쿠팡 리뷰 분석을 통한 스팸 리뷰 판단과 타 쇼핑몰과의 리뷰 생태계 비교

3조 도지
이윤재, 김정호, 박성하, 조민호

목차

Part 1,

프로젝트 개요

Part 1

프로젝트 개요

기획 배경

78.6%

제품 구매 시 항상
소비자 리뷰를
확인하는 편이다

58.4%

지인이 추천을 한 제품이라도
소비자 리뷰를
확인하고 구매한다

69.4%

소비자 평가가
부정적일 경우
해당 제품을 구매하지 않는다

소비자는 다른 소비자의 상품 리뷰에 영향을 많이 받는다

Part 1

프로젝트 개요

기획 배경

이브닝뉴스 조현용 기자

판치는 가짜

뉴스데스크

입력 2013-09-11 17:39 |

[오늘 이 뉴스] 돈 받고 "맛있어요"...가짜 리뷰 첫 실형 '철퇴'

10개 중 4개

입력 2021-05-26 20:41 | 수정 2021-05-26 20:42

[김성민의 실밸 레이더]
넘쳐나는 가짜 리뷰에
핵심 경쟁력이 장애물로

결국 A씨는 '배민 모니터링단'에 덜미가 잡혔고, 법원은 리뷰 조작에는 처음으로 징역 10개월의 실형을 선고했습니다.

Part 1

프로젝트 개요

기획 배경

In a statement, Amazon spokesperson Patrick Graham said customer trust is a top priority.

"In 2019, we invested over $500 million and have more than 8,000 employees protecting our store from fraud and abuse. We have robust proactive and reactive systems in place to protect our store and our customers." Graham also noted that the company uses a combination of machine learning tools and human investigators to analyze millions of reviews per week, with a goal of stopping fake reviews before they are published.

- 아마존 대변인은 고객과의 신뢰가 최우선이라고 말한다.
- 2019년 5억달러 이상과 8000명 이상의 자원이 거짓 탐지에 투자되었다.
- 머신러닝과 인력을 통해 매주 수백만개의 리뷰를 검사한다.
- 가짜 리뷰가 공개되기 전에 막는 것을 목표로 한다.

➢ 쿠팡 리뷰 데이터를 통한 스팸 리뷰 판단

Part 1

프로젝트 개요

목표

< 쿠팡 리뷰 데이터 분석을 통한 스팸 리뷰 판단 시스템 개발 >

1. 파생 변수 분석을 통한 스팸 리뷰 판단
2. 리뷰 유사도 분석을 통한 스팸 리뷰 작성자 판단

< 쿠팡 리뷰 판단 시스템을 사용하여 타 쇼핑몰 리뷰 판단 >

1. 쿠팡 리뷰 분석을 통해 만들어진 리뷰 판단 시스템을 사용한
   타 쇼핑몰 (11번가)의 리뷰 판단
2. 판단 결과를 통한 쿠팡과 타 쇼핑몰의 스팸 리뷰 생태계 차이 확인

Part 1

프로젝트 개요

구성원 및 역할

공통 역할

- 쿠팡 리뷰 크롤링, 추가 변수 생성

개인 역할

- 이윤재 (팀장) : 데이터 마트 구축, 단어 출현 빈도 기반 리뷰 분석 및 스팸 리뷰 판단
- 김정호 : 리뷰 생태계 분석을 위한 타 쇼핑몰 리뷰 크롤링
- 박성하 : 감성분석과 Clustering을 통한 리뷰 분석 및 스팸 리뷰 판단
- 조민호 : 리뷰 유사도를 통한 스팸 작성자 판별

