한국지역지리학회지 제23권 제2호(2017) 376-387

토픽 모델링을 이용한 트위터 데이터의 공간 분포 패턴 분석*

우현지** · 김영훈***

Spatial Distribution Patterns of Twitter Data with Topic Modeling*

Woo, Hyun Jee** · Kim, Young Hoon***

요약 : 본 연구는 트위터를 대상으로 트윗 공간 데이터에서 지리적 의미를 탐색하기 위한 방법을 모색하였다. 트윗 공간 데이터의 구축 과정 및 지리적 분석의 프레임워크를 정립하고 지리적 연구 방법론을 제안하였다. 이를 위해 본 연구는 제주도의 GPS 좌표 참조 트윗(geotweet)을 대상으로 트윗의 내용적 특성과 트윗 발생 위치의 공간 분포 특성을 확인하였다. 제주도 좌표 참조 트윗에서는 지명 또는 장소명이 많이 출현하였는데, 이는 자신의 위치를 알리고자하는 의도로 파악하였다. 트윗의 공간 분포는 제주공항을 중심으로 한 일부 관광지 주변으로 핫스팟이 확인되었고, 이는 제주도 유동인구 핫스팟과 유사한 패턴을 보였다. 주제 중심의 트윗 분석을 위해 본 연구에서는 토픽 모델링 알고리즘을 이용하여 분석하였다. 분석 결과, 주제의 지리적 위치와 트윗의 내용은 서로 관련이 있음을 알 수 있었다. 마지막으로 본 연구는 토픽 모델링 분석을 통해 방대한 트윗 데이터의 내용에 상응하는 지역 분포 특성을 직관적으로 확인하는데 유용하게 활용될 수 있다는 것을 확인하였다.
주요어 : 트위터, 트윗 공간 데이터, 좌표 참조 트윗, 주제 중심 트윗, 토픽 모델링

Abstract : This paper attempts to analyze the geographical characters of Twitter data and presents analysis potentials for social network analysis in geography. First, this paper suggests a methodology for a topic modeling-based approach in order to identify the geographical characteristics of tweets, including an analysis flow of Twitter data sets, tweet data collection and conversion, textural pre-processing and structural analysis, topic discovery, and interpretation of tweets' topics. GPS coordinates referencing tweets(geotweets) were extracted among sampled Twitter data sets because it contains the tweet place where it was created. This paper identifies a correlated relationship between some specific topics and local places in Jeju. This correlation is closely associated with some place names and local sites in Jeju Island. We assume it is the intention of tweeters to record their tweet places and to share and retweet with other tweeters in some cases. A surface density map shows the hotspots of tweets, detecting around some specific places and sites such as Jeju airport, sightseeing sites, and local places in Jeju Island. The hotspots show similar patterns of the floating population of Jeju, especially the thirty-year age group. In addition, a topic modeling algorithm is applied for the geographical topic discovery and comparison of the spatial patterns of tweets. Finally, this empirical analysis presents that Twitter data, as social network data, provide geographical significance, with topic modeling approach being useful in analyzing the textural features reflecting the geographical characteristics in large data sets of tweets.
Key Words : Twitter, tweet spatial data, GPS-referenced tweets, topic-oriented tweets, topic modelling

1. 서 론

새로운 소셜 미디어(social media) 환경의 등장으로 기존의 공간 개념이 재구성되고 있다. 기존의 공간이 오프라인 공간을 의미했다면 이제는 온라인 공간으로 그 범위가 확장되고 있다. 즉, 초고속 정보통신망을 통해 형성되는 미디어의 공간은 고정적이지는 않지만 실제의 공간을 반영하고 있으며, 사용자간 연결을 기반으로 하는 네트워크의 끊임없는 상호작용으로 구성·확장되고 있다(이병혁 등, 2005). 따라서 지금의 IT 환경과 연관된 새로운 공간구조의 변화를 해석하려는 연구가 필요하다.

특히 모바일 환경을 기반으로 이루어지는 SNS의 사용이 급등함에 따라 개인들은 언제 어디서나 미디어에 접근할 수 있으며 개인들의 일상은 시공간의 흐름에 따라 기록된다. 온라인 상에서 개인들이 실시간으로 자유롭게 정보를 생산, 전달, 공유하려는 변화가 발생하게 되었고 이제는 개인이 기존 미디어의 정보

* 이 논문은 2017년 한국지역지리학회 동계학술대회에서 발표한 내용을 수정·보완한 연구임.
** 한국교원대학교 지리교육과 박사수료(Ph.D. candidate, Department of Geography Education, Korea National University of Education)(whyunjee@gmail.com)
*** 한국교원대학교 지리교육과 교수(Professor, Department of Geography Education, Korea National University of Education)(gis@knue.ac.kr)

전달 및 여론 형성의 역할을 담당하게 되었다(박유경, 2014). 이러한 데이터의 상당수는 사용자의 위치 정보를 함께 포함하고 있으며, 사용자의 특성 및 사회적 관계에 따라 다양한 스케일의 공간 범위에서 지리적 분석을 가능하게 하였다. 즉 대규모의 자발적 지리 정보(big volunteered geographic Information)[1] 로의 소셜빅데이터를 통해 지리적 공간에 투영되는 대중들의 행태를 즉각적으로 포착해 낼 수 있다(Goodchild, 2007; 신정엽, 2014). 또한 다양한 주제에 따른 지리적 특성을 분석함으로서 온라인에서 기능하는 공간이 중요한 지리적 연구 대상이 될 수 있음을 보여준다.

본 연구는 소셜빅데이터[2] 중 하나인 트위터를 대상으로 공간 정보 데이터 구축 및 지리적 분석의 프레임워크를 정립하고 트윗 공간 데이터의 지리적 연구 방법론을 제안하는 것을 목적으로 하였다. 트윗 데이터는 사용자들의 일상, 의견, 감정, 상황 등이 기록된 텍스트적 정보와 그것이 발생한 위치적 정보가 실시간으로 수집되는 공간 데이터이다. 이러한 트윗의 내용적 특성과 트위터 사용자들의 지리적 위치와의 관련성을 확인하고 온라인 공간에서의 지역 특성을 발견하려는 방법을 연구한다.

트위터와 같은 빅 데이터의 가치는 그 속에서 중요한 정보를 추출하여 의미가 있는 정보로 변환되었을 때 발생한다. 그렇기 때문에 트윗 데이터에 활발하게 논의되는 주제들을 도출하기 위해 토픽 모델링과 같은 텍스트 마이닝 기법을 이용하였다. 토픽 모델링의 결과는 각 트윗마다 할당된 주제의 확률 분포와, 각 주제에 대한 단어 확률 분포로서 나타난다. 도출된 주제의 확률 구간별로 트윗의 비율을 산출하고, 주제에 따른 공간 분포 및 지역 특성을 확인하였다.

