데이터 분석 인턴 활동 정리

안동수

목차

- 제가 일한 곳: 경제인문사회연구회
- 인턴에서 경험한 것들
    - 교육
    - 업무 및 프로젝트

경제인문사회연구회는 무엇을 하는 곳인가?

경제인문사회연구회는 경제・인문사회분야의 정부출연연구기관을 지원・육성하고, 체계적으로 관리함으로써 국가의 연구사업정책의 지원 및 지식산업 발전에 이바지하기 위하여 설립된 정부출연연구기관으로 국무총리(국무조정실) 산하의 기타공공기관이다.

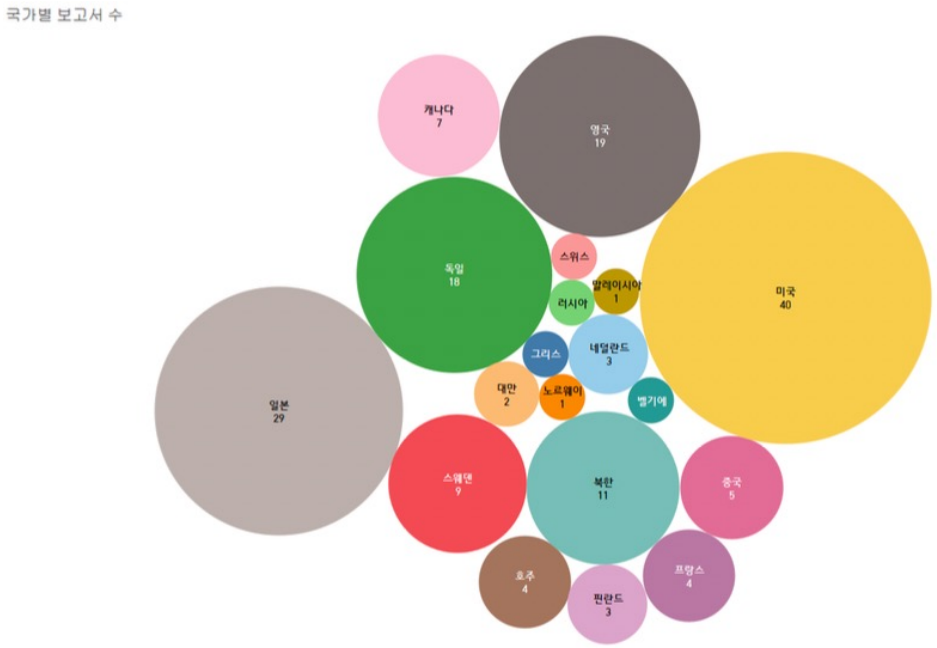
-> 즉, 정책 연구 및 제안하는 곳

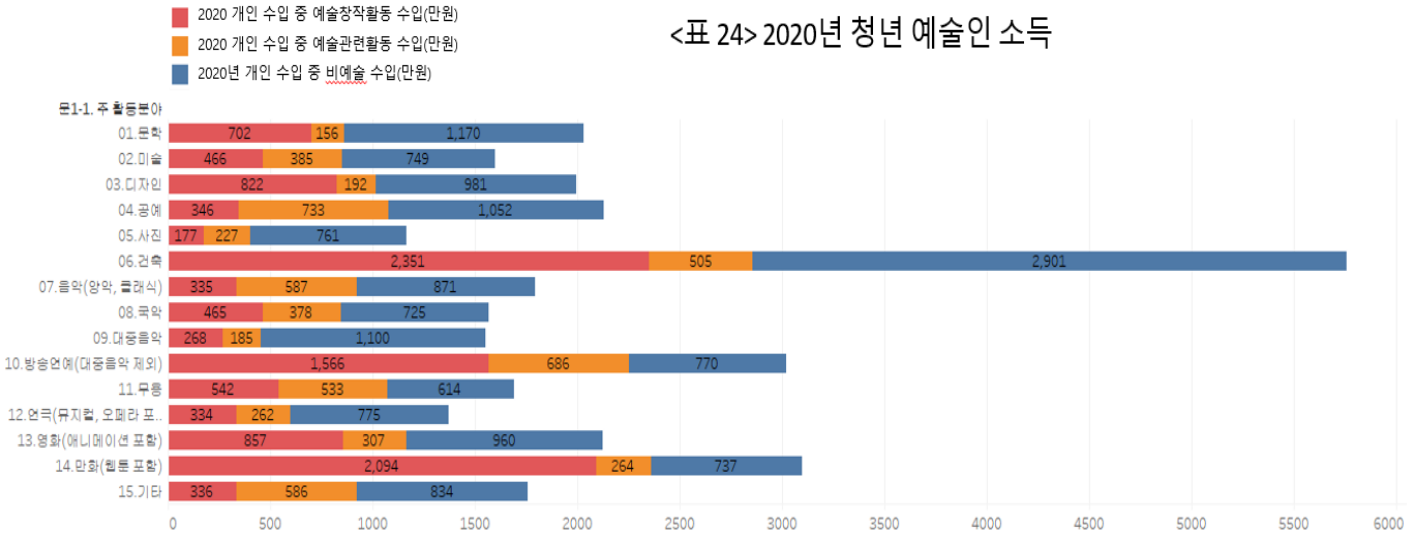
인턴 활동을 하면서 받은 교육

- 텍스트 데이터 교육
  - 텍스트 데이터 수집하는 방법(웹 데이터 크롤링 등)
  - 정규표현식 *정규 표현식
  - 텍스트 데이터 정제 및 전처리 방법
  - 텍스트 데이터 시각화(워드 클라우드, 트리맵 등)
  - 텍스트 데이터 모델링 (텍스트 데이터 감정분석 등)
- ‘태블로’라는 시각화 툴
- Qgis 지리 데이터 다루는 프로그램 등등

