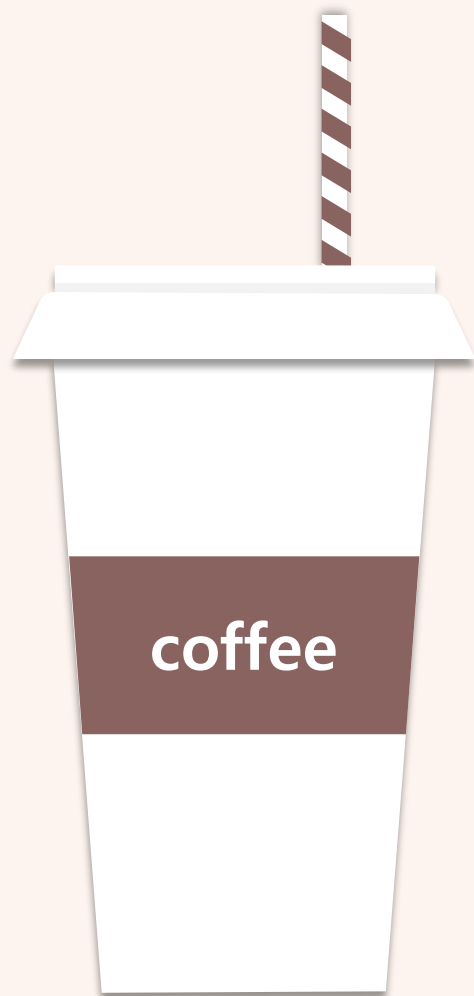


카페 리뷰 텍스트 분석



〈이성규〉

Contents



1 서론

2 데이터 수집

3 전처리

4 단어 빈도

5 감성 분석

6 주제 분석

7 결론

1. 서론

참고) 한국경제신문

<대한민국 커피점 90%, 7만5520곳이 '동네 카페'>

<동네 카페 13% 폐업했는데...스벅 메가커피 뭐가 다르길래>

- 현황 및 문제점


전국 카페 수는 2020년 7월을 기준 8만 3692개로, 이 중 90%인 7만 5520개가 개인 카페다. 그러나 매출은 [프랜차이즈 37.8% : 비프랜차이즈 62.2%]로, 매장 수 대비 매출의 '부익부 빈익빈'이 존재한다. 더군다나 전국에서 가장 카페가 많은 서울 (1만 8535개)에서만 최근 3년간 카페가 22% 늘었고, 주요 프랜차이즈 카페의 서울 내 매장 수가 13.6% 많아지는 데 비해, 개인 카페는 23.4%나 늘어나며 개인 카페의 매출 경쟁은 더 치열해졌다.

또한, 주요 프랜차이즈 카페의 폐업률이 1~2% 내외인 것에 비해, 개인 카페의 폐업률은 계속 상승해 작년 대비 올해 개인 카페 폐업률은 12.8%로, 10개의 개인 카페 중 1개가 문을 닫았다. 이런 배경에서 프랜차이즈와 달리 딱히 매뉴얼이 없는 개인 카페들이 장기적인 카페 운영에 있어서 어떤 점을 고려하면 좋을 지 고객들의 리뷰를 통한 텍스트분석으로 알아보려고 한다.

- 왜 '리뷰'인가?

검색포털에 '카페'를 검색하면 정확도순과 인기도순으로 카페들을 상위페이지에 나열해주는데, 이때 리뷰가 우선순위 반영에 큰 영향을 미친다. 상위페이지에 노출되는 것이 잠재적 고객들을 확보하는데 상대적으로 유리하기 때문에, 카페 노출 목적에서 리뷰의 중요성을 볼 수 있다. 그리고 '카페 리뷰'를 검색하면 카페리뷰와 관련된 마케팅 컨설팅 회사들의 광고가 나열되는데, 이는 카페의 리뷰가 카페 운영에 중요하다는 것을 시사한다. 마케팅 컨설팅 회사들이 검색포털에 광고료를 내면서까지 카페 리뷰에 대한 컨설팅을 광고하는 것은 그만큼 리뷰에 대해 다수의 카페들이 신경을 쓰고 있다는 것이다.