본 연구는 사례 지역으로 제주시를 선정하였다. 제주시는 인구 대비 트윗 수가 전국에서 가장 높은 지역으로서, 트윗 사용자의 대부분은 제주도의 거주민이 아닌 방문자로 추정된다.[3] 제주시의 트윗 중에서 GPS 좌표가 참조된 트윗을 대상으로 제주도 트윗의 내용 및 공간 분포 특성을 파악하고, 토픽 모델링 과정을 통해 도출된 주제에 따라 지리적 특성을 확인하였다.

2. 이론적 배경 및 관련 연구

본 연구의 대상인 트위터는 사용자의 생각 또는 의견을 140자 한도의 짧은 글에 담아낸 메시지인 트윗을 교환하여 사용자들이 소통하도록 하는 소셜 네트워크 서비스(Social Network Service)이자 마이크로블로그(microblog) 서비스(http://ko.wikipedia.org/wiki/트위터)이다. 이러한 짧은 트윗 텍스트 이면에는 더 많은 메타데이터가 존재한다. 트윗 관련 정보로서 트윗 고유 ID, 트윗 생성 일자, 텍스트, 작성 위치 등이 있고, 사용자 관련 정보로서 사용자 공유 ID, 팔로워 수 등을 포함한다. 사용자의 필요에 따라 위치 정보의 참조 여부를 결정할 수 있는데, 모바일의 GPS에서 부여되는 좌표를 참조하는 경우 지리학 연구에서 유용하게 활용될 수 있다(박재희, 2013).

반면 사용자-선택적인 위치정보의 포함으로 인하여 데이터의 위치 정보 획득이 중요하게 되었다. 실제 트윗에서 좌표가 참조된 위치 정보를 포함한 비율은 1% 내외이다(강애띠, 2015). 따라서 데이터에서 위치 정보의 정확도를 높이고 사용자의 위치를 유추하는 연구가 진행되고 있다. 강애띠·강영옥 등(2015)은 트윗 데이터의 위치 희박 문제의 대안으로 사용자의 거주 지역을 유추하는 방법을 고안하였다. 트위터 사용자의 이동궤적과 사용자의 언어에서 일상생활패턴을 확인하였다. 오효정 등(2014)은 트윗 텍스트의 내용을 분석하여 사용자의 연령이나 성별뿐만 아니라 사용자 선호지역을 추출하는 방법을 고안하였다.

이러한 위치 정보를 포함하는 데이터들은 각 지역별로 발생하는 트위터 사용 빈도라는 지표로서 공간적 분석이 가능하며(Sui and Goodchild, 2011; 홍일영, 2015), 각 지역별로 나타나는 트위터 내용적 특성에 따른 차이를 분석하는 연구들이 있다. Li *et al.* (2013)은 미국에서 위치가 태그된 트윗과 플리커(Flickr) 이미지를 대상으로 시공간적 분포 패턴을 비교하고 지역의 사회경제적 특성과의 관계를 파악하였다. 또한 트위터와 플리커의 공간 밀도를 종속 변수로 하여 PLSR(partial Least Squares Regression)을 적용하여 교육과 인종과 같은 요인이 영향을 미치는 것을 확인하였다. 신정엽(2014)은 트윗 데이터에 대한 논의를 정보 격차를 중심으로 고찰하였다. 미국 킹 카운티를 사례로, 트윗 데이터가 시공간에 따라 집중적으로 분포하고 도시-농촌간 정보 격차가 나타나고 있음을 탐지하였다. 또한 트윗 데이터의 분포는 사회인구학적 변수 중 젊은 층 인구, 소득 등의 변수와 일부 관련성을 가지는 것을 확인하였다.

비정형의 방대한 텍스트 자료에서 다양하게 논의되는 내용들을 파악하고 트렌드를 파악하기 해서 텍스트 마이닝 방법들이 활용된다. 그 중 토픽 모델링(topic modelling)은 텍스트 마이닝의 대표적인 기법이다. 이중에서 Blei(2003)가 제안한 LDA(Latent Dirichelt Allocation) 알고리즘은 문서에서 주제를 추론해내는 절차적 확률 분포 모델로서, 토픽 모델링 연구에 있어 표준 도구로 자리잡았다(Blei, *et al.*, 2003; 배정환 등, 2013). LDA를 적용한 지리학에서의 연구는 모델은 결과를 이용해 특정 주제에 대한 내용을 감지하는 연구(강애띠, 2016), LDA를 변형하거나 추가적 정보를 더하여 모델을 개선하여 LDA 자체의 성능을 높이는 연구(Yin *et al.*, 2011; Zhao *et al.*, 2011; Hong and Davison, 2010; 김자연, 2016) 등이 있다. 강애띠(2016)는 트위터에서 사용자가 스트레스에 대해 표현하고 있는 트윗들을 추출하여 LDA 알고리즘을 적용하여 15개의 토픽을 추출하였다. 그 토픽들을 스트레스 원인, 결과, 해소방법이라는 3가지 주제로 분류하고 지역별로 스트레스에 대한 주제와 감성이 차이가 있음을 확인하였다.

그러나 국내에서 소셜 공간 데이터를 행정 구역별로 지도화하고 공간 분포 특성을 확인하는 일부 연구 외에 이러한 데이터가 포함하는 다양한 잠재적인 주제들에 상응하는 지리적 분포와 그것 사이의 관계를 밝히려는 연구는 많지 않다. 다시 말해 공간적으로 응집되어 있는 지리적 주제는 고정된 경계와 다르게 분류될 수 있으며(Yin *et al.*, 2011), 토픽 모델링을 이용하여 유의미한 지리적 분류를 통한 새로운 지리적 공간을 탐색하려는 연구가 요구된다.

3. 트윗 데이터의 지리적 분석을 위한 방법론적 접근

1) 연구 개요

본 연구는 트윗 데이터의 위치 정보에 상응하는 트위터 사용자들의 언어적 특성을 주목하였다. 이는 유사한 내용적 정보를 공유하는 사용자의 집합은 그들의 지리적 특성을 반영하는 성향과 관련이 깊다. 이러한 사용 행태를 통해 드러나는 트위터의 지리적 공간 탐색을 분석하기 위해 다음과 같은 방법으로 연구를 진행하였다.

