

〔 텍스트 마이닝을 이용한 카페 평점 예측 〕



4조 유미리 서상혁 송기훈

Table of Contents

목 차

001 하 나 의 카 페 , 다 른 시 각 ?

002 데 이 터 설 명

003 분 석 과 정

004 분 석 결 과



A photograph of a cafe interior. The background is a rustic brick wall. On the left, there is a wooden shelf with various items, including a coffee machine and boxes. A yellow stool is in the foreground on the left. In the center, a wooden table is surrounded by yellow chairs. On the right, two wooden surfboards are leaning against the wall. The text "Part 1" and "하나의 카페, 다른 시각?" is overlaid on the image.

Part 1

하나의 카페, 다른 시각?

하 나 의 카 페 , 다 른 시 각 ?

하 나 의 카 페 , 왜 평 점 이 다 를 까 ?



카페온화 성수점

카페 · 후기 2.8 (54) · 리뷰 269



지도



길찾기



로드뷰



공유하기



카페온화 성수점 카페

★ 4.32/5 · 방문자리뷰 513 · 블로그리뷰 755

출발

도착



리뷰



거리뷰



공유



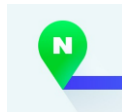
카페온화 성수

4.0 ★★★★★ 리뷰 265개

카페



Kakao Map



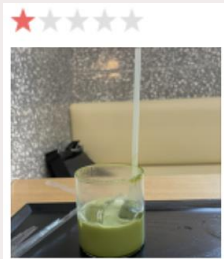
Naver Map



Google Map

하 나 의 카 페 , 다 른 시 각 ?

리뷰를 통한 평점 파악



칠천오백원



양이 너무 창렬이에요....

♡ 좋아요

수플레 맛집이라더니 맞네요 ♥ 딸기도 신선했고 몽글몽글하니 존맛탱이었음!! 대신.. 가격이 꽤 비싼편 ㅠ 커피는 보니까 직접 로스팅하시던데~ 로스... ∨



디저트가 맛있어요 +2

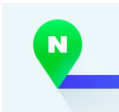
★★★★☆ 5달 전

매장이 넓고 쾌적합니다.

수플레팬케이크는 계란 냄새가 좀 나는 것 같아요.



Kakao Map



Naver Map



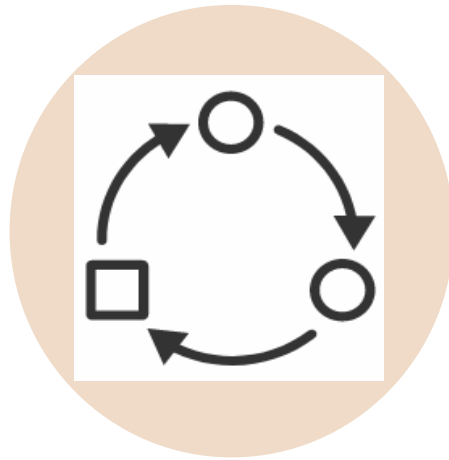
Google Map

하나의 카페, 다른 시각?

프로젝트 진행 과정



Data
Collection



Preprocessing
&
Labeling



Text Mining



Modeling

A close-up photograph of three hands holding coffee cups in a toast. The top hand holds a white cup with latte art. The bottom-left hand holds a dark cup with ice and coffee. The bottom-right hand holds a white cup with latte art. The background is blurred, showing a wooden table and coffee-making equipment.

Part 2

데이터 설명

데이터 설명

출처, 데이터 규모

Collection -> 카카오 지도, 다이닝 코드, 식신

약 25,000개의 리뷰와 평점 크롤링

Data Scale -> 카카오 지도 : 14,000, 다이닝 코드 : 8,000, 식신 : 3,000

Location -> 카페 밀집도가 많은 서울시 행정구역
[ex) 성수동, 서초동, 송파동, 사당동, ...

데이터 설명

데이터 Features / Target

Target

★★★★☆ 5달 전

Feature

매장이 넓고 쾌적합니다.
수플레팬케이크는 계란 냄새가 좀 나는 것 같아요.

Preprocessed feature

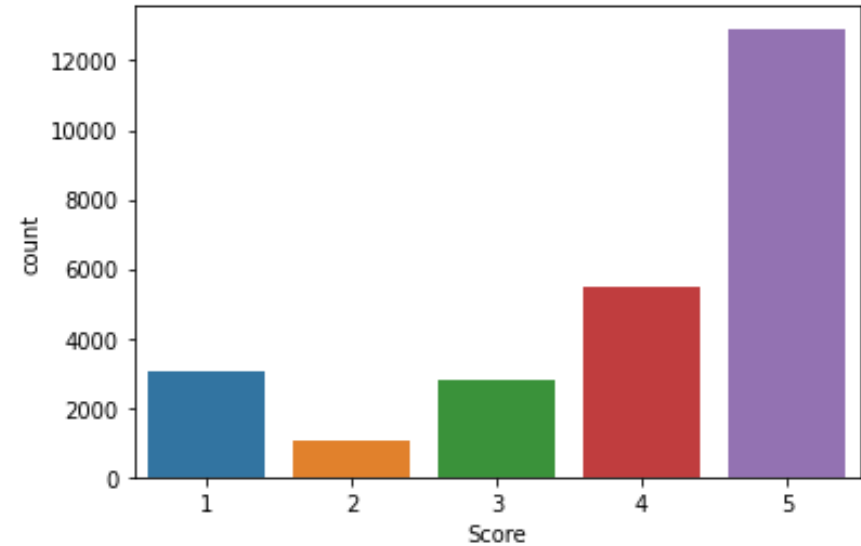
['매장', '넓다', '쾌적하다', '플레', '팬케이크', '계란', '냄새', '같다']

