

텍스트에 따른 뉴스 카테고리 분류

AI빅데이터융합경영학과
20182797 김진호



CONTENTS



서론

데이터수집

웹 스크래핑

전처리

불용어 처리
및 전처리

단어빈도

단어빈도표
단어구름

추가분석

모델 학습 및 평가
가중치 분석

주제분석

LDA를 이용한
주제분석

결론



서론 1

서론

N 뉴스 | 연예 | 스포츠 | 날씨 | 프리미엄

언론사별 정치 경제 사회 생활/문화 IT/과학 세계 랭킹 신문보기 오피니언 TV 팩트체크

dM 뉴스 연예 스포츠

홈 사회 정치 경제 국제 문화 IT 연재 포토 팩트체크

많은 뉴스들을 한번에 파악할 수 있는 방법은 없을까?




데이터 수집

2

웹 스크래핑



웹 스크래핑



뉴스 연예 스포츠

홈 사회 정치 경제 국제 문화 IT 랭킹 연재 포토 TV


전체기사 < 2021. 04. 28 > 오늘

최신 사회 정치 경제 국제 문화 연예 스포츠 IT 칼럼 보도자료 자동생성기사

전체기사 건강 생활정보 공연/전시 책 여행레저 문화생활일반 날씨 뷰티/패션 가정/육아 음식/맛집 종교


전동석, 코로나19 확진..음성 받고 자가격리 중 양성 ... 중앙일보 · 21:47

뮤지컬 배우 전동석이 신종 코로나바이러스 감염증(코로나19) 양성 판정을 받았다. 전동석 소속사 빅보스엔터테인먼트는 28일 공식 인스타그램을 통해 "지난 23일 코로...



[날씨] 내일 충청 이남 황사 영향권..일부 지역 우박 YTN · 21:46

불청객 황사가 또다시 전국을 덮쳤습니다. 이에 따라 미세먼지 농도가 3~5배까지 짙게 나타나기도 했는데 내일도 황사의 영향으로 충청과 전북, 경북 지역의 미세먼지 ...



다음 뉴스의 카테고리가 좌측 사진과 같이 8개 정도가 있었습니다. 8개는 너무 많다고 판단되어 이 중 5개인 '정치', '경제', '국제', '문화', 'IT'를 뽑아 카테고리를 분류하기로 결정했습니다.

웹 스크래핑 방법으로는 requests를 사용하여 2022년 5월 전체 기사를 일별로 3 페이지씩 뽑아 카테고리별로 기사를 크롤링 했으며 중복되는 기사가 있을 수 있다고 생각하여 완전 똑같은 기사들은 제거해주었습니다.



전처리 3

불용어 처리 및 전처리



불용어 처리 및 전처리

RANKS NL >> Request Free Plan Demo Plans & Pricing More ▾

Korean Stopwords

Home > Resources > Stopwords > Korean

Korean Stopwords

아	어찌됐든	하기보다는
휴	그위에	차라리
아이구	게다가	하는 편이 낫다
아이쿠	점에서 보아	흐흐
아이고	비추어 보아	놀라다
어	고려하면	상대적으로 말하
나	하게될것이다	자면
우리	일것이다	마치
저희	비교적	아니라면
따라	좀	싫
의해	보다더	그렇지 않으면
을	비하면	그렇지 않다면
를	시키다	안 그러면
에	하게하다	아니었다면
의	할만하다	하든지
가	의해서	아니면
으로	연이서	이라면
로	이어서	좋아
에게	잇따라	알았어
뿐이다	뒤따라	하는것도
의거하여	뒤이어	그만이다
근거하여	결국	어쩔수 없다
입각하여	의지하여	하나
기준으로	기대여	일
예하면	통하여	일반적으로
예를 들면	자마자	일단
예를 들자면	더욱더	한편으로는

▲ 불용어사전 출처 : [Korean Stopwords \(ranks.nl\)](https://ranks.nl/)

기호처리 출처 : [\[Python\] Bag of Words + Sentiment Analysis - Replet \(textmining.kr\)](https://textmining.kr/)

좌측 사진은 [Korean Stopwords \(ranks.nl\)](https://ranks.nl/)에서 제공하는 한국어 불용어 사전입니다. 이 사이트에서 한국어 불용어를 가져와 엑셀파일에 저장한 후 리스트로 변환하여 불용어 처리를 했습니다.