먼저, 트위터의 Open API를 활용하여 트윗 데이터를 수집한다. 수집된 데이터는 연구 목적에 적합한 필드를 정제하고 GIS 분석이 가능한 형식으로 변환하는 데이터 파싱(parsing)과정을 거친다. 수집된 트윗의 내용은 형태소 분석을 통해 문장을 구성하는데 직접적 영향을 미치는 자질을 추출하는 자연어 처리(NLP; Natural Language Processing)과정을 수행한다. 이렇게 구축된 데이터는 연구 목적, 분석 시점, 분석의 공간적 스케일 등을 조건으로 전처리한다. 제외어 및 포함어 리스트를 통한 트윗 필터링 – 시간 조건의 설정 – 공간 스케일의 지정의 일련의 과정을 통해 추출되는 트윗 데이터를 분석용 초기 자료로 구축

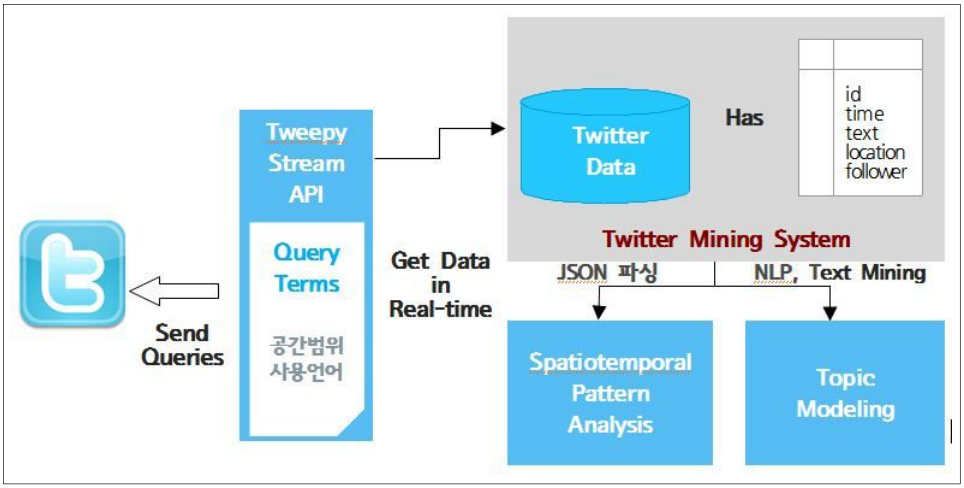

그림 1. 트윗 데이터 지리적 분석 시스템

한다. 다음으로 트윗 데이터에서 토픽을 추출하기 위해 토픽 모델링을 수행한다. 토픽 모델링의 결과는 각 트윗마다 할당된 토픽의 리스트와, 각 토픽에 대한 단어 확률 분포로서 나타난다. 트윗의 토픽별 확률 가중치에 기반하여 지리적 패턴 및 지역별 특성을 분석한다.

2) 트윗 데이터 구축

트위터는 데이터의 효율적인 수집을 위해 트위터 API를 제공하고, 다양한 프로그래밍 언어를 통해 트위터의 API에 접근할 수 있다. 본 연구에서는 트위터의 Streaming API를 통해 국내에서 실시간으로 발생하는 위치 정보가 포함된 트윗들을 수집하였다. 그렇게 하기 위해 남한 전체를 커버하는 직사각형의 공간적 범위에서 발생하는 트윗을 조건으로 데이터를 호출하였다.

3) 트윗 텍스트의 형태소 분석

수집된 트윗은 바로 내용적 분석이 불가능하기 때문에 의미 있는 텍스트의 자질을 추출하는 자연어 처리 과정이 필요하다. 트윗의 평균 글자수는 51개 남짓으로(김진만, 2015), 짧은 텍스트에서 말하고자 하는 의미를 파악하기 어렵다. 텍스트의 의미론적 기능에 기반하여 문장에서 각각의 단어를 명사, 동사, 형용사와 같은 카테고리로 품사 분류(POS Tagging)하여 실질적인 의미를 지니는 단어의 원형으로 텍스트를 변환하여야 한다.

어떤 형태소를 사용할지는 연구의 주제에 따라 결정되어야 하는데, 찬반 의견이나 감정 등을 분류하기 위해서는 명사 이외에 동사나 형용사도 함께 추출해야 분석 의도가 잘 드러난다. 이차적으로 단어의 음절 길이와 별도의 불용어 리스트에 의해 다시 한 번 형태소를 정제하여 트윗 고유 ID에 해당하는 형태소 분석 파일을 저장한다. 본 연구에서 일련의 형태소 분석 과정은 KoNLPy 모듈에서 제공하는 트위터와 같은 대용량의 문서에 대한 처리시간이 빠른 Twitter 분석기를 사용하였다.

4) 트윗 데이터 전처리(pre-processing)

데이터 구축 단계에서 1차적으로 수집한 트윗 데이터는 불필요한 데이터를 제거하고 분석에 사용될 데이터만을 추출하여 데이터의 정확도를 높이는 과정이 요구된다. 트윗 데이터 전처리의 첫 단계는 데이터 정확도에 크게 영향을 미치는 광고성 또는 악의성 트윗을 제거하는 것이다. 실제 동일한 단어나 URL을 포함한 내용의 트윗이 지속적으로 수집되는 경우, 특정 사용자가 유사한 내용으로 연속하여 트윗을 작성하는 경우, 특정 매체에서 수집한 트윗을 소수의 사용자가 독점하는 경우 등은 데이터의 신뢰성에 큰 악영향을 미친다(하수욱 등, 2012). 두 번째 단계는 분석 목적에 적합한 트윗을 추출해내는 것이다. 연구자의

트윗 ID	트윗 작성 시간	트윗 내용	사용자 ID	X (위도)	Y (경도)	장소 ID	장소명	팔로워 수	…

그림 2. 트윗 데이터의 1차적 자료 변환 결과 테이블(예시)

그림 3. 트윗 데이터의 필터링 과정 예시

관심, 연구 목적, 연구하려는 시점에 대한 조건을 설정하여 원하는 트윗을 추출한다. 세 번째 단계는 트윗 데이터에 연계된 공간 정보를 재조직화 하는 것이다. 수집된 데이터들은 시스템 상에서 GPS 기반으로 등록된 위치와 지역명을 참조한 위치로 기록된다. 본 연구에서 사용한 전처리 과정의 예시는 다음의 〈그림 3〉과 같다.

5) 트윗의 토픽 모델링

이 단계에서 사용된 텍스트 마이닝 기법은 Blei (2003)에 의해 제안된 토픽 모델링 기법 중 대표적인 LDA 알고리즘이다(Blei, *et al.*, 2003). 이 알고리즘은 문서에서 주제를 추론해내는 확률 분포 모델로서, 하나의 텍스트 내의 단어들의 사용 빈도를 확률적으로 분석하여 잠재적 변수인 주제를 학습한다. LDA는 문서가 가질 수 있는 주제들의 확률 분포와 각 주제에 대한 단어들의 확률 분포를 추론한다(조태민·이지형, 2015). 다시 말해, 전체 문서 집합에서 K개의 토픽을 추출할 수 있으며, 토픽에 포함된 단어들을 바탕으로 토픽의 의미를 유추할 수 있다.

토픽 모델링 분석을 위해서는 각 문서와 단어들의 관계를 기반으로 하는 문서-용어 행렬(Document-Term Matrix)를 구성한다. 각각의 문서들은 단어들의 빈도와 같은 특성을 나타내는 Bag-Of-Words(BOW)로 저장, 관리되는데, 이것은 '문서에 단어가 n번 존재한다'의 정보로서 표현되는 방법이다. 그리고 나서 BOW를 이용하여 문서에서 단어들의 상대적 중요도를 나타내는 TF-IDF(Term Frequency-Inverse Document) 가중치를 계산한다(원진영·김대곤, 2014). 이는 문서의 핵심 단어를 추출하고, 문서군에서 공통적으로 등장하는 단어의 중요도를 낮추는 과정이다. TF는 문서에서 특정 단어가 출현한 빈도로서 이 값이 높다면 중요도가 높다고 간주한다. IDF는 DF의 역수로서 DF는 특정 단어가 등장하는 문서의 수로 이 값이 높다면 중요도가 낮다고 간주되므로 IDF가 클수록 문서에서 잘 등장하지 않는 단어를 의미한다.

