한남대학교 빅데이터

어디 지역(=분야) 출신이니?

뉴스기사분류

Contents

01 개발배경

02 개발 개요

03 데이터 수집

04 데이터 전처리

05 모델 학습

06 서비스 개발

07 활용방안 및 기대효과

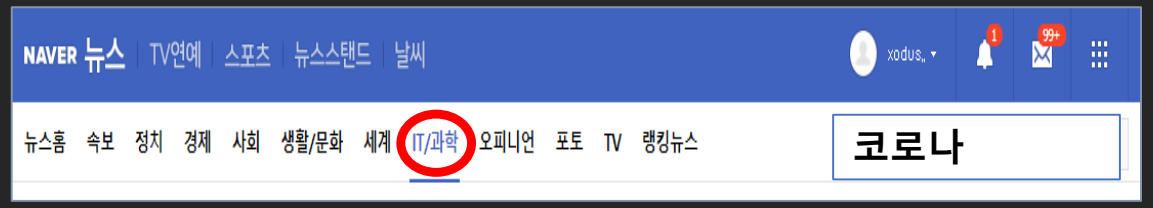
개발배경

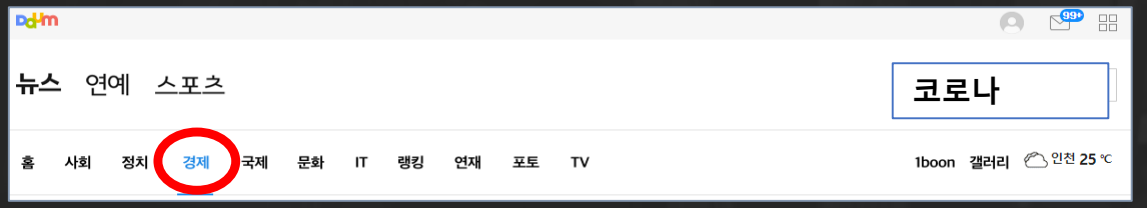
1. 개발 배경 – 주제 선정 동기

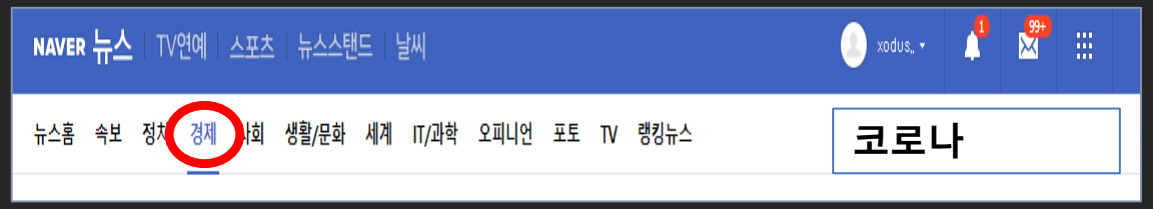
방대한 데이터를 찾아보는데 있어서 신속성과 정확성을 갖음과 동시에
원하는 데이터를 찾을 수 있다는 장점을 부가하기 위해서

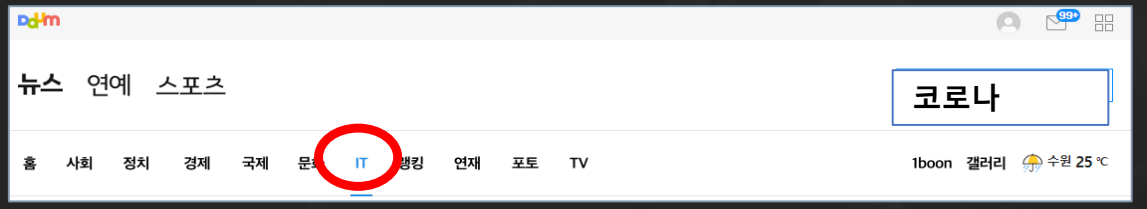
1. 개발 배경 – 주제 선정 동기

- 대표 포털사이트의 뉴스사이트에서 특정 단어를 검색 했을 때 카테고리 별로 검색되어 나오지 않음





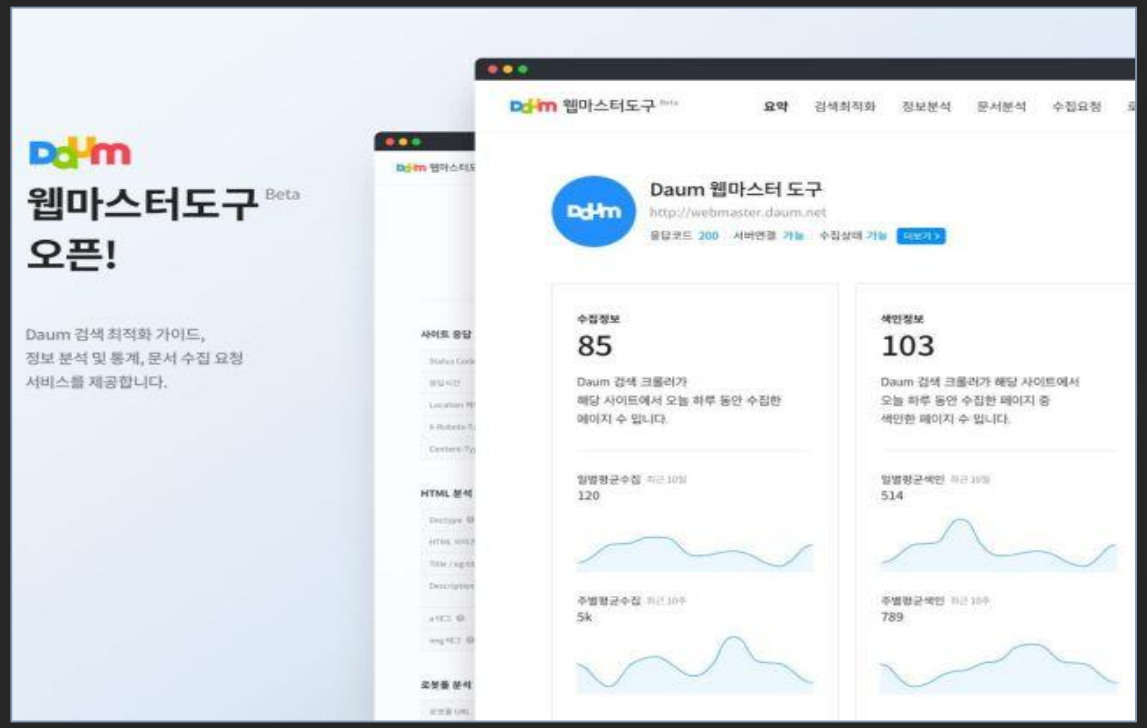




1. 개발 배경 – 주제 선정 동기

- **검색 속도 최적화를 돕는 ‘ 다음 웹마스터 도구 베타‘ 공개**

카카오, Daum은 검색 엔진 기술력을 활용해 검색 서비스를 향상시키기 위해 ‘Daum 웹마스터 도구 beta’ 를 오픈하여 검색 선순화 활성화로 콘텐츠 공급자와 이용자 모두의 만족도를 높이는데 큰 역할을 할 것으로 기대하고 있다.



