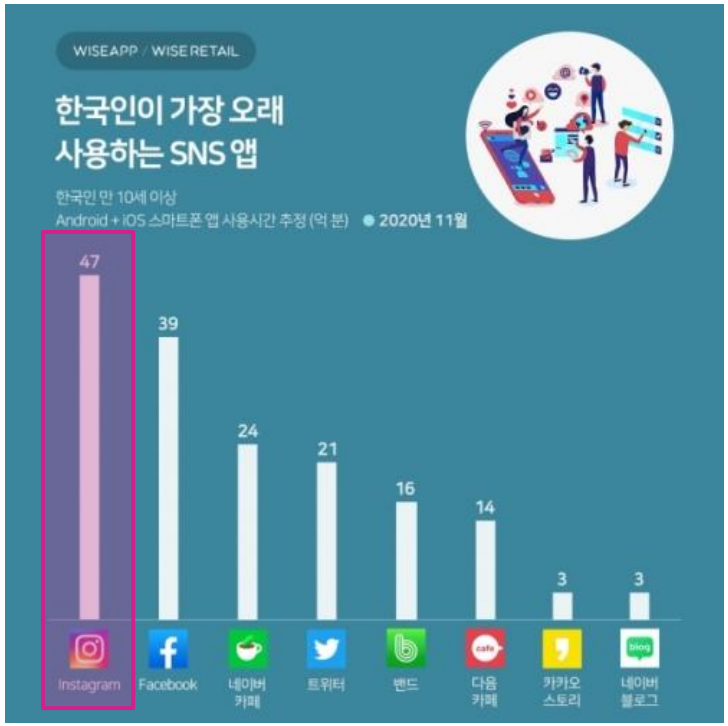
빅데이터 분석 미니프로젝트

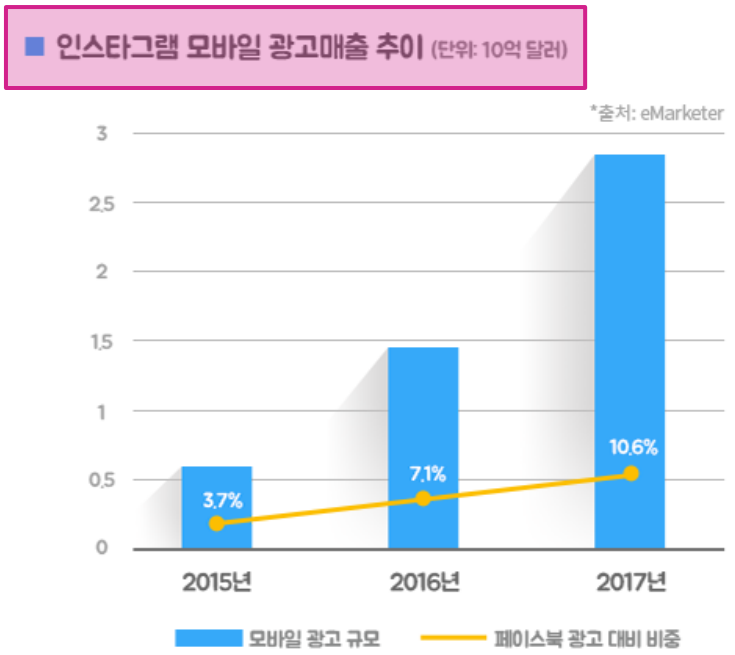
인스타그램 텍스트 시각화와 네이버 검색량과의 상관관계

2조: 최서영, 강용

2020 가장 오래 사용하는 앱

인스타그램은 2020년 한국인이 가장 많이 사용하는 앱으로 선정되었으며, 국내 1위 포털 기업인 네이버의 광고 검색량과 비교하여 인스타그램의 광고 효과에 대해서 알아보고자 합니다.





프로젝트 일정

일 정 (7일)	내 용
2021.03.17 – 03.18	주제선정
2021.03.18 – 03.20	데이터 수집
2021.03.21 – 03.22	데이터 전처리
2021.03.22 – 03.23	데이터 분석
2021.03.23 – 03.23	데이터 시각화

