LDA 토픽 모델링을 활용한 서울시 도시공원 이용자의 인식 및 행태 연구†

최혜영*, 강민우, 정윤경***, 이준*****
*성균관대학교 공과대학 건설환경공학부 부교수, **성균관대학교 소프트웨어융합대학 소프트웨어학과 석사과정,
***성균관대학교 소프트웨어융합대학 소프트웨어학과 부교수, ****성균관대학교 일반대학원 조경학과 석사과정

1. 서론

2020년, COVID-19가 시작되면서 직면한 사회적, 물리적 거리두기는 많은 영역에서 사람들이 가지고 있던 기존의 인식 및 태도, 규범의 변화를 가져왔다. 특히 공공 공간의 역할에 대해 다시 생각하게 되었으며, 거리두기에 비교적 자유롭고 안전한 공원 및 녹지의 중요성을 깨닫게 되었다. 공원으로의 방문 빈도는 2022년 10월 기준, 팬데믹 이전과 비교해 약 52% 늘어났으며(Google Mobility Report, 2022) 도시공원에 관한 연구 또한 활발히 일어나게 되었다. 특히, 사람들의 도시공원 이용 패턴 변화와 같은 행태 연구나 도시공원에 대한 만족도 등에 관한 연구가 다수 수행되었다. 기술의 발전과 함께 팬데믹 이후 비대면이 선호되면서 연구자가 직접 현장 조사를 하지 않고도 다량의 데이터를 수집, 분석하여 결과를 도출하는 방식이 적극적으로 사용되기 시작했다(채진해 등, 2020). 특히 사회관계망서비스(SNS) 데이터는 공원 이용자가 연구자의 개입 없이 자발적으로 작성한 데이터로 이용자 관점에서 객관적으로 도시공원을 바라볼 수 있다는 장점이 있다(Park et al., 2022).

본 연구는 SNS 중 인스타그램(Instagram) 플랫폼을 활용하여 2017년 1월부터 2023년 8월까지의 텍스트 데이터를 확보한 후 LDA 토픽 모델링을 통해 서울시 공원 전체에 대한 사람들의 인식 및 행태를 살펴보는 것을 목표로 삼았다. 연구의 결과를 공원의 물리적, 경험적 환경 분석과 접목한다면 도시공원의 계획 방향에 대한 시사점을 도출할 수 있을 것으로 기대된다. 특히 인스타그램은 SNS 중에서도 MZ 세대들이 많이 사용하는 플랫폼으로 미래세대의 공원에 대한 인식 및 이용 행태를 알아볼 수 있다는 점에서 연구 가치가 있다.

2. 연구방법

2.1 데이터 수집

파이썬(Python)을 통해 2017년 1월부터 2023년 8월까지 인스타그램에서 “공원”, “서울”의 키워드로 포스트를 수집하였다. 서울의 특정 공원이 아닌 서울 전역의 공원을 중립적으로 살펴보기 위해 먼저 “공원”의 키워드가 포함된 포스트를 크롤링하였다. 이 중 서울의 공원과 연관된 게시글 데이터를 확보하기 위해 “서울” 키워드를 포함한 데이터로 범위를 좁혔다. “공원”과 “서울” 키워드가 모두 포함된 포스트 개수는 총 17,043개(COVID-19 이전의 포스트 개수: 9,274개, COVID-19 이후의 포스트 개수: 7,769개)다. 수집한 데이터 중 포스트에 포함된 해시태그와 본문 텍스트 전체가 본 연구에 사용되었다.

2.2 데이터 전처리

수집한 데이터를 분석 가능한 가장 작은 단위로 나누는 토큰화(tokenization)를 수행하였으며 그중 명사만을 추출하여 사용하였다. 선행 테스트를 통해 불용어(예: 인스타, 스타, 일리, 맞팔, 선팔 등)를 선정한 뒤 분석에서 배제하였다. 이 과정을 통해 총 151,482개의 토큰을 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 토픽 모델링에 최종 활용하였다.

