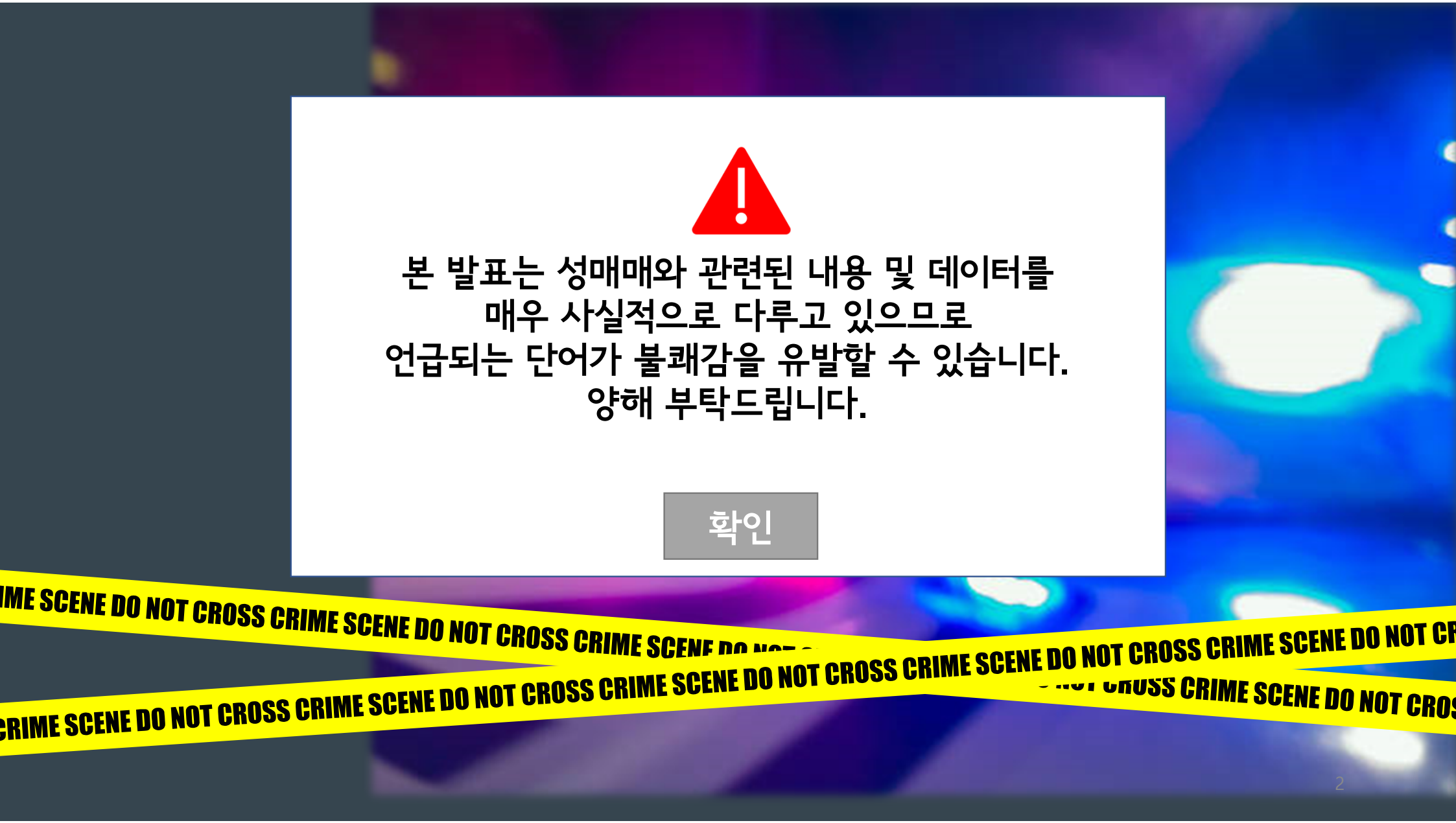
SNS 상의 성매매 게시물 분석

-성매매 트윗과 일반 트윗의 특성 차이 중심-

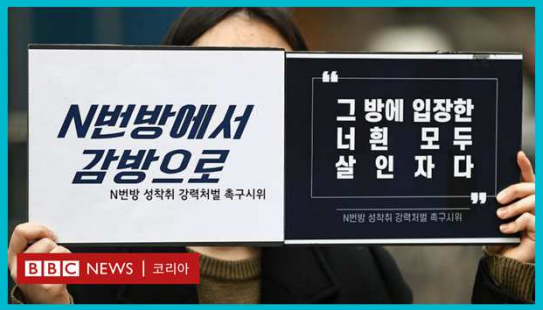
DScover 사회2조_
노최유하, 경정원, 김해니, 선해라


본 발표는 성매매와 관련된 내용 및 데이터를
매우 사실적으로 다루고 있으므로
언급되는 단어가 불쾌감을 유발할 수 있습니다.
양해 부탁드립니다.
확인
IME SCENE DO NOT CROSS CRIME SCENE DO NOT CROSS CRIME SCENE DO NOT CROSS CRIME SCENE DO NOT CROSS CRIME SCENE DO NOT CR
CRIME SCENE DO NOT CROSS CRIME SCENE DO NOT CROSS CRIME SCENE DO NOT CROSS CRIME SCENE DO NOT CROSS CRIME SCENE DO NOT CROS
2

차례

분석 목적

N번방 등의 성매매 및 착취 범죄 증가



1

많은 '성매매알선행위에관한 법률' 위반 사례

출처: 국가통계포털 KOSIS

2

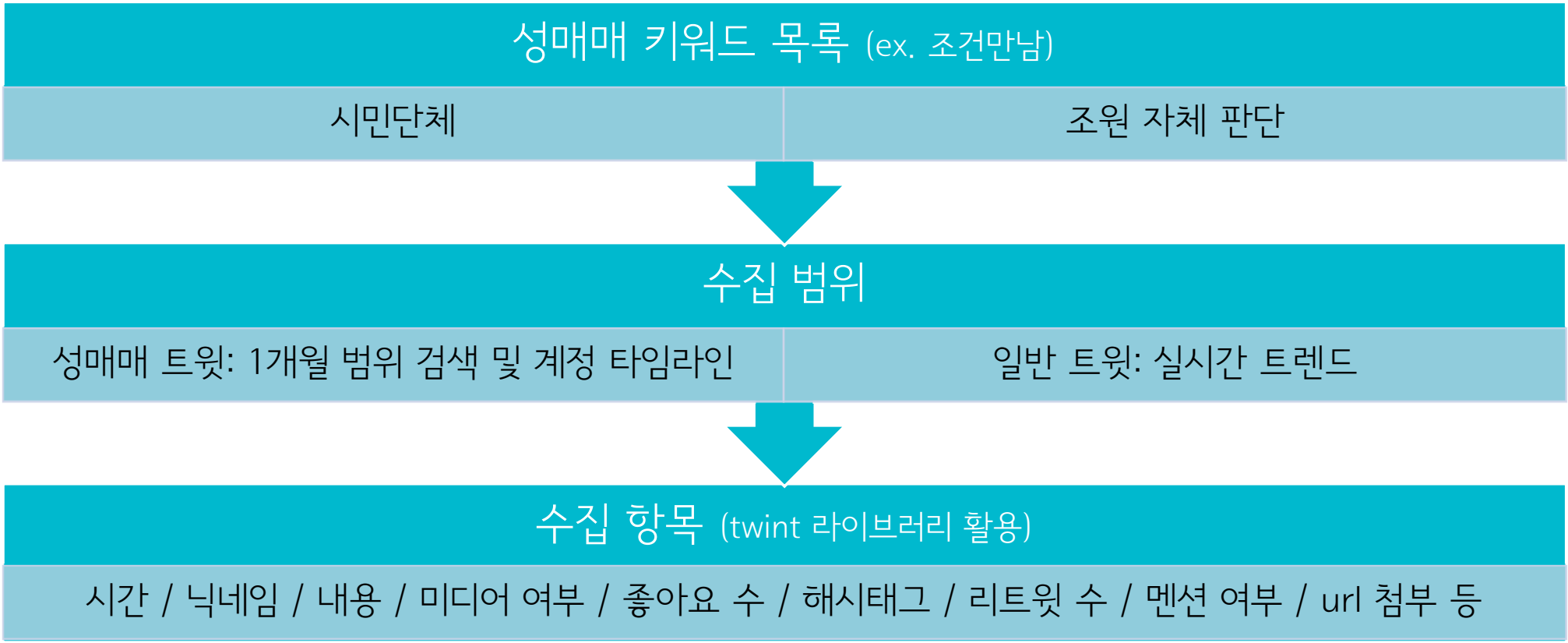
SNS상의 성매매 알선 및 홍보 게시물

3

twitter 64.4%

성매매 트윗의 특성 파악하여 SNS상의 직간접적인 노출 방지

데이터 수집



[참고 논문]
송봉규, & 김예정. (2020). 성매매알선웹사이트 트위터 (Twitter)의 연결 실태와 대책. *한국범죄심리연구*, *16*, 75-86.
윤현식, 윤영호, 박현재 (2020). 머신러닝을 통한 SNS 상의 성매매 알선 홍보 글 탐지의 효율성 제고방안 연구. 한국지역정보화학회지, 23(3), 43-65. 장정현, 나스리디노프 아지즈 (2017). 유해 해시태그 비율 기반의 유해 정보 판단 및 수집 시스템. 한국정보과학회 학술발표논문집, 273-275

