

PROJECT1

OLIST E-COMMERCE DATA EDA 프로젝트

프로젝트 개요 및 데이터셋 소개

프로젝트 목적: E-COMMERCE 셀러를 위한 통합 셀링 현황 보고서 제작

데이터셋 소개 및 선정 이유:

KAGGLE에 있는 E-COMMERCE 데이터셋 중

- 10만 건에 이르는 주문 데이터(필요한 데이터 남기고 결측치 제거 후 96455건) 존재
- 아이템, 셀러, 고객, 지불 수단, 리뷰 데이터 등 E-COMMERCE에서 핵심적인 데이터가 갖춰져 있음



olist 소개:

브라질의 이커머스 솔루션 업체로, 온라인 판매를 원하는 사업자가 고객

판매 사이트 개설, 상품 자동 분류, ERP, 모니터링 대시보드, 배송 현황 관리, 데일리 이슈 리포트 등 제공

진행 환경 및 분석, 시각화 도구:

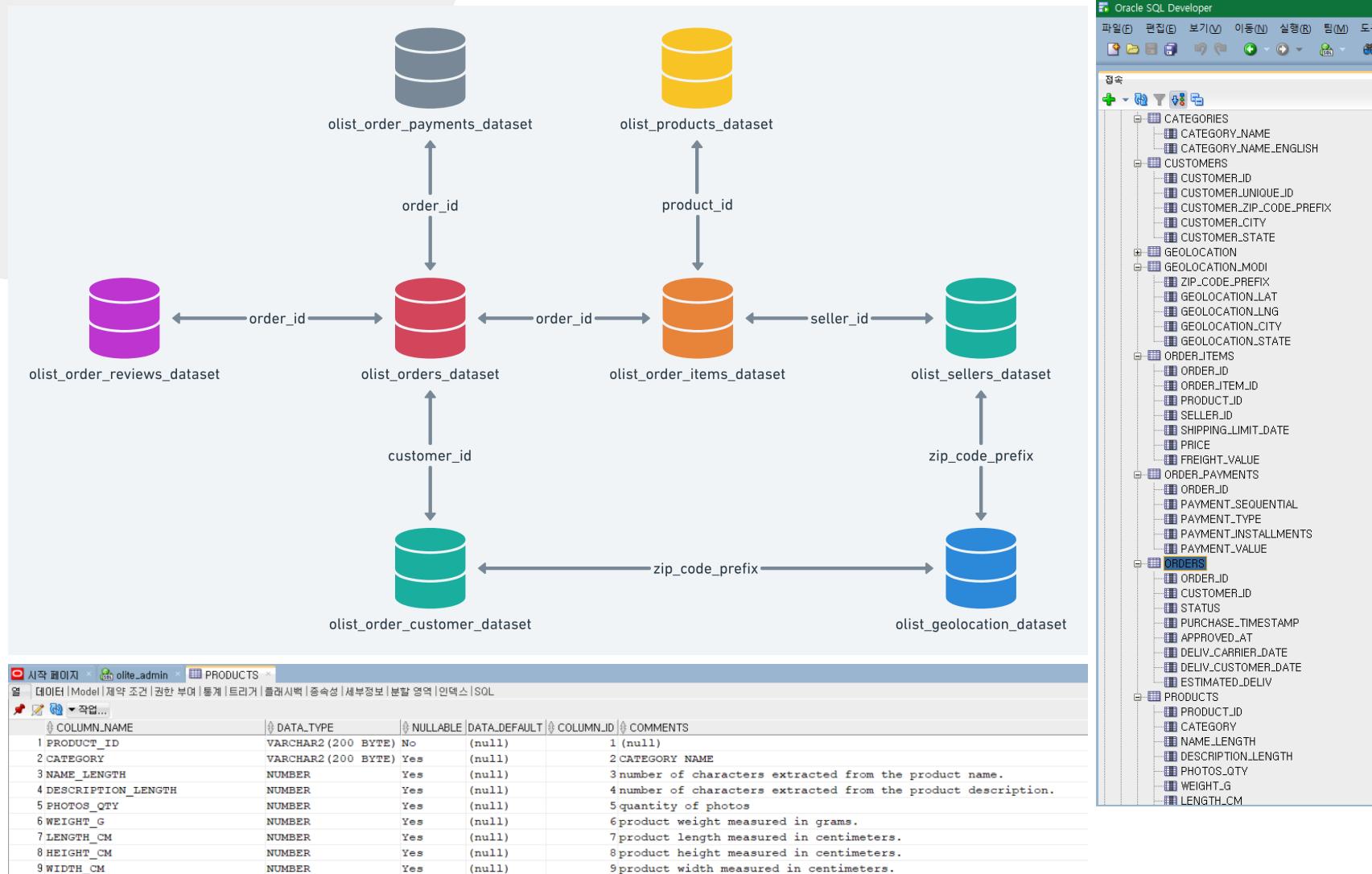
JUPYTER NOTEBOOK, NOTION / PYTHON, SQL, EXCEL / NUMPY, PANDAS, WORDCLOUD
/ SEABORN, PLOTLIB, FOLIUM 등