2. 데이터 수집

사이트	웹 스크랩 방법	대상 지역	데이터 수	카페 수
 kakaomap	Selenium	서울특별시 + 6대광역시 + 강릉, 남양주, 파주	40,730개	약 4,000개

- Why use the kakaomap ?

카페리뷰를 수집할 수 있는 사이트는 카카오맵을 제외하고도 망고플레이트, 트립어드바이저, 네이버 플레이스 등으로 다양하다. 그러나 이들은 평균 평점이 5점 만점에 최소 4점 이상인 카페들을 보여주기 때문에, 데이터 수집시 리뷰의 긍정·부정의 비율을 어느정도 유지할 수 없기 때문에, **긍정·부정의 리뷰와 평점이 이들**에 비해 상대적으로 다양한 카카오맵을 웹 스크랩 사이트로 선정하였다.

- How use the WebScrap?

카카오맵은 동적 웹 스크랩을 해야하므로 Selenium 을 사용해서 '카페 목록을 가져오고 해당 페이지 카페들의 데이터를 모두 수집하면 페이지를 넘겨주는 **브라우저1**'과 '상세보기의 리뷰와 평점을 페이지를 넘기며 수집하는 **브라우저2**'를 구성하였다. 각 브라우저마다 다른 페이지징 코드와 크롤링 함수를 만들어서 적용하였고, 검색어마다 첫 페이지와 페이지징이 다르기 때문에, 첫 페이지 확인 후 필요시 코드를 수정해서 데이터를 수집하였다. cf) 구분을 위해 크롤링 지역마다 ipynb를 생성.

3. 전처리

1

중복행 제거

크롤링 과정에서
발생한 중복된 행
제거.

- 4732개

2

같은 리뷰 제거

‘최고예요’, ‘좋아요’,
‘보통이에요’, ‘별로예요’
자동화된 중복 리뷰 제거.

- 4551개

3

한글 아닌 리뷰 제거

분석에 불필요한
특수문자나 영어인
리뷰 제거.

- 8개

4

결측값 제거

앞선 전처리에서
처리되지 못한
결측값 제거.

- 1개

5

형태소 분석

토큰화를 하기 전에, ‘아메리카노’와 ‘라떼’, ‘맛집’ 등 리뷰에 따라 형태소가 다르게 분석되는 단어들을 txt파일에 품사와 점수를 넣어 저장하고, `kiwi.load_user_dictionary`를 이용하여 사용자 사전을 추가하는 전처리를 함. 이어서 분석에 불필요한 의존명사를 제외한 명사를 추출하는 형태소 함수(tokenizer)를 키위(Kiwi)를 사용하여 만듦.

4. 단어빈도 - (1) 과정 설명

- 단어문서행렬(TDM) (Korean_stopwords + Tokenizer)

한글 불용어를 처리하기 위해 **한국어 사용자 불용어사전**을 다운받아, 카페 리뷰 분석에서 불필요한 단어들을 사전에 추가하여 CountVectorizer에 **stop_words** 파라미터를 적용하였다.

그리고 앞서 전처리 과정에서 만든 (사용자사전이 추가되고 의존명사를 제외한 명사를 추출해주는) 형태소 분석 함수인 extract_keywords를 **tokenizer** 파라미터에 적용하여 **단어문서행렬(TDM)**을 만들었다.