네이버, '실검' 일부 개편...내일부터 '연령별 맞춤' 순위 표출

입력 2019.10.30 14:16 | 수정 2019.10.30 14:16

네이버, 실검 첫 화면 '연령별 맞춤' 순위로 변경

네이버 "집중도 분산 효과 기대"

플랫폼 전면 개편 나선 카카오에 비해 소극적이라는 평

실시간 인기

1 삼성바이오로직스 위탁

2 한·중·일 배터리 삼국지,

- **검색 서비스의 불편사항을 개선하는 네이버**

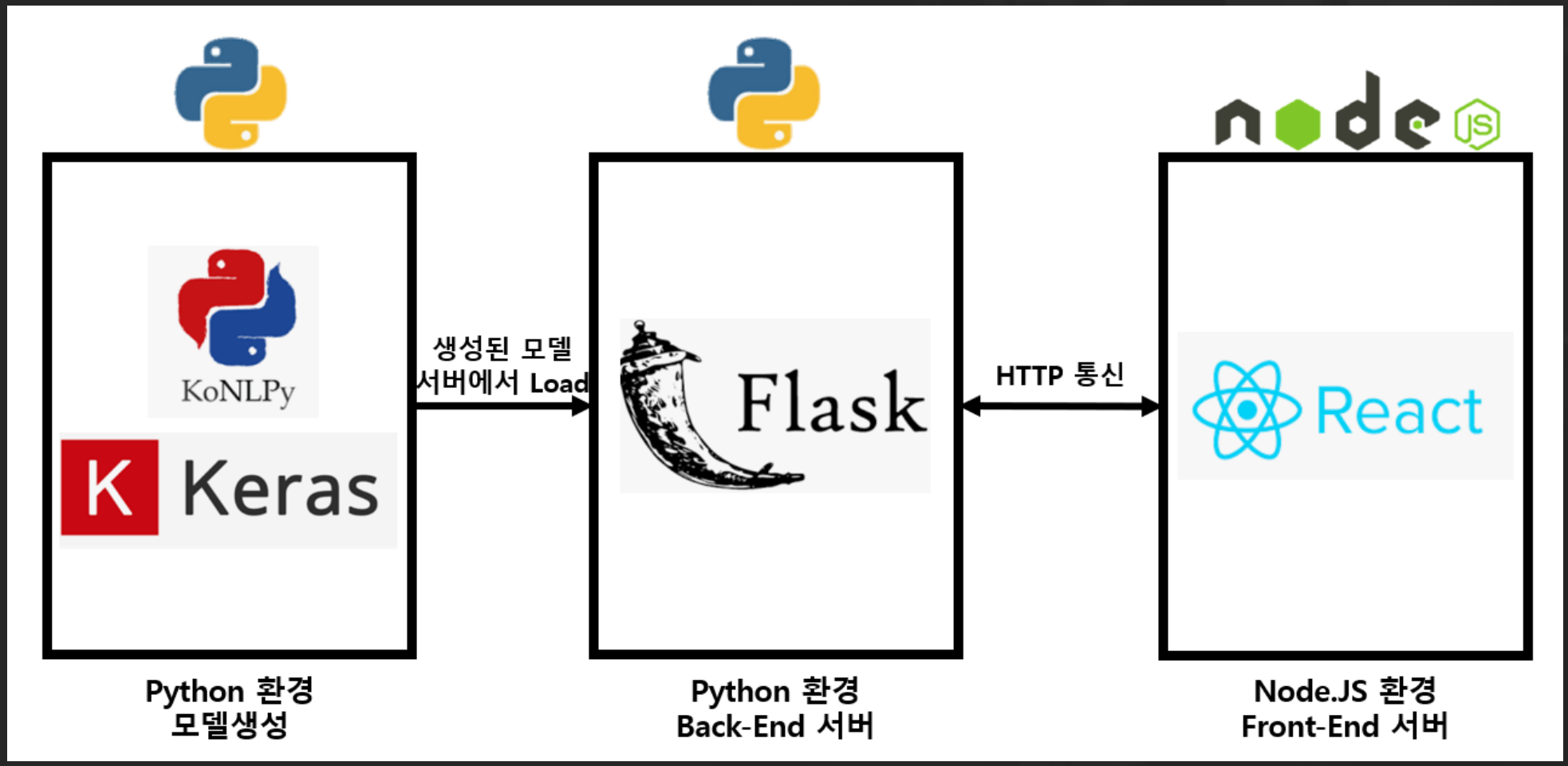
네이버 관계자는 오래된 실시간 검색 방식을 개편하여 모든 사람이 같은 실시간 검색어를 보는 현재의 집중도를 분산시키는 효과를 볼 수 있을 것으로 기대하고 있다.

이는 소비자들의 불편, 불만 사항인 실시간 검색 조작의 가능성을 배제하도록 하기 위한 소통의 결과물이라고 볼 수 있다.

출처 : 한국경제

개발개요

02. 개발개요 – 사용 라이브러리



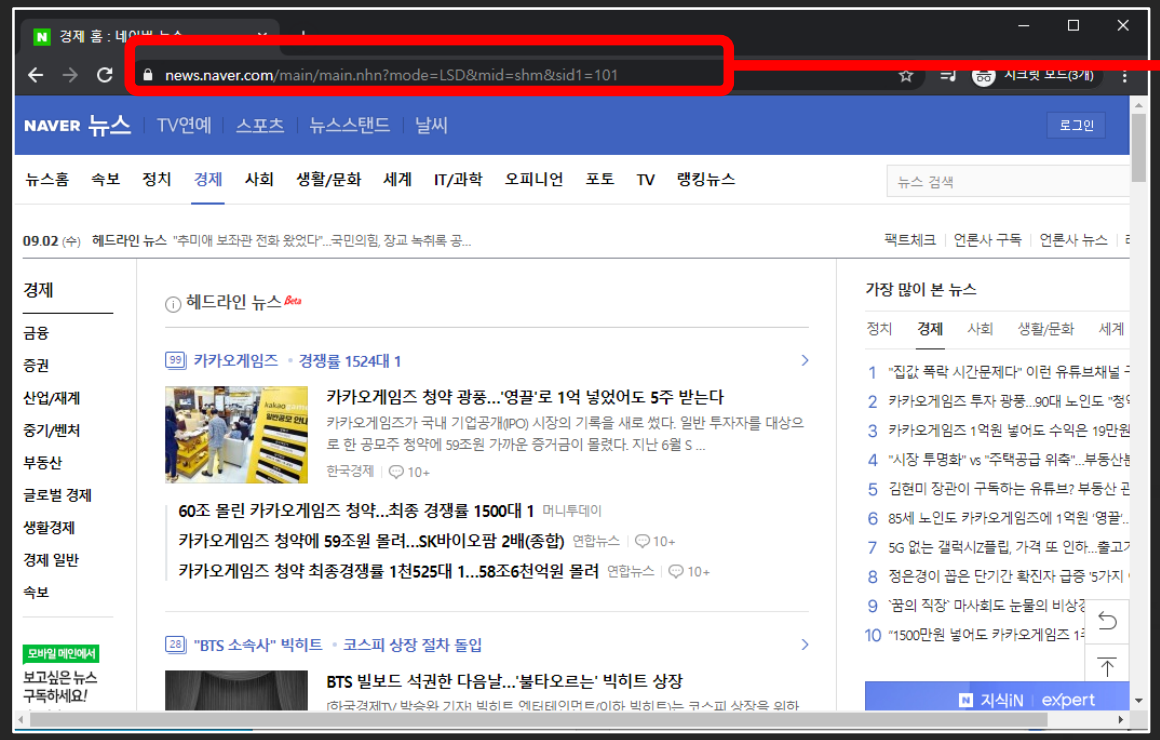
02. 개발개요 – 프로젝트 일정

	21	22	23	24	25	26	27	28	29	30	1	2	3
Data collection	■	■	■	■									
Data preprocessing				■	■								
Professor Counseling						■							
Data modeling							■	■					
Service development									■	■	■		
Document												■	■