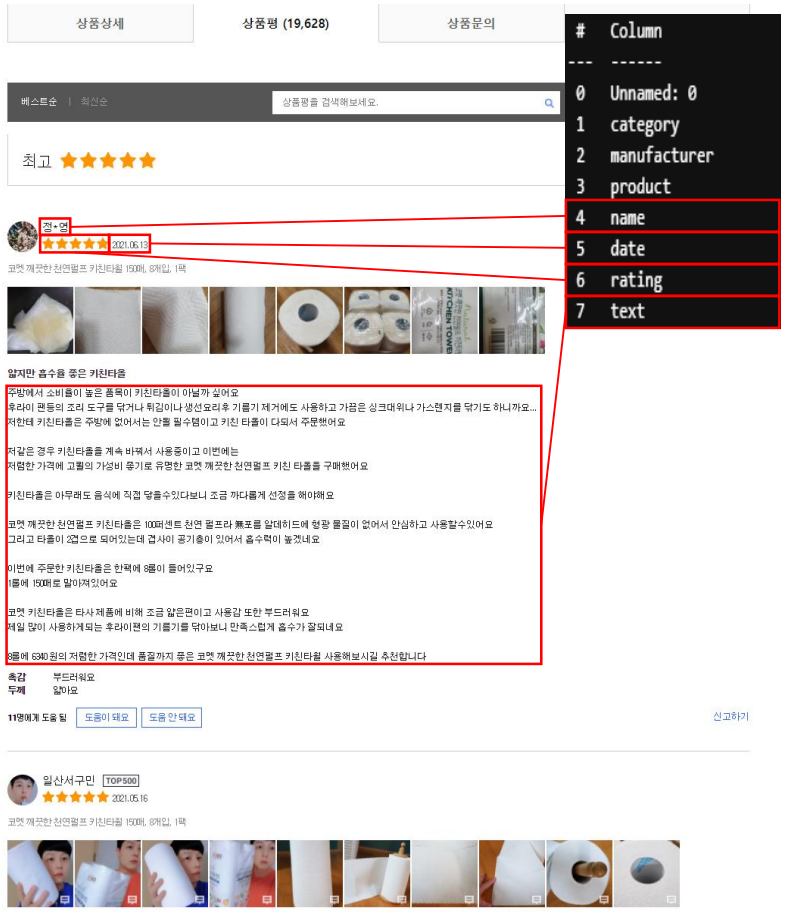
Part 2,

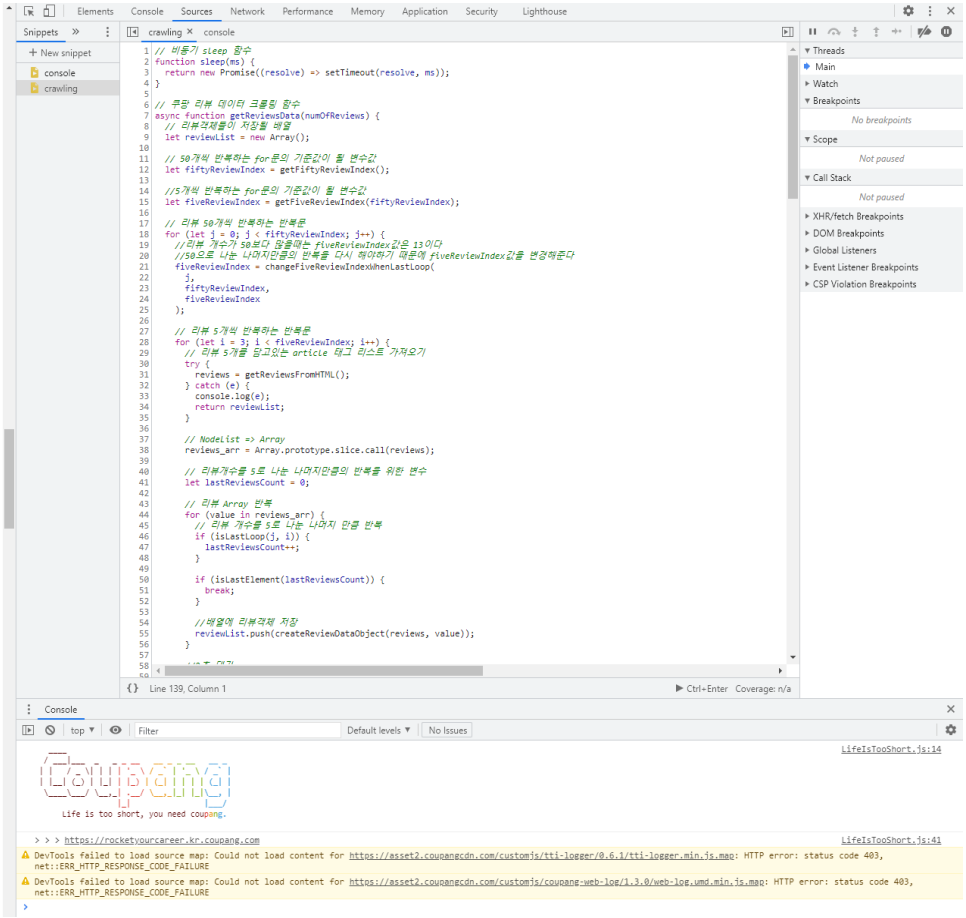
프로세싱

Part 2

프로세싱

데이터 수집

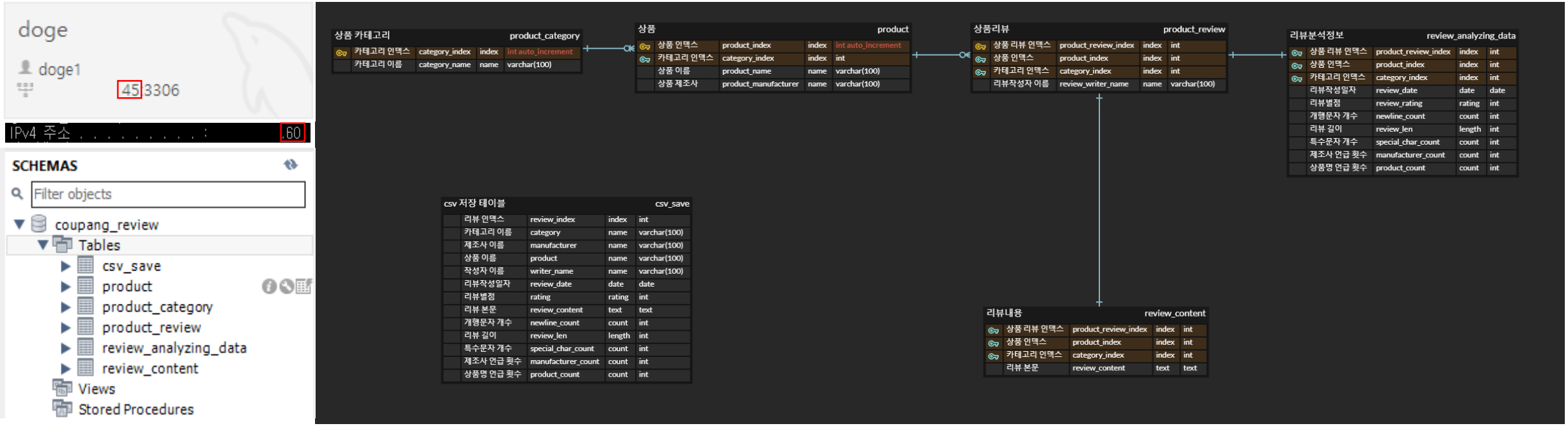




- 크롤링 위한 JavaScript Code 작성후 Chrome 개발자 도구 내 Snippets 기능 사용
- 작성자 명, 내용 등 크롤링
- 카테고리 별 약 10개 제품 1~5점 리뷰 최대 500개

프로세싱

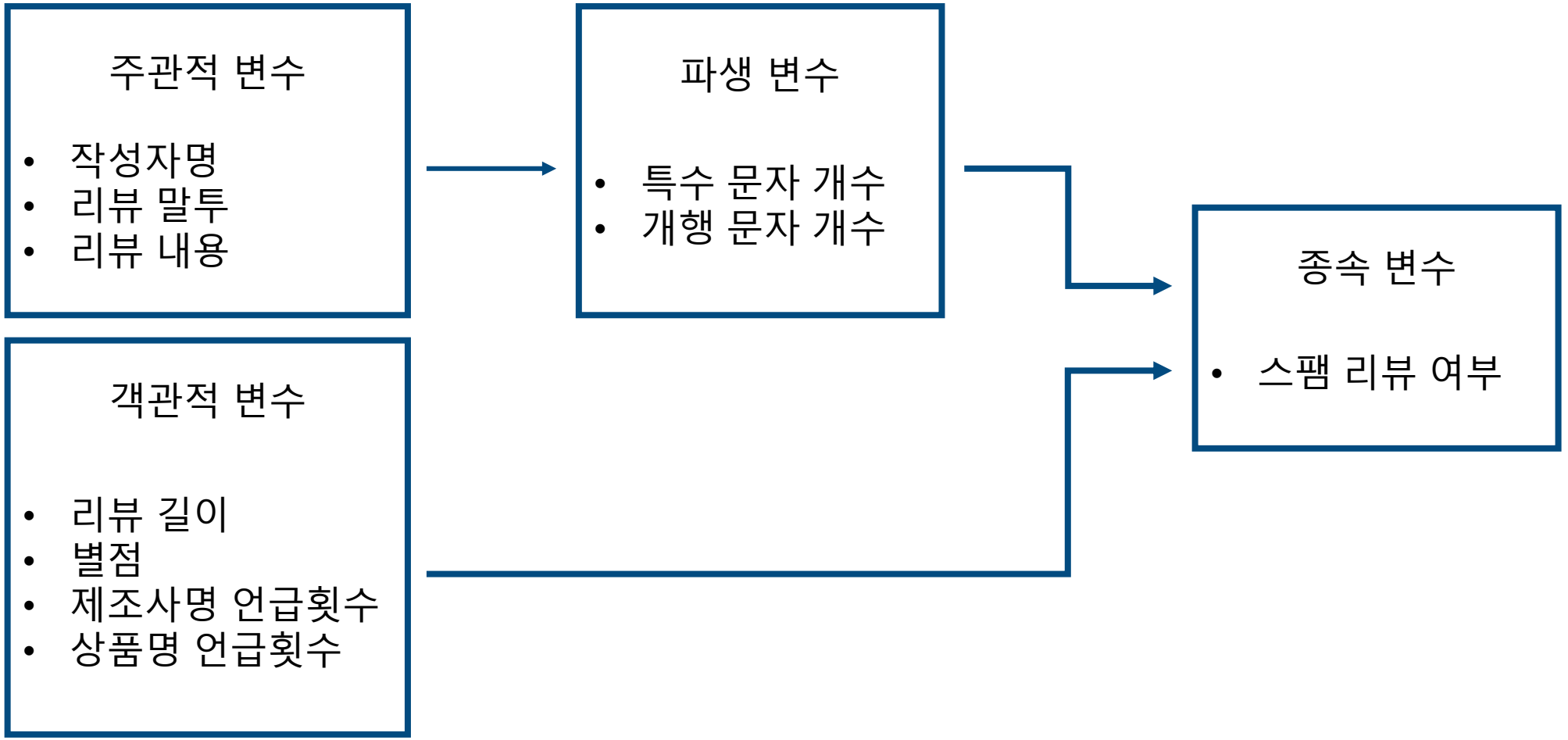
데이터 마트 구축



- 특정 PC DB에 데이터 마트 구축 후 해당 DB Host 지정
- Host로 부터 계정 발급 후 해당 DB로 접속하여 데이터 사용

프로세싱

연구 모형


주관적 변수
작성자명
리뷰 말투
리뷰 내용
파생 변수
특수 문자 개수
개행 문자 개수
객관적 변수
리뷰 길이
별점
제조사명 언급횟수
상품명 언급횟수
종속 변수
스팸 리뷰 여부