- 단어빈도 (데이터프레임 + WordCloud + StyleCloud)

만든 TDM을 통해 단어들의 빈도를 계산하여 데이터프레임으로 나타내었다. 이를 빈도를 기준으로 내림차순으로 정렬해서 가장 많이 나온 10개의 단어를 통해 어떤 리뷰가 주를 이루는지 알아보고 키워드들을 크게 3가지 분류로 나누어 분석하였다. 이어서 전체적으로 어떤 키워드들이 많이 나오는지 시각적으로 확인하기 위해 **StyleCloud**를 사용해서 **커피잔 모양**으로 표현하였다.

4. 단어빈도 - (2) 단어구름 & 해석 및 설명

- 단어빈도

순위	많이 나온 단어
1	맛
2	커피
3	친절
4	카페
5	분위기
6	직원
7	음료
8	가격
9	빵
10	사람

- 단어구름 (StyleCloud)



분홍 동그라미를 친 단어들은 전체적으로 카페의 **메뉴와 맛**과 관련된 주제에서 높은 빈도로 언급되었을 것으로 보인다.

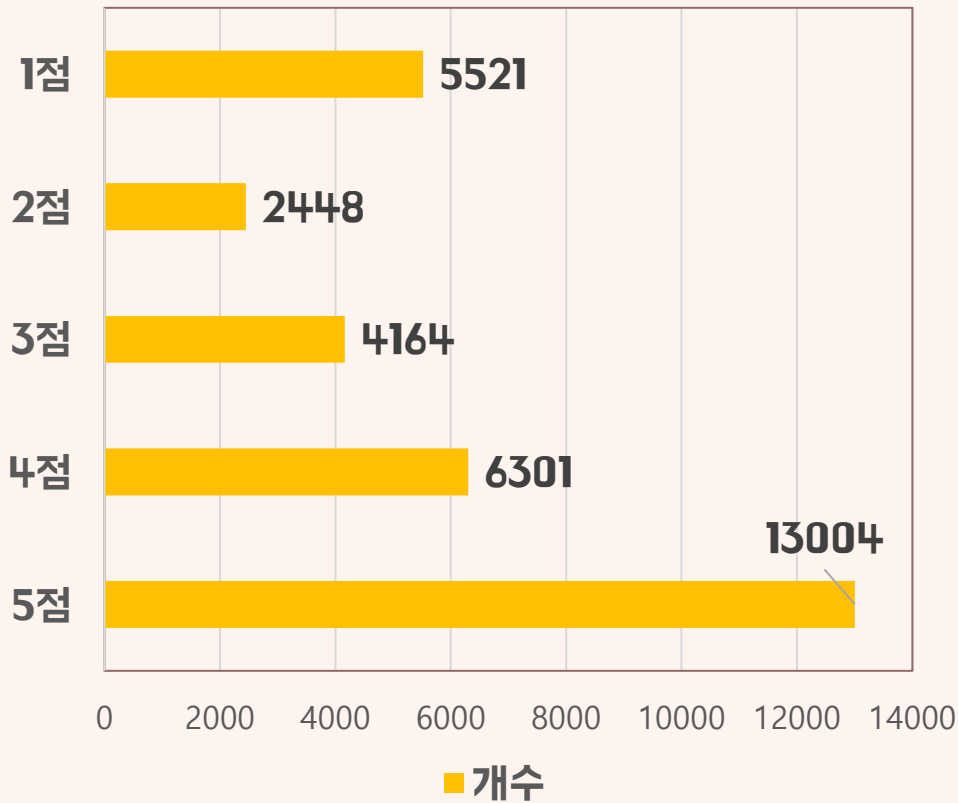
주황 동그라미를 친 단어들은 전체적으로 카페 **근무자의 태도**와 관련된 주제에서 높은 빈도로 언급되었을 것으로 보인다.

초록 동그라미를 친 단어들은 전체적으로 카페의 **공간적인 분위기**와 관련된 주제에서 높은 빈도로 언급되었을 것으로 보인다.