데이터 수집

03. 데이터 수집- 뉴스 크롤링

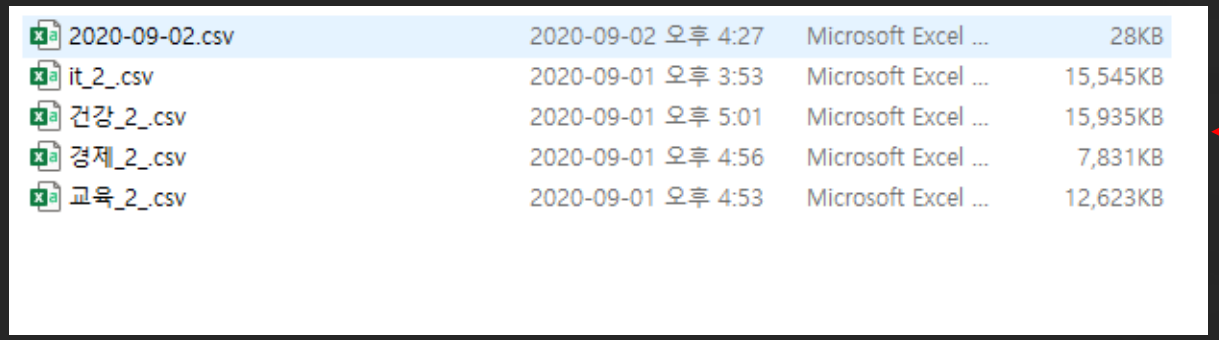
원하는 뉴스페이지를 찾아서 URL 복사




```python
from bs4 import BeautifulSoup
from urllib.request import Request, urlopen
import pandas as pd
```

```python
for i in range(1,11):
    url = '사이트 주소'%(i)

    headers = {'User-Agent' : 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML

    request = Request(url, headers = headers)
    response = urlopen(request)
    html = response.read()
    soup = BeautifulSoup(html, 'html.parser')
    name = soup.select('#main_content > div.list_body.newsflash_body > ul > li > dl > dt')
#    duration = soup.select('#main_content > div.list_body.newsflash_body > ul> li > dl > dd > spa

#score = soup.select('body > div.rankingDetailBody > div > table > tr > td > b')


    for n in name :
        name_l.append(n.text.replace('\n\n', '').replace('\n\r\n\t\t\t\t\t\t\t\t', '').replace('\r
        score_l.append(n.a.get('href'))

    for s in score_l:
        url=s
        request2 = Request(url, headers = headers)
        response2 = urlopen(request2)
        html2 = response2.read()
        soup2 = BeautifulSoup(html2, 'html.parser')
        content = soup2.select('#articleBodyContents')


        for c in content :
            duration_l.append(c.text.replace('\n\n\n\n\n','').replace('\n\t\n',''))
```

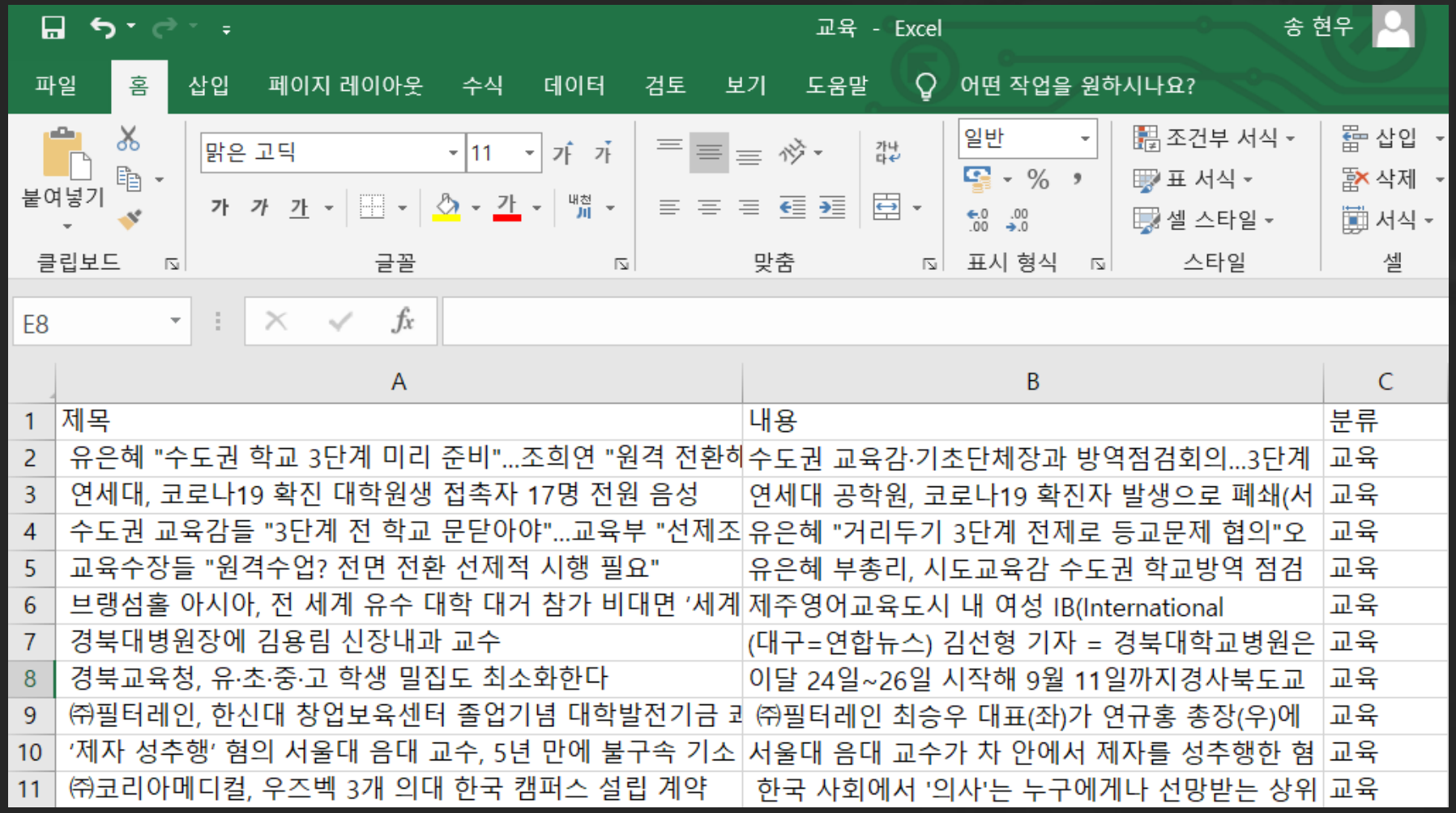
뉴스기사 파일 생성



2020-09-02.csv	2020-09-02 오후 4:27	Microsoft Excel ...	28KB
it_2_.csv	2020-09-01 오후 3:53	Microsoft Excel ...	15,545KB
건강_2_.csv	2020-09-01 오후 5:01	Microsoft Excel ...	15,935KB
경제_2_.csv	2020-09-01 오후 4:56	Microsoft Excel ...	7,831KB
교육_2_.csv	2020-09-01 오후 4:53	Microsoft Excel ...	12,623KB

03. 데이터 수집- 뉴스 크롤링

CSV로 저장된 뉴스데이터



교육 - Excel　　송 현우

파일 | 홈 | 삽입 | 페이지 레이아웃 | 수식 | 데이터 | 검토 | 보기 | 도움말 | 어떤 작업을 원하시나요?

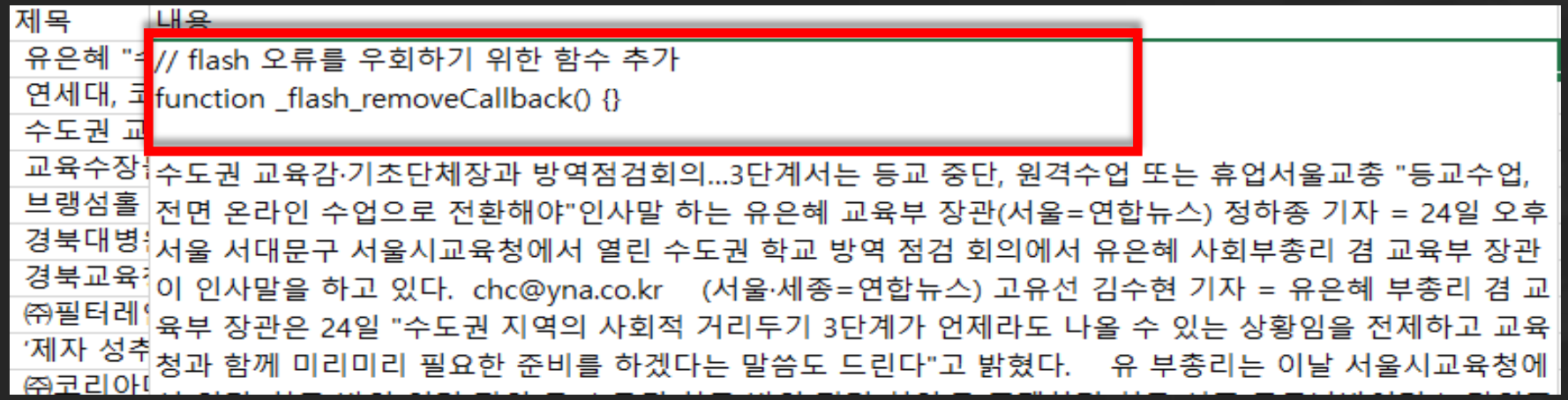
붙여넣기 | 클립보드 | 맑은 고딕 | 11 | 글꼴 | 맞춤 | 일반 | 표시 형식 | 조건부 서식 | 표 서식 | 셀 스타일 | 스타일 | 삽입 | 삭제 | 서식 | 셀

