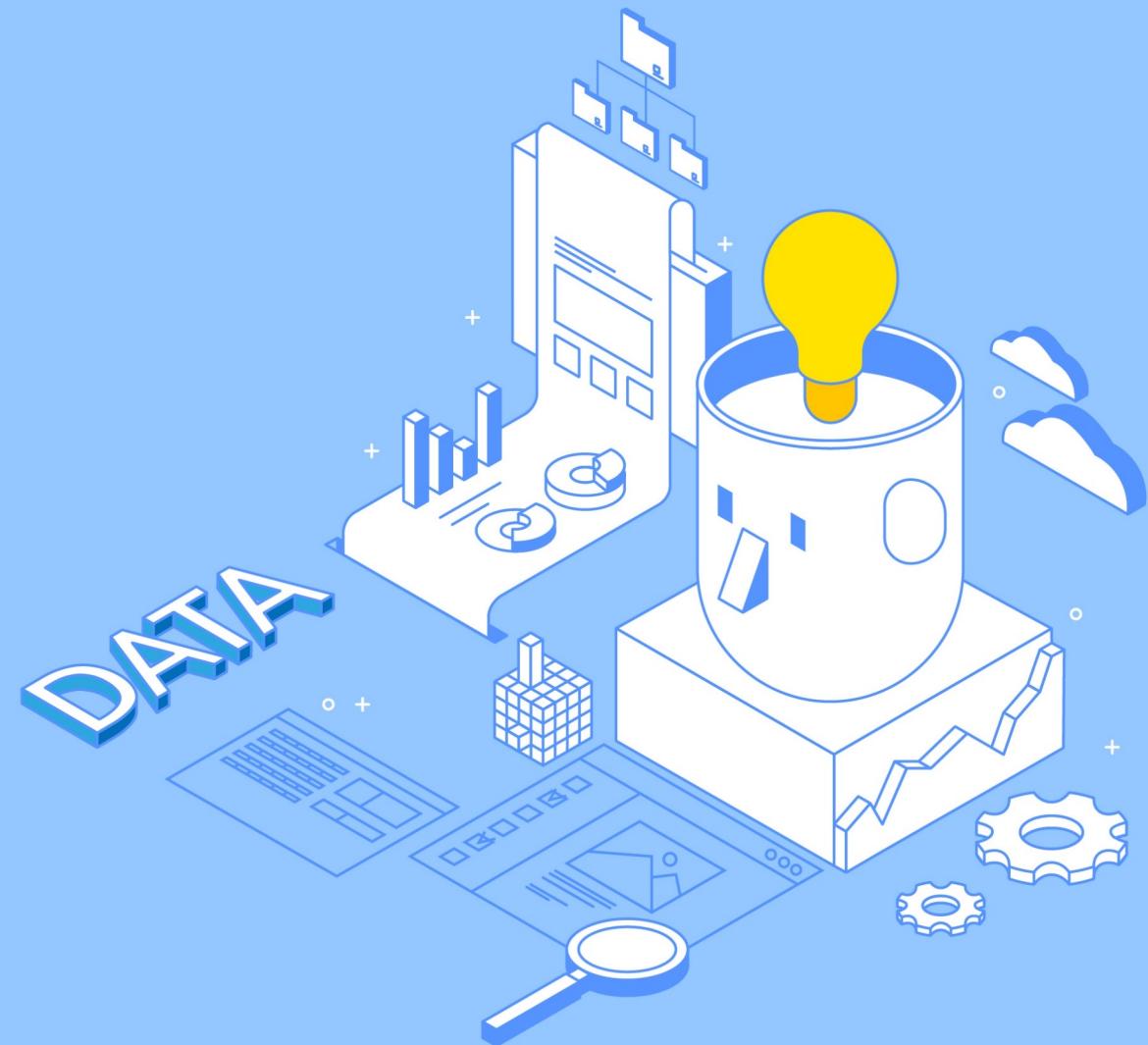




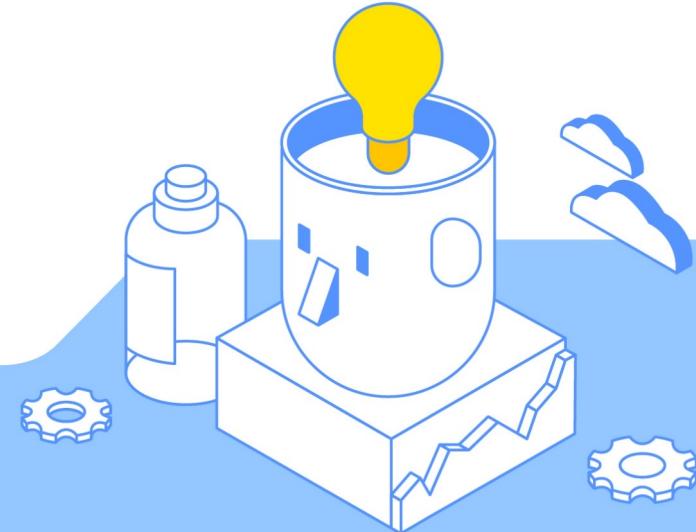
STAT323 mini project

깁스 샘플링을 이용한 국내 코로나 인식 분석

2016150463 이지현



Contents

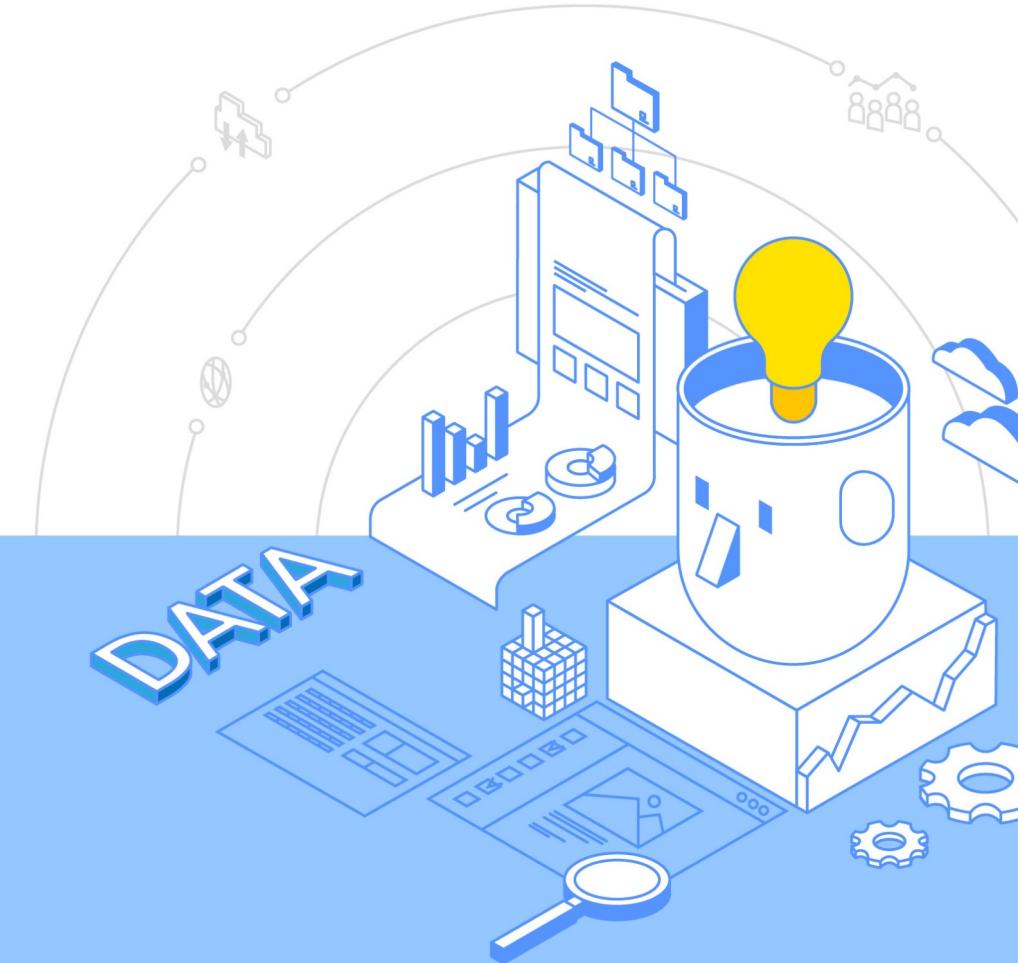


- I 프로젝트 배경 및 목적
- II 데이터 수집 및 전처리
- III 데이터 분석
- IV 결론

Chapter I

프로젝트 배경 및 목적

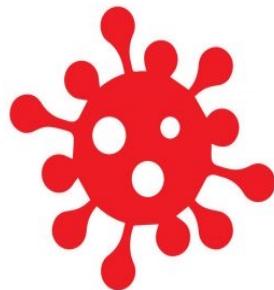
1. 주제 선정 배경
2. 탐구 목적



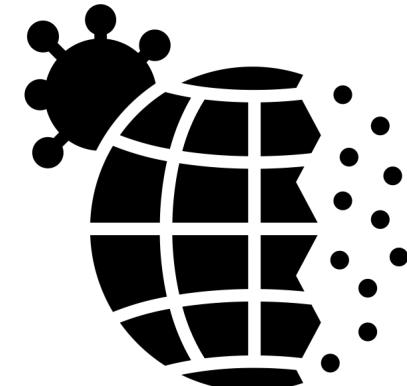
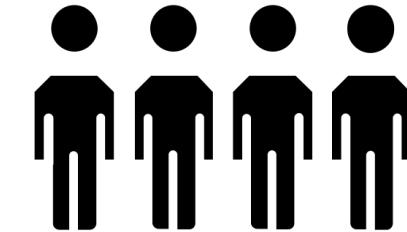
1. 프로젝트 배경 및 목적



