토픽 모델링 기반 뉴스기사 분석을 통한 서울시 이슈 도출

*권민지
서울기술연구원
*mjkwon@sit.re.kr

Identifying Seoul city issues based on topic modeling of news article

*Kwon, Min-Ji
Seoul Institute of Technology

요약

대중들에게 정보를 빠르고 정확하게 제공하는 대표 매체인 뉴스 기사는 일 평균 1만 5천 건 이상이 보도되고 있다. 특정 주제 또는 분야에 대한 전반적인 동향을 파악하고자 대량의 텍스트 데이터를 수집하여 텍스트 마이닝(Text mining)과 머신러닝 등을 적용하는 연구들이 활발하게 수행되고 있다. 본 연구에서는 서울시의 이슈 및 문제를 파악하고자 약 5년간 뉴스 기사를 수집하여 키워드 분석 및 토픽 모델링을 적용하였다. 분석 결과 5년간의 뉴스 기사에서 빈번하게 출현하는 키워드들을 도출하였고 연도별로 도출된 키워드들을 비교분석하였다. 또한 토픽 모델링 적용 결과 뉴스 기사를 구성하는 20개의 주제를 도출하였으며 이를 기반으로 서울시의 주요 이슈들을 파악할 수 있다. 본 연구는 연도별, 분야별 세부 내용 및 시계열 분석, 다른 도시들의 이슈 및 문제를 도출하는데 활용될 것으로 기대된다.

1. 서론

뉴스 기사는 우리 사회에서 발생하는 사건과 이슈를 담고 있으며 대중들에게 정보를 제공하는 주요 매체이다. 하루에 보도되는 뉴스 기사의 양은 사람이 직접 처리하고 분석할 수 있는 수준을 초월한다. 뉴스 빅데이터 분석 플랫폼 빅카인즈[1]에서는 하루 동안 54개 주요 언론사의 약 1만 5천 건의 뉴스 기사가 수집되고 있으며 30년간의 약 6천만 건의 뉴스를 축적하고 있다. 이에 특정 주제 또는 분야에 대한 뉴스 기사 분석에 데이터 기반의 방법론이 요구된다.

따라서 본 연구에서는 서울시 이슈를 도출하고자 약 5개년 중앙지[2] 뉴스 기사를 수집하여 키워드 분석과 토픽 모델링을 수행하였다. 본 연구 결과로 향후 장기적 시계열 데이터를 분석하여 보다 심도 높은 서울시의 이슈를 모니터링하고 다른 도시들의 문제 및 이슈 또한 도출할 수 있을 것으로 기대된다.

2. 선행연구

학계, 기술, 산업 등의 영역에서 논문, 특허, 뉴스 기사 등의 데이터를 기반으로 동향을 파악하는 연구들이 다수 존재한다. 주로 대량의 관련 텍스트 데이터를 수집하고 군집으로 분류하여 연도별 이슈를 비교하고 새로운 분야의 성장 및 쇠퇴, 분야 간 융합을 분석한다. [1]에서는 신재생에너지 동향을 파악하기 위해 언론기사를 수집하여 토픽 모델링 분석을 수행하였다. [2]에서는 논문 데이터에 토픽 모델링을 적용해 기존의 학문 분류체계와 비교분석하였다. [3]에서는 뉴스기사 데이터에 동시 등장한 회사들을 동일 산업군으로 가정하고 SIC 산업 코드와 비교분석하였다.

3. 데이터 수집 및 방법론

본 연구에서는 뉴스 기사를 수집하여 키워드 분석 및 LDA 토픽 모델링을 적용해 서울시 이슈를 도출하는 것을 목표로 한다. 또한 도출된 결과를 기반으로 연도별 이슈 변화를 비교분석한다.

3.1 데이터 수집

분석 대상 데이터로는 2015.01.01.~2019.09.16.(데이터 수집일 기준) 중앙지 뉴스기사 중 “서울시”를 포함하는 약 5개년 기사를 수집하였다. 빅카인즈 상세검색을 통해 분석 대상이 되는 기사의 URL 데이터를 수집하고 Python Newspaper 라이브러리를 통해 뉴스기사 본문을 수집하였다. 수집된 전체 뉴스기사 건수는 총 80,609건이며 연도별 건수는 **Fig. 1.**과 같다.

