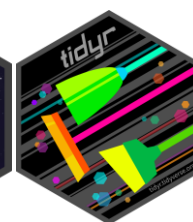
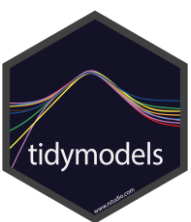
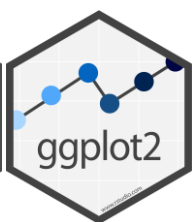


논문작성을 위한

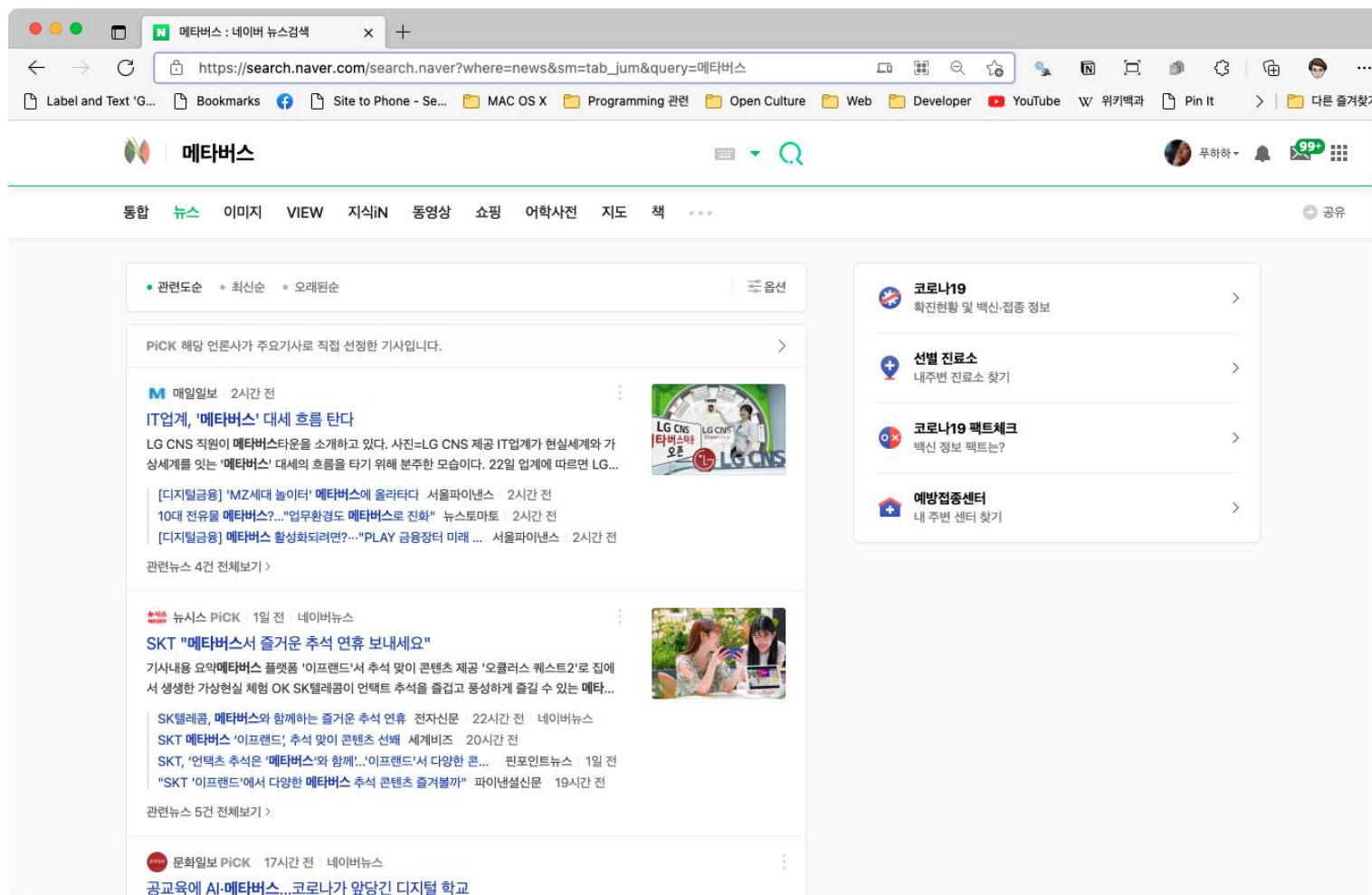


통계분석

빅데이터와 기계학습에 슬쩍...



네이버 뉴스 검색 결과



- 네이버 뉴스 검색 결과
- 같은 뉴스 서로 다른 문서

<SK텔레콤이 언택트 추석을 위해 메타버스(Metaverse) 서비스를 고객에게 제공한다.>

SK텔레콤이 언택트 추석을 즐겁고 풍성하게 즐길 수 있는 **메타버스(Metaverse)** 서비스를 마련해 고객에게 제공한다.

SK텔레콤은 메타버스 공간인 이프랜드에서 아바타를 통해 즐길 수 있는 다양한 콘텐츠를 준비했다. 메타버스 대중화를 견인하기 위해 모집한 '이프렌즈' 중심으로 다양한 모임이 개최된다.

이프렌즈는 다양한 주제를 가지고 메타버스 세상에서 즐길 수 있는 콘텐츠를 만들고 이용자와 실시간으로 소통하며 메타버스 대중화를 견인할 인플루언서 그룹이다.

2021. 10. 14일(목) 15:00 ~ 16:00

5 중형다그림 리스크에 비트코인 휘청... 5000만원선 위태
23시간전

LINCOLN AVIATOR
여유로운 일상을 비행하다
자세히 보기 >

본야별 주요뉴스
홍준표 "이재명, 대장동 의혹 '특검' 수용해야"
대장동 달려간 홍준표 "이재명, 후보 사퇴 아니라 감옥 가라"
이재명 "대장동 개발 결정 당시엔 감질-황포라는 비난까"
추석 연휴 글로벌 증시 강타한 '헝다 쇼크'... 중국황 위기
"약자 지배하려는 강대국 시도에 반대"... 바이든-시진핑,...

SK텔레콤이 언택트 추석을 즐겁고 풍성하게 즐길 수 있는 메타버스(Metaverse) 서비스를 마련해 고객에게 제공한다.

SK텔레콤은 메타버스 공간인 이프랜드에서 아바타를 통해 즐길 수 있는 다양한 콘텐츠를 준비했다. 메타버스 대중화를 견인하기 위해 모집한 '이프렌즈' 중심으로 다양한 모임이 개최된다.

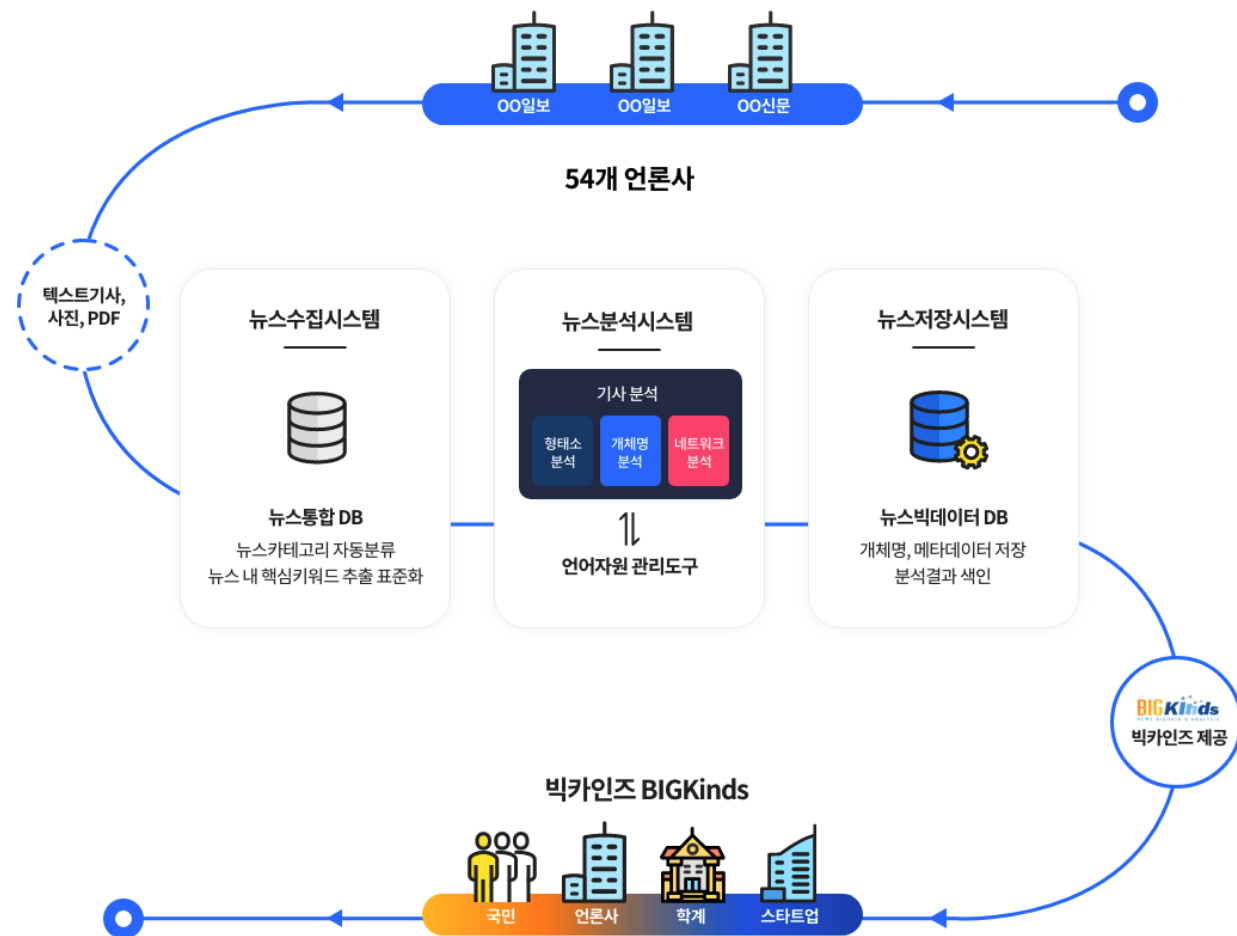
- 네이버 뉴스 검색 결과
 - OpenAPI

필드	설명
title	개별 검색 결과이며, title, originallink, link, description, pubDate 를 포함한다.
originallink	검색 결과 문서의 제공 언론사 하이퍼텍스트 link를 나타낸다.
link	검색 결과 문서의 제공 네이버 하이퍼텍스트 link를 나타낸다.
description	검색 결과 문서의 내용을 요약한 패시지 정보이다. 문서 전체의 내용은 link를 따라가면 읽을 수 있다. 패시지에서 검색어와 일치하는 부분은 태그로 감싸져 있다.
pubDate	검색 결과 문서가 네이버에 제공된 시간이다.

```
<rss version="2.0">
  <channel><title>Naver Open API - news :: '주식'</title>
    <link>http://search.naver.com</link>
    <description>Naver Search Result</description>
    <lastBuildDate>Mon, 26 Sep 2016 11:01:35 +0900</lastBuildDate>
    <total>2566589</total>
    <start>1</start>
    <display>10</display>
    <item>
      <title>국내 <b>주식</b>형펀드서 사흘째 자금 순유출</title>
      <originallink>http://app.yonhapnews.co.kr/YNA/Basic/SNS/r.aspx?c=AKR20160926019000008&did=1195m</originallink>
      <link>http://openapi.naver.com/l?AAAC2NSwvCMBCEf832WJK06e0Qg+kDLAqCXjyGJqUFk9i0Kv57t0VYdr+ZgZ35ZcJXQFPBIQFZbVBIK0toDGYQ47o+ITkAa3Gc+SyxU28T4t5bNKyaHJ5glI7d6CBprdcGkvp0rYFldtLIi+mRl0lTFJRQFH4PyM7qz6TISZbmPFoFTf004JyVnJE8smLADn3sBjlfmvMITFKF63Hbusuha+++xxc/Bc8nKskAAAA=</link>
      <description>국내 <b>주식</b>형 펀드에서 사흘째 자금이 빠져나갔다. 26일 금융투자협회에 따르면 지난 22일 상장지수펀드(ETF)를 제외한 국내 <b>주식</b>형 펀드에서 126억원이 순유출됐다. 472억원이 들어오고 598억원이 펀드... </description>
      <pubDate>Mon, 26 Sep 2016 07:50:00 +0900</pubDate>
    </item>
    ...
  </channel>
</rss>
```

