


텍스트에 따른 뉴스 카테고리 분류

2020.04.29

빅데이터경영통계전공

20192784 윤경서



CONTENTS



서론	데이터수집	전처리	단어빈도	감성분석	주제분석	결론
	웹 스크래핑	불용어 처리 및 전처리	단어빈도표 단어구름	모델 학습 및 평가 가중치 분석	LDA를 이용한 주제분석	



서론 1

서론

네이버, 다음과 같은 사이트에 접속하면 다양한 카테고리의 기사들이 뜹니다.

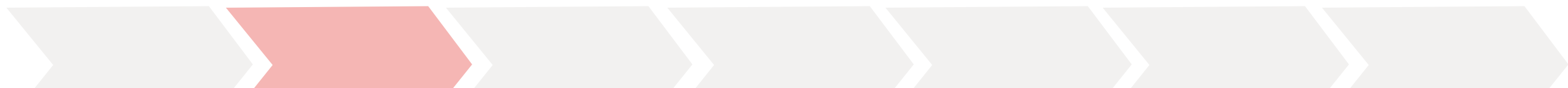
이러한 기사들이 어떠한 기준을 가지고 카테고리가 나누어지는지 보고 싶어 다음기사의 카테고리 분류 분석을 하게 되었습니다.

단어빈도표와 단어구름으로 4월의 각 카테고리별로 빈도 수에 따라 빈번하게 등장했던 단어 수를 알아보고, 감성분석으로 위 단어들에 따라 얼마나 잘 분류가 되는지 알아본 후, 주제 분석을 통해 각 카테고리별로 주제가 잘 분류 되는지 알아보려 합니다.




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 ...



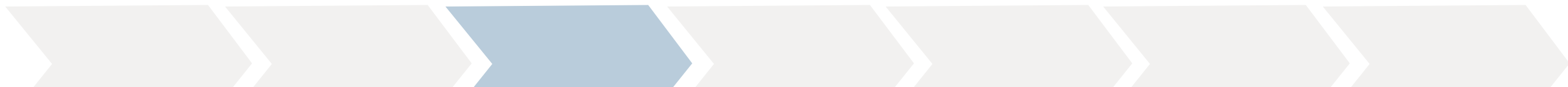
다음 뉴스의 카테고리가 좌측 사진과 같이 8개 정도가 있었습니다. 8개는 너무 많다고 판단되어 이 중 5개인 ‘사회’, ‘경제’, ‘국제’, ‘문화’, ‘IT’를 뽑아 카테고리를 분류하기로 결정했습니다.

웹 스크래핑 방법으로는 requests를 사용하여 4월1일부터 22일까지의 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	싹
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/stopwords/korean)

기호처리 출처 : [\[Python\] Bag of Words + Sentiment Analysis - Replet \(textmining.kr\)](https://textmining.kr/2016/04/python-bag-of-words-sentiment-analysis-replet/)

좌측 사진은 [Korean Stopwords \(ranks.nl\)](https://ranks.nl/stopwords/korean)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/2016/04/python-bag-of-words-sentiment-analysis-replet/)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, stanza를사용하여 명사만 추출하는 함수를 사용하였습니다.



단어 빈도

4

단어빈도표

단어구름



단어 빈도표

CountVectorizer를 사용하여 빈도순으로 단어빈도표를 만들었습니다.

- 사회 기사

	단어	빈도
98	확진자	404
75	지역	352
30	발생	287
40	사업	284
83	코로	279
33	백신	277
71	조사	260
69	접종	259
76	지원	254
8	경찰	222

사회 단어빈도표

- 경제 기사

	단어	빈도
48	시장	532
94	한국	396
38	사업	393
19	기업	391
61	원	376
88	투자	374
66	정부	360
44	서울	355
29	미국	347
0	가격	312

경제 단어빈도표

- 국제 기사

	단어	빈도
63	일본	1411
73	중국	1387
67	정부	747
25	미국	665
32	백신	494
54	오염수	416
82	코로	398
61	인도	392
92	현지	373
90	한국	354

국제 단어빈도표

- 문화 기사

	단어	빈도
48	서울	356
81	지역	310
35	백신	301
19	내일	293
14	기온	287
96	확진자	280
32	발생	278
89	코로	274
71	접종	212
69	전국	199

문화 단어빈도표

- IT 기사

	단어	빈도
17	기술	631
45	서비스	604
30	미국	530
18	기업	526
54	시장	500
31	반도체	476
12	국내	461
87	투자	405
76	지원	401
3	개발	394

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.