E8

	A	B	C
1	제목	내용	분류
2	유은혜 "수도권 학교 3단계 미리 준비"...조희연 "원격 전환히	수도권 교육감·기초단체장과 방역점검회의...3단계	교육
3	연세대, 코로나19 확진 대학원생 접촉자 17명 전원 음성	연세대 공학원, 코로나19 확진자 발생으로 폐쇄(서	교육
4	수도권 교육감들 "3단계 전 학교 문닫아야"...교육부 "선제조	유은혜 "거리두기 3단계 전제로 등교문제 협의"오	교육
5	교육수장들 "원격수업? 전면 전환 선제적 시행 필요"	유은혜 부총리, 시도교육감 수도권 학교방역 점검	교육
6	브랭섬홀 아시아, 전 세계 유수 대학 대거 참가 비대면 '세계	제주영어교육도시 내 여성 IB(International	교육
7	경북대병원장에 김용림 신장내과 교수	(대구=연합뉴스) 김선형 기자 = 경북대학교병원은	교육
8	경북교육청, 유·초·중·고 학생 밀집도 최소화한다	이달 24일~26일 시작해 9월 11일까지경사북도교	교육
9	㈜필터레인, 한신대 창업보육센터 졸업기념 대학발전기금 쾌	㈜필터레인 최승우 대표(좌)가 연규홍 총장(우)에	교육
10	'제자 성추행' 혐의 서울대 음대 교수, 5년 만에 불구속 기소	서울대 음대 교수가 차 안에서 제자를 성추행한 혐	교육
11	㈜코리아메디컬, 우즈벡 3개 의대 한국 캠퍼스 설립 계약	한국 사회에서 '의사'는 누구에게나 선망받는 상위	교육

데이터 전처리

04. 데이터 전처리 - 잡음제거



제목	내용
유은혜 "수	// flash 오류를 우회하기 위한 함수 추가 function _flash_removeCallback() {}
연세대, 코	
수도권 교	
교육수장들	수도권 교육감·기초단체장과 방역점검회의…3단계서는 등교 중단, 원격수업 또는 휴업서울교총 "등교수업,
브랭섬홀	전면 온라인 수업으로 전환해야"인사말 하는 유은혜 교육부 장관(서울=연합뉴스) 정하종 기자 = 24일 오후
경북대병원	서울 서대문구 서울시교육청에서 열린 수도권 학교 방역 점검 회의에서 유은혜 사회부총리 겸 교육부 장관
경북교육청	이 인사말을 하고 있다. chc@yna.co.kr (서울·세종=연합뉴스) 고유선 김수현 기자 = 유은혜 부총리 겸 교
㈜필터레이	육부 장관은 24일 "수도권 지역의 사회적 거리두기 3단계가 언제라도 나올 수 있는 상황임을 전제하고 교육
'제자 성추	청과 함께 미리미리 필요한 준비를 하겠다는 말씀도 드린다"고 밝혔다. 유 부총리는 이날 서울시교육청에
㈜코리아디	

