

team project :

카페 리뷰 분석

19101967 강채원
17101951 김혜인
19101984 문대정



목차

01 연구 배경 / 목적

02 데이터 수집

03 데이터 분석 / 해석

01

연구 배경/목적

1.1 연구 배경

AI 탑재한 네이버 '맛집 추천'...하루 85만명 찾는다

오대석 기자 > 홍성용 기자 >

입력 2020/08/25 17:42 | 수정 2020/08/25 17:42

0

개인 선호도·취향 적극 반영
네이버 "광고는 100% 배제"
카카오도 AI 알고리즘 추가
SK·망고플레이트등 경쟁 가열

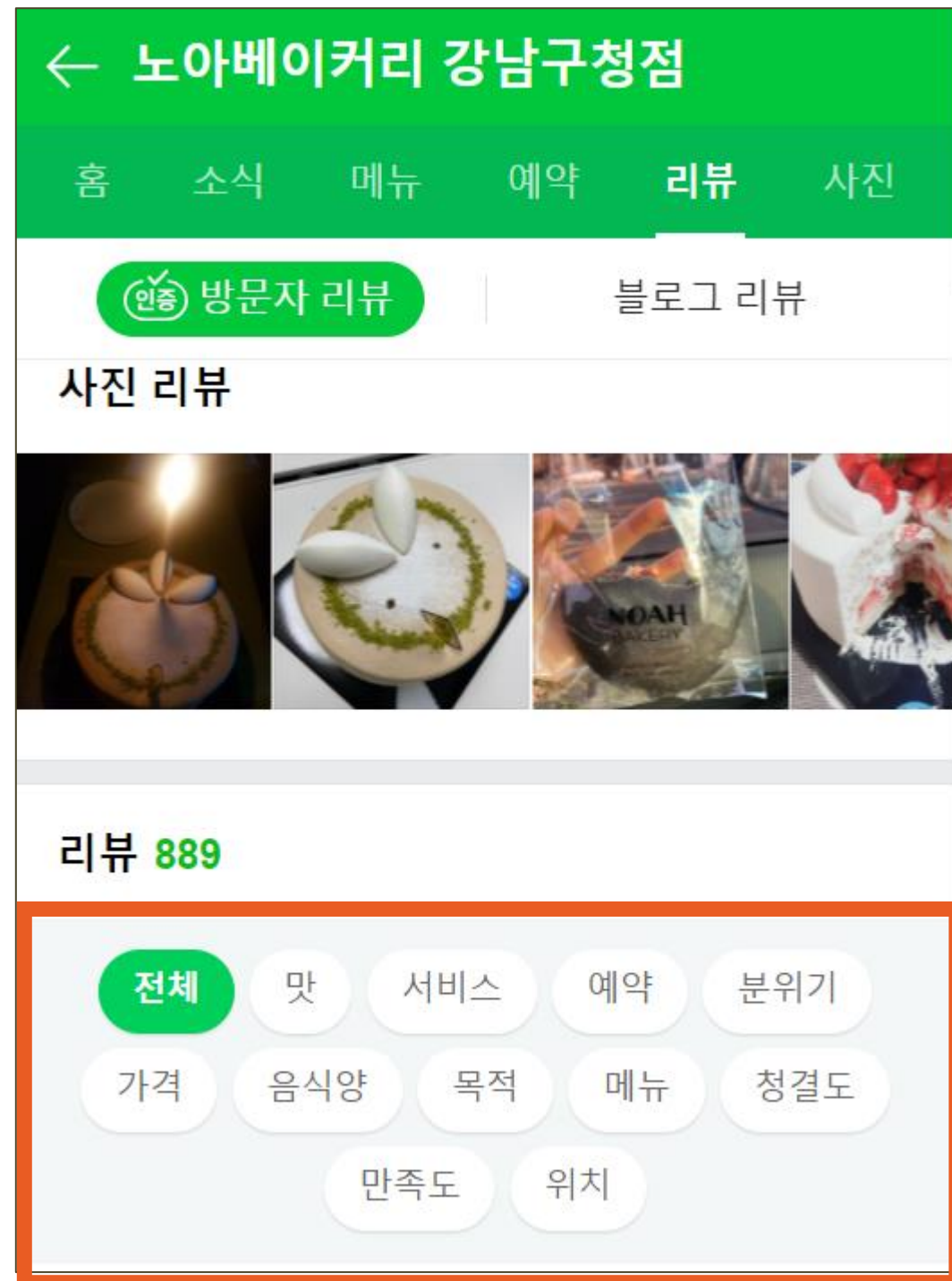
네이버 영수증 첨부 맛집 리뷰, 10개월 만에 1억
건 돌파

하루 평균 리뷰 작성수 40만 건

백봉삼 기자 | 입력 :2020/08/28 08:44 -- 수정: 2020/08/28 09:20 | 인터넷

리뷰의 영향력 증가

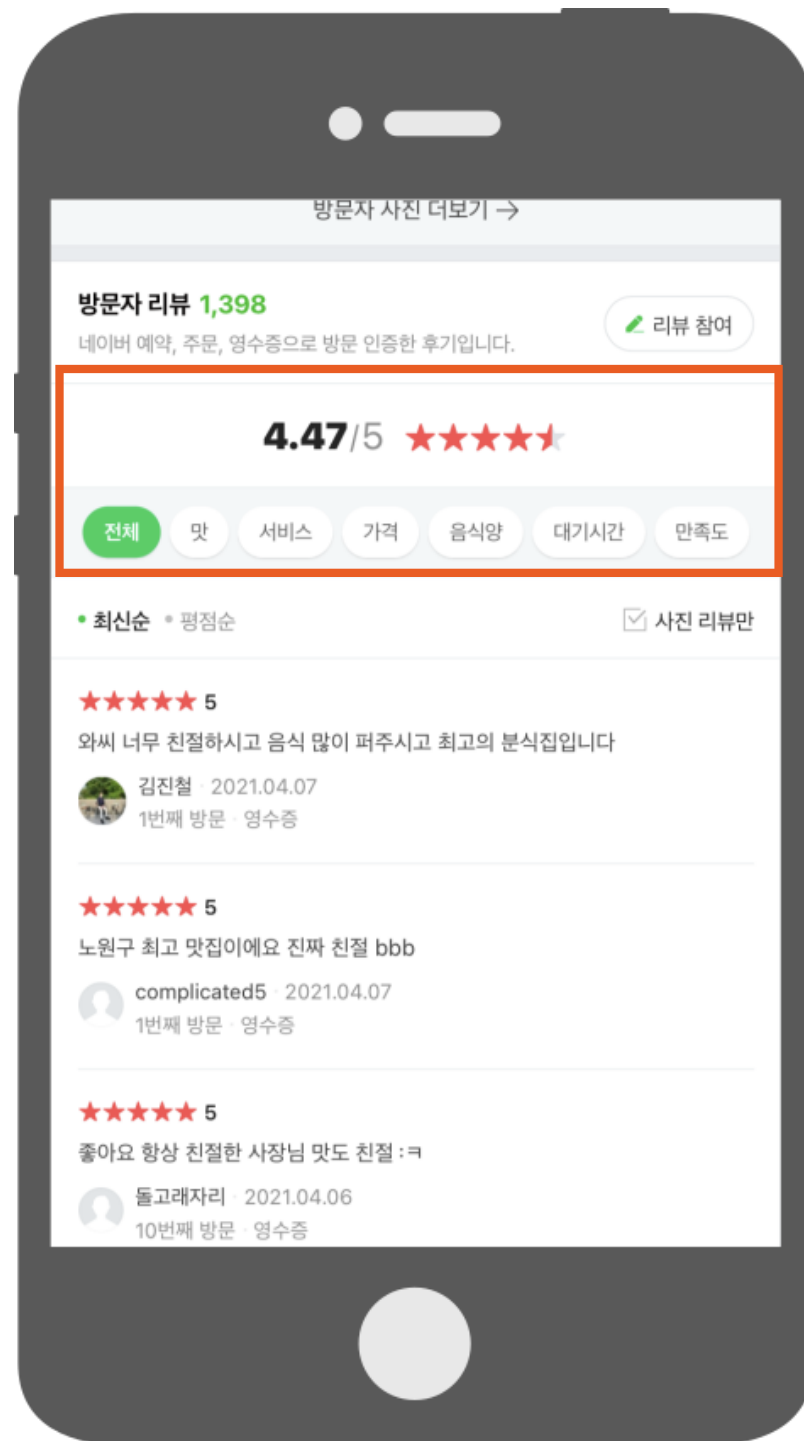
1.1 연구 배경



다양한 기준의 리뷰가 존재



1.1 연구 배경



종합 평점과 분야별 평가의
불일치 가능성

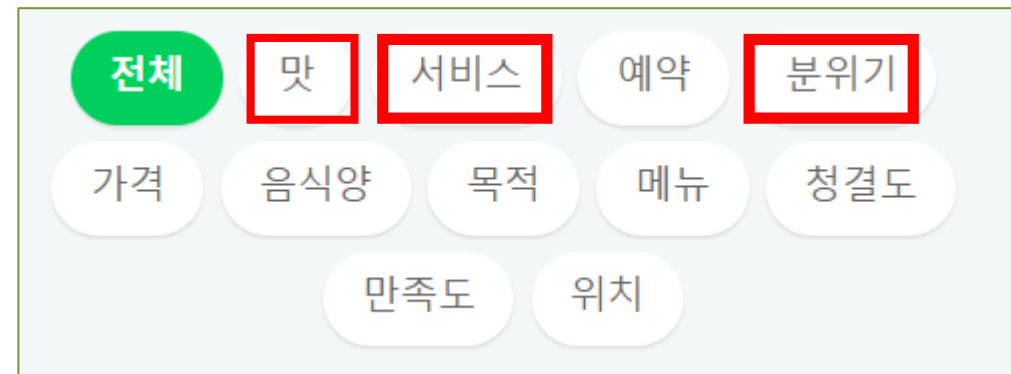
+

방대한 양의 리뷰

"키워드"로 분야와 긍정/부정을 분류하여
키워드에 맞는 리뷰를 보기 편하도록
분리하면 어떨까?