마지막으로 TF-IDF 가중치를 통해 불용어를 제거하는데, 문서 집합 내에서 단어의 최소 출현횟수가 일정 값 이하인 경우 단어의 정보를 학습할 맥락(context)이 부족하므로 제외하고, DF가 높은 단어들은 자주 사용되어 노이즈로 작용하므로 제외한다. 그렇지만 단순히 이런 방식으로만 처리하지 않고, 독립적으로 의미를 지니지 못하는 형식 형태소들은 별도의 불용어 리스트를 작성하여 처리하는 것이 토픽 모델링 결과의 정확성을 높일 수 있다. 토픽 모델링 과정의 각 단계마다 별도의 파일로 저장한다.

6) 토픽 모델링 결과 해석

토픽 모델링의 결과로서 최종적인 말뭉치(corpus)[4]로부터 문서마다 할당된 주제의 리스트가 생성되고 리스트 형태는 [주제 ID, 주제 확률]로 표현된다. 각 주제들은 단어에 대한 다차원 분포로서 각 주제의 단어마다 [단어 ID, 단어 확률]의 리스트가 생성된다. 주제에서 확률이 높은 단어는 낮은 확률의 단어보다 주제와 더 관련이 있고, 높은 가중치를 갖는 단어 순으로 주제를 요약하는 방법이 일반적이다(Coelho and Richert, 2015). 이 과정에서 연구자가 적절한 주제의 수를 조절하여야 하는데, 문서의 주제와 주제에 대한 개념 구조를 잘 이해할 필요가 있다(박자현·송민, 2013).

도출된 주제의 해석을 위해 주제 구성 단어들의 의미적 연관성을 고려하여 주제의 내용을 유추하고 토픽명을 부여한다. 토픽에서 높은 확률을 갖는 단어를 중심으로 주제를 요약하는데, 단순한 단어의 배열만으로 의미를 파악하기 어렵다. 그렇기 때문에 한 문서에서 동시에 출현한 키워드의 쌍을 생성하고, 키워드의 동시 출현 빈도와 단어 간의 조직적인 연결 관계를 통해 문서의 의미 구조를 유추한다. 즉 문서 내에서 연결정도 중심성이 큰 키워드와 키워드 간 군집구조를 확인하여 토픽 모델링의 결과와 비교할 수 있다(진설아 등, 2013)[5].

4. 제주도 좌표 참조 트윗의 공간 분포 및 주제별 지역 특성 분석

1) 제주도 좌표 참조 트윗의 공간 분포 분석

트윗 데이터의 공간 분포를 분석하기 위한 가장 기초적인 방법은 개별 트윗의 작성 위치를 지도상에 나타내주거나, 행정구역 등을 기반으로 트윗 작성 빈도를 단계구분도의 형태로 나타내는 것이다. 이는 트윗의 발생 빈도가 기대 이상으로 높은 지역을 탐색하려

표 1. 제주도 트윗 텍스트 내 키워드 출현 빈도(상위 20개)

제주(좌표 참조 트윗 954건)							
순위	명사	출현 빈도(건)	발생비율 (출현빈도/총트윗수)	순위	명사	출현 빈도(건)	발생비율 (출현빈도/총트윗수)
1	제주	178	18.7%	11	국제공항	20	2.1%
2	제주도	65	6.8%	12	굿바이	18	1.9%
3	제주시	54	5.7%	13	하우스	16	1.7%
4	특별자치도	54	5.7%	14	유채꽃	16	1.7%
5	집	49	5.1%	15	커피	16	1.7%
6	벚꽃	35	3.7%	16	공항	15	1.6%
7	카페	30	3.1%	17	꽃	15	1.6%
8	섬	23	2.4%	18	성산일출봉	13	1.4%
9	우도	22	2.3%	19	녹산(로)	13	1.4%
10	서귀포시	21	2.2%	20	수월봉	12	1.3%

는 것이다. 이러한 방법은 트윗 데이터의 전반적인 시공간 분포의 패턴을 파악하는데 용이하다.

이 장에서는 제주도에서 발생된 GPS 좌표가 참조된 트윗(geotweet)을 대상으로 트윗 데이터의 공간 분포 패턴을 분석하였다. 분석에 앞서 트윗을 작성하려는 의도에 따라 위치 정보의 참조 유형이 달라질 것으로 가정하고, 관광객이 많은 제주도는 그들의 정확한 좌표를 참조하여 방문 장소와 관광 행태를 드러내는 트윗을 작성할 것으로 예상하였다. 시기적으로 국회위원 총선과 봄꽃의 개화 시점인 2016년 4월 1일에서 4월 8일까지의 제주도 좌표 참조 트윗 954건을 대상으로 키워드의 출현 빈도를 분석하여 어떤 내용이 주로 언급되었는지 분석하였다. 좌표 참조 트윗의 경우에는 지명 또는 장소명이 많이 출현하였는데, 이는 좌표를 첨부하여 자신의 위치를 드러내고자 하는 의도가 반영된 것으로 해석하였다(표 1 참조).

