**데이터 마이닝**



**진행사항**

**기업 부실 평가 요인 분석  
(기업 별 부정적인 의견을 야기하는 키워드 도출)**

**5조 팀원 : 201311244 전석규**

**201411192 김기연**

**201611220 이소형**

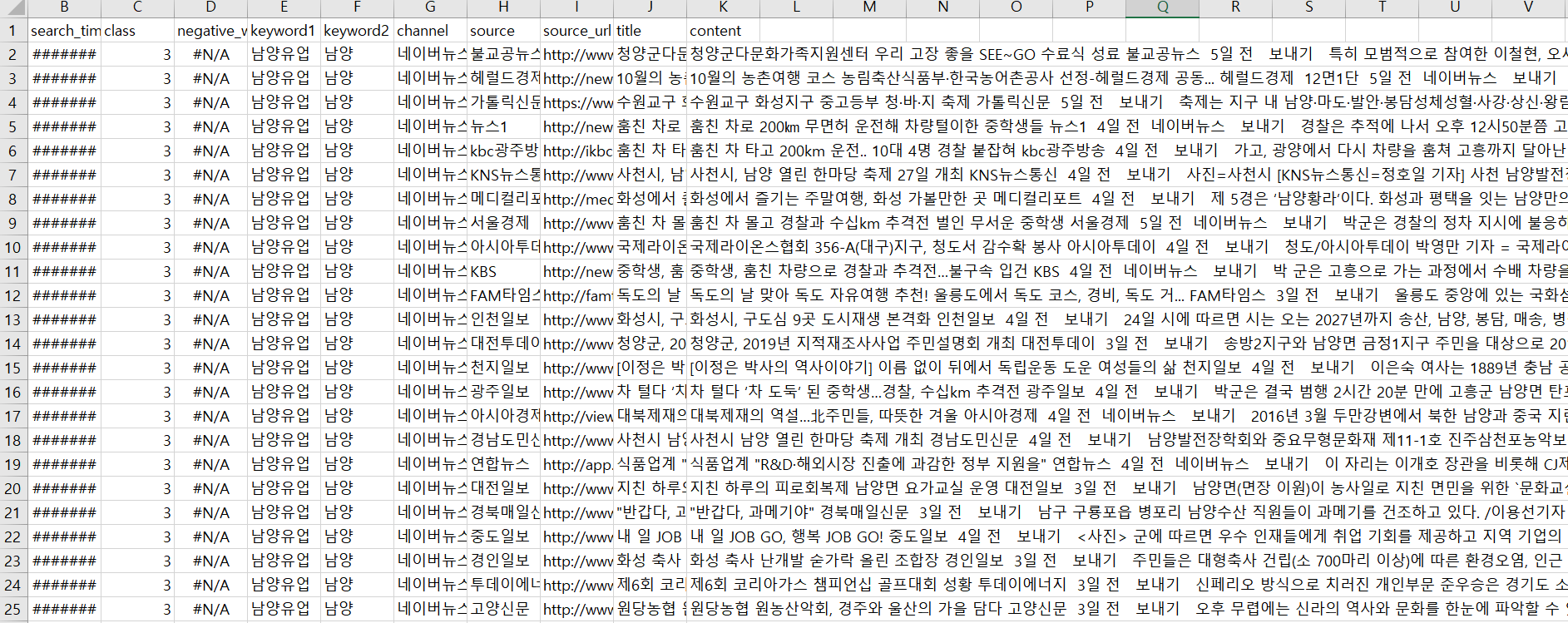
**201611225 이홍인**

**제출일 : 2018-11-14**

1. **데이터 획득**

다니는 회사의 도움을 받아 ‘남양유업’ 검색했을 때 기준 네이버 게시물 크롤링 데이터를 획득했습니다.

챕쳐 이미지)



Search\_time : 본래 크롤링 했을 때의 서버시간이지만 이 프로젝트에서 사용할 것 같지 않아서 깨진 채로 놔뒀습니다.

class : 속성 값으로 사용될 수도 있어서 넣었습니다.

keyword : 프로그램을 사용할 기업 이름입니다.

keyword2 : 해당 단어로 검색한 결과입니다.

channel : 매체(뉴스, 블로그, 카페) 종류입니다.

source : 해당 게시글을 작성한 작성자 정보입니다.

source\_url : 해당 게시글의 url 입니다.

title : 해당 게시글의 제목입니다.

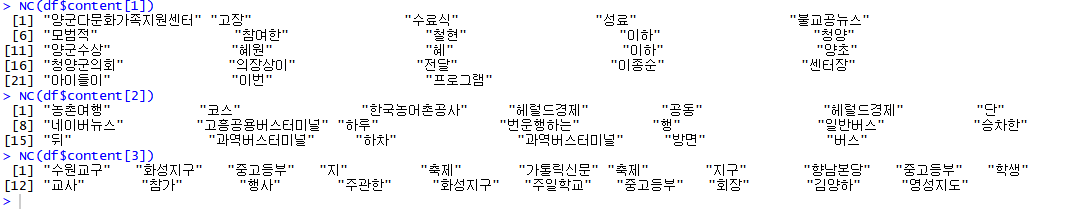
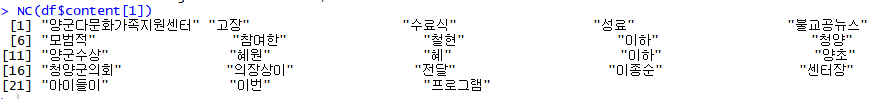
content : 해당 게시글의 글 내용입니다.