데 이 터 설 명

데 이 터 Features / Target

Target : 5

of Reviews : 25,569



< Target 의 분 포 >



Part 3

분석 과정

Crawling

Plan



: 다이닝 코드



: 네이버지도



: 구글 지도



: 카카오 지도



: 식신



: 망고 플레이트

Actual



: 다이닝 코드



: 카카오 지도



: 식신



Crawling : 카카오 지도

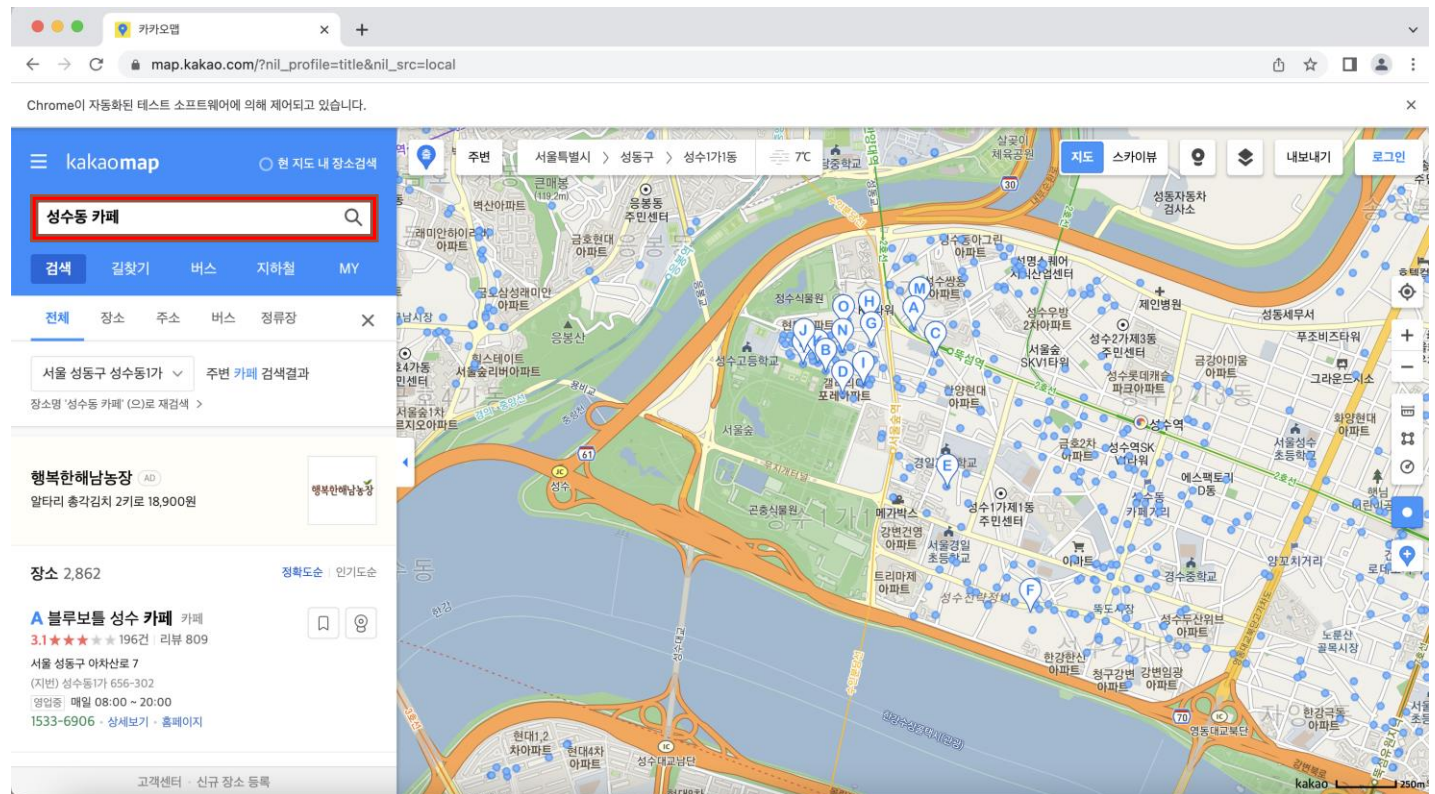


1. 사이트 접속 & 키워드 입력

- XPath로 검색창 위치 찾기

```
/html/body/div[5]/div[2]/div[1]/div[7]/div[5]/a
```