제주도의 트윗 분포의 공간적 밀도를 파악하기 위해 커널 밀도 분석을 수행하였다. 〈그림 4〉에서 보는

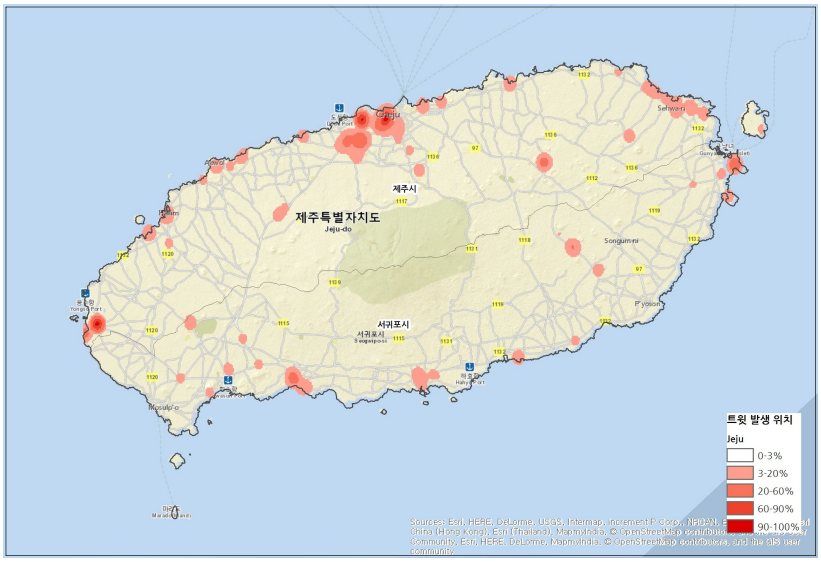

그림 4. 제주도 트윗 밀도

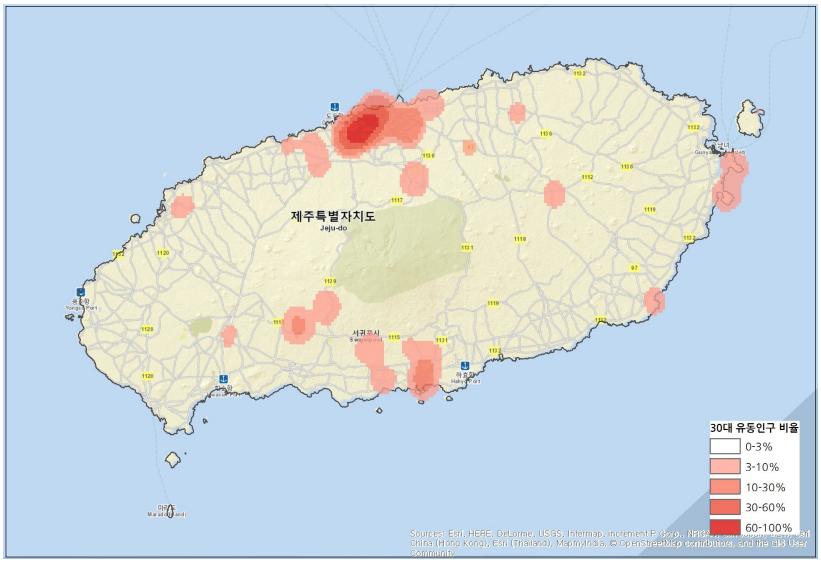

그림 5. 제주도 30대 남자 유동인구 밀도

것처럼 제주공항 주변 지역 및 서귀포 시청 인근 지역, 수월봉, 섭지코지 등의 관광지가 위치한 해안을 따라 핫스팟이 분포하였다. 제주도 총 49,112개의 지점에서의 유동인구 비율을 가중치로 하여 유동인구 밀도의 핫스팟을 확인하면, 제주 공항이 가장 높고, 제주시의 노형동, 연동 주변지역 및 서귀포시 월드컵 경기장 인근과 서귀포시 성산읍 인근에서 나타났다(그림 5 참조). 결과적으로, 트윗 밀도의 핫스팟과 유동인구 비율의 핫스팟은 유사한 지역적 패턴을 보이는 것을 확인할 수 있었다. 이러한 현상은 다른 선행연구에서도 확인된 바 있는데(구자용, 2015), 특히 제주도는 관광객의 유동인구가 많은 장소가 트윗 생성량이 많다는 것을 보여준다.

2) 좌표 참조 트윗의 토픽 모델링 분석

트윗 공간 데이터에 잠재하는 다양한 주제들은 그것의 지리적 분포와 관련이 있다(Yin *et al.*, 2011). 그렇게 때문에 지리적 분포에 따라 관심 주제를 나타내는 것이 가능하다.이 장은 트윗 데이터에서 주제를 발견하고 주제에 따른 지리적 분포 패턴을 분석하였다. 사용된 데이터는 2016년 7월 5일 정오부터 9월 14일 정오까지 수집된 4093개의 제주도 GPS 좌표 참조 트윗이다. 트윗의 공간 분포를 분석한 데이터와 다른 데이터를 사용한 이유는 분석을 수행한 기간이 다르고, 토픽 모델링 분석은 충분한 데이터의 양이 확보되어야 하기 때문에 장기간 데이터 수집이 요구되었기 때문이다. 앞 장에서 제주도의 좌표 참조 트윗은 주로 관광객이 작성하여 지명 및 장소에 대한 내용의 비율이 높다는 것을 확인하였다. 즉, 제주도 트윗에서 추출되는 주제의 내용은 관광 목적, 관광지 특성, 여행자의 행태를 반영할 것으로 가정하였다.

그렇다면 관광객들이 어떤 목적과 내용으로 트윗을 작성하며 트윗이 발생한 장소와 어떤 관련이 있는지 확인하기 위해 토픽 모델링 분석을 위한 TF-IDF 행렬을 작성하였다. 최종적으로 출현빈도가 높은 '제주(1323회)', '제주도(508회)', '특별자치도(418회)'의 단어와 출현 횟수가 10회 미만으로 모델링 학습을 위한 맥락이 부족한 다수의 단어들을 제거하고 최종적 코퍼스를 생성하였다.

그 다음 문서 집합에서 추출될 토픽의 수를 조절하면서 토픽에 할당되는 단어의 확률과 단어 간의 관계를 파악하였다. 5개의 토픽 수에서 순차적으로 값을 늘려가면서 모델링 결과를 확인하여 15개의 토픽과 토픽별로 확률 값이 높은 상위 10개의 단어를 추출하였다.

토픽에 속한 단어들의 확률적 배열을 토대로 토픽명을 유추하는 방법이 일반적인데, 이러한 방식으로 토픽의 내용을 파악하기가 쉽지 않다. 그렇기 때문에 각 트윗에서 특정 토픽의 확률 가중치가 0.7 이상인

표 2. 토픽 모델링 결과(제주 좌표 참조 트윗 4903건 대상)

topic #1		topic #2		topic #3		topic #4		topic #5	
word	score	word	score	word	score	word	score	word	score
세상	0.177	서귀포시	0.632	선물	0.104	함덕	0.121	올레	0.155
토요일	0.127	바다	0.103	사진	0.076	커피	0.117	민박	0.139
바람	0.071	국수	0.024	환상	0.073	서우	0.111	국제공항	0.120
날씨	0.067	해변	0.023	게스트	0.071	해변	0.083	제주시	0.075
인생	0.037	쇠소깍	0.015	이름	0.069	생각	0.079	센터	0.060
하늘	0.024	시작	0.012	게스트하우스	0.060	파크	0.067	여행자	0.052
과물	0.023	여름	0.008	주신	0.057	코스	0.063	카페	0.048
푸른	0.023	고기국수	0.008	카페	0.057	올레	0.058	매일	0.017
힐링	0.022	관광	0.007	여행	0.041	카페	0.041	한잔	0.016
해안	0.022	자매	0.006	아침	0.041	망고	0.015	협재해수욕장	0.015
topic #6		topic #7		topic #8		topic #9		topic #10	
word	score	word	score	word	score	word	score	word	score
가을	0.287	기분	0.084	터미널	0.152	공항	0.377	돌집	0.133
마음	0.079	오픈	0.066	도착	0.117	오름	0.099	테디베어	0.132
해수욕장	0.075	예약	0.059	근처	0.116	가족	0.096	지엄	0.125
사진	0.074	이야기	0.055	제주시	0.094	성산읍	0.085	맥주	0.067
자연	0.068	노을	0.054	국제공항	0.082	감동	0.059	라면	0.027
여름	0.042	서울	0.035	마음	0.080	호텔	0.040	해물	0.025
함덕	0.020	동쪽	0.034	점심	0.029	스낵	0.018	제주시	0.022
오후	0.015	마을	0.033	느낌	0.024	숲길	0.016	음식	0.020
협재	0.014	박물관	0.028	한림	0.020	사려	0.015	구경	0.017
롯데	0.013	애월	0.026	산방산	0.017	차귀도	0.012	우도	0.016
topic #11		topic #12		topic #13		topic #14		topic #15	
word	score	word	score	word	score	word	score	word	score
김녕	0.086	정리	0.189	공원	0.144	제주시	0.150	기념	0.063
전복	0.069	금능	0.166	안녕	0.085	잔잔	0.065	돼지	0.058
공원	0.065	대한민국	0.127	제주시	0.054	우도	0.058	메뉴	0.046
식당	0.061	휴식	0.048	가격	0.047	서귀포	0.054	도로	0.035
남자	0.060	아쿠아플라넷	0.016	성산일출봉	0.037	사랑	0.032	애월	0.022
백퍼	0.057	한화	0.014	다방	0.035	모습	0.023	몽상	0.021
맛집	0.041	친구	0.014	문화	0.027	하늘	0.021	카페	0.018
돼지	0.032	캐빈	0.010	국제공항	0.024	정상	0.012	리조트	0.011
섭지코지	0.022	동문	0.010	고양이	0.016	우동	0.012	비치	0.007