팀 역할

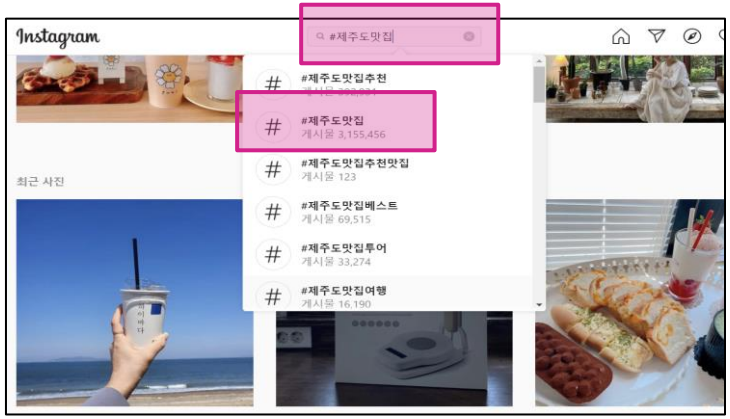
팀 원	키워드	역 할
강용	#제주도맛집	데이터수집, 데이터 전처리, 데이터 분석, 데이터 시각화
최서영	#제주카페	데이터수집, 데이터 전처리, 데이터 분석, 데이터 시각화

웹 크롤링을 이용한 데이터 수집

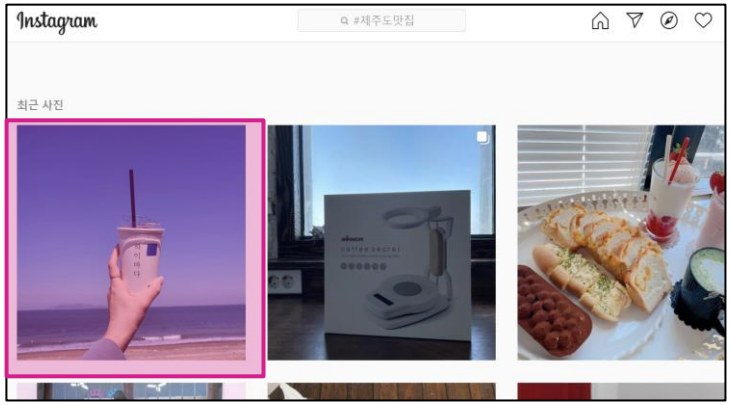
#제주도맛집, **#제주카페** 2개의 키워드를 인스타그램에서 검색하여 **id, 본문내용, 좋아요 수, 댓글, 대댓글**의 정보를 웹 크롤링하여 수집

- 데이터수집 방법: 웹 크롤링
- 데이터수집 키워드: #제주도맛집, #제주카페
- 데이터수집 날짜: 2021. 03. 14 ~ 03. 19
- 데이터수집 정보: id, 본문내용(해시태그 포함), 좋아요수, 조회수, 댓글, 대댓글
- 분석 개발 환경: Rstudio
- 사용 라이브러리: Rselenium, httr, rvest