① (1) 주제 선정 배경



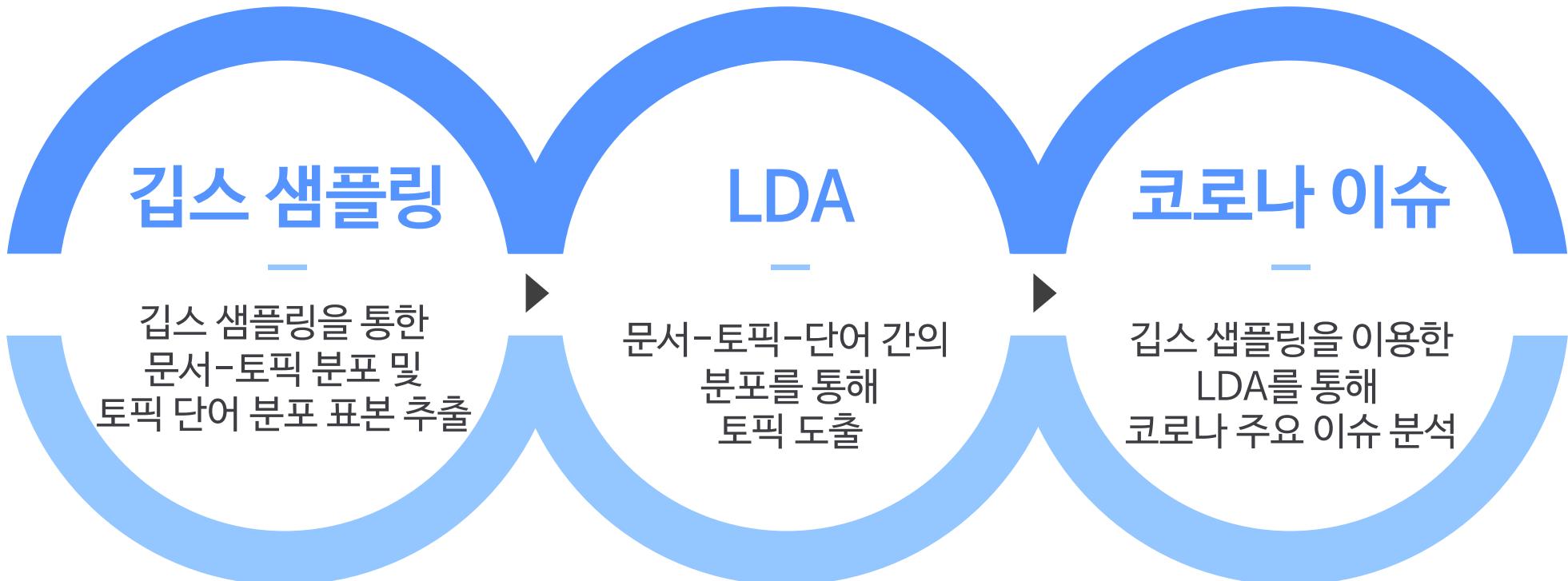
COVID-19
Coronavirus



1. 프로젝트 배경 및 목적



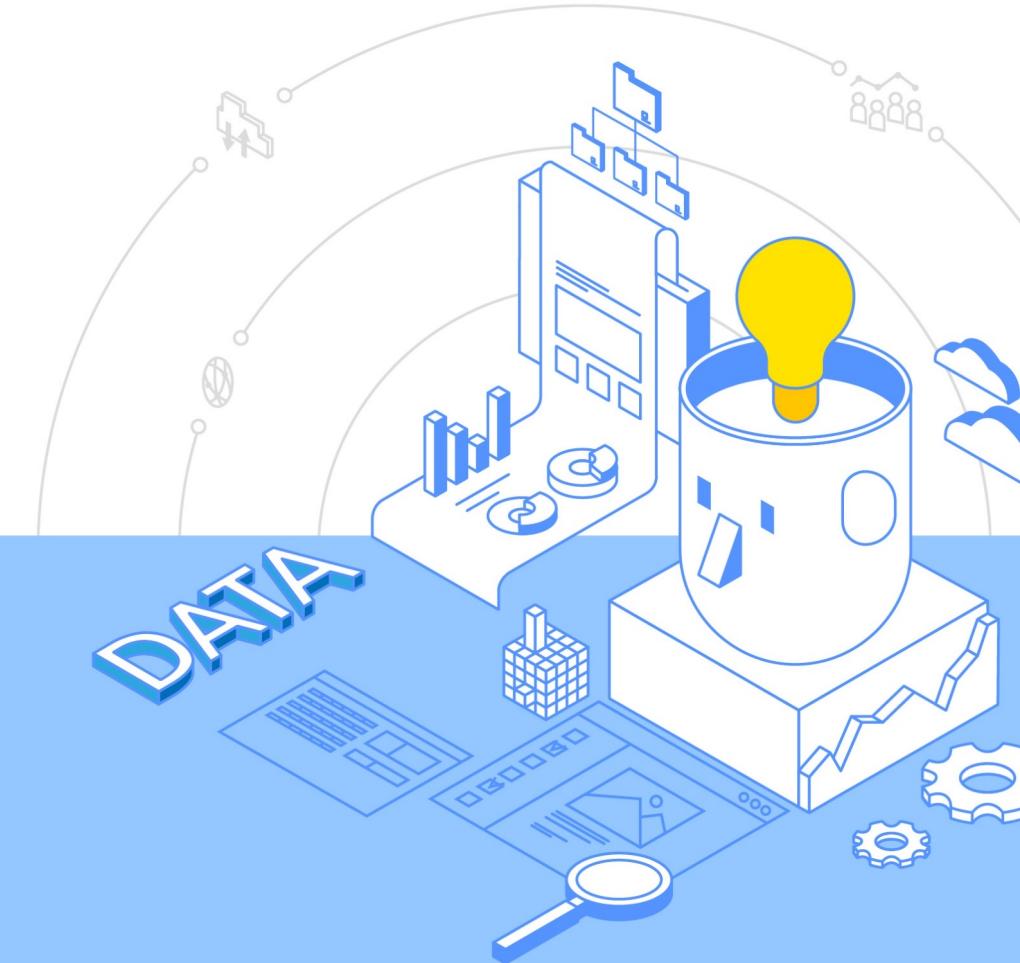
① (2) 탐구목적



Chapter II

데이터 수집 및 전처리

1. 데이터 선정
2. 데이터 수집
3. 데이터 전처리



2. 데이터 수집 및 전처리



① (1) 데이터 선정



- 2021년 1월 1일 ~ 6월 25일 3사 코로나 19 관련 뉴스 수집

2. 데이터 수집 및 전처리



(2) 데이터 수집

검색필터 초기화

기간 +

2021 (197,554)

언론사 +

경향신문 (3,231)

국민일보 (5,469)

내일신문 (1,540)

동아일보 (4,007)

문화일보 (2,323)

펼쳐보기 ▾

통합분류 +

정치(19,773)

경제(53,060)

뉴스 인용문 사설 최신순 10건씩 보기 결과 내 재 검색

분석제외 취소 (0) 검색식 저장 뉴스분석 리포트 생성 < < 1 / 2000 >

"코로나19"
뉴스 검색 결과 197,554 건입니다. (2021-03-27 ~ 2021-06-27 기준)

"Science First" '과학기술강국포럼 29일 창립
과학기술강국포럼' 창립식. 웹포스터. 국민의힘 김영식 국회의원 제공...
부산일보 정치>국회_정당 | IT_과학>IT_과학일반 | 정치>행정_자치 2021/06/27 송현수

출근 재개 앞두고...美 기업들 "백신 접종 의무화"
[아시아경제 김지희 기자] 미국 기업과 지방정부들이 코로나19 백신 접종을 의무화하고 나섰다....
아시아경제 국제>미국_북미 | 경제>국제경제 | 사회>의료_건강 2021/06/27 김지희

"감염경로 불명" 4명 중 1명...학교·학원 집단감염 일파만파
2주간 감염경로 불명자 25.3% 달해...
서울신문 지역>부산 | 지역>강원 | 사회>의료_건강 2021/06/27 강주리

- 한국언론진흥재단 '빅카인즈' 사이트 통해 뉴스 URL 데이터 수집

2. 데이터 수집 및 전처리



(2) 데이터 수집

	뉴스 식별자	일자	언론사	기고자	제목	통합 분류1	통합 분류2	통합 분류3	사건/사고 분류1	사
1	08100101.20210627121101001	20210627	KBS	박민철	확진자 닷새째 600명 대...지역별 거리두기 단계 오후 발표	사회>의료_건강				
2	08100101.20210627121100001	20210627	KBS	김용준	코로나19 신규 확진 614명..."주말 감소 양상 없어"	사회>의료_건강	사회>미디어			
3	08100201.20210627120316001	20210627	MBC	박진주	내일부터 30세 미만 사회필수인력 등 미접종자 사전예약	사회>사회일반	지역>경기	지역>울산		
4	08100201.20210627120315001	20210627	MBC	박진주	닷새 연속 600명대..."학원 등 집단감염 속출"	사회>의료_건강	사회>교육_시험	지역>울산		