프로세싱

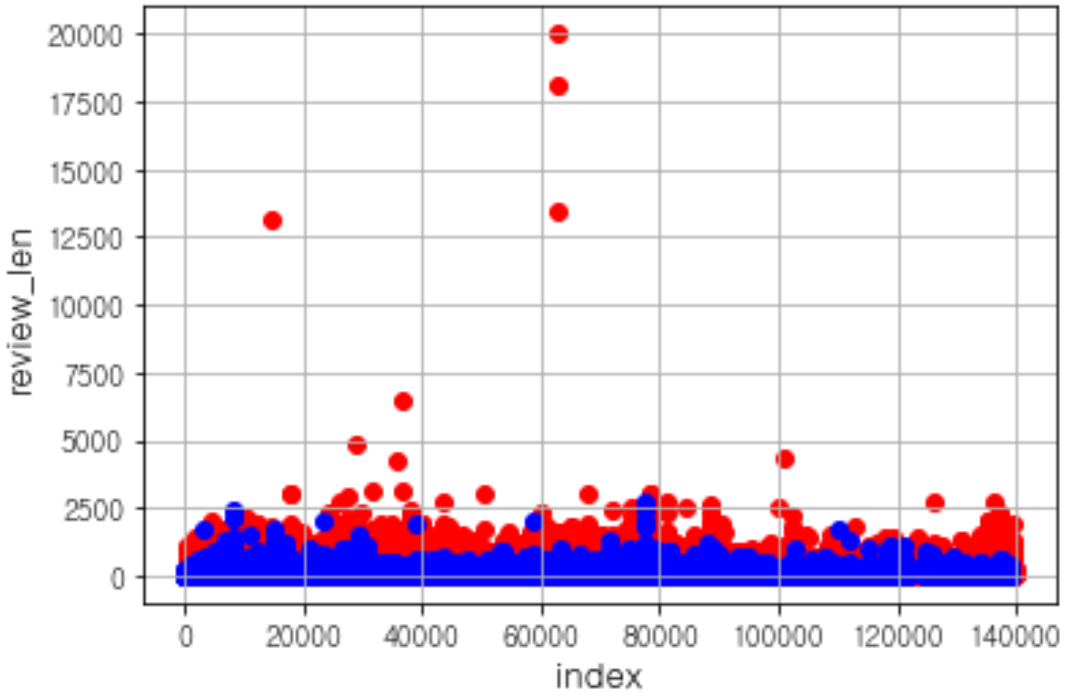
데이터 변수

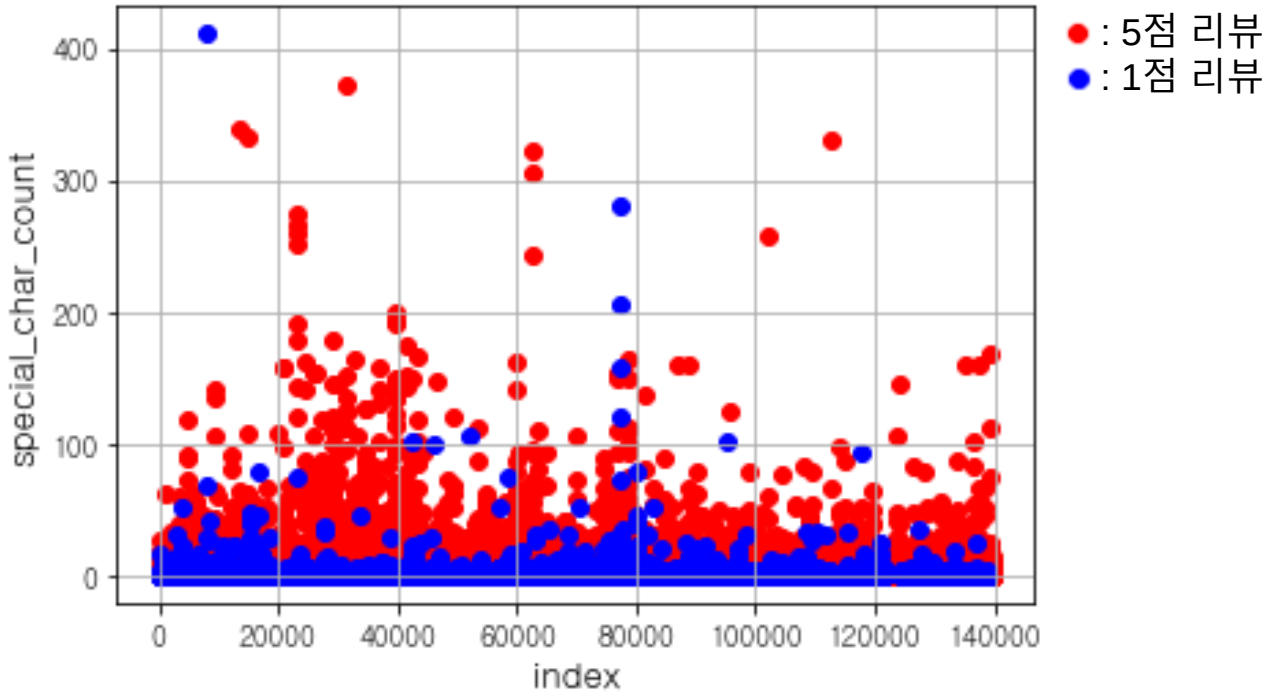
기존 변수	변수 설명
review_index	리뷰 고유 ID
category	쿠팡 내 카테고리
manufacturer	제조업체명
product	제품명
writer_name	리뷰 작성자 명
review_date	리뷰 작성 일자
rating	리뷰 별점
review_content	리뷰 내용

파생 변수	변수 설명
newline_count	개행 문자 개수
review_len	리뷰 길이
special_char_count	특수 문자 개수
manufacturer_count	제조업체명 언급 횟수
product_product	제품명 언급 횟수

프로세싱

데이터 분석 : 1. 단어 출현 빈도 기반





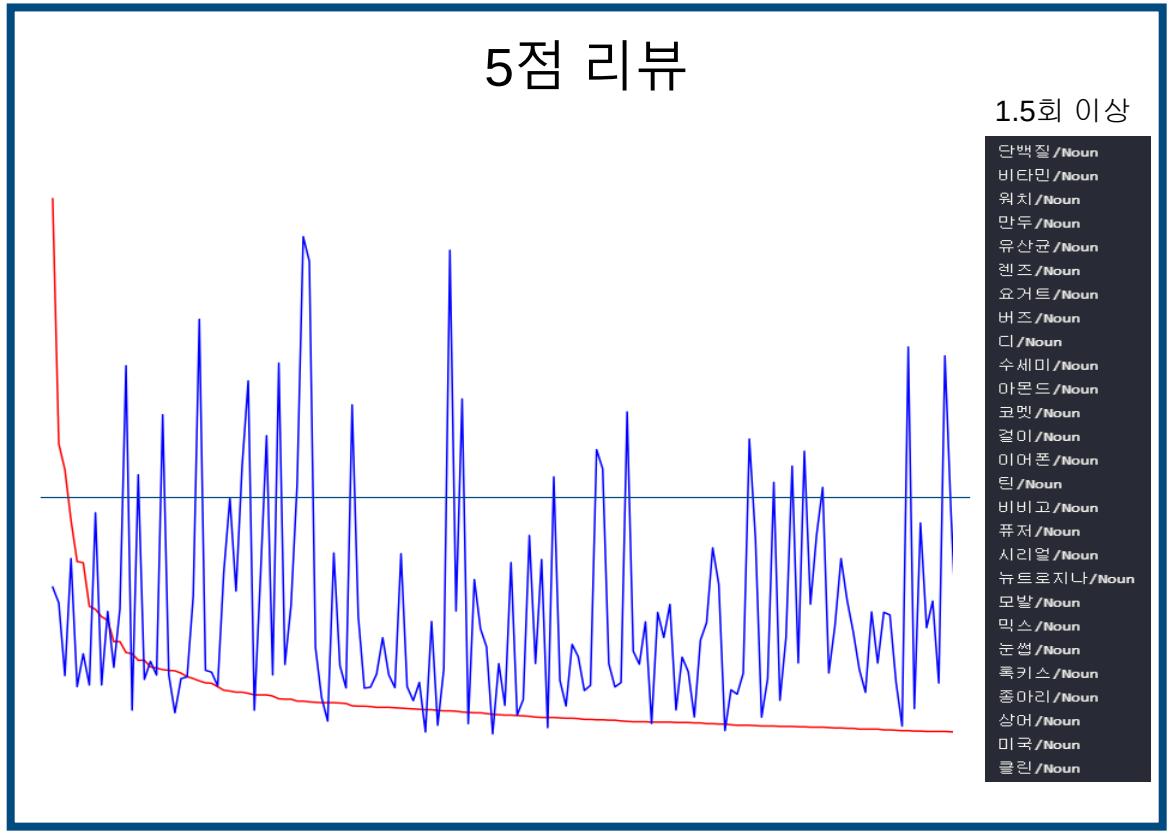
- 5점 리뷰가 1점리뷰보다 리뷰의 길이가 더 긴 양상을 보임
  → 정보 전달을 위한 스팸 리뷰가 포함되어 있을 가능성 多