- 사회 기사



◀ pixabay에서 가족 사진을 가져와
단어구름에 적용했습니다.

앞 페이지에서 만들었던 단어빈도표의 결과와 같이 ‘확진자’, ‘지역’, ‘발생’ 등의 단어가 크게 그려진 것을 볼 수 있습니다.

- 경제 기사



◀ pixabay에서 지폐 사진을 가져와
단어구름에 적용했습니다.

앞 페이지에서 만들었던 단어빈도표의 결과와 같이 ‘시장’, ‘한국’ 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



◀ pixabay에서 지구 사진을 가져와
단어구름에 적용했습니다.

앞 페이지에서 만들었던 단어빈도표의 결과와 같이 ‘일본’, ‘중국’, ‘정부’ 등의 단어가 크게 그려진 것을 볼 수 있습니다.

- IT 기사



◀ pixabay에서 컴퓨터 사진을 가져와
단어구름에 적용했습니다.

앞 페이지에서 만들었던 단어빈도표의 결과와 같이 ‘기술’, ‘서비스’, ‘미국’, ‘기업’ 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.

- 문화 기사



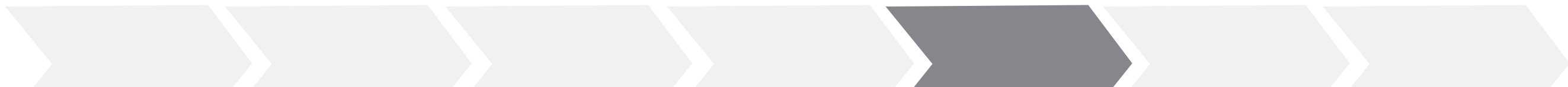
◀ pixabay에서 चु추는 사진을 가져와
단어구름에 적용했습니다.

앞 페이지에서 만들었던 단어빈도표의 결과와 같이 ‘서울’, ‘지역’,
‘백신’ 등의 단어가 크게 그려진 것을 볼 수 있습니다.



감성 분석 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 0.7841의 정확도를 보였습니다.

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

```
19/19 [=====] - 0s 2ms/step - loss: 1.2013 - accuracy: 0.7841
```

```
[1.201332688331604, 0.7841105461120605]
```

가중치 분석

카테고리가 5개이기 때문에 각 카테고리 별로 가중치 분석 결과를 보았습니다.
5개의 카테고리 모두 최고 가중치가 0.4를 넘지 못하였습니다.

- 사회 기사

	토큰	가중치_0
18794	소중	0.233155
1316	검색	0.235723
11065	메일	0.241715
40279	혐의	0.243284
31157	조사	0.244482
15407	불	0.247600
39187	한상욱	0.252801
1929	경찰	0.267875
25374	울산	0.269161
12697	바삭바삭	0.316723

- 경제 기사

	토큰	가중치_1
26552	은행	0.217192
7705	달러	0.220249
4737	금융	0.220273
21179	실적	0.220598
1090	거래	0.221012
25484	원	0.221372
32426	지수	0.223293
508	감소	0.225523
38670	하락	0.260982
17041	상승	0.275246

- 국제 기사

	토큰	가중치_2
41046	회견	0.272297
8048	대변인	0.280195
8714	도쿄	0.287155
9893	로이터	0.302842
30517	정부	0.319359
36613	통신	0.332228
40210	현지	0.338982
31946	중국	0.351237
14439	보도	0.361730
28217	일본	0.380925