5	08100201.20210627073222001	20210627	MBC	신정연	도쿄 올림픽 멜타 변이 '확산' 기폭제 되나?	국제>일본	사회>의료_건강	지역>울산		
6	08100201.20210627073221002	20210627	MBC	박윤수	신규 확진 600명 안팎...개편 거리두기 발표	사회>의료_건강	사회>의료_건강	지역>울산		
7	08100101.20210627061038002	20210627	KBS	박예원	1500만명 1차 접종 초과 달성하고 상반기 접종 마무리	사회>의료_건강	사회>여성	지역>울산		
8	08100101.20210627061038001	20210627	KBS	한승연	집단감염 속 새로운 거리두기 '지역별 단계' 오늘 발표	사회>교육_시험	사회>여성	지역>울산		
9	08100101.20210626234349002	20210626	KBS	이정	중학교 등 코로나19 9명 신규 확진...누적 2,799명	지역>울산	지역>경남	지역>강원		
10	08100101.20210626234349001	20210626	KBS	이정	동행세일 첫 주말 상권 '활기'...업종따라 '온도차'	경제>서비스_쇼핑	경제>유통	지역>울산		
11	08100101.20210626215443002	20210626	KBS	김가람	제주, 백신 접종 80대 사망신고 접수..."역학조사 중"	지역>제주	사회>사회일반	지역>경기		
12	08100101.20210626215443001	20210626	KBS	김가람	제주, 이틀 동안 확진자 11명..."백신 접종완료 8.9%"	지역>제주	사회>의료_건강	지역>경남		
13	08100101.20210626215128001	20210626	KBS	김문영	올해 춘천레저대회 개막식 취소...대회 분산 개최	지역>강원	지역>경남	지역>부산		
14	08100101.20210626214829002	20210626	KBS	천현수	창원경상대병원, 이동형 에크모 국내 2번째 도입	지역>강원	사회>의료_건강	지역>울산		
15	08100101.20210626214812001	20210626	KBS	송현준	경남 코로나19 신규 확진 10명...다음달 거리두기 방안 발표	사회>여성	국제>아시아	지역>울산		
16	08100101.20210626214659002	20210626	KBS	천춘환	충북 귀농귀촌 증가세 3.1%↑	지역>충북	지역>대전	지역>강원		
17	08100101.20210626214500001	20210626	KBS	박준형	지역 소비자 심리지수 6달 연속 상승	경제>외환	지역>대전	지역>강원		
18	08100101.20210626214459001	20210626	KBS	이종영	경북도, 9월까지 도민 70% 1차 접종 목표	지역>경남	지역>충남	지역>경기		
19	08100101.20210626214350001	20210626	KBS	성용희	대전·세종·충남 코로나19 확진자 37명 추가	지역>대전	지역>강원	지역>부산		
20	08100101.20210626214349001	20210626	KBS	유진환	지난해 충남 인구 자연감소 4천 명...농업의 위기	지역>충남	지역>전북	지역>부산		