1.2 연구 목적



NaiveBayes

카페를 선정할 때 가장 중요하게 생각하는
맛, 서비스, 분위기 세가지 요소로 리뷰 분류

+

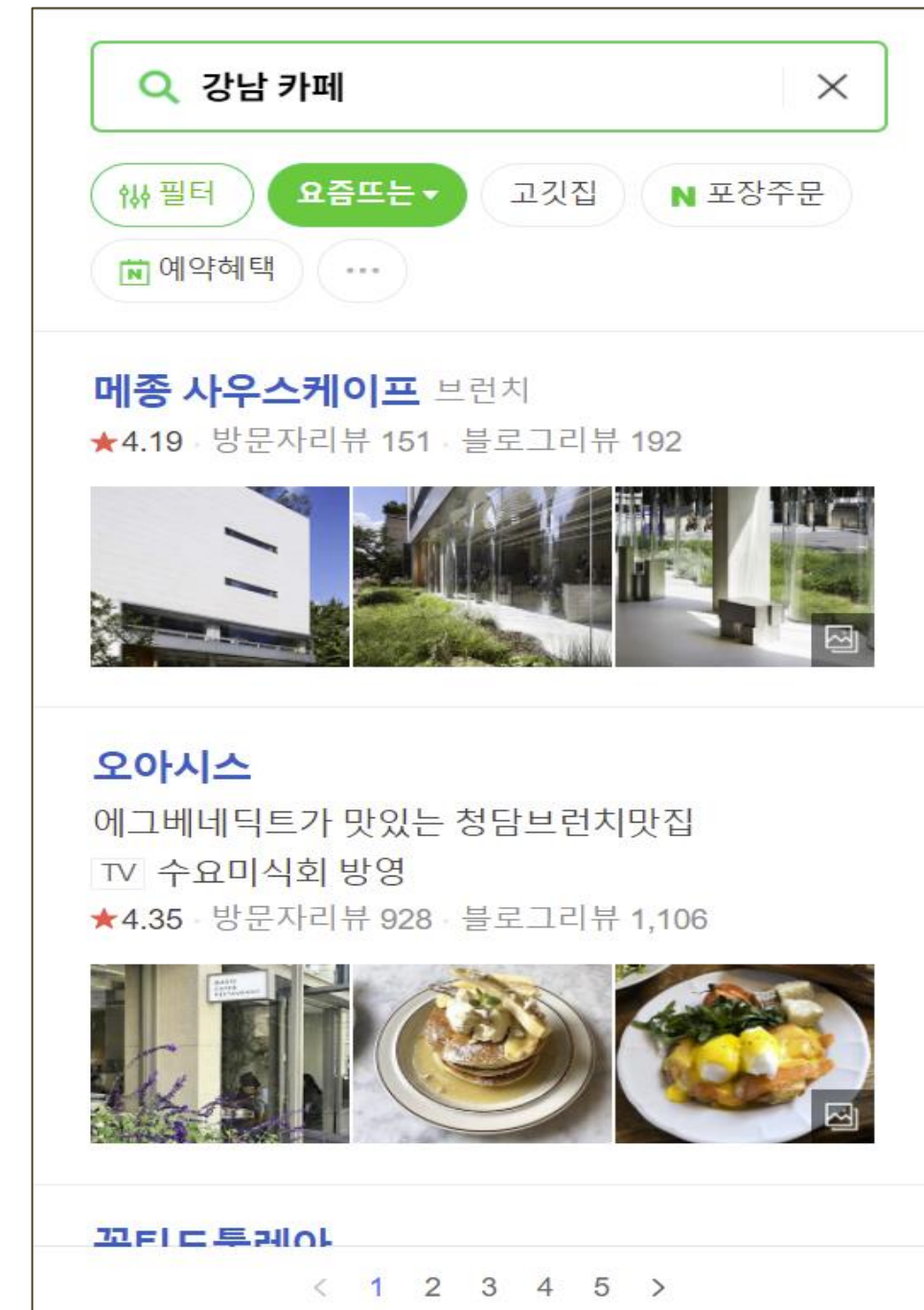
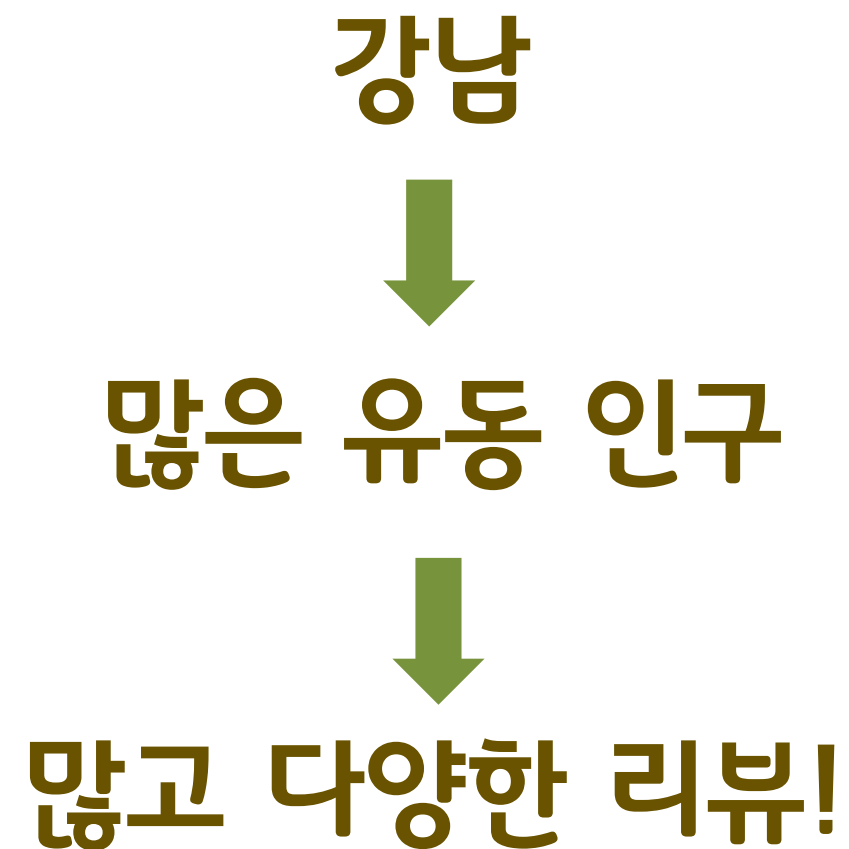
NaiveBayes를 이용하여
각 요소별 긍정/부정 리뷰 시각화!



02 데이터 수집

2.1 데이터 수집 범위

리뷰 수집 범위를 강남의 카페로 설정.



Source : <https://map.naver.com/>

2.1 데이터 수집 범위

서울 열린데이터 광장에서 강남의 카페 리스트를 가져옴.