업무 : 논문 자료 시각화

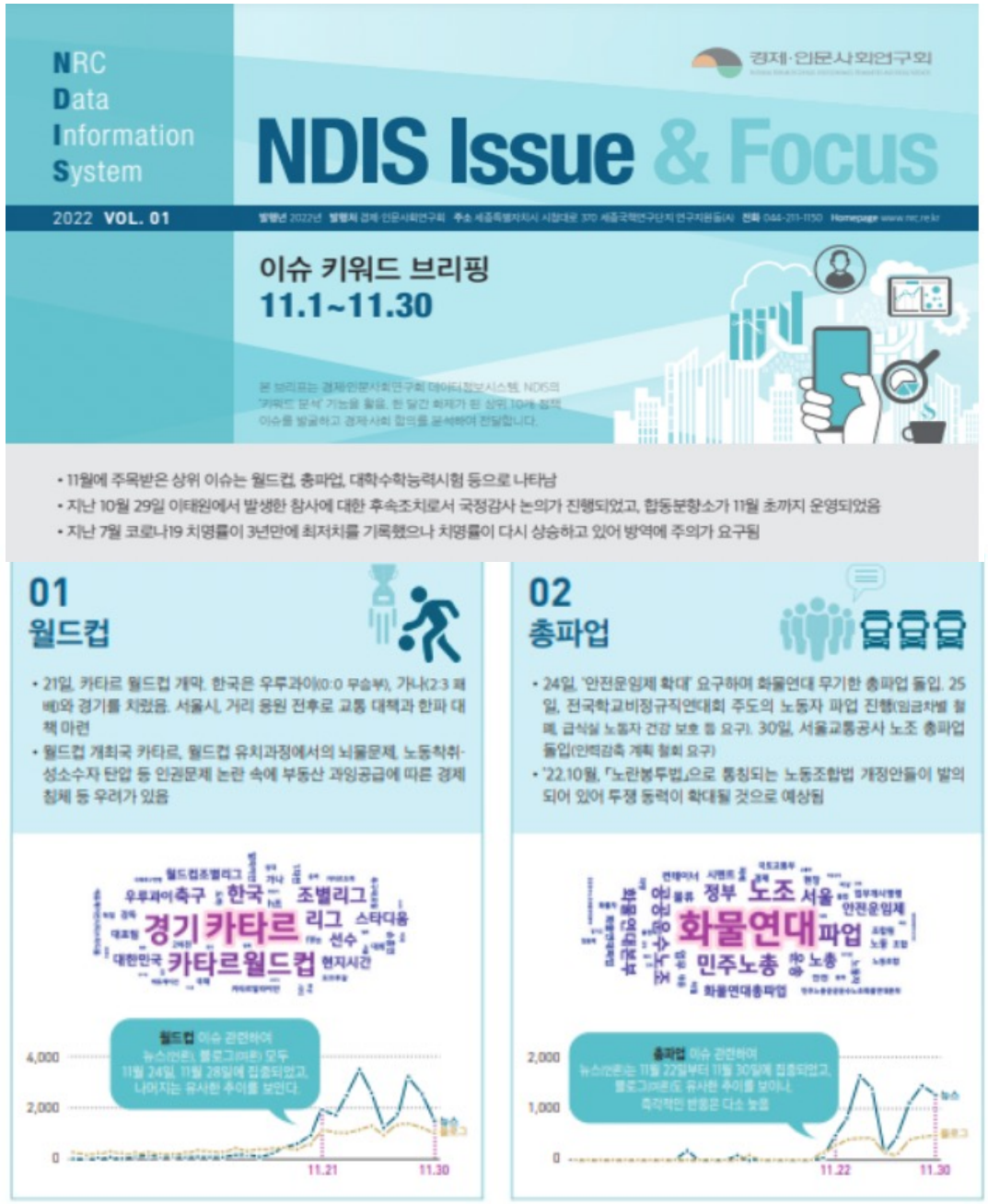
- 연구진 논문자료 태블로 시각화





업무 : 월간지 이슈 발간


전략
합동참모본부
대륙
대응
한국
처음
북방한계선
storm
도발
비질런트
핵
대규모
한반도
미국
북한
한미
스톰
탄도
탄도미사일
미사일
훈련
동해
대륙간탄도미사일
발사
뉴스
한미연합공중훈련
400
200
0
2022.11.01.
뉴스
블로그


NRC
Data
Information
System
경제·인문사회연구회
NDIS Issue & Focus
2022 VOL. 01
이슈 키워드 브리핑
11.1~11.30
• 11월에 주목받은 상위 이슈는 월드컵, 총파업, 대학수학능력시험 등으로 나타남
• 지난 10월 29일 이태원에서 발생한 참사에 대한 후속조치로서 국정감사 논의가 진행되었고, 합동분향소가 11월 초까지 운영되었음
• 지난 7월 코로나19 치명률이 3년만에 최저치를 기록했으나 치명률이 다시 상승하고 있어 방역에 주의가 요구됨
01
월드컵
• 21일, 카타르 월드컵 개막. 한국은 우루과이(0:0 무승부), 가나(2:3 패배)와 경기를 치렀음. 서울시, 거리 응원 전후로 교통 대책과 한파 대책 마련
• 월드컵 개최국 카타르, 월드컵 유치과정에서의 뇌물문제, 노동착취·성소수자 탄압 등 인권문제 논란 속에 부동산 과잉공급에 따른 경제침체 등 우려가 있음
우루과이
축구
한국
조별리그
경기
카타르
리그
스타디움
선수
대한민국
카타르월드컵
현지시간
4,000
2,000
0
11.21
11.30
02
총파업