• 빅카인즈

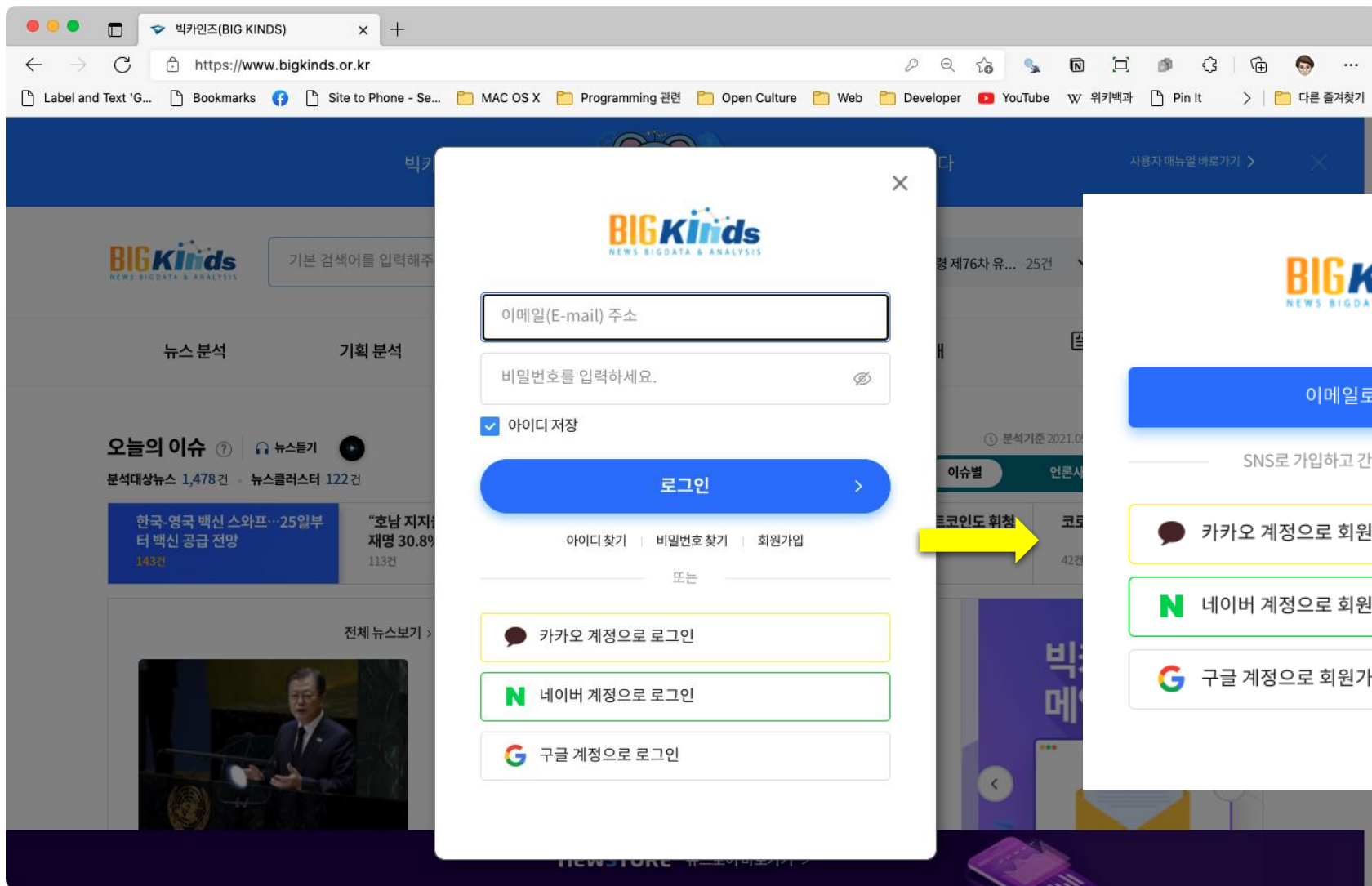
- 한국언론진흥재단이 구축한 서비스
- 뉴스수집시스템, 분석시스템, 저장시스템 등으로 구성돼 있으며, 저장된 뉴스 분석 정보는 국민, 언론사, 학계, 스타트업 등이 활용할 수 있는 뉴스빅데이터 분석서비스
- 비정형 데이터를 정형 데이터로 제공



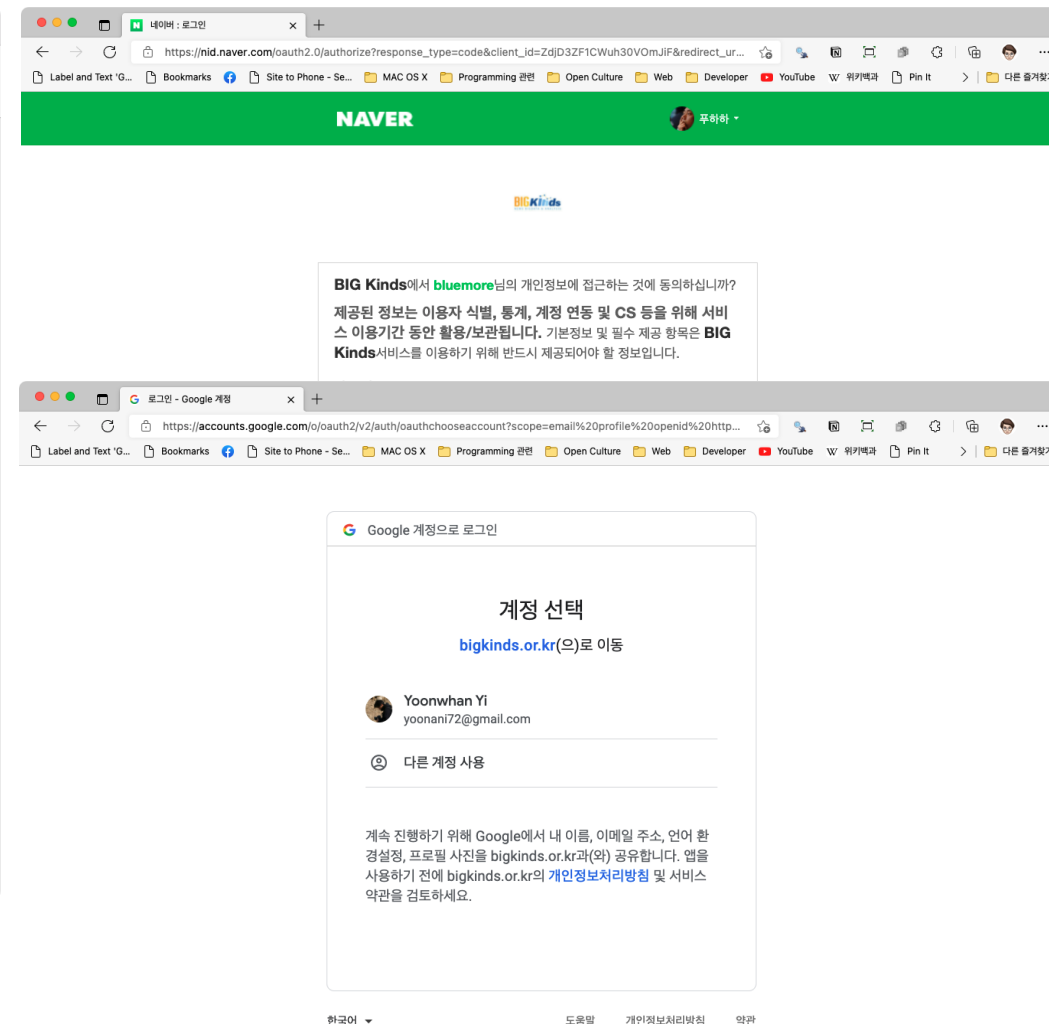
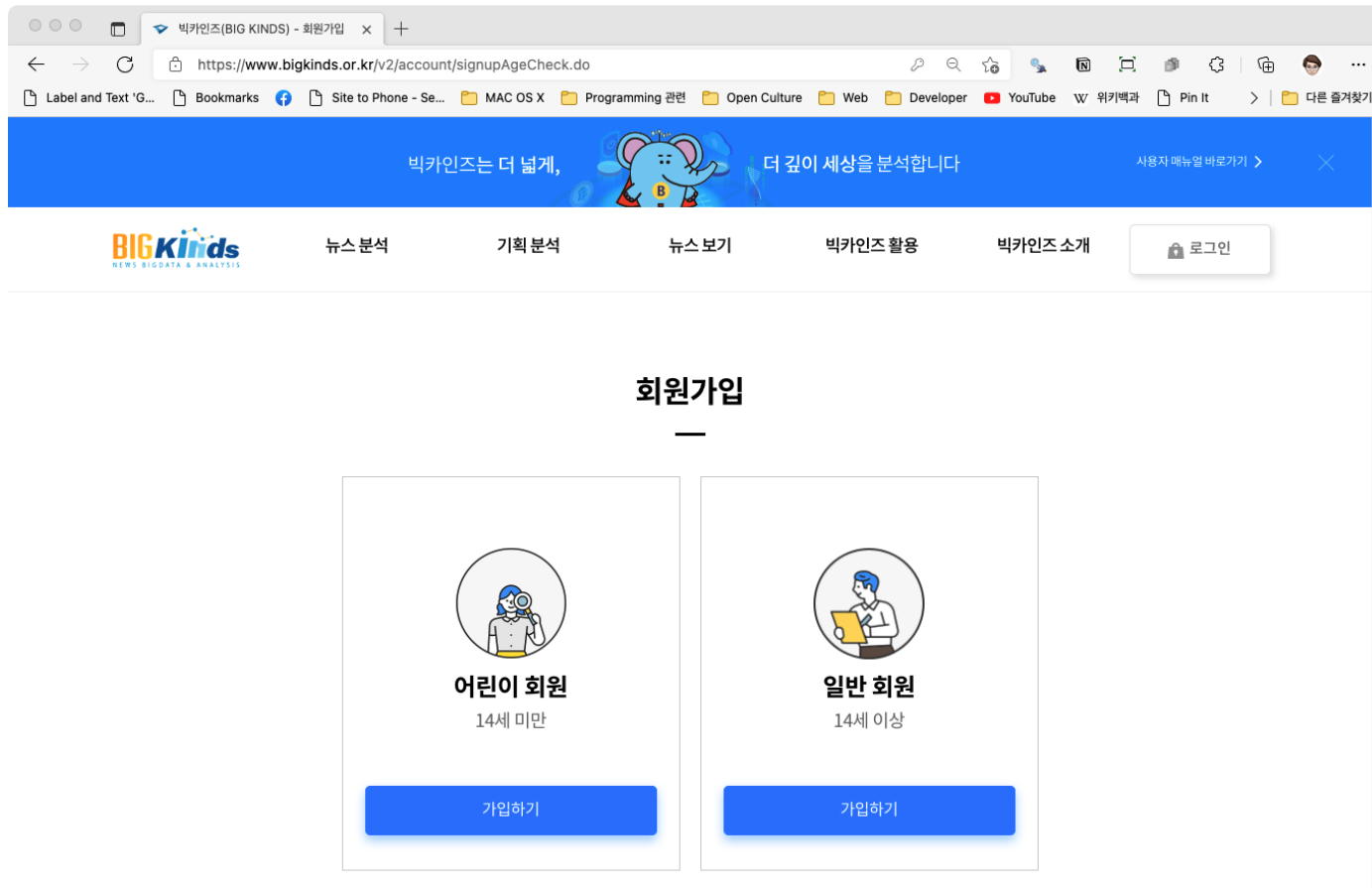
- 정형데이터와 비정형데이터
 - 정형 데이터
 - R에서 데이터 프레임과 같은 구조를 갖는 데이터
 - 관찰대상으로부터 속성을 관찰하여 테이블 구조로 저장
 - 열 구조와 저장한 데이터를 통해 정보 도출
 - 관찰하고자 하는 속성 정의가 구조를 정의하는 과정
 - 비정형 데이터
 - 사전에 정의된 구조가 없는 데이터
 - 글(텍스트)이 대표적인 형태로 하나의, 데이터 속에서 정보를 도출하여 과정이 복잡함
 - 텍스트 속 유추 가능한 구조는 문법 혹은 사람들의 표현 방법(확률 계산 등으로)
 - 사진의 경우 색상, 채도, 명도의 분포 등을 파악
 - 통계교육원 : 비정형 데이터의 가치를 캐다
 - http://sti.kostat.go.kr/window/2017b/html/2017_win_1.html