서울 열린데이터 광장
SEOUL OPEN DATA PLAZA

| 개방자 | 관리번호 | 인허가일 | 인허가처 | 영업상태 | 영업상태 | 상세영 | 상세영 | 폐업일자 | 휴업일자 | 휴업종류 | 재개업일 | 전화번호 | 소재지 | 소재지 | 지번주소 | 도로명 | 도로명 | 사업장명 | 최종수정 | 데이터 | 데이터 | 업태구분 |
|---------|-----------|----------|------|------|-------|-----|-----|------|------|------|------|-----------|--------|--------|-------|-------------------------------|--------------|------------------|----------|---------|---------|------|
| 3220000 | 3220000-1 | 20011110 | | 1 | 영업/정상 | 1 | 영업 | | | | | 2.3E+08 | 261 | 135839 | 서울특별시 | 서울특별시 | 6192 | 스타벅스 선릉 세화빌딩점 | 2.02E+13 | U | 40:00.0 | 커피숍 |
| 3220000 | 3220000-1 | 20060928 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 723606 | 99.64 | 135190 | 서울특별시 | 서울특별시 | 6376 | 커피와쟁이 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20070125 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 555.84 | 135897 | 서울특별시 | 서울특별시 | 6018 | (주)커피빈코리아압구정로데오점 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20070911 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 34540 | 39.67 | 135910 | 서울특별시 | 서울특별시 | 6131 | 이디야 역삼역점 | 2.02E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20090417 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 48.99 | 135841 | 서울특별시 | 서울특별시 | 강남구 역포마카페강남점 | 2.02E+13 | I | 59:59.0 | 커피숍 | |
| 3220000 | 3220000-1 | 20100210 | | 1 | 영업/정상 | 1 | 영업 | | | | | 25480552 | 271.24 | 135090 | 서울특별시 | 강남구 삼성동 159-(주)커피빈코리아 도심공항타워점 | 2.02E+13 | I | 59:59.0 | 커피숍 | | |
| 3220000 | 3220000-1 | 20100302 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 48.51 | 135859 | 서울특별시 | 서울특별시 | 6269 | 카페따아모 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100308 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 75.75 | 135860 | 서울특별시 | 서울특별시 | 6259 | 에이름커피 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100413 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 121.67 | 135820 | 서울특별시 | 서울특별시 | 6056 | 더쇼 | 2.02E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100422 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 129.65 | 135923 | 서울특별시 | 서울특별시 | 6235 | 해머스미스커피 역삼태헤란로점 | 2.02E+13 | U | 01:00.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100426 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 518 22 | 14.08 | 135935 | 서울특별시 | 서울특별시 | 6240 | 사르르커피 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100506 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 260 | 135815 | 서울특별시 | 서울특별시 | 6046 | 제이스커피나인 학동점 | 2.02E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100511 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 758 83 | 343.08 | 135860 | 서울특별시 | 서울특별시 | 6258 | 스타벅스 뱅뱅사거리 | 2.02E+13 | U | 40:00.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100601 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 43 | 135996 | 서울특별시 | 서울특별시 | 6105 | 베스트빈 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100812 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 22.16 | 135884 | 서울특별시 | 서울특별시 | 6349 | 카페 더 단골 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100827 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 556 84 | 186.94 | 135936 | 서울특별시 | 서울특별시 | 6243 | 파스쿠찌(강남태극당) | 2.02E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100827 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 34569 | 15.4 | 135502 | 서울특별시 | 서울특별시 | 6174 | 카페쉴루 | 2.02E+13 | U | 07:00.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100830 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 548 45 | 43 | 135900 | 서울특별시 | 서울특별시 | 6001 | 소호앤노호 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100830 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 30151 | 347.54 | 135954 | 서울특별시 | 서울특별시 | 6014 | 스타벅스 압구정로데오역 | 2.02E+13 | U | 40:00.0 | 커피숍 |
| 3220000 | 3220000-1 | 20101020 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 4.5 | 135951 | 서울특별시 | 서울특별시 | 6065 | 비전케어 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20101104 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 33 | 135943 | 서울특별시 | 서울특별시 | 6338 | 커피날에 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20101112 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 567 39 | 59.4 | 135927 | 서울특별시 | 서울특별시 | 6219 | 모모웨이 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20101129 | | 1 | 영업/정상 | 1 | 영업 | | | | | 7.09E+10 | 33.15 | 135848 | 서울특별시 | 서울특별시 | 6186 | 재단법인 아름다운커피 세정점 | 2.02E+13 | U | 40:00.0 | 커피숍 |
| 3220000 | 3220000-1 | 20101130 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 501 38 | 47.9 | 135842 | 서울특별시 | 서울특별시 | 6190 | 커피보보르 | 2.02E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20100916 | | 1 | 영업/정상 | 1 | 영업 | | | | | 02 443 16 | 43.93 | 135240 | 서울특별시 | 서울특별시 | 6322 | ABOUT 책과 머피 | 2.01E+13 | I | 59:59.0 | 커피숍 |
| 3220000 | 3220000-1 | 20111201 | | 1 | 영업/정상 | 1 | 영업 | | | | | | 12.5 | 135957 | 서울특별시 | 서울특별시 | 6075 | (주)아워홈 린컴퍼니점카페 | 2.02E+13 | U | 40:00.0 | 커피숍 |

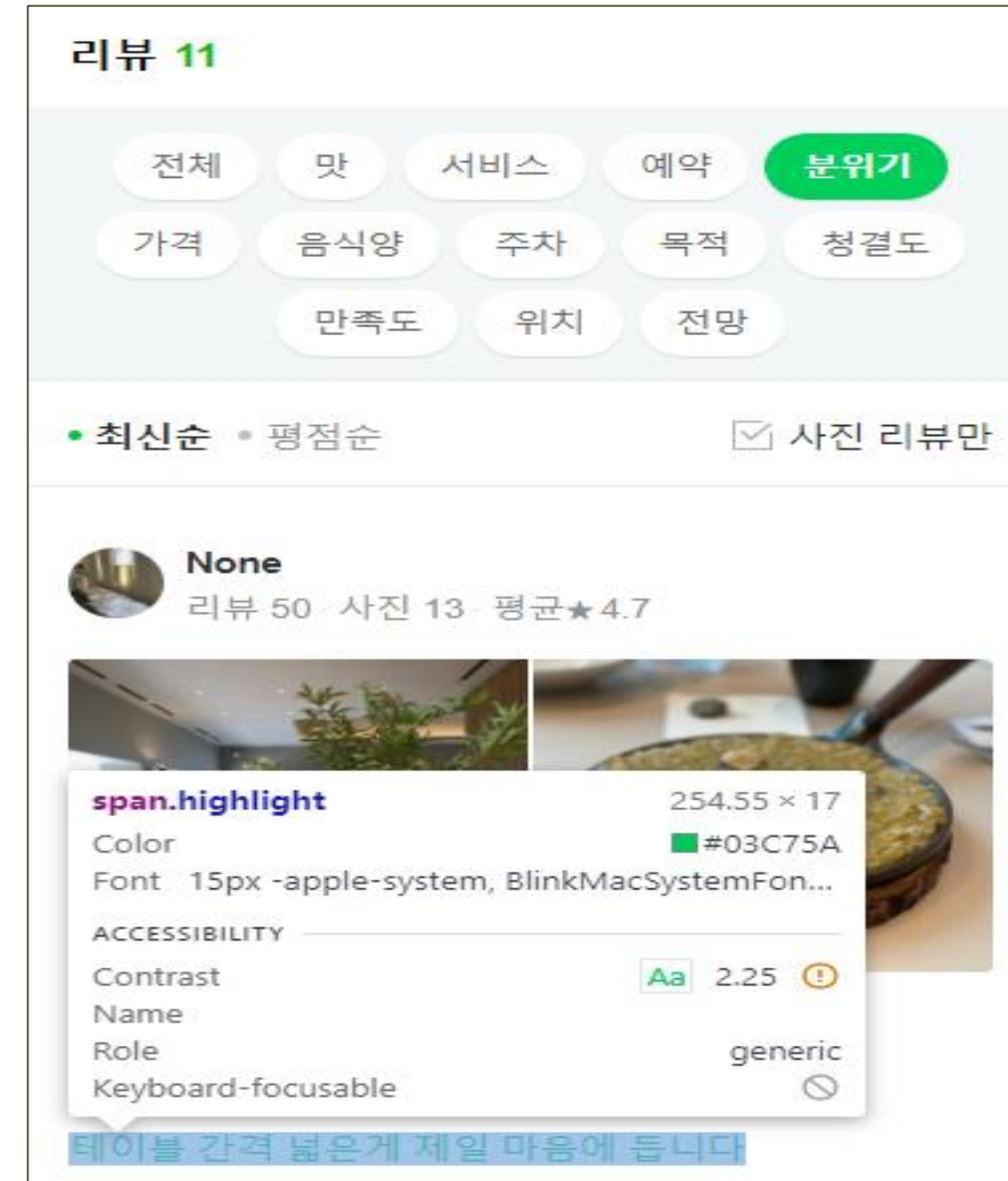
Source : 서울 열린데이터 광장(<http://data.seoul.go.kr/dataList/datasetList>)

2.2 데이터 크롤링

카테고리별 리뷰 일부분이
Highlight tag로 표시됨



원하는 카테고리별로
구분하여 크롤링이 가능



2.2 데이터 크롤링

전체 **맛** 서비스 예약 분위기

가격 음식양 주차 배달 목적

메뉴 청결도 대기시간 만족도


위치

최신순 • 평점순

☒ 사진 리뷰만

이다금

리뷰 11 · 사진 1 · 평균★4.6



★★★★★ 5

2021.05.05 | 3번째 방문 | 영수증

3번째 방문입니다! 오늘은 어린이날이라고 풍선이랑
곰돌이 꽃이(?)도 주네요 ㅎㅎ **항상 맛있게 먹고갑니**
다

전체 맛 서비스 예약 **분위기**

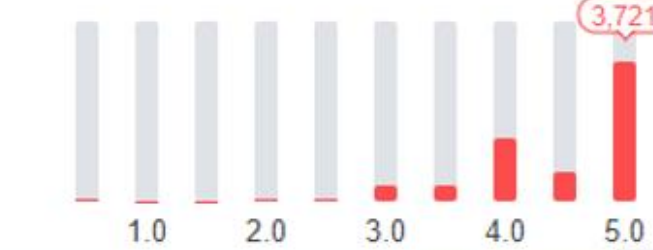
가격 음식양 주차 배달 목적


메뉴 청결도 대기시간 만족도

위치

★★★★★ 4.42/5

7,245개 (6,056명 참여)





★★★★★ 0.5

2021.04.24 | 1번째 방문 | 영수증

아 정말 테이크아웃 줄 너무 이상한 방식이에요 **그 앞**
마당 너무 좁은데 50명이 다닥다닥 붙어서 코로나 걸
릴것 같아요. 거기다 같이 주문한 동행은 도... ▼