5. 감성분석 - (1) 라벨링

- 라벨링

카페 평점 별 개수



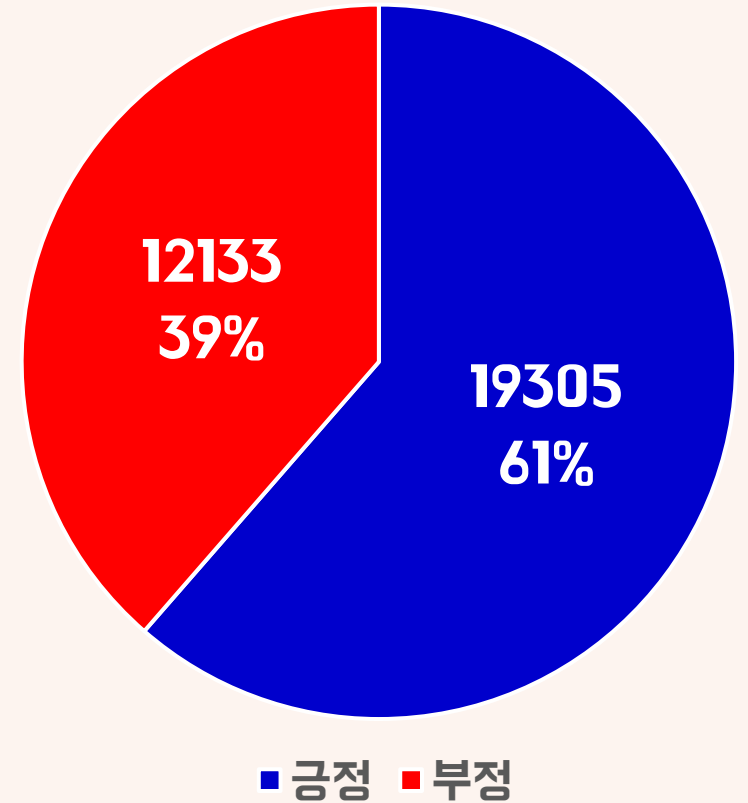
평점 3점 리뷰는
긍정적인 리뷰보다 부정적인 리뷰가
대부분이므로 0으로 라벨링.

Labeling

1,2,3점 -> 0

4,5점 -> 1

라벨링 후 긍정·부정



5. 감성분석 - (2) 과정 설명

- 감성분석 과정

- ① 만들어 둔 문서단어행렬(TDM)을 X로, 리뷰 평점을 라벨링한 값(0, 1)을 Y로 대입한다.
- ② sklearn 패키지의 train_test_split을 이용해 X와 Y를 각각 훈련데이터와 평가데이터로 나눈다.
- ③ 훈련데이터는 70%, 평가데이터는 30%로 설정하였다.
- ④ Keras 모델에 시그모이드(로지스틱 함수)를 적용하고 최적화 방법으로 adam을 사용하였다.
- ⑤ **EarlyStopping**을 설정하여 학습데이터에서 10%를 분할해서 만든 검증데이터로 val_accuracy가 더 이상 개선되지 않는 지점까지 학습을 반복하였다.
- ⑥ 모델을 학습시키고 평가해보니 **약 73%의 성능**이 나왔다.
- ⑦ 학습시킨 모델에서 얻은 파라미터로 단어에 가중치를 주어 긍정적인 단어와 부정적인 단어를 나타냈다.
- ⑧ **긍정적인 단어와 부정적인 단어를 분석·해석**하고 이를 시각적으로 한눈에 확인하기 위해서 WordCloud를 사용하여 **단어구름**으로 표현하였다.

5. 감성분석 - (3) 긍정 리뷰

- 긍정 리뷰(상위 20개) 분석·해석

순위	긍정적인 단어	순위	긍정적인 단어
1	최고	11	완벽
2	존맛탱	12	굿
3	행복	13	인생
4	감사	14	감동
5	만족	15	드라이브
6	번창	16	환상
7	짬	17	맛집
8	힐링	18	정성
9	대박	19	강
10	최애	20	예술

순위	많이 나온 단어
1	맛
2	커피
3	친절
4	카페
5	분위기
6	직원
7	음료
8	가격
9	빵
10	사람

대부분의 긍정적인 리뷰들은 카페의 전반적인 **긍정적 평가**와 관련 있었다.