가중치 분석

- 문화 기사

	토큰	가중치_3
6135	끝	0.239300
35215	카툰	0.245662
6566	내륙	0.246843
12669	바람	0.248868
5047	기사문	0.255183
6542	낮	0.255194
5175	기온	0.275155
21760	아침	0.277749
6376	날씨	0.279201
6621	내일	0.279559

- IT 기사

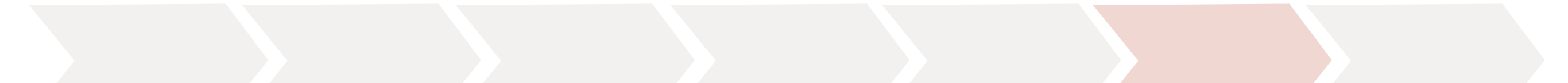
	토큰	가중치_4
30500	정보	0.249023
40957	활용	0.256570
4645	글로벌	0.258017
3890	국내	0.260890
5014	기반	0.265069
23817	연합	0.290374
34693	출시	0.293720
9373	디지털타임스	0.297160
17525	서비스	0.298629
5080	기술	0.302853



주제 분석

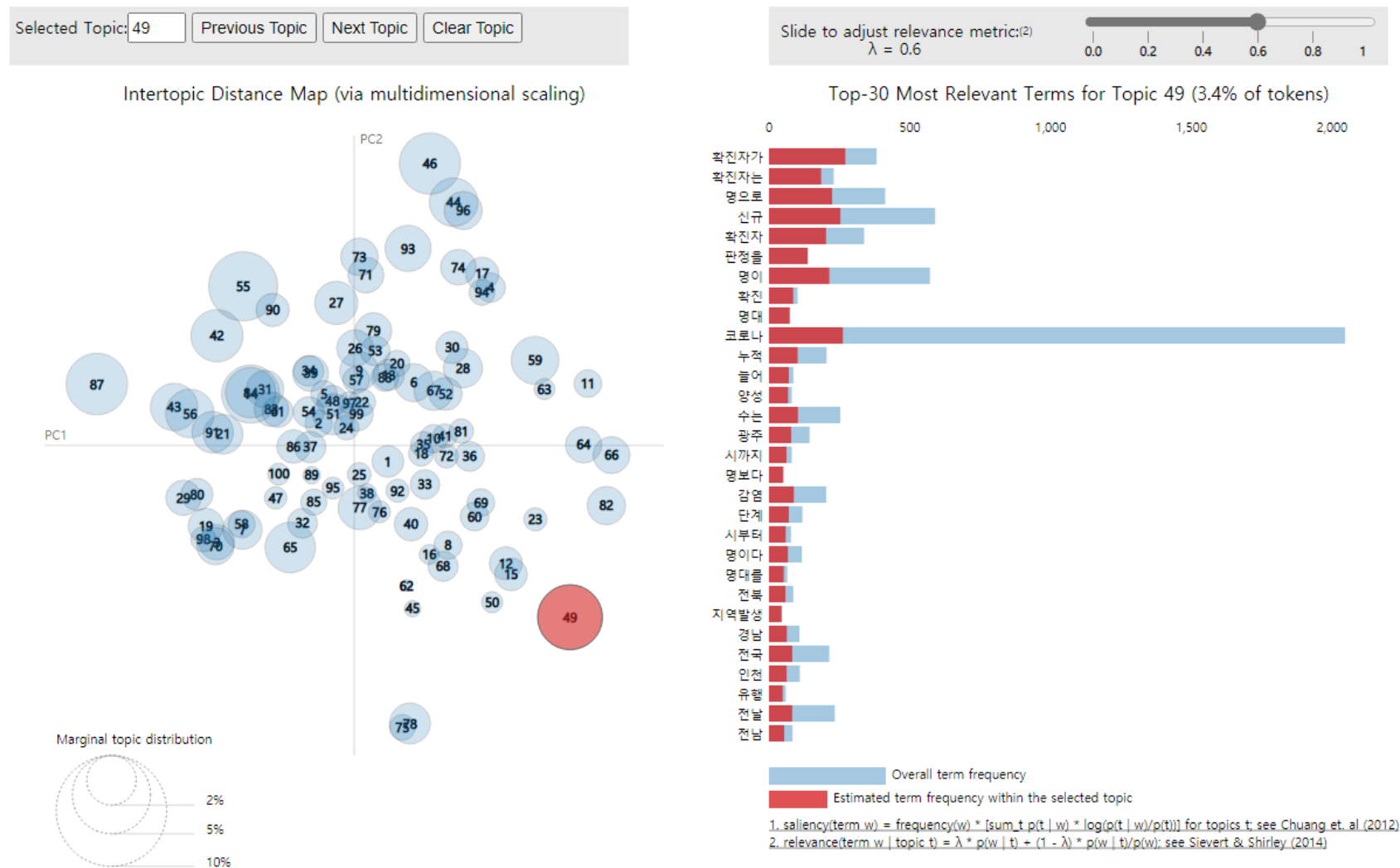
6

LDA를 이용한 주제 분석



LDA 를 이용한 주제 분석

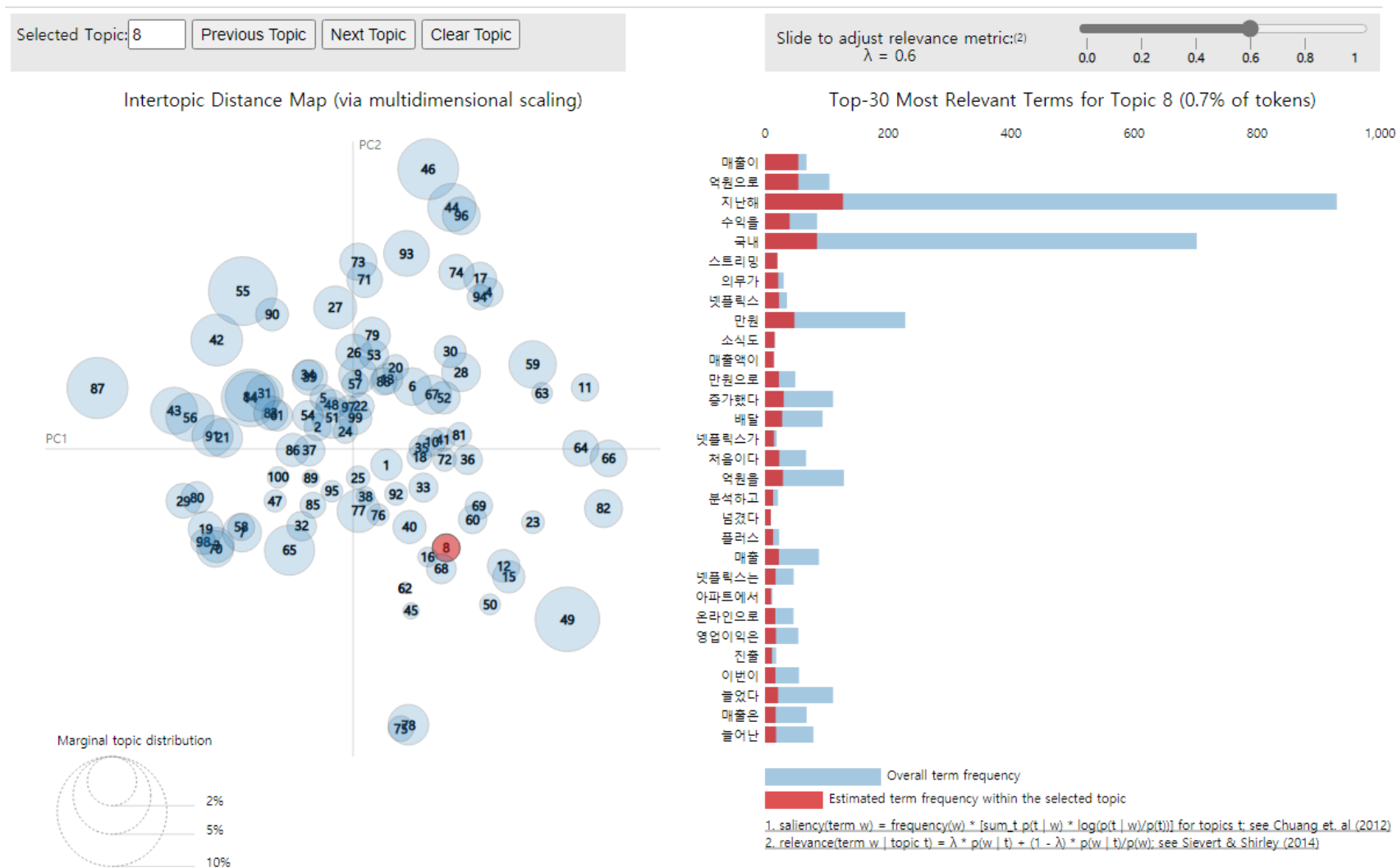
- 사회 기사



확진자, 지역, 발생과 같은 단어들이 주를 이루었던 사회기사의 단어 빈도표를 참고한 결과 주제 중 49번 주제가 코로나와 관련된 사회 기사일 가능성이 높다고 판단됩니다.