- Selenium의 Keys 모듈 이용해서 keyword 넣고 검색하기



분석 과정 : Crawling



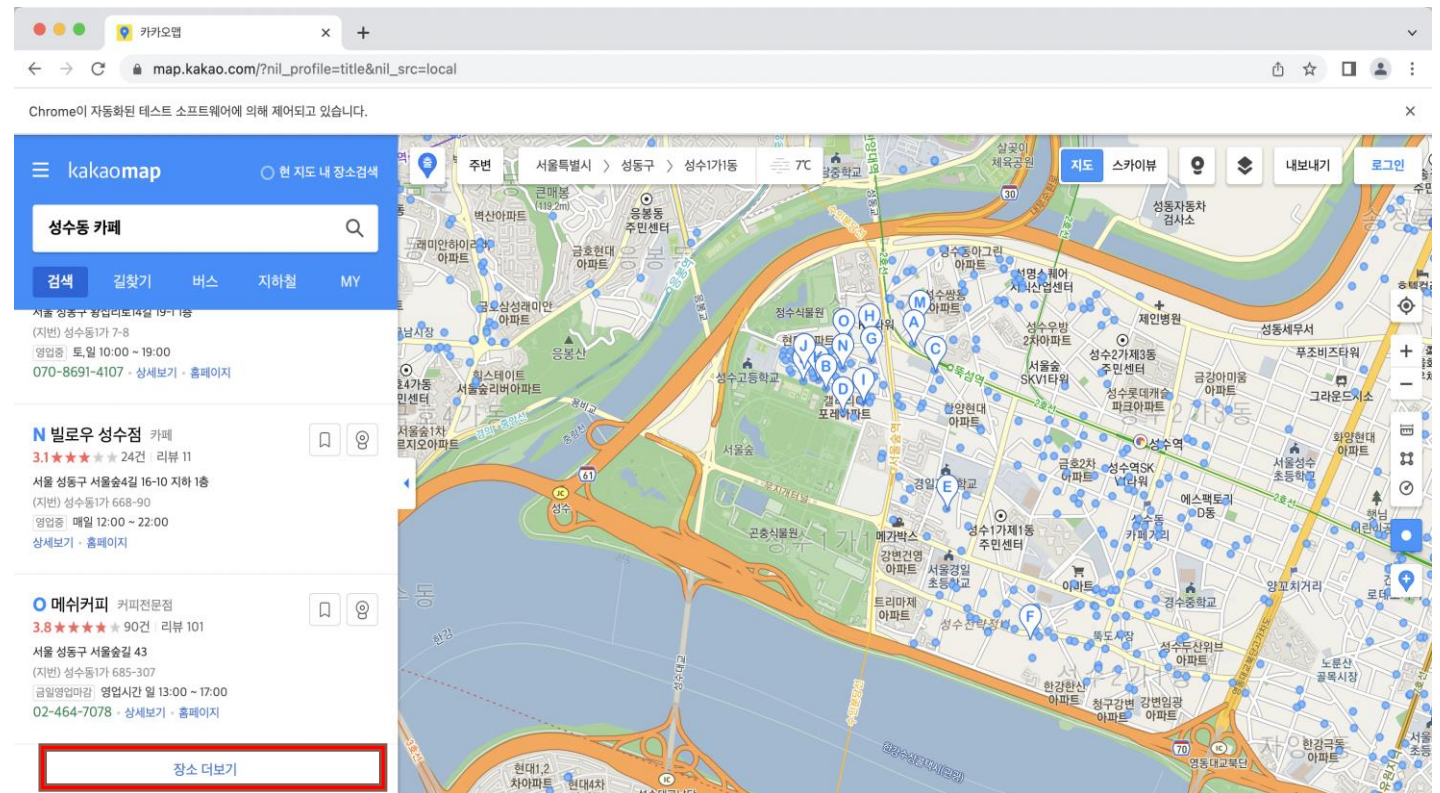
Crawling : 카카오 지도

2. 장소 더보기 클릭

- XPATH로 장소 더보기 위치
가져 오기

```
/html/body/div[5]/div[2]/div[1]  
/div[7]/div[5]/a
```

- Selenium으로 클릭하기





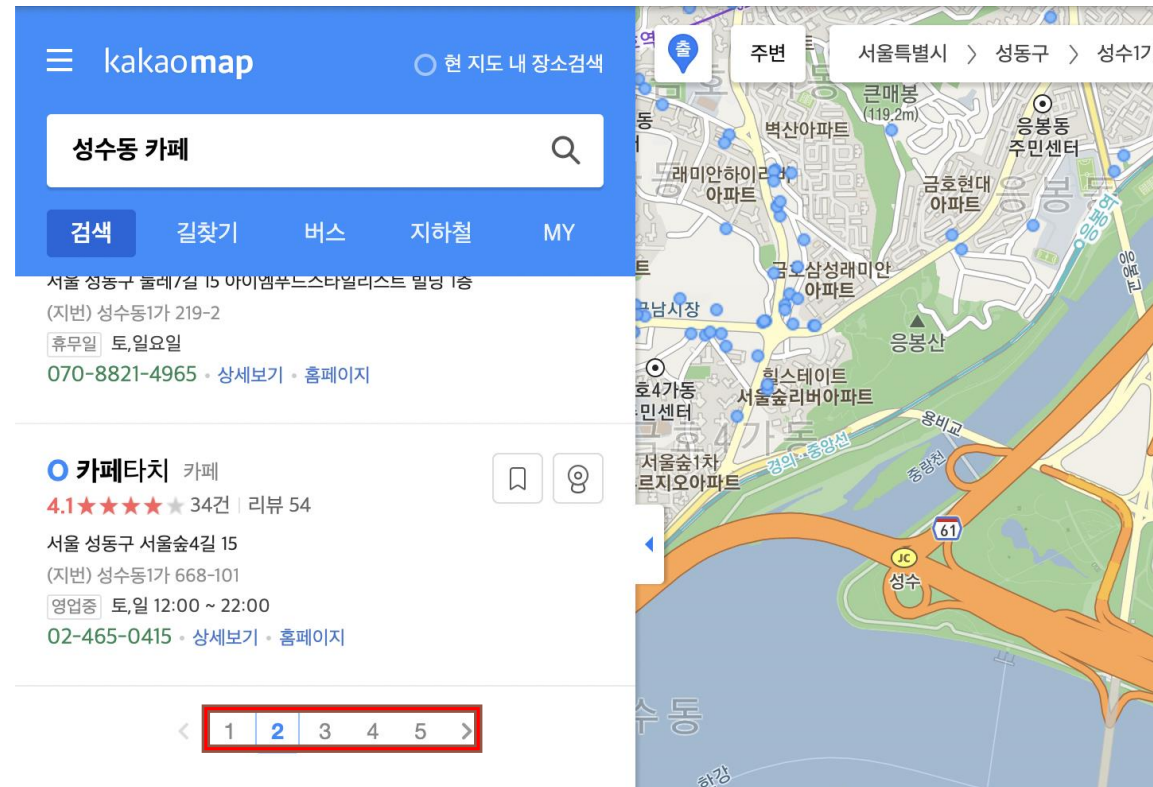
Crawling : 카카오 지도

3. 페이지 버튼 클릭하기

- XPATH로 페이지 버튼 위치 찾기

```
/html/body/div[5]/div[2]/div[1]  
/div[7]/div[6]/div/a
```