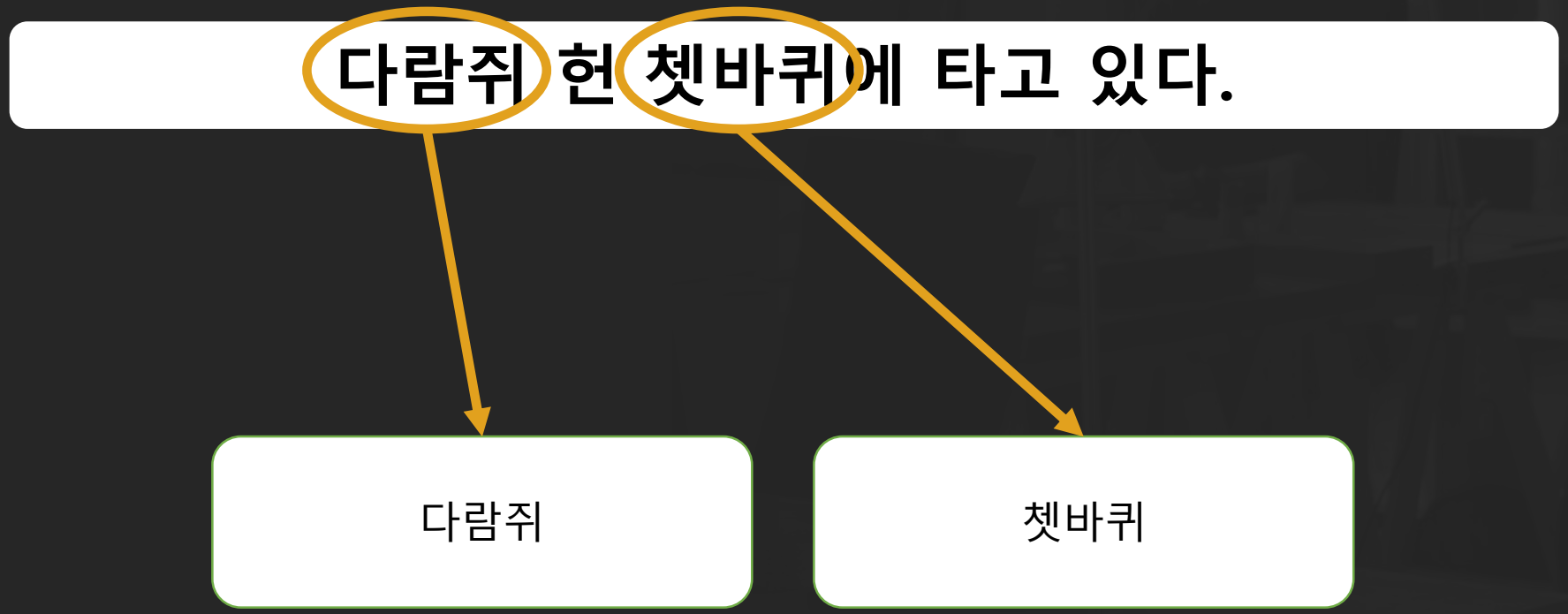
- **기사 내용과 관련 없는 내용 제거**

  - 공백, // flash 오류 우회 함수 추가 → function _flash_removeCallback() {}

  - 특수문자 , 이메일 크롤링

04. 데이터 전처리 – 형태소 분석

- 기사 분류에 관련 없는 동사, 형태소를 제거하고 명사만을 추출하여 정확한 기사 분류에 도움을 줌



04. 데이터 전처리 – 형태소 분석

기사내용

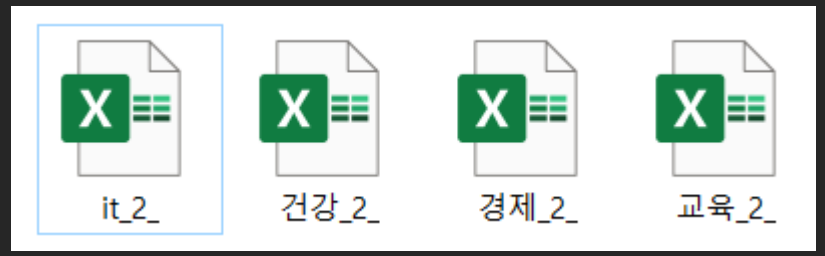
온라인으로 진행...1500명 이상 참여 예정(지디넷코리아=방은주 기자)한국심리학회(회장 조현섭 총신대학교 교수)는 '2020 제74차 한국심리학회 연차학술대회'를 20~22일 사흘간 온라인으로 개최한다고 19일 밝혔다.이번 학술대회는 1개 프리세션과 대외 심포지엄, 6개 세션 특별 심포지엄, 11개 세션 분과 심포지엄, 6개 세션 분과 워크숍, 청소년 심리학교실, 한·중·일 심포지엄 등 총 111편의 포스터 발표가 온라인으로 열린다. 1500명 이상이 참가할 예정이다. 주최 측은 참가 인원수 제한에도 200명이 넘는 중·고등학생이 청소년심리학교실에 참가 신청을 했다고 밝혔다.1일차(20일)에 개최할 프리세션은 '국민을 위한 심리서비스 법제화의 현실적 필요성과 실현 방안'이라는 주제로 국내 심리서비스 운영의 문제점을 진단하고 개선 방향을 제안한다.이후 대외심포지엄에서 '위기의 지구, 재난, 심리학'에서 길을 찾다'는 주제로 최근 발생한 코로나19 대유행을 중심으로 재난이 우리의 삶과 심리에 미치는 부정적 영향을 이해하고, 이에 대한 대응 방안을 모색하는 자리를 마련했다. 특히, 심리학을 기반으로 의학, 보건학, 의료사회학 등 다양한 인접 학문 관점에서 재난의 피해를 해결하는데 도움이 되는 연구 성과를 논의할 예정이다. 또한 미국심리학회를 대표해 산드라 L. 슐먼(Sandra L. Shullman) 회장이 발표자로 참여한다.2일차(21일)에는 심포지엄과 워크숍을 통해 우리 사회의 다양한 문제를 다룬다. 한국형 심리방역체계 구축, 중독, 자살, 심리적 외상, 디지털 성범죄, 재범 방지 등과 같은 우리 사회가 직면한 현안들을 심리학적으로 진단하고 그 해결책을 모색한다.특히 정서장애의 근거 기반치료, 게슈탈트 치료, 코칭 프로그램 등과 같이 우리가 현대사회를 살아

명사 추출

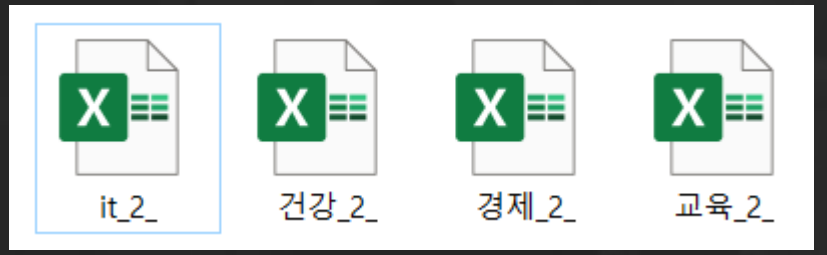
['온라인', '진행', '1500명', '이상', '참여', '예정(지디넷코리아=방은주', '기자)한국심리학회(회장', '조현섭', '총신대학교', '교수', '2020', '제74차', '한국심리학회', '연차학술대회', '20', '22일', '사흘간', '개최', '19일', '이번', '학술대회', '1개', '프리세션', '대외', '심포지엄', '6개', '세션', '특별', '11개', '분', '워크숍', '청소년', '심리학교', '한·중·', '등', '111편', '포스터', '발표가', '참가', '예정', '주최', '측은', '인원수', '제한', '200명', '중·고등학생', '청소년심리학교실', '신청', '1일차(20일)', '국민', '심리서비스', '법제화', '현실적', '필요성', '실현', '방안', '주제', '국내', '운영', '문제점', '진단', '개', '방향', '제안', '이후', '대외심포지엄', '위기', '지구', '재난', '심리학', '찾다', '발생', '코로나19', '대유행', '중심', '우리', '삶', '심리', '부정적', '영향', '이해', '이', '대응', '방안', '모색', '자리', '마련', '특히,', '기반', '의학', '보건학', '의료사회학', '다양한', '인접', '학문', '관점', '피해', '해결', '도움', '연구', '성과', '논의', '미국심리학회', '대표', '산드라', '슐먼(Sandra', '회장', '발표자', '2일차(21일)', '사회', '문제', '한국형', '심리방역체계', '구축', '중독', '자살', '심리적', '외상', '디지털', '성범죄', '재범', '방지', '직면', '현안들', '심리학적', '해결책', '정서장애', '근거', '기반치료', '게슈탈트', '치료', '코칭', '프로그램', '현대사회', '다양', '맥락', '문제들', '실용적', '프로그램들', '다수', '준비', '마지막', '3일차(22일)', '청소년들', '응용', '수',

모델 학습

05. 모델학습 – 데이터 쌓기



형태소 분석이 끝난 파일



카테고리 정수화 : 0 1 2 3

```
w_list = []
for item in data_0.내용:
    item = item[1:-1]
    temp = item.split(',')
    w_list.append(temp)

w_list_result = []
for n in w_list:
    news = []
    for x in n:
        x = x.strip()
        if(len(x) > 1):
            news.append(x)
    w_list_result.append(news)
    y_category.append(0)

#IT Append
```

```
w_list = []
for item in data_1.내용:
    item = item[1:-1]
    temp = item.split(',')
    w_list.append(temp)

for n in w_list:
    news = []
    for x in n:
        x = x.strip()
        if(len(x) > 1):
            news.append(x)
    w_list_result.append(news)
    y_category.append(1)
#건강 Append
```

```
w_list = []
for item in data_2.내용:
    item = item[1:-1]
    temp = item.split(',')
    w_list.append(temp)

for n in w_list:
    news = []
    for x in n:
        x = x.strip()
        if(len(x) > 1):
            news.append(x)
    w_list_result.append(news)
    y_category.append(2)
#경제 Append
```

```
w_list = []
for item in data_3.내용:
    item = item[1:-1]
    temp = item.split(',')
    w_list.append(temp)

for n in w_list:
    news = []
    for x in n:
        x = x.strip()
        if(len(x) > 1):
            news.append(x)
    w_list_result.append(news)
    y_category.append(3)
#교육 Append
```

입력 데이터 저장 (X data)
결과 데이터 저장 (Y data) 저장

05. 모델학습 – 토큰화, 시퀀스화

Tokenizer(Keras)
빈도수 잦은 단어
토큰화

➡

Texts_to_Sequences
함수사용 단어
시퀀스화

➡

결과데이터 (Y data)
Sequence Matrix
변환

```
[[   0    0    0 ...   49 2776 2827]
 [   0    0    0 ...  509  320  384]
 [   0    0    0 ... 2176 2170 2177]
 ...
 [   0    0    0 ...   33    5    3]
 [   0    0    0 ... 3695 4919 2565]
 [   0    0    0 ... 3695 4919 2565]]
[    0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0  691  438  107  461  438  609   74  946
  4188 4663  204   18 3459 3229  250  138  453   59  342   62 3828  342
   946  839  468   62  242  566  302  468  138  638  347  438  536 2973
  1619  302  838  695  838  831  438  136   51  982  672  113  908  292
   838 3863  895   11  982  471   61   66   18  224 3817  118  438   74
   283 1367   62  146  415   54 1995  253   12  138  342   62  972  453
  2474 1030   42  438  838 3863  111  838  471  111  630  322  438  998
   838  505  838  138  471 3474   62   11  926  536 2973   66  463   82
   175  940   62   11  223   65  267    9  950  419 3818 2945   11  338
  3863  895  204 4626   36 1399 1311   18 2936 1399  403 4626 3863 1399
    99  438 4626 2364   82 3863  167   62  773   85  857   59  520 2202
  1697   85  198  368 2036   36 1710 3818 2556 1685  992 2086  471 4882
  3818  471  438   59    9   36  946  302  992  471   51  259   59  950
     9  438  672  578    9 3818  282  382  204  672 3818  229 2945   40
  3818 1445  302  471  342   62   82 1877   40   39  181  415 2191 1140
  1271    4 3012  382  204  313   74  438 1771  438 3140  283 1771   82
   258 2037  698  313  453  217    2   49 2776 2827]
500
```