- 한국언론진흥재단 '빅카인즈' 사이트 통해 뉴스 URL 데이터 수집

엑셀 다운로드

2. 데이터 수집 및 전처리



② (2) 데이터 수집

일자	언론사	기고자	본문	URL
20210627	KBS	박민철	[앵커] 다음 달 새 거리두기 시행을 앞두고, 각 지역별로 적용될	https://news.kbs.co.kr/news/view.do?ncd=5219291&ref=DA
20210627	KBS	김용준	[앵커] 코로나19 신규 확진자 닷새째 600명대를 나타냈습니다.	https://news.kbs.co.kr/news/view.do?ncd=5219283&ref=DA
20210627	MBC	박진주	상반기 백신 접종 대상자 가운데 아직 접종을 받지 못한 11만명	https://imnews.imbc.com/replay/2021/nw1200/article/6281880_34908.htm
20210627	MBC	박진주	◀ 앵커 ▶ 주말이라 코로나19 검사건수가 줄었는데도 불구하고,	https://imnews.imbc.com/replay/2021/nw1200/article/6281879_34908.htm
20210627	MBC	신정연	◀ 앵커 ▶ 전파력이 가장 강한 것으로 알려진 코로나19 델타 변	https://imnews.imbc.com/replay/2021/nwtoday/article/6281844_34943.htm
20210627	MBC	박윤수	◀ 앵커 ▶ 오늘 0시 기준 코로나19 신규 환자 수는 600명 안팎이	https://imnews.imbc.com/replay/2021/nwtoday/article/6281843_34943.htm
20210627	KBS	박예원	[앵커] 코로나19 예방 접종을 한 차례 이상 받은 사람이 1500만	https://news.kbs.co.kr/news/view.do?ncd=5219215&ref=DA
20210627	KBS	한승연	[앵커] 코로나19 신규확진자 668명, 나흘째 600명대를 기록했습니다	https://news.kbs.co.kr/news/view.do?ncd=5219214&ref=DA
20210626	KBS	이정	[KBS 울산]오늘 울산에서는 9명의 코로나19 신규 확진자가 발생	https://news.kbs.co.kr/news/view.do?ncd=5219201&ref=DA
20210626	KBS	이정	[KBS 울산][앵커]코로나19 여파로 침체된 내수경기를 살리기 위함	https://news.kbs.co.kr/news/view.do?ncd=5219200&ref=DA
20210626	KBS	김가람	[KBS 제주]제주에서 코로나19 백신을 맞은 80대가 사망했다는 신	https://news.kbs.co.kr/news/view.do?ncd=5219185&ref=DA
20210626	KBS	김가람	[KBS 제주]제주 지역 코로나 상황 전해드립니다.어제 제주에서	https://news.kbs.co.kr/news/view.do?ncd=5219184&ref=DA
20210626	KBS	김문영	[KBS 춘천]다음 달 열릴 예정이던 춘천레저대회 개막축전이 취소	https://news.kbs.co.kr/news/view.do?ncd=5219178&ref=DA
20210626	KBS	천현수	[KBS 창원]창원경상대병원이 국내 2번째로 이동형 에크모 장비를	https://news.kbs.co.kr/news/view.do?ncd=5219182&ref=DA
20210626	KBS	송현준	[KBS 창원] [앵커] 경남에서는 10명이 코로나19 신규 확진 판정을	https://news.kbs.co.kr/news/view.do?ncd=5219130&ref=DA
20210626	KBS	천춘환	[KBS 청주]충북 지역의 귀농귀촌 가구가 늘었습니다.충청북도는	https://news.kbs.co.kr/news/view.do?ncd=5219160&ref=DA
20210626	KBS	박준형	[KBS 대구]대구경북의 소비자 심리지수가 두 달 연속 기준치 100	https://news.kbs.co.kr/news/view.do?ncd=5219158&ref=DA
20210626	KBS	이종영	[KBS 대구]경상북도는 오는 9월 말까지 전 도민의 70%인 185만	https://news.kbs.co.kr/news/view.do?ncd=5219153&ref=DA
20210626	KBS	성용희	[KBS 대전]대전과 세종, 충남에서 코로나19 확진자 37명이 더 나	https://news.kbs.co.kr/news/view.do?ncd=5219149&ref=DA

- 빅데이터는 뉴스 본문 데이터가 제공되지 않는 한계

2. 데이터 수집 및 전처리



→ (2) 데이터 수집

```
def KBS(url):
    request = requests.get(url)
    bs = BeautifulSoup(request.text, "html.parser")
    article = bs.find('div', {'id':'cont_newstext'}).text.strip()
    return(article)

def MBC(url):
    driver.get(url)
    article = driver.find_element_by_css_selector(
        '#content > div > section.wrap_article > article > div.news_cont > div.news_txt').text.strip()
    return(article)

def SBS(url):
    request = requests.get(url)
    bs = BeautifulSoup(request.text, "html.parser")
    div = bs.find('div', class_='text_area').text.strip()
    return(div)
```