- Selenium이용해서 페이지 버튼
클릭하기



분석 과정 : Crawling

Crawling : 카카오 지도

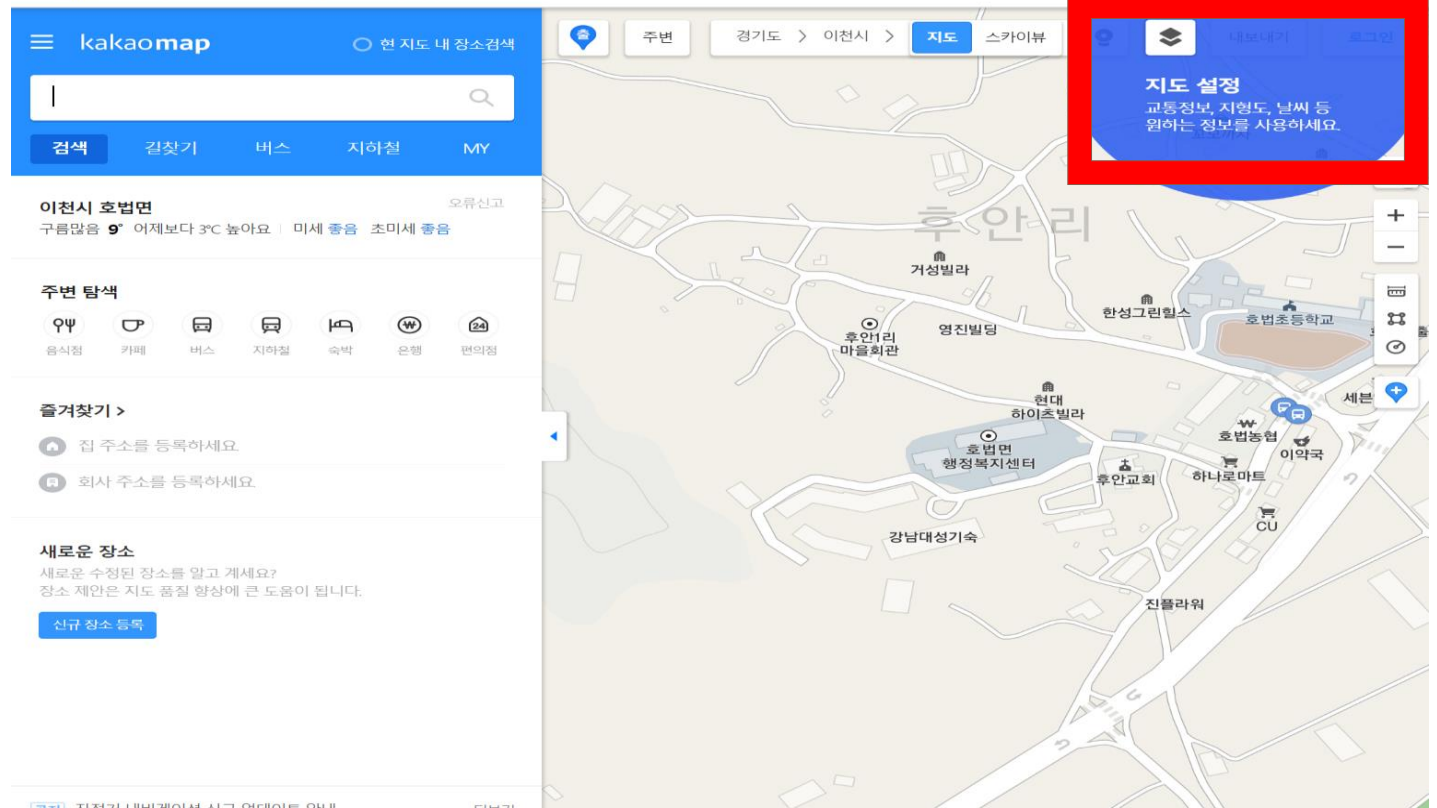
장소 더보기 클릭 에러 (문제 및 해결 방안)

- 상단 팝업으로 인한 장소 더보기 element 가져오는데 error

- 해결 방안

- For 문을 통해 2번 url 접속하기

Chrome이 자동화된 테스트 소프트웨어에 의해 제어되고 있습니다.