데이터 수집

초기 데이터셋의 데이터 수

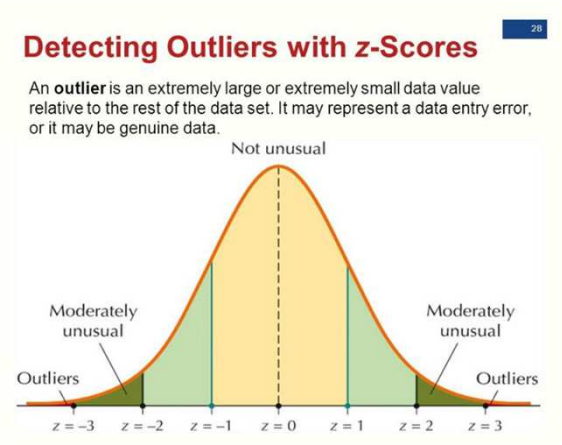
keyword	search_word	count	keyword	search_word	count	keyword	search_word	count	search_word	count
성매매 유형 문구		145	성매매 행위 문구		1227	트렌드	더블헤더	19	ATEEZ투표완료	1000
		743			198		미세먼지	549	StrayKidsRUN	1095
		1183			1780		보건실침대	3	THEBOYZFINALVOTE	1000
		210			6315		수업시간몰폰	2	닉네임에받침을빼면귀엽다	1000
		8415			1414		어버이날	1129	더보이즈	4225
		30617					피어리스	48	비투비피날레응원	1118
		2988					허벅지씨름	4	스트레이키즈	4249
		46068					호흡곤란	66	실버라이트	1173
		298							에스에프	2080
		21970							에이티즈	4000
		29469							짱보이즈	3141
		14								
		27022								
		8273								
		31910								
		30717								
		1975								
합계		242017			10934			1820		24081

※성매매 관련 단어를 유포하지 말라는 센터 측의 당부가 있어 검색어를 가립니다.

전처리 과정

1) 일반 트윗 중 성매매 트윗 **필터링**: 눈으로 보이는 특성 활용
2) **열 추가**

   트윗 길이, 물음표 개수, 공백 개수, 공백 비율, 마지막 단어의 특징적 형태 존재 여부, 닉네임 길이, 닉네임 내 물음표 개수
3) NULL **값**이 대부분인 열 삭제: non-NULL ≤ 4
4) **결측치** 채우기
5) **고유 값이 1개인 열** 삭제
6) 트윗 내 **명사 비율** 열 추가(KoNLPy 패키지 사용, 형태소 분석)
7) **성매매 트윗 여부** 판별 열 추가(일반: 0, 성매매: 1)
8) **이상치** 처리: 박스 플롯 & z-score의 절댓값 3 이상인 행 4만 6천 개 삭제.



전처리 과정

최종 데이터셋의 데이터 수

keyword	search_word	count	keyword	search_wo	count	keyword	search_word	count
성매매 유형 문구	[redacted]	500	성매매 행위 문구	[redacted]	433	트렌드	더블헤더	19
		807			169		미세먼지	549
		1252			473		보건실침대	3
		196			1893		수업시간몰폰	2
		8227			1190		어버이날	1129
		25126					피어리스	48
		2661					허벅지씨름	4
		39851					호흡곤란	66
		243						
		18162						
		23921						
		9						
		23839						
		5790						
		28523						
		27520						
		679						
합계		207306			4158			1820

※ 성매매 관련 단어를 유포하지 말라는 센터 측의 당부가 있어 검색어를 가립니다.

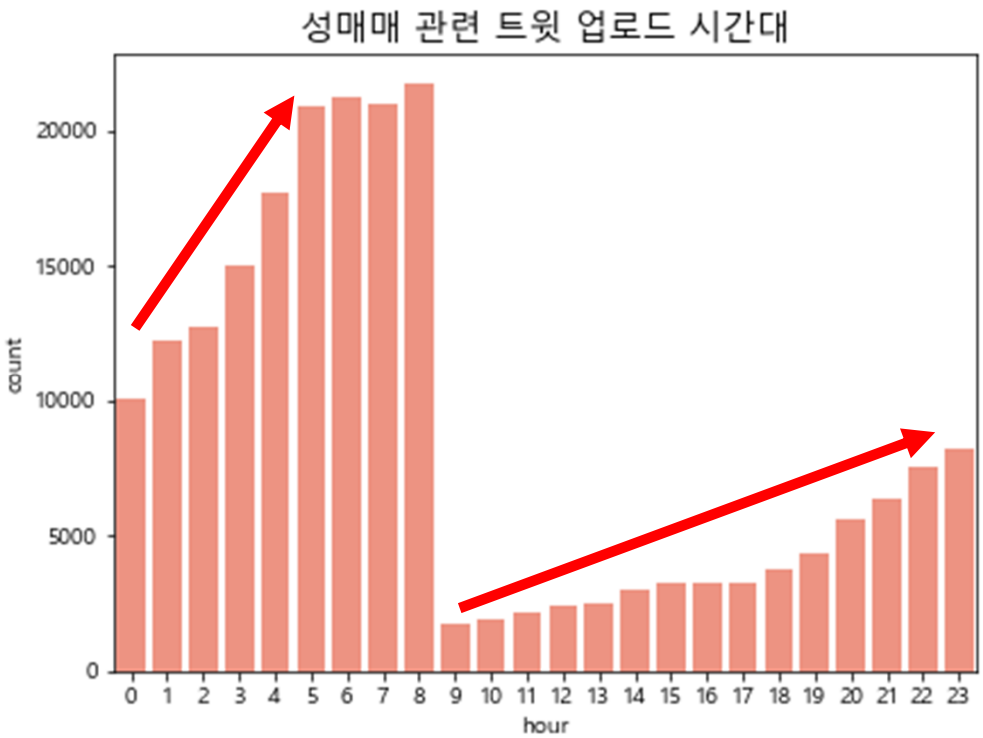
전처리 과정

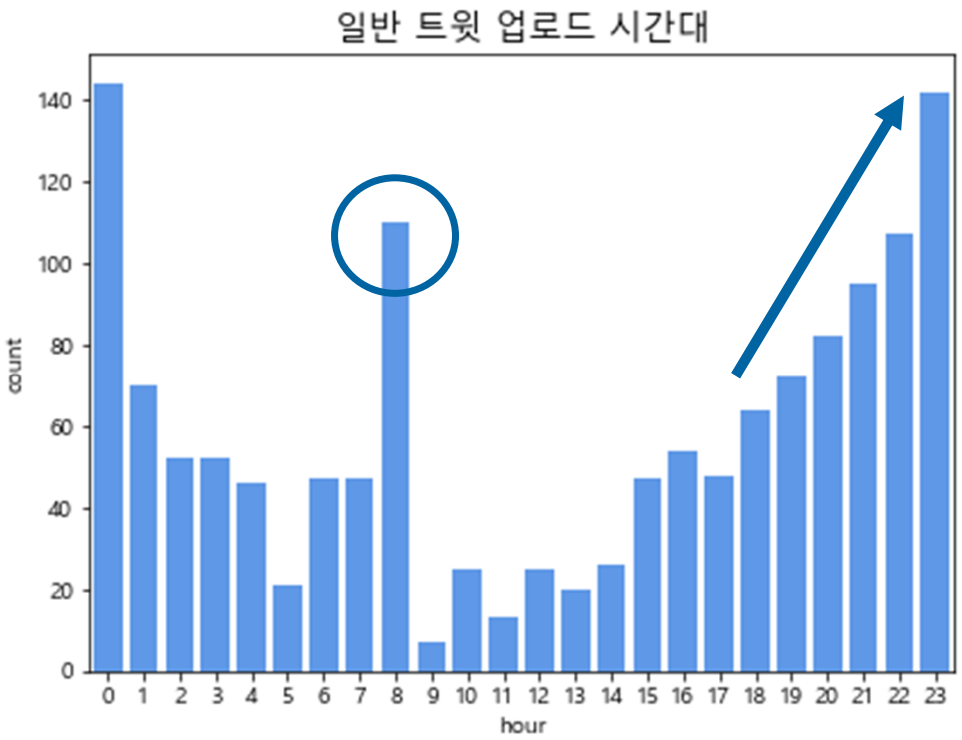
성매매 트윗과 일반 트윗의 특성 평균치

특성	성매매 트윗	일반 트윗
텍스트 길이	127.77자	86.62자
리트윗 된 수	0.00개	0.10개
좋아요 수	0.00개	0.52개
해시 태그 수	6.18개	4.83개
이미지 첨부 여부	0.80%	0.22%
텍스트 중 명사 비율	0.74%	0.44%
텍스트 길이 대비 공백 비율	0.13%	0.18%
닉네임 길이	4.36자	5.18자

※ 리트윗 수, 해시태그 수 등은 성매매 트윗과 일반 트윗 모두 거의 없으므로 발표에서는 다루지 않을 예정.