- 파이썬 bs4, selenium 패키지 통해 각 언론사 뉴스 데이터 수집

2. 데이터 수집 및 전처리



—. (3) 데이터 전처리

```
from ckonlp.tag import Twitter
from ckonlp.tag import Postprocessor
from ckonlp.utils import load_wordset

#dictionary에 단어 추가
twitter.add_dictionary('확진자', 'Noun')
twitter.add_dictionary('촬영기자', 'Noun')
twitter.add_dictionary('지원금', 'Noun')
twitter.add_dictionary('키워드', 'Noun')
twitter.add_dictionary('접종수', 'Noun')

#불용어 리스트 불러오기
stopwords = load_wordset('stopwords.txt')

postprocessor = Postprocessor(
    base_tagger = twitter,
    stopwords = stopwords, # 불용어 제거
    passtags = 'Noun', # 명사만 선택
)

documents = []
for txt in text:
    txt = re.sub(r'[0-9]+', '', txt) #숫자 제거
    txt = re.sub('[-=+,#/?:^$.@*\"※~&%·!』＼＼'|\(\)\[\]\<\>`＼＼...]', '', txt) #특수문자 제거
    words = postprocessor.pos(txt)
    new = ""
    for word in words:
        if len(word[0]) > 1: #두글자 이상 단어 선택
            new += " " + word[0]
    documents.append(new)
```

- 불용어 및 특수문자, 1글자 단어 제거
- 분석에 필요한 단어 사전에 재정의 및 명사만 추출

2. 데이터 수집 및 전처리



(3) 데이터 전처리

content

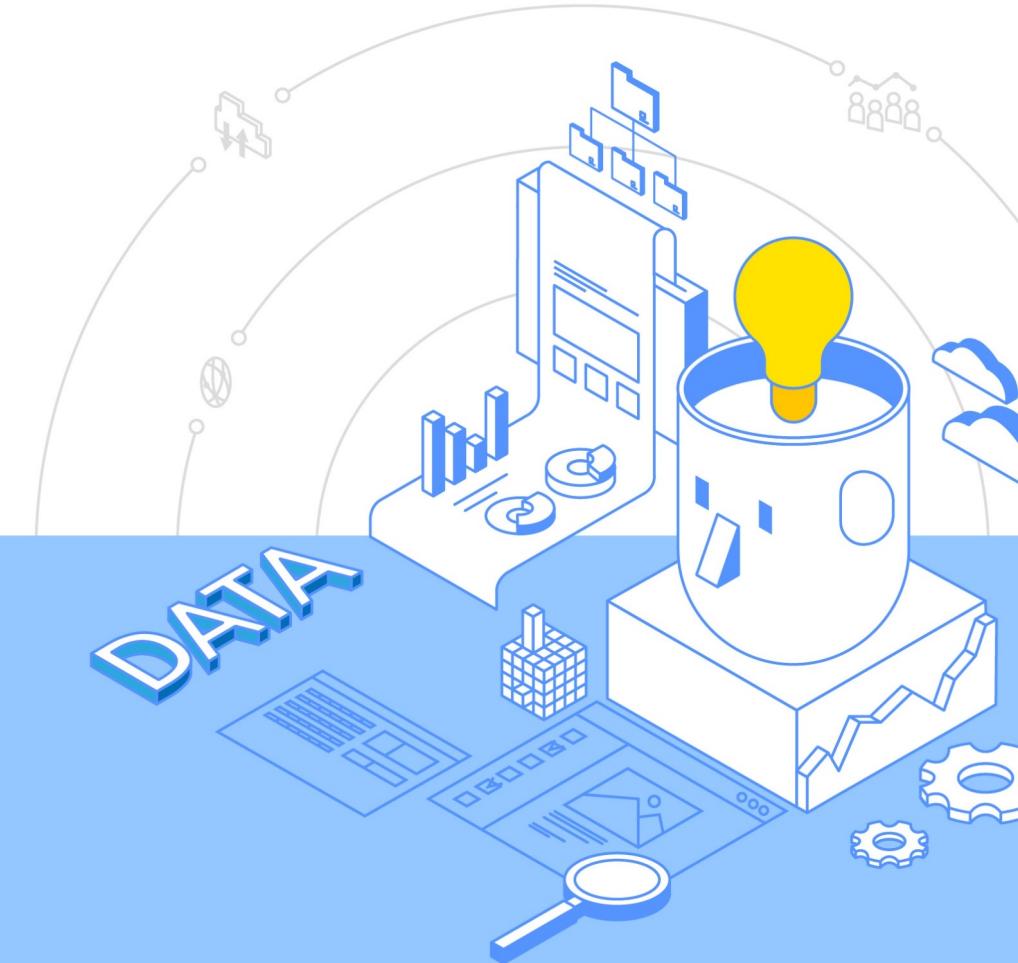
오늘 강원도 코로나 신규 확진자 별로 원주 춘천 평창 동해 속초 삼척 강릉 태백 확진자 대부분 기준 확진자 접촉 감염 평창 원주 춘천 동해 확진자 감염 경로 확인
강원도 곳곳 피기 시작 다예 봄꽃 축제 상춘객 시기 올해 코로나 여파 축제 개최 취재 리포트 삼척 맹방 유채꽃 다예 이맘 무리 장관 올해 성합 맹방 유채꽃 축제 지난해 올해 취
울산 학급 평균 학생 수가 학교 조사 다예 전보 코로나 시대 이하 여야 목소리 지고 보도 박영하 리포트 대학입시 농어촌 특별 전형 인기 울주군 범서 고등학교 학급 평균 학생 을
국내 코로나 신규 확진자 사백 스물 여덟 하루 사백 명대 백신 돌파 접종 기록 가운데 다음 부터 일흔 다섯 이상 고령 접종 사용 화이자 백신 만회 분량 도착 화이자 백신 접종 이
태연 강원도 의원 오늘 도의 도정 질문 도내 코로나 방역 사용 공간 방역 가운데 암모늄 화합물 공통 발견 인체 영향 대해 조사 주장 대해 강원도 방역 제의 정확 성분 유해 여부
부산 공동 어시장 오늘 부터 어획 바닥 분류 판하 기준 경매 진행 다지 코로나 확진자 발생 보름 부산 공동 어시장 확진자 작업 별도 관련 천백 대한 전수 조사 모두 음성 판정
오늘 부산 코로나 신규 확진자 서구 냉장 사업 직원 부산시 보건 당국 직원 대해 전수 조사 감염 경로 확인 집단 감염 발생 해운대구 향목 교회 추가 확진자 다다 일부 이상 노인
전교조 전북 지부 학교 학급 학생 최대 제한 하자 요구 학급 학생 축하 내용 교육기본법 중등교육 개정안 아직 법안 사소 위원회 통과 국회 제화 촉구 학급 학생 이면 코로나 시대
공약 검증 오늘 키워드 재난 지원금 정부 재난 지원금 지급 준비 일부 지자체 다양 형태 긴급 지원금 지급 부산시 후보 재난 지원금 대해 입장 인지 정리 리포트 정부 차례 코로나
한편 오늘 프로야구 정규 시즌 코로나 통합 매뉴얼 발표 확진 선수 더라도 곧바로 리그 중단 한국 야구 위원회 코로나 확진 선수 발생 해당 선수 밀접 접촉 선수 제외 대체 선수
전국 시도 교육감 의회 교원 업무 부담 사기 저하 이유 올해 교원능력개발평가 시행 유예 요구 교육감 의회 장문 교원 평가 부담 코로나 극복 교육활동 전념도록 교원능력개발
방송 영화 만화 게임 영상 문화 기업 한자리 모이 국내 최대 영상 문화 산업 단지 사업 마침내 경기도 부천 협약 궤도 보도 박재우 리포트 가상현실 웹툰 게임 버스 트럭 이용 아
제주도 오늘 부터 요양 병원 요양시 이상 소자 사자 대상 분기 코로나 예방접종 전체 분기 접종 대상자 특수교육 사자 유치원 어린이집 초등학교 학년 교사 보건 교사 경찰 소방
제주도 오늘 오후 기준 코로나 신규 확진자 오지 제주 사흘 확진자 발생 누적 확진자 유지 다도 격리 치료 오늘 퇴원 현재 읍읍 병상 입원 중인 확진자 확진자 접촉 비롯 해외 입
코로나 여파 차례 연기 광주 비엔날레 다음 마음 이하 영혼 주제 이번 비엔날레 개국 참여 작품 올해 코로나 상황 감안 공식 누리집 유튜브 작품 감상 온라인 관람 시스템 구축 ?
오늘 광주 코로나 신규 확진자 추가 광주시 전남 확진자 동전 노래방 이용자 관련 확진자 전북 확진자 접촉 명과 해외 유입 확진자 추가 전남 확진자 관련 광주 지역 누적 확진자