The screenshot shows the BigKinds website interface. At the top, there's a navigation bar with the BigKinds logo and a search bar. Below the navigation bar, there's a section for '오늘의 이슈' (Today's Issues) with a table of news items. The table has columns for '이슈별' (Issue), '언론사별' (Media), and '나의 관심뉴스' (My Interest News). The first row of the table shows '한국-영국 백신 스와프... 25일부터 백신 공급 전망' with 143 articles. The second row shows '“호남 지지율, 이낙연 38.5%-이재명 30.8%”' with 113 articles. The third row shows '신규 확진 1,729명...1차 접종 인구 대비 71.1%' with 58 articles. The fourth row shows '중 형다 위기에 비트코인도 위청' with 50 articles. The fifth row shows '코로나19 확진자 추가 발생' with 42 articles. Below the table, there's a section for '전체 뉴스보기' (View All News) with a list of news items. The first item is '文, 마지막 유엔무대에서 종전선언 승부수...北미사일 언급 없었다'. The second item is '바이든, 유엔총회에서 “한반도 비핵화 위한 진지한 외교 추구”'. The third item is '문 대통령, 내년 화이자 백신 추가 도입 논의'. The fourth item is '文, 임기 마지막 유엔총회 연설...한반도 평화 노력 지지 요청'. The fifth item is '美 싸이티바, 한국에 백신 원부자재 생산시설 투자'. On the right side of the page, there's a 'NEWSTORE' banner for '뉴스토어' (News Store) with a promotion for '회원가입만 하면 선물이 팡!팡!' (Just sign up and get gifts!).

<https://www.bigkinds.or.kr/>



로그인 클릭 후 회원가입



회원가입의 시간입니다 ☺

- 빅카인즈를 이용한 뉴스 기사 수집
 - 검색어 : 메타버스
 - 검색어 입력후 "상세검색" 클릭



BIGKinds
NEWS BIGDATA & ANALYSIS

메타버스

상세 검색

검색도움말

뉴스 분석 기획 분석 뉴스 보기 빅카인즈 활용

오늘의 이슈 ? | 뉴스듣기

분석대상뉴스 1,478 건 · 뉴스클러스터 122 건

한국-영국 백신 스와프...25일부터 백신 공급 전망 143건	“호남 지지율, 이낙연 38.5%-이재명 30.8%” 113건	신규 확진 1,729명...1차 접종 인구 대비 71.1% 58건
---------------------------------------	---------------------------------------	---

- 빅카인즈를 이용한 뉴스 기사 수집
 - 상세검색 설정

메타버스

Q

상세 검색

검색도움말

8 문 대통령·BTS 유엔 연설... 20건

▼

로그인

기간

—

언론사

+

통합 분류

—

사건사고 분류

—

상세검색

—

☐ 서울
 ☐ 경기
 ☐ 강원
 ☐ 충청
 ☐ 경상
 ☐ 전라
 ☐ 제주

☐ 중앙지
 ☐ 경제지
 ☐ 지역종합지
 ☐ 방송사
 ☐ 전문지

경향신문	국민일보	내일신문	동아일보	문화일보	서울신문
세계일보	조선일보	중앙일보	한겨레	한국일보	매일경제
머니투데이	서울경제	아시아경제	아주경제	파이낸셜뉴스	한국경제
헤럴드경제	강원도민일보	강원일보	경기일보	경남도민일보	경남신문
경상일보	경인일보	광주매일신문	광주일보	국제신문	대구일보

2021-06-22 ~ 2021-09-22 ×

초기화

적용하기

- 빅카인즈를 이용한 뉴스 기사 수집
 - 상세검색 설정
 - 기간 : “1개월”을 선택해 봅시다

기간	+	언론사	—	통합 분류	—	사건사고 분류	—	상세검색	—
1일	1주	1개월	3개월	6개월	1년	전체			
직접입력									
2021-08-22			📅	~	2021-09-22			📅	
2021-08-22 ~ 2021-09-22 ✕									
								초기화	적용하기

- 빅카인즈를 이용한 뉴스 기사 수집
 - 상세검색 설정
 - 언론사 : “강원”을 선택해 봅시다(강원도민일보와 강원일보가 선택됩니다).

기간	언론사	통합 분류	사건사고 분류	상세검색																																			
<input type="checkbox"/> 서울 <input type="checkbox"/> 경기 <input checked="" type="checkbox"/> 강원 <input type="checkbox"/> 충청 <input type="checkbox"/> 경상 <input type="checkbox"/> 전라 <input type="checkbox"/> 제주	<table border="1"> <tr> <td><input type="checkbox"/> 중앙지</td> <td>경향신문</td> <td>국민일보</td> <td>내일신문</td> <td>동아일보</td> <td>문화일보</td> <td>서울신문</td> </tr> <tr> <td><input type="checkbox"/> 경제지</td> <td>세계일보</td> <td>조선일보</td> <td>중앙일보</td> <td>한겨레</td> <td>한국일보</td> <td>매일경제</td> </tr> <tr> <td><input type="checkbox"/> 지역종합지</td> <td>머니투데이</td> <td>서울경제</td> <td>아시아경제</td> <td>아주경제</td> <td>파이낸셜뉴스</td> <td>한국경제</td> </tr> <tr> <td><input type="checkbox"/> 방송사</td> <td>헤럴드경제</td> <td>강원도민일보</td> <td>강원일보</td> <td>경기일보</td> <td>경남도민일보</td> <td>경남신문</td> </tr> <tr> <td><input type="checkbox"/> 전문지</td> <td>경상일보</td> <td>경인일보</td> <td>광주매일신문</td> <td>광주일보</td> <td>국제신문</td> <td>대구일보</td> </tr> </table>				<input type="checkbox"/> 중앙지	경향신문	국민일보	내일신문	동아일보	문화일보	서울신문	<input type="checkbox"/> 경제지	세계일보	조선일보	중앙일보	한겨레	한국일보	매일경제	<input type="checkbox"/> 지역종합지	머니투데이	서울경제	아시아경제	아주경제	파이낸셜뉴스	한국경제	<input type="checkbox"/> 방송사	헤럴드경제	강원도민일보	강원일보	경기일보	경남도민일보	경남신문	<input type="checkbox"/> 전문지	경상일보	경인일보	광주매일신문	광주일보	국제신문	대구일보
<input type="checkbox"/> 중앙지	경향신문	국민일보	내일신문	동아일보	문화일보	서울신문																																	
<input type="checkbox"/> 경제지	세계일보	조선일보	중앙일보	한겨레	한국일보	매일경제																																	
<input type="checkbox"/> 지역종합지	머니투데이	서울경제	아시아경제	아주경제	파이낸셜뉴스	한국경제																																	
<input type="checkbox"/> 방송사	헤럴드경제	강원도민일보	강원일보	경기일보	경남도민일보	경남신문																																	
<input type="checkbox"/> 전문지	경상일보	경인일보	광주매일신문	광주일보	국제신문	대구일보																																	
2021-08-22 ~ 2021-09-22 × 강원도민일보 × 강원일보 × 초기화 적용하기																																							

- 빅카인즈를 이용한 뉴스 기사 수집
 - 상세검색 설정
 - 통합분류 : IT외 다른 분야의 관점을 알아보려고 하니 ‘정치’, ‘사회’, ‘문화’, ‘지역’, ‘스포츠’ 선택