분석 결과: 형태적 분석



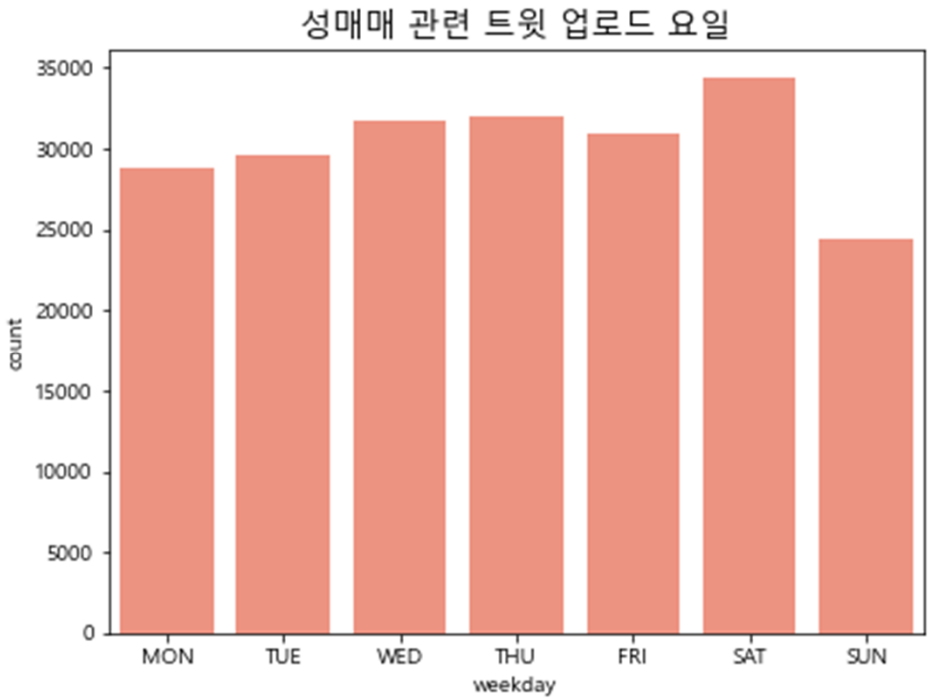


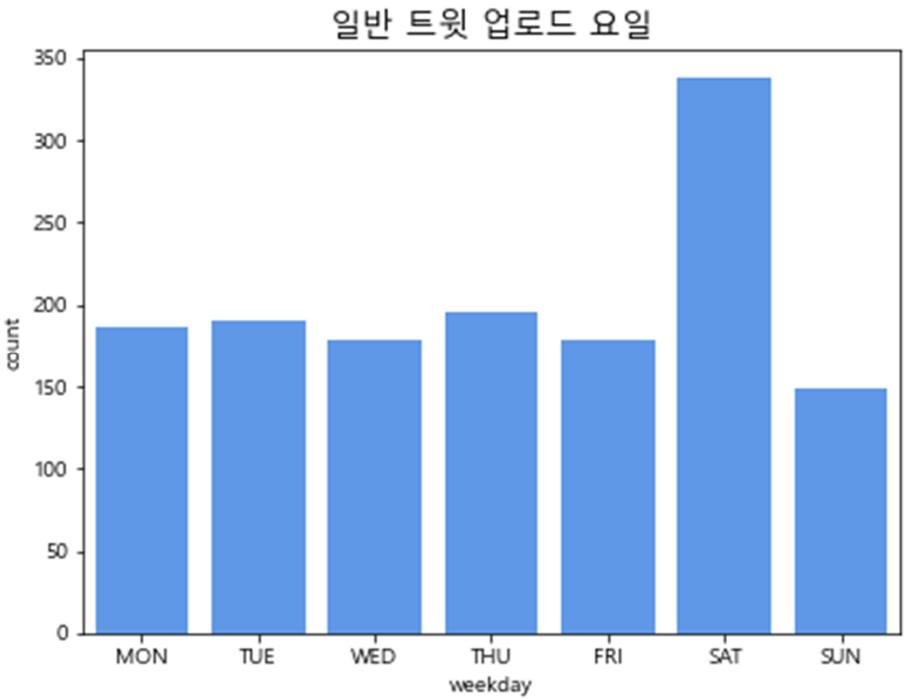
[업로드 시각 시계열 분석]

성매매 트윗: 오전 9시부터 꾸준히 상승

일반 트윗: 출근·등교 시간대에 높음. 퇴근 시간부터 취침 시간까지 상승

분석 결과: 형태적 분석



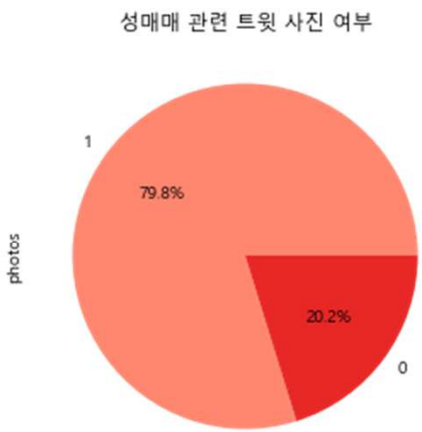


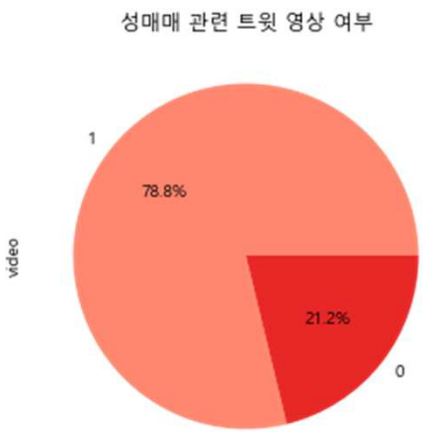
[업로드 요일 분석]

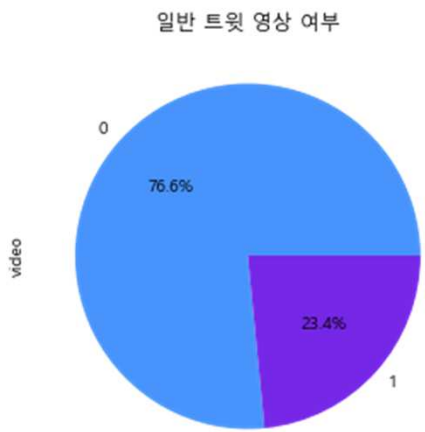
성매매 트윗: 전체적으로 비슷한 비율. 일요일에 가장 적음.

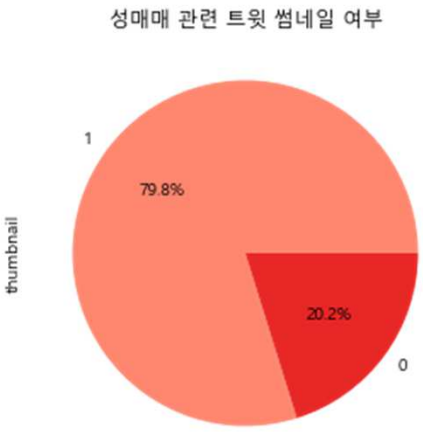
일반 트윗: 토요일에 많이 올라옴.

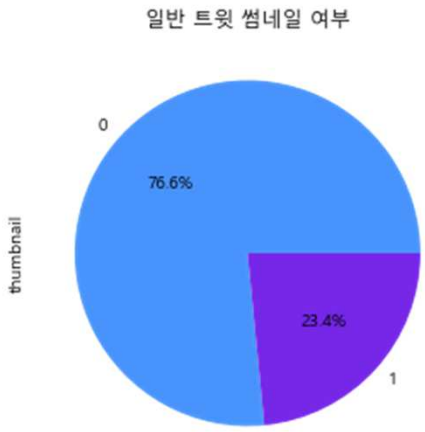
분석 결과: 형태적 분석











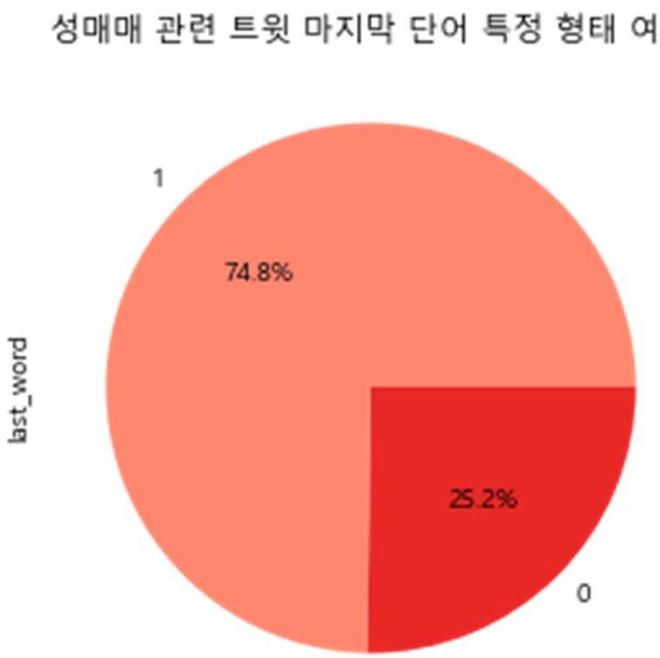
[트윗 내 미디어 분석]