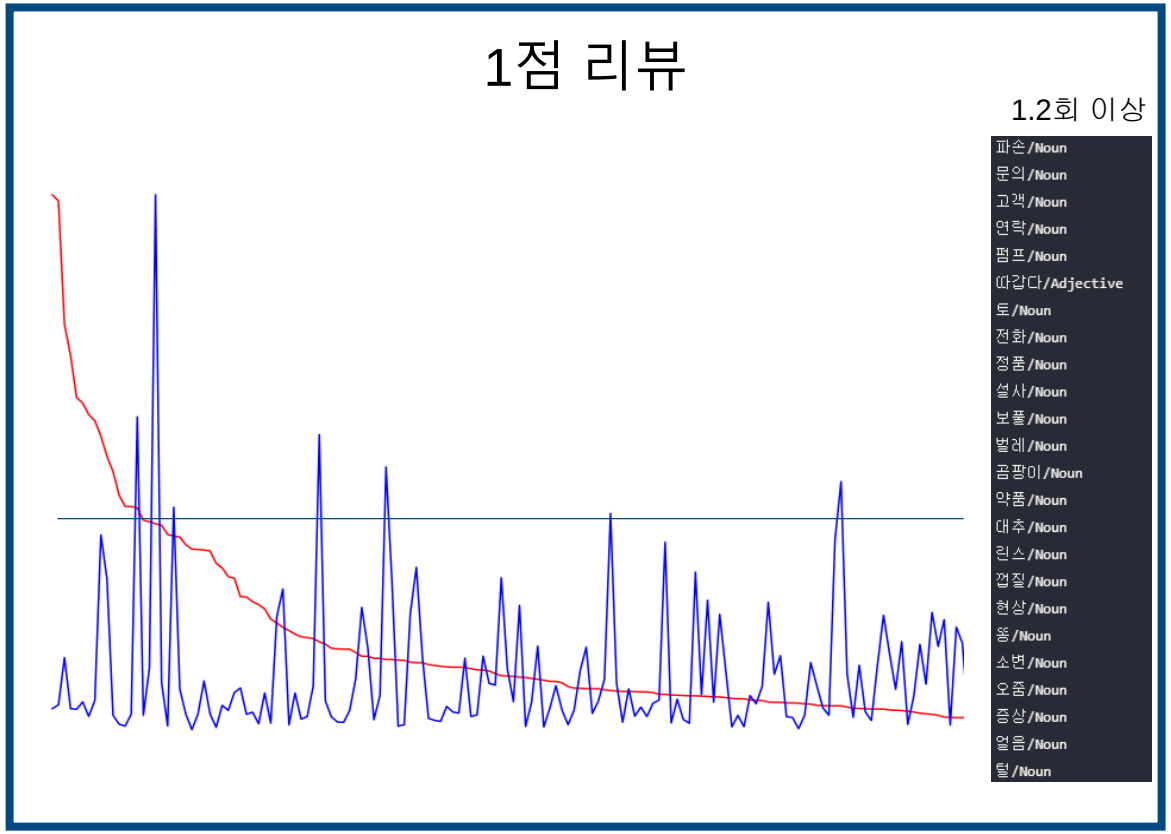
- 5점 리뷰가 전체적으로 특수문자를 많이 포함하고 있음

Part 2

프로세싱

데이터 분석 : 1. 단어 출현 빈도 기반

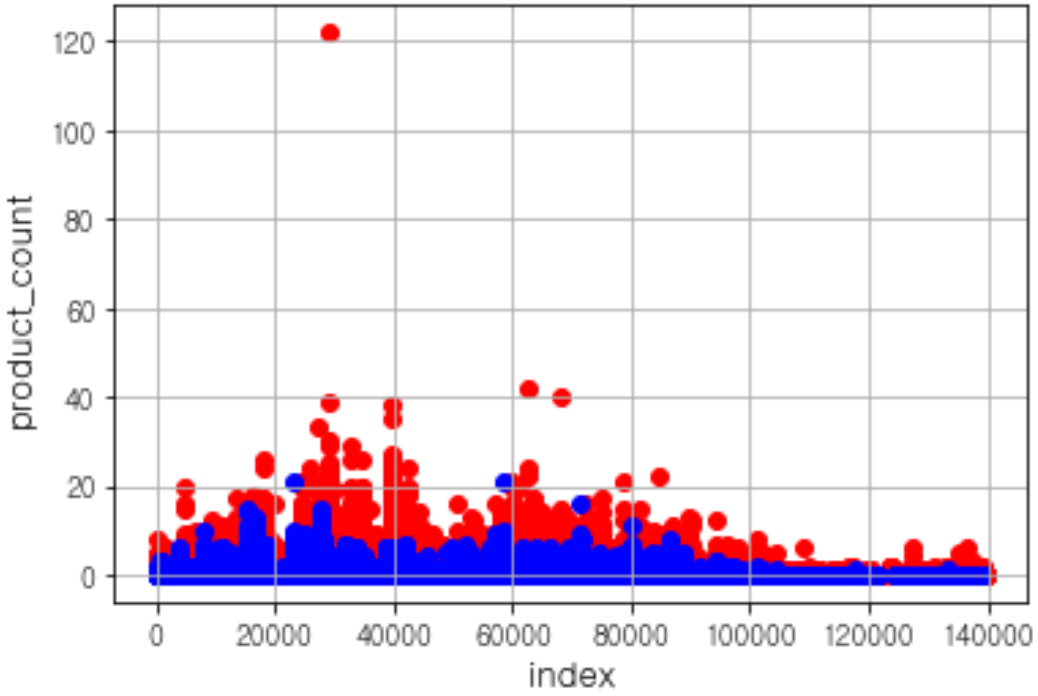


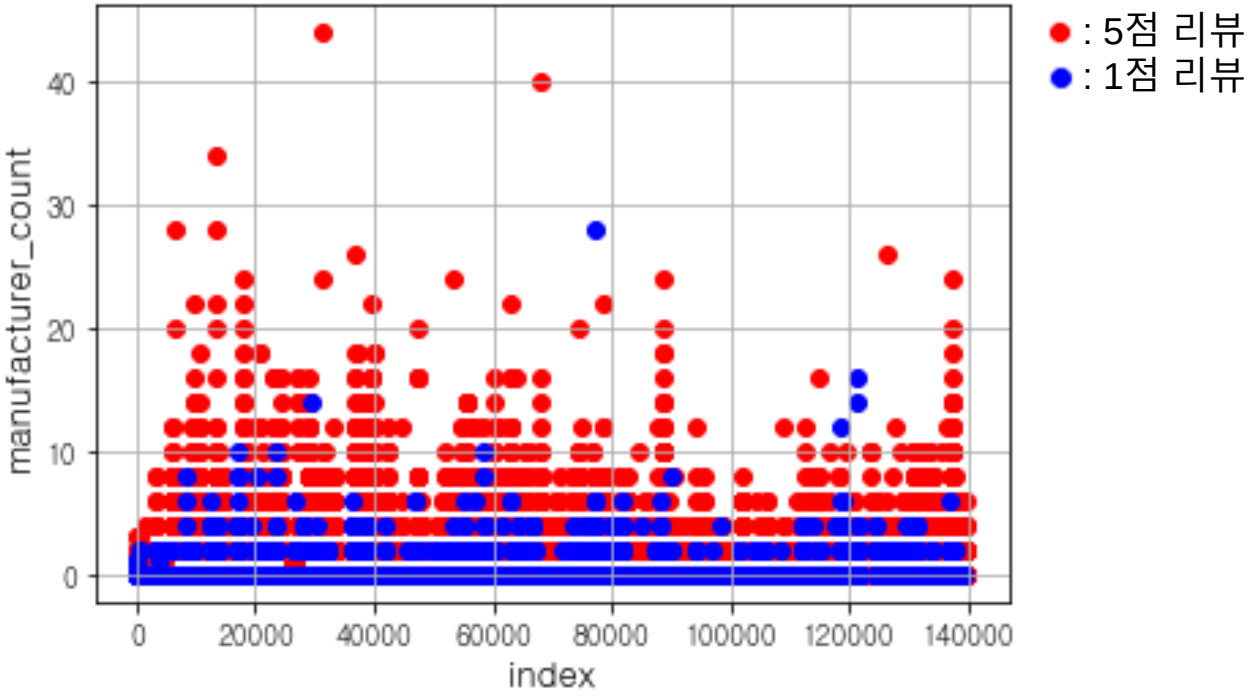


- 5점 리뷰에서는 제품명과 제조사명, 1점 리뷰에서는 단순 부정적 단어 다수 출현

프로세싱

데이터 분석 : 1. 단어 출현 빈도 기반

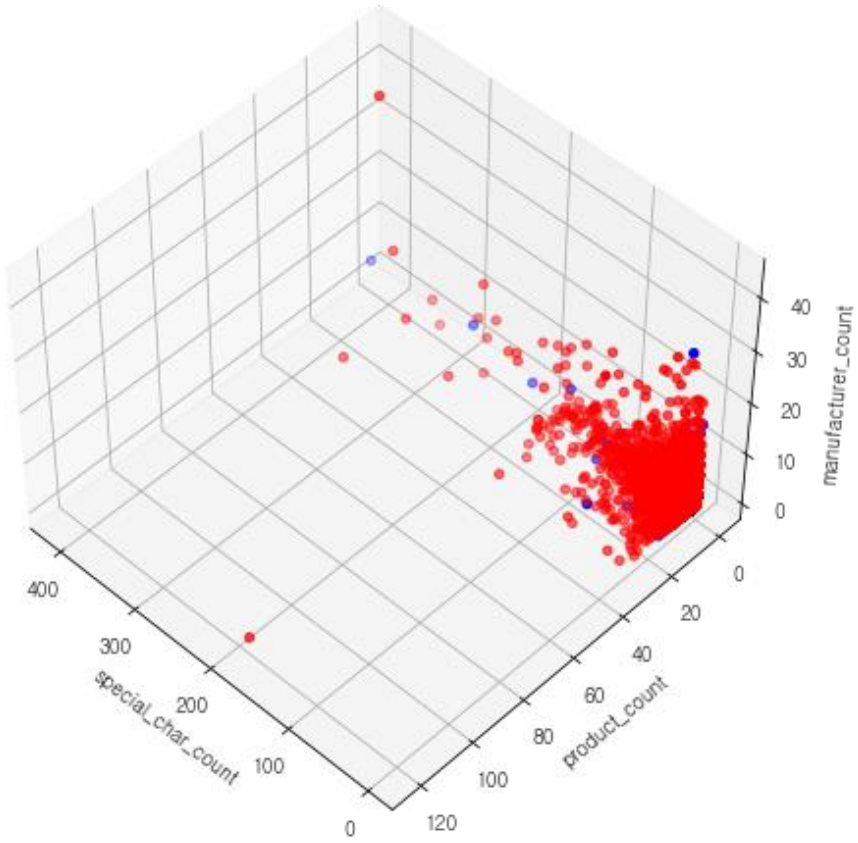




- 상품명 언급횟수와 제조사명 언급횟수 역시 1점리뷰보다 5점리뷰에서 크게 앞서
  결론적으로 유의미한 지표라는 것을 확인

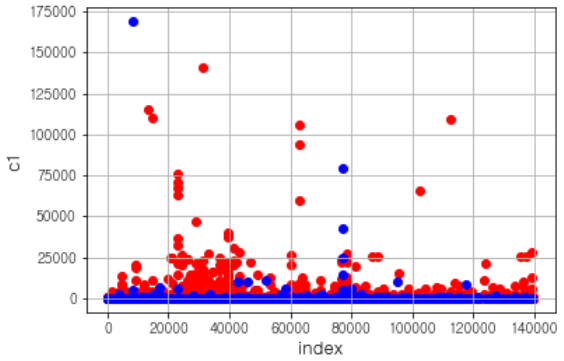
프로세싱

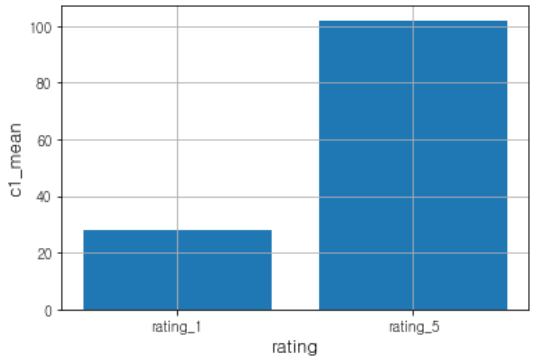
데이터 분석 : 1. 단어 출현 빈도 기반



특수 문자 개수 ✚ 제조사 언급 횟수 ✚ 제품명 언급 횟수