- 최종분석 데이터

Chapter III

데이터 분석

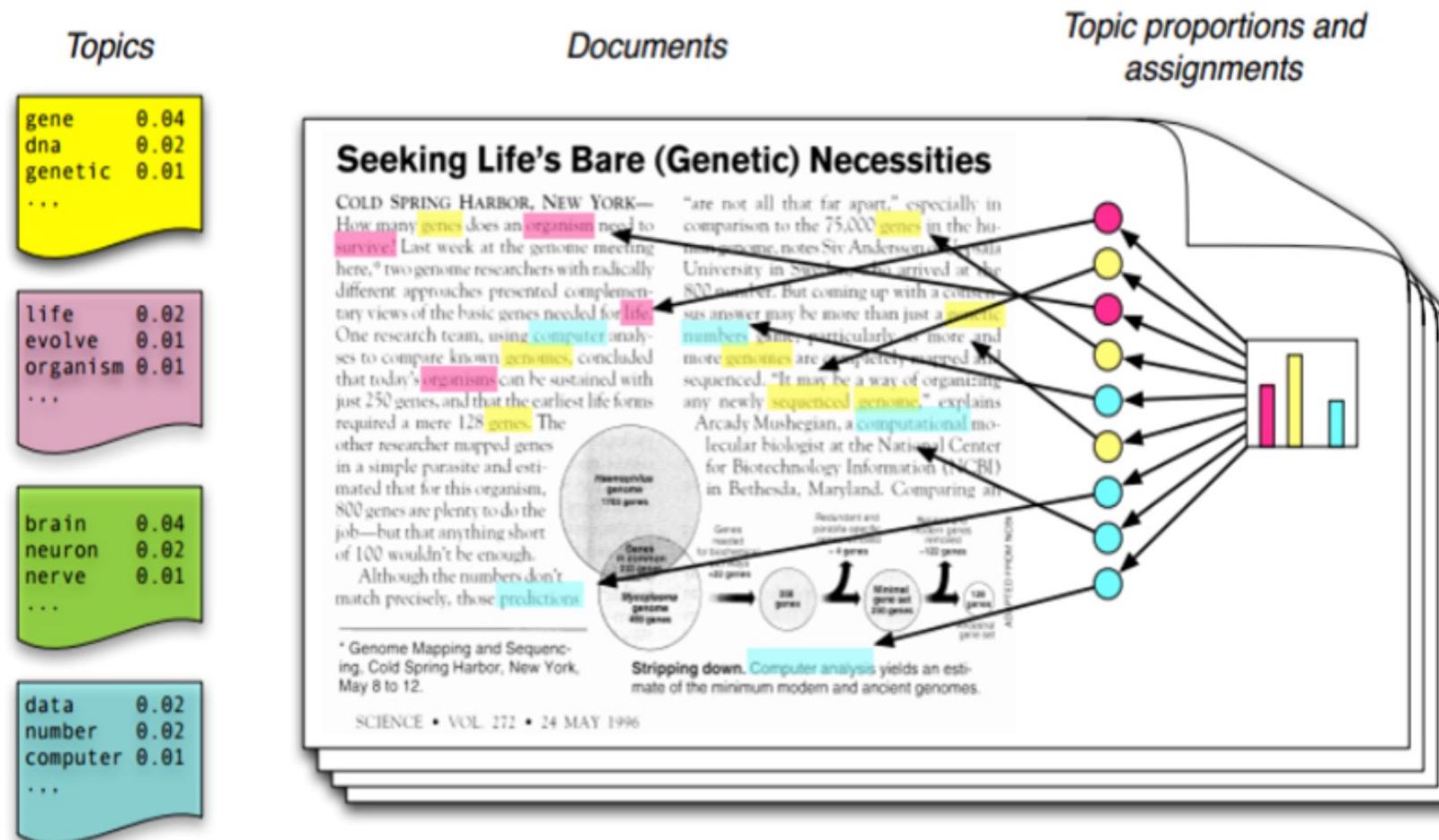
1. 김스샘플링을 이용한 LDA 구현
2. 패키지를 이용한 분기별 주요 뉴스 토픽



