Korean Journal of Sport Science
2019, Vol. 30. No. 3, 501-512.
https://doi.org/10.24985/kjss.2019.30.3.501

체육과학연구

Introduction of topic modeling for extracting potential information from unstructured text data: Issue analysis on news article of dementia-related physical activity

Hyo-Jun Yun, Jae-Hyeon Park, & Jiwun Yoon*

Korea National Sport University

[Purpose] The purpose of this study is to introduce the basic concepts and procedures for topic modeling and to explain topic modeling to news articles about dementia-related physical activities. And it is also to discuss the possibility of using topic Modeling in the field of physical education. **[Methods]** In this study, the LDA algorithm of topic modeling is explained and the analysis procedure is summarized step by step by text preprocessing, text formatting, and topic number determination. The application cases were selected from 274 news articles about dementia-related physical activities reported in 13 major daily newspapers from 2000 to 2018. **[Results]** When the number of topics is 3, the Coherence Score figure is the highest. Topic 1 is about welfare services for dementia patients, Topic 2 is about prevention of dementia, and Topic 3 is about dementia research. The ratio by each subject is Topic 2 (46.0%), Topic 3 (33.2%) and Topic 1 (20.8%) in order of high ratio. **[Conclusion]** Topic modeling is an effective methodology to extract potential information excluding subjectivity of researchers. It is expected to be used when searching for information in massive texts in the field of physical education.

Key words: Textmining, Topic modeling, LDA algorithm

서 론

비정형 데이터란 다양한 형태의 데이터 중에서 문자, 그림, 영상 등의 형태로 생성되는 데이터를 의미한다. 최근 스마트폰의 대중화와 인터넷의 발달로 전자신문의 기사, 소셜미디어, 유튜브 등에서 생성되는 비정형 데이터의 수는 헤아릴 수가 없을 정도이다. 이에 데이터 분석의 트렌드는 정형데이터에서 비정형 데이터로 옮겨가고 있는 경향이 뚜렷하다. 전 세계적으로 생산되는 일일 데이터의 양은 약 2.3조 기가바이트에 이르며 이 중 80%가 비정형 데이터라는 IBM 보고(Bae, 2019)가 이를 증명한다.

비정형 데이터 중에서 '텍스트 정보를 어떻게 분석하고 이해해야 할 것인가?'는 데이터사이언스 분야에서 관심 있게 다루고 있는 주제 중 하나이다(DiMaggio et al., 2013). 뉴스, 블로그, SNS 등 텍스트 정보가 다루는 특정 주제에 대하여 하루에도 수만 건 이상 쏟아지는 비정형 데이터 문서가 담고 있는 의미(semantics)를 분석해 내는 작업이기 때문(Park, 2016)이다. 예컨대 뉴스 기사에서 다루는 텍스트는 특정 문제에 대한 사회적 관점을

논문 투고일 : 2019. 04. 26.
논문 수정일 : 2019. 06. 15.
게재 확정일 : 2019. 07. 12.
* 교신저자 : 윤지운(woona80@gmail.com).
* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A5B5A02025759)

담고 있으며(Iyengar, 1994), 트위터, 페이스북, 기사 댓글 등의 텍스트는 대중들의 생각과 경험 등이 내포되어 있다(Kim & Lee, 2014; Rapp et al., 2013). 따라서 비정형 텍스트 분석은 뉴스기사를 통해 특정 사회현상에 대한 의미를 부여(Crossman et al., 2007)할 수 있으며 트위트, 페이스북, 기사댓글 등을 통해 대중들의 생각과 사회적 요구를 파악하거나 특정 상품에 대한 품평을 확인(Myung et al., 2008)하여 기업의 비즈니스 모델개발에 활용하기도 한다(Hahm & Lee, 2016).

비정형 텍스트 분석에 대한 중요성이 강조되면서 다양한 분야에서 데이터마이닝을 적용한 연구들이 적지 않게 수행되고 있다(Kang & Kang, 2018; Kim & Kim, 2016; Newman et al., 2006). 비정형 텍스트 자료로부터 토픽을 추출하기 위한 시스템 설계 및 구현(Park, 2016), 질적 연구의 내용분석을 위한 의미연결망(Yoon & Park, 2015) 등 텍스트 자료 분석 방법론적 측면에서의 다루고 있는 연구를 찾을 수 있고, SNS 상품리뷰 분석(Kim & Kim, 2016), 신문기사를 통한 기술트렌트 분석(Kang et al., 2013) 등 텍스트 자료의 출처에 따른 연구도 찾아볼 수 있다.

비정형 텍스트 문서에서 텍스트가 가진 주요한 의미를 추출하려는 방법으로는 의미연결망과 토픽모델링 분석이 대표적이다. 의미연결망은 네트워크 이론에 기반을 둔 분석방법으로써 노드(node)로 규정하는 개별 단어 그리고 개별 단어와 단어 간의 관계(relationship)를 기반으로 네트워크 구조에서 단어 간의 상호연관 관계를 통해 문서가 전달하고자 하는 중요한 의미, 즉 메시지를 찾고자 한다(Kim & Jang, 2016). 의미연결망 분석에서 단어(node)들이 형성한 네트워크는 특정 의미를 내포하고 있음(Paranyushkin, 2010)을 전제하기 때문에 텍스트의 집합체에서 전달하고자 하는 의미를 찾아낼 수가 있다(Park & Chung, 2013)는 것이다. 특히 의미연결망 분석은 텍스트로부터 단어를 추출하고 단어의 연결 관계를 근거로 주요 의미를 추출하기 때문에 의미 해석과정에 객관성을 확보할 수 있다는 장점이 있다. 이에 따라 텍스트에서 숨겨진 의미파악을 목적으로 하는 연구에서 주로 활용되고 있다. 그러나 의미연결망 분석에서 단어를 추출하고 정제하는 과정에서 연구자의 주관성이 개입된다는 제한(Chung et al., 2013; Yoon & Park, 2015)도 배제할 수는 없다.