c1





프로세싱

데이터 분석 : 2. 감성 분석 기반 Clustering

- KoNLPy

오늘 받았습니다.
컬러는 정말 예쁜데 가루날림이 너무 심해요.

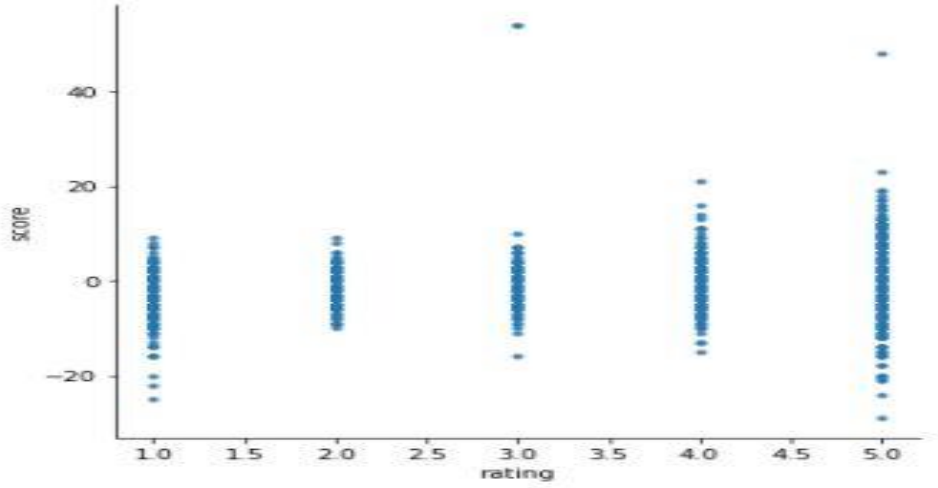
"컬러" "정말" "예쁘다" "가루날림" "너무" "심하다"

- KNU 한국어 감성사전

속성/감성어
컬러: 예쁘다(+1), 환상적이다(+2), …
가루날림: 없다(+1), 있다(-1), 심하다(-2), …

강조부사
정말(+2), 너무(+2), …

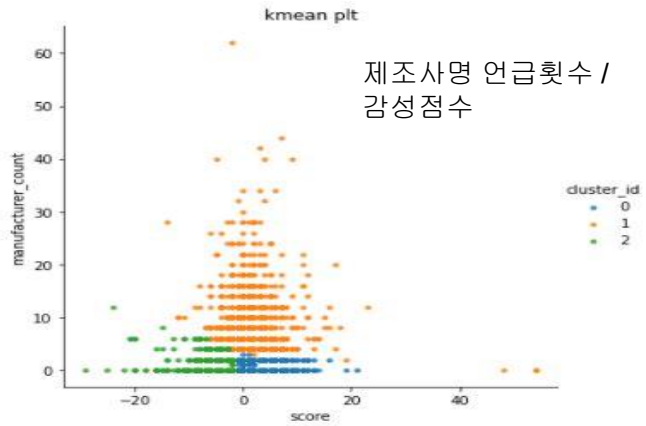
속성	관련 단어
컬러	"정말"(+2) "예쁘다"(+1)
가루 날림	"너무"(+2) "심하다"(-2)



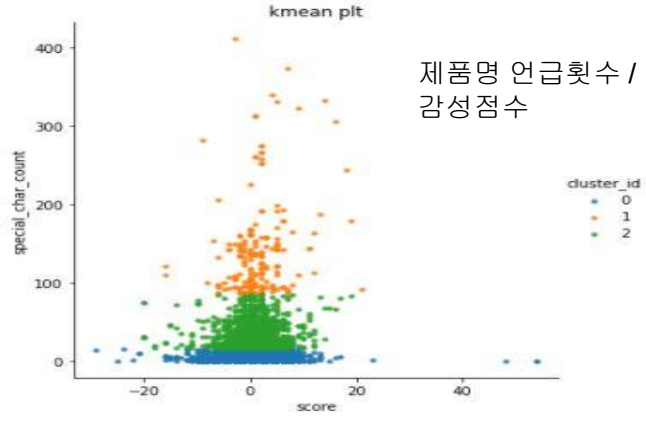
- 별점이 높을수록 감성 점수가 높은 추세를 보임
- 높은 별점에서도 다수의 낮은 감성 점수 검출
  "부정적인 상황에서 해당 제품으로 긍정적으로 변화" ,
  "부정적인 리뷰에 걱정했으나 긍정적" 등의 리뷰로 판단