전반적인 긍정적 평가는 ‘최고, 행복, 만족, 번창, 짬’ 등에서 볼 수 있었고 평가에 대한 원인으로서는 주로 **메뉴와 맛** (존맛탱, 환상, 맛집, 예술)과 **공간적 모습**(드라이브, 강)에서 볼 수 있었다.

긍정적인 리뷰를 통해 고객들이 카페 선정에 있어서 **메뉴와 맛, 좋은 경치**에 대해 초점을 둔다는 것을 유추했다.

5. 감성분석 - (4) 부정 리뷰

- 부정 리뷰(상위 20개) 분석·해석

순위	부정적인 단어	순위	부정적인 단어
1	최악	11	말투
2	실망	12	장사
3	엉망	13	공지
4	교육	14	손님
5	인스타	15	벌레
6	싸가지	16	정신
7	시장통	17	태도
8	편의점	18	알바
9	돈	19	냄새
10	별로	20	안감

순위	많이 나온 단어
1	맛
2	커피
3	친절
4	카페
5	분위기
6	직원
7	음료
8	가격
9	빵
10	사람

대부분의 부정적인 리뷰들은 카페의 전반적인 **부정적 평가**와 관련 있었다.

전반적인 부정적 평가는 대체로 ‘최악, 실망, 별로, 안감’ 등에서 볼 수 있었고 평가에 대한 원인으로는 주로 **근무자의 태도**(교육, 싸가지, 말투, 태도, 손님, 알바)와 **공간적 분위기**(시장통, 벌레, 냄새)에서 볼 수 있었다.

부정적인 리뷰를 통해 고객들이 카페 선정에 있어서 근무자의 **친절한 태도와 차분한 분위기, 청결**에 대해 초점을 둔다는 것을 유추했다.

5. 감성분석 - (5) 긍정부정 리뷰 단어구름

- 긍정 단어구름



- 부정 단어구름



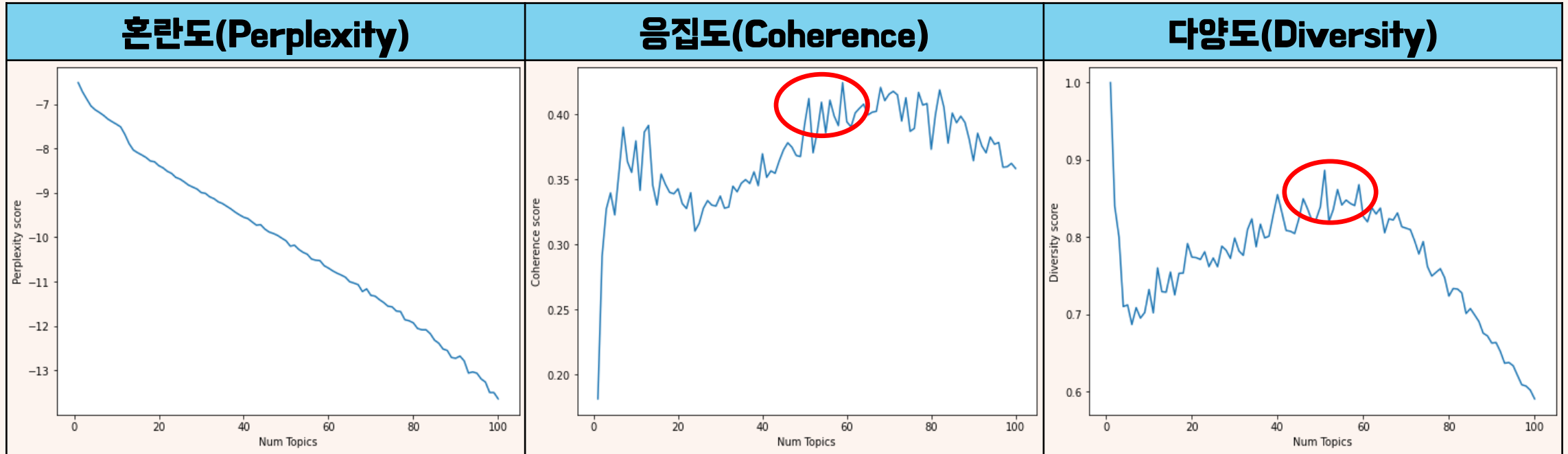
6. 주제분석 - (1) Latent Dirichlet Allocation