2.3 LDA 토픽 모델링

LDA 토픽 모델링을 통해 텍스트 데이터를 분석하였으며 파이썬 Gensim 라이브러리를 사용하였다. LDA는 2003년 David M. Blei에 의해 고안된 토픽 모델링의 방법(Blei et al., 2003)으로, 대량의 데이터를 차원 축소를 통해 잠재적(latent)으로 의미 있는 토픽을 찾아내는 데이터 분석 방법이다. 토픽 모델링은 데이터가 토픽들의 혼합으로 이루어져 있으며, 토픽들은 확률 분포에 기반을 두어 단어를 만들어 낸다고 가정한다. 즉, 데이터 내에 어떤 토픽이, 어떤 비율로 구성되어 있는지 분석하는 기법이다. LDA 토픽 모델링은 “토픽의 단어 비중과 문서의 토픽 비중이라는 두 가지 변수의 결합 확률분포에 따라 문서의 토픽을 찾는 과정”이다(문성현 등, 2018).

†본 발표 논문은 성균관대학교의 2023학년도 AI융합연구비에 의하여 연구되었음.

3. 분석 및 결과

3.1 전체 기간 분석 결과

Coherence Score(일관성 지수)가 높을수록 토픽이 의미론적으로 일관성이 높다고 판단한다(Newman et al., 2010). 하지만 토픽의 수가 너무 많을 경우 각 토픽의 설명력이 떨어진다고 판단하여 토픽 개수 10개를 최대로 제한한 뒤 Coherence Model을 통해 토픽 최적화 테스트를 진행하였다. 그 결과 전체 기간에 해당하는 데이터의 경우 6개의 토픽으로 구분되었다. 토픽 1은 19.1%, 토픽 2는 14.4%, 토픽 3은 13.4%, 토픽 4는 16.9%, 토픽 5는 19.7%, 토픽 6은 16.4%의 설명력을 가진다. 토픽 1의 상위 빈도 20위 내 단어에는 재미(0.120)[1], 한복(0.036), 홍대(0.031), 웨딩(0.022), 선유도(0.022), 치킨(0.009), 자전거(0.009), 맥주(0.009) 등이 등장했다. 토픽 2의 경우 올림픽(0.039), 박물관(0.037), 자유(0.033), 축제(0.031), 숲길(0.024), 경의선(0.024), 혼자(0.017), 아침(0.014), 홀로(0.014) 등이 상위 빈도 20위 내 단어에 포함되었다. 토픽 3에는 편안(0.132), 산모(0.063), 휴식(0.031), 남산(0.026), 반포(0.024), 저녁(0.017), 평일(0.016) 등의 단어가, 토픽 4에는 맛집(0.045), 운동(0.028), 피크닉(0.027), 놀이터(0.018), 커피(0.020), 무료(0.017), 어린이(0.016), 행복(0.015), 육아(0.015), 소풍(0.014) 등의 단어가 상위 빈도 20위 내 단어에 해당되었다. 토픽 5는 벚꽃(0.078), 다리(0.051), 사람(0.029), 강아지(0.028), 구경(0.027), 시민(0.027), 호수(0.026), 셀피(0.023), 애견(0.015), 일요일(0.014), 토요일(0.013)을 포함하며, 촬영(0.069), 여유(0.042), 카메라(0.028), 용산(0.027), 스튜디오(0.022), 노을(0.022), 추억(0.019), 화보(0.019), 커플(0.016), 인물사진(0.015)의 단어가 토픽 6을 설명하는 상위 빈도 단어로 분석되었다. 이용자들의 인식 및 행태가 드러나는 각 토픽별 성격을 살펴보면, 토픽 1은 '재미있는 활동', 토픽 2는 '자유로움', 토픽 3은 '일상 휴식', 토픽 4는 '가족 소풍', 토픽 5는 '반려동물과 함께하는 여가 활동', 토픽 6은 '사진 찍기'로 구분할 수 있다.