결과값

```
4
[[1. 0. 0. 0.]
 [1. 0. 0. 0.]
 [1. 0. 0. 0.]
 ...
 [0. 0. 0. 1.]
 [0. 0. 0. 1.]
 [0. 0. 0. 1.]]
```

05. 모델학습 – 모델 생성

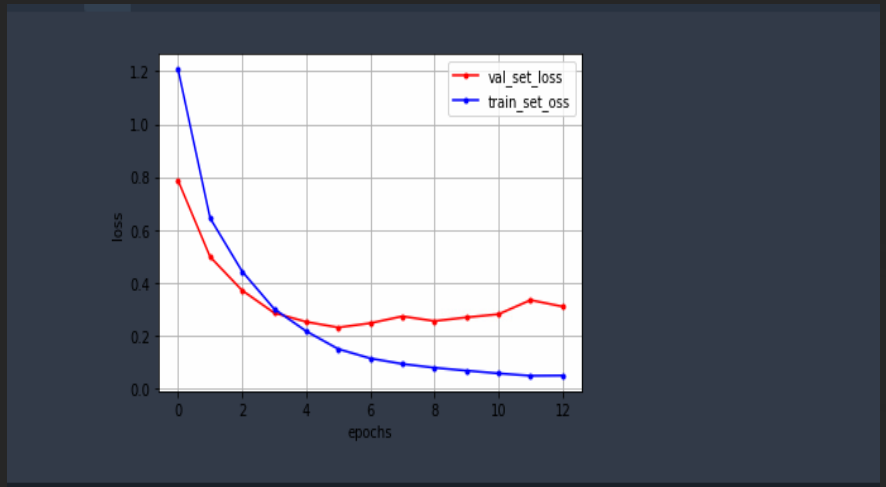

훈련데이터
검증데이터
분리
단어 임베딩
LSTM 모델링
Dense Layer 생성
훈련설정
Batch size : 500
Epochs : 20
모델 학습

05. 모델학습 - 학습결과

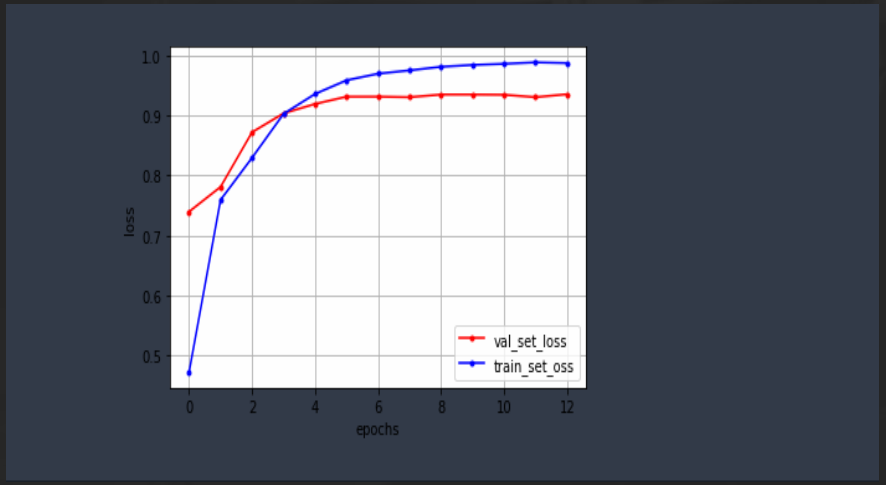
학습 결과
Accuracy: 93%

```
In [40]:  1  print("정확도 : %.4f" % (model.evaluate(X_test, y_test)[1]))

223/223 [==============================] - 15s 69ms/step - loss: 0.2903 - accuracy: 0.9368
정확도 : 0.9368
```