프로세싱

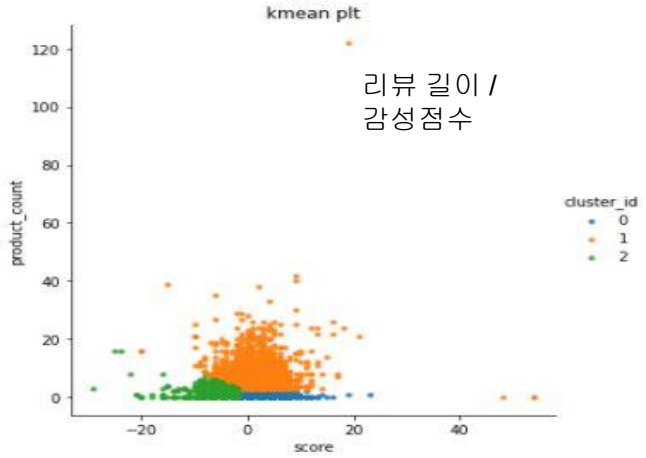
데이터 분석 : 2. 감성 분석 기반 Clustering



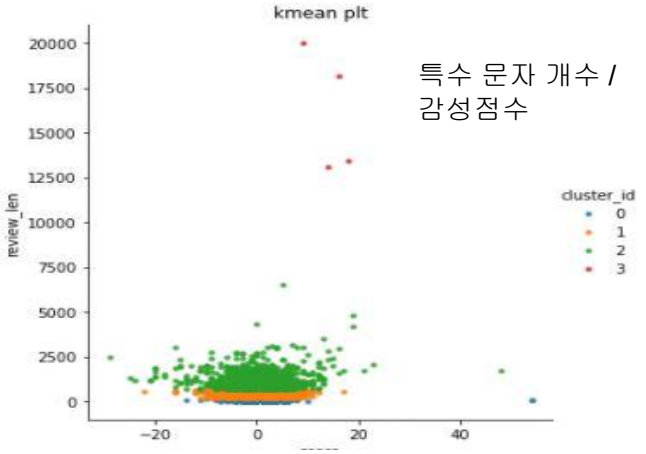
제조사명 언급횟수 / 감성점수



제품명 언급횟수 / 감성점수



리뷰 길이 / 감성점수



특수 문자 개수 / 감성점수

Clustering 경계값

- 감성점수 > 0
- 제조사명 언급횟수 > 3
- 제품명 언급횟수 > 3
- 리뷰 길이 > 300
- 특수 문자 개수 > 10

프로세싱

데이터 분석 : 3. 리뷰 유사도 기반

➢ KorSTS new benchmark

- kakao brain에서 제공한 한국어 문장 유사도 데이터
- 한국어 문장 유사도 분석을 위한 새로운 benchmark
- 영어 문장 번역본을 포함한 8600여 개의 데이터
- 각 데이터 마다 2개의 문장과 0~5점으로 구성

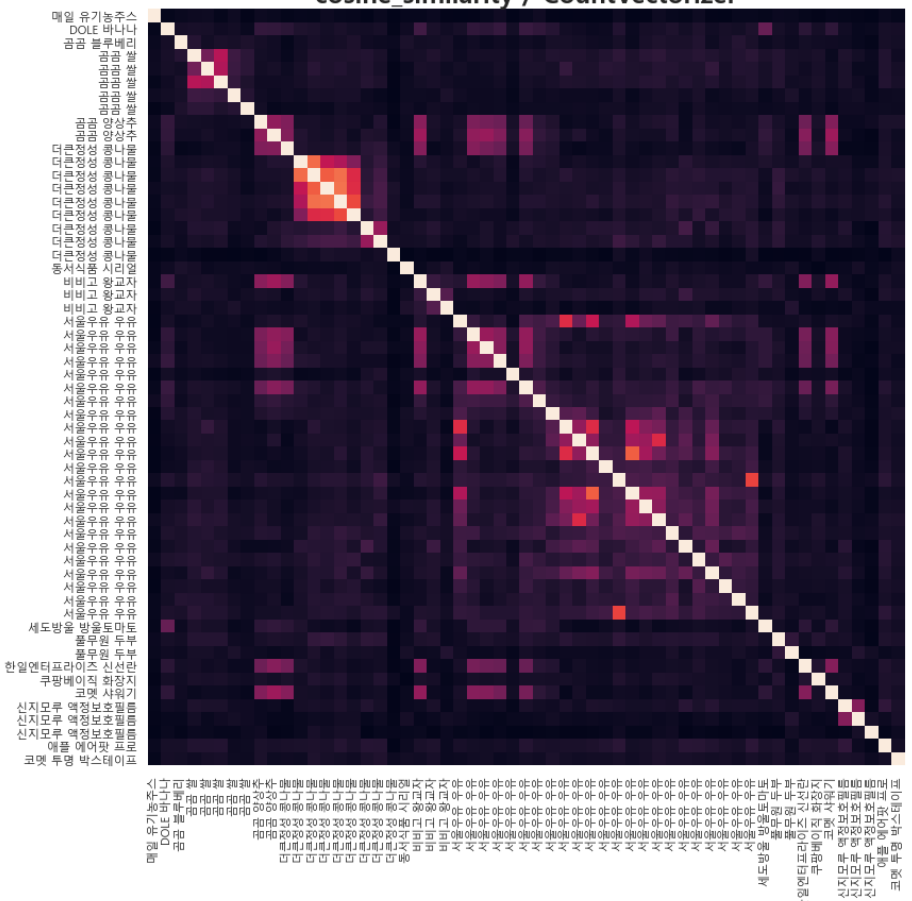
5.000	비행기가 이륙하고 있다.	비행기가 이륙하고 있다.
3.800	한 남자가 큰 플루트를 연주하고 있다.	남자가 플루트를 연주하고 있다.
3.800	한 남자가 피자에 치즈를 뿌려놓고 있다.	한 남자가 구운 피자에 치즈 조각을 뿌려놓고 있다.
2.600	세 남자가 체스를 하고 있다.	두 남자가 체스를 하고 있다.
4.250	한 남자가 첼로를 연주하고 있다.	자리에 앉은 남자가 첼로를 연주하고 있다.
4.250	몇몇 남자들이 싸우고 있다.	두 남자가 싸우고 있다.
0.500	남자가 담배를 피우고 있다.	남자가 스케이트를 타고 있다.
1.600	남자가 피아노를 치고 있다.	남자가 기타를 연주하고 있다.
2.200	한 남자가 기타를 치고 노래를 부르고 있다.	한 여성이 어쿠스틱 기타를 연주하고 노래를 부르고 있다.
5.000	사람이 고양이를 천장에 던지고 있다.	사람이 고양이를 천장에 던진다.
4.200	그 남자는 다른 남자를 막대기로 때렸다.	그 남자는 다른 남자를 막대기로 때렸다.
4.600	한 여성이 아기를 안아서 캥거루를 안는다.	한 여성이 아기를 안아서 팔에 캥거루를 안는다.
3.867	남자가 플루트를 연주하고 있다.	남자가 대나무 플루트를 연주하고 있다.
4.667	사람이 종이 한 장을 접고 있다.	누군가가 종이를 접고 있다.
1.667	한 남자가 도로를 달리고 있다.	판다 개가 도로에서 달리고 있다.
3.750	개가 베이컨을 등에서 떼려고 하고 있다.	개가 등에 있는 베이컨을 먹으려고 하고 있다.
5.000	북극곰이 눈 위에서 미끄러지고 있다.	북극곰이 눈 위로 미끄러져 가고 있다.
0.500	여자가 글을 쓰고 있다.	여자가 수영을 하고 있다.