그와 함께 [\[Python\] Bag of Words + Sentiment Analysis - Replet \(textmining.kr\)](https://textmining.kr/)의 '텍스트 전처리 프로세스'를 참고하여 BeautifulSoup 패키지를 사용한 전처리를 했습니다. get_text함수로 태그나 마크업 기호를 빼주었고, re 패키지를 사용하여 한글이 아닌 것을 공백으로 바꾸어주었습니다.

그 후, stanza를 사용하여 명사만 추출하는 함수를 사용하였습니다.



단어 빈도 4

단어빈도표

단어구름



단어 빈도표

CountVectorizer를 사용하여 빈도순으로 단어빈도표를 만들었습니다.

- 정치 기사

	단어	빈도
13911	윤석열	274
5017	대통령	266
14319	이렇게	217
2336	국민의힘	202
13888	윤	177
14595	이준석	172
14083	의원	134
14124	의원은	133
5066	대통령이	125
16595	제가	125

정치 단어빈도표

- 경제 기사

	단어	빈도
5686	미국	124
18050	화물연대	93
11219	원	85
9747	안전운임제	84
10848	올해	78
1894	국내	78
18053	화물연대는	74
3899	대비	73
2185	금리	67
2173	글로벌	66

경제 단어빈도표

- 국제 기사

	단어	빈도
1948	미국	89
5002	중국	53
1574	러시아	52
3918	우크라이나	48
5575	코로나	48
1116	뉴시스	45
2018	바이든	33
4368	일본	31
1961	미국의	31
442	공감언론	31

국제 단어빈도표

- 문화 기사

	단어	빈도
22169	코로나	96
11788	소나기가	70
19562	주	61
21801	치료	54
7987	미국	48
13906	알코올	48
6509	로또	48
21351	최고	47
17448	일부터	46
1187	경기	45

문화 단어빈도표

- IT 기사

	단어	빈도
4138	누리호	249
7394	발사	220
18769	차	133
7462	발사체	119
2311	국내	106
15173	이송	104
979	게임	86
6837	문제가	86
2940	기상	81
688	갤럭시	80

IT 단어빈도표

단어구름

Pixabay(<https://pixabay.com/ko/>) 의 무료이미지 중 카테고리별로 사진을 뽑아와 단어구름에 색과 모양을 적용했습니다.

단어구름에 색과 모양을 적용하기 위해 <https://pinkwink.kr/1029> 를 참고했습니다.

- 정치 기사



앞 페이지에서 만들었던 단어빈도표의 결과와 같이 '윤석열', '대통령', '국민의힘' 등 단어가 크게 그려진 것을 볼 수 있습니다.

- 경제 기사



앞 페이지에서 만들었던 단어빈도표의 결과와 같이 '미국', '화물연대' 등의 단어가 크게 그려진 것을 볼 수 있습니다.

단어구름

Pixabay(<https://pixabay.com/ko/>) 의 무료이미지 중 카테고리별로 사진을 뽑아와 단어구름에 색과 모양을 적용했습니다.

단어구름에 색과 모양을 적용하기 위해 <https://pinkwink.kr/1029> 를 참고했습니다.

- 국제 기사



앞 페이지에서 만들었던 단어빈도표의 결과와 같이 '미국', '러시아', '우크라이나' 등의 단어가 크게 그려진 것을 볼 수 있습니다.

- 문화 기사



앞 페이지에서 만들었던 단어빈도표의 결과와 같이 '코로나', '치료', 등의 단어가 크게 그려진 것을 볼 수 있습니다.

단어구름

Pixabay(<https://pixabay.com/ko/>) 의 무료이미지 중 카테고리별로 사진을 뽑아와 단어구름에 색과 모양을 적용했습니다.
단어구름에 색과 모양을 적용하기 위해 <https://pinkwink.kr/1029> 를 참고했습니다.