3. 데이터 분석



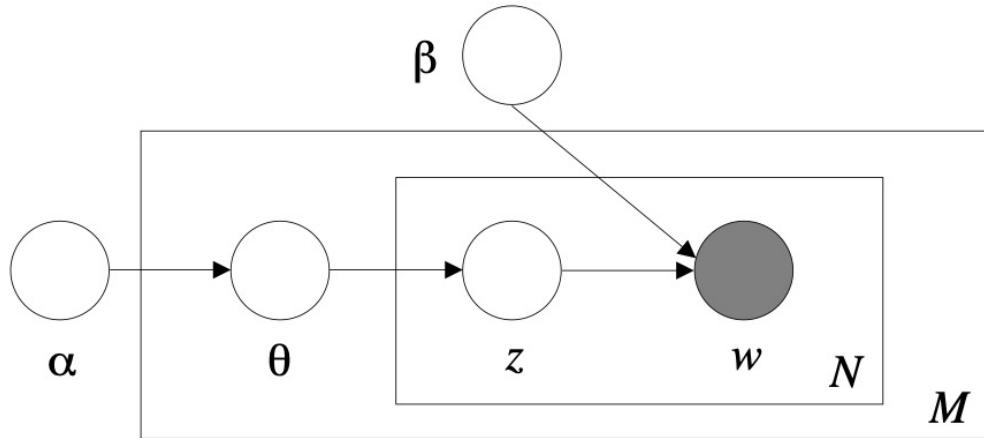
(1) 김스샘플링을 이용한 LDA 구현



3. 데이터 분석



• (1) 갑스샘플링을 이용한 LDA 구현



$$\begin{aligned}
 p(z_i | \mathbf{z}^{(-i)}, \mathbf{w}) &= \frac{p(\mathbf{w}, \mathbf{z})}{p(\mathbf{w}, \mathbf{z}^{(-i)})} = \frac{p(\mathbf{z})}{p(\mathbf{z}^{(-i)})} \cdot \frac{p(\mathbf{w} | \mathbf{z})}{p(\mathbf{w}^{(-i)} | \mathbf{z}^{(-i)}) p(w_i)} \\
 &\propto \prod_d \frac{B(n_{d,\cdot} + \alpha)}{B(n_{d,\cdot}^{(-i)} + \alpha)} \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(n_{k,\cdot}^{(-i)} + \beta)} \\
 &\propto \frac{\Gamma(n_{d,k} + \alpha_k) \Gamma(\sum_{k=1}^K n_{d,k}^{(-i)} + \alpha_k)}{\Gamma(n_{d,k}^{(-i)} + \alpha_k) \Gamma(\sum_{k=1}^K n_{d,k} + \alpha_k)} \cdot \frac{\Gamma(n_{k,w} + \beta_w) \Gamma(\sum_{w=1}^W n_{k,w}^{(-i)} + \beta_w)}{\Gamma(n_{k,w}^{(-i)} + \beta_w) \Gamma(\sum_{w=1}^W n_{k,w} + \beta_w)} \\
 &\propto (n_{d,k}^{(-i)} + \alpha_k) \frac{n_{k,w}^{(-i)} + \beta_w}{\sum_{w'} n_{k,w'}^{(-i)} + \beta_{w'}} \tag{11}
 \end{aligned}$$

3. 데이터 분석



• (1) 갑스샘플링을 이용한 LDA 구현

Input: words $w \in \text{documents } d$

Output: topic assignments z and counts $n_{d,k}$, $n_{k,w}$, and n_k

begin

 randomly initialize z and increment counters

foreach iteration **do**

for $i = 0 \rightarrow N - 1$ **do**

$word \leftarrow w[i]$

$topic \leftarrow z[i]$

$n_{d,topic} = 1$; $n_{word,topic} = 1$; $n_{topic} = 1$

for $k = 0 \rightarrow K - 1$ **do**

$p(z = k | \cdot) = (n_{d,k} + \alpha_k) \frac{n_{k,w} + \beta_w}{n_k + \beta \times W}$

end

$topic \leftarrow \text{sample from } p(z | \cdot)$

$z[i] \leftarrow topic$

$n_{d,topic} += 1$; $n_{word,topic} += 1$; $n_{topic} += 1$

end

end

return $z, n_{d,k}, n_{k,w}, n_k$

end

Algorithm 1: LDA Gibbs Sampling

3. 데이터 분석



(1) 갑스샘플링을 이용한 LDA 구현

```
#Procedure
#Gibbs sampling
for (iter in 1:iter.max){
  for (doc in 1:length(docs)){
    for (word in 1:length(docs[[doc]])){
      z_old = assign_topic[[doc]][word]
      index = docs[[doc]][word]

      # 두 분포에서 z 관련 정보를 제거
      document_topic[doc, z_old] = document_topic[doc, z_old] - 1
      topic_word[z_old, index] = topic_word[z_old, index] - 1
      prob_topic = rep(0, K)
      for (k in 1:K){
        #P(Z|W) 특정 단어가 특정 주제에 속할 확률 값
        prob_topic[k] = (document_topic[doc, k] + alpha)*
          (topic_word[k, w.index]+beta)/rowSums(topic_word+beta)[k]
      }

      # 구한 확률을 이용해 새로운 주제(topic) 할당
      z_new = sample(1:K, 1, prob = prob_topic)
      if(z_new != z_old) assign_topic[[doc]][word] = z_new

      topic_word = assign_word_topic(assign_topic, docs, topic_word)
      document_topic = make_document_topic(assign_topic, K, document_topic)
    }
  }
}
```



	Jan1	Jun1
1	코로나	코로나
2	새해	접종
3	확진자	백신
4	백신	확진자
5	뉴스	학생
6	감염	예약
7	지난해	오늘
8	확진	뉴스
9	관련	부터
10	지역	방역

3. 데이터 분석



(2) 패키지를 이용한 분기별 주요 뉴스 토픽

1분기(1월-3월)

"대통령"	"접종"	"확진자"	"지원금"	"답변"
"미국"	"백신"	"감염"	"지원"	"배달"
"바이든"	"아스"	"확진"	"지급"	"어요"
"중국"	"병원"	"검사"	"사업"	"시장"
"수업"	"의료"	"집단"	"공인"	"고요"
"일본"	"화이자"	"방역"	"소상"	"가격"
"학교"	"요양"	"판정"	"후보"	"온라인"
"학생"	"예방접종"	"신규"	"민주당"	"아이"
"북한"	"환자"	"접촉"	"재난"	"판매"
"등교"	"바이러스"	"당국"	"일자리"	"축제"