트윗의 내용을 참조하였다. 다시 말해, 트윗의 내용이 70% 이상 특정 토픽을 대변한다고 해석할 수 있는 트윗의 내용으로부터 토픽을 해석하였다. 보조적으로 동시 발생 키워드 네트워크 분석을 통해 군집으로 확인되는 단어들의 연결 관계와 구조를 참조하였다.

각각의 트윗에서 가장 높은 확률 가중치를 갖는 토픽을 그 트윗의 대표 토픽으로 간주하고, 각 토픽에 해당하는 트윗 수와 토픽의 확률 가중치를 4개의 구간으로 나누어 구간별로 해당 트윗의 수를 집계하고 이러한 값을 비율로 환산하였다. 전체 토픽 중에서 토픽 1은 해당 트윗 수 1157건, 트윗 비율 28.3%로 전체 트윗에서 가장 비중이 높은 토픽으로 확인되었다. 그러나 토픽 1에 해당된 트윗의 85% 정도가 토픽 확률 가중치 0.3 이하 구간에 분포한다. 그렇기 때문에 토픽 1의 경우는 어떤 주제에 특화되어 있다고 할 수 없다. 반면 토픽 4와 같은 경우는 해당 트윗 수 216건, 트윗 비율 5.3%의 토픽으로, 해당 트윗의 36% 정도가 토픽 확률 가중치 0.7을 초과하는 구간에 포함되어 있다. 이러한 토픽은 해당 트윗들이 더욱 해당 주제를 잘 반영하고 있고, 이들 트윗의 내용을 확인하면 토픽

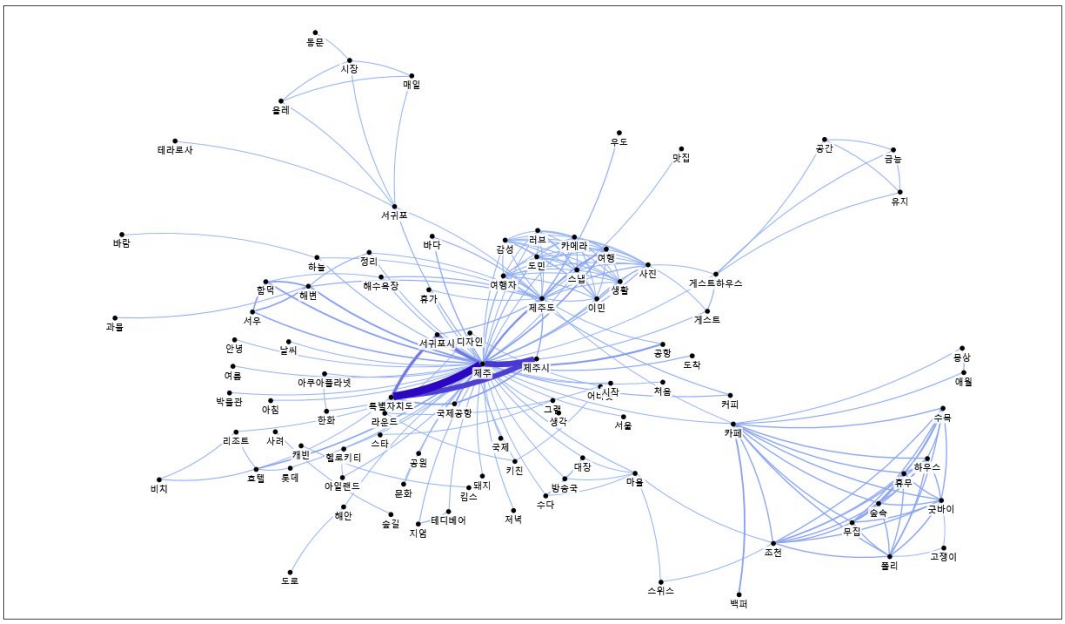

그림 6. 제주도 좌표 참조 트윗의 동시 출현 단어 네트워크

의 의미를 쉽게 유추할 수 있다.

실제로 토픽 4의 경우는 '함덕', '커피', '서우', '해변', '카페' 순으로 단어가 분포되어 있는데, 토픽의 확률 가중치가 0.7 이상인 트윗의 내용을 확인한 결과, '함덕 서우봉 해변'과 관련된 내용이 포함된 트윗이 주로 등장하였다. 토픽의 확률 가중치가 0.5~0.7 구간에서는 '바람 커피', '커피 동굴'과 같은 '커피' 관련 내용이 다수 언급되었다. 이러한 키워드 간 연결고리는 '함덕 해변 스노클링 후 커피 한 잔', '황홀한 함덕 일몰. #커피동굴 #출장커피 #제주사진한장 #바람커피

표 3. 각 토픽의 확률 가중치에 따른 구간별 트윗 수 및 트윗 비율

토픽 구분	트윗 수(건)					트윗 비율(%)				
	ALL	≦0.3	≦0.5	≦0.7	≦1.0	ALL	≦0.3	≦0.5	≦0.7	≦1.0
Topic #1	1157	978	53	123	3	28.3	84.5	4.6	10.6	0.3
Topic #2	276	7	63	200	6	6.7	2.5	22.8	72.5	2.2
Topic #3	276	14	55	181	25	6.7	5.1	19.9	65.6	9.1
Topic #4	216	8	45	86	78	5.3	3.7	20.8	39.8	36.1
Topic #5	182	12	106	48	16	4.4	6.6	58.2	26.4	8.8
Topic #6	225	10	76	113	25	5.5	4.4	33.8	50.2	11.1
Topic #7	230	15	63	131	21	5.6	6.5	27.4	57.0	9.1
Topic #8	105	5	34	58	8	2.6	4.8	32.4	55.2	7.6
Topic #9	157	3	20	133	1	3.8	1.9	12.7	84.7	0.6
Topic #10	123	7	29	84	3	3.0	5.7	23.6	68.3	2.4
Topic #11	260	8	45	186	21	6.4	3.1	17.3	71.5	8.1
Topic #12	180	7	46	123	4	4.4	3.9	25.6	68.3	2.2
Topic #13	187	15	40	122	10	4.6	8.0	21.4	65.2	5.3
Topic #14	158	8	21	124	6	3.9	5.1	13.3	78.5	3.8
Topic #15	361	5	35	313	8	8.8	1.4	9.7	86.7	2.2
합계	4093	1102	731	2025	235	100				

