5, 8 여가장소: 4, 7 부가적 요소: 2, 4
	Not Defined (Irregular)	Group 3 부가적 요소 : 1. 5		Group 4 여가장소: 2 2차적 요소: 2, 5
	Negative 1, 2점(+) 4, 5점(-)	Group 5 활동장소 : 6, 7 여가장소 : 6 부가적 요소 : 3		Group 6 활동장소: 4 여가장소: 5 2차적 요소: 1

결과 정리 - 집단 구분

긍정 주제

1) 비(非) 솔로
입장료(활), 고궁 투어(여),
기념품(2차), 다양한 가게(2차)

2) 솔로
시점/시간(활), 시점/시간(여), 산책(여),
번화가(2차), 인기 관광지(2차), 결제 시스템(부가)

부정 주제

1) 비(非) 솔로
가족(활), 공원(여), 시설(부가)

2) 솔로
외국인 관광객(2차)

Topic number		Sentiment score		
		Positive (Over 3 point)	Not Defined	Negative (Under 3 point)
Frequency ratio	Positive 1, 2점(-) 4, 5점(+)	Group 1 활동장소: 1*, 1 여가장소: 1*, 1, 2 2차적 요소: 2*, 3*, 2, 3 부가적 요소: 1	Group 2 활동장소: 4	Group 3 여가장소: 2* 부가적 요소: 4, 5
	Not Defined (Irregular)	Group 4 활동장소: 3 부가적 요소: 4*	Group 5 부가적 요소: 2*	Group 6 활동장소: 2 부가적 요소: 3*
	Negative 1, 2점(+) 4, 5점(-)	Group 7 활동장소: 3* 여가장소: 3 2차적 요소: 1* 부가적 요소: 5*, 2, 3	Group 8 활동장소: 4*	Group 9 활동장소: 2* 여가장소: 3* 2차적 요소: 1 부가적 요소: 1*

결과 정리

긍정 주제

1) 전체 : 자식 교육(활), 고궁 풍경(여), 도심 공원(여), 쇼핑 지역(2차), 전통 기념품(2차), 화장품(2차)
2) 비(非) 솔로 : 입장료(활), 고궁 투어(여), 기념품(2차), 다양한 가게(2차)
3) 솔로 : 시점/시간(활), 시점/시간(여), 산책(여), 번화가(2차), 인기 관광지(2차), 결제 시스템(부가)

부정 주제

1) 전체 : 사진(활), 가이드 투어(여), 외국인 쇼핑(2차)
2) 비(非) 솔로 : 가족(활), 공원(여), 시설(부가)
3) 솔로 : 외국인 관광객(2차)

- ✓ 공통 : 깔끔히 조성된 관광지에 산책 및 방문 선호 및 쇼핑 지역 또는 상품에 대하여 긍정적 반응.
- ✓ 비(非)솔로 : 활동 장소에 대한 '입장료' 에 대해 긍정적으로 생각하지만, '가족'의 존재가 관람에 부정적 영향 확인 가능.
- ✓ 비(非)솔로 : 부가적 요소에서 시설이 만족스럽지 못함을 추측 가능.
- ✓ 솔로 : '시점/시간'과 같은 시간적 요소가 1차적 요소에서 공통적으로 긍정 요소임.
- ✓ 공통, 솔로 : 2차적 요소에서 외국인과 관련한 부정적 영향 확인 가능.

시사점 및 한계점, 향후 과제

1. 시사점

✓ 한국어 감성사전 구축 전, 리뷰 데이터와 평점 데이터를 이용하여 감성 분석이 가능함을 확인할 수 있음.

→ 관광지 뿐만 아니라 호텔, 여행사, 항공사, 외식업계 등 다양한 분야에서 속성을 추출하여 긍정, 부정을 확인할 수 있음.

✓ 관광지 유형 및 동반자 유무에 따른 속성을 추출하였으므로 관광지 개발 계획 또는 관리에 있어 방향성을 제시할 수 있음.