전체 맛 **서비스** 예약 분위기

가격 음식양 주차 목적 청결도



만족도 위치 전망

최신순 • 평점순

☒ 사진 리뷰만

cka0826

리뷰 32 · 사진 43 · 평균★3.8



★★★★★ 4

2021.05.03 | 1번째 방문 | 영수증

맛, 분위기, 서비스별 highlight tag 크롤링

* 수집된 데이터

| store_name | aspect | star | review | | | | | | |
|------------|--------|------|----------------------------------|--|--|--|--|--|--|
| 스타벅스커피 | mood | 0.5 | 스벅 직원들 분위기가 영 아닙니다 | | | | | | |
| 토이서 | taste | 0.5 | 밀크티. | | | | | | |
| 투썸플레이스 | taste | 0.5 | 분리돼있던 케 이크를 당연스럽게 꺼내길래 새로운걸로 달라고 | | | | | | |
| 투썸플레이스 | taste | 2 | 밀크티프라페가 안돼서 | | | | | | |
| 공차 대치점 | taste | 2 | 밀크티에 밀크폼을 시키고 | | | | | | |
| 스타벅스커피 | taste | 2 | 쉬림프&해초샐러드 너무 비리고 | | | | | | |
| 탐앤탐스커피 | taste | 2 | 맛이없네 | | | | | | |
| 밀크다방(부산점) | taste | 2 | 커피가 부드러워요 | | | | | | |
| 스타벅스커피 | taste | 2.5 | 커피맛은 좋아요 | | | | | | |
| 파스쿠찌커피 | mood | 3 | 분위기 좋은 커피숍 | | | | | | |
| 써브웨이커피 | taste | 3 | 서브웨이는 어딜가도 다 똑같아서 맛은 있지만. | | | | | | |
| 투썸플레이스 | mood | 3 | 에어컨이 고장났는지 별로 안 시원하네요 | | | | | | |
| 카페 파스쿠찌 | taste | 3 | 커피맛은 쏘쏘 | | | | | | |
| 제네럴 스타벅스 | taste | 3 | 커피가 독특하니 맛있어요 | | | | | | |
| 탐앤탐스커피 | taste | 3 | 맛있어요 | | | | | | |
| 셀렉토커피 | taste | 3 | 커피가 쓴맛이 심해요 | | | | | | |
| 스타벅스커피 | mood | 3 | 분위기 좋네요 | | | | | | |
| 빌리엔젤커피 | taste | 3 | 맛있어요! | | | | | | |
| 더조이(TH) | taste | 3 | 구슬아이스크림이 맛있어요 | | | | | | |
| 이디야역점 | taste | 3.5 | 과자 짭 | | | | | | |
| 탐앤탐스커피 | taste | 3.5 | • 강 믹스커피 맛이네요 | | | | | | |
| 스타벅스커피 | mood | 3.5 | 분위기 좋아요 | | | | | | |
| 더조이(TH) | taste | 3.5 | 아이스크림도 츠러스도 좋아어요 | | | | | | |

NaiveBayes
(지도학습)



데이터에 **긍/부정 라벨** 필요



**라벨 생성을 위한
긍/부정 사전의 필요성**

2.3 사전 : Text vs 형태소

```
text = '안녕하세요. 오래간만이네요~~. 어제 재미있었어요.'
print(oka.morphs(text))    # stem, default False
```

[제목 없음]

```
> ['안녕하세요', '.', '오래간만', '이네요', '~~.', '어제', '재미있었어요', '.']
```

VS

```
text = '안녕하세요. 오래간만이네요~~. 어제 재미있었어요.'
print(oka.pos(text))
> [('안녕하세요', 'Adjective'), ('.', 'Punctuation'), ('오래간만', 'Adverb'), ('이네요', 'Verb'),
  ('~~.', 'Punctuation'), ('어제', 'Noun'), ('재미있었어요', 'Adjective'), ('.', 'Punctuation')]

print(oka.pos(text, join=True))
> ['안녕하세요/Adjective', './Punctuation', '오래간만/Adverb', '이네요/Verb', '~~./Punctuation',
  '어제/Noun', '재미있었어요/Adjective', './Punctuation']
```

2.3 사전 : Text vs 형태소

* 실제 리뷰 예시

맛있어요!
쉬폰과 크림이 엄청부드럽고 맛있었어요.
케익도 너무 맛있어요~

- TEXT 로 사전 구축할 경우

[엄청부드럽고] 를 사전에 추가함.

=> [부드럽다], [매우부드럽다] 등 중복된
단어들을 또 사전에 추가해야 함

- 형태소로 사전 구축할 경우

[엄청], [부드럽다] 로 분리

=> 사전을 효율적, 효과적으로 작성 가능

2.4 형태소 분석

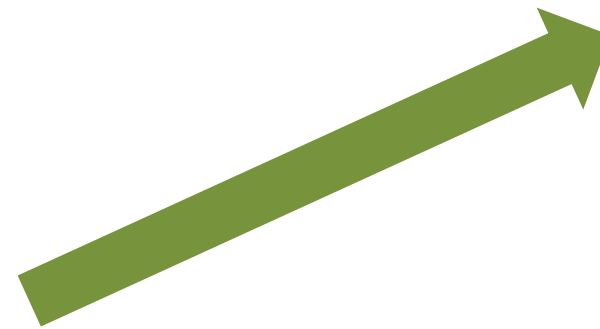
| | store_name | aspect | star | review | processed | processed_review_v2 | | | | | | | |
|------|------------|--------|------|---------|------------------------|--|--|--|--|--|--|--|--|
| 0 | 스타벅스커피 | mood | 0.5 | 스벅 직원들 | ['스벅', '직원', '들'] | ['스벅', '직원', '들', '분위기', '가', '영', '아니다'] | | | | | | | |
| 1 | 토이셔 | taste | 0.5 | 밀크티. | ['밀크', '티'] | ['밀크', '티', '.'] | | | | | | | |
| 2 | 투스셈플레이 | taste | 0.5 | 분리돼있던 | ['분리', '돼'] | ['분리', '돼다', '케', '이크', '를', '당', '연', '스럽게', '꺼내다', '새롭다', '달라', '고'] | | | | | | | |
| 3 | 투스셈플레이 | taste | 2 | 밀크티프리 | ['밀크', '티', '프리'] | ['밀크', '티', '프리', '페', '가', '안', '돼다'] | | | | | | | |
| 4 | 공차 대치소 | taste | 2 | 밀크티에 | ['밀크', '티'] | ['밀크', '티', '에', '밀크', '폼', '을', '시키다'] | | | | | | | |
| 5 | 스타벅스커피 | taste | 2 | 쉬림프&해물 | ['쉬', '림프', '&', '해물'] | ['쉬', '림프', '&', '해물', '샐러드', '너무', '비리', '고'] | | | | | | | |
| 6 | 탐앤탐스커피 | taste | 2 | 맛이없네 | ['맛', '이'] | ['맛', '이', '없다'] | | | | | | | |
| 7 | 밀크다방커피 | taste | 2 | 커피가 부드럽 | ['커피', '가'] | ['커피', '가', '부드럽다'] | | | | | | | |
| 8 | 스타벅스커피 | taste | 2.5 | 커피맛은 | ['커피', '맛'] | ['커피', '맛', '은', '좋다'] | | | | | | | |
| 9 | 스타벅스커피 | taste | 2.5 | 커피맛은 | ['커피', '맛'] | ['커피', '맛', '은', '좋다'] | | | | | | | |
| 10 | 스타벅스커피 | taste | 2.5 | 커피맛은 | ['커피', '맛'] | ['커피', '맛', '은', '좋다'] | | | | | | | |
| 11 | 파스쿠찌커피 | mood | 3 | 분위기 좋음 | ['분위기', ''] | ['분위기', '좋다', '커피숍'] | | | | | | | |
| 12 | 써브웨이커피 | taste | 3 | 서브웨이는 | ['서브웨이'] | ['서브웨이', '는', '어딜', '가도', '다', '똑같다', '맛', '은', '있다', '.'] | | | | | | | |
| 13 | 투스셈플레이 | mood | 3 | 에어컨이 | ['에어컨', ''] | ['에어컨', '이', '고장', '나다', '별로', '안', '시원하다'] | | | | | | | |
| 14 | 카페 파스쿠찌 | taste | 3 | 커피맛은 | ['커피', '맛'] | ['커피', '맛', '은', '쏘다', '쏘다'] | | | | | | | |
| 15 | 제네럴 스타벅스 | taste | 3 | 커피가 독특 | ['커피', '가'] | ['커피', '가', '독특하다', '맛있다'] | | | | | | | |
| 16 | 탐앤탐스커피 | taste | 3 | 맛있어요 | ['맛있어요'] | ['맛있다'] | | | | | | | |
| 17 | 셀렉토커피 | taste | 3 | 커피가 쓴 | ['커피', '가'] | ['커피', '가', '쓴맛', '아', '심해', '요'] | | | | | | | |
| 4517 | 스타벅스커피 | mood | 5 | 매장 분위기 | ['매장', '분'] | ['매장', '분위기', '깔끔하다'] | | | | | | | |
| 4518 | 파스쿠찌커피 | taste | 5 | 직원분도 | ['직원', '분'] | ['직원', '분', '도', '친절하다', '음료', '도', '맛있다', '.'] | | | | | | | |
| 4519 | 파스쿠찌커피 | taste | 5 | 커피맛 굿 | ['커피', '맛'] | ['커피', '맛', '굿'] | | | | | | | |
| 4520 | 파스쿠찌커피 | taste | 5 | 맛있어요 | ['맛있어요'] | ['맛있다'] | | | | | | | |
| 4521 | 파스쿠찌커피 | taste | 5 | 맛있어요 | ['맛있어요'] | ['맛있다'] | | | | | | | |
| 4522 | 파스쿠찌커피 | taste | 5 | 맛있었어요 | ['맛있었어요'] | ['맛있다'] | | | | | | | |
| 4523 | 투스셈플레이 | taste | 5 | 민트초콜릿 | ['민트', '초'] | ['민트', '초코', '라떼', '달', '고', '맛있다', '????'] | | | | | | | |
| 4524 | 커피랑 차 | taste | 5 | 넘 맛있어요 | ['넘', '맛있'] | ['넘다', '맛있다'] | | | | | | | |
| 4525 | 커피랑 차 | taste | 5 | 가격대비 | ['가격', '대'] | ['가격', '대비', '커피', '맛있다'] | | | | | | | |
| 4526 | 커피랑 차 | taste | 5 | 친절하고 | ['친절하고'] | ['친절하다', '맛있다', '~'] | | | | | | | |
| 4527 | 커피랑 차 | taste | 5 | 커피맛이 | ['커피', '맛'] | ['커피', '맛', '이', '좋다'] | | | | | | | |
| 4528 | 커피랑 차 | taste | 5 | 매일 향만 | ['매일', '향'] | ['매일', '향', '만', '음미', '하다', '~', '커피', '맛', '좋다', '~'] | | | | | | | |
| | store_name | aspect | star | review | processed | processed_label | | | | | | | |