- IT 기사

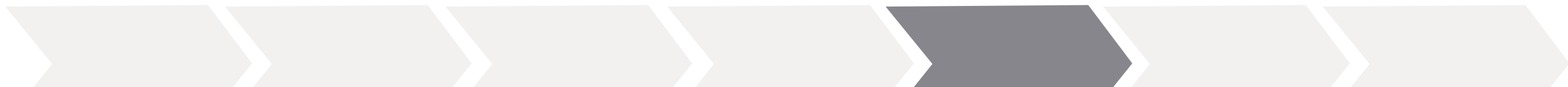


앞 페이지에서 만들었던 단어빈도표의 결과와 같이
'누리호', '발사체' 등의 단어가 크게 그려진 것을 볼 수 있습니다.



추가분석 5

모델 학습 및 평가
가중치 분석



모델 학습 및 평가

Tf-idf로 만든 tdm으로 감성분석을 진행했습니다.

그 후 <https://wikidocs.net/22933> 을 참고하여 감성분석을 진행했습니다.

카테고리가 5개라 다중분류이기 때문에 tensorflow를 이용하여 label을 one-hot 인코딩 해준 뒤 softmax함수로 변경해주었습니다.

또, loss함수를 categorical_crossentropy로 변경해주었습니다.

그리고 과적합을 막기 위한 EarlyStopping과, 정확도가 이전보다 좋아질 경우에만 모델을 저장하도록 하는 ModelCheckpoint를 사용하여 callbacks를 지정해주었습니다.

감성분석 결과는 다음과 같이 0.6423의 정확도를 보였습니다.

```
1 model.evaluate(x_test.A, y_test)
```

```
9/9 [=====] - 0s 8ms/step - loss: 1.3495 - accuracy: 0.6423
```

```
[1.3494668006896973, 0.6423357725143433]
```

가중치 분석

카테고리가 5개이기 때문에 각 카테고리 별로 가중치 분석 결과를 보았습니다.

5개의 카테고리 모두 최고 가중치가 0.2를 넘지 못하였습니다.

- 정치 기사

	토큰	가중치_0
44366	윤석열	0.197258
6912	국민의힘	0.168744
40320	열린	0.166397
14785	대통령	0.165915
47906	입장을	0.158061
44833	의원	0.156356
37676	앞서	0.151696
26666	북한의	0.151629
44875	의원	0.151203
14838	대통령이	0.150693

- 경제 기사

	토큰	가중치_1
43736	유가	0.135520
14405	대비	0.134349
65233	한국경제	0.133098
67941	화물연대는	0.128151
65235	한국경제티브이	0.125723
21520	미국	0.125371
7183	국토교통부와	0.123501
40081	연방준비제도	0.122411
46678	인상	0.120966
17184	따른	0.120768

- 국제 기사

	토큰	가중치_2
42965	워싱턴	0.090781
21543	미국의	0.090146
17666	러시아	0.088777
62320	특파원	0.088717
67288	현지	0.085917
54161	중국	0.079556
25183	보도했다	0.079235
8934	기사내용	0.079021
42182	요약	0.075201
21540	미국으로	0.074935

가중치 분석

- 문화 기사

	토큰	가중치_3
67040	헬스조선	0.140308
1292	강원	0.134253
58275	최고	0.133443
32200	소나기가	0.128175
15342	도로	0.124150
8971	기상캐스터	0.117041
60844	클립아트코리아	0.116978
36792	아침	0.116332
39829	연구	0.112512
4655	곳곳에	0.112024

- IT 기사

	토큰	가중치_4
12697	누리호	0.165815
23097	발사	0.164112
23173	발사체	0.146999
4530	고흥	0.146197
65449	한국항공우주연구원	0.143313
23098	발사가	0.141991
12703	누리호를	0.141487
52653	조립동에서	0.138250
8955	기상	0.137137
12700	누리호는	0.136980