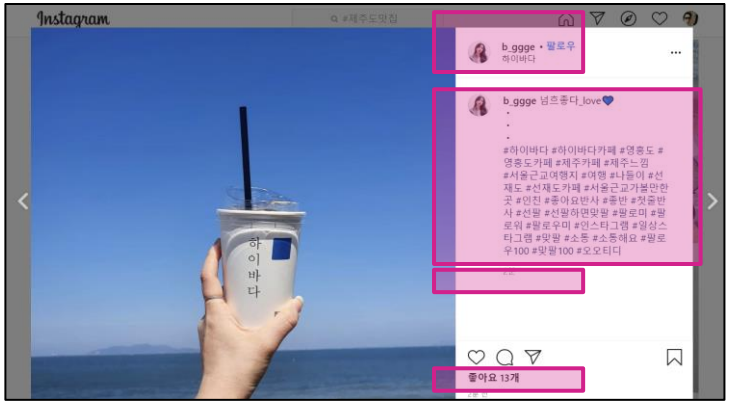
인스타그램 크롤링 작업 순서



키워드 검색



첫번째 컨텐츠 클릭



필요한 데이터 속성값 찾아서 크롤링

우측 이동 버튼을 클릭하여 반복적으로 수행

크롤링 -> csv파일 출력

	id	txt	date	like	video	reply	re_reply
12597	leejxxmxx	우➿야➿호 💛	2021년 3월 14일	좋아요 63개	NA	❤️	@leejxxmxx #
12598	leejxxmxx	치킨마요 참치마요 내마음홈치지마요 💛	2021년 3월 14일	좋아요 65개	NA	❤️	@leejxxmxx #
12599	frito__jeju	제발 먹어주세요 ㅠㅜ 너무 맛있어요😭 . #제주맛집#먹스타...	2021년 3월 14일	좋아요 21개	NA	❤️	#귀덕맛집 #귀
12600	hokkom_	눈,앞 에 탁트인 제주바다를 바라보며 떡볶이, 돌트라왕새우,...	2021년 3월 14일	NA	조회 18회	NA	NA
12601	mustache_jeju_island	. 🌈 알록달록 무지개빛 수제 똥카롱 🦒 100% 유기농 제주 ...	2021년 3월 14일	좋아요 4개	NA	NA	NA
12602	frito__jeju	#마이떠 너무 맛있어서 출추면서 먹음. 제주바다 보면서. 길...	2021년 3월 14일	좋아요 568개	NA	😉👍👍👍	@lesondeleau
12603	jejudostar	제주애월해안도로 꼭 들르길 잘한 카레맛집!! 감성있고 오션...	2021년 3월 14일	좋아요 3개	NA	NA	NA
12604	jejudostar	제주애월해안도로 꼭 들르길 잘한 카레맛집!! 감성있고 오션...	2021년 3월 14일	좋아요 5개	NA	NA	NA
12605	the_help_7	<광고> 비싼 문어가 통으로 들어간 짬뽕!! 맛도 맛이지만 창...	2021년 3월 14일	NA	조회 137회	😍😍😍	NA
12606	jeju9754	샌드위치&브런치&커피 제주여행만족😍💔 오션뷰 카페...	2021년 3월 14일	가장 먼저 좋아요 를 눌러보세요	NA	NA	NA
12607	jeju9754	샌드위치&브런치&커피 제주여행만족😍💔 오션뷰 카페...	2021년 3월 14일	좋아요 1개	NA	NA	NA
12608	jeju9754	모닝커피&브런치맛집 제주여행만족😍💔 오션뷰 카페🍮 ...	2021년 3월 14일	가장 먼저 좋아요 를 눌러보세요	NA	NA	NA
12609	frito__jeju	사랑스러운 만찬 !! ❤️ 완전 내스타일. 넘넘 맛있음 😍 . Rep...	2021년 3월 14일	좋아요 18개	NA	❤️	#귀덕맛집 #귀
12610	kcloset_j	케이애드는 제주도에서 활동하는 인스타그램 마케팅 전문 ...	2021년 3월 14일	가장 먼저 좋아요 를 눌러보세요	NA	NA	NA
12611	NA	NA	NA	NA	NA	NA	NA
12612	NA	NA	NA	NA	NA	NA	NA
12613	NA	NA	NA	NA	NA	NA	NA
12614	NA	NA	NA	NA	NA	NA	NA
12615	NA	NA	NA	NA	NA	NA	NA
12616	NA	NA	NA	NA	NA	NA	NA

Showing 12,597 to 12,616 of 12,641 entries, 7 total columns

약 **12,616개** 행과 **7개**의 열로 구성하는 **데이터프레임 생성**

데이터 타입 변경, 새로운열 추가, 중복된 데이터 제거

```
'data.frame':   12641 obs. of  8 variables:
 $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ id       : chr  "jeju_party24_" "mustache_je
 $ txt      : chr  "최고가 아니면 모시지않습니다.\n
간 빈백 "| __truncated__ "check!<U+0001F447><U+
때 #녹자밭 #제주도 #제주도여행 #"| __truncated__ .
 $ date     : chr  "2021년 3월 19일" "2021년 3월
 $ like     : chr  "가장 먼저 \n좋아요\n를 눌러보세.
 $ video    : chr  NA NA NA NA ...
 $ reply    : chr  NA NA "<U+0001F9F7>" NA ...
 $ re_reply: chr  NA NA "@youngdumbyouth_sale
```

before

```
'data.frame':   8667 obs. of  8 variables:
 $ id         : chr  "jeju_party24_" "mustache_
 $ txt        : chr  "최고가 아니면 모시지않습니다
바다"| __truncated__ "check     skirt38 000
...
 $ date       : chr  "2021년 3월 19일" "2021년
 $ like       : num  0 1 2 12 0 2 NA 3 1 0 ...
 $ video      : num  NA NA NA NA NA NA 2 NA NA
 $ reply      : chr  NA NA "" NA ...
 $ re_reply   : chr  NA NA "  #빈티지 #빈티지패션
 $ like_video: num  0 1 2 12 0 2 2 3 1 0 ...
```