Loss Graph

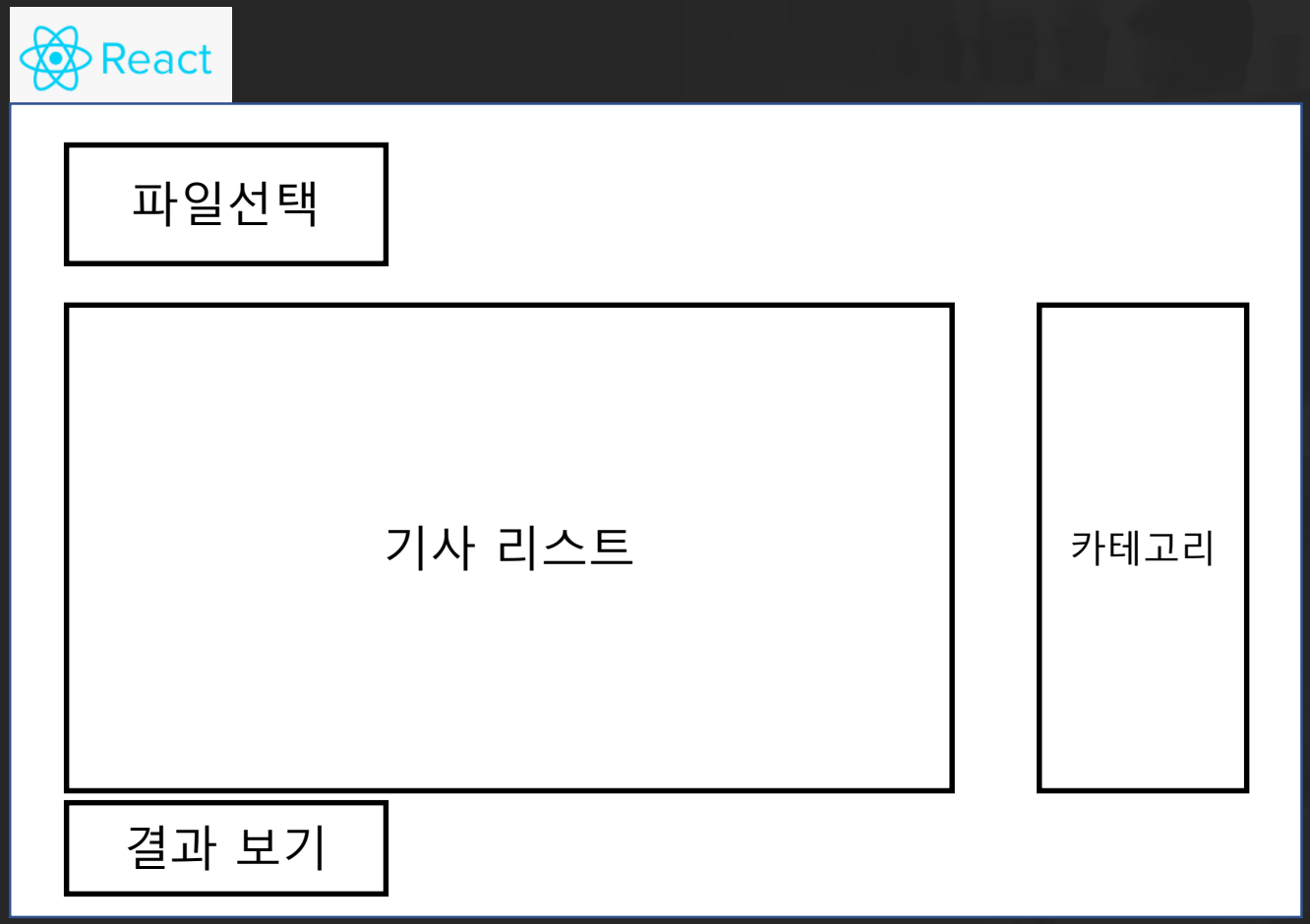


Accuracy Graph

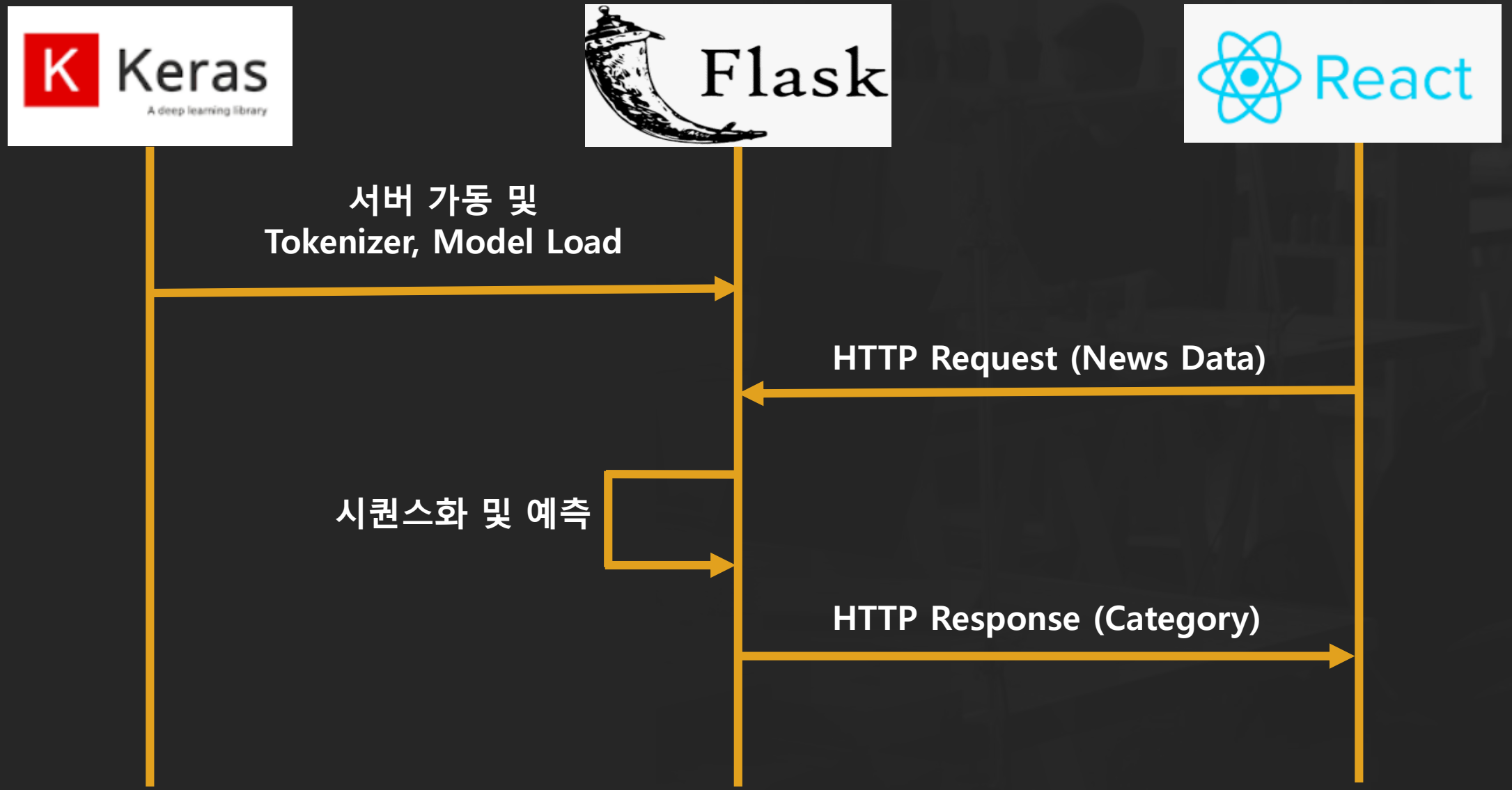


서비스 개발

06. 서비스 개발 – 초기 디자인



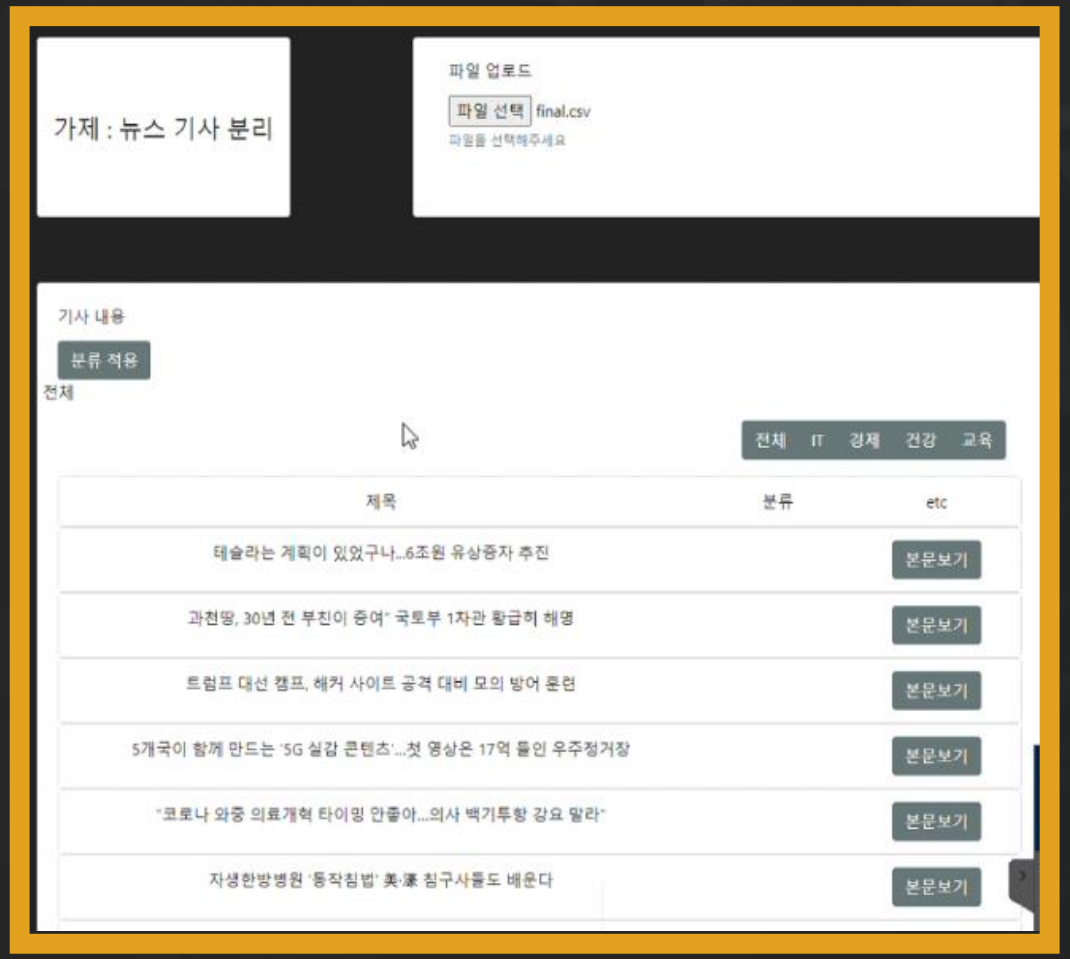
06. 서비스 개발 – Back-End 서버



06. 서비스 개발 – 서비스 사용

(1) 기사 파일 선택

(2) 기사 로드 완료



06. 서비스 개발 – 서비스 사용

(3) 분류 적용 버튼 클릭

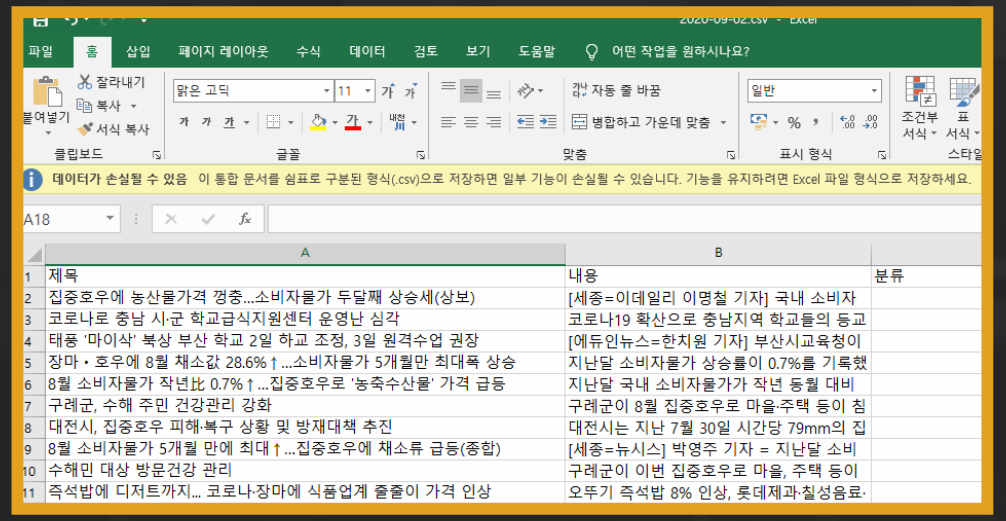
06. 서비스 개발 – 서비스 사용

(4) 분류 완료 및 분야별 모아 보기

06. 서비스 개발 – 검색활용(집중호우)