기간	언론사	통합 분류	사건사고 분류	상세검색
		<div> <div> + <input checked="" type="checkbox"/> 정치 + <input type="checkbox"/> 경제 + <input checked="" type="checkbox"/> 사회 + <input checked="" type="checkbox"/> 문화 + <input type="checkbox"/> 국제 + <input checked="" type="checkbox"/> 지역 + <input checked="" type="checkbox"/> 스포츠 </div> </div>		
<div> 2021-08-22 ~ 2021-09-22 × 강원도민일보 × 강원일보 × 행정_자치 × 북한 × 국회_정당 × 외교 × 정치일반 × 선거 × 청와대 × 정치 × 의료_건강 × 환경 × 사건_사고 × 여성 × 장애인 × 날씨 × 노동_복지 × 사회일반 × 미디어 × 교육_시험 × 사회 × 영화 × 문화일반 × 미술_건축 × 음악 × 방송_연예 × 학술_문화재 × 종교 × 요리_여행 × 생활 × 전시_공연 × 출판 × 문화 × 충남 × 대전 × 경남 × 제주 × 대구 × 경기 × 지역일반 × 울산 × 광주 × 강원 × 전북 × 충북 × 부산 × 경북 × 전남 × 지역 × 야구 × 한국프로야구 × 메이저리그 × 일본프로야구 × 스포츠일반 × 축구 × 해외축구 × 국가대표팀 × 한국프로축구 × 올림픽_아시안게임 × 농구_배구 × 골프 × 월드컵 × 스포츠 × </div> <div> 초기화 <div>적용하기</div> </div>				

- 빅카인즈를 이용한 뉴스 기사 수집
 - 검색화면
 - 뉴스, 인용문, 사설별로 활용
 - 뉴스(기본) 클릭 후,
하단의 STEP 03 클릭

빅카인즈(BIG KINDS) - 뉴스 검색

https://www.bigkinds.or.kr/v2/news/search.do

STEP 01 · 뉴스 검색 - 메타버스

STEP 02 · 검색 결과

검색필터 초기화

기간 +

2021 (20) ☐

언론사 +

경향신문 (0) ☐

국민일보 (0) ☐

내일신문 (0) ☐

동아일보 (0) ☐

문화일보 (0) ☐

펼쳐보기

통합분류 +

정치(0) ☒

경제(1) ☐

사회(1) ☒

뉴스 인용문 사설

최신순 10건씩 보기 결과 내 재검색

분석제외 취소 (0) 검색식 저장 뉴스분석 리포트 생성

1 / 2

체크 버튼을 선택하면 분석에서 제외됩니다.

"메타버스"

뉴스 검색 결과 20 건입니다.

2021-08-22 ~ 2021-09-22 기준

추석이란 무엇인가? 정답은 당신 마음 속에

집콕 추석 연휴 잘 지내는 법예측·교육·전시 프로그램 다채유튜브로 보름달 공개 관측회메타버스 활용 어린이 콘텐츠 김...

강원도민일보 문화>방송_연예 | 문화>미술_건축 | 문화>전시_공연 2021/09/17 김여진

[강릉]인터포 지역특화산업육성 공모 선정

[강릉]강릉 소재 IT솔루션 전문기업 (주)인터포가 최근 중소벤처기업부에서 주관하는 '2021년 지역특화산업육성+(R&D...

강원일보 IT_과학>모바일 | 지역>울산 | 지역>충북 2021/09/16 고달순

[언종언]메타버스의 미래

'현실세계와 같은 사회·경제·문화 활동이 이뤄지는 3차원 가상세계.' 최근 혁신적인 기술로 돌풍을 일으키고 있는 메타버...

강원일보 문화>문화일반 | IT_과학>모바일 | IT_과학>콘텐츠 2021/09/10

- 빅카인즈를 이용한 뉴스 기사 수집
 - 분석 및 시각화
 - 데이터 다운로드
 - 엑셀로 다운로드 받아 data 폴더로 이동합니다.
 - 파일명은 “news.xlsx”로 변경

STEP 03 · 분석 결과 및 시각화

데이터 다운로드	관계도 분석	키워드 트렌드	연관어 분석	정보 추출
----------	--------	---------	--------	-------

검색한 뉴스의 메타데이터(언론사, 기고자, 제목 등)와 개체명(인물, 기관, 장소 등) 분석 데이터를 엑셀파일로 제공하는 서비스입니다.



데이터 다운로드는 최대 20,000건의 데이터가 다운로드 됩니다. 미리보기는 최대 20개까지 보여집니다.

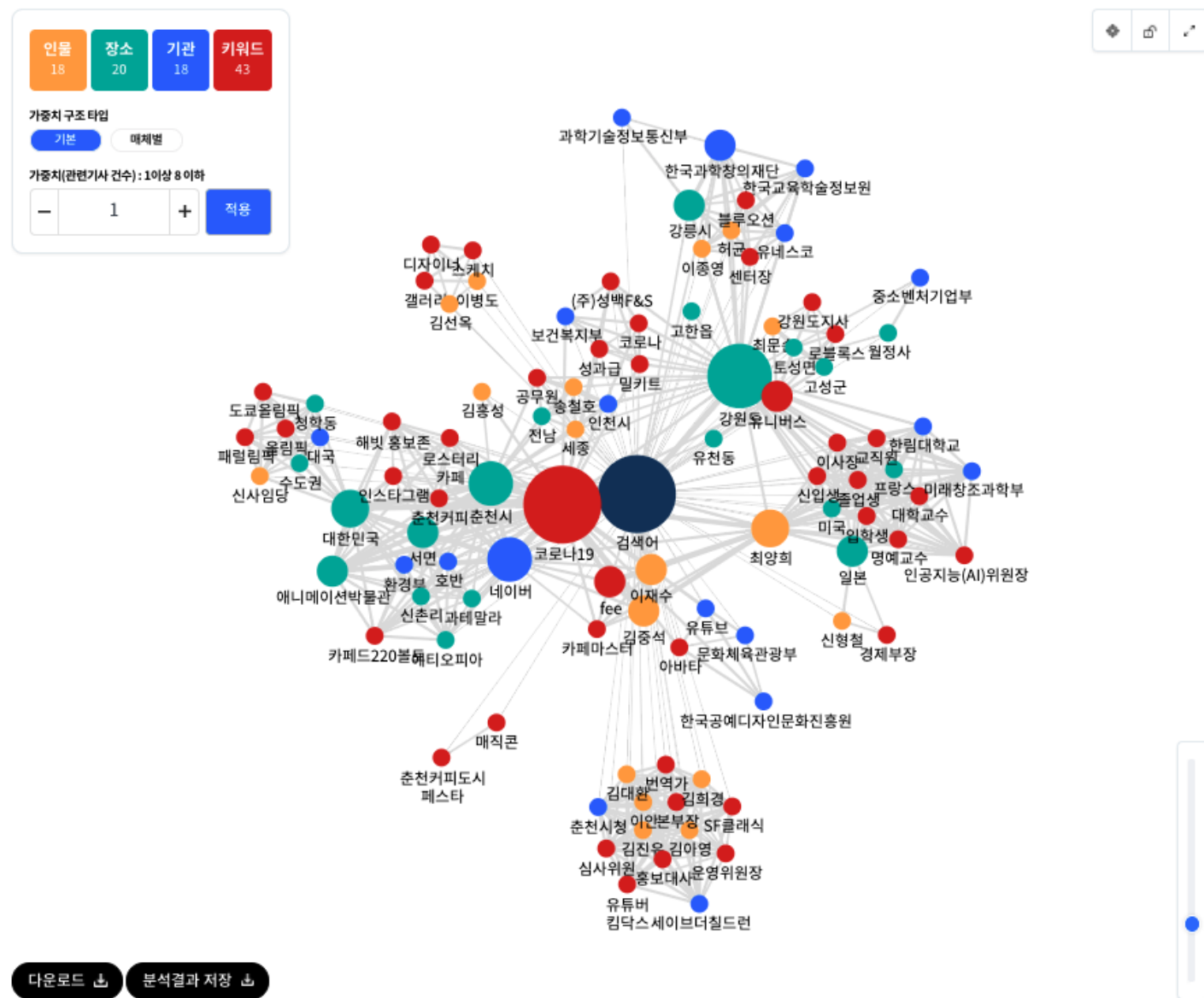
'키워드' 항목은 본문 내에서 추출된 키워드 중 단순 숫자(1, 2, 2018, 2019 등), 이메일 주소, 시간을 뜻하는 단어(밤, 낮, 새벽 등)를 제외한 결과가 표시됩니다.

	뉴스 식별자	일자	언론사	기고자	제목	통합 분류1	통합 분류2
1	01300101.20210917001654001	20210917	강원도민일보	김여진	추석이란 무엇인가? 정답은 당신 마음 속에	문화>방송_연예	문화>미술_2
2	01300201.20210916024411001	20210916	강원일보	고달순	[강릉]인터포 지역특화산업육성 공모 선정	IT_과학>모바일	지역>울진
3	01300201.20210910024237001	20210910	강원일보		[언중언]메타버스의 미래	문화>문화일반	IT_과학>모바일
4	01300201.20210910024231001	20210910	강원일보		[The 초점]이제는 메타버스 올림픽이다	문화>출판	문화>미술_2
5	01300201.20210909024417001	20210909	강원일보	장현정	[춘천]메타버스 도시'로 급부상하는 춘천시	IT_과학>콘텐츠	지역>울진
6	01300201.20210908024413001	20210908	강원일보	최기영	내년 강원세계산림엑스포 '메타버스' 플랫폼 구축	지역>전남	지역>강원
7	01300201.20210907024352002	20210907	강원일보	김도균	[강릉]"강릉 문화자원 - 메타버스 연계 시장 선점"	지역>경남	지역>경북
8	01300201.20210906024112001	20210906	강원일보	김수빈	춘천의 커피향기 한 폭에 담아	문화>전시_공연	문화>미술_2
9	01300101.20210906001612003	20210906	강원도민일보	오세현	국내 첫 메타버스 활용 '커피도시 춘천' 선포	지역>경남	지역>대전
10	01300101.20210905161913001	20210905	강원도민일보	이성찬	메타버스·로봇바리스타와 함께하는 '춘천커피도시 페스타' 개막	지역>경남	지역>제주
11	01300101.20210904162622001	20210904	강원도민일보	오세현	춘천커피도시 페스타 개막...메타버스로 춘천커피 '한 눈에'	지역>경남	지역>대전
12	01300201.20210903030142001	20210903	강원일보	김도균	[강릉]메타버스 추진 토론회	지역>경남	IT_과학>IT_과학
13	01300101.20210903001623006	20210903	강원도민일보	박지은	정선 폐광촌 골목길 '3차원 가상세계' 구축, Z세대 끌어모은다	지역>경남	지역>전북
14	01300201.20210902024137001	20210902	강원일보	이현정	[춘천]춘천영화제 'SF영화제'로 30일 개막	문화>영화	문화>전시_2
15	01300201.20210901024316002	20210901	강원일보	최기영	'메타버스에 올라타라' 지자체 선정 경쟁	지역>경남	지역>충남
16	01300101.20210831001610001	20210831	강원도민일보	이연제	유천동에는 '메타버스' 카페·클럽이 있다	지역>지역일반	지역>충남
17	01300201.20210830024509001	20210830	강원일보	장현정	[특집]호반의 도시 춘천 '커피도시'로 새로운 도약 꿈꾼다	지역>경남	지역>제주
18	01300201.20210830024419001	20210830	강원일보	오석기	최양희 한림대총장 "4세대 대학은 열린대학...지역과 협력해 사회발전 중심에 서겠다"	사회>교육_시험	IT_과학>IT_과학
19	01300101.20210830001552001	20210830	강원도민일보	오세현	막국수·달걀비 이어 지역 대표할 '춘천커피' 새 출발 알린다	IT_과학>콘텐츠	지역>경남
20	01300201.20210826024366001	20210826	강원일보	최영재	[강원경제인 대상 수상기업]지역주민 108명 채용...도 프랜차이즈 최초 가맹점 120호 돌파	경제>취업_창업	경제>유통