after

특수문자 제거

id	txt
jeju_party24_	최고가 아니면 모시지않습니다. 정직한 주대 ! 사소한 부분까...
mustache_jeju_island	. 제주 핫플레이스 콧수염 카페 😡 여심 저격 피크닉 옥상 빈...
youngdumbyouth_sale	check👇 •skirt38,000->15,000 skirt 허리 33(밴딩), 총 기장 6...
brojeans_pic	삼각대와 리모컨 본전뽑기 . . . . . #오설록 #오설록티뮤지엄 ...
farvento	한치오일파스타+새우크림파스타🤩 신창휴게소 맞은편 언...
youngdumbyouth_sale	check👇 •cardigan 25,000->10,000 free 구매 문의 dm♡
sugartop	*NA*

before

id	txt
jeju_party24_	최고가 아니면 모시지않습니다 정직한 주대 사소한 부분...
mustache_jeju_island	제주 핫플레이스 콧수염 카페 여심 저격 피크닉 옥상 빈백...
youngdumbyouth_sale	check skirt38 000 15 000 skirt 33 63 ...
brojeans_pic	삼각대와 리모컨 본전뽑기 #오설록 #오설록티뮤지...
farvento	한치오일파스타 새우크림파스타 신창휴게소 맞은편 언덕에...
youngdumbyouth_sale	check cardigan 25 000 10 000 free dm
justcallme_chongchong	녹두삼계탕 몸이 따뜻해지고 피로가 풀린다 제주도의 3월은...

after

본문, 댓글, 답글에서 6번 이상 출현한 데이터 선별

제주카페 485	제주여행 231	제주도 206	제주맛집 149	제주도여행 127
제주도카페 107	좋아요 81	데일리 77	좋아요반사 76	서귀포카페 74
제주카페추천 72	제주핫플 72	카페투어 67	팔로우 67	제주카페투어 65

네이버에서의 해당 키워드 검색량 확인

연관키워드 조회 결과 (234개)

전체추가	연관키워드 ?	월간검색수 ?	
		PC	모바일
추가	모알보알	1,110	4,360
추가	제주모알보알	1,220	15,300

출현 빈도, 검색량을 변수로 하는 데이터프레임 생성

```
> head(searchdata)
  X.U.FEFF.num  pc mobile pc_mobile
1          313 170   1780      1950
2          243 380   1590      1970
3          369 140   2270      2410
4          153  30    420       450
5          116 110    830       940
6           24 870   7790      8660
```

텍스트빈도와 검색량간의 상관분석

```
> cor(searchdata$X.U.FEFF.num, searchdata$pc)
[1] -0.1534083
> cor(searchdata$X.U.FEFF.num, searchdata$mobile)
[1] -0.1533433
> cor(searchdata$X.U.FEFF.num, searchdata$pc_mobile)
[1] -0.1535431
```