- LDA 주제분석

- ① TfidfVectorizer에 앞서 만든 stop_words와 tokenizer 파라미터를 적용해서 TDM을 만들었다.
- ② Tfidf를 적합해서 만든 words를 사전형태로 id2token을 만들었다.
- ③ review를 한글 불용어 사전을 추가한 extract_keywords를 적용해 토큰화한 docs를 만들었다.
- ④ 토큰형식의 docs를 doc2bow를 적용해 corpus 형식으로 만들었다.
- ⑤ 최적의 topic 수를 찾기위해서 num_topics를 1~100까지 LDA 모델에 반복해서 적용하여 각 num_topics 별 혼란도, 응집도, 다양도를 구하고 시각화함.
- ⑥ (응집도, 다양도)와 (혼란도, 응집도) 두가지 경우를 고려하여 최적의 Topic 수를 구했다.
- ⑦ 최종 LDA 모델을 pyLDAvis로 시각화하고 주제를 분석하였다.
- ⑧ 분석한 주제의 결과를 해석하였다.

6. 주제분석 - (2) 응집도와 다양도

- 응집도와 다양도를 고려하여 최적의 Topic 수 찾기

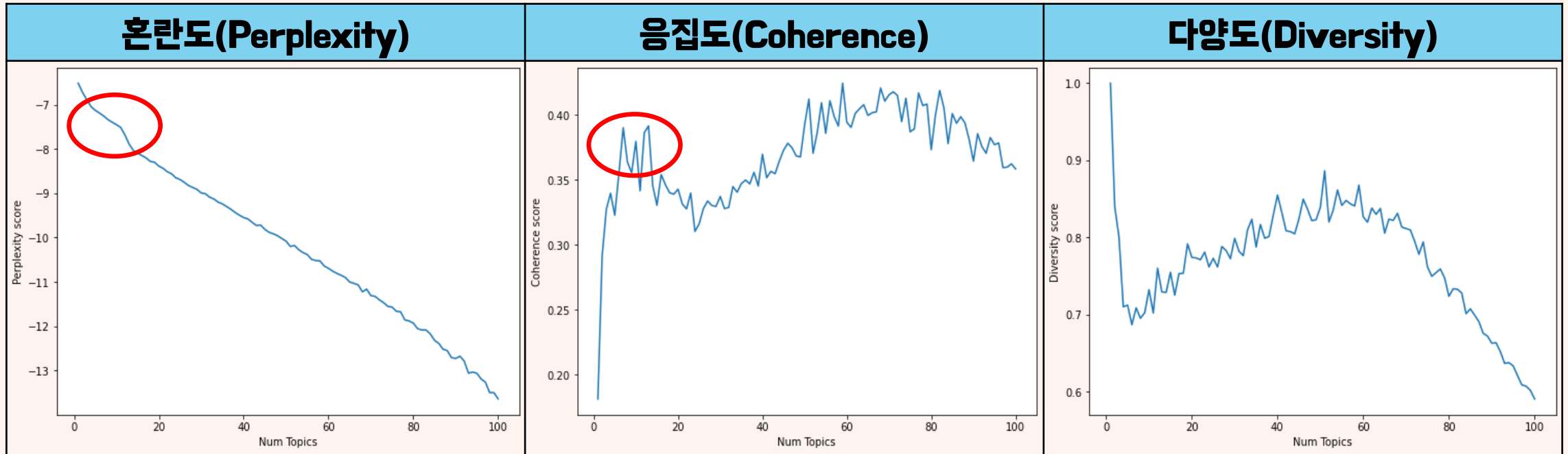


응집도 기준 최적의 Topic 수는 59개
다양도 기준 최적의 Topic 수는 51개
두 Topic 수의 응집도, 다양도 차이가 미미하여
혼란도까지 고려했을 때의 **최적의 Topic 수는 59개**

Topic의 수 \ 지표	혼란도	응집도	다양도
	Topic의 수		
51개	-10.200	0.412	0.886
59개	-10.642	0.424	0.867

6. 주제분석 - (3) 혼란도와 응집도


- 혼란도와 응집도를 고려하여 최적의 Topic 수 찾기



혼란도는 Topics 수가 많을수록 감소하기 때문에 Topic 수 대비 높은 응집도를 가지는 **최적의 Topic 수는 13개**이며, 앞서 선정한 59개에 비해 분석이 용이할 것이라 판단해 **최종 Topic 수를 13개로** 선정했다.

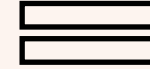
지표 Topic의 수	혼란도	응집도	다양도
	13개	59개	