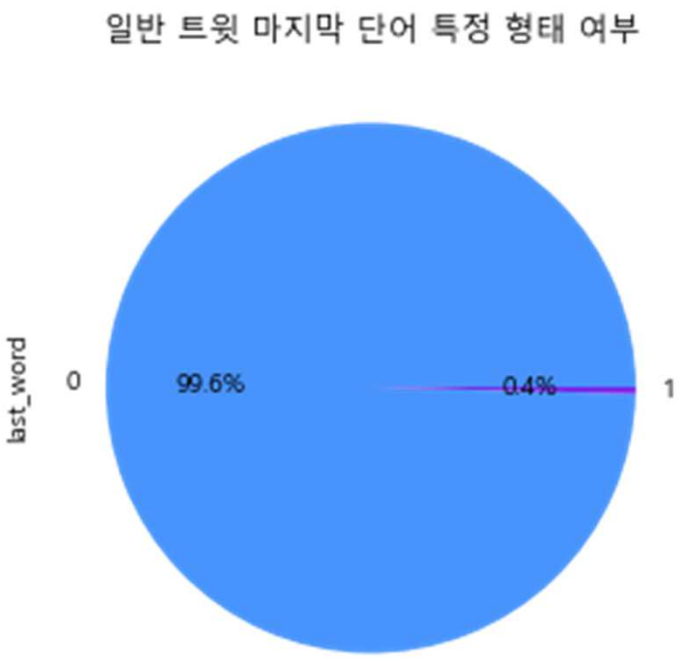
성매매 트윗: 약 80%가 미디어 포함.

자극적인 이미지로 시선을 끌려는 것으로 보임.

일반 트윗: 약 20%만 미디어 포함.

분석 결과: 형태적 분석

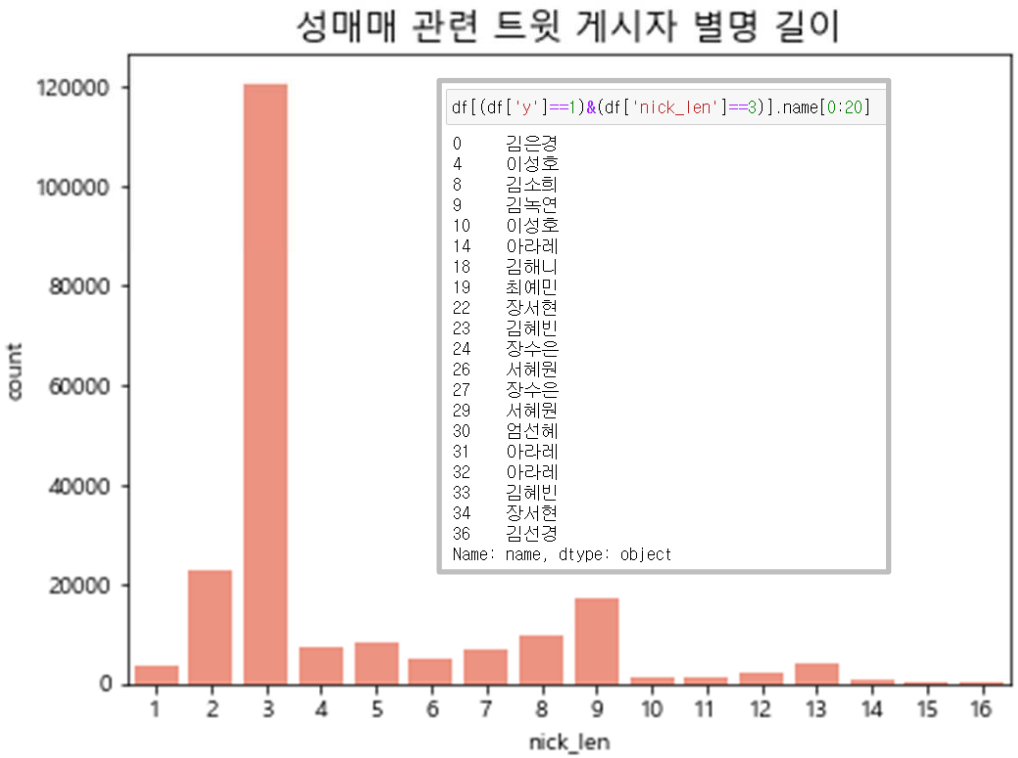


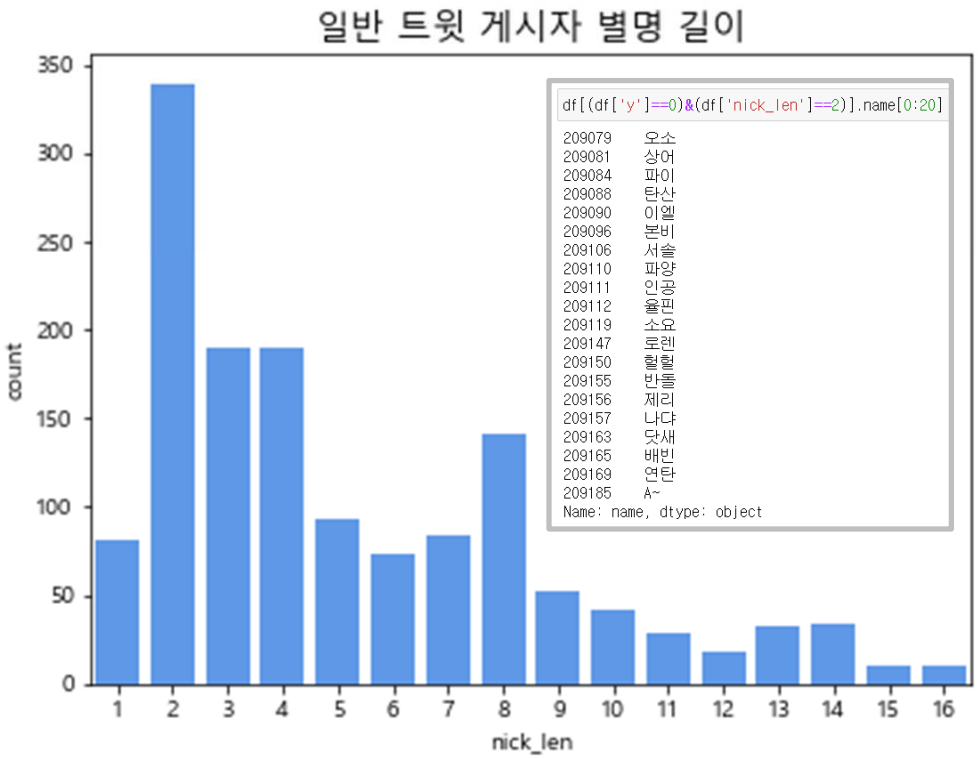


[마지막 단어 형태 분석]

성매매 트윗의 약 75%가 영어로 시작해서 숫자로 끝나는 특성을 가짐.
일반 트윗의 0.4%는 url을 포함한 경우이므로 이를 제외하면
마지막 단어의 특정 형태는 성매매 트윗 고유의 특성.

