2022년도 경남빅데이터센터 분석 계획

토픽 모델링을 활용한 태풍 힌남노 경남지역 뉴스토픽 분석

☐ 연구배경 및 목적

○ 태풍 힌남노 관련 경남지역 뉴스의 토픽 모델링을 통해 주요 토픽을 분석

☐ 연구수행

○ 기　간 : 2022. 9. 14 ~ 2022. 9. 30

○ 데이터

데이터명	기준 (날짜)	설명	출처
뉴스 메타데이터	2022. 08.~09.	· 경상남도 내 힌남노 관련 뉴스 메타데이터 (언론사, 기고자, 제목 등)와 개체명(인물, 기관, 장소 등)	한국언론진흥재단

○ 기대효과

- 주요 토픽을 통해 경상남도 지역별 피해 사례 및 복구 방안 파악

☐ 연구 주요내용

○ 태풍 힌남노 관련 경남지역 뉴스 데이터 수집

○ 데이터에 분포된 태풍 힌남노 관련 주요 키워드 수집 및 분석

○ 토픽 모델링을 통해 키워드 내에 잠재된 의제(토픽) 파악

☐ 분석방법

○ 워드클라우드를 활용한 뉴스 주요 키워드 시각화

- 전처리 된 뉴스 텍스트 데이터의 키워드 추출
- 추출된 키워드의 출현 빈도를 바탕으로 키워드 시각화

○ LDA(Latent Dirichlet Allocation) 기반 토픽 모델링

- 토픽 모델링에 이용되는 대표적인 알고리즘
- 모델링 결과에 따른 뉴스데이터 키워드의 비율을 참고하여 토픽 파악

☐ 전처리

○ 불용어 및 특수문자 제거

- 워드클라우드 및 토픽 모델링에 큰 의미가 없는 단어 및 특수문자 제거
- '취재','기자','앵커','KBS','[' ,'' 등

○ 분석 단어 전처리

- 분석 결과의 의미를 명확히 하기 위해 명사로 된 단어만 선정
- 최소 2글자 이상의 단어만을 선정

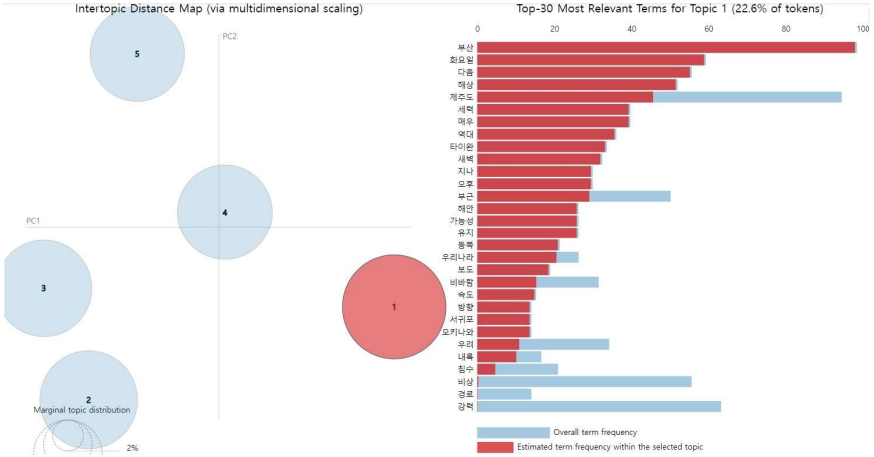
☐ 그래프 예시(워드클라우드)

○ 뉴스데이터 워드클라우드 시각화(힌남노 발생 ~ 한반도 상륙 전)

○ 뉴스데이터 워드클라우드 시각화(한반도 상륙 및 이후)

☐ 그래프 예시(토픽 모델링)

○ 태풍 발생 ~ 한반도 상륙 전 뉴스토픽 모델링 시각화(5개의 잠재된 토픽)



○ 태풍 한반도 상륙 및 이후 뉴스토픽 모델링 시각화(9개의 잠재된 토픽)