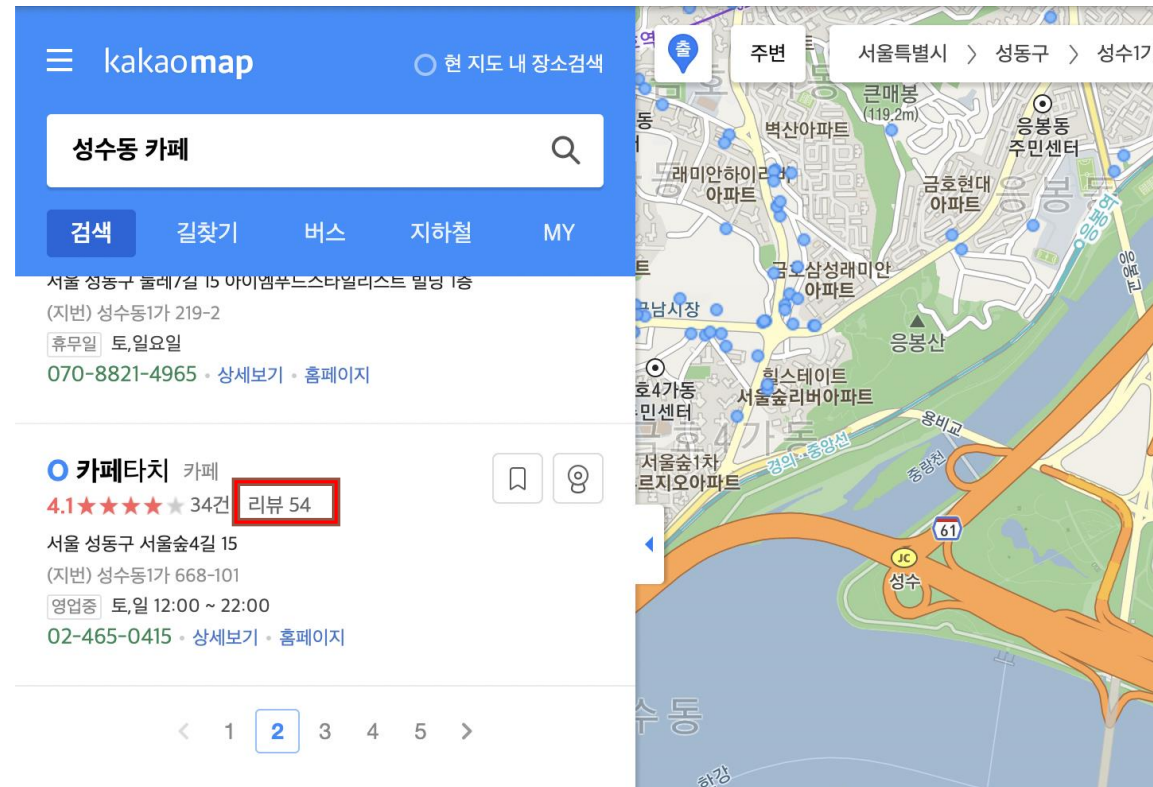




Crawling : 카카오 지도

4. BeautifulSoup으로 리뷰 페이지 링크 수집

- HTML 태그로 페이지로 페이지 링크 수집
- a.numberofscore
- 태그의 href로 링크 수집





Crawling : 카카오 지도

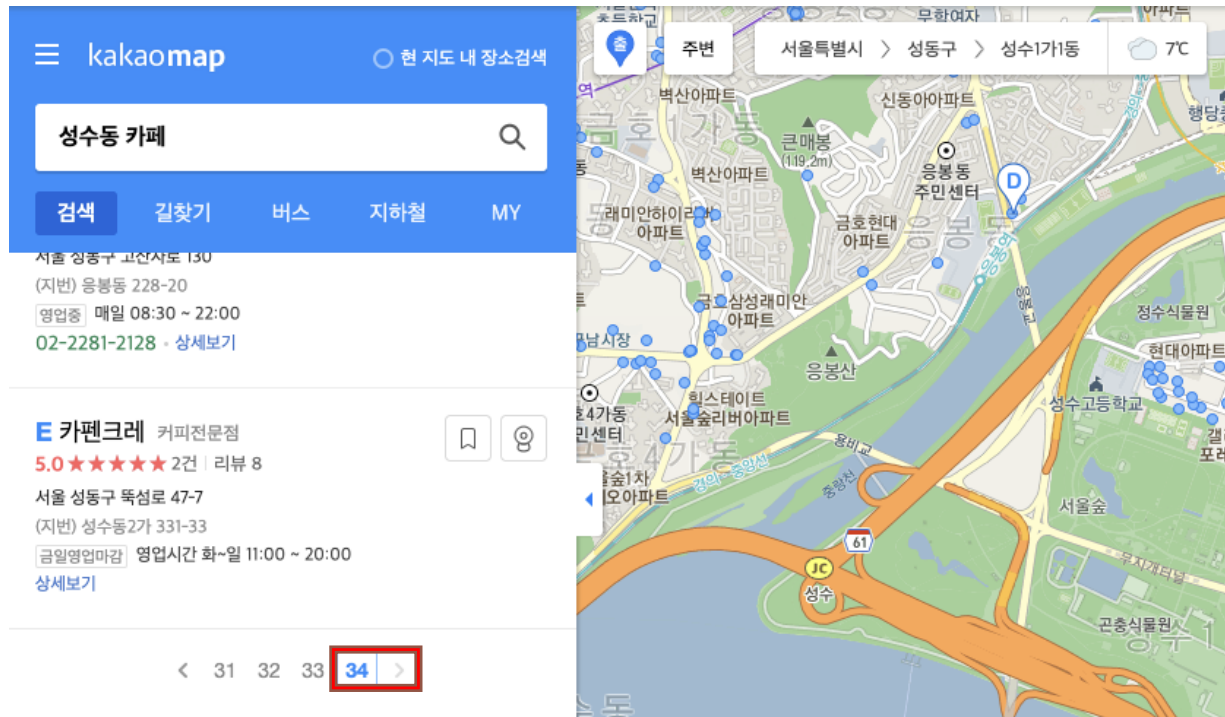
5. 마지막 페이지까지 이동하면서
모든 카페들의 리뷰 링크 가져오기 (문제 및 해결 방안)

- While 문으로 무한루프

1. 5 번째 페이지로 접속 후
2. 페이지 내의 수집 후
3. > (다음 페이지) 버튼을 클릭
4. 무한 반복

- 끝 페이지인 경우

1. 5 개의 페이지 버튼 없는 경우
에러 발생
2. Try - except 구문으로 예외처리
3. 다음 페이지가 버튼의 class 가
'next disabled'로 변경
4. 조건문 if를 활용해 무한루프를
끝내기





Crawling : 카카오 지도

6. 리뷰 페이지 내의 후기 더보기 클릭

- XPATH로 후기 더보기 element 가져오기
`/html/body/div[2]/div[2]/div[2]/div/div/a`

- Selenium으로 클릭하기

- While문 무한 반복 클릭

- 마지막에 '후기 접기'를 이용해서 if문으로
후기 더보기 일때만 클릭하기

전체 29

4.3점 ★★★★★

맛 2

친절 2

분위기 2



나아영 | 후기 55 | 별점평균 4.1 | 2023.01.14.

★★★★★

맛

친절

분위기



마카롱 마이썸

커피값은 성수동 평균이라긴하지만 그래도 좀비싼느낌.

♡ 좋아요



요호 | 후기 83 | 별점평균 4.5 | 2023.01.13.

★★★★★

- 마카롱 식감도 맛도 너무너무 좋은데 종류가 너무 한결 같아서 아쉬움. 특히 맛들 중 최애도 없어서

- 기본적인 맛이 더 많았으면 좋겠음 레드벨벳이라던지 일그레이라던지, 솔티드 카라멜이라던지 녹차라던지 지금은 맛이 딱하...

♡ 좋아요



ㅋ | 후기 85 | 별점평균 3.9 | 2022.09.18.

★★★★★

♡ 좋아요

후기 더보기



Crawling : 카카오 지도

7. 리뷰 페이지 내의 모든 리뷰, 평점 가져오기

- BeautifulSoup 라이브러리 사용
- HTML 태그로 페이지로 리뷰, 평점 가져오기
- 리뷰 태그: `p.txt_comment span`
- 평점 태그:
`div.evaluation_review`
`span.ico_star.inner_star`
- 수집한 링크들은 for문 사용해 위의 과정들을 반복

전체 29

4.3점 ★★★★★

맛 2

친절 2

분위기 2



나아영 | 후기 55 | 별점평균 4.1 | 2023.01.14.

★★★★★

맛

친절

분위기



마카롱 마이썰

커피값은 성수동 평균이라긴하지만 그래도 좀비싼느낌.

♡ 좋아요



요호 | 후기 83 | 별점평균 4.5 | 2023.01.13.

★★★★★

- 마카롱 식감도 맛도 너무너무 좋은데 종류가 너무 한결 같아서 아쉬움. 특히 맛들 중 최애도 없어서

- 기본적인 맛이 더 많았으면 좋겠음 레드벨벳이라던지 일그레이라던지, 솔티드 카라멜 이라던지 녹차라던지 지금은 맛이 딱하...

♡ 좋아요



후기 95 | 별점평균 3.9 | 2022.09.18.

