

뉴스 마스터

뉴스분석과 키워드 추출

뉴스마스터

18 박현규 20 정은선
20 김하늘 18 김태원
20 이아영

목차

개발환경

프로젝트 구현방법

프로젝트 대표 구현방법

중간 결과

크롤링,wordtovec,chatbot
구현결과

계획

지금까지 진행한 것들
남은 목표



개발환경

어떤 환경에서 작업



뉴스의 동향 파악 향상
의사 결정에 대한 정보 효과적 제공



프로젝트 구현 방법

주요 기술들



크롤링(crawling)

웹사이트(website), 하이퍼링크(hyperlink), 데이터(data), 정보 자원을 자동화된 방법으로 수집, 분류, 저장하는 것.

텍스트 마이닝(text mining)

특정 목적에 맞게 유의미한 정보를 추출하는 분석 및 처리 과정

챗봇(chat bot)

채터봇(Chatterbot)은 음성이나 문자를 통한 인간과의 대화를 통해서 특정한 작업을 수행하도록 제작된 컴퓨터 프로그램이다

중간 결과

크롤링 실행결과

크롤링

	date	title	source	contents	link						
0	1시간 전	LGU+ 반려동물	ZDNet Korea	LG유플러스는	https://zdnet.co.kr/view/?no=20230517085639						
1	23시간 전	반려동물 축적	뉴스시스언론사	기사내용 요약	http://www.newsis.com/view/?id=NISX20230516_0002304551&cID=10401&pID=10400						
2	23시간 전	반려동물 보살	매일경제언론	이통식 반려	https://www.mk.co.kr/article/10737007						
3	19시간 전	구팡, 반려동물	뉴스1	구팡은 반려	https://www.news1.kr/articles/5048505						
4	20시간 전	양평군, 동물보호	연합뉴스	경기 양평군은	https://www.yna.co.kr/view/AKR20230516103100061?input=1195m						
5	20시간 전	반려동물 수의	데일리한국	사진= 구팡 저	https://daily.hankooki.com/news/articleView.html?idxno=957529						
6	19시간 전	서대문구, 반려	이투데이	이투데이=김기	https://www.etoday.co.kr/news/view/2249586						
7	7시간 전	임실에서 반려	동아일보	국내 반려동물	https://www.donga.com/news/Society/article/all/20230516/119316963/1						
8	20시간 전	미넷, 반려동물	이데일리	이번 미넷 가	http://www.edaily.co.kr/news/newspath.asp?newsid=02876566635609576						
9	18시간 전	충북대 수의대	뉴스1	충북대학교는	https://www.news1.kr/articles/5048523						
10	20시간 전	미넷, 반려동물	이데일리	이번 미넷 가	http://www.edaily.co.kr/news/newspath.asp?newsid=02876566635609576						
11	2시간 전	친환경·안전	뉴스시스	특히 어린아이	http://www.newsis.com/view/?id=NISX20230516_0002305458&cID=10433&pID=13000						
12	1일 전	염록스 백신	코메디닷컴	정부는 의심	https://kormedi.com/?p=1590405						
13	18시간 전	대구보건대	뉴스노컷뉴스	대구보건대 보	https://www.nocutnews.co.kr/news/5944446						
14	1일 전	바이오노트, 노	머니S	점점 높아지	http://moneys.mt.co.kr/news/mwView.php?no=2023051608494965186						
15	1일 전	아이·반려동물	한국경제TV	특히 어린아이	http://www.wowtv.co.kr/NewsCenter/News/Read?articleId=A202305160414&t=NN						
16	1일 전	"반려동물과"	레이디경향	더불어 동물	https://lady.khan.co.kr/khlady.html?mode=view&artid=202305160726011&code=11						
17	3시간 전	"배를 많이	할	국민일보	구팡이 반려	https://news.kmb.co.kr/article/view.asp?arclid=0018267618&code=61171811&cp=nv					
18	20시간 전	영진전문대	동	한국대학신문	곽지현 애니	https://news.unn.net/news/articleView.html?idxno=546739					
19	14시간 전	영진전문대	동	NSP통신	곽지현 애니	https://www.nspna.com/news/?mode=view&newsid=637572					
20	1분 전	LGU+, 반려	동	IT비즈니스	염상필 LG유플	https://www.itbiznews.com/news/articleView.html?idxno=99270					
21	47분 전	마이넷닥터,	뉴스	시선뉴스	그러면서 "마	https://www.sisunnews.co.kr/news/articleView.html?idxno=183634					
22	17시간 전	대구보건대	이	베리타스알파	대구보건대는	http://www.veritas-a.com/news/articleView.html?idxno=457194					
23	17시간 전	충북대, 반려	충북일보	오는 21일 청	https://www.inews365.com/news/article.html?no=765124						
24	1일 전	서대문구, 반려	전국매일신문	이 사업은 반	https://www.jeonmae.co.kr/news/articleView.html?idxno=957858						
25	19시간 전	파주시-파주시	충북일보	파주시는 다	http://www.joongboo.com/news/articleView.html?idxno=363592368						
26	23시간 전	구팡, 반려	동물	디스커버리뉴	▲ 구팡, 반려	http://www.discoverynews.kr/news/articleView.html?idxno=1012002					
27	20시간 전	구팡, 수의사	디지털투데이	수의사가 실	http://www.digitaltoday.co.kr/news/articleView.html?idxno=476559						
28	18시간 전	제주시, 무허가	뉴스제주	제주에서 무	http://www.newsjeju.net/news/articleView.html?idxno=389031						
29	16시간 전	대구보건대	보	뉴스프리존	대구보건대학	http://www.newsfreezone.co.kr/news/articleView.html?idxno=469723					
30	19시간 전	서울호서직업	뉴스	쿠키뉴스	서울호서직업	http://www.kukinews.com/newsView/kuk202305160168					
31	20시간 전	반려동물	강남	코프타임즈	반려동물의	http://www.thegoodtimes.co.kr/news/articleView.html?idxno=58016					