➡ 약 4500개의 리뷰로
각각의 형태소 추출

2.5 긍/부정 사전

맛, 분위기 서비스별 긍/부정 사전 제작

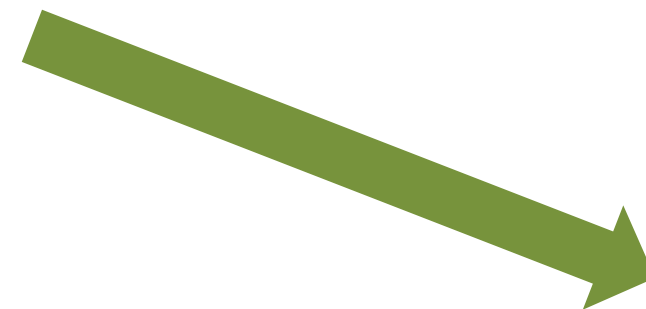
| | | |
|-------------------------------------|--|--|
| ['분위기', '도', '좋다'] | | |
| ['노래', '가', '좋다'] | | |
| ['맛있다'] | | |
| ['투썸', '커피', '맛있다'] | | |
| ['커피', '맛', '무난', '하다'] | | |
| ['레드', '벨벳', '케익', '너무', '맛', '있다'] | | |
| ['커피', '가', '맛있다'] | | |
| ['케이크', '먹다', '엇', '늘다', '맛', '있다'] | | |
| ['커피', '랑', '케이크', '맛있다'] | | |
| ['커피', '도', '맛', '나', '요'] | | |
| ['친절', '매장', '깨끗하다', '좋다'] | | |
| ['직원', '이', '친절하다'] | | |
| ['커피', '가', '맛있다'] | | |
| ['맛', '도', '좋다'] | | |
| ['맛있다'] | | |
| ['커피', '가', '저렴하다', '맛있다'] | | |
| ['저렴하다', '아메리카노', '~', '좋다'] | | |
| ['콜드', '브루', '라떼', '좋다'] | | |
| ['커피집', '중', '에', '최고'] | | |
| ['커피', '맛', '나다'] | | |



Taste



Mood



Service

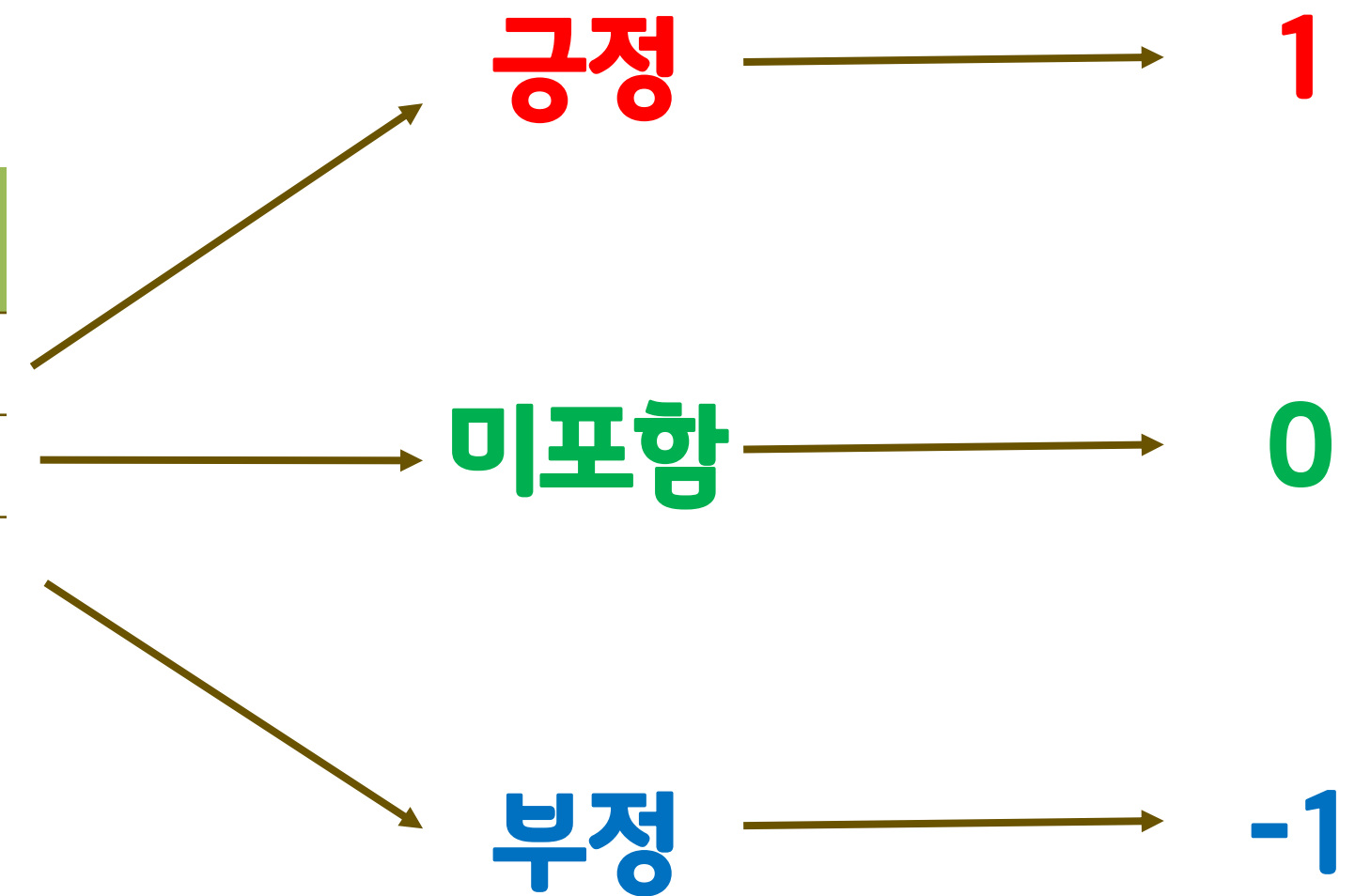
2.5 긍/부정 사전

완성된 카테고리별 긍정 부정 사전

| taste_pos | taste_neg | mood_pos | mood_neg | service_pos | service_neg | |
|-----------|-----------|----------|----------|-------------|-------------|--|
| 맛있다 | 특이하다 | 좋다 | 좁다 | 친절하다 | 그닥 | |
| 아주 | 나쁘다 | 시원하다 | 어수선 | 좋다 | 않다 | |
| 신선하다 | 않다 | 넓다 | 없다 | 빠르다 | 불친절하다 | |
| 친절하다 | 차다 | 좋다 | 아쉽다 | 친절하다 | 느리다 | |
| 따뜻하다 | 그냥 | 고급스럽다 | 덥다 | 좋다 | 불친절 | |
| 진리 | 그렇다 | 예쁘다 | 더워 | 괜찮다 | 아깝다 | |
| 존맛임 | 맛없다 | 편안하다 | 고장 | 친절 | 지저분하다 | |
| 특이하다 | 약하다 | 쾌적 | 별로 | 착하다 | 불친절하다 | |
| 짱 | 약간 | 이쁘다 | 나쁘다 | 웬만하다 | 별로 | |
| 많다 | 싱겁다 | 쾌적하다 | 시끄럽다 | 저렴하다 | 아쉽다 | |
| 신기하다 | 강하다 | 저격 | 아쉽다 | 깔끔하다 | 안좋다 | |
| 좋다 | 없다 | 조용하다 | 별로 | 최고 | 무표정 | |
| 시원하다 | 살짝 | 좋다 | 불편 | 빨리 | 불편하다 | |
| 진하다 | 별로 | 넓다 | 춥다 | 상냥하다 | 깡 | |
| 정말 | 아쉽다 | 아늑하다 | 불편하다 | 굿 | 어수선하다 | |
| 부드럽다 | 다르다 | 좋아 | 시끌벅적하다 | 웃음 | 속상하다 | |
| 이쁘다 | 맹맹하다 | 널찍 | 어수선하다 | 편하다 | | |
| 최고 | 떨어지다 | 쾌적 | 너무 | | | |
| 다양하다 | 밍밍 | 최고 | | | | |
| 굿굿 | 최악 | 깔끔 | | | | |
| 괜찮다 | | 친절 | | | | |
| 맛있다 | | 조용하다 | | | | |
| 아쉽다 | | 좋다 | | | | |

2.6 데이터 라벨링

| 원래 리뷰 | 형태소 추출 |
|--------------|-------------------------|
| 궁합이 정말 좋았습니다 | ['궁합', '이', '정말', '좋다'] |
| 맛이 오묘해여 | ['맛', '이', '오묘하다'] |
| 커피맛은 아쉬웠네요 | ['커피', '맛', '은', '아쉽다'] |