★★★★★

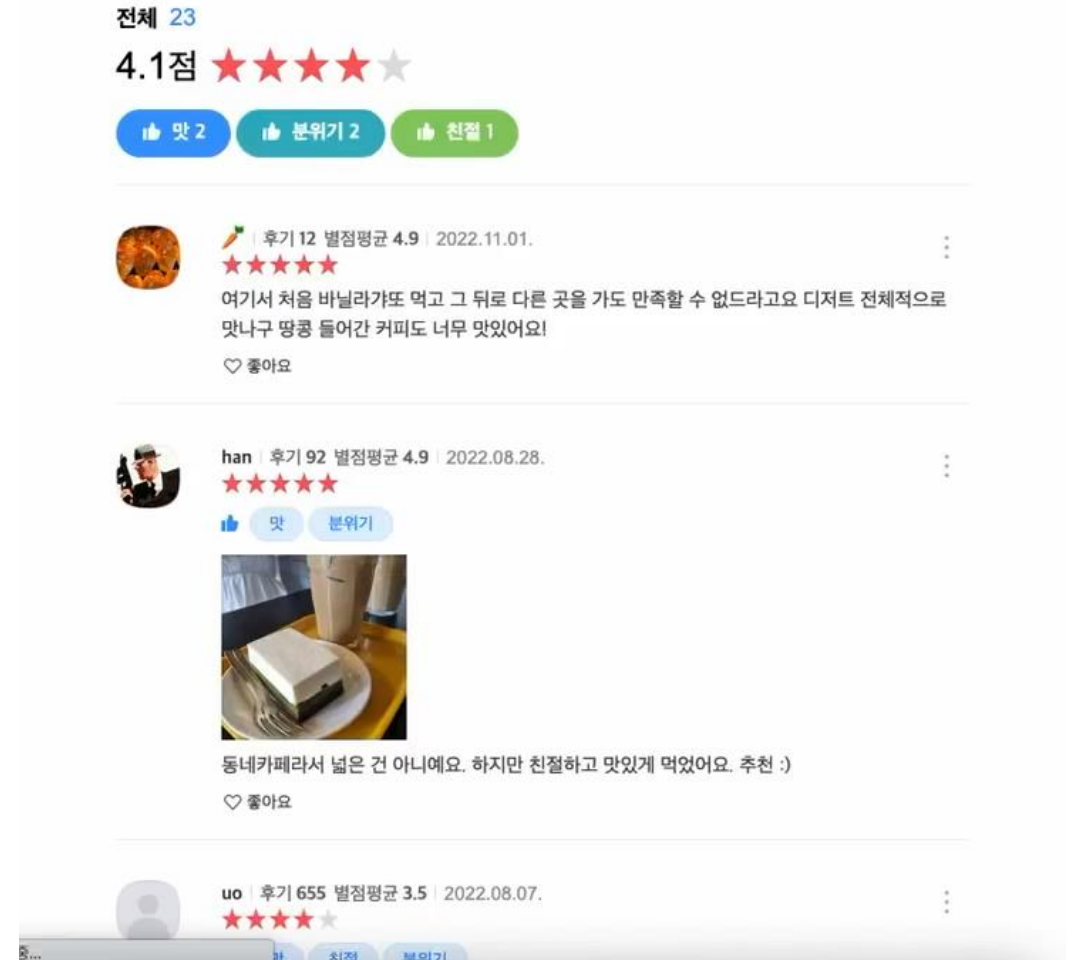
♡ 좋아요

후기 더보기 ~



Crawling : 카카오 지도

- 리뷰 페이지 링크 수집
- 페이지 버튼 클릭



- 후기 더보기 클릭

PreProcessing

- 전 처리 (Preprocessing) 함수 생성

1. 한글, 공백을 제외한 문자 제거

2. Okt 객체를 이용하여 형태소 토큰화, 품사 태깅

3. 품사 - 명사, 동사, 형용사만 남기기

4. 노이즈, 불용어 제거

● ● ●  stop_words.txt

그저 그렇다 나라 이다 딱하다 굿즈 아스 왜왜 샌프란시
스코 개봉 도쿄 블루 보틀 육천원 거마 여기 싶다 스타
디 때문 댓글 여자 들여다보다 남자 여자 계시다 치다
드립커피 있다 에이 들다 착즙 가세 이드 다크 상그리
장님 니당다 프린트 불다 트램벌린 도넛 밀가루 나서다
감다 녹차 말차 헤드셋 마이크 이나 자체 보드 마우스
사양 로지텍 일산 라이언 출입 말다 어차피 예정 가야
리도 악세다 다방 슈페 아인 한류 북한 개다 인터넷 공
용 문턱 코코넛 바안 보체 샘플러 글라스 칵테일 위스키
계단 비상 지큐 조리사 애오개 파다 스카 법도 일본인
야하다 고여 정체 나다 쓰시 시반 후시 스위스 다임 그
게 갈다 만요 장님 여사 친전 읍니 낭시 전과 거려 다해
대서 사보다 달구 가게 절기 두다 꾸다 돼다 런가 로나
호점층 이드 도리 오잉 오예 프라하 상치 그루 나루 허
니 시험 리코 밤톨 어어 토요일 오시 카이 전당 제스 격인
아메 소담 편이 사람인 아바 머거다 펫프렌 리해 놀숄

< Okt 불용어 사전 >

단어 선정 기준


: 평점에 영향을 미치지 않는 단어들

PreProcessing

1. 한글, 공백을 제외한 문자 제거

- 이모티콘, 특수문자 등 의미가 없는 문자들이 포함되어 있어
한글, 공백을 제외한 문자를 제거함

Example) :

“맛  다른 리뷰처럼 라떼 진짜 맛있어용 시럽 안 넣었는데도 맛 있는!
점심 시간 엄청 북적거림 오후 세시쯤은 책 읽어도 될 만큼 여유 있는 분위기!”



“ 맛 다른 리뷰처럼 라떼 진짜 맛있어용 시럽 안 넣었는데도 맛 있는
점심 시간 엄청 북적거림 오후 세시쯤은 책 읽어도 될 만큼 여유 있는 분위기”

PreProcessing

2. Okt 객체를 이용하여 형태소 토큰화, 품사 태깅

- 한국어 형태소 분석을 위한 라이브러리인 KoNLPy의 형태소 분석기인 Okt 객체를 생성한 후, Okt.pos() 함수를 이용하여 토큰화 및 품사 저장

Example) :

“맛 다른 리뷰처럼 라떼 진짜 맛있어용 시럽 안 넣었는데도 맛 있는
점심 시간 엄청 북적거림 오후 세시쯤은 책 읽어도 될 만큼 여유 있는 분위기”



['맛', '다른', '리뷰', '처럼', '라떼', '진짜', '맛있다', '시럽', '안', '넣다', '맛', '있다',
'점심', '시간', '엄청', '북적', '거림', '오후', '세시', '쯤', '은',
'책', '읽다', '되다', '만큼', '여유', '있다', '분위기']

[('맛', 'Noun'), ('다른', 'Noun'), ('리뷰', 'Noun'), ('처럼', 'Josa'), ('라떼', 'Noun'), ('진짜', 'Noun'), ('맛있다', 'Adjective'),
('시럽', 'Noun'), ('안', 'Noun'), ('넣다', 'Verb'), ('맛', 'Noun'), ('있다', 'Adjective'), ('점심', 'Noun'), ('시간', 'Noun'),
('엄청', 'Adverb'), ('북적', 'Noun'), ('거림', 'Noun'), ('오후', 'Noun'), ('세시', 'Noun'), ('쯤', 'Suffix'), ('은', 'Josa'),
('책', 'Noun'), ('읽다', 'Verb'), ('되다', 'Verb'), ('만큼', 'Noun'), ('여유', 'Noun'), ('있다', 'Adjective'), ('분위기', 'Noun')]