LDA 를 이용한 주제 분석

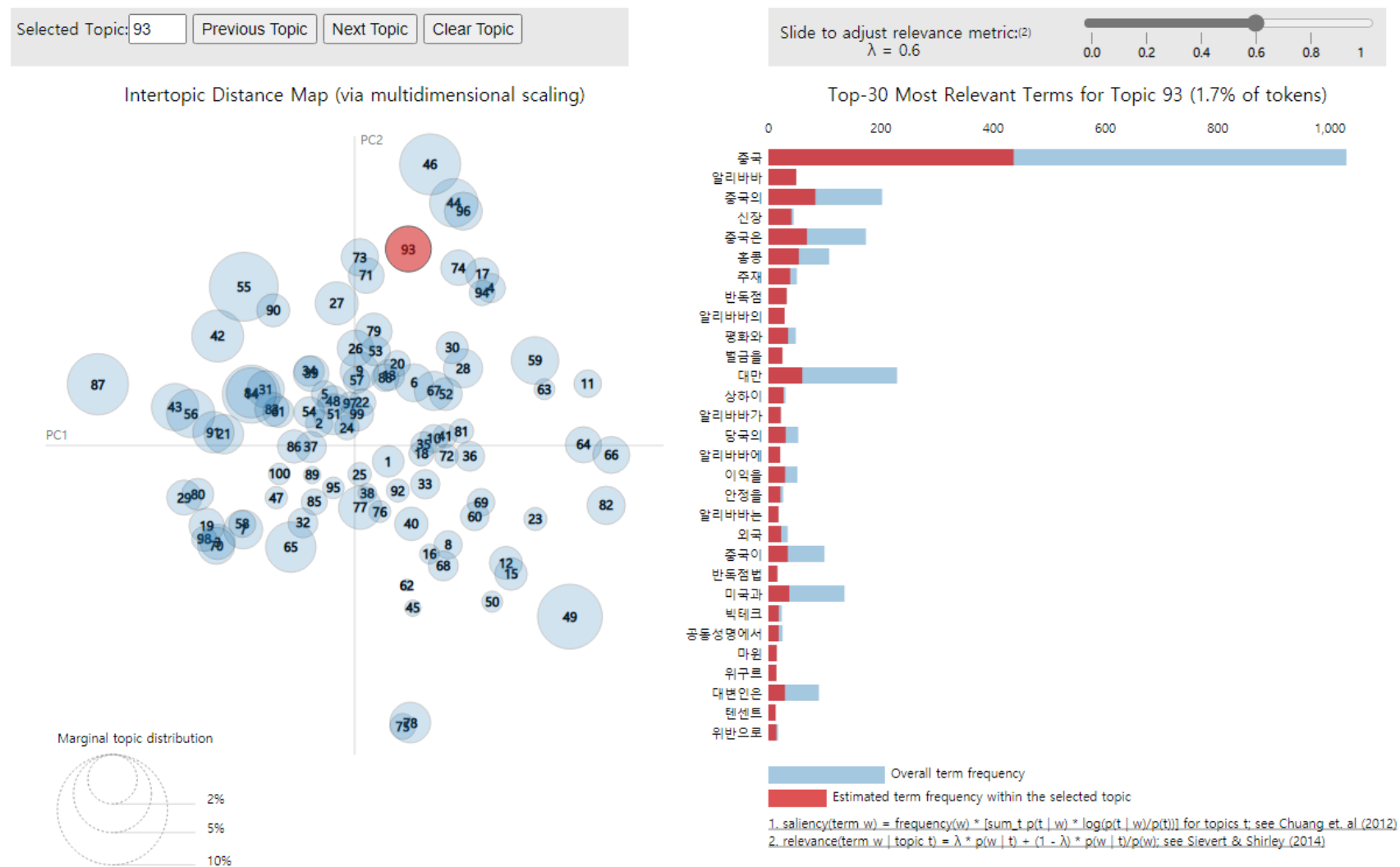
- 경제 기사



시장, 사업, 원과 같은 단어들이 주를 이루었던 경제기사의 단어 빈도표를 참고한 결과 주제 중 8번 주제가 매출, 수익률과 관련된 경제 기사일 가능성이 높다고 판단됩니다.