2.6 데이터 라벨링

* 라벨링 코드

```
make_label.py > ...
1  import pandas as pd
2  from konlpy.tag import Okt
3
4  SAVE_PATH = 'C:/Users/max53/(Assignment)/Datamining_class/team_project'
5  CRAWL_CSV_NAME = 'datamining_processed_review_result_ANSI_TOTAL_v2.csv'
6  DICT_NAME = 'posneg_dict_recent_rev.xlsx'
7
8  def preprocess(raw_review):
9
10     processed_review = raw_review.split(',')
11     print(processed_review)
12     return processed_review
13
14
15  def make_label(df, posneg_dict):
16
17     label = []
18     processed_review = df['processed_review_v2']
19     print(processed_review)
20     aspect_values = df['aspect'].values
21
22     for i, one_review in enumerate(processed_review):
23
24         # 이상한 샘플 직접 뜯어 보기
25         # if i == 35:
26         #     break
27
28         flag = False
29         aspect = aspect_values[i]
30
31         pos_dict = posneg_dict[f'{aspect}_pos'].values
32         neg_dict = posneg_dict[f'{aspect}_neg'].values
33
34
35         one_review = preprocess(one_review)
36
37         for j, word in enumerate(one_review):
38
39             # word 출력해보면 one_review에 list로 담겨 있는게 아니라, char로 되어버려서, ',', [ 같은 기호 슬라이싱으로 삭제
40             if j == len(one_review)-1:
41                 word = word[2:-2]
42             else:
43                 word = word[2:-1]
44             # print(word)
```

```
45
46     if word in pos_dict:
47         # print('pos')
48         flag = True
49         label.append(1)
50         break
51     elif word in neg_dict:
52         # print('neg')
53         flag = True
54         label.append(-1)
55         break
56     else:
57         # print('netural')
58         pass
59
60     if flag == False:
61         label.append(0)
62
63     return label
64
65
66  df = pd.read_csv(f'{SAVE_PATH}/{CRAWL_CSV_NAME}', encoding='ANSI')
67  posneg_dict = pd.read_excel(f'{SAVE_PATH}/{DICT_NAME}')
68
69  label = make_label(df, posneg_dict)
70  print(len(label))
71
72  # 라벨 컬럼 추가
73  df['label'] = label
74  df.to_csv(f'{SAVE_PATH}/review_label2.csv', encoding='ANSI')
```

2.6 데이터 라벨링

| | A | B | C | D | E | G | H |
|----|----|-----------------|---------|------|-----------------|--|-------|
| 1 | | store_name | aspect | star | review | processed_review_v2 | label |
| 2 | 0 | 이디야강남세곡점 | taste | 4 | 딸기쥬스 맛이 약간 약해요 | ['딸기', '쥬스', '맛', '이', '약간', '약하다'] | -1 |
| 3 | 1 | 이디야강남세곡점 | taste | 5 | 달고나라떼 맛나요 | ['달고나', '라떼', '맛', '나', '요'] | 1 |
| 4 | 2 | 이디야강남세곡점 | taste | 4 | 커피보단 셰이크가 짱 | ['커피', '보단', '셰이크', '가', '짱'] | 1 |
| 5 | 3 | 이디야강남세곡점 | mood | 4 | 내부도 시원하고 | ['내부', '도', '시원하다'] | 1 |
| 6 | 4 | 이디야강남세곡점 | mood | 4 | 이야기하기 좋아요 | ['이야기', '하다', '좋다'] | 1 |
| 7 | 5 | 이디야강남세곡점 | service | 4 | 친절하시구 좋아요~ | ['친절하다', '좋다', '~'] | 1 |
| 8 | 6 | 이디야강남세곡점 | service | 5 | 직원분이 너무 친절하세용~ | ['직원', '분', '이', '너무', '친절하다', '~'] | 1 |
| 9 | 7 | 이즈니생메르 | taste | 5 | 맛있게 잘 먹었어요 ㅎㅎ | ['맛있다', '자다', '먹다', 'ㅎ'] | 1 |
| 10 | 8 | 이즈니생메르 | taste | 5 | 에그샌드위치도 크루아상 버터 | ['에그', '샌드위치', '도', '크루아상', '버터', | 1 |
| 11 | 9 | 이즈니생메르 | taste | 5 | 빵 정말 맛있습니다 | ['빵', '정말', '맛있다'] | 1 |
| 12 | 10 | 이즈니생메르 | taste | 4.5 | 맛있어요 | ['맛있다'] | 1 |
| 13 | 11 | 이즈니생메르 | taste | 5 | 신기한 빵들이 많아서 좋았어 | ['신기하다', '빵', '들', '이', '많다', '좋다'] | 1 |
| 14 | 12 | 이즈니생메르 | service | 3 | 점원분이 너무 일하시는게 느 | ['점원', '분', '이', '너무', '일', '하다', '느리다'] | -1 |
| 15 | 13 | 이즈니생메르 | service | 5 | 사장님 친절하고 | ['사장', '님', '친절하다'] | 1 |
| 16 | 14 | 스타벅스 스타필드코엑스몰R점 | taste | 5 | 맛있어요~ | ['맛있다', '~'] | 1 |
| 17 | 15 | 스타벅스 스타필드코엑스몰R점 | taste | 4 | 맛있어요 | ['맛있다'] | 1 |
| 18 | 16 | 스타벅스 스타필드코엑스몰R점 | taste | 4 | 맛있습니다요잉 ㅎㅎ | ['맛있다', '요', '잉', 'ㅎ'] | 1 |
| 19 | 17 | 스타벅스 스타필드코엑스몰R점 | taste | 5 | 커피맛은 진리 | ['커피', '맛', '은', '진리'] | 1 |
| 20 | 18 | 스타벅스 스타필드코엑스몰R점 | taste | 5 | 아주 맛있어요 | ['아주', '맛있다'] | 1 |
| 21 | 19 | 스타벅스 스타필드코엑스몰R점 | taste | 5 | 맛있어요 | ['맛있다'] | 1 |
| 22 | 20 | 스타벅스 스타필드코엑스몰R점 | taste | 5 | 맛있어요 | ['맛있다'] | 1 |
| 23 | 21 | 스타벅스 스타필드코엑스몰R점 | taste | 5 | 잘 먹었습니다 | ['자다', '먹다'] | 1 |
| 24 | 22 | 스타벅스 스타필드코엑스몰R점 | taste | 5 | 말차라떼 맛나요 | ['말차', '라떼', '맛', '나', '요'] | 1 |
| 25 | 23 | 스타벅스 스타필드코엑스몰R점 | taste | 5 | 친절하고 음식도 따뜻하고 맛 | ['친절하다', '음식', '도', '따뜻하다', '맛있다'] | 1 |
| 26 | 24 | 스타벅스 스타필드코엑스몰R점 | mood | 5 | 분위기 좋고 | ['분위기', '좋다'] | 1 |

데이터에
라벨 생성

03 데이터 분석/해석

3.1 데이터 분석 과정

1) 필요 패키지 & 데이터 불러오기

```
In [1]: import pandas as pd
import numpy as np
import random
import os
from sklearn.naive_bayes import MultinomialNB, BernoulliNB
from sklearn.model_selection import train_test_split
from sklearn import metrics
import matplotlib.pyplot as plt
from sklearn.model_selection import StratifiedKFold
%matplotlib inline

data=pd.read_csv('C:/Users/max53/(Assignment)/Datamining_class/team_project/final_review_data_label_added.csv', encoding='ANS
```

```
In [2]: data.head()
```

Out[2]:

| | Unnamed: 0 | store_name | aspect | star | review | processed_review | processed_review_v2 | label |
|---|------------|------------|--------|------|----------------|----------------------------------|----------------------------------|-------|
| 0 | 0 | 이디야강남세곡점 | taste | 4.0 | 딸기쥬스 맛이 약간 약해요 | [딸기, '쥬스', '맛', '이', '약간', '약해요] | [딸기, '쥬스', '맛', '이', '약간', '약하다] | -1 |
| 1 | 1 | 이디야강남세곡점 | taste | 5.0 | 달고나라떼 맛나요 | [달고나, '라떼', '맛', '나', '요] | [달고나, '라떼', '맛', '나', '요] | 1 |
| 2 | 2 | 이디야강남세곡점 | taste | 4.0 | 커피보단 쉐이크가 짱 | [커피, '보단', '쉐이크', '가', '짱] | [커피, '보단', '쉐이크', '가', '짱] | 1 |
| 3 | 3 | 이디야강남세곡점 | mood | 4.0 | 내부도 시원하고 | [내부, '도', '시원하고] | [내부, '도', '시원하다] | 1 |
| 4 | 4 | 이디야강남세곡점 | mood | 4.0 | 이야기하기 좋아요 | [이야기, '하기', '좋아요] | [이야기, '하다', '좋다] | 1 |

3.1 데이터 분석 과정

2) 전처리 후 Document Term Matrix 형태로 만들기

```
In [10]: import re

stopwords_raw = '아 휴 마이구 마이쿠 마이고 어 나 우리 저희 따라 의해 을 를 에 의 가 으로 로 에게 뿐이다 익거하여 근거하여 믿
stopwords_raw2 = 'ㅇ ㅋ ㅎ 은 는 이 가 네 요 커피 남자 여자 직원 알바생 사장 매장 분위기 아메리카노 포장 빵 밀크티 자리 우유
stopwords = stopwords_raw.split(' ')
print(stopwords)
def clean_text(texts):
    corpus = []
    for i in range(0, len(texts)):
        # print(str(texts[i]))
        modi_sent = []
        review = re.sub(r'[@%&*+()/~#&#+?%xc3%xa1%-#|%.#;#;!%-#,%-#~#$%'"', '', str(texts[i])) #remove punctuation
        for word in review.split(' '):
            if word not in stopwords:
                modi_sent.append(word)
        corpus.append(' '.join(modi_sent))
    return corpus
```