PreProcessing

3. 품사 - “동사, 형용사, 명사”만 남기기

- 실질적인 의미를 담고 있는 품사라고 생각되는 동사, 형용사, 명사를 제외한 품사는 제거

Example) :

['맛', '다른', '리뷰', '처럼', '라떼', '진짜', '맛있다', '시럽', '안', '넣다', '맛', '있다',
'점심', '시간', '엄청', '북적', '거림', '오후', '세시', '쯤', '은',
'책', '읽다', '되다', '만큼', '여유', '있다', '분위기']



['맛', '다른', '리뷰', '라떼', '진짜', '맛있다', '시럽', '안', '넣다', '맛', '있다',
'점심', '시간', '북적', '거림', '오후', '세시',
'책', '읽다', '되다', '만큼', '여유', '있다', '분위기']

PreProcessing

4. 노이즈, 불용어 제거

- 불용어 사전 제작 후, 학습에 방해가 될 것으로 보이는 불용어에 포함된 단어와, 한 글자의 단어는 제거

Example) :

['맛', '다른', '리뷰', '라떼', '진짜', '맛있다', '시럽', '안', '넣다', '맛', '있다',
'점심', '시간', '복적', '거림', '오후', '세시',
'책', '읽다', '되다', '만큼', '여유', '있다', '분위기']



['다른', '리뷰', '라떼', '진짜', '맛있다', '시럽', '넣다', '있다',
'점심', '시간', '복적', '거림', '오후', '세시',
'읽다', '되다', '만큼', '여유', '있다', '분위기']

Vectorization : TF-IDF

TF-IDF

- 단어들의 출현 빈도와 단어의 문서 빈도수를 통해 희귀성을 고려하여 정확하게 단어의 중요성을 수치화한 기법

```
1 tfidf_train.toarray()
```

```
array([[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.]])
```

<TF-IDF 행렬>

```
1 tfidf_train.toarray().shape  
(25244, 4630)
```

<TF-IDF의 형태>

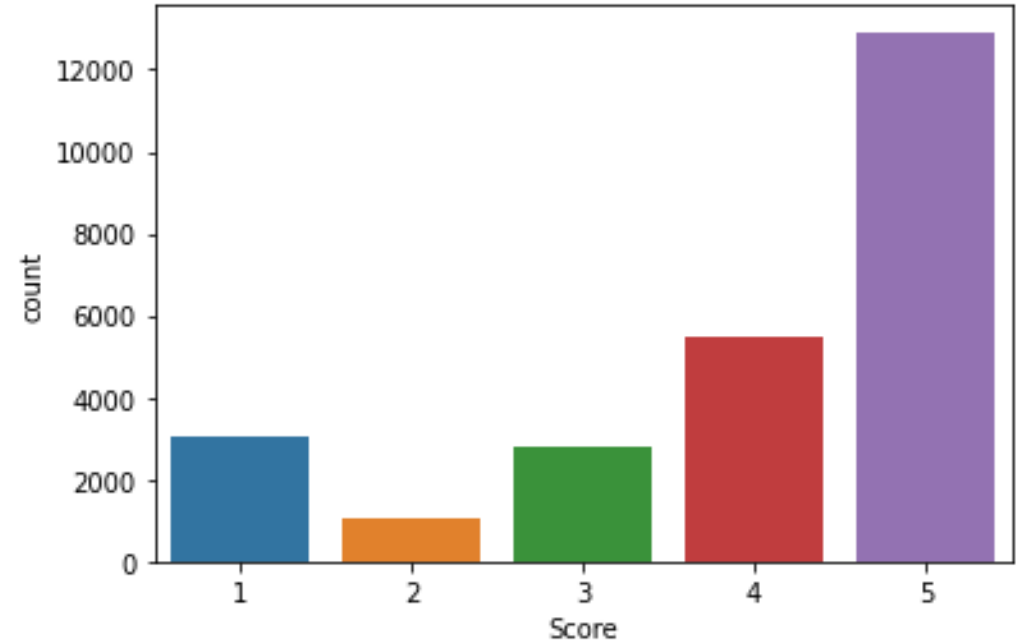
A top-down view of four white ceramic cups filled with coffee, each featuring a different latte art design. The cups are arranged in a square pattern on a dark, textured wooden tray. The background is dark and out of focus. The text 'Part 4' and '분석 결과' is centered over the image, flanked by two large, thin, white curly braces.

Part 4

분석 결과

데이터 Target

< Target 의 분포 >



< Modeling 평가 기준 > : Recall

Why?

- Target 데이터의 분포 비대칭성이 존재하기 때문에 데이터 분포의 비대칭을 고려하지 않은 accuracy는 적합하지 않다.
- 즉, 실제 값이 Positive인 것 중에 예측한 값이 Positive인 비율을 나타내는 Recall을 쓰는 것이 적합하다.
- Ex) 스팸 메일을 정확하게 스팸으로 예측하는 것

모델 평가 기준 : Recall

Macro : 각 Target들의 Confusion matrix로부터 구한 평균 값

Weighted : 실제 Target의 개수를 고려하여 가중치를 곱하여 구한 평균 값

Micro : 각 Target들의 Confusion matrix를 종합한 새로운 Confusion matrix로부터 구한 평균 값

	precision	recall	f1-score	support
1	0.68	0.64	0.66	608
2	0.26	0.05	0.08	213
3	0.32	0.06	0.11	650
4	0.39	0.26	0.31	1261
5	0.61	0.88	0.72	2690
accuracy			0.58	5422
macro avg	0.45	0.38	0.37	5422
weighted avg	0.52	0.58	0.52	5422

Counter({5: 2690, 1: 608, 4: 1261, 3: 650, 2: 213})

<classification의 report>

Analyze : Modeling

1 .



- X G B o o s t 모델에 다중 분류를 위한 `Multi:softmax`, `Multi:softprod` 파라미터를 활용하기 위함

2 .



- L i g h t G B M 모델에 다중 분류를 위한 `application의 multiclass` 파라미터를 활용하기 위함

Analyze : Modeling

1.

dmlc
XGBoost

recall score : 0.34181301892977994

-> Underfit

- 데이터 복잡도 ↓
- 모델 복잡도 ↑

2.