LDA 를 이용한 주제 분석

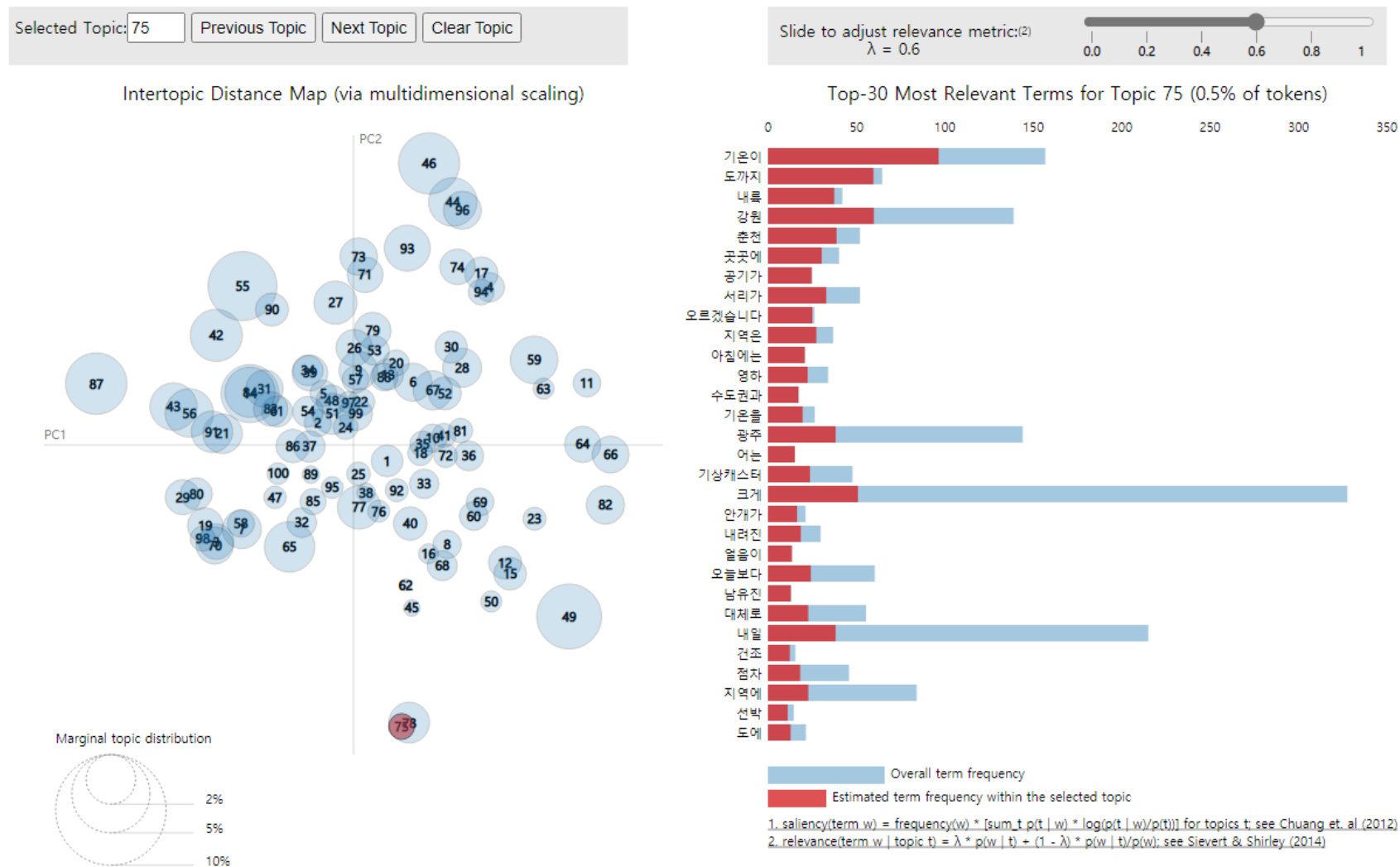
- 국제 기사



중국, 정부 등과 같은 단어들이 주를 이루었던 국제기사의 단어 빈도표를 참고한 결과 주제 중 93번 주제가 중국과 관련된 국제 기사일 가능성이 높다고 판단됩니다.