- 빅카인즈를 이용한 뉴스 기사 수집
 - 분석 및 시각화
 - 관계도 분석
 - 가중치 조절
 - 데이터와 이미지파일 다운로드

기사 데이터 (Excel)
 그래프 데이터 (Excel)
 이미지 파일 (PNG)
 이미지 파일 (JPG)

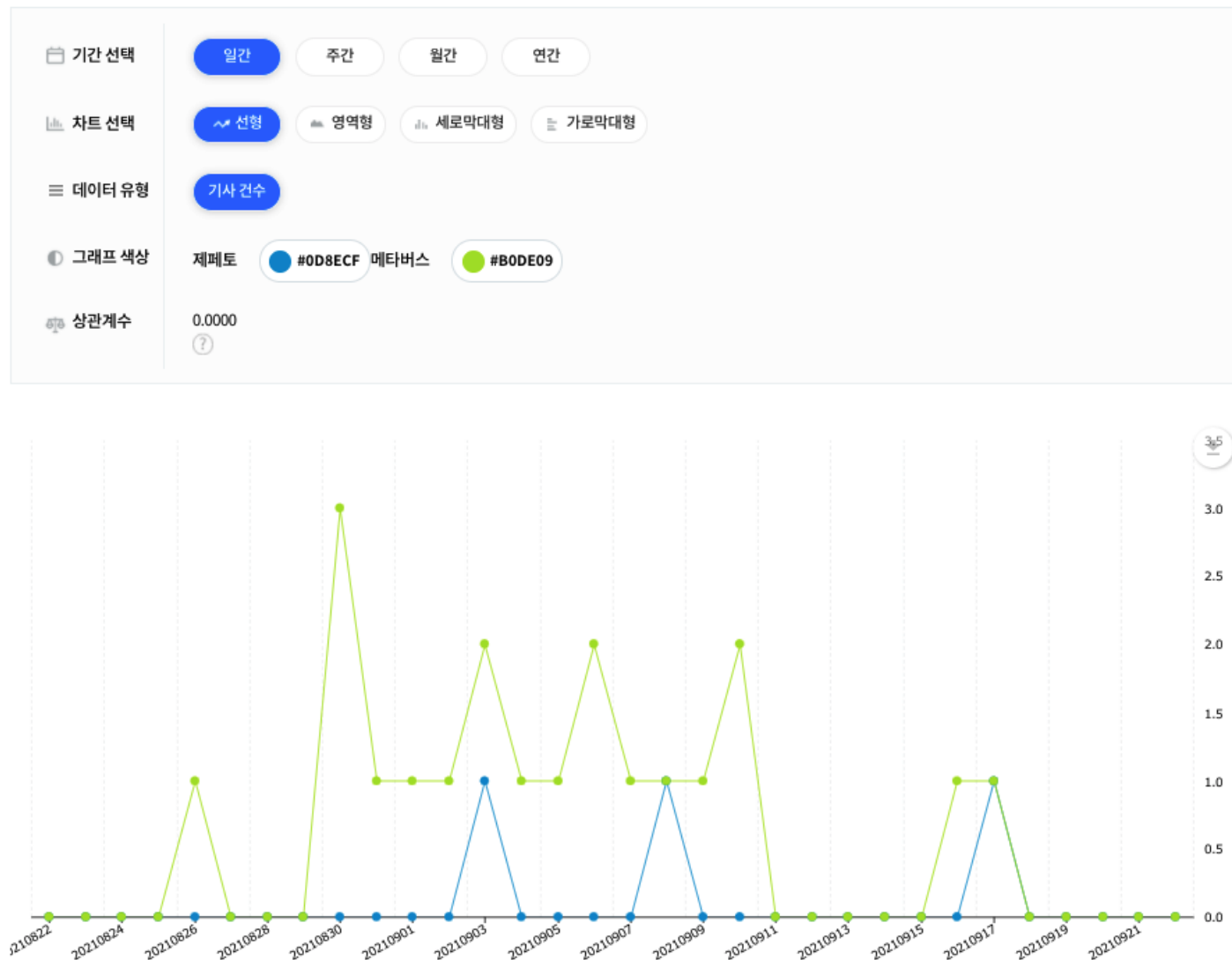
다운로드 
 분석결과 저장 



- 빅카인즈를 이용한 뉴스 기사 수집
 - 분석 및 시각화
 - 키워드 트렌드
 - 일간, 주간, 월간, 연간



- 빅카인즈를 이용한 뉴스 기사 수집
 - 분석 및 시각화
 - 키워드 트렌드
 - 키워드를 콤마(,)로 연결하여 검색하면, 두 개의 키워드에 대한 결과를 보여줍니다.



- 빅카인즈를 이용한 뉴스 기사 수집
 - 분석 및 시각화
 - 연관어 분석
 - 단어구름

분석 뉴스 건수: 100 300 500 800 1,000

차트 선택: 워드클라우드 막대그래프

데이터 유형: 가중치 ? 기사 건수



- R을 이용하여 뉴스 기사를 분석하기 위한 준비
 - 필요한 패키지 : 전희원님이 개발하신 KoNLP를 사용합니다.
 - KoNLP를 위해 Java 실행 환경(JRE)이 필요합니다.
 - 설치전에 기존에 설치한 패키지들을 업데이트 합니다.
 - RStudio가 아닌 R 에서 실행합니다. (Rstudio는 꺼주세요)
 - 다음의 명령을 입력합니다.

```
install.packages( installed.packages()[ , 1], repos="https://cran.seoul.go.kr" )
```

- Rtools를 설치합니다.
 - Windows : <https://cran.seoul.go.kr/bin/windows/Rtools/>
 - Installing Rtools 에서 64비트용 다운로드 후 실행
 - 설치후 R에서 다음을 실행

```
writeLines('PATH="${RTTOOLS40_HOME}\\usr\\bin;${PATH}"', con = "~/.Renviron")
```

- Rtools를 설치합니다.
 - Mac : <https://mac.r-project.org/tools/>
 - Apple 에서 제공하는 Xcode를 설치합니다(앱스토어)
 - 위 링크에서 gfortran-x.y-Mojave.dmg 를 다운로드 받은 후 설치합니다.
 - 이후 다음을 터미널(응용 프로그램) 에서 다음을 실행합니다.
`export PATH=$PATH:/usr/local/gfortran/bin`
- Rtools 설치 이후 R을 종료 후 다시 시작합니다.
- KoNLP 설치
 - Java 설치 : 박찬영님이 개발하신 multilinguer 패키지 이용
 - 다음을 실행합니다. 실행 후 필요한 패키지를 R이 설치합니다.
> `install.packages("multilinguer")`
> `library(multilinguer)`
> `install_jdk()`

- KoNLP 설치
 - 필요 패키지 설치
 - 다음을 실행합니다.

```
install.packages( c("hash", "tau", "Sejong", "RSQLite", "devtools", "bit",  
"rex", "lazyeval", "htmlwidgets", "crosstalk", "promises", "later",  
"sessioninfo", "xopen", "bit64", "blob", "DBI", "memoise", "plogr", "covr",  
"DT", "rcmdcheck", "rversions"), type = "binary")
```

- KoNLP 설치
 - github에 있는 KoNLP를 설치하기 위해 remotes 패키지를 설치합니다.
 - KoNLP를 설치합니다.

```
> install.packages( "remotes" )  
> remotes::install_github('haven-jeon/KoNLP', INSTALL_opts=c("--no-multiarch"))
```

- 분석을 위해 필요한 패키지 연결하기

```
library( tidyverse )  
library( readxl )  
library( KoNLP )
```

- 데이터 읽어오기 : 빅카인즈에서 다운로드 받은 파일 이용
 - KoNLP에서는 문자열 벡터를 요구하므로 “본문”열의 데이터를 벡터로 저장합니다.

```
news <- read_excel("./data/news.xlsx")  
articles <- news$본문
```

- 품사를 구분하기 위한 사전 읽어오기

```
useNIADic()
```

- 분석을 위한 전처리
 - 품사 구분시 불필요한 단어 제거
 - 한글만 사용하기
 - `str_replace_all()` : 문자열 벡터에서 패턴을 찾아 원하는 문자열로 변경합니다.
 - `swords` 는 문자열 들을 `|` 로 구분한 것으로 패턴에서 OR의 역할을 합니다.
 - 패턴 `[^가-힣]` 은 한글 가부터 힣 즉, 모든 한글과 띄어쓰기를 제외한 문자를 뜻합니다.

```
swords <- "해서|하면|하지|들이|당시|생각|경우|하게|때문|하기"  
atc <- str_replace_all( articles, swords, " ")
```

```
articles[1:3 ]  
atc <- str_replace_all( atc, "[^가-힣]", " ")  
atc[1:3]
```

- 명사 추출하기
 - KoNLP의 `extractNoun()` 함수를 이용합니다.
 - 문자열 벡터의 각 원소로부터 찾은 명사들을 리스트로 저장합니다.

```
atc_nouns <- extractNoun( atc )  
str( atc_nouns )
```

List of 20

```
$ : chr [1:45] "콧" "추석" "연휴" "예술" ...  
$ : chr [1:45] "강릉" "강릉" "소재" "솔루션" ...  
$ : chr [1:36] "현실" "세계" "사회" "경제" ...  
$ : chr [1:39] "강릉" "문자" "원" "풍부" ...  
$ : chr [1:47] "메타버스" "플랫폼" "활용" "축제" ...  
$ : chr [1:49] "강원" "세계" "산림" "엑스포" ...  
$ : chr [1:40] "강릉" "강릉시" "미래" "신산업" ...
```

- 리스트를 하나의 벡터로 만듭니다

```
n_words <- simplify( atc_nouns )
```


- 단어의 출현 빈도 세기
 - 두 글자 이상으로 된 단어만 분석에 사용합니다.
 - `str_length()` 함수는 글자수를 세어 줍니다.
 - `dplyr`의 `count()` 함수는 데이터 프레임의 특정 열의 값별로 갯수를 세어줍니다.
 - `sort` 인수에 `TRUE`를 전달하면 내림 차순으로 정렬합니다.

```
data.frame( word = n_words ) %>%  
filter( str_length( word ) > 1 ) %>%  
count( word, sort=TRUE )
```

	word	n
1	춘천	27
2	메타버스	25
3	도시	19
4	커피	19
5	강릉	13
6	플랫폼	11
7	세계	10
8	춘천시	8
9	가상	7
10	강원	7