3.2 데이터 전처리

데이터 전처리는 다음 순서로 진행하였다. 첫째, 수집된 뉴스기사 본문에서 KoNLP 라이브러리를 사용해 명사를 추출한다. 둘째, gensim 라이브러리의 bigram을 사용해 빈번하게 출현하는 2단어를 하나의 단어로 처리한다. 셋째, 빈번히 등장하나 의미가 없는 불용어를 제거한다. 그 예로는 “그”, “이”, “곳”, “이제”, “를”, “때문”, “통해” 등이 있다.

1) 빅카인즈(BIGKinds)는 뉴스빅데이터 분석 서비스로 종합일간지, 경제지, 지역일간지 등 54개 언론사의 뉴스기사 메타데이터 제공(https://www.bigkinds.or.kr/)
2) 경향신문, 국민일보, 내일신문, 동아일보, 문화일보, 서울신문, 세계일보, 조선일보, 중앙일보, 한겨레, 한국일보

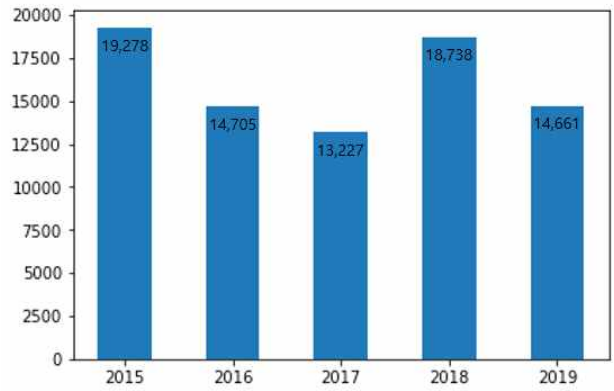

Fig. 1. 연도별 수집된 뉴스기사 건수

3.3 토픽 모델링

토픽 모델링은 대규모 텍스트 데이터를 분류하는 머신러닝 기법으로 단순 키워드 분석과 달리 키워드들의 군집으로 형성되는 문서의 주제를 도출할 수 있다. 이 논문에서는 토픽 모델링의 대표적인 모델인 Latent Dirichlet Allocation(LDA) 기법을 사용하였다. LDA는 잠재변수인 주제별 단어 분포와 문서별 주제 분포를 기반으로 문서가 생성된다고 가정한다.[4]

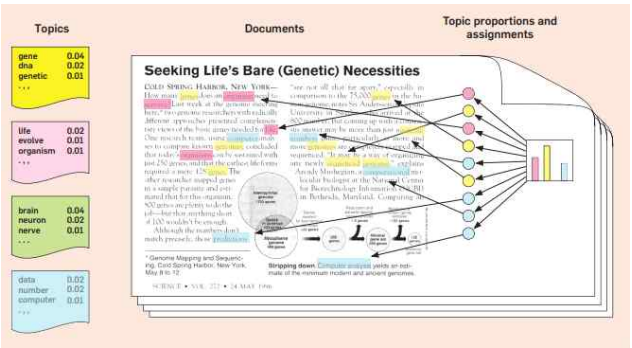

Fig. 2. LDA Topic modeling

4. 분석 결과

전처리한 데이터에서 빈번하게 출현하는 키워드 1,000건을 워드클라우드(wordcloud)로 출력한 결과는 **Table 1.**과 같다. 2015년도부터 2019년도까지 전체 뉴스 기사에서 시민, 정부, 정책, 사회, 주민, 교육 등의 단어가 빈번히 출현하는 것을 확인할 수 있다. 또한 전체 기사에서 다수 출현한 단어들을 제외하고 각 연도별로 빈번히 도출된 키워드를 도출하였을 때 2019년도의 경우, 광화문 광장, 차량, 택시, 미세먼지 등의 키워드가 빈번히 출현하였고 2015년도에는 병원, 분양, 메르스, 환자 등의 키워드가 다수 출현하였다.

전체 기사에 토픽 모델링을 적용하여 20개의 도시 이슈를 도출한 결과 토픽과 토픽을 구성하는 단어들은 **Table 2.**와 같다. Topic 1은 환경 및 이동수단, Topic 3은 광화문 광장 집회, Topic 4는 주택 및 부동산, Topic 5는 택시 업계, Topic 6은 교육, Topic 7은 문화, Topic 8은 종교, Topic 9는 도시 인프라, Topic 12는 공공시설, Topic 13은 청년 및 취업, Topic 15는 의료, Topic 18은 스포츠, Topic 19는 근로자 및 임금, Topic 20은 범죄 및 수사와 관련된 주제로 추론해볼 수 있다.