분석 결과: 형태적 분석



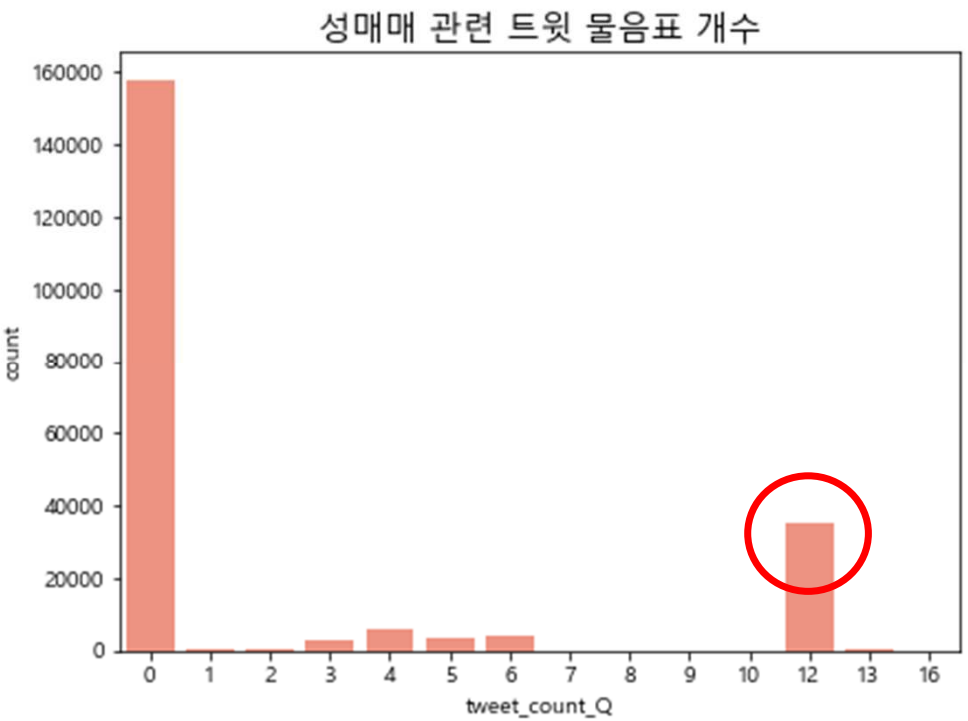


[닉네임 길이 분석]

성매매 트윗: 3글자인 경우가 대부분임. 실명 사용.

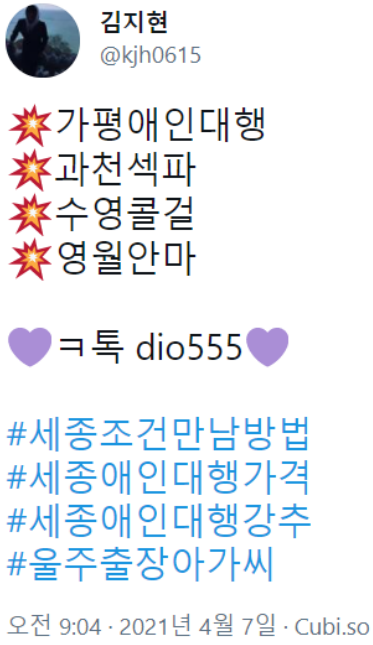
일반 트윗: 2글자가 많음. 별명 사용.

분석 결과: 형태적 분석




```
df[(df['y']==1)&(df['tweet_count_Q']==12)].tweet[0:10]
```

```
9477        ??가평애인대행 ??과천섹파 ??수영콜걸 ??영월안마  ??ㅋ톡 dio555??
#세종조건만남방법 #세종애인대행가격 #세종애인대행강추 #울주출장아가씨
9478      ??보은애인대행 ??영동섹파 ??옥천콜걸 ??음성안마  ??ㅋ톡 dio555??  #
광산애인대행후기 #광주동구출장샵 #광주동구출장아가씨강추 #광주동구출장샵
9479              ??상주애인대행 ??영천섹파 ??영주콜걸 ??구미안마  ??ㅋ톡 di
o555??  #청양콜걸샵 #청양콜걸강추 #청양조건만남 #청양애인대행
9480        ??영주애인대행 ??양평섹파 ??동두천콜걸 ??과천안마  ??ㅋ톡 tw567??
#성북구조건만남 #성북구조건만남강추 #성북조건만남강추 #성북조건만남
9494          ??고령애인대행 ??성주섹파 ??칠곡콜걸 ??예천안마  ??ㅋ톡 tw56
7??  #오산애인대행추천 #오산애인대행
9495    ??구로애인대행 ??강서섹파 ??
진안애인대행후기 #경남출장샵강추 #경
9496         ??고양애인대행 ??용인섹
#광진콜걸샵최고 #광진콜걸샵정보 #광진
9506          ??광진애인대행 ??동대문
5??  #구리시조건만남  #구리애인대행기
9507          ??속초애인대행 ??삼척
55??  #도봉콜걸 #도봉구조건만남강추
9508          ??의령애인대행 ??양산섹
#함안출장서비스 #함안출장업소강추 #함
Name: tweet, dtype: object
```

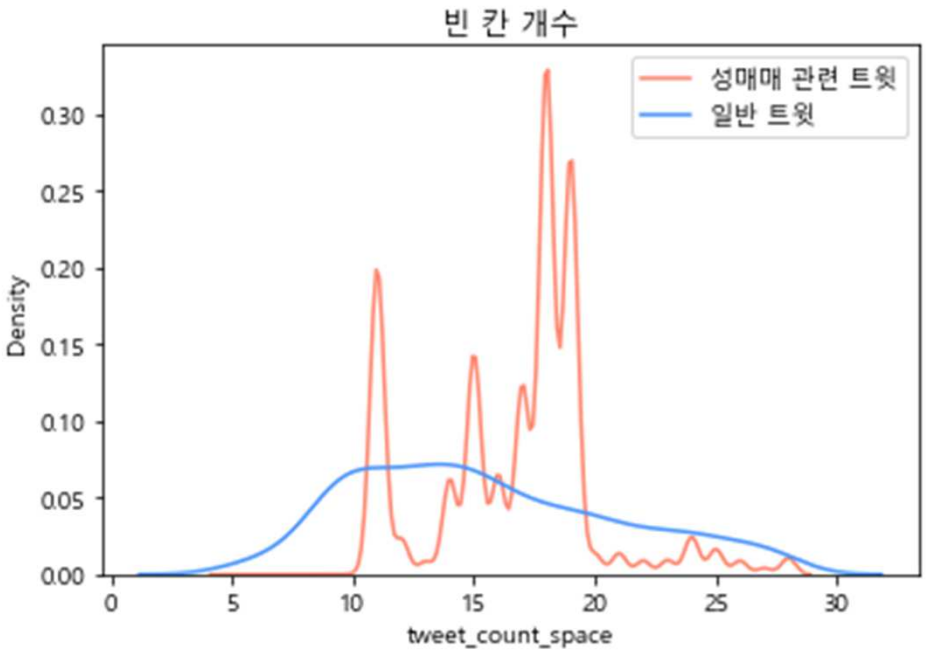


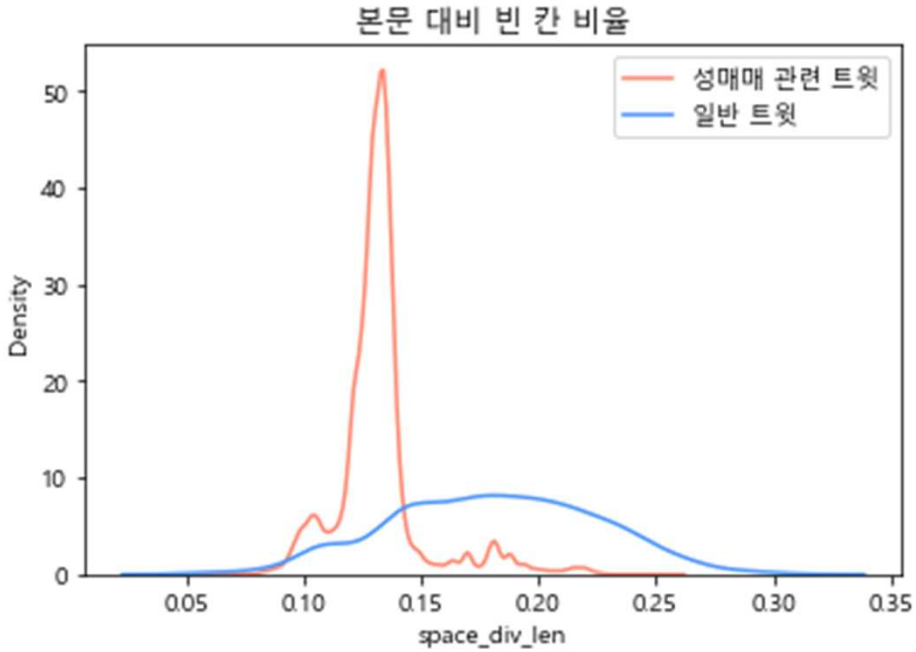
[물음표 개수 분석]

성매매 트윗: 이모티콘이 6개 사용된 경우로 보임.(인코딩 오류)

일반 트윗: 질문 형태의 트윗이라 1~2개 정도가 사용됨.