- 워드클라우드 만들기
 - wordcloud2 패키지의 wordcloud2() 함수를 사용합니다.

```
library( wordcloud2 )
```

```
data.frame( word = n_words ) %>%
  filter( str_length( word ) > 1 ) %>%
  count( word, sort=TRUE ) %>%
  wordcloud2()
```



- 인접 단어를 통한 연관 분석
 - 인접한 두 단어를 하나로 묶어 보시다.
 - `lead()` 함수는 바로 다음 벡터의 원소를 알려줍니다.

```
data.frame( word = n_words ) %>%  
  filter( str_length( word ) > 1 ) %>%  
  mutate( nxt_word = lead(word) )
```

	word	nxt_word
1	추석	연휴
2	연휴	예술
3	예술	교육
4	교육	전시
5	전시	프로그램

- 인접 단어를 통한 연관 분석
 - 인접한 두 단어의 빈도를 구해 봅시다.
 - count() 함수 이용을 위해 두 단어를 하나의 단어로 합칩니다
 - paste() 함수는 문자열을 sep 인수로 전달한 문자를 통해 서로 연결합니다.
 - 문자열을 합치기 전에 앞뒤의 단어가 동일한 단어쌍은 제거합니다.

```
data.frame( word = n_words ) %>%  
  filter( str_length( word ) > 1 ) %>%  
  mutate( nxt_word = lead(word) ) %>%  
  filter( word != nxt_word ) %>%  
  mutate( bi_gram = paste( word, nxt_word, sep='-' ) ) %>%  
  count( bi_gram, sort = TRUE)
```

	bi_gram	n
1	커피-도시	14
2	춘천-커피	9
3	도시-페스	7
4	메타버스-플랫폼	6
5	애니메이션-박물관	6

- 인접 단어를 통한 연관 분석
 - 네트워크 도표를 그리기 위해 합친 단어들을 다시 분리합니다
 - dplyr의 separate() 함수는 기존 열의 값을 분리하여 새로운 열에 저장하는 역할을 합니다.
 - 앞서 paste() 함수의 예제를 보기 위해 사용하였으나, 기존 두 열을 하나의 열로 합치는, 즉 separate() 와 반대 역할을 수행하는 unite() 함수도 있습니다.
 - 분리한 값을 갖는 데이터 프레임을 저장합니다 (할당연산자의 방향을 바꿔서 사용할 수 있습니다).

```
data.frame( word = n_words ) %>%  
  filter( str_length( word ) > 1 ) %>%  
  mutate( nxt_word = lead(word) ) %>%  
  filter( word != nxt_word ) %>%  
  mutate( bi_gram = paste( word, nxt_word, sep="-" ) ) %>%  
  count( bi_gram, sort = TRUE ) %>%  
  separate( bi_gram, c("word", "nxt_word"), sep="-" ) -> bi_gram_df
```

- 인접 단어를 통한 연관 분석
 - 네트워크 도표를 그리기 위해 필요한 패키지를 설치하고 연결합니다.
 - ggraph 패키지의 ggraph() 함수는 네트워크 도표를 그리는데 사용합니다.
 - tidygraph 패키지의 as_tbl_graph() 함수는 네트워크 분석을 위한 자료를 만듭니다(graph)
 - 가장 많이 등장 한 상위 20개 단어쌍을 사용하고자 합니다.

```
library( ggraph )  
library( tidygraph )
```

```
# 그래프 형태로 표현  
bi_gram_df %>%  
  head( 20 ) %>%  
  as_tbl_graph()
```

```
# A tbl_graph: 27 nodes and 20 edges  
#  
# A directed simple graph with 8 components  
#  
# Node Data: 27 × 1 (active)  
  name  
  <chr>  
1 커피  
2 춘천  
3 도시  
4 메타버스  
5 애니메이션
```

```
# Edge Data: 20 × 3  
  from    to    n  
  <int> <int> <int>  
1      1      3  14  
2      2      1   9  
3      3     11   7  
# ... with 17 more rows
```


- 인접 단어를 통한 연관 분석
 - tbl_graph 살펴보기

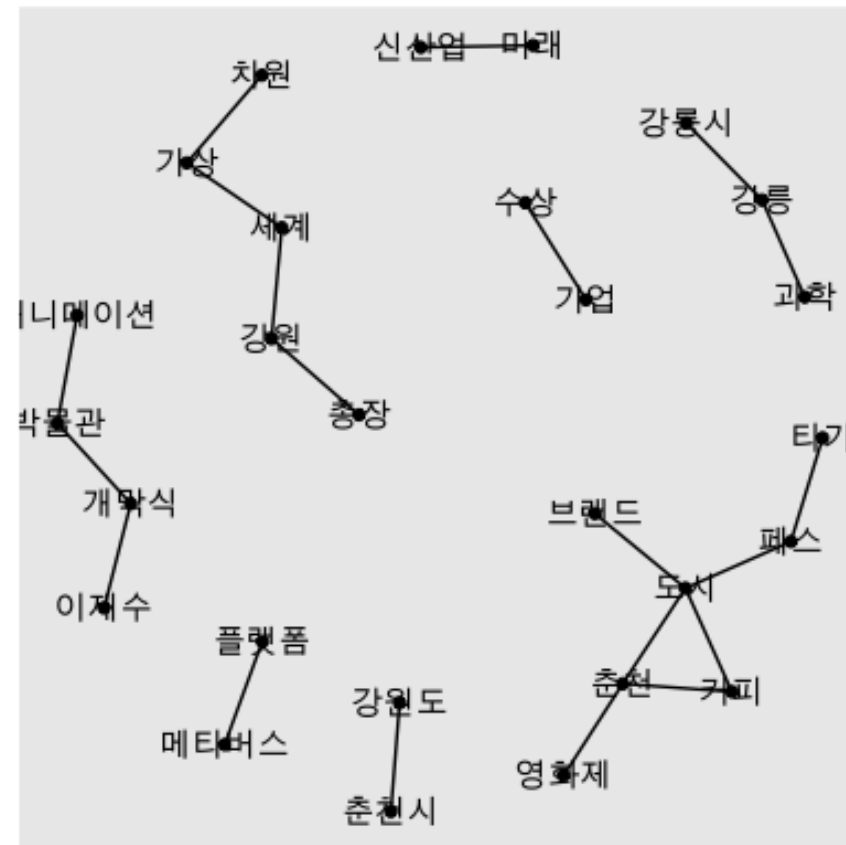
```
bi_gram_df %>%
  head( 20 ) %>%
  as_tbl_graph() -> grp_1

grp_1 %>% data.frame()
grp_1 %>% activate( edges ) %>%
data.frame( )
```

	name		from	to	n
1	커피	1	1	3	14
2	춘천	2	2	1	9
3	도시	3	3	11	7
4	메타버스	4	4	16	6
5	애니메이션	5	5	8	6
6	가상				
7	미래				
8	박물관				
9	수상				
10	차원				

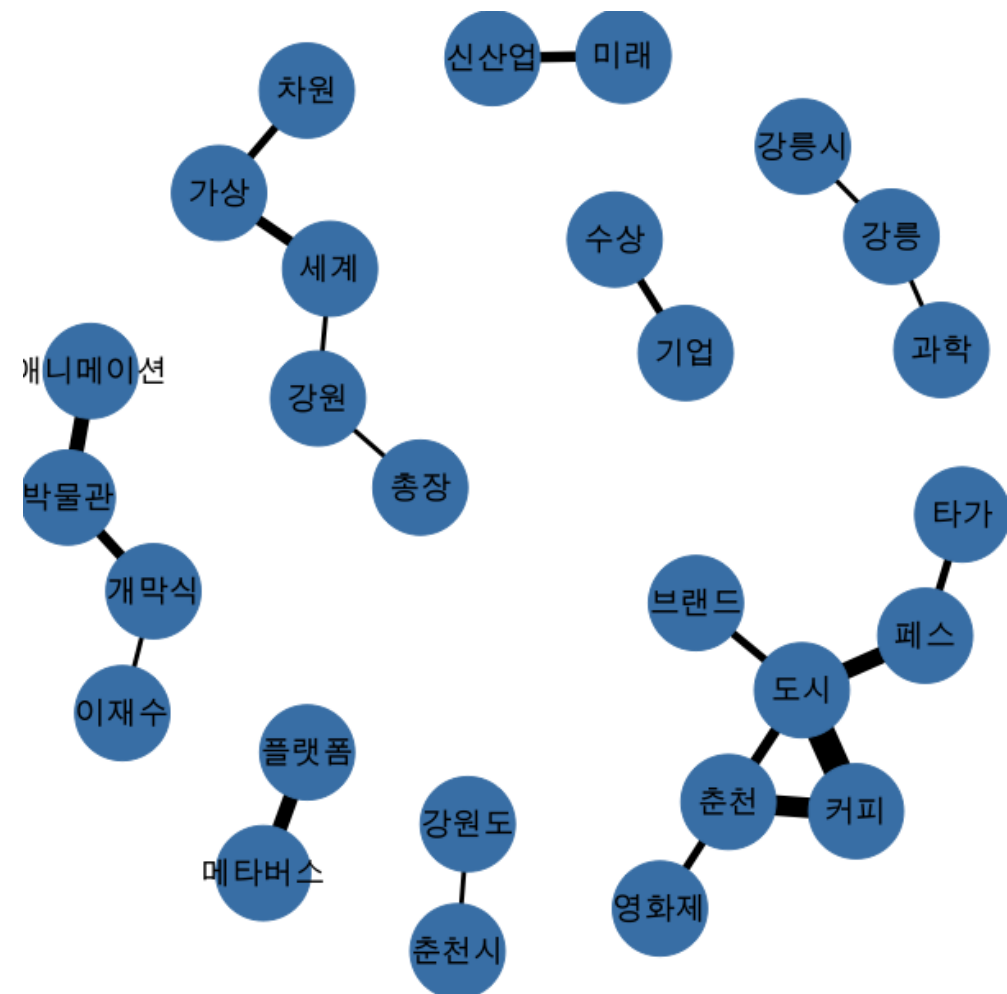
- 인접 단어를 통한 연관 분석
 - ggraph() 함수로 네트워크 도표를 그립니다.
 - node : 여기서는 각 단어를 나타냅니다.
 - geom_edge_link()
 - link : 각 노드를 연결하는 선입니다.
 - geom_node_point(), geom_node_text()

```
bi_gram_df %>%
  head( 20 ) %>%
  as_tbl_graph() %>%
  ggraph( layout = "kk") +
    geom_edge_link() +
    geom_node_point() +
    geom_node_text( aes(label=name))
```



- 인접 단어를 통한 연관 분석
 - 의미 파악을 위해 조금 더 꾸며봅시다.
 - link 의 두께를 출현수의 제곱근 값으로 표현
 - 노드의 크기 확대와 색상 변경
 - 바탕화면 색을 없애기 위해 테마 변경

```
bi_gram_df %>%
  head( 20 ) %>%
  as_tbl_graph() %>%
  ggraph( layout = "kk") +
  geom_edge_link(
    aes( width = sqrt( n ) ), show.legend = FALSE
  ) +
  geom_node_point( size=20, color="steelblue" ) +
  geom_node_text( aes(label=name), size = 5) +
  theme_void()
```