[아, 휴, 마이구, 마이구, 마이고, 어, 나, 우리, 저희, 따라, 의해, 을, 를, 에, 의, 가, 으로,
로, 에게, 뿐이다, 의거하여, 근거하여, 입각하여, 기준으로, 예하면, 예를
소인, 소생, 저희, 지말고, 하지마, 하지마라, 다른, 물론, 또한, 그리고 In [21]: review_DTM
다, 뿐만, 아니라, 만이, 아니다, 만은, 아니다, 막론하고, 관계없이, 그치지
지만, 등간에, 논하지, 았다, 따지지, 았다, 설사, 비록, 더라도, 아니면, 만
다, 물론하고, 향하여, 향해서, 향하다, 쪽으로, 틈타, 이용하여, 타다, 오르
미, 밖에, 하여야, 비로소, 한다면, 몰라도, 외에도, 이곳, 여기, 부터, 기
다, 하려고하다, 미리하여, 그리하여, 그렇게, 함으로써, 하지만, 일때, 할때,
으로써, 로써, 까지, 해야한다, 일것이다, 반드시, 할줄알다, 할수있다, 할수있
면, 등, 등을, 제, 겨우, 단지, 다만, 할뿐, 덩동, 엉그, 대해서, 대하여,
마만큼, 얼마큼, 남짓, 여, 얼마간, 약간, 다소, 조금, 다수, 몇, 얼마
러나, 그렇지만, 하지만, 이외에도, 대해, 말하자면, 뿐이다, 다음에, 반대로,
대로, 바꾸어서, 말하면, 바꾸어서, 한다면, 만약, 그렇지않으면, 까악, 퉁,
거리다, 짜당, 응당, 해야한다, 에, 가서, 각, 각각, 여러문, 각종, 각자,
과, 그러므로, 그래서, 고로, 한, 까닭에, 하기, 때문에, 거기나, 이지만,
연, 실로, 아니나다를가, 생각한대로, 진짜로, 한적이있다, 하곤하였다, 하, 하하
오, 왜, 어째서, 무엇때문에, 어찌, 하겠는가, 무슨, 어디, 어느곳, 더군다나
때, 언제, 야, 미봐, 어미, 여보시오, ㅎㅎ, 흥, 휴, 헉헉, 혈혈혈역, 영차
아야, 앓, 아야, 팔팔, 졸졸, 짹짹, 뚝뚝, 주룩주룩, 샅, 우르르, 그래도,
바꾸어말하자면, 혹은, 혹시, 답다, 및, 그에, 따르는, 때가, 되어, 즉, 지
도, 할지라도, 일지라도, 지든지, 몇, 거의, 하마터면, 인젠, 이젠, 된바에야,
그위에, 게다가, 접에서, 보마, 비추며, 보마, 고려하면, 하게될것이다, 일것이
비하면, 시키다, 하게하다, 할만하다, 의해서, 연이서, 이어서, 잇따라, 뒤따라
기대여, 통하여, 자마자, 더욱더, 불구하고, 얼마든지, 마음대로, 주저하지, 앞

```
In [21]: review_DTM
```

Out[21]:

[illegible]

7046 rows × 1002 columns

3.1 데이터 분석 과정

3) 학습데이터, 검증데이터로 분리 후 NaiveBayes 적용

```
In [22]: trn, test=train_test_split(review_DTM, test_size=0.2, shuffle=True, random_state=42)
```

```
In [23]: trn
```

Out [23]:

| | 05 | 1000원 | 1일 | 25311 | 30분 | 3월 | 500원 | 5월 | 7시 | 86 | ... | 환불 | 환하다 | 황금 | 회사 | 후식 | 출름하다 | 옥임자 | 옥화 | 힐링 | label |
|------|-----|-------|-----|-------|-----|-----|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|-----|-------|
| 949 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 4374 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 251 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 4136 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3150 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4426 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 466 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3092 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3772 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 860 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

3623 rows × 1027 columns

```
In [24]: trn_x = trn.drop('label', axis=1)
trn_y = trn['label']
```

```
In [25]: multinB = MultinomialNB()
multinB.fit(trn_x, trn_y)
```

Out [25]: MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)

```
In [26]: multinB.score(trn_x, trn_y)
```

Out [26]: 0.9712945073143804

```
In [27]: test_x = test.drop('label', axis=1)
test_y = test['label']

pred_y = multinB.predict(test_x)
```

```
In [28]: multinB.score(test_x, test_y)
```

Out [28]: 0.9646799116997793

3.2 초기 결과 - Precision

```
from sklearn.metrics import classification_report  
print(classification_report(test_y, pred_y, target_names=['부정', '중립', '긍정']))
```

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 부정 | 0.74 | 0.52 | 0.61 | 33 |
| 중립 | 0.87 | 0.57 | 0.68 | 23 |
| 긍정 | 0.97 | 0.99 | 0.98 | 850 |
| accuracy | | | 0.96 | 906 |
| macro avg | 0.86 | 0.69 | 0.76 | 906 |
| weighted avg | 0.96 | 0.96 | 0.96 | 906 |

전체적으로 높은
PRECISION 값



예측 자체는 정교함

3.2 초기 결과 - Recall

```
In [29]: from sklearn.metrics import classification_report  
print(classification_report(test_y, pred_y, target_names=['부정', '중립', '긍정']))
```

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 부정 | 0.74 | 0.52 | 0.61 | 33 |
| 중립 | 0.87 | 0.57 | 0.68 | 23 |
| 긍정 | 0.97 | 0.99 | 0.98 | 850 |
| accuracy | | | 0.96 | 906 |
| macro avg | 0.86 | 0.69 | 0.76 | 906 |
| weighted avg | 0.96 | 0.96 | 0.96 | 906 |

$$Recall = \frac{TP}{TP + FN}$$

낮은 RECALL 값



‘부정’ 데이터 수 부족
or
긍정으로 잘못 예측 과다

3.3 성능 향상을 위한 원인 분석

1) 테스트 리뷰를 넣어, 어떻게 분류하는지 확인해보기

```
: from konlpy.tag import Okt
  okt = Okt()

: test_review_list = ['영업시간이 짧네요', '별로네요~', '점원분이 너무 일하시는데 느려서']

test_review_list_after_okt = [okt.morphs(review, norm=True, stem=True) for review in test_review_list]
print('형태소 분석기 거친 후 : ')
print(test_review_list_after_okt)

test_review_list_processed = [' '.join(review) for review in test_review_list_after_okt]
test_review_list_processed = clean_text(test_review_list_processed)
print('\n추가 전처리 거친 후 : ')
print(test_review_list_processed, '\n')

# test_DTM_vector = CountVectorizer()
# K_test_DTM = test_DTM_vector.fit_transform(test_review_list_processed)
# print(test_DTM_vector.get_feature_names())
# print(K_test_DTM.toarray(), '\n')

size = len(test_review_list)
test_review_DTM = pd.DataFrame(np.zeros([size, trn_x.shape[1]]), columns = DTM_vector.get_feature_names())
test_review_DTM

형태소 분석기 거친 후 :
[['영업', '시간', '이', '짧다'], ['별로', '네', '요', '~'], ['점원', '분', '이', '너무', '일', '하다', '느리다']]

추가 전처리 거친 후 :
['영업 짧다', '별로', '점원 분 느리다']

:
   05  1000원  1일  25311  30분  3월  500원  5월  7시  86  ...  화자  환불  환하다  황금  회사  후식  출렁하다  욕임자  욕화  힐링
0  0.0    0.0  0.0    0.0  0.0  0.0    0.0  0.0  0.0  0.0  ...  0.0  0.0    0.0  0.0  0.0    0.0  0.0  0.0  0.0
1  0.0    0.0  0.0    0.0  0.0  0.0    0.0  0.0  0.0  0.0  ...  0.0  0.0    0.0  0.0  0.0    0.0  0.0  0.0  0.0
2  0.0    0.0  0.0    0.0  0.0  0.0    0.0  0.0  0.0  0.0  ...  0.0  0.0    0.0  0.0  0.0    0.0  0.0  0.0  0.0

3 rows x 1020 columns
```

```
pred_y = multinNB.predict(test_review_DTM.iloc[:, :trn_x.shape[1]])
pred_y
```

```
array([0, 1, 1], dtype=int64)
```

```
np.set_printoptions(suppress=True)
```

```
pred_y = multinNB.predict_proba(test_review_DTM.iloc[:, :trn_x.shape[1]])
pred_y
```

```
array([[0.20922277, 0.66187071, 0.12890653],
       [0.3797847 , 0.05459929, 0.56561601],
       [0.3082807 , 0.12190473, 0.56981457]])
```

- 순서대로 -1(부정)일 확률, 0(중립)일 확률, 1(긍정)일 확률

3.3 성능 향상을 위한 원인 분석

2) '부정' 라벨을 달고 있지만 긍정으로 예측한 샘플 확인

```
temp[:20]
```

```
array(['딸기쥬스 맛이 약간 약해요', '점원분이 너무 일하시는게 느려서', '타피오카 맛이 약해서', '향이 강해여',  
      '맛이없어', '딸기마보카도라테 맛없음', '불친절합니다.', '자몽에이드가 좀 싱겁다', '불친절해요',  
      '불친절 매장도 지저분하다', '불친절 0.5점도 아까움', '가게 내부는 조금 좁아요',  
      '갈적마다 불친절한 서비스에 기분이 썩 좋지는 않아요.', '아쉽~~그래도 맛있음',  
      '곳들 중에 직원분들이.. 가장 불친절해요.', '기본 커피를 묻지도 않고', '불친절해여', '불친절해요',  
      '불친절해요.', '스벅하나뿐이라 그런지 항상 자리가 없네요.'], dtype=object)
```

3.3 성능 향상 과정

1) 라벨 자체가 잘못 매겨진 경우

은 음.. 멍멍해요', '빵 바삭하지 않아요', '직원이 홀신질함', '다
좀 불편', '불친절하네요~', '서비스가 별로여서.', '불친절하고
맛이 너무 강해요', '바닐라라떼는 너무 달지않고', '파니니는
모든 직원이 불친절해서 안갔어요.', '세상제일 불친절', '투썸
친절해서 기분이 나쁘네요.', '???? 불친절 끝판', '실내가 너무
이 별로 덜 친절하셨음', '마카롱 나쁘지않아요', '불친절해요!
불편하고', '조금 시끌벅적한 분위기지만.괜찮아여', '알바생
과 양이 너무 아쉽다.', '에이드도 멍멍해요', '에어컨이 너무
좀 불편해요', '맛이 진짜없음', '빵 바삭하지 않아요', '맛이 없