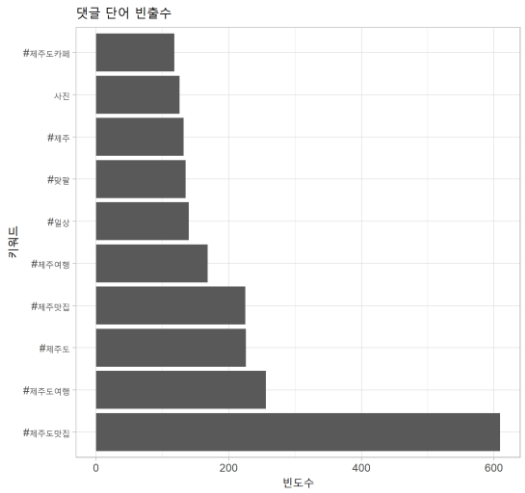
ExtractNoun, 워드클라우드

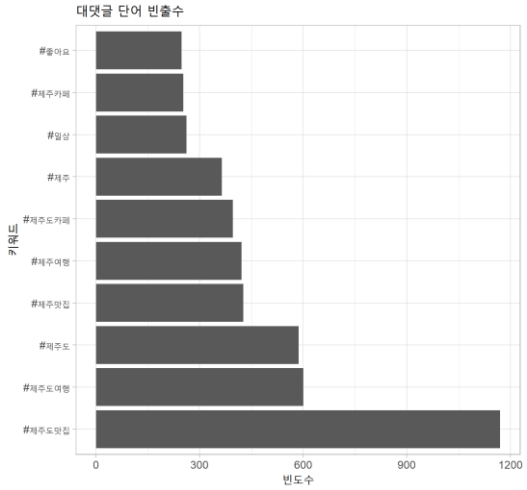
댓글에서 추출

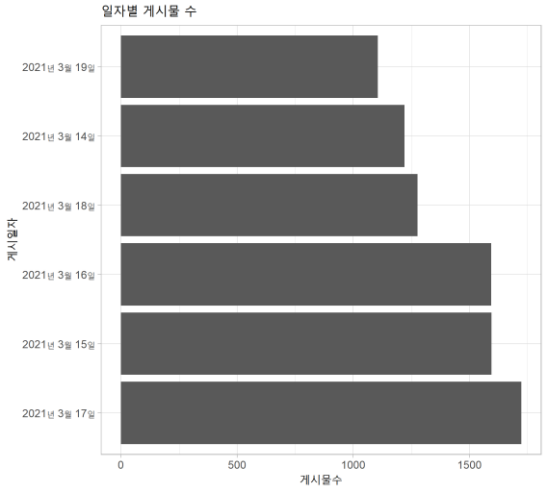
대댓글에서 추출

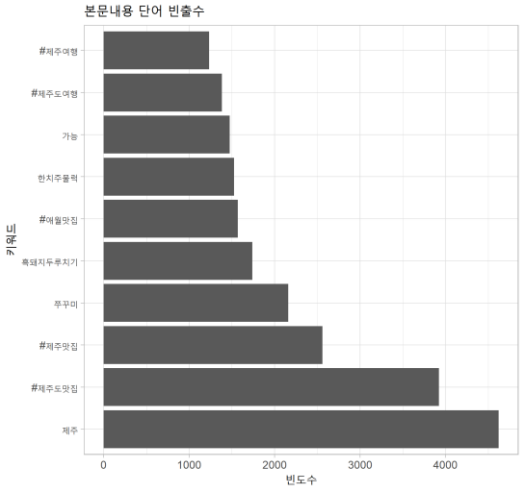
본문내용에서 추출

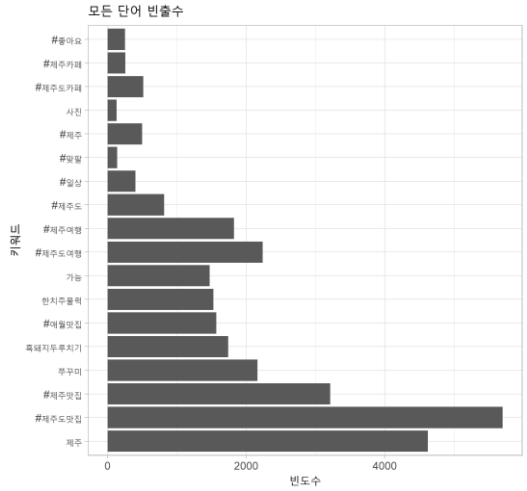
ggplot 시각화
상위 10개의 텍스트 빈도수, 일별 게시물 수











프로젝트 개발 및 분석 후기

최서영

- 동적 웹페이지를 크롤링하는 부분에서 중간에 오류가 난 채 내버려두거나, 반복문이 끝난 후에는 selenium 서버가 꺼져서, 일일이 데이터 수집 과정에 집중해야 하는 부분이 어려웠다.
- 일정기간(일주일)이 지난 인스타그램 피드는 보이지 않아서 월간 검색량에 비해 짧은 데이터만 수집할 수 있어서 아쉬웠다.
- 동적 웹페이지의 자료, 특히 많은 사람들이 사용하고 있는 인스타그램 데이터를 크롤링해서 이용했다는 것이 뿌듯했다.

강용

- **웹크롤링**

컨텐츠의 필요한 정보 반복문을 통해서 동일한 작업을 계속 수행할 경우 인스타그램에서 어느 시점에 정보 제공을 막도록 설정하고 있어서 크롤링을 지속할 수가 없었습니다. 구글링으로 유사 프로젝트를 검색해보았지만 인스타그램을 만개 이상으로 크롤링한 기록은 찾아보기 어려웠고 3만개의 추출하기로한 목표설정을 달성하지 못해 아쉬웠습니다.

- **데이터전처리**

수집한 데이터의 정보가 대부분 텍스트로 구성되어 있으며, 여러가지 특수문자와 외국어로 인하여 전처리 작업이 많은 시간을 할애하게 되어서 분석을 충분히 하는데 제한이 되었습니다.

- **분석**

마찬가지로 데이터가 텍스트로 구성되어 있어서 다각도로 분석하는게 제한이 되었습니다.

- **총평**

이번 프로젝트로 크롤링만큼은 충분히 자신감이 생겼습니다!