 **LightGBM**

recall score : 0.38031217778997173

-> Underfit

- 데이터 복잡도 ↓
- 모델 복잡도 ↑

Analyze : Modeling



Before

데이터 핸들링 :

(25244, 4630)

모델 핸들링 :

max_depth=3, n_estimators=200

재현율: 0.38031217778997173

After

(25244, 300)

, max_depth=9, n_estimators=400

재현율: 0.3786981337771428

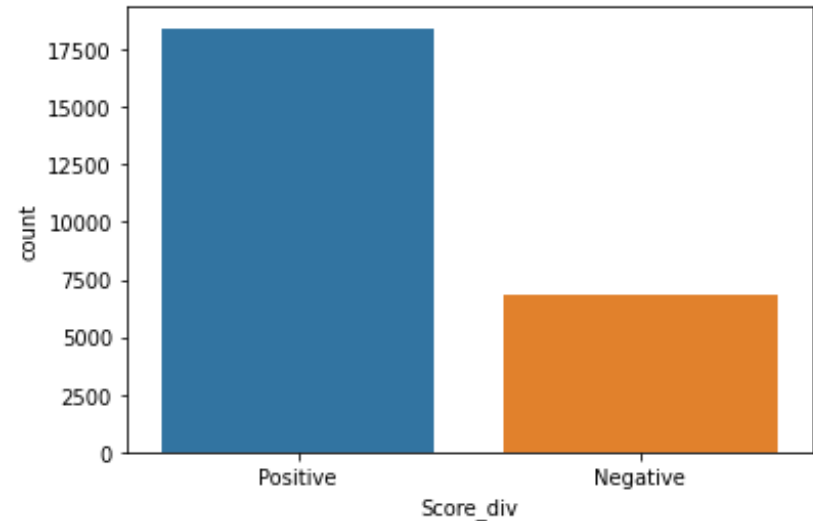
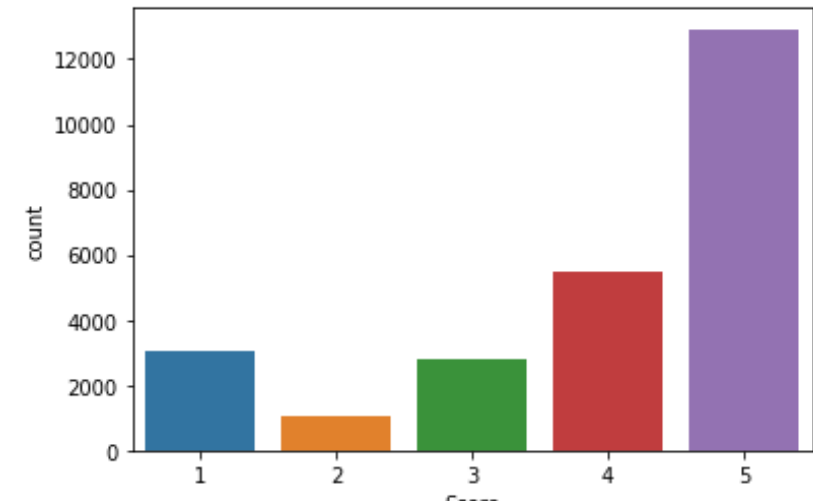
Analyze : 한계점 & 개선사항

< 리뷰 평점 데이터 분포도 한계 >

- 리뷰 데이터 특성상 1점 ~ 3점 평점이 4점, 5점 평점에 비해 비중이 적기 때문에 1~3점 평점 데이터 수집에 한계가 존재
- 즉, 모델의 성능에 문제가 있는 것이 아닌, 치우쳐져 있는 데이터의 분포도로 인해 새로운 가설을 성립

< 모델 성능 개선을 위한 새로운 가설 생성 > : 다중분류 → 이진분류

- Negative : 1 ~ 3 점
- Positive : 4 ~ 5 점



Analyze : 한계점 & 개선사항



다중분류

Train 재현율: 0.4238248421038482
Test 재현율: 0.37174541605412315

이진분류

Train 재현율: 0.8355089844511726
Test 재현율: 0.7356708980840572



- 다중분류 문제를 이진분류 문제로 바꿈으로써 모델의 성능 상승
- Train data > Test data = 과적합 발생
즉, 이진분류 문제는 다중분류에 비해 복잡도가 낮은 Logistic을 쓰는 것이 적합

Analyze : 한계점 & 개선사항



이진 분류

Train 재현율: 0.8355089844511726
Test 재현율: 0.7356708980840572



이진 분류

→ Train 재현율: 0.7825694927756734
Test 재현율: 0.7531868171484799

- LightGBM에서 Logistic Regression으로 바꿈으로써 모델 복잡도를 낮춤

⇒ 과적합 해결

분석 결과

Analyze : 최종 분석 결과



이진 분류

Train 재현율: 0.7825694927756734
Test 재현율: 0.7531868171484799

Review : 여기서 음료 두잔이랑 허니브레드 시켰는데 12시에 와서 4시반까지 공부했다고 눈치졌어요
Actual Score : Negative
Predicted Score : Negative

Review : 맛이 미치셨나요?여기 제 방앗간입니다 주변가면 못 지나치고 들려요
Actual Score : Positive
Predicted Score : Positive

분석 결과

Analyze : 개선사항

< 다중분류위해 필요한 점 >

1. 1 ~ 3 평점의 더 많은 데이터를 확보 -> 균등한 label들의 분포

2. 더 다양한 플랫폼들의 리뷰들을 크롤링

Analyze : 최종 분석 결과



우 리 모 델 의 예 측 평 점

4.2



1. 블루보틀 성수점

카페, 아메리카노
#차모임, #무료주차, #시끌벅적한

56점 | ★ 3.6 (58명) | ♡ 30



블루보틀 성수 카페

3.8 ★★★★★ 리뷰 2,315개 · ~~₩₩₩₩~~



A 블루보틀 성수 카페 카페

3.1 ★★★★★ 196건 | 리뷰 815



블루보틀 성수 카페 블루보틀

★ 4.4/5 · 방문자리뷰 16,653 · 블로그리뷰 3,639

A top-down view of four white ceramic coffee cups arranged in a square on a dark, textured wooden tray. Each cup contains a latte with intricate white foam art. The cups have dark blue handles. A semi-transparent, light beige circle is centered over the image, containing the text 'Thank You' and '감사합니다' in a dark, sans-serif font.

Thank You
감사합니다