전체 IT 경제 건강 교육

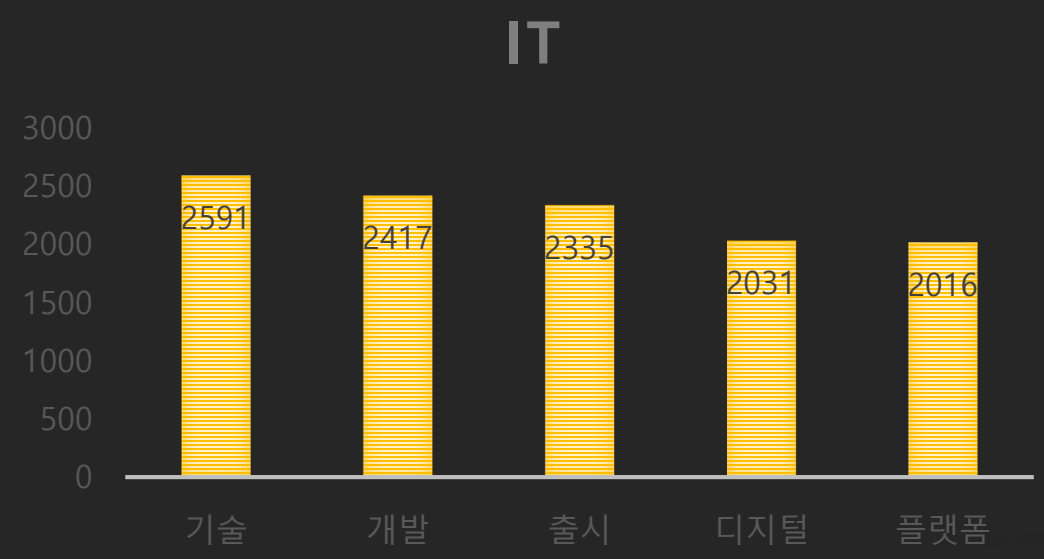
제목	분류	etc
집중호우에 농산물가격 껑충…소비자물가 두달째 상승세(상보)	경제	본문보기
코로나로 충남 시·군 학교급식지원센터 운영난 심각	경제	본문보기
태풍 '마이삭' 북상 부산 학교 2일 하교 조정, 3일 원격수업 권장	교육	본문보기
장마・호우에 8월 채소값 28.6%↑…소비자물가 5개월만 최대폭 상승	경제	본문보기
8월 소비자물가 작년比 0.7%↑…집중호우로 '농축수산물' 가격 급등	경제	본문보기
구례군, 수해 주민 건강관리 강화	건강	본문보기
대전시, 집중호우 피해·복구 상황 및 방재대책 추진	경제	본문보기
8월 소비자물가 5개월 만에 최대↑…집중호우에 채소류 급등(종합)	경제	본문보기
수해민 대상 방문건강 관리	경제	본문보기
즉석밥에 디저트까지… 코로나·장마에 식품업계 줄줄이 가격 인상	IT	본문보기

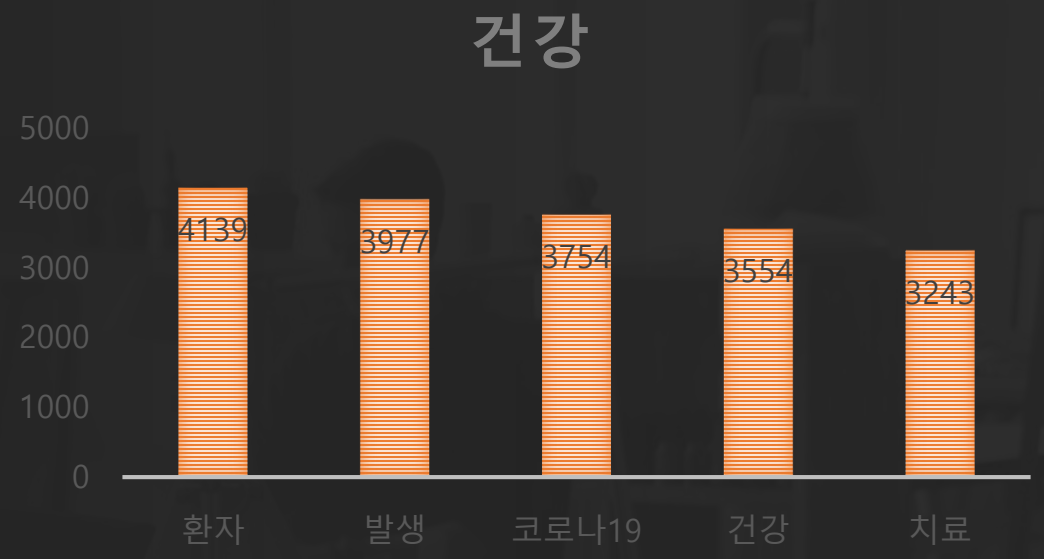


	A	B	분류
1	제목	내용	
2	집중호우에 농산물가격 껑충…소비자물가 두달째 상승세(상보)	[세종=이데일리 이명철 기자] 국내 소비자	
3	코로나로 충남 시·군 학교급식지원센터 운영난 심각	코로나19 확산으로 충남지역 학교들의 등교	
4	태풍 '마이삭' 북상 부산 학교 2일 하교 조정, 3일 원격수업 권장	[에듀인뉴스=한치원 기자] 부산시교육청이	
5	장마・호우에 8월 채소값 28.6%↑…소비자물가 5개월만 최대폭 상승	지난달 소비자물가 상승률이 0.7%를 기록했	
6	8월 소비자물가 작년比 0.7%↑…집중호우로 '농축수산물' 가격 급등	지난달 국내 소비자물가가 작년 동월 대비	
7	구례군, 수해 주민 건강관리 강화	구례군이 8월 집중호우로 마을·주택 등이 침	
8	대전시, 집중호우 피해·복구 상황 및 방재대책 추진	대전시는 지난 7월 30일 시간당 79mm의 집	
9	8월 소비자물가 5개월 만에 최대↑…집중호우에 채소류 급등(종합)	[세종=뉴시스] 박영주 기자 = 지난달 소비	
10	수해민 대상 방문건강 관리	구례군이 이번 집중호우로 마을, 주택 등이	
11	즉석밥에 디저트까지… 코로나·장마에 식품업계 줄줄이 가격 인상	오뚜기 즉석밥 8% 인상, 롯데제과·칠성음료·	

2020-09-02

네이버 검색 키워드 “집중호우”
뉴스 리스트


IT
3000
2500
2000
1500
1000
500
0
2591
2417
2335
2031
2016
기술
개발
출시
디지털
플랫폼


건강
5000
4000
3000
2000
1000
0
4139
3977
3754
3554
3243
환자
발생
코로나19
건강
치료


경제
4000
3000
2000
1000
0
3335
3222
2555
2344
2231
미국
수준
3554억달러
미술작품
프로젝트


교육
4000
3000
2000
1000
0
3699
3631
3069
2927
2865
학생
학교
진행
제공
교육부

활용 방안 및 기대사항

07. 활용 방안 및 기대 사항

검색엔진

포털 검색엔진에 이용 가능

자동화

- 지식인/블로그에 해시태그
자동화 기능 추가 가능
- 세분화 되고 전문성 있는 정보 획득

연관성

키워드 검색 시 분야별 기사
포커스 와 입장 등을 알 수 있다

Q&A

감사합니다.