LDA 를 이용한 주제 분석

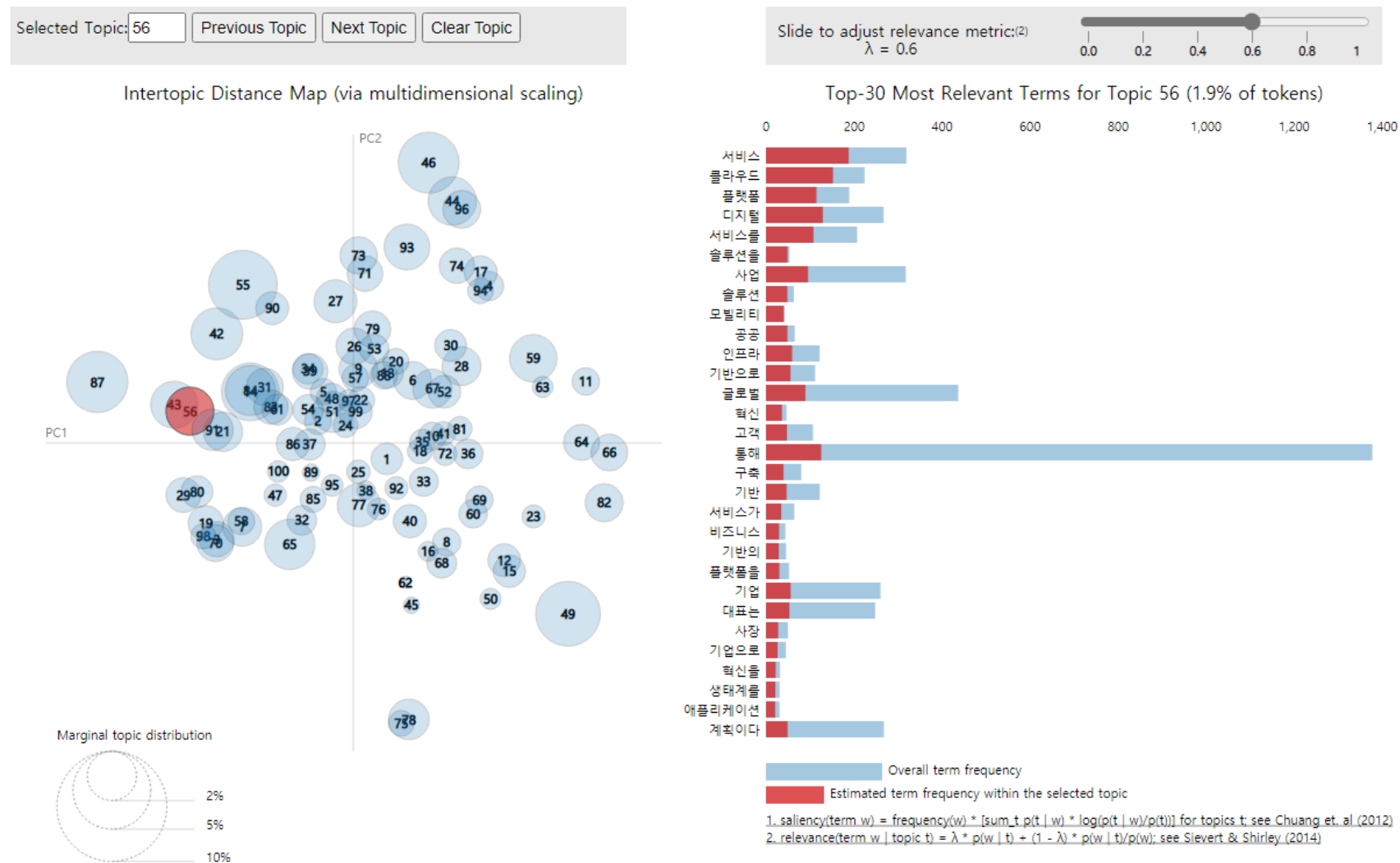
- 문화 기사



지역, 기온, 내일 등과 같은 단어들이 주를 이루었던 문화기사의 단어 빈도표를 참고한 결과 주제 중 75번 주제가 기상정보와 관련된 문화기사일 가능성이 높다고 판단됩니다.

LDA 를 이용한 주제 분석

- IT 기사



기술, 서비스, 기업 등과 같은 단어들이 주를 이루었던 IT기사의 단어 빈도표를 참고한 결과 주제 중 56번 주제가 클라우드 플랫폼 서비스와 관련된 IT기사일 가능성이 높다고 판단됩니다.

LDA 를 이용한 주제 분석

전체 기사를 모두 합하여 주제 분석을 진행하였습니다.

Loss가 -17.2924정도인 모델을 사용하였습니다.

주제분석 결과 응집도는 0.4449였으며, 주제별 다양도는 0.6152를 보였습니다.

앞에서 카테고리별로 1개씩을 뽑아 보여드린 것 과 같이 카테고리별로 주제가 나뉘지는 것을 볼 수 있었습니다.



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.

자기 평가표

01

서론 : 2점 - 다루고자 하는 주제와, 각 목차별로 어떤 것을 알아볼 지 명확히 제시하였다.

02

데이터수집 : 3점 - 다양한 카테고리를 한번에 크롤링할 수 있도록 코드를 정리하고 통일시켰다.

03

전처리 : 3점 - 한국어 불용어 사전과 태그, 마크업 기호제거 등을 하기 위해 다양한 시도를 하였다.

04

단어빈도 : 3점 - 단어구름을 만드는 과정에서 카테고리에 맞는 모양과 색을 사진에 따라 넣어주기 위해 다양한 시도를 하였다

05

감성분석 : 3점 - 다중분류에 맞는 함수를 찾고 모델 실행 시 제약조건을 걸어주는 등 다양한 시도를 하였다.

06

주제분석 : 2점 - LDA를 사용하여 카테고리별 주제를 잘 파악하였다.

07

결론 : 2점 - 서론에서 보고자 했던 것을 되돌아보며 아쉬웠던 점과 잘했던 점을 돌아보았다.



감사합니다