• 24일, '안전운임제 확대' 요구하며 화물연대 무기한 총파업 돌입. 25일, 전국학교비정규직연대회 주도의 노동자 파업 진행(임금차별 철폐, 급식실 노동자 건강 보호 등 요구). 30일, 서울교통공사 노조 총파업 돌입(인력감축 계획 철회 요구)
• '22.10월, 「노란봉투법」으로 통칭되는 노동조합법 개정안들이 발의되어 있어 투쟁 동력이 확대될 것으로 예상됨
정부
노조
서울
화물연대
파업
안전운임제
민주노총
노총
화물연대총파업
2,000
1,000
0
11.22
11.30
뉴스
블로그

프로젝트1

(텍스트 데이터 분석, 토픽 모델링을 활용하여)

윤석열 대통령 연설문에서 대통령의 핵심 가치 및 주제찾기

프로젝트1 과정

Lda 토픽모델링이란?

LDA(Latent Disrichlet Allocation) 토픽 모델링은 확률 기반의 모델링 기법을 통해 방대한 양의 문서 데이터를 분석함으로써 문서 내에 어떤 토픽이, 어떤 비율로 구성되어 있는지를 분석합니다.

→ 즉, 큰 방대한 텍스트 데이터에서 핵심토픽을 추출해줌. -> 57개 연설문에서 대통령이 말하고자 하는 핵심.주제(가치)들 도출 가능

대통령실 홈페이지에서 대통령 연설문 57개 python으로 웹_크롤링

Mecab

명사
동사
형용사

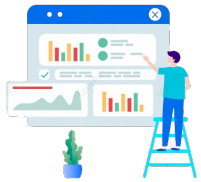
Python 형태소 분석기인 mecab을 이용해서 문장을 명사, 동사, 형용사 형태로 추출

불용어(=분석하는데 필요없는 단어) 사전 활용하여 불용어 제거 Ex) 의존명사와 같은 한글자 명사들. 척, 뿐, 것 등등

python

Coherence model

일관성 점수로 토픽 갯수 지정
(7개일때, 일관성 점수가 가장 높았음)