PROJECT1

OLIST E-COMMERCE DATA EDA 프로젝트

ERD 탐색

ERD 탐색 및 DB 적재 (ORACLE DB)



ERD 탐색 결과 – 데이터셋의 한계점

1. 불분명한 카테고리 분류 기준:

카테고리 간의 비교 분석이 어려움

2. 제품에 제품명 없이고 유일 코드만 존재:

카테고리 내 아이템 간의 비교 분석이 어려움

3. 희박한 재구매율:

FREQUENCY가 충분하지 않아, 고객 RFM 분석이 어려움

4. 3년 이상 지난 데이터:

2018년 10월 이후 데이터가 없어서 남미 이커머스 마켓의 성장이나 COVID 펜데믹 같은 글로벌 이슈가 반영되지 않았기 때문에 특히 수요 관련 예측 분석 시 타당성이 부족함

PROJECT1

OLIST E-COMMERCE DATA EDA 프로젝트

전처리 및 EDA 과정

1. 서비스 출시 및 데이터 수집 중단 시기

데이터가 현저히 적은 첫 3개월과 마지막 1개월 데이터 제외



2. 지리정보 전처리

셀러 및 고객 주소 정보가 부정확하거나 결측값인 경우

GEOLOCATION 데이터로 수정 및 대체, GEOLOCATION.CSV에도
결측데이터가 존재하여 일부 데이터는 웹에서 수집하여 대체

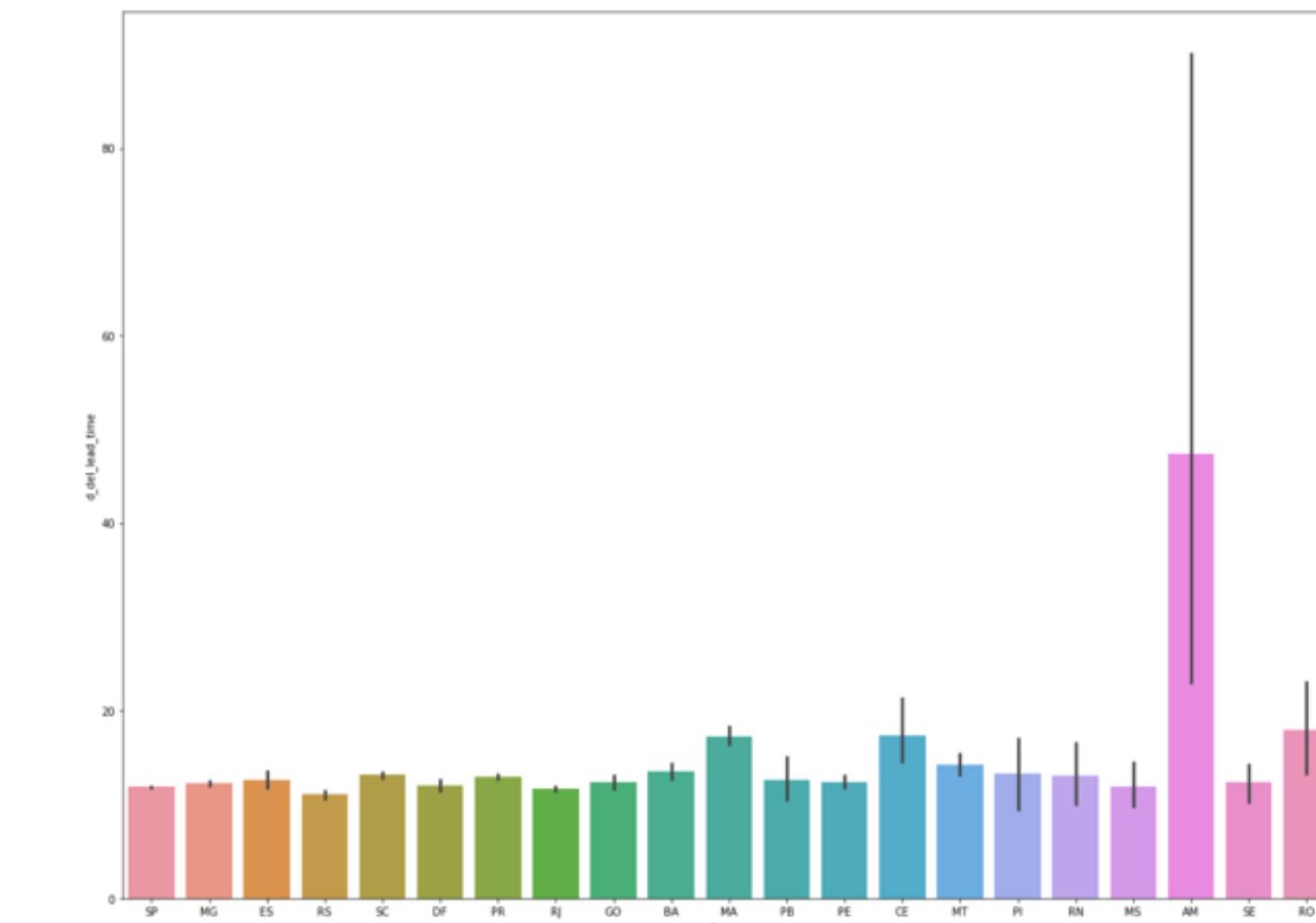
(2) 개별 테이블-데이터 설정

The screenshot shows a database interface with two tables: 'geolocation_city' and 'geolocation_state'. The 'geolocation_city' table lists various cities with their coordinates and state IDs. The 'geolocation_state' table lists states with their names and state codes. A blue arrow points from the 'geolocation_state' table towards a list of city names on the right, suggesting a mapping or lookup process for city names.