직접 라벨 수정

3.3 성능 향상 과정

1) 데이터가 너무 작은 경우

아메리카노 최악이었어요 => '최악'이다 => '최악'이 데이터에 하나밖에 없어서 학습 때 보지 못해서 긍정 비율이 68%
포장도 너무 느려요 => '느리다' => '느리다'가 데이터에 두개밖에 없는데 둘 다 검증용 데이터라 학습 때 한번도 보지 못해서 긍정 비율이 68%



StratifiedKFold 로 모델을 학습

* StratifiedKFold

과적합

교차 검증



StratifiedKFold

```
: skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
skf.get_n_splits(X, y)
```

```
: 5
```

```
: multinB = MultinomialNB()

for train_index, test_index in skf.split(X, y):
#     print("TRAIN:", train_index, "TEST:", test_index)
#     print(len(train_index))
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y[train_index], y[test_index]

    multinB.fit(X_train, y_train, alpha=2.0)
#     multinB.score(y_train, y_test)
```


3.3 성능 향상 과정

1) 데이터가 너무 작은 경우



강남구와 인접한 서초구의
카페 리뷰 데이터 추가 수집

'ㅇㅋㅎ은는이가네요 커피 남자 여자 직원 알바생
사장 매장 분위기 아메리카노 포장 빵 밀크티 자리 우유
초콜릿 크로와상 케익 케익 푸딩 디저트 공간 실내
인테리어 층 분들 얼음 스콘 테이블 간격 알바 카페 모카
분 드립커피 자몽 요거트 음료 에 미소 로 도 메뉴 망고
그린 애플 시즌 샐러드 처음 서브웨이 명 밀크 티 그린티
라떼 에어컨 테이크아웃 주문 포장 대추차 샌드위치
아이스크림 홍차 폼 브라운 슈가 쥬얼리 뿐만 아니라'

3.3 성능 향상 과정

2) 중립 샘플 제거

| | | | | | | |
|------|-------------|-------|-----|------------|----------------------------------|----|
| 4122 | 스타벅스커피 | mood | 5 | 음악이 너무 좋아요 | ['음악', '이', '너무', '좋아', '요'] | -1 |
| 4229 | 85번가 | taste | 5 | 할수없을 것 같아요 | ['할수', '없', '하다', '없', '아요'] | -1 |
| 4246 | 스타벅스커피 | taste | 5 | 커피가 좀 짜요 | ['커피', '가', '커피', '가', '요'] | -1 |
| 4253 | 스타벅스커피 | taste | 5 | 커피가 좀 짜요 | ['커피', '가', '커피', '가', '요'] | -1 |
| 4266 | 빌리엔젤 | taste | 5 | 당근케이크예요 | ['당근', '케', '당근', '케', '예요'] | -1 |
| 4311 | 잠바주스 | mood | 5 | 매장은 좀 짜요 | ['매장', '은', '매장', '은', '요'] | -1 |
| 4409 | 선릉중앙점 | taste | 5 | 아이스크림이에요 | ['아이스크', '아이스크', '아이스크', '예요'] | -1 |
| 4469 | 달콤커피 | taste | 5 | 구 커피도 좋아요 | ['구', '커피', '구', '커피', '요'] | -1 |
| 1 | 이디야강남점 | taste | 5 | 달고나라떼예요 | ['달고나', '라떼', '달고나', '라떼', '예요'] | 0 |
| 21 | 스타벅스 | taste | 5 | 잘 먹었습니다 | ['잘', '먹었', '자다', '먹', '습니다'] | 0 |
| 22 | 스타벅스 | taste | 5 | 말차라떼예요 | ['말차', '라', '말차', '라', '예요'] | 0 |
| 59 | 수분(Soobong) | taste | 4 | 파스타는 맛있어요 | ['파스타', '맛', '파스타', '맛', '있어요'] | 0 |
| 62 | 투썸플레이스 | taste | 5 | 커피가맛나요 | ['커피', '가', '커피', '가', '요'] | 0 |
| 79 | 투썸플레이스 | taste | 2.5 | 커피가 텃어요 | ['커피', '가', '커피', '가', '요'] | 0 |
| 80 | 투썸플레이스 | taste | 0.5 | 생크림은 맛있어요 | ['생크림', '맛', '생크림', '맛', '있어요'] | 0 |
| 83 | 투썸플레이스 | taste | 4 | 커피도 맛나요 | ['커피', '도', '커피', '도', '요'] | 0 |
| 102 | 이디야커피 | taste | 4.5 | 커피가 맛나요 | ['커피', '가', '커피', '가', '요'] | 0 |
| 126 | (주)한스케 | taste | 4 | 옛생각나는 맛이에요 | ['옛', '생각', '옛', '생각', '있어요'] | 0 |
| 132 | (주)한스케 | taste | 5 | 딸기케익이에요 | ['딸기', '케', '딸기', '케', '예요'] | 0 |
| 135 | (주)한스케 | taste | 4.5 | 치즈수플러예요 | ['치즈', '수', '치즈', '수', '예요'] | 0 |

중립 라벨 제거 후
긍정, 부정 라벨만 있을 때 확률 관찰

3.4 성능 향상 결과

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 부정 | 0.85 | 0.53 | 0.65 | 32 |
| 중립 | 0.89 | 0.40 | 0.56 | 84 |
| 긍정 | 0.93 | 1.00 | 0.96 | 790 |
| accuracy | | | 0.93 | 906 |
| macro avg | 0.89 | 0.64 | 0.72 | 906 |
| weighted avg | 0.92 | 0.93 | 0.91 | 906 |

[성능향상 전]



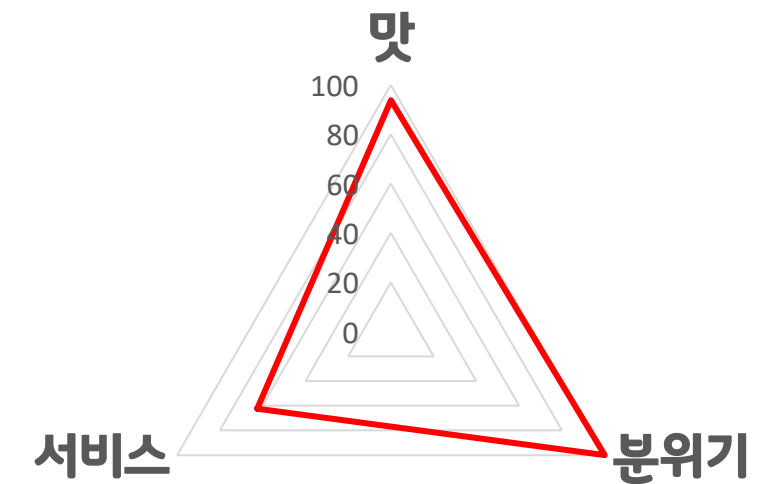
| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 부정 | 0.85 | 0.81 | 0.83 | 63 |
| 긍정 | 0.99 | 0.99 | 0.99 | 1346 |
| accuracy | | | 0.99 | 1409 |
| macro avg | 0.92 | 0.90 | 0.91 | 1409 |
| weighted avg | 0.98 | 0.99 | 0.98 | 1409 |

[성능향상 후]

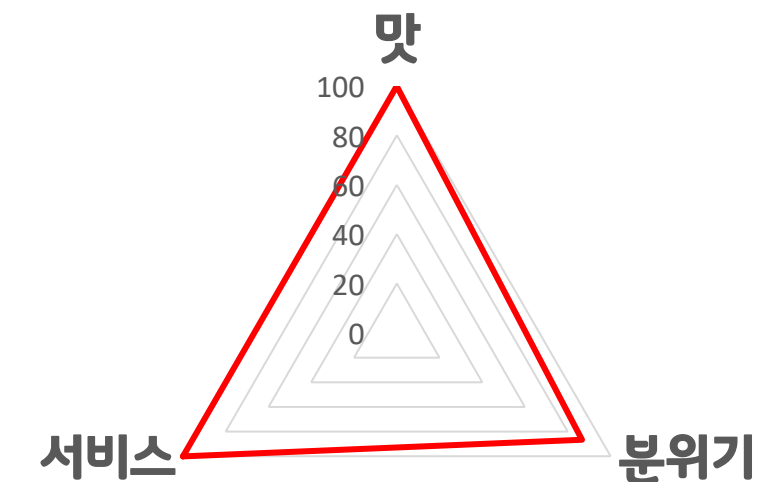
- 전반적 평가지표 모두 개선됨
- 이전보다 정교해진 리뷰 분석을 통한 카페 긍/부정 분류

3.5 결과

[투썸플레이스 압구정로데오역점] 매장 taste 요소의 긍정 리뷰 비율은 93.75% 입니다.
[투썸플레이스 압구정로데오역점] 매장 mood 요소의 긍정 리뷰 비율은 100.0% 입니다.
[투썸플레이스 압구정로데오역점] 매장 service 요소의 긍정 리뷰 비율은 62.5% 입니다.



[스타벅스 강남오거리점] 매장 taste 요소의 긍정 리뷰 비율은 100.0% 입니다.
[스타벅스 강남오거리점] 매장 mood 요소의 긍정 리뷰 비율은 86.67% 입니다.
[스타벅스 강남오거리점] 매장 service 요소의 긍정 리뷰 비율은 100.0% 입니다.



발표글

마칩니다.

Thank you!