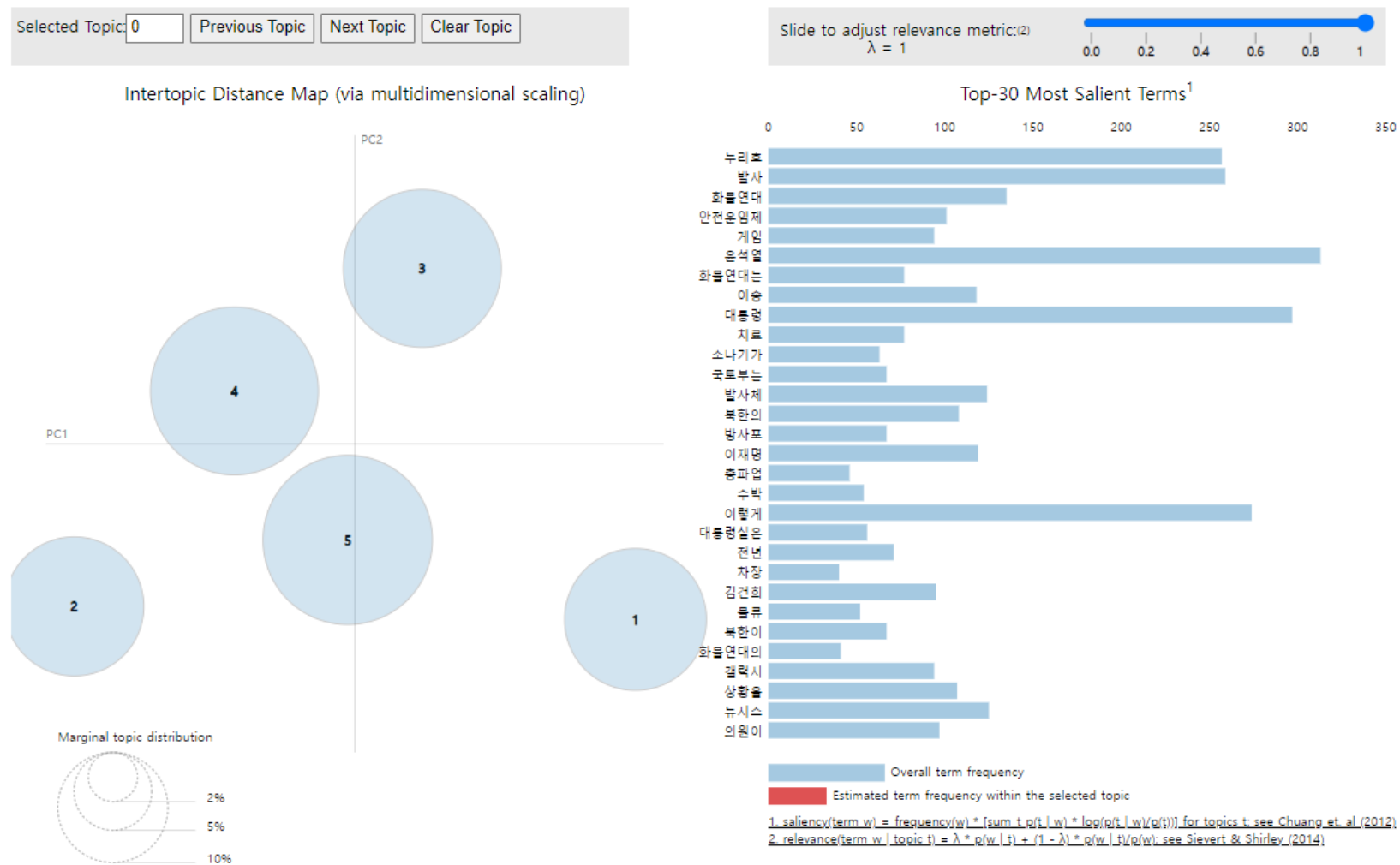
주제 분석

6

LDA를 이용한 주제 분석



LDA 를 이용한 주제 분석



분류가 5개였기 때문에 LDA 토픽수를 5개로 임의로 지정해서 모델링



결론 7

결론

전반적으로 보았을 때, 단어구름과 단어빈도표를 통한 카테고리별 빈도수에 따른 단어를 명확히 확인할 수 있었습니다.

그리고, 감성분석으로 카테고리를 분류해본 것도 적당한 정확도를 보이며 잘 분류되었습니다.

그러나 가중치 분석에서 정확도에 비해 가중치가 매 카테고리에서 모두 낮게 나왔다는 점이 아쉬웠습니다. 주제분석을 통해 알아본 카테고리별 주제는 카테고리의 특성이 잘 드러나게 주제가 분류되었던 것을 볼 수 있었습니다.



감사합니다