13개	-7.889	0.391	0.729
59개	-10.642	0.424	0.867

6. 주제분석 - (4) pyLDAvis를 통한 주제 분석

Topic	1번 (5.7%)	2번 (6.6%)	3번 (7.5%)	4번 (13.1%)	5번 (8.7%)	6번 (7.7%)	7번 (8.5%)
주제	티 카페	음식점 카페	베이커리 카페	커피 맛	카페 근무자의 친절한 태도	불만족스러운 카페 구성품	좌석 부족 & 소음
키워드	차 버터 잔 독특 브라우니 바게트 초콜릿 무화과 오렌지 코코넛	공간 주차 음식 메뉴 피자 파스타 맥주 풍경 루프탑 창 시그니처 고급	케이크 크림 와플 아이스크림 샌드위치 마들렌 라떼 아메리카노	커피 맛 원두 향 산미 특별 냄새 분위기 드립 잔 아메리카노	친절 사장 직원 설명 방문 의사 청결 제공	주문 천 물 컵 테이크아웃 수준 짜증 사이즈 기본 별로 보통 값	자리 불편 테이블 의자 좌석 실내 사람 소리 가족
Topic	8번 (5.5%)	9번 (5.4%)	10번 (9.8%)	11번 (6.4%)	12번 (9.0%)	13번 (6.2%)	
주제	다양한 음료 & 제과	불만족스러운 카페 시스템	웨이팅	카페 근무자의 불친절한 태도	동네 인스타 감성 카페	긍정적 평가는 주로 맛 관련	
키워드	밀크티 얼그레이 바닐라라떼 말차 베이글 팬케이크 입맛 맛 취향저격	기분 이용 응대 말투 후기 전반 문제 계산 안내 와이파이 장사	사람 줄 관리 오픈 밖 문 대기 눈치 마감	알바 직원 서비스 최악 기대 부족 카운터 태도 진심 싸가지 교육 전화	빵 가격 스콘 디저트 동네 인스타 감성 대비 카페 집	최고 맛집 이유 추천 인생 진짜 가성비 인정 커피 맛 신맛 단맛	

6. 주제분석 - (5) 분석 결과 해석

1번 (5.7%)	2번 (6.6%)	3번 (7.5%)	12번 (9.0%)
티 카페	음식점 카페	베이커리 카페	동네 인스타 감성 카페



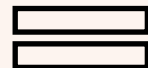
메뉴에 따라 카페의 종류가 세분화되고 있으며,
특히 2번과 12번 주제는 **공간적 분위기**의
키워드가 많이 언급되며 더욱 세분화되었다.