3. 셀러주별 배송소요시간 평균

실제 브라질 국토적인 제약이 드러나는 것으로 확인됨

이를 통해 배송 데이터 분석을 통하여 합당한 배송예상시간 책정과 상세한 배송 정보를 보고서에 넣고자 함.



PROJECT1

OUST E-COMMERCE DATA EDA 프로젝트

배송 만족도 상승 방안 제안을 위한 지역 구분별로 소요 시간 분석

1. 가설설정 및 지표 설정: 프로세스 타임스탬프 별로 차이를 구하면 지역 구분별로 차이가 있지 않을까?

1. 지표 설정

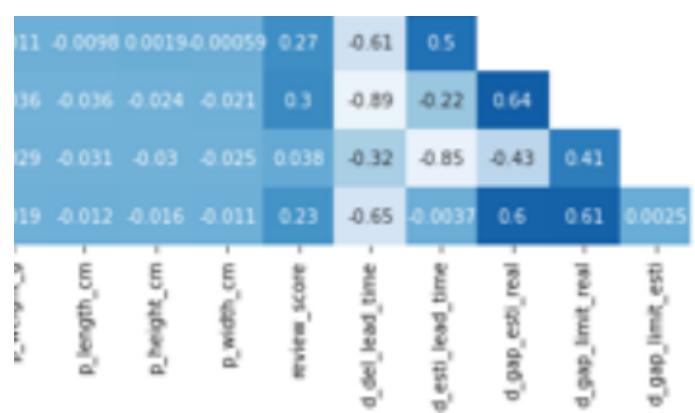
실 소요 시간 (del_lead_time)	배달완료 - 구매완료시간 (order_delivered_customer_date - order_purchase_timestamp)
예상 소요 시간 (esti_lead_time)	예상배달완료 - 구매완료시간 (order_estimated_delivery_date - order_purchase_timestamp)
실제 지역 시간 (gap_esti_real)	예상배달완료 시간 - 배달완료 시간 (order_estimated_delivery_date - order_delivered_customer_date)
남은 지역 시간 (gap_limit_real)	배송limit시간 - 배달완료시간 (shipping_limit_date - order_delivered_customer_date)
추가 limit 시간 (gap_limit_esti)	배송limit시간 - 예상배달완료 시간 (shipping_limit_date - order_estimated_delivery_date)

shipping_limit_date(order_items table) : 배송 최대 limit 기간

order_estimated_delivery_date (orders table) : 구매자에게 고시한 예상 배송완료 시간

2. 시간차 피처 생성 후 배송소요시간 지표들과, 배달완료 후 리뷰작성시간까지 걸린 시간까지의 상관관계 분석:

실제로 배송이 고지된 예상시간보다 느리게 도착할 수록 고객은 더 빨리 리뷰를 작성한 것으로 분석 결과 드러남.