모델링 후, 토픽 결과에 대한 해석

프로젝트1 과정

연구 이유 : 윤석열 대통령 연설물을 통한 의도 파악 및 연구주제 발굴

대한민국 대통령실 **대통령의 말과 글**

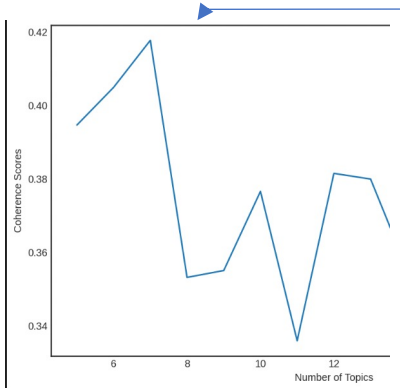
연설문57개에서 모든 문장들 수집

명사 : 개혁, 비상, 국무회의
동사 : 지키다, 만들다
형용사 : 어렵다

문장 형태소 분석

'척', '뿐', '때문', '저는'과 같은 단어 등 제거

토픽 모델링 분석에 유의미하지 않은 단어 제거

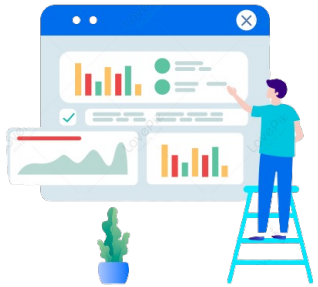


토픽 모델링 하기 위해 토픽 갯수 지정 필요
- 몇개의 주제(토픽)이 모델링 하는데 적합한지, 일관성 점수 Coherence model 라이브러리를 이용해 토픽 개수 7개 지정

LDA 토픽모델링

DATA MODELING

모델링 실행



토픽 모델링 결과 분석

프로젝트1 결과

Ⅲ 토픽 모델링(1)

토픽의 수 : 7개 선정

토픽 1	
자유	1.226
평화	0.492
가치	0.427
시민	0.363
세계	0.306
보편	0.305
인권	0.255
연대	0.248
지키다	0.242
번영	0.225

자유_보편적 인권 가치

토픽 2	
혁신	0.876
규제	0.786
민간	0.374
성장	0.315
도약	0.273
추진	0.183
과학	0.167
개선	0.160
필요	0.152
기업	0.150

규제혁신_민간성장(기업 중심 도약 추진 필요)

토픽 3	
산업	0.813
국가	0.482
기술	0.482
정부	0.301
우주	0.300
안보	0.292
미래	0.275
전략	0.273
육성	0.264
핵심	0.253

미래산업_전략육성

토픽 4	
국민	2.238
지키다	0.354
나라	0.354
책임	0.330
안전	0.317
생명	0.303
만들다	0.256
정부	0.252
재산	0.241
서다	0.239

국민안전(생명 및 재산)_국가책임

Ⅲ 토픽 모델링 (2)

토픽 5	
피해	0.964
지원	0.543
복구	0.432
신속	0.407
입다	0.308
크다	0.273
집중호우	0.270
이번	0.250
일상	0.243
최선	0.227

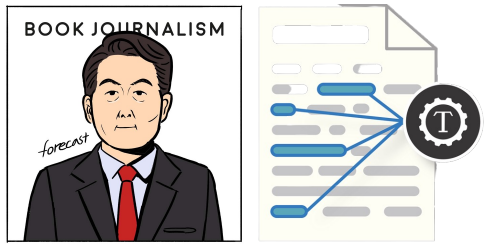
집중호우_피해복구

토픽 6	
반갑다	1.661
빕다	0.142
위원	0.109
이렇다	0.093
존경	0.090
충남	0.075
도민	0.071
임원	0.043
연합회	0.042
선거	0.041

첫 인사

토픽 7	
감사	2.237
진심	0.075
자리	0.071
헌신	0.037
의료진	0.035
역사	0.029
빌리다	0.026
노력	0.023
일깨우다	0.021
교직원	0.020

감사 인사



57개 연설문 -> 토픽 7개로 압축
토픽별 구성하고 있는 단어 분포 모습 및 결과해석

프로젝트1 결론

윤석열 대통령 연설문 등 텍스트 분석 결과,
1) 산업기술, 2)경제안보, 3)국민안전, 4) 규제혁신, 5) 균형발전 등 연구주제 발굴

No	주요 주제	중장기 목표(주요내용)	근거
1	산업기술	초격차 미래성장 기반 구축	토픽3
2	경제안보	국내외 경제안보 협력 체계 구축	단어 의미 분석(안보, 협력)
3	국민안전	국민의 생명,재산 보호를 위한 국가 책임 강화	토픽4
4	규제혁신	민간(기업) 중심 성장을 위한 규제 완화	토픽2
5	균형발전	지역(로컬) 브랜드 개발을 통한 균형발전	단어 의미 분석(지역)

윤석열 대통령 57개연설문을 통해 윤석열 정부가 추구하는 주요 주제들, 토픽들 추정