4번 (13.1%)	8번 (5.5%)	13번 (6.2%)
커피 맛	다양한 음료 & 제과	긍정적 평가는 주로 맛 관련



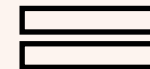
다양한 메뉴와 커피 맛에 대한 주제들로,
특히 4번 주제는 13개의 주제들 중 가장 비중이 크기 때문에
'커피 맛'에 대한 키워드들이 전체 리뷰에서 비중이 크며,
13번 주제와 감성분석의 긍정적인 키워드를 보면,
리뷰의 긍정적인 평가는 주로 커피 맛과 관련되어 보인다.

5번 (8.7%)	11번 (6.4%)
카페 근무자의 친절한 태도	카페 근무자의 불친절한 태도



근무자의 태도에 대한 주제들로, 친절·불친절에 따라 감성분석에서의
긍정·부정 키워드를 다루고 있다. 특히 친절은 리뷰에서 크게 부각되지 않지만,
불친절은 부정적인 리뷰로 직결되기 때문에 더 신경써야 할 것으로 보인다.

6번 (7.7%)	7번 (8.5%)	9번 (5.4%)	10번 (9.8%)
불만스러운 카페 구성품	좌석 부족 & 소음	불만스러운 카페 시스템	웨이팅



고객의 불편을 도출해 **불만을 유발**하는 요인에
대한 주제들로, 부정적인 리뷰관리에 있어서
크게 신경 쓰고 참고해야 할 주제들이다.

7. 결론

- 분석결과 요약정리

단어빈도 분석에서 리뷰가 '메뉴와 맛', '근무자의 태도', '공간적 분위기' 3가지 큰 틀로 분류되었고, 3가지 키워드 분류가 감성분석과 주제분석에서도 적용되었다. 감성분석에서는 메뉴와 맛·공간적 분위기에 의해 긍정적 평가를 내포하는 키워드들이 나왔으며, 근무자의 태도·공간적 분위기에 의해 부정적 평가를 내포하는 키워드들이 나왔다. 이를 통해 고객들이 카페 선정에 있어서 메뉴와 맛, 좋은 경치, 친절한 태도, 차분한 분위기, 청결 등에 초점을 두고 있다는 것을 알 수 있었다. 따라서 카페 운영에 있어서 긍정적인 주제는 부각하고, 부정적인 주제는 해결하거나 발생하지 않게끔 계속해서 신경을 기울여야 한다.

- 문제 해결 방안

매뉴얼이 없는 개인 카페들은 주제분석의 첫번째 해석을 참고하여 주메뉴를 정하고 공간적 분위기를 고려해 카페 종류를 세분화한다. 또한 두번째 해석을 볼 때, 커피 맛이 리뷰에서 비중이 크고 다른 주제들에 비해 긍정적인 리뷰 도출에 크게 관련 있기 때문에 커피 맛을 신경써야 한다. 세번째 해석과 함께 감성분석을 참고하면 근무자의 친절은 크게 부각되지 않지만 불친절은 부정적인 주제와 리뷰 도출에 직결되므로, 불친절에 초점을 두고 근무자를 지속적으로 교육하고 모니터링해서 불필요한 부정적 리뷰 생성을 방지해야 한다. 마지막으로 네번째 해석을 보면, 카페의 구성품, 운영시스템, 좌석부족, 소음, 웨이팅은 긍정적인 키워드에서 부각되지 않지만, 위 주제들이 고객들의 기대에 비해 부족하다면 고객들의 불만을 유발하여 불친절과 마찬가지로 부정적 리뷰 도출에 직결되므로, 이와 관련된 리뷰가 나온다면 문제가 되는 주제를 바로 해결해야 한다.



The END