Part 2

프로세싱

데이터 분석 : 3. 리뷰 유사도 기반

cosine_similarity / CountVectorizer



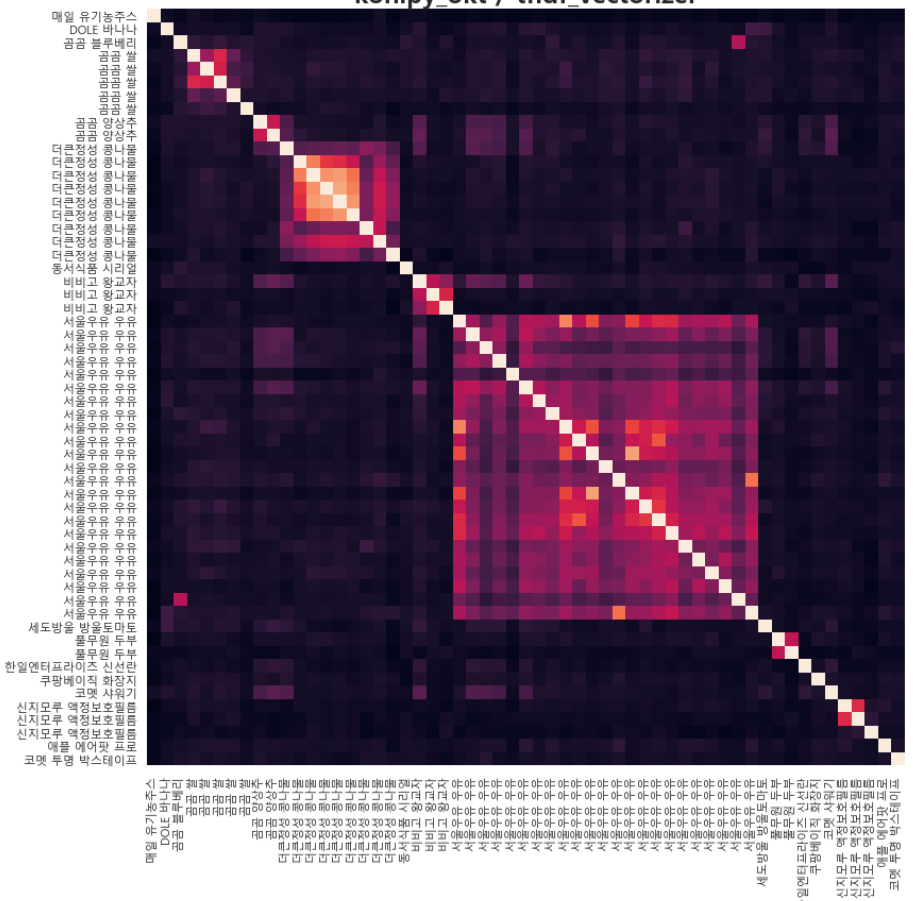
코사인 유사도 검사

- 단순 token화
- 단어 출현 횟수 count

형태소 분석기 사용
유사도 검사

- 형태소 별로 token화
- 출현 빈도 높은 단어 패널티 부여

konlpy_okt / tfidf_vectorizer



프로세싱

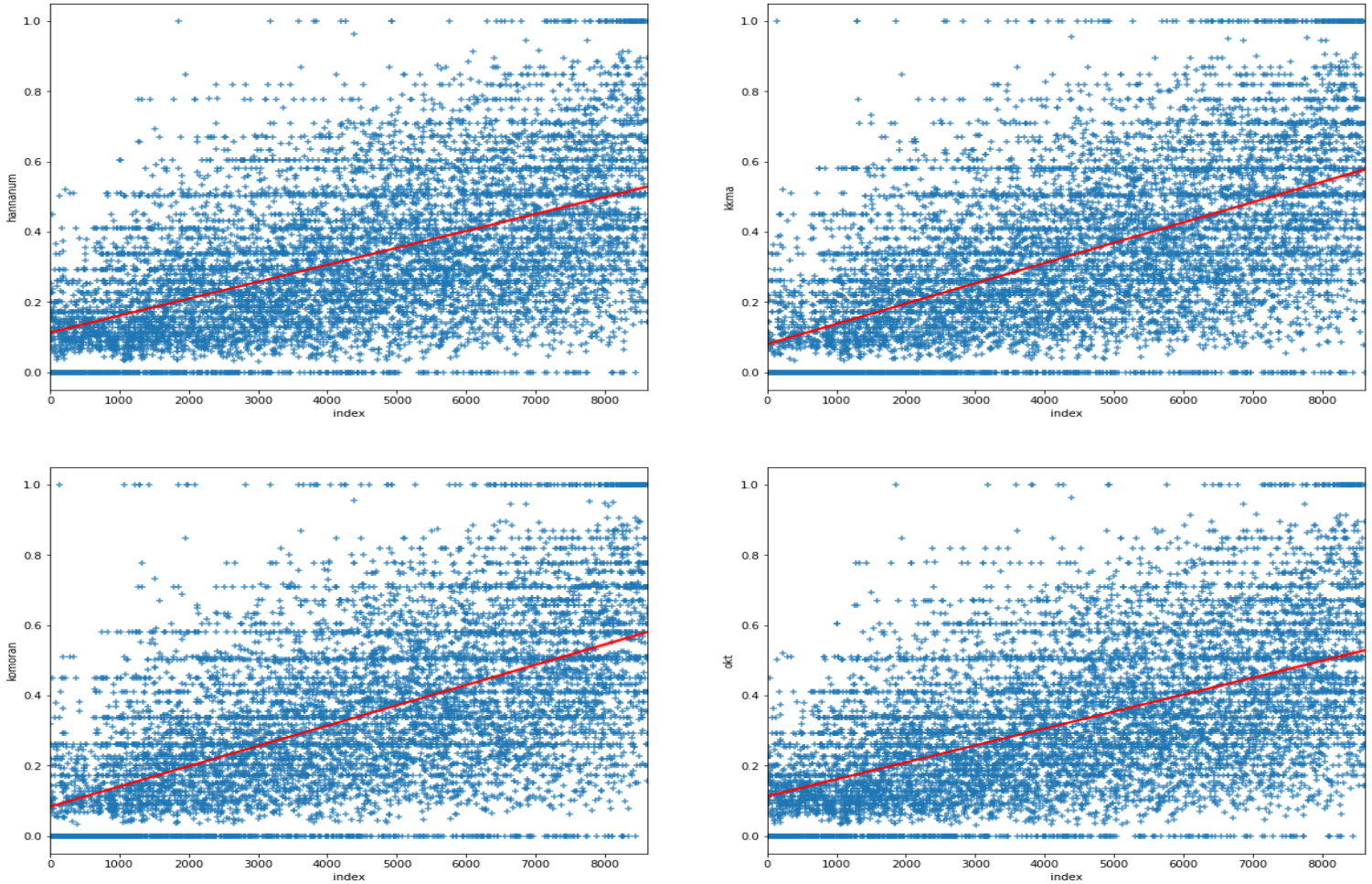
데이터 분석 : 3. 리뷰 유사도 기반

3.75	레드 더블 데커 버스	후드가 달린 빨간색 더블 데커 버스
3.75	흑백 고양이가 담요 위에 누워 있다.	흑백 고양이가 고양이 침대에 누워 있다.
3.765	여자가 반죽을 굴리고 있다.	한 여성이 밀가루를 뿌리고 반죽을 굴리고 있다.
3.769	Teya Ryan은 CNN에서의 최신 최고급 셔플에서 미국 프로그래밍의 총책임자로 자리를 옮길 것이라고 이 네트워크는 발표했다.	CNN의 최근 시청률을 회복시키기 위한 움직임에서, 이 네트워크는 미국 프로그래밍의 총책임자인 테야 라이언을 축출했다.
3.777777778	우리가 백인과 흑인이 %라고 가정해보자.	'백색과 % 검정색'이 있는 곳을 상상해 보세요.
3.777777778	우리가 백인과 흑인이 %라고 가정해보자.	'백색과 % 검정색'이 있는 곳을 상상해 보세요.
3.786	그러나 2.6.0 커널의 상업적 사용은 대부분의 고객에게 여전히 몇 달이 걸리고 있습니다.	주요 리눅스 유통업체에 의한 2.6 커널의 상업적 출시는 여전히 몇 달 남았다.
3.8	티베트 수도사는 중국 통치에 항의하여 불을 지른 후 사망한다.	10대 티베트 수도승은 중국에서 불을 지른 후 사망한다.
3.8	그리고 공격에 대해 후회 나 죄책감을 느꼈는지 묻는 질문에 그의 대답은 "아니오"라는 확고한 대답이었다.	10월 12일 공격에 대해 후회를 느꼈는지 물었습니다. 대답은 "아니오"라는 확고한 대답이었습니다.
3.8	1994년 쇼 시작 이후 존 카터 박사를 연기한 배우 노아 와일은 2004-2005시즌을 통해 쇼와 계약을 연장했다.	1994년 시리즈가 시작된 이래로 존 카터 박사를 연기해온 노아 와일은 프로듀서 워너 브라더스와의 계약에 1년 연장을 사인...
3.8	게다가, 1차 진료 신탁(PCT)은 올해 처음으로 스타 등급을 받았다.	1차 진료 신탁과 정신 건강 신탁 또한 올해 처음으로 공식적으로 평가되었다.
3.8	29세의 레이 브렌트 마시는 여러 건의 장례 사기 혐의를 받고 있으며, 거짓 진술, 사체 남용, 절도 혐의를 받고 있다.	29세의 레이 브렌트 습지는 또한 신체 학대와 절도 혐의를 받고 있다.