3. 셀러와 구매자의 위치 차이에 따른 분석:

실제로 같은 도시보다 같은 주에 있을 경우 더 오래 걸리고, 같은 주일 때보다 다른 주에 있을 경우 더 오래 걸림을 확인

df_same_city

df_same_city.describe()				
del_lead_time_insamecity	esti_lead_time_insamecity	gap_esti_real_insamecity	gap_limit_real_insamecity	gap_limit_esti_insamecity
count	4978	4978	4978	4978
mean	5 days 19:06:33.120329449	14 days 12:01:22.940136601	8 days 16:54:49.819807151	0 days 06:47:40.020691040
std	4 days 23:57:34.216918207	6 days 19:03:46.506443936	7 days 03:25:56.254175071	4 days 21:46:11.188201222
min	0 days 12:48:07	2 days 00:15:03	-105 days +10:21:44	-112 days +20:31:44
25%	2 days 22:31:21.50000	11 days 01:32:38.25000	5 days 04:38:10	-1 days +01:24:08.50000
50%	4 days 17:14:23	13 days 09:34:46	8 days 06:01:54	0 days 17:46:26.50000
75%	7 days 00:43:32	17 days 05:24:31.25000	12 days 05:16:21.50000	2 days 17:49:15.50000
max	117 days 03:38:43	96 days 06:23:22	48 days 06:57:09	24 days 13:47:23
				5 days 00:13:00

df_same_state

df_same_state.describe()				
del_lead_time_insamestate	esti_lead_time_insamestate	gap_esti_real_insamestate	gap_limit_real_insamestate	gap_limit_esti_insamestate
count	29545	29545	29545	29545
mean	8 days 06:41:37.329734303	18 days 01:26:51.808630902	9 days 18:45:14.478896598	-2 days +04:57:15.317617195
std	6 days 10:55:57.982725113	6 days 22:50:53.049203734	7 days 19:52:22.684805871	6 days 08:59:59.184867884
min	0 days 18:45:10	2 days 00:11:32	-176 days +03:08:29	-166 days +13:39:29
25%	4 days 10:28:31	13 days 09:23:27	6 days 03:27:13	-4 days +00:12:24
50%	6 days 22:48:11	17 days 23:33:22	10 days 00:27:39	-1 days +05:38:20
75%	10 days 03:20:49	21 days 11:49:34	13 days 09:01:38	1 days 07:43:14
max	191 days 11:07:30	149 days 14:12:53	146 days 00:23:13	41 days 07:07:50
				15 days 20:19:00

df_diff_states

df_diff_states.describe()				
del_lead_time_indiffstates	esti_lead_time_indiffstates	gap_esti_real_indiffstates	gap_limit_real_indiffstates	gap_limit_esti_indiffstates
count	61454	61454	61454	61454
mean	15 days 03:33:37.828278061	27 days 04:57:05.249796596	12 days 01:23:27.421518534	-9 days +11:55:57.070036125
std	10 days 01:10:47.174829253	7 days 15:47:17.959682460	11 days 05:37:30.083647359	10 days 21:24:14.635119233
min	0 days 20:43:20	2 days 12:02:20	-189 days +00:35:53	-206 days +08:54:21
25%	8 days 21:39:56.75000	22 days 05:39:13.50000	7 days 05:26:29	-12 days +03:24:00.50000
50%	12 days 18:51:36	26 days 03:52:21	13 days 03:06:14.50000	-7 days +17:31:10.50000
75%	18 days 08:34:32	31 days 02:55:02	18 days 03:07:27.50000	-3 days +08:39:04.25000
max	209 days 15:05:12	155 days 03:15:04	139 days 09:32:15	1035 days 08:59:06
				911 days 22:35:00

PROJECT1

OLUST E-COMMERCE DATA EDA 프로젝트

한계점 및 디벨롭 방안

한계점 1: 언어적 한계

- 카테고리 구분에 문화적 차이가 있을 수 있었으며, 이를 반영한 분석을 할 수 없었음
- 실제 리뷰를 분석할 시에 단어 빈도수 정도만 확인할 수 있었음 (영어로 번역을 시도하였으나 의미상 분석이 불가한 상태가 많았음)

한계점 2: 능력적 한계

- 데이터셋이 다방면으로 고루 갖춰져 있어 분석 진행 시 함께 다뤄야 할 데이터가 방대하였음에 반해 분석 툴에 능숙하지 않아 전처리 및 EDA에 시간이 많이 들었음
- 또한 지리정보를 함께 사용할 수 있는 FOLIUM 사용법이 미숙하여 원하는 분석을 시도하지 못함