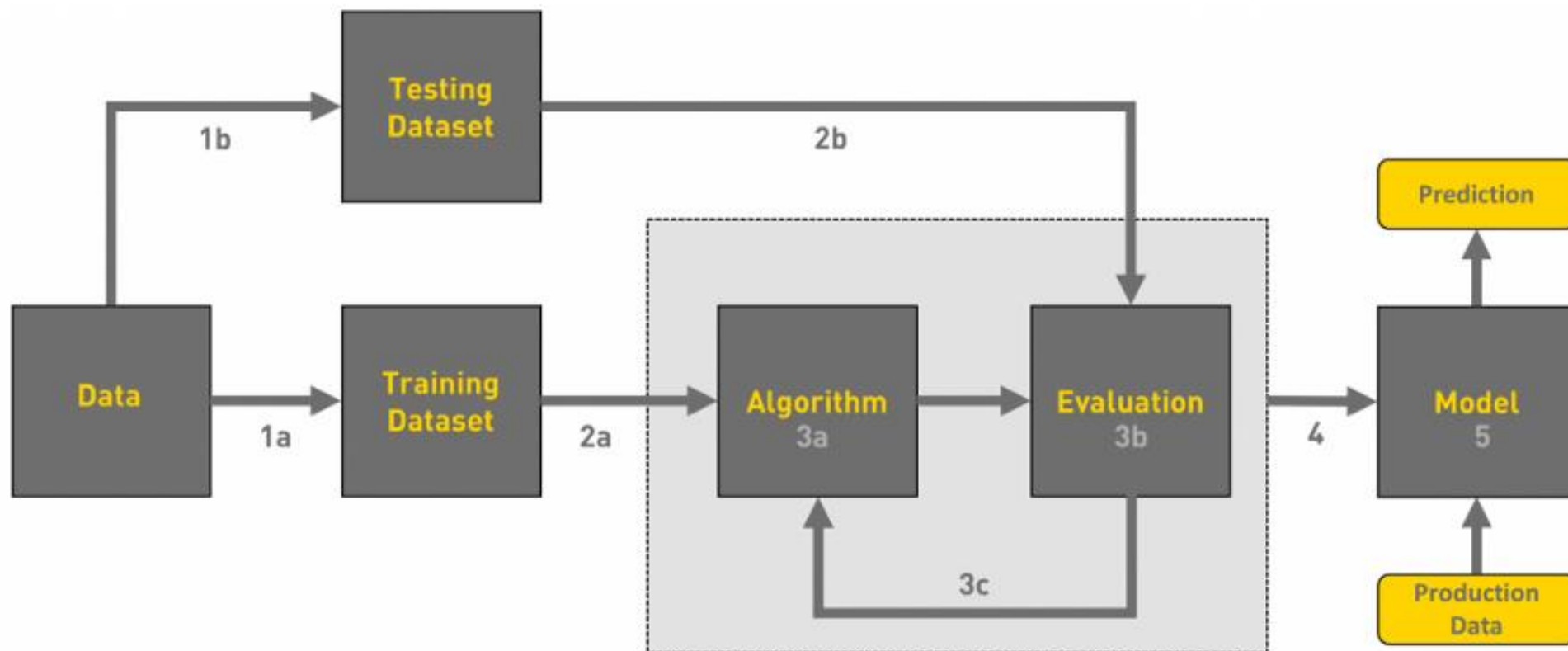


- 성능 향상을 위해 필요한 것
 - 전처리 강화(stopwords, 오타자, 춘천시와 춘천 등)
 - 사용자 정의 사전 구축

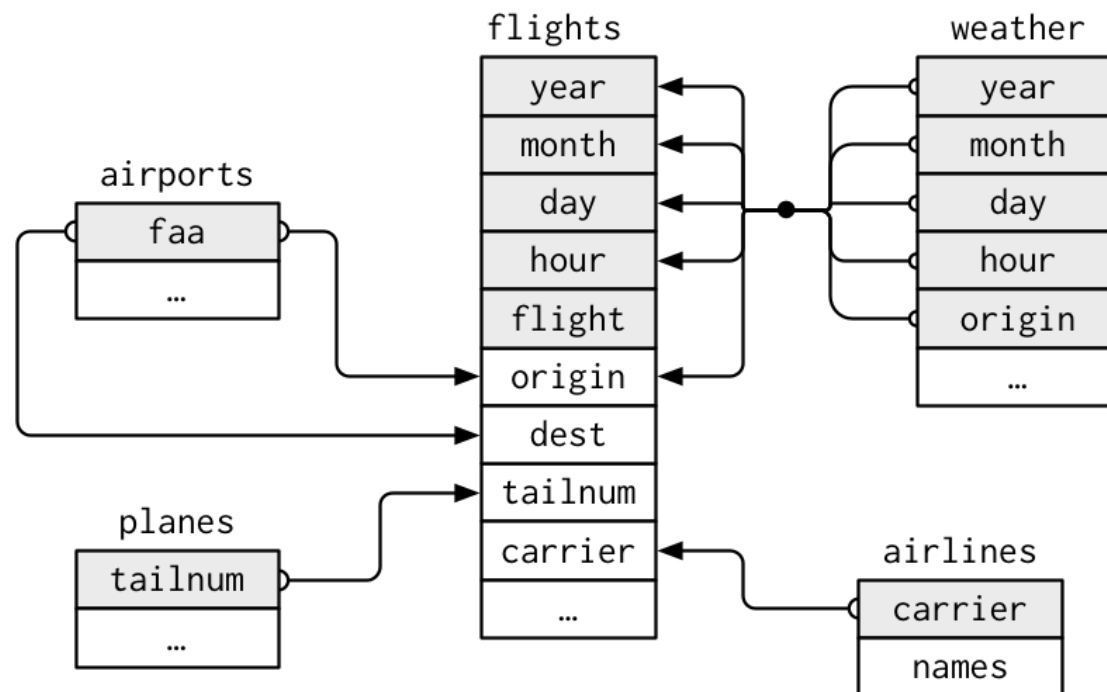
```
user_dict <- data.frame( term=c( " 페스티벌", "애니메이션 박물관",  
                                "사회적경제", "커먼즈필드"),  
                          tag="ncn" )  
dics <- c("sejong", "woorimalsam")  
category <- c("political")  
buildDictionary( ext_dic=dics,  
                 category_dic_nms = category,  
                 user_dic = user_dict,  
                 replace_usr_dic = TRUE)  
get_dictionary("user_dic")
```

- 기계학습(Machine Learning), 위키피디아
 - 경험을 통해 자동으로 개선하는 컴퓨터 알고리즘의 연구이다.
 - 기계가 일일이 코드로 명시하지 않은 동작을 데이터로부터 학습하여 실행할 수 있도록 하는 알고리즘을 개발하는 연구 분야 – 아서 사무엘(Arthur Samuel) 1959, IBM R&D
 - 표현(representation)과 일반화(generalization)
 - 표현 : 데이터의 평가
 - 일반화 : 아직 알 수 없는 데이터에 대한 처리
 - 기계 학습의 유형
 - 감독학습 (Supervised Learning)
 - 훈련 데이터(Training Data)로부터 하나의 함수를 유추해내는 방법
 - 분류(Classification)와 회귀(Regression)
 - 자율학습, 비지도학습 (Unsupervised Learning)
 - 정답 라벨이 없는 데이터를 비슷한 특징끼리 군집화 하여 새로운 데이터에 대한 결과를 예측하는 방법
 - 군집화(clustering), 차원 축소(Dimensionality Reduction)
 - 강화학습 (Reinforced Learning)
 - 행동심리학에서 영감을 받았으며, 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법
 - 게임과 실시간 결정 등

- 기계학습 과정
 - 데이터 수집과 정리
 - 데이터 분리 : Train data, Test data
 - 모델/알고리즘 수립
 - 평가
 - 배포



- 모델링과 기계학습을 위한 R 패키지 : tidymodels
 - The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles.
 - 홈페이지 : <https://www.tidymodels.org/>
 - 오늘의 예제 : <https://www.tidymodels.org/start/recipes/>
 - 필요한 패키지
 - `library(tidyverse)`
 - `library(tidymodels)`
 - `library(nycflights13)`
 - `library(skimr)`
 - 예제 데이터
 - flights : 항공기의 이착륙 데이터
 - weather : 날짜별 일기 상황



- 데이터 준비하기

```
flight_data <- flights %>%  
  mutate(  
    arr_delay = ifelse(arr_delay >= 30, "late", "on_time"),  
    arr_delay = factor(arr_delay),  
    date = lubridate::as_date(time_hour)  
  ) %>%  
  inner_join(weather, by = c("origin", "time_hour")) %>%  
  select(dep_time, flight, origin, dest, air_time, distance,  
         carrier, date, arr_delay, time_hour) %>%  
  na.omit() %>%  
  mutate_if(is.character, as.factor)
```

	year	month	day	carrier	flight	origin	time_hour
1	2013	1	1	UA	1545	EWR	2013-01-01 05:00:00
2	2013	1	1	UA	1714	LGA	2013-01-01 05:00:00
3	2013	1	1	AA	1141	JFK	2013-01-01 05:00:00
4	2013	1	1	B6	725	JFK	2013-01-01 05:00:00
5	2013	1	1	DL	461	LGA	2013-01-01 06:00:00

	year	month	day	origin	time_hour
1	2013	1	1	EWR	2013-01-01 01:00:00
2	2013	1	1	EWR	2013-01-01 02:00:00
3	2013	1	1	EWR	2013-01-01 03:00:00
4	2013	1	1	EWR	2013-01-01 04:00:00
5	2013	1	1	EWR	2013-01-01 05:00:00
6	2013	1	1	EWR	2013-01-01 06:00:00

- 다음을 실행해 봅시다
 - 데이터 살펴보기

```
glimpse( flight_data )
```

- 지연과 비지연의 비율 살펴보기 : 로지스틱 모형의 반응변수

```
flight_data %>%  
  count( arr_delay ) %>%  
  mutate( prop = n / sum(n) )
```

- skimr::skim() : 요약

```
flight_data %>%
  skim(dest, carrier)
```

```
— Data Summary —
Name                Values
Number of rows      325819
Number of columns    10
```

```
-----
Column type frequency:
  factor                2
```

```
-----
Group variables      None
```

```
— Variable type: factor —
  skim_variable n_missing complete_rate ordered n_unique
1 dest          0           1 FALSE         104
2 carrier        0           1 FALSE          16
```

```
top_counts
1 ATL: 16771, ORD: 16507, LAX: 15942, BOS: 14948
2 UA: 57489, B6: 53715, EV: 50868, DL: 47465
```

- 데이터 나누기

```
# 난수의 초기값 : 서로 동일한 난수 생성  
set.seed(222)
```

```
# 데이터 분리하기, 비율은 훈련 데이터가 75%가 되도록  
data_split <- initial_split(flight_data, prop = 3/4)
```

```
# 훈련 데이터와 검정 데이터로 나누어 저장하기  
train_data <- training( data_split )  
test_data  <- testing( data_split )
```