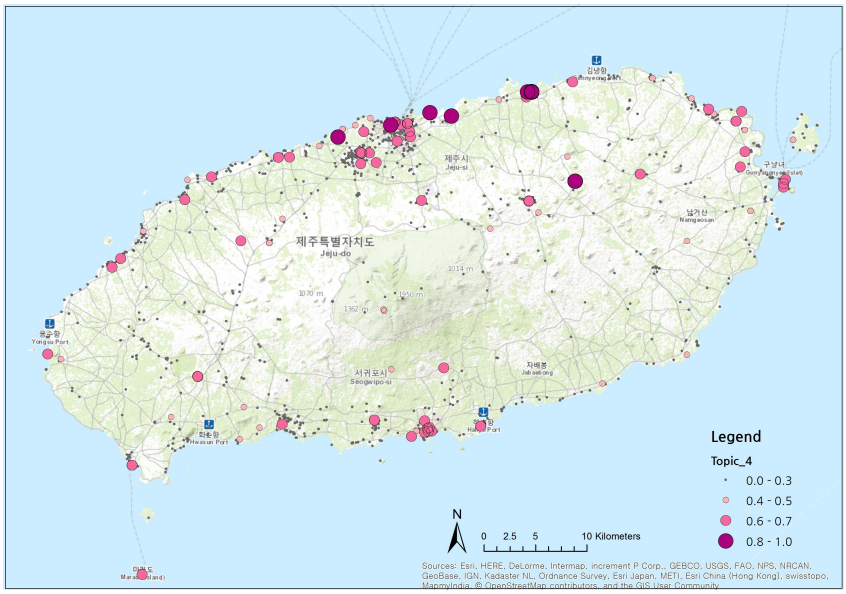

그림 7. 토픽 4의 확률 구간별 트윗의 분포 예시

@ 제주 함덕 서우봉 해변'과 같은 내용의 트윗으로 인해 같은 토픽으로 할당되었다고 여겨진다. 동시 출현 단어 네트워크에서도 토픽 4에 할당된 '함덕', '서우', '해변'의 관계성을 확인할 수 있다. 즉, 토픽 4는 '함덕 서우봉 해변'과 '커피' 관련 주제로 함축된다.

트윗에서 토픽 4에 할당된 토픽의 확률 값의 구간별로 지리적 분포를 확인해 보면 높은 확률 구간의 트윗들은 주제를 대표하는 지명 또는 위치 상에 분포하고 있는 것을 확인할 수 있다(그림 7 참조). 결과적으로 트윗에서 특정 값 이상의 가장 높은 확률 가중치가 할당된 토픽이 그 트윗을 대표하는 토픽이 되고, 이러한 트윗의 분포는 토픽이 설명하는 지리적 위치와 더욱 관련이 있었다. 그렇기 때문에 트윗 데이터의 토픽 모델링을 통해 트윗 내용에 상응하는 지역 분포와의 관련성을 직관적으로 확인할 수 있었다. 이것은 방대한 소셜빅데이터의 주제 및 내용에 따른 지리적 분석이 가능하다는 의미이다. 부가적으로 토픽에 할당된 단어들의 네트워크에서의 키워드 간의 연결 구조를 확인하면 토픽 모델링과 상호 보완적인 기능을 수행하여 보다 깊이 있는 토픽 분석이 이루어 질 수 있다.

5. 요약 및 결론

본 연구는 대중들의 생각의 속도가 실시간으로 반영되는 소셜빅데이터인 트위터를 대상으로, 그것의 발생하는 위치에 근거하여 내용적 특성에 따른 지역 특성을 파악하고자 하였다. 이를 위해 트윗 공간데이터의 지리적 분석 방법론의 프레임워크를 정립하고, 트윗 데이터의 토픽 모델링을 통해 트윗의 주제에 따른 공간 분포를 확인하였다. 사례 연구로서 이러한 분석 방법론을 적용하여 제주도를 대상으로 지역적 특성을 분석하였다.

먼저, 트윗 공간 데이터의 지리적 분석 방법의 절차는 다음과 같다. 트위터의 Open API를 활용하여 실시간으로 트윗 데이터 호출·수집한다. 수집된 데이터는 연구에 사용될 필드를 정제하고 데이터 파싱(parsing)과정을 통해 분석이 가능하도록 변환한다. 트윗의 내용은 형태소 분석을 통해 문장의 의미에 직접적인 자질만을 추출한다. 이렇게 구축된 데이터는 제외어 및 포함어 리스트를 통한 필터링 – 시간 조건의 설정 – 공간 스케일의 지정의 일련의 전처리 과정을 통해 분석용 초기 자료로 구축한다. 다음으로 트윗 데이터에서 주제를 추출하기 위해 문서와 단어 간의 TF-IDF 행렬을 구성하고 공통어를 포함한 불용어를 제거한다. 마지막으로 토픽 모델링을 통해 도출된 트윗의 토픽별 확률 가중치에 기반하여 지리적 패턴 및 지역별 특성을 분석한다.

위의 과정에 근거하여 제주도에서 발생한 좌표 참

조 트윗을 대상으로 공간 분포와 토픽에 따른 지리적 분포 패턴을 확인하였다. 첫 번째 좌표 참조 트윗을 작성하는 목적을 분석하기 위해 키워드 출현 빈도를 확인하였는데 지명 또는 장소명이 많이 출현하였다. 이는 좌표를 첨부하여 자신의 위치를 드러내고자 하는 의도가 반영된 것이다. 반면 지역명 참조 트윗의 경우 정치적 키워드의 출현 빈도가 높았는데, 이런 경우 트윗은 사용자의 사회적 관심사에 관한 내용에 편향적인 것으로 해석하였다. 두 번째, 제주도 트윗의 분포의 공간적 밀도는 제주공항을 중심으로 해안의 관광지를 따라 핫스팟이 나타났다. 이것은 제주도 유동인구의 핫스팟과 유사한 패턴을 보였다. 세 번째, 제주도 트윗의 토픽 모델링 분석 결과, 트윗에서 특정 값 이상의 높은 확률 가중치가 할당된 토픽이 그 트윗의 대표 토픽으로서, 이러한 트윗의 분포는 토픽이 설명하는 지리적 위치 및 내용과 더욱 관련이 있었다. 그렇기 때문에 토픽 모델링을 통해 트윗 데이터의 내용에 상응하는 지역 분포와의 관련성을 직관적으로 확인하는데 유용하게 활용될 수 있다는 점을 확인하였다. 이는 좌표 참조된 트윗을 올린 사용자의 시기적인 관심이나 특정 장소에 대한 선호를 토픽 모델링을 통해 직관적으로 파악하고, 그들의 위치에 대한 분포나 밀도를 확인하여 특정 지역에 대한 정보의 수집이나 마케팅의 수단으로 활용될 수 있다.