→ 특히, 집단을 구분한 속성 또한 추출하였기 때문에 여행 상품 개발에 있어 무엇을 우선순위로 어떤 방향을 추구해야 하는지 알 수 있음.

2. 한계점 및 향후 과제

✓ 오역으로 인한 한국어 데이터는 분석에 쓰여지지 않았기 때문에 데이터 손실 발생

→ 네이버 플레이스, SNS 등에서 덜 손상된 한국어 데이터의 확보가 필요

✓ 원천 데이터를 확보한 출처에 따라 분석을 진행

→ 블로그, 카페와 같은 웹에서의 리뷰와 Instagram, Facebook과 같은 SNS에서 추출되는 속성은 다를 것으로 예상됨.

출처

국립국어원 (2020). 『말뭉치 감성 분석 및 연구』.

김수연. 정유경 & 송민 (2015). 한글 감성 분류를 위한 감성 사전 구축에 관한 실험적 연구. 『KLISS 2015 Proceedings of the Winter International Conference』, 143-150.

김지연. 조우용. 최정혜 & 정혜림 (2016). 온라인상의 기업 및 소비자 텍스트 분석과 이를 활용한 온라인 매출 증진 전략.『한국경영과학회지』, 41(2), 81-100.

남승주. 이현철 (2019). LDA 토픽 모델링을 활용한 항공승객 유형 별 특성 분석.『경영과학』, 36(3), 81-100.

박경열. 안희자(2019). 텍스트 마이닝을 활용한 DMZ관광 이슈의 토픽 모델링 분석. 『관광레저연구』, 31(4), 143-159.

박영욱. 정규엽 (2021). DMR(Dirichlet Multinomial Regression) 토픽모델링을 이용한 온라인 리뷰 빅데이터 기반 고객감성 분석에 관한 연구 : 국내 5성급 호텔의 외국인 이용객 리뷰를 중심으로. 『호텔경영학연구』, 30(2), 1-20.

박은정. 조성준 (2014). KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지.『제 26회 한글 및 한국어 정보처리 학술대회』, 133-138.

심준식. 김형중 (2017). LDA 토픽 모델링을 활용한 판례 검색 및 분류 방법. 『전자공학회논문지』, 54(9), 67-75.

이현주 (2017). 빅데이터를 활용한 경복궁 방문 경험 분석. 『관광연구』, 32(2), 297-318.

임영희. 김홍범 (2019). 호텔 온라인 리뷰 빅데이터를 활용한 감성분석에 관한 연구. 『호텔경영학연구』, 28(7), 105-123.

조민경. 이병주 (2021). 토픽모델링을 통한 국내 대형항공사들의 서비스품질 비교:트립어드바이저 리뷰를 중심으로. 『호텔관광연구』, 23(1), 152-165.

조수민 (2020). 관광 홍보의 뉴 미디어 트렌드.『한국관광정책』, 80(1), 109-113.

최자영. 김현아 & 김용범 (2020). 온라인 리뷰가 매출에 미치는 영향력 분석:텍스트기반 감성지수를 중심으로. 『유통연구』, 25(3), 1-21.

편집부 (2020). 빅데이터 마케팅 전쟁. 『마케팅』, 54(10), 16-27.

한국관광 데이터랩 (2021). 『한국관광 데이터랩 브리프』.

홍태호. 니우한잉. 임강 & 박지영 (2018). LDA를 이용한 온라인 리뷰의 다중 토픽별 감성분석- TripAdvisor 사례를 중심으로. 『정보시스템연구』, 27(1), 89-110.

D. Blei, A. Y. Ng & M. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Newman, D., Lau, J. H., Grieser, K. & Baldwin, T. (2010). Automatic evaluation of topic coherence. *HLT10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* 100-108.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41(6), 391-407.

Jansen-Verbake, M. (2008). Inner city tourism, resources, tourists, and promoters. *Annals of Tourism Research,* 13(1), 79-100.