크롤링 코드를 실행한 결과, 기사가 올라온 시간, 기사 제목, 기사 본문, 기사 링크를 추출하는 csv파일이 출력됨
뉴스크롤링을 사용하기 위해 html파일을 추가적으로 봐야했다. html에 익숙하지 않아서 XPath와 같은 경로의 의미를 인터넷으로 공부함

중간 결과 텍스트마이닝 실행결과



반려동물	98
동물	25
쿠팡	21
닥터	21
서비스	20

wordtovec

기사본문의 명사들을png파일으로 명사추출
왼쪽사진은 반려동물 뉴스기사 검색 실행결과
NLTK를 활용하여 키워드 추출을 하려고했으나
한글지원이안되어 막혔고, konlpy 를 찾아 키워드 추출

한글 형태소 분석을 지원하는 konlpy 패키지에
서 okt모듈을 사용하여 크롤링한 뉴스기사 중 본문
의 명사를 뽑아 워드클라우드에 사용함

중간결과 link 제공 실행결과

```
(py36) C:\python_test>python links.py
83      https://www.xportsnews.com/article/1716975
Name: link, dtype: object
58      https://sports.donga.com/article/all/20230426/119022124/1
Name: link, dtype: object
63      http://enews.imbc.com/News/RetrieveNewsInfo/380327
Name: link, dtype: object
```

검색한 기사의 url이 실행결과로 뜸

정보를 가져와서 단어 많은 순으로 url 결과를 출력했다

텍스트마이닝에서 뽑은 명사를 모두 링크를 추출하는데 사용하기에는 빈도수가 높은 5개 정도의 단어를 제외하고는 본문에 한 두 번 정도 나오기 때문에 필요하지 않다고 느꼈다

프로젝트 기대효과

우리는 어떤 프로그램을 만들 것인가



다양한 산업 분야에서의 응용이 가능하며, 뉴스에 포함된 다양한 정보들을 적극 활용하여 사람들에게 더욱 효과적으로 정보를 전달하고, 의사 결정에 활용할 수 있도록 도와줄 것임

사회문제에 대한 넓은 인식으로 자기 의사를 결정할 수 있다.