1. **진행사항**

먼저 엑셀파일의 데이터를 R로 불러들여서 데이터 프레임에 넣었습니다.

해당 데이터 프레임의 content 컬럼에 있는 데이터(글 내용)에서 simplePos22를 이용하여 보통명사만 추출했습니다.

추출한 사진)

해당 데이터들을 쉼표로 구분하여 한 컬럼으로 합쳐서 raw data 1500줄을 뽑았습니다.

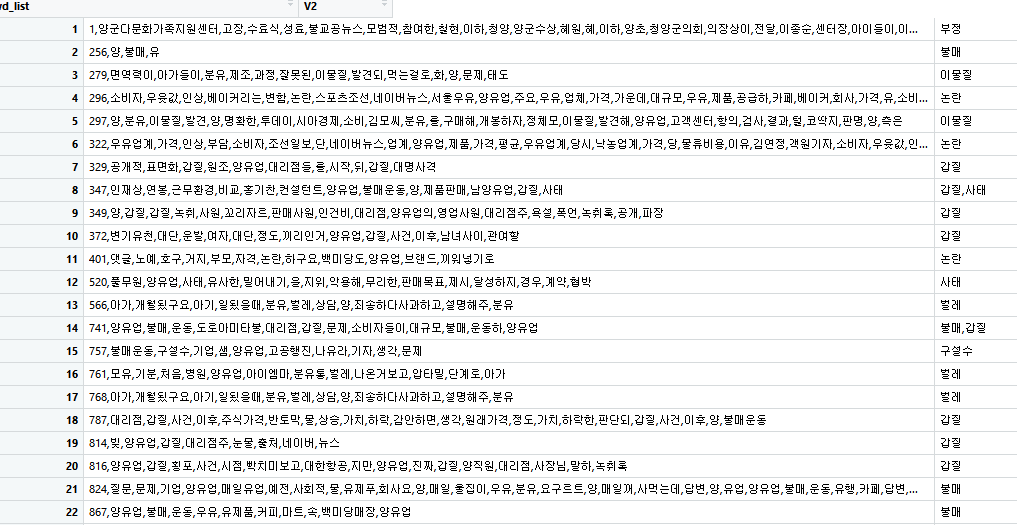
raw data 사진)

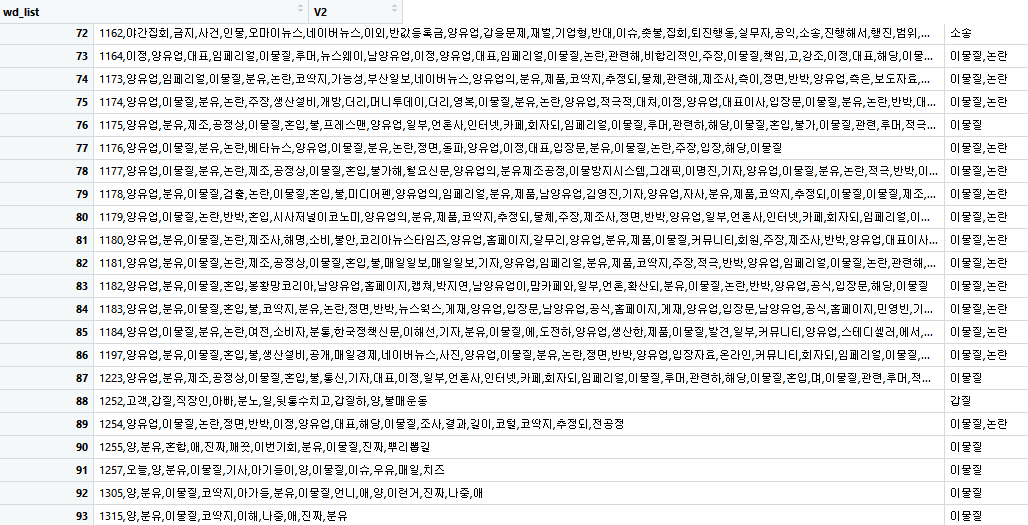


…



raw data에서 부정어("불매", "갑질", "사태", "불량", "소송", "불공정", "이물질", "논란", "구설수", "벌레", "과징금")가 포함돼 있다면 부정 게시물로 판단 후 몇 번째 글이고, 어떤 단어로 인해 부정게시물이 됐는지 정보와 함께 부정게시물 meta data를 만들었습니다.





1500개에서 96개가 나온 모습입니다.

현재 여기까지 진행되었습니다. 앞으로는 부정 단어들을 git에서 뽑아내서 좀 더 다양하게 넣어보고 부정게시물을 다양하게 뽑아서 제대로 된 meta data 생성 후 연관규칙 알고리즘을 통해 같이 자주 나온 새 부정 단어를 추출해볼 예정입니다.

모두 추출됐다면, 매체별로 정리도 해보고 새로 추출된 부정 단어를 다시 넣어 새 부정게시물을 뽑고 또 그와 관련된 새 부정어를 뽑는 방식으로 반복도 해 볼 예정입니다.