분석 결과: 형태적 분석



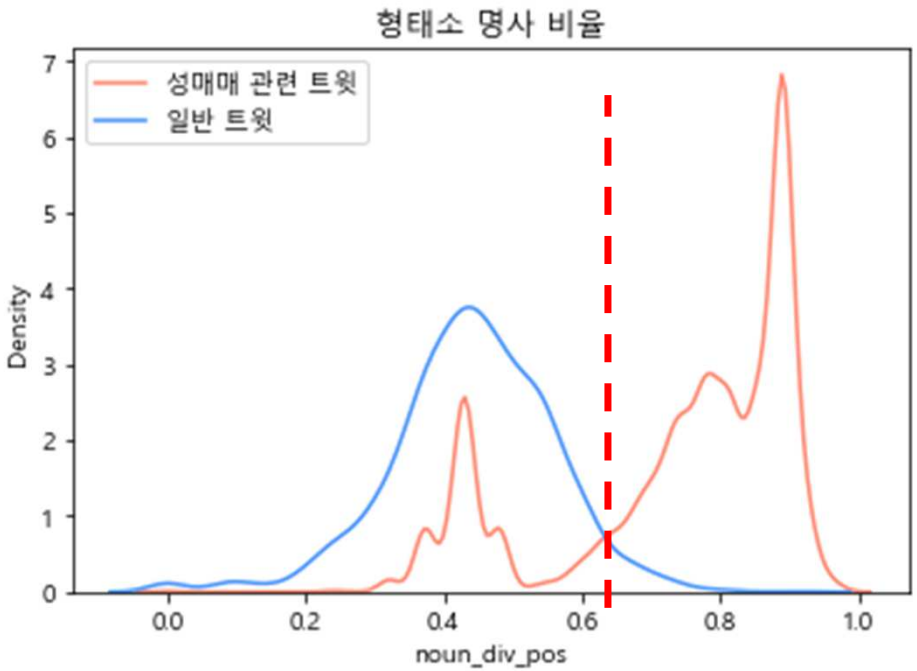


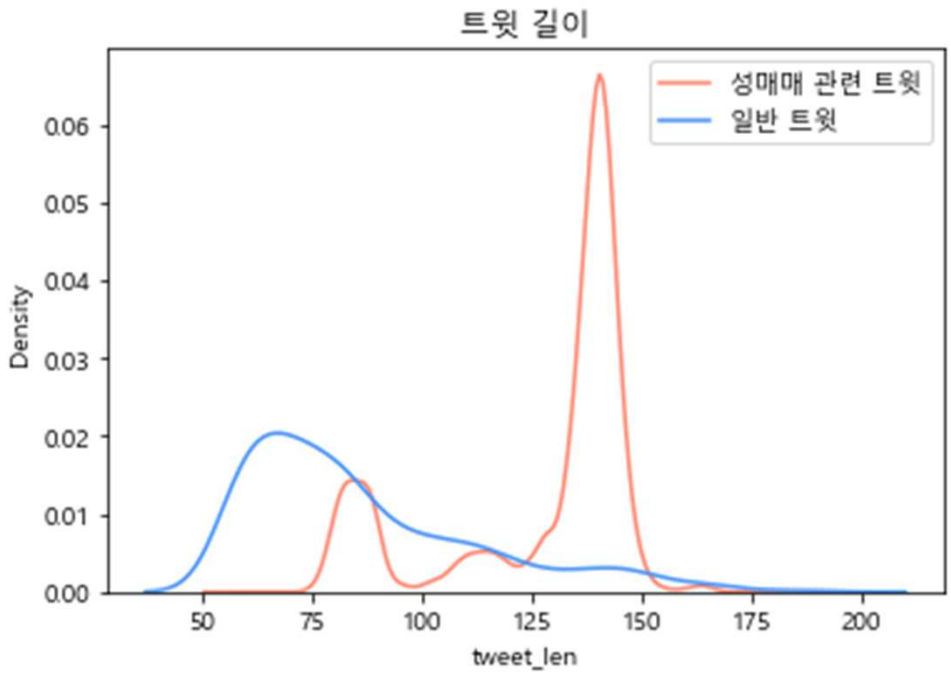
[공백 개수 및 비율 분석]

성매매 트윗: 특정 수치에서 높은 분포를 나타냄. → 단어 나열 형태이기 때문으로 보임.

일반 트윗: 완만한 분포 곡선을 보임. → 문장을 적으므로 공백 개수가 골고루 분포.

분석 결과: 형태적 분석





[명사의 비율]

성매매 트윗: 대부분 명사로 이루어짐.

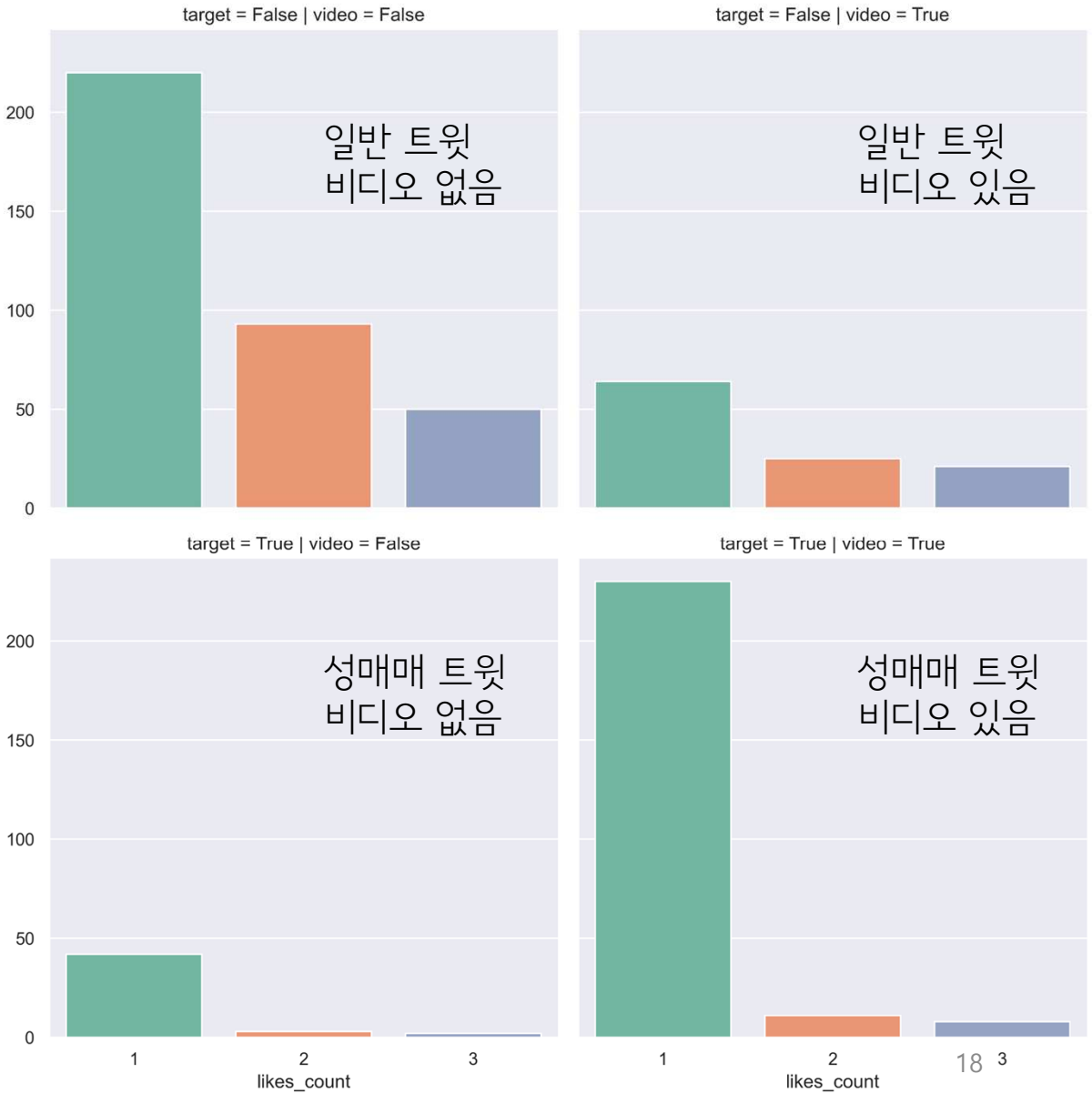
일반 트윗: 정규분포의 모습과 유사함.

[트윗 길이]

성매매 트윗: 140자 주변에 분포하는데 이는 최대 글자수.

일반 트윗: 길지 않음. 한두 문장인 것으로 보임.