- tidymodels의 3단계

recipe() → prep() → bake()

Defines the
preprocessing

(returns a recipe)

Calculates
statistics from
the training set

(returns a recipe)

Applies the
preprocessing
to data sets

(returns a tibble)

- 요리방법 정의하기 : recipe()

```
flights_rec <- recipe(arr_delay ~ ., data = train_data)
```

반응변수는 arr_delay

설명변수는 나머지 모든 변수

- tidymodels 사용하기
 - 열(변수)의 역할 지정하기 : 행을 고유하게 구별하는 ID 역할(필요한 경우)

```
flights_rec <- recipe(arr_delay ~ ., data = train_data) %>%  
  update_role(flight, time_hour, new_role = "ID")
```

- 확인해 보기

```
summary( flights_rec )
```

```
# A tibble: 10 × 4
```

	variable	type	role	source
	<chr>	<chr>	<chr>	<chr>
1	dep_time	numeric	predictor	original
2	flight	numeric	ID	original
3	origin	nominal	predictor	original
4	dest	nominal	predictor	original
5	air_time	numeric	predictor	original
6	distance	numeric	predictor	original
7	carrier	nominal	predictor	original
8	date	date	predictor	original
9	time_hour	date	ID	original
10	arr_delay	nominal	outcome	original

- tidymodels 사용하기
 - recipe 정의하기 : 영향을 주는 요인(feature) 및 데이터 전처리

```
recipe(arr_delay ~ ., data = train_data) %>%  
  update_role(flight, time_hour, new_role = "ID") %>%  
  step_date(date, features = c("dow", "month")) %>%  
  step_holiday(date, holidays = timeDate::listHolidays("US"),  
               keep_original_cols = FALSE)
```

Data Recipe

Inputs:

	role	#variables
ID		2
outcome		1
predictor		7

Operations:

Date features from date
Holiday features from date

- tidymodels 사용하기
 - recipe 정의하기 : 영향을 주는 요인(feature) 및 데이터 전처리
 - 명목형(factor) 자료의 가변수(더미변수) 만들기

```
recipe(arr_delay ~ ., data = train_data) %>%  
  update_role(flight, time_hour, new_role = "ID") %>%  
  step_date(date, features = c("dow", "month")) %>%  
  step_holiday(date, holidays = timeDate::listHolidays("US"),  
               keep_original_cols = FALSE) %>%  
  step_dummy(all_nominal_predictors())
```

- tidymodels 사용하기
 - recipe 정의하기 : 영향을 주는 요인(feature) 및 데이터 전처리
 - 빈도가 적은 명목형 자료의 경우 훈련 데이터에 없는 경우 발생 : dest 열의 LEX

```
test_data %>%  
  distinct(dest) %>%  
  anti_join(train_data)
```

```
flights_rec <- recipe(arr_delay ~ ., data = train_data) %>%  
  update_role(flight, time_hour, new_role = "ID") %>%  
  step_date(date, features = c("dow", "month")) %>%  
  step_holiday(date,  
               holidays = timeDate::listHolidays("US"),  
               keep_original_cols = FALSE) %>%  
  step_dummy(all_nominal_predictors()) %>%  
  step_zv( all_predictors() )
```


- tidymodels 사용하기
 - 모델 구축 : parsnip 패키지
 - 로지스틱 회귀모델
- ```
lr_mod <- logistic_reg() %>%
 set_engine("glm")
```
- 모델을 recipe에 적용하기 : workflow 패키지의 workflow() 함수 사용
  - 앞서 정의한 recipe로 훈련데이터 처리하기
  - 훈련데이터에 적용하기
  - 검정데이터에 적용하기의 역할 수행

```
flights_wflow <-
 workflow() %>%
 add_model(lr_mod) %>%
 add_recipe(flights_rec)
```

- tidymodels 사용하기
  - 모델을 recipe에 적용하기 : workflow 패키지의 workflow() 함수 사용

```
> flights_wflow
== Workflow ==
Preprocessor: Recipe
Model: logistic_reg()
— Preprocessor —
4 Recipe Steps
• step_date()
• step_holiday()
• step_dummy()
• step_zv()

— Model —
Logistic Regression Model Specification (classification)
Computational engine: glm
```

- tidymodels 사용하기
  - 훈련데이터에 recipe를 적용하고 **모델 훈련(계수 추정)하기** – 시간이 조금 걸립니다 ☺

```
flights_fit <- flights_wflow %>%
 fit(data = train_data)
```

```
flights_fit
```

- 결과를 모아 봅시다 : 계수 추출하기

```
flights_fit %>%
 extract_fit_parsnip() %>%
 tidy()
```

- 결과를 모아 봅시다 : recipe 추출하기

```
flights_fit %>%
 extract_recipe() %>%
 tidy()
```

- tidymodels 사용하기
  - 지금까지의 과정 : 30분 이상 도착하는 비행기 예측하기
    - 모델 수립 : lr\_mod
    - 전처리 recipe 생성 : flights\_rec
    - 모델과 recipe 결합 : flights\_wflow (workflow 이용)
    - 모델 수립(훈련) : fit( ) 이용 flight\_fit
  - 훈련 모델로 부터 예측하기 : 검정데이터 이용

```
predict(flights_fit, test_data)
```

```
A tibble: 81,455 × 1
 .pred_class
 <fct>
1 on_time
2 on_time
3 on_time
```

- tidymodels 사용하기
  - 훈련 모델로 부터 예측하기 : 세부 내용(확률) 파악하기

```
predict(flights_fit, test_data, type="prob")
```

```
A tibble: 81,455 × 2
```

|   | .pred_late | .pred_on_time |
|---|------------|---------------|
|   | <dbl>      | <dbl>         |
| 1 | 0.0547     | 0.945         |
| 2 | 0.0515     | 0.949         |
| 3 | 0.0361     | 0.964         |

```
augment(flights_fit, test_data) %>%
```

```
 select(arr_delay, time_hour, flight, .pred_class, .pred_on_time, .pred_late)
```

```
A tibble: 81,455 × 5
```

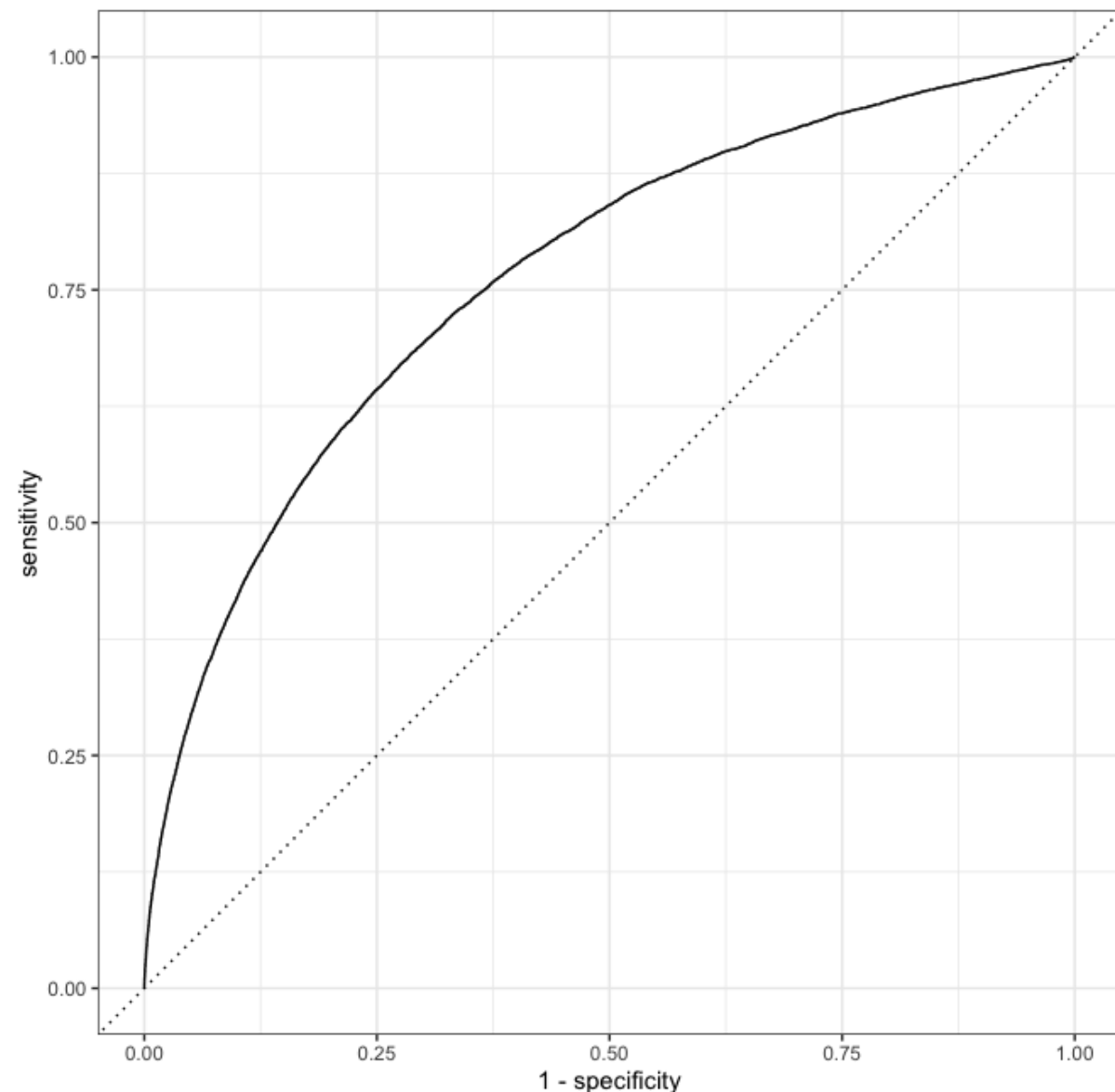
|   | arr_delay | time_hour           | flight | .pred_class | .pred_on_time |
|---|-----------|---------------------|--------|-------------|---------------|
|   | <fct>     | <dtm>               | <int>  | <fct>       | <dbl>         |
| 1 | on_time   | 2013-01-01 05:00:00 | 1545   | on_time     | 0.945         |
| 2 | on_time   | 2013-01-01 05:00:00 | 1714   | on_time     | 0.949         |
| 3 | on_time   | 2013-01-01 06:00:00 | 507    | on_time     | 0.964         |

- tidymodels 사용하기
  - 성능측정 : roc\_curve()

```
flights_aug %>%
 roc_curve(truth = arr_delay, .pred_late) %>%
 autoplot()
```

- Area Under Curve

```
flights_aug %>%
 roc_auc(truth = arr_delay, .pred_late)
```



**수고하셨습니다.**

ps : 궁금하신 것은 언제든지 [yoonani72@gmail.com](mailto:yoonani72@gmail.com) 으로 문의주세요.