개발 방안 1: 데이터를 이해하는 시간을 들이기

- 명확한 분석 목표를 세우고 필요한 데이터가 무엇일지 고민 후 계획 세우기
- 데이터셋이 풍성할 수록 살펴보아야 하는 스키마가 많은 것을 인지하고 분석 목표에 맞춰 활용 계획 세우기

개발 방안 2: 배송에 리뷰가 영향을 끼친다는 명확한 관계성 확인하기

- 리뷰 WORDCLOUD 결과 배송 관련 어휘가 많았을 뿐 실제로 고객 만족도에 영향을 끼치는지 확인되지 않은 상태에서 분석을 진행하였기 때문에 분석의 타당성이 부족함
- NLP를 배웠으므로 번역 후 리뷰 분석에 활용하여 SCORE나 시간 정보와 함께 고객 만족도 분석을 진행할 것



PROJECT2

NLP를 활용한 네이버 영화 평점 조작여부 판별 프로젝트

프로젝트 개요

프로젝트 목적: 네이버 영화 리뷰의 조작여부를 판별 및 조작 평점을 제외하여 실제 평점 도출을 통한 네이버 영화 리뷰 신뢰도 제고

1. 문제정의:

- 악의적으로 생산된 거짓리뷰에 의해 과대, 과소 평가된 영화가 있고 실제로 대중들 사이에 신뢰도가 낮음
- 네이버 영화는 영화 데이터베이스가 넓은 편임에도 영화 관련 정보 통합 플랫폼으로서의 기능을 발휘하지 못함
- 컨텐츠 고관여 유저들이 늘어나는 트렌드에 반하여 네이버 영화는 이용자 손실 가능성이 높은 상태
- 추가적으로 네이버 시리즈온 유입 및 매출 증대의 기회를 잃고 있음

2. 프로젝트 수행단계 계획 및 현실적 목표 설정

계획: 영화 리뷰 데이터 수집 > 거짓 리뷰 라벨링 > 반지도학습을 사용하여 정황적 정보와 자연어 데이터(리뷰)로 거짓 판별

수정 목표: 조작 평점을 제외한 평점 도출 과정 설계

진행환경 및 분석, 자연어처리, 시각화 도구:

JUPYTER NOTEBOOK, NOTION / PYTHON, SQL / NUMPY, PANDAS, KONLPY, WORDCLOUD, TENSORFLOW / SEABORN 등



PROJECT2

NLP를 활용한 네이버 영화 평점 조작여부 판별 프로젝트

단계적 데이터셋 수집

영화 표본 추출을 위한 요인 검정:

어휘 사용 현황을 고려한 표본 추출 및 분석을 통하여 최종 데이터셋에 존재할 수 있는 편향을 제거하고자 함

1. 장르별로 특정 어휘의 빈도수나 어감이다를 수 있음

- 공포, 드라마, 액션, 코미디, 판타지, 애니메이션 각 장르별 5만개의 리뷰 데이터 추출

2. 개봉 시기별로 특정 어휘의 빈도수나 어감이다를 수 있음

- 스마트폰 보급 확대 전 (~2009), 스마트폰 보급 확대 시기(2010~2015), OTT 서비스 출시 및 이용자수 증가 시기(2016~) 세 시기 별로 5만개의 리뷰 데이터 추출

토큰화 및 불용어 처리 진행 후 WORDCLOUD 진행한 결과

요인별로 큰 차이가 없음을 확인

이후 장르나 개봉 시기별 비중에 상관없이 영화 선정 및 크롤링(조사적 편의를 위하여 거짓 리뷰가 존재할 가능성이 큰 영화 중심으로 선정):

약 250만개의 리뷰 데이터 추출



PROJECT2

NLP를 활용한 네이버 영화 평점 조작여부 판별 프로젝트

EDA 및 전처리

1. 정황적 정보:

개봉일, 작성일자, 평점, 관람여부, 스포일러 여부

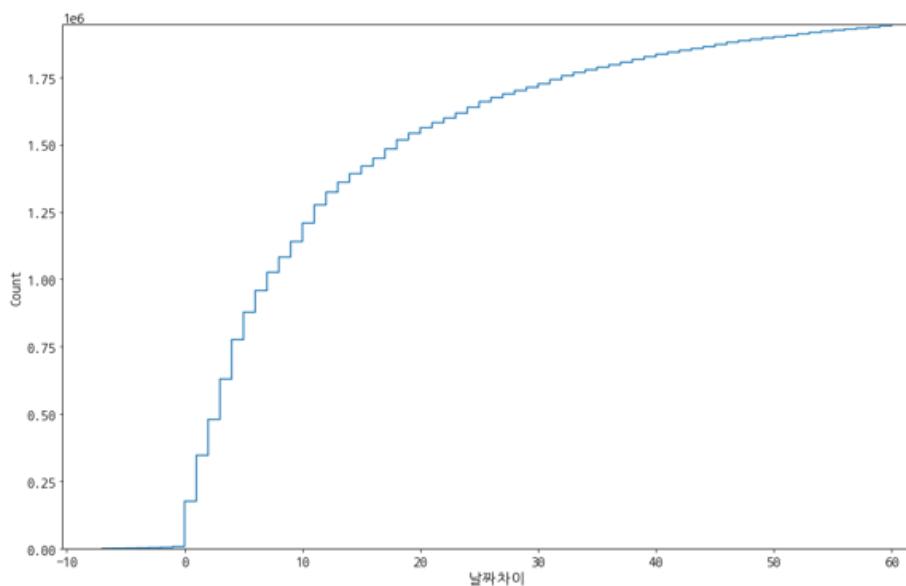
2. 피처 추가:

개봉일과 작성일과의 차이(날짜차이), 한글 형태 댓글길이

3. 조작리뷰로 사료되는 리뷰 데이터 총 564개 선정 및 라벨링

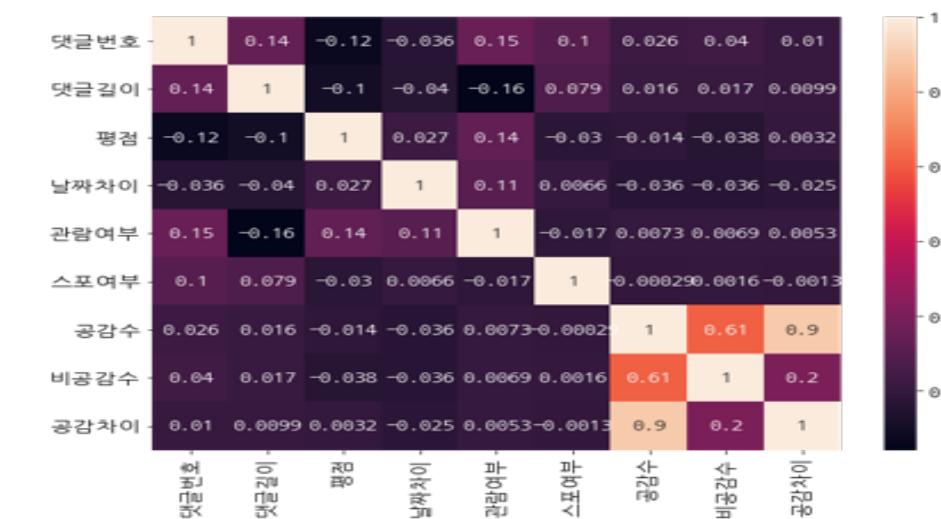
(1) EDA를 통한 기간 제한:

위그래프 (날짜차이 = 개봉일 - 댓글작성일) 결과 날짜차이별 댓글수는 개봉 일을 시점으로 꾸준히 하락하며, 특히 약 20일을 기점으로 상승하는 폭이 많이 줄어들기 때문에, 개봉 후 3주인 21일 이전까지의 댓글만을 사용



(2) 정황적 정보간의 상관관계:

공감수에 관련한 컬럼은 앞에서 제외하기로 결정했으므로 나머지 5개 컬럼은 서로가 관련이 없음을 확인하고 이후 분석을 진행함



3. 자연어처리:

(1) 정규표현식을 통해 특수문자, 영어 등 제거

(2) 형태소분석기 중 OKT(MORPHS)를 활용한 토큰화:

자연어처리에서 일반적으로 자주 사용되는 MECAB과 비교하였으나, MECAB의 경우 의미를 잃을 수 있는 수준까지 단어를 자르는 것으로 확인되어 OKT 선택