- 3.8점보다 점수가 낮은 경우에도 의미상 유사한 문장 데이터 다수 포함
- 유사도 기준을 엄격하게 하기위해 임계값 3.8점으로 설정

프로세싱

데이터 분석 : 3. 리뷰 유사도 기반



- KoNLPy 내부
  4개의 형태소 분석기 비교
- kkma와 komoran에 비해
  hannanum과 okt 형태소 분석기가
  더 나은 결과를 보여줌
- 오타에 대해 비교적 잘 처리한다고
  알려진 okt 형태소 분석기를 채택

프로세싱

데이터 분석 : 3. 리뷰 유사도 기반

- 회귀선의 3.8점 지점을 확인한 결과 유사도 값 0.41을 확인
- okt 형태소 분석기를 사용한 결과 유사도가 0.41이상인 경우 3.8점 이상이라고 간주

해당 작성자 스패머 판단

Part 2

프로세싱

데이터 분석 결과

➢ 단어 출현 빈도 기반

- 쿠팡 : 139757개 리뷰 중 12911개 (9.24%)
- 11번가 : 5320개 리뷰 중 338개 (6.35%)

➢ 감성 분석 기반 Clustering

- 쿠팡 : 139757개 리뷰 중 238개 (0.17%)
- 11번가 : 5320개 리뷰 중 21개 (0.39%)

➢ 리뷰 유사도 기반

- 쿠팡 : 5172명 중 925명 (17.97%)
- 11번가 : 935명 중 11명 (1.18%)

Part 2

프로세싱

데이터 분석 결과

- 단어 출현 빈도 기반

"곰곰거 가격만 저렴하고 품질이나 맛은 별로~ " 라는 선입견을 없애주는 제품....
샤워 후에도\n외출 할 때 핸드 로션으로\n얼굴 로션이 떨어지면\n화장 전 얼굴...
★ ● 자연스러운 눈썹 펜슬 !! 쌩얼에도 만족 ● ★ ...
쿠팡의 시장 침투력은 어마어마 하네요.\n\n더블에이 밀크 사기에는 돈이 아깝고\n...

- 감성 분석 기반 Clustering

뉴트로지나 딥클린 포밍 클렌저, 100g, 3개입\n2020.04.08 배송\n\n...
이야..\n다들 향이 좋다고하셔서\n얼마나 좋은가 하고 시켰는데\n상상이상이네요 ㅋ...
— 록키스 티트리 필링젤 . 120 ml —\n\n\n— 사용하고 나니,...
▶상품명\n록키스 티트리 필링젤\n\n▶용량\n120ml×1개\n\n▶유효기간...

- 리뷰 유사도 기반

안녕하세요, 쿠팡 리뷰어 일산서구민 입니다.\n\n\n그래요, 또 접니다..\n\n...
안녕하세요, 쿠팡 리뷰어 일산서구민 입니다.\n\n\n두달 반 만에, 또 재구매 하...
안녕하세요 쿠팡 이용자 여러분\n쿠팡 리뷰어 일산서구민 입니다.\n\n\n고소하고 ...
안녕하세요 쿠팡이용자 여러분\n쿠팡리뷰어 일산서구민 입니다.\n\n\n빙그레 바나나...

★★★★★ 100% 내돈내산 후기입니다.~^^ 비비고 왕교자(냉동) 1.4kg x 1개~! ●구매동기.. 만두를 왜 만두라고 하는지 아시나요? ...
★★★★★ 100% 내돈내산 후기입니다.~!!! ●구매동기.. 집에서나 캠핑장에서 간식과 술안주로 먹기에 대박 맛나는 비비고 왕교자(냉...
★★★★★ 대한민국 대표만두라고 할 정도로 만두 매니아들은 다 알만한 만두~!!! 비비고 왕교자 만두를 또 다시 맛보네요~^^ ●구매...
두께도 좋구요~ 깨끗하게 잘 닦입니다~^^ 가성비도 짱 입니다~^^
★★★★★ 100% 내돈내산 후기입니다.~!!! "쿠팡베이직 네추럴 3겹 롤 화장지~!" ●구매동기.. 화장실용 롤화장지를 많은 브랜드를 써...
★★ 100% 내돈내산 후기입니다.~!!! ●구매동기.. 차량용 실내향수가 필요하여, 이것저것 찾아보다가 코멧 설레는 향 가득 차량용 디퓨저...

Part 3,

개선 사항 및 기대효과

Part 3

개선 사항 및 기대효과

한계점 및 개선사항

- 단어 출현 빈도 기반
  - c1 계산식에서의 각 feature 가중치에 대한 고찰
  - 필터링 기준이 된 이상치 값의 타당성
- 감성 분석 기반 Clustering
  - KNU 감성 사전의 한계점 : 강조어구에 대한 판단 불가
  - 전체 데이터 양에 비해 극도로 적은 필터링 결과 리뷰 수
- 리뷰 유사도 기반
  - 작성자 IP가 아닌 이름으로 구별하여 익명 처리된 다른 작성자를 같은 작성자로 간주 가능 (ex) 김*규, 이*민)
  - 데이터 내부 동일 작성자 리뷰 2개 미만인 경우 적용 불가능

Part 3

개선 사항 및 기대효과

기대효과

< 쿠팡 내 스팸리뷰 일부 필터링 가능 >

➢ 스팸 리뷰 처리가 미흡한 쿠팡 내 제품 구매시 리뷰 필터링을 통한 제품 선택

< 스팸리뷰 판단 미실시 소규모 쇼핑몰에 적용 가능 예상 >

➢ 스팸 리뷰 필터링 시스템이 잘 갖춰지지않은 소규모 쇼핑몰에 적용 가능

< 쇼핑몰 외 온라인 구매 플랫폼에 확장 적용 가능 >

➢ 연구 데이터 변화를 통한 다양한 플랫폼 적용

Part 4,

개발후기 및 느낀점

개발후기 및 느낀점

개발후기 및 느낀점

- 이윤재
  - 일반적인 방법으로 크롤링이 안되는 상황과 논리적 결함을 고치는 과정에서 다양한 시도를 해볼 수 있었기 때문에 좋은 경험이 되었다.
- 김정호
  - 많은 것을 배우고 많은 부족함을 알게 되어서 남은 교육 기간동안 더 열심히 하겠다.
- 박성하
  - 기본기가 부족해 처음에 헤매었지만 시행착오를 겪으며 많이 배울 수 있는 좋은 기회가 되었다.
- 조민호
  - 팀원들과 각자 다양한 방법으로 분석을 하면서 나의 방법과 다른 팀원의 방법을 비교해보며 다양한 지식을 얻을 수 있어서 좋았고 쉽지 않은 주제였지만 목표한 만큼의 결과가 나와서 뿌듯하다.

References 및 데이터 출처

- 윤덕환, 채선애, 송으뜸, 김윤미. "2017 소비자 리뷰 영향력 조사." 리서치보고서, 2017.7 (2017): 1-28.
- 오하영. "동시출현 단어분석 기반 스팸 문자 탐지 기법." 정보보호학회논문지 26.3 (2016): 693-700.

- 쿠팡(Coupang). (https://www.coupang.com/)
- 11번가. (https://www.11st.co.kr/main)
- kakaobrain. KorNLUDatasets (https://github.com/kakaobrain/KorNLUDatasets/tree/master/KorSTS)

Q & A