3.2 COVID-19 이전과 이후의 비교

COVID-19 이전은 총 7개의 토픽, COVID-19 이후는 총 9개의 토픽으로 구분되었다. COVID-19 이전의 경우 전체 기간 토픽 분석 결과와 크게 다르지 않았다. 다만 전체 기간 토픽 분석의 내용에서 카페, 맛집, 여유, 피크닉, 패션, 화보, 셀피 등 소비와 문화 중심의 단어가 별도의 토픽(17.7%)으로 집중되는 것을 확인할 수 있었다. COVID-19의 발생 후 현재까지의 분석 결과는 COVID-19 발생 전과 다소 다른 결과를 보여준다. 9개의 토픽 중 16.9%로 가장 비중이 높은 토픽의 경우 보호(0.035), 코로나(0.029), 사람0.028), 시간(0.027), 생활(0.024), 서비스(0.020), 낙산(0.020), 마스크(0.016), 주변(0.015), 바람(0.015), 풍경사진(0.015) 등의 단어가 상위 빈도 20위 내에 올랐다. 코로나와 관련된 단어들이 "공원", "서울" 키워드와 함께 등장하는 것을 알 수 있었다. 두 번째로 높은 비중(15.3%)을 차지하는 토픽은 취미(0.034), 신도시(0.031), 리프트(0.027), 출사(0.021), 수도권(0.021), 등산(0.021), 레이스(0.019), 추억(0.017), 잠실(0.017), 코스(0.016) 등의 단어를 포함하였다. 사회적 거리두기가 가능한 활동에 대한 언급이 늘어난 것으로 보인다.

4. 연구의 의의 및 한계

사람들이 자발적으로 작성한 대량의 텍스트 데이터를 확보하여 분석에 이용함으로써 연구의 질과 객관성을 높인 점, 서울시 전체 공원을 대상으로 삼은 점, 직접 조사 수행이 어려울 경우나 불특정 다수의 의견을 살펴보고자 할 때 대안적 방법이 될 수 있다는 점에서 본 연구의 의의가 있다. 동시에 몇 가지 점에서 한계 또한 가진다. 첫째, 부산, 제주도 등 다른 지역의 명칭이 등장하는 토픽이 발생하는 점을 토대로 "공원", "서울"의 키워드로 수집한 데이터가 모두 '서울의 공원'과 관련 있다고 보기에 어려움이 있다. 둘째, "서울숲", "한강시민공원" 등 서울의 대표 공원은 그 자체로 상당수의 게시글이 검색되는데 반해 "공원", "서울"의 키워드로는 약 17,000여 건에 불과하여 데이터의 완결성이 부족하다고 볼 수 있다. 연구에서 사용한 데이터가 서울시 전체의 공원을 대표할 수 있는지 검증이 필요하다. 셋째, 일반적, 보편적 성격의 연구 결과로 인해 도시공원의 계획 및 설계로 연결시키기에 구체성이 부족하다. 후속 연구를 통해 본 연구 결과의 한계를 극복할 수 방향으로 다듬어나갈 필요가 있다.

참고문헌

1. 문성현, 정세환, 지석호(2018) Latent dirichlet allocation 기법을 활용한 해외건설시장 뉴스기사의 토픽 모델링(topic modeling). Journal of the Korean Society of Civil Engineers 38(4): 595-599.
2. 채진해, 조민준, 김복영(2020) 텍스트 빅데이터 분석을 통한 COVID-19 전후 서울시 주요 도시공원의 시민 이용행태 및 관심도 변화. 서울연구원 서울연구논문 공모전 종간콘텐츠.
3. Blei, D. M., A. Y. Ng and M. I. Jordan(2003) Latent dirichlet allocation. Journal of Machine Learning Research 3(Jan): 993-1022.
4. Google(2022) COVID-19 Community Mobility Reports.
5. Newman, D., J. H. Lau, K. Grieser and T. Baldwin(2010, June) Automatic evaluation of topic coherence. In Human language technologies. The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: 100-108.
6. Park, S., S. Kim, J. Lee and B. Heo(2022) Evolving norms: Social media data analysis on parks and greenspaces perception changes before and after the COVID 19 pandemic using a machine learning approach. Scientific Reports 12(1): 13246.

1) 단어 옆 숫자는 각 토픽과 가까운 정도, 즉 해당 토픽을 잘 설명하는 정도를 의미한다.