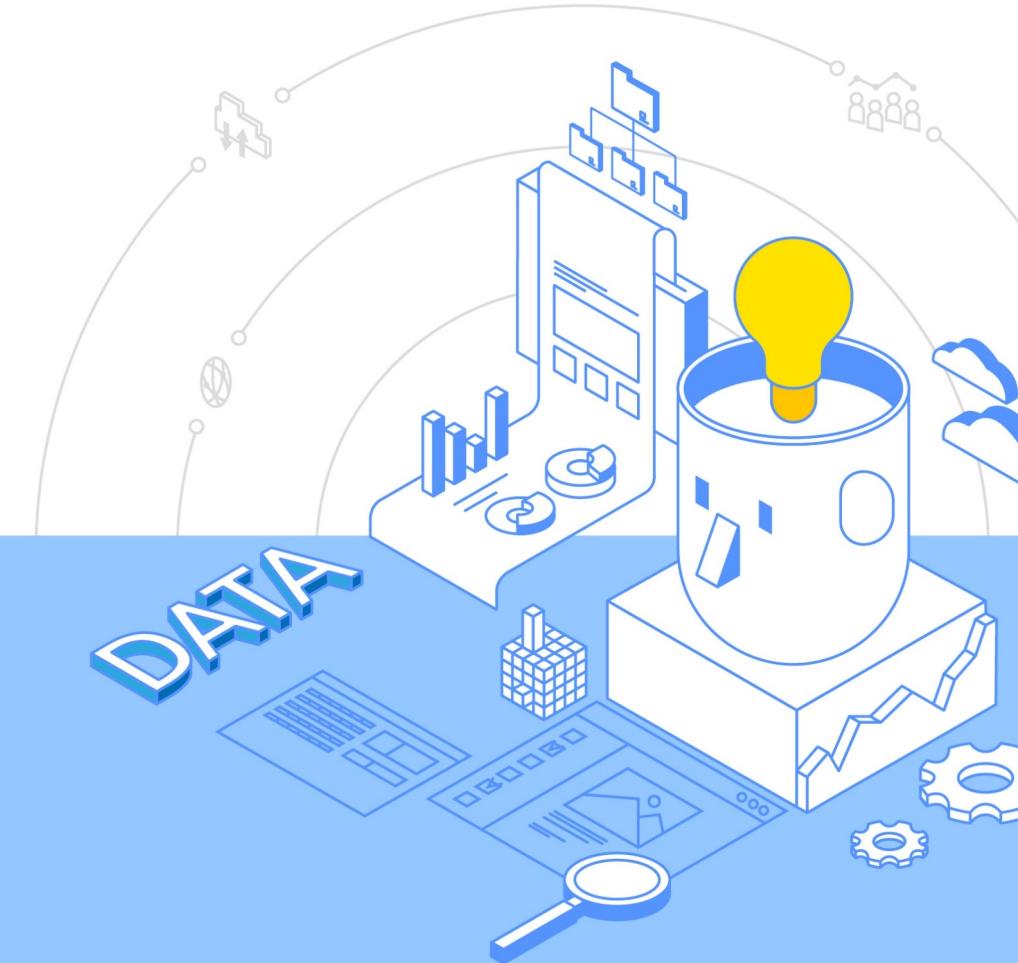
2분기(4월-6월)

"확진자"	"어요"	"지원"	"접종"	"대통령"
"감염"	"답변"	"지역"	"백신"	"미국"
"확진"	"경찰"	"사업"	"아스"	"후보"
"검사"	"가격"	"학생"	"화이자"	"민주당"
"방역"	"고요"	"학교"	"예약"	"바이든"
"신규"	"소비자"	"제주"	"이상"	"국민"
"판정"	"시장"	"수업"	"정부"	"일본"
"발생"	"거래"	"교육"	"예방접종"	"중국"
"접촉"	"배달"	"지급"	"바이러스"	"북한"
"집단"	"래서"	"관광"	"확진자"	"총리"

- R의 lda 패키지의 lda.collapsed.gibbs.sampler 함수 사용

Chapter IV

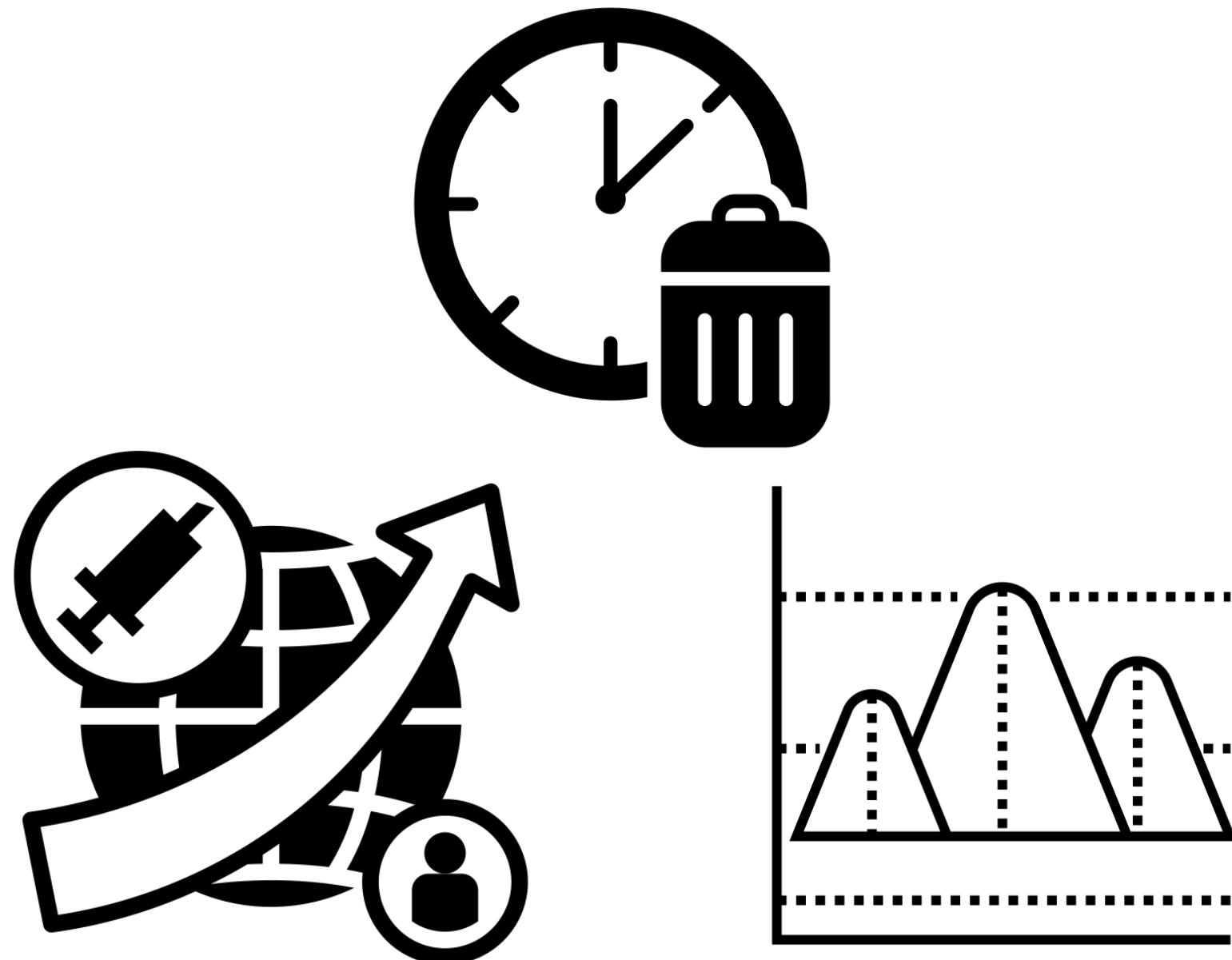
결론



4. 결론



66





감사합니다.