분석 결과: 형태적 분석



[트윗 유형과 비디오 존재 여부에 따른 좋아요 수 분석]

일반 트윗

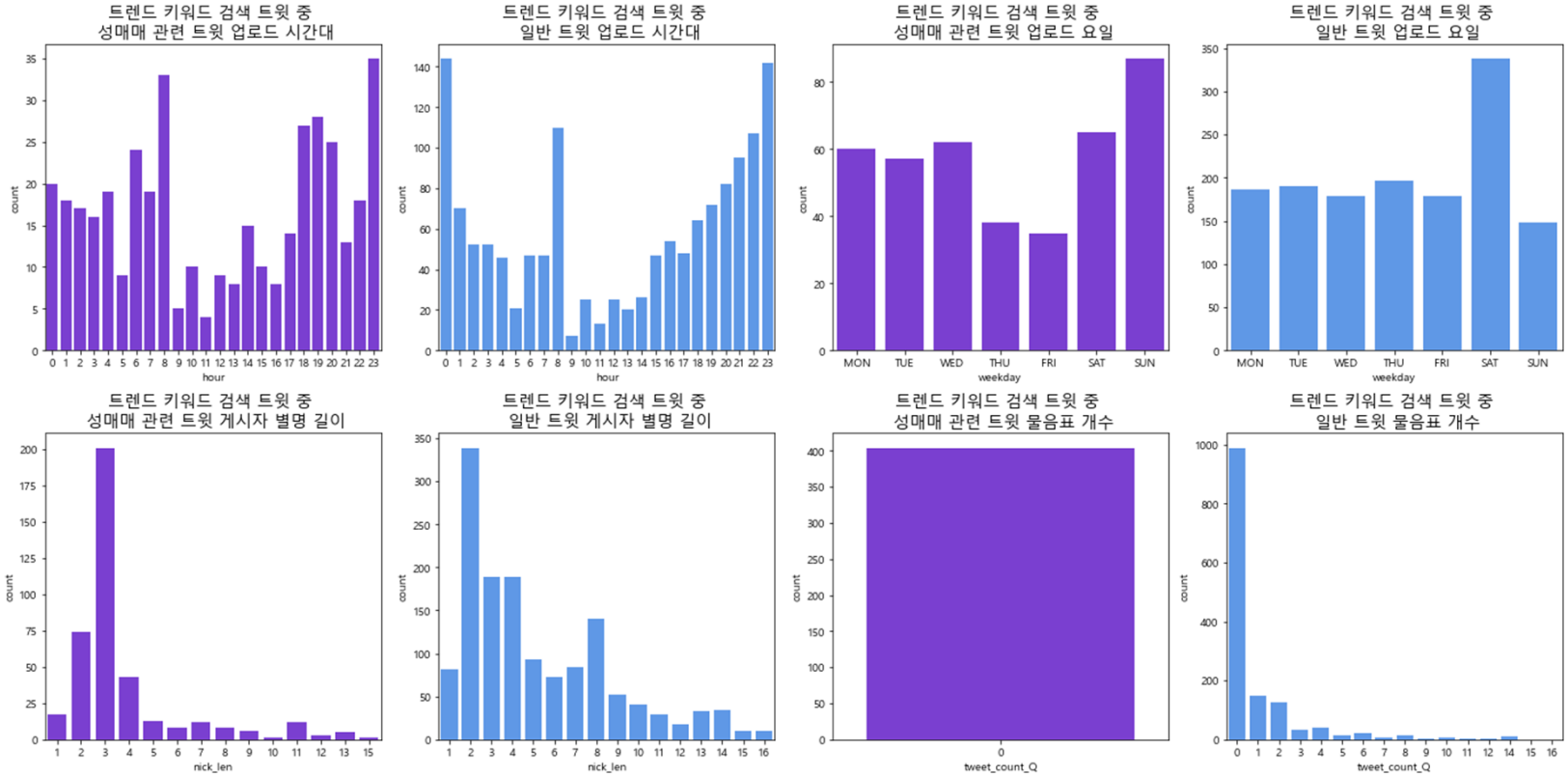
좋아요가 두 개 이상인 트윗이 꽤 있음.

비디오가 없는 일상적 트윗이 좋아요가 더 많음.

성매매 트윗

비디오가 있는 경우 좋아요가 한 개일 때가 특히 많은데 이용자의 유입보다는 인위적인 것일 가능성이 높아 보임.

분석 결과: 형태적 분석



[실시간 트렌드에서 수집한 트윗 내 성매매 트윗: 약 22%]

시간대: 전형적인 성매매 트윗에 비해 불규칙적.

별명 길이: 전형적인 트윗과 유사함.

요일: 일요일에 높은 분포를 보임.

물음표 개수(이모티콘 수): 사용하지 않음.

분석 결과: 내용적 분석

[Tf-Idf 분석] ※특정 문서 내에서 어떤 단어가 얼마나 중요한지 나타냄.

	word	count
1	출장	638842
2	만남	404814
3	콜걸	296808
4	안마	203736
5	조건	183907
6	대행	155016
7	섹파	154527
8	마사지	153722
9	애인	149774
10	오피	145383
11	페이	136444
12	걸	98481
13	여행	84390
14	훈녀	83893
15	노래방	83611
16	맛집	81829
17	휴게소	81605
18	횟집	80860
19	아가씨	62367
20	룸싸롱	59617
21	호텔	52857
22	톡	48500
23	후기	48427
24	모텔	33357
25	강남	32171
26	룸	31626
27	서울	31243
28	부산	29768
29	성남	28945
30		

출현 빈도가 높으면서
성매매와 가장 연관이 높은 단어 추림.
부사, 어간 등 불(不)용어 제외.

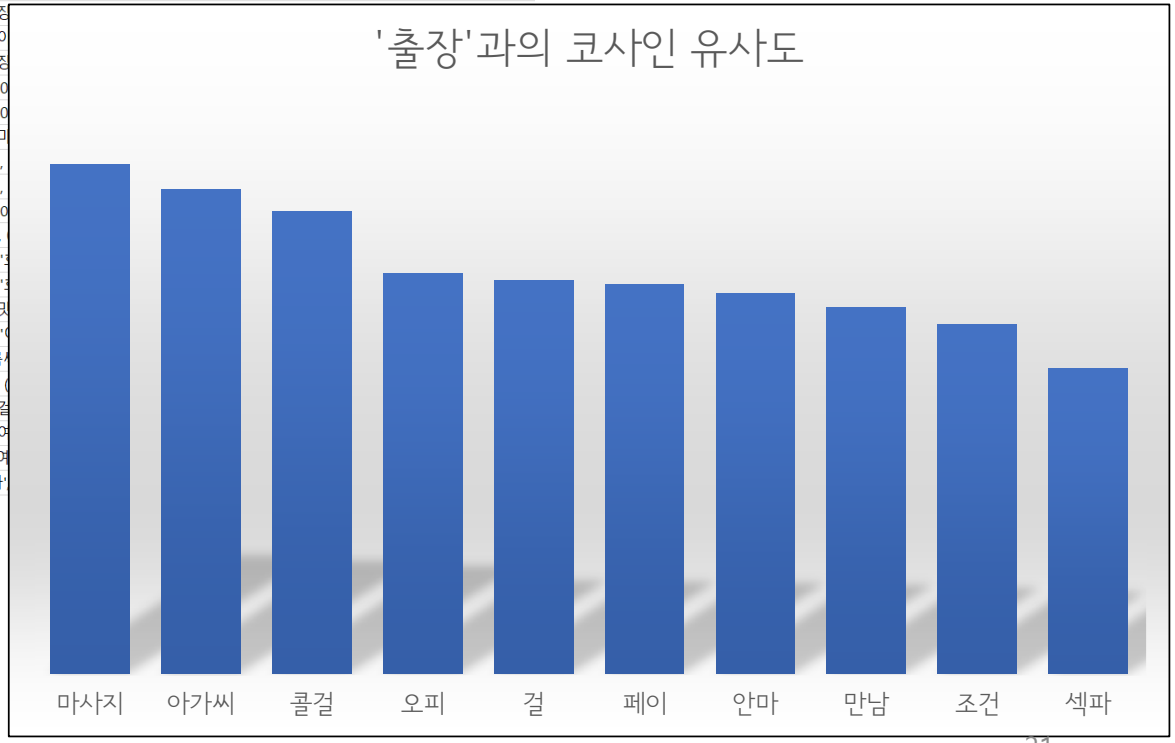
지역명이나
성매매가 이루어지는 장소,
일상적으로 사용하는 용어
등이 주로 나타남.

분석 결과: 내용적 분석

[코사인 유사도 분석]