마지막으로 본 연구 내용을 바탕으로 시급하게 연구되어야 할 내용으로는 트윗 데이터의 신뢰성과 정확성을 높이는 기법 개발이 필요하다. 예를 들어 필터링 과정에서 더욱 세분화된 조건을 주어 불필요한 트윗을 제거할 수 있는 기법이나 알고리즘이 개발되어야 한다. 또한 지역성을 잘 드러낼 수 있는 주제를 선정하여 트윗 데이터의 공간적 구조에 대해 보다 심도 있는 연구가 필요하다.

주

1) 자발적 지리정보(Volunteered Geographic Information; VGI)는 주로 정부에 의해 생성·관리되는 전통적인 지리 정보와는 달리, IT 환경에서 다양한 유형의 사용자 및 기관들에 의해 자발적으로 그들이 원하는 시간과 장소에서 쉽게 생성·공유된다는 특징이 있다(Goodchild, 2007; 신정엽, 2014).
2) 소셜 미디어가 생성하는 빅데이터에 대한 정의는 학계에서 구체적으로 언급하고 있지는 않지만, 소셜 미디어에 의해 생성된 정형과 비정형의 데이터가 혼합된 데이터를 소셜빅데이터라 할 수 있다(최선화·배병걸, 2013).
3) 연구자의 분석 결과, 제주도의 인구만명당 트윗수는 8.6건으로 전국 시도 대비 가장 높았다. 서울시는 전체 트윗 데이터의 52.9%가 발생하였지만 인구만명당 트윗수가 6.8건으로 2위를 차지하였다.
4) 말뭉치(corpus)는 구조화된 텍스트 집합이다. 자연어 연구를 위해 언어 표본을 추출하여 언어의 빈도와 분포를 확인할 수 있는 자료이며(https://ko.wikipedia.org/wiki/말뭉치), 데이터 마이닝 절차 중 정제, 통합, 선택, 변환의 과정을 거친 구조화된 단계로서 어떤 추가적 절차 없이 실험에 사용할 수 있는 상태를 의미한다(https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf; 강애띠, 2016 재인용).
5) 일련의 토픽 모델링 과정은 파이썬 기반의 'Gensim' 모듈을 사용하고, 'NodeXL'로 동시 출현 단어의 네트워크를 확인하였다.

문 헌

강애띠, 2016, 트윗에서 추출한 스트레스 감성과 토픽의 공간적 특성 연구, 이화여자대학교 박사학위논문.

강애띠·강영옥, 2015, 타임라인데이터를 이용한 트위터 사용자의 거주 지역 유추방법, 한국공간정보학회지, 23(2), 69-81.

구자용, 2015, 공간정보 빅 데이터의 지도화와 공간적 분포 특성에 관한 연구 –서울시 지역의 트윗 데이터를 사례로, 국토지리학회지, 49(3), 349-360.

김자연, 2016, Multi-depth LDA 모델을 이용한 마이크로블로그에서의 토픽분석, 고려대학교 석사학위논문.

김진만, 2015, 소셜네트워크서비스에서 추출된 감정의 핫스팟 분석: 한국어 트위터와 하둡에코시스템을 중심으로, 상명대학교 박사학위논문.

박유경, 2014, 공공기관 소셜미디어의 메타데이터 표준요소 분석, 경북대학교 석사학위논문.

박자현·송민, 2013, 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석, 정보관리학회지, 30(1), 7-32.

박재희, 2013, SNS데이터의 도시정책지표로서의 활용 가능성 연구: 트윗 데이터의 주거환경 만족에 대한 공간적 특성, 이화여자대학교 석사학위논문.

배정환·손지은·송민, 2013, 텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석, 지능정보연구, 19(3), 141-156.

신정엽, 2014, 정보 격차의 맥락에서 트윗 데이터의 이론적 고찰과 실증적 공간 탐색: 미국 킹 카운티를 사례로, 한국지도학회지, 14(2), 89-106.

오효정·윤보현·최남현·유철중·김용, 2014, 소셜 빅데이터 내용 분석 기반 사용자 그룹별 선호지역 및 이동패턴 시각화, 한국정보기술학회지, 12(12), 195-203.

원진영·김대곤, 2014, 텍스트마이닝을 활용한 사회위험 이슈 도출, 한국위기관리논집, 10(7), 33-52.

이병혁·이기현·윤지영, 2005, IT와 공간구조의 재구성, 정보통신정책연구원.

전철욱(편역), 2016, Building Machine Learning Systems with Python 한국어판(개정판), 에이콘, 서울(Coelho, L. P. and Richert, W., 2013, *Building Machine Learning Systems with Python*, Packt Publishing, Ltd., Birmingham, UK)

조태민·이지형, 2015, LDA모델을 이용한 잠재 키워드 추출, 한국지능시스템학회 논문집, 25(2), 180-185.

진설아·허고은·정유경·송민, 2013, 트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구, 정보관리학회지, 30(1), 285-302.

최선화·배병걸, 2013, 소셜 빅데이터 재난관리 운영방안 및 이슈 탐지기법 연구, 국립재난안전연구원.

하수욱·남광우·류근호, 2012, 마이크로 블로그기반의 공간 지식 추출 기법연구, 한국공간정보학회지, 20(2), 129-136.

홍일영, 2015, 국내 지오트윗의 공간분포, 한국지도학회지, 15(2), 93-101.

Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003, Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3, 993-1022.

Hong, L., and Davison, B. D., 2010, Empirical study of topic modeling in twitter, *Proceedings of the First Workshop on Social Media Analytics(SOMA)*, 80-88.

Li, L., Goodchild, M. F., and Xu, B., 2013, Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr, *Cartography and Geographic Information Science*, 40(2), 61-77.

Sui, D., and Goodchild, M., 2011, The convertgence of GIS and social media: challenges for GIScience, *International Journal of Geographical Information Science*, 29(1), 1737-1748.

Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T., 2011, Geographical topic discovery and comparison, *Proceedings of the 20th International Conference on World Wide Web*, 247-256.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., and Li, X., 2011, Comparing twitter and traditional media using topic models, *European Conference on Information Retrieval*, Springer Berlin Heidelberg, 338-349.

• **교신** : 김영훈, 28173, 충북 청주시 흥덕구 강내면 태성탑연로 250, 한국교원대학교 지리교육과(이메일: gis@knue.ac.kr, 전화: 043-230-3641)

Correspondence : Kim, Young-Hoon, 28173, Department of Geography Education, Korea National University of Education, 250, Taeseongtabyeon-ro, Heungdeok-gu, Cheongju-si, Chungbuk, Korea(E-mail: gis@knue.ac.kr, Tel: +82-43-230-3641)

(접수: 2017.04.22, 수정: 2017.05.15, 채택: 2017.05.20)