향상된
의사결정

효과적인
커뮤니케이션

뉴스를 쉽게 제공해 사람들의 사회 인지도가 향상된다

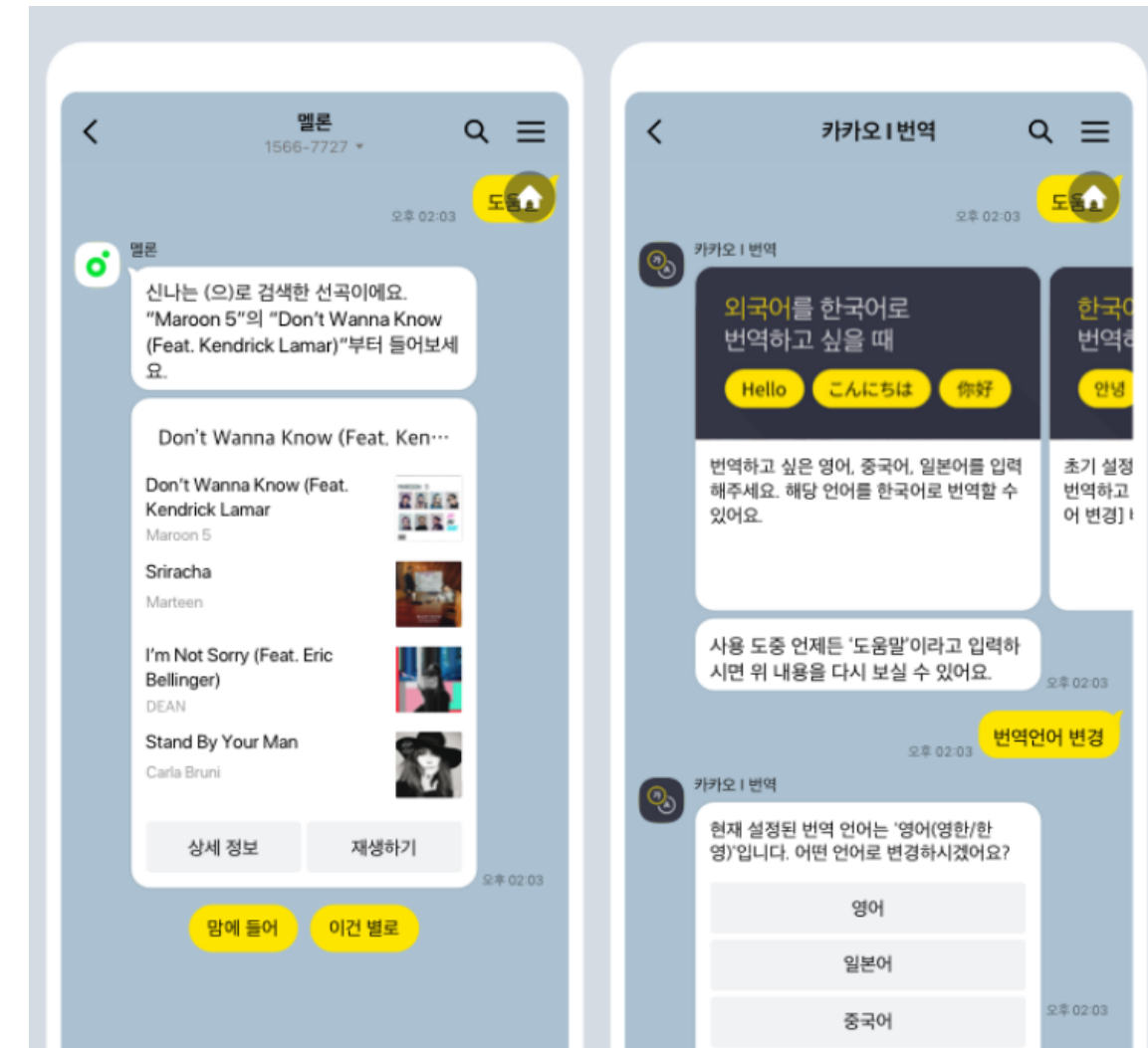
인지도
향상

현재 상황, 계획

남은 목표들, 진행 현황

Details	Period of execution(Month) (Plan : ⇨)																Etc		
	03					04				05					06				
	1	2	3	4	5	1	2	3	4	1	2	3	4	5	1	2		3	4
An idea conference																			
Find source code																			
Modifying source code																			
Finding Crawling Sites																			
Text Mining Test																			
Chatbot Implementation																			
Catbot Test																			
Program Test																			
Creating a result report																			

현재 텍스트마이닝을 하는 목표까지 완료한 상태이며 챗봇을 구현하고 테스트하는 과정이 남은
대략 2주동안 챗봇으로 뉴스데이터를 보여주는 것을 실현할 것임



감사합니다

질문이 있다면 말씀해주세요.