2015-2019				2019			
시민	37,034	정부	27,453	광화문광장	4,007	차량	2,567
정책	25,655	사회	23,539	택시	2,507	미세먼지	2,181
2018				**2017**			
미세먼지	2,973	차량	2,915	여성	1,868	광장	1,339
여성	2,701	택시	2,426	역사	1,314	일자리	1,270
2016				**2015**			
청년	4,982	역사	1,727	병원	3,058	분양	2,323
청년수당	1,617	일자리	1,335	메르스	2,220	환자	2,088

Table 1. 연도별 뉴스기사 키워드 워드 클라우드

	Characterizing Words
Topic 1	미세먼지, 환경, 에너지, 안전, 사고, 자전거, 차량, 자동차
Topic 2	동물, 어린이집, 유치원, 반려동물, 고양이
Topic 3	광화문 광장, 정치, 후보, 집회, 개혁, 반대, 운동
Topic 4	주택, 건물, 아파트, 부동산, 재개발, 건축, 재건축
Topic 5	택시, 택시기사, 승객, 택시 업계, 승차 거부, 요금
Topic 6	학교, 교육, 대학, 청소년, 교사, 수업, 전형
Topic 7	주민, 마을, 도서관, 공연, 문화, 예술, 음악
Topic 8	교회, 한국 교회, 목사, 기독교, 예배
Topic 9	한강, 도로, 지하철, 버스, 운행, 터널, 승객
Topic 10	단지, 건설, 분양, 상가, 오피스텔, 부지, 완공
Topic 11	대통령, 북한, 청와대, 국정원, 위원, 국장
Topic 12	공원, 도시, 조성, 전시, 역사, 디자인, 관광, 박물관
Topic 13	청년, 정책, 청년 수당, 여성, 일자리, 취업, 소득
Topic 14	정부, 지자체, 예산, 공무원, 자치구, 위원회, 의회
Topic 15	병원, 환자, 의료, 치료, 사망, 증상, 건강, 진료
Topic 16	가게, 영화, 회원, 작가, 광고, 작품
Topic 17	제로 페이, 결제, 가맹, 환급, 할인, 인하, 공제
Topic 18	스포츠, 선수, 팬, 감독, 경기장, 경기, 구단, 리그
Topic 19	노동자, 노조, 생활임금, 근로자, 고용, 채용, 최저임금
Topic 20	경찰, 검찰, 수사, 의혹, 변호사, 재판, 징계, 처벌

Table 2. Topic - Characterizing Words

5. 결론

본 연구에서는 뉴스 기사 키워드 분석 및 토픽 모델링 적용을 기반으로 서울시의 이슈를 도출하였고 연도별로 비교분석하였다. 향후 서울시의 특정 분야와 관련된 문제 및 이슈에 대한 세부 분석과 키워드들 간의 관계를 도출하는 네트워크 분석을 수행해 볼 수 있다.

이 연구를 기초자료로 서울시의 연도별 이슈 변화 분석 및 다른 도시들의 문제 및 이슈 모니터링 또한 도출이 가능할 것으로 기대된다.

ACKNOWLEDGEMENT

본 논문은 서울기술연구원(18-4-4, 서울형 혁신융합 미래도시 기획 연구)의 지원을 받아 수행된 연구임.

Reference

[1] KyuSik Shin, HoeRyeon choi, HongChul Lee. “Topic Model Analysis of Research Trend on Renewable Energy”. Journal of the Korea Academia-Industrial cooperation Society, 16(9), 6411-6418. 2015.

[2] Chyi-Kwei Yau, Alan Porter, Nils Newman, and Arho Suominen. “Clustering scientific documents with topic modeling”. Scientometrics, 100(3), 767-786. 2014.

[3] Namil Kim, Hyeokseong Lee, Wonjoon Kim, Hyunjong Lee, and Jong Hwan Suh. “Dynamic patterns of industry convergence: Evidence from a large amount of unstructured data”. Research Policy, 44(9), 1734-1748. 2015.

[4] David M. Blei. “Probabilistic topic models”. Commun. ACM, 55(4), 77-84. 2012.