※ 출현 빈도가 높은 단어를 중심으로 한 트윗 내에서 자주 쓰이는 단어 비율

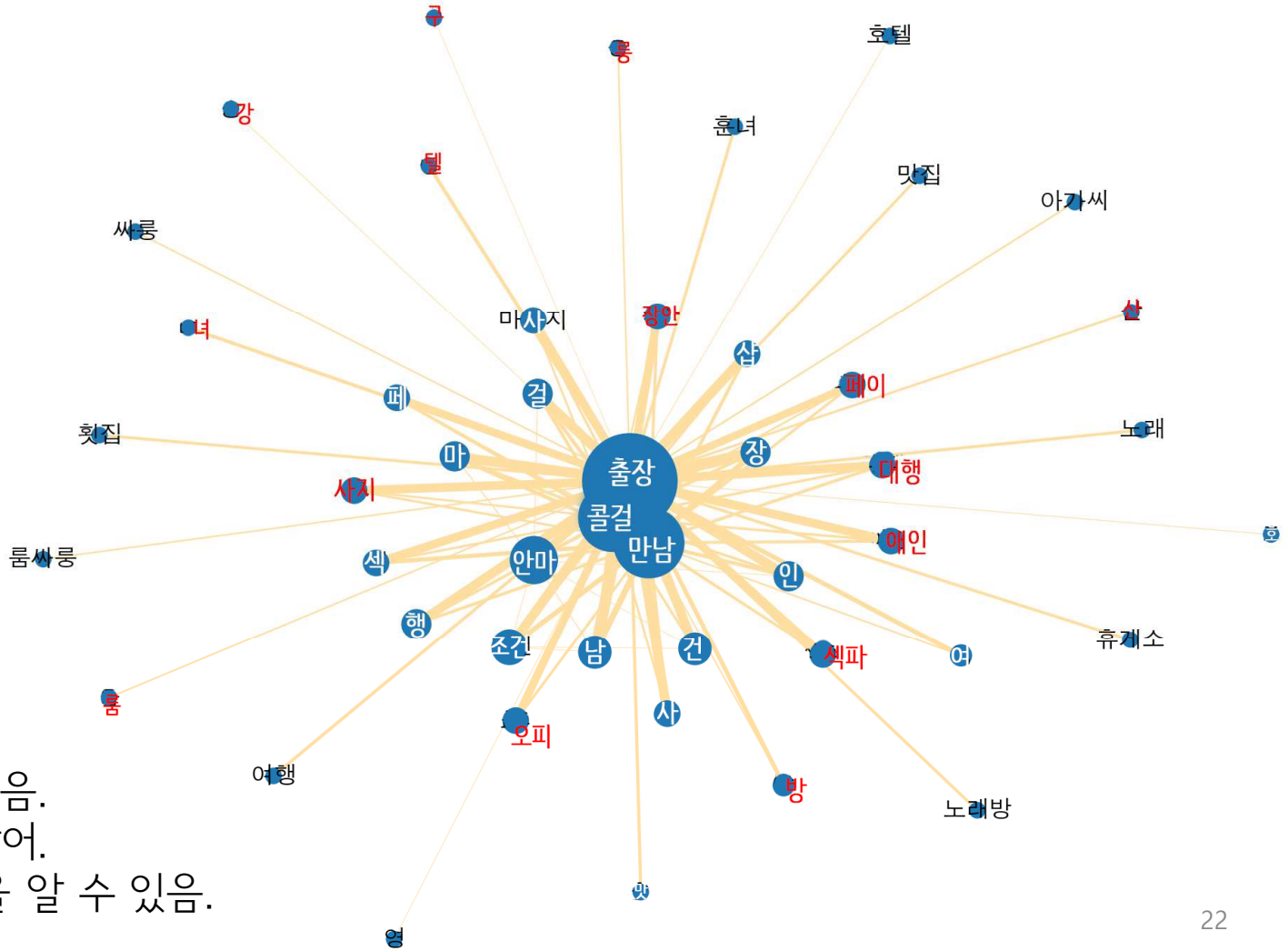
base_word	similar_words
출장	[('마사지', 0.8964150852092784), ('아가씨', 0.8532777011528654), ('콜걸', 0.8133497217353909), ('오피', 0.704724709409685), ('걸', 0.6926344455666812), ('페이', 0.6846136228725918), ('안마', 0.67009098
만남	[('조건', 0.9172036507003127), ('페이', 0.747146490766263), ('오피', 0.7352410114310644), ('출장', 0.6443694214913613), ('안마', 0.630766137096904), ('걸', 0.5268505010002674), ('콜걸', 0.5120752411044
콜걸	[('대행', 0.8549946004726225), ('애인', 0.8420490127471817), ('섹파', 0.8414736569581801), ('안마', 0.8314811785928824), ('출장
안마	[('섹파', 0.8455590188351392), ('대행', 0.8386564422261789), ('콜걸', 0.8314811785928824), ('애인', 0.8246452559755854), ('페이
조건	[('만남', 0.9172036507003127), ('페이', 0.7647100496826337), ('오피', 0.7526935162614034), ('안마', 0.6429894553322567), ('출장
대행	[('애인', 0.9824575569819365), ('섹파', 0.939964509725158), ('콜걸', 0.8549946004726225), ('안마', 0.8386564422261789), ('걸', 0
섹파	[('대행', 0.939964509725158), ('애인', 0.9263186909591159), ('안마', 0.8455590188351392), ('콜걸', 0.8414736569581801), ('걸', 0
마사지	[('출장', 0.8964150852092784), ('아가씨', 0.8178156274612681), ('콜걸', 0.7557878382287598), ('걸', 0.5915827020427039), ('안마
애인	[('대행', 0.9824575569819365), ('섹파', 0.9263186909591159), ('콜걸', 0.8420490127471817), ('안마', 0.8246452559755854), ('툭',
오피	[('페이', 0.9120572946085316), ('안마', 0.7804309036550425), ('조건', 0.7526935162614034), ('만남', 0.7352410114310644), ('걸',
페이	[('오피', 0.9120572946085316), ('안마', 0.8044297407199837), ('조건', 0.7647100496826337), ('만남', 0.747146490766263), ('걸', 0
걸	[('오피', 0.7239927901479479), ('출장', 0.6926344455666812), ('페이', 0.685075246698222), ('안마', 0.678452340032911), ('콜걸',
여행	[('훈녀', 0.854214927925054), ('횟집', 0.8416027989732321), ('노래방', 0.8051284685722635), ('룸싸룽', 0.7976440536575073), ('호
훈녀	[('횟집', 0.9006174033740971), ('노래방', 0.8707999442125088), ('룸싸룽', 0.8587499075808376), ('여행', 0.854214927925054), ('호
노래방	[('횟집', 0.9387578064505028), ('훈녀', 0.8707999442125088), ('룸싸룽', 0.8458764535912366), ('여행', 0.8051284685722635), ('맛
맛집	[('횟집', 0.8444824877369028), ('훈녀', 0.8147157056730153), ('휴게소', 0.8131695359541061), ('노래방', 0.8043357809421696), ('여
휴게소	[('맛집', 0.8131695359541061), ('횟집', 0.7342793224376791), ('노래방', 0.7014359244457257), ('훈녀', 0.7001458402751117), ('룸싸
횟집	[('노래방', 0.9387578064505028), ('훈녀', 0.9006174033740971), ('룸싸룽', 0.8782307606262117), ('맛집', 0.8444824877369028), (
아가씨	[('출장', 0.8532777011528654), ('마사지', 0.8178156274612681), ('콜걸', 0.7192860107188435), ('안마', 0.5387106724233168), ('걸
룸싸룽	[('호텔', 0.8960521048378954), ('횟집', 0.8782307606262117), ('훈녀', 0.8587499075808376), ('노래방', 0.8458764535912366), ('여
호텔	[('룸싸룽', 0.8960521048378954), ('횟집', 0.8353733939427704), ('훈녀', 0.82011157217683), ('노래방', 0.7980810542068546), ('여
툭	[('안마', 0.5485736895508121), ('애인', 0.5454312953617408), ('콜걸', 0.541724595280131), ('대행', 0.5358500159682833), ('섹파',



빈도수가 높을수록 더 노골적인 단어가 함께 사용됨.

분석 결과: 내용적 분석

[네트워크 분석]



간선이 굵으면 동시 출현 빈도가 높음.
노드가 크면 중심성이 높은 핵심 단어.
직접적인 성적 단어가 중심 단어임을 알 수 있음.

분석 결과: 내용적 분석

[네트워크 분석]

군집1 성매매 업소 / 유형 관련	군집2 지역 관련	군집3 직접적인 관련이 없는 기타 단어
'풀싸롱', '마사지', '술', '담배', '대리', '노래방', '부산', '룸싸롱', '알바', '출장', '스타', '가이드', '브라', '추천', '전용', '여성', '마곡', '서산', '아가씨', '문의', '술집', '대구', '휴게텔', '정보', '사이트', '오프', '세종', '강원', '금', '울산', '광주대', '제주', '구매', '만남', '훈남', '서면', '밤', '대구인', '판매', '청주', '핸섬가이', '강서구', '발산'	'강남', '셔츠', '레깅스', '서울', '코로나', '란제리', '강남역', '가라오케', '슈퍼스타', '서초', '예방', '반포', '방', '동미러', '선릉역', '신사', '경기', '수원', '논현', '김포', '부천', '파주', '안산', '의정부', '종로', '분당', '인천', '홍대', '일산', '안양', '신림'	'더킹', '선릉', '강남풀싸롱', '가격', '야구장', '위치', '쓰리아워', '역삼', '줄', '반사', '좋아요', '섹트', '지진', '직업', '쯔위', '목마', '지수', '지마켓', '주성치', '쫄볶', '주식', '환영', '아유' ※ 검색어에 노출되기 위한 것으로 보임.

의의

1	일반 트윗과 대조되는 성매매 트윗의 특성 파악
2	성매매 관련 단어 데이터 베이스 구축
3	트윗을 업로드하는 매크로 시스템에 대응 가능
4	데이터에 기반한 트윗 신고 및 차단을 통한 사용자 불쾌감 감소

한계

1	실시간 트렌드에서 수집한 일반 트윗의 개수 보완
2	일반 트윗에 섞여 있는 성매매 트윗에 관한 분석
3	성매매 트윗 판별 머신 러닝 모델을 구축하지 못한 점


THE TRUE ADMINISTRATION OF JUSTICE IS THE FIRMEST PILLAR OF GOOD GOVERNMENT

감사합니다.