(3) 불용어처리: 토큰화 후 확인하여 의미를 잃은 단어 제외

(4) 분석 가능한 글이 남아있지 않은 로우 삭제



PROJECT2

NLP를 활용한 네이버 영화 평점 조작여부 판별 프로젝트

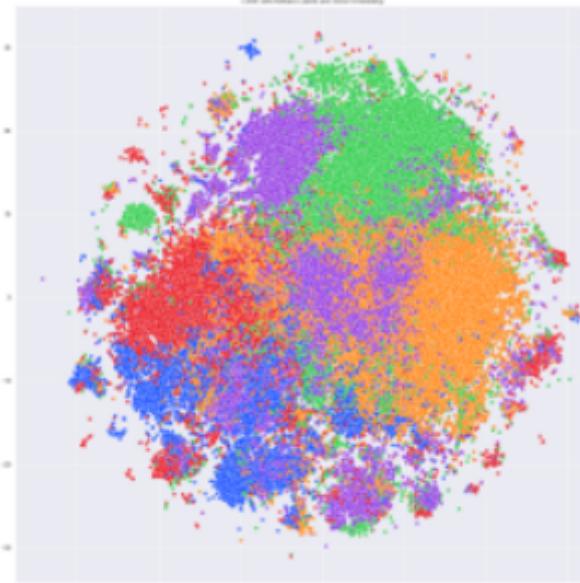
데이터 마이닝

1. WORD2VEC(SKIP-GRAM)으로 워드 임베딩 진행:

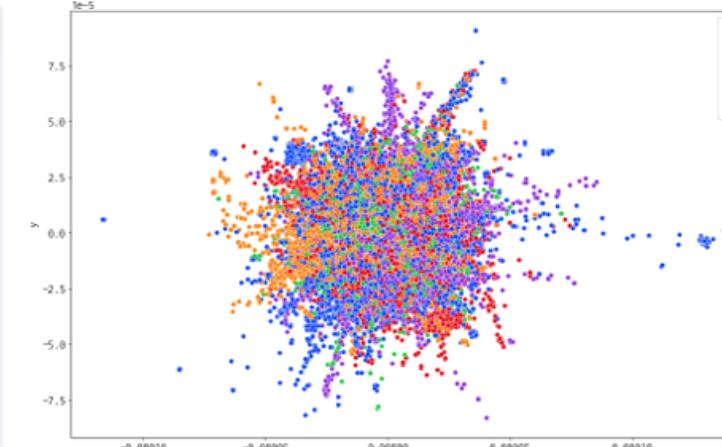
CBOW의 경우 비슷한 위치나 의미를 가진 단어 위주로 높은 값을 보여서 조작 댓글로 보이는 댓글들의 문장 자체를 보는 것에 적합하지 않음

2. 문장 단위 벡터화: WORD2VEC 값을 이용하여 SENT2VEC 진행, K-MEANS로 클러스터링

댓글 문장을 벡터화하여 시각화 및 확인 결과 가짜평점의 분포가 눈에 띄게 나타나지 않음



문장단위 클러스터링: 20만개 클러스터링 결과(좌) 및 거짓 리뷰를 포함한 60만개의 데이터 클러스터링 결과(우)

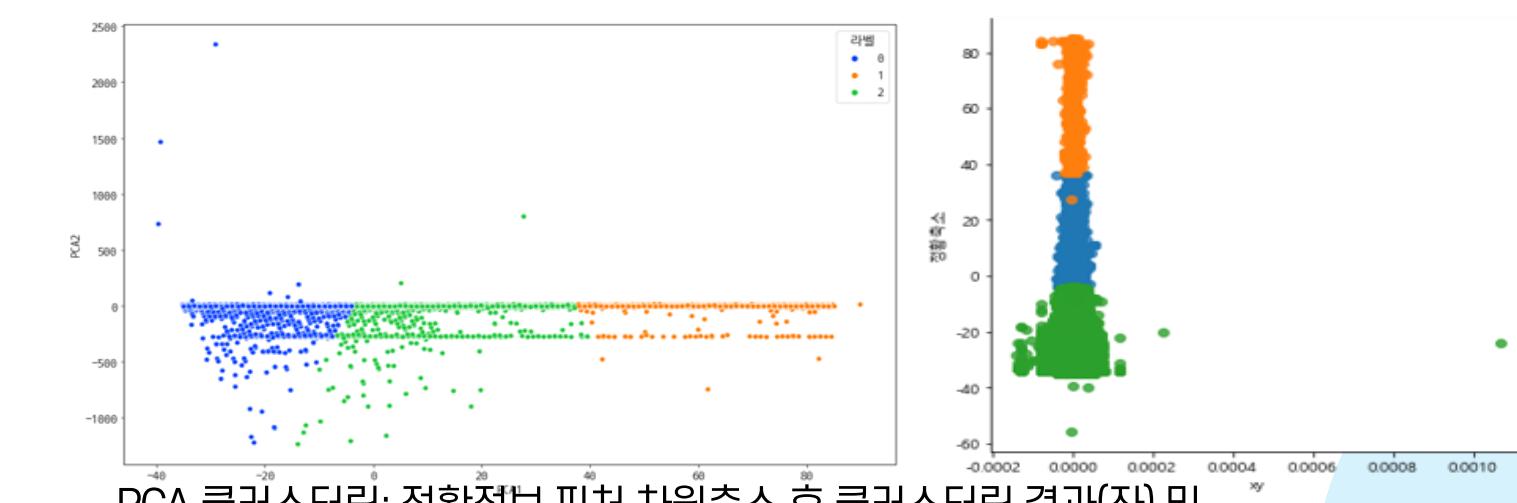


3. 정황정보 피처 PCA 진행 후 K-MEANS 클러스터링:

실제로 가짜리뷰로 정의한 560여개의 댓글 중 75%에 해당하는 댓글이 같은 군집에서 나타남

4. 정황정보 피처 PCA 값과 문장 벡터 PCA 값 두 가지 속성으로 K-MEANS 클러스터링:

정황정보 피처 개수로 인한 영향을 줄이기 위함이었음
실제로 정황정보만 클러스터링 했을 시와 비슷한 형태를 띨
결국 정황정보를 빼고 가짜판별은 어렵다고 판단



PCA 클러스터링: 정황정보 피처 차원축소 후 클러스터링 결과(좌) 및 정황정보와 문장벡터 함께 차원축소 후 클러스터링 결과(우)



PROJECT2

NLP를 활용한 네이버 영화 평점 조작여부 판별 프로젝트

프로젝트 결과

1. 문장은 정황정보와 함께 두면 유의미해 질 수 있음

: 워드 임베딩을 통한 문장 단위 임베딩만으로는 가짜에 가까운 데이터를 가짜로 확정 짓기 어려운 형태였음

164973	걸캅스	재밌어요시사회로 봤었는데 또 보려구요	10	0	0	18	2
350832	오블리비언	오랜만에 볼만한 대작이 나왔다 년 들어 본 영화중에 최고	10	0	0	31	2
94332	건축학개론	토율날 보고왔는데 여운이 살라지질 않는다 그래서 또 봐야 겠다	10	0	0	31	2

문장과 정황정보를 함께 두고 클러스터링 후 분류된 댓글들이 실제로 선정했던 거짓 리뷰와 유사하게 나온 상태

2. 댓글길이가 짧을수록 거짓 댓글과 비슷한 정황일 확률이 높음

	댓글길이 평균	댓글길이 최대	댓글길이 최소
가짜 댓글이 많은 클러스터	18.21	32	1
그렇지 않은 군집1	99.81	126	74
그렇지 않은 군집2	46.74	73	11

3. 관람객이어도 조작 댓글일 확률이 낮지 않음

	관람여부
가짜 댓글이 많은 클러스터	242565
그렇지 않은 군집1	19187
그렇지 않은 군집2	68954

PROJECT2

NLP를 활용한 네이버 영화 평점 조작여부 판별 프로젝트

프로젝트 한계점 및 활용 기대효과, 디벨롭 방안

1. 한계점:

- (1) 한국어 형태소 분석기의 한계: 사소한 뉘앙스 파악할 수 있을 만큼 고도화된 형태소 분석이 가능한 모델이 부재
- (2) 유저 정보 수집의 한계: 아이디의 생성일시나, 특정 유저의 다계정 여부 등을 알아볼 수 있는 데이터가 있었다면 조작을 위해 만들어진 아이디인지 판별하여 활용 할 수 있었을 것
- (3) 반지도학습 연구 정보 부족의 한계: 자연어 반지도학습을 통한 모델링 관련 정보가 부족하여 분석 진행에 어려움이 존재하였음
- (4) 거짓으로 정의내린 것이 진실로 거짓리뷰가 아닐 가능성이 있다는 것의 한계

2. 프로젝트 활용 방안: 댓글과 정황 함께 분석하여 네이버 영화 리뷰 속 조작평점을 판별하는 기준 수립 및 실시간 어뷰징 적용

활용 기대 효과:

조작평점 제외 후 실제 평점을 도출하여 네이버 영화 평점의 신뢰도 상승

- ▶ 네이버 영화 이용자수 증가
- ▶ 시리즈온 유입 가능성 증가 및 매출 상승 기대

3. 프로젝트 종료 이후 디벨롭 계획:

- (1) 다른 자연어처리 모델 활용을 시도하여 댓글 분석 고도화
 - (2) 라벨링 데이터 추가 생성 및 반지도학습 수행
- ▶ KNN을 활용한 조작 리뷰 분류 안티치트 모델 설계 기획 중