토픽모델링은 대규모의 문서로 만들어진 원천 텍스트(original text)에서 문서가 전달하고자 하는 핵심 주제를 도출하고 주제에 상응하는 문서를 추론하여 제시하는 방법론이다(Kang et al., 2013; Park & Oh, 2017; Steyver & Griffiths, 2007). 토픽모델링은 생성·확률모델(generative probabilistic model)인 잠재디리클레할당(Latent Dirichlet Allocation; LDA) 알고리즘을 적용(Blei, 2012)한다. 특히 토픽모델링에서는 텍스트 문서집합 내 잠재적인 토픽(latent topic)들이 불규칙하게 혼합되어 있다는 가정하에, 문서상에 분포하고 있는 관측가능한 단어들의 패턴을 파악하여 토픽을 추론해낸다(Yoon & Suh, 2018). 또한, 이 과정에서 수학적이고 통계적인 방법을 활용하기 때문에 문서에서 토픽을 추출하는데 있어 연구자의 주관성은 배제된다. 따라서 의미연결망으로 텍스트 자료를 분석할 때 의미 추출과정에서 발생하는 연구자의 주관성 개입에 따른 제한점을 극복하는 것이 가능할 뿐 아니라 대량의 자료에서 잠재된 토픽(주제)을 찾아낼 수 있다는 것이 장점이다(Park & Song, 2013).

국내 체육학 분야에서 토픽모델링에 관한 연구를 살펴보면, Lee & Park(2017)은 한국체육측정평가학회지의 연구 동향을 분석하면서 토픽모델링을 적용한 바 있으며, Park 등(2018)은 여가·레크리에이션 학술연구의 특징을 분석하기 위한 목적으로 토픽모델링을 활용한 바 있다. 이 연구들은 토픽 분석을 활용하여 체육학 분야에 토픽모델링을 소개한 바 있다. 그러나 토픽모델링 방법이 비정형 텍스트 자료 분석을 위해 적용된 것이 비교적 최근이며, 많은 후속연구를 유도해 낼 수 있는 유용한 분석알고리즘이라는 점에서 선행연구는 토픽모델링을 사례에 적용하여 수행하였을 뿐 튜토리얼 성격의 정보로써 활용할 수 없다는 아쉬움이 적지 않았다. 이에 토픽모델링의 이론적 틀과 분석방법에 관한 정보를 체계적으로 소개하여 후속 연구자들이 토픽모델링을 효과적으로 적용할 수 있는 활용지침서의 필요성이 제기되었다.

한편, 뉴스 기사는 매체를 통한 언론 보도의 한 형태로서 우리 사회에서 특정 현상에 대한 의미를 붙이기 위한 핵심역할(Crossman et al., 2007)을 수행한다. 이는 뉴스 기사에서 보도하고 있는 이슈들은 다양한 정책 설정에 지대한 영향을 줄 수 있음을 의미한다(Han et al., 2016). 그뿐만 아니라 뉴스 기사에서 보도를 통하여 대중

의 공감 정도와 인식을 제고(Choi & Kweon, 2014)하는데 유용하게 활용된다. 이는 뉴스 기사를 중심으로 한 텍스트 분석을 통하여 분야별 이슈를 파악하고 정책 패러다임의 변화를 확인할 수 있음을 시사한다. 따라서 이 연구에서 토픽모델링을 소개함에 있어 뉴스 기사를 대상으로 하였고 주제어는 치매 관련 신체활동으로 정하였다. 최근 치매와 신체활동에 대한 관련성은 학계 및 현장에서 주요 이슈 중 하나이다. 신체활동 참여가 치매 질병의 진행속도를 감소시킴과 동시에 뇌 기능을 강화할 수 있다고 보고(Chang et al., 2010; Hamer & Chida, 2009; Um et al. 2011)되었기 때문이다. 이 연구에서는 뉴스 기사에서 보도하는 치매 관련 신체활동에 대한 토픽을 파악하여 어떤 주제어들로 토픽들이 구성되는지를 확인할 수 있도록 계획하였다. 치매 관련 신체활동 기사에서 주요 주제를 파악하는 것은 체육학 분야의 연구 방향 설정, 체육 정책 설정을 위한 토대 구축에 주요한 역할을 할 수 있을 것이다.

따라서 이 연구는 방법론적 측면에서 토픽모델링에 대한 기본개념 및 절차를 소개하고 체육학 분야에서 적용 가능성을 살펴보는 것이 목적이다. 방법론에 적용한 사례로는 치매 관련 신체활동에 대한 뉴스 기사에 토픽모델링을 적용하여 분석함으로써 그 가능성을 확인하였다.

토픽모델링

토픽모델링의 개념

정보 검색 분야에서 문헌 내의 잠재적 의미구조를 파악하려는 시도들은 지속해서 이루어져 왔다. 특히 토픽모델링은 텍스트마이닝 기법의 하나로 비구조화된 문헌 집합에서 잠재된 토픽들을 추출하는 확률적 알고리즘이다(Blei et al., 2003). 일반적인 군집화(clustering) 기법은 하나의 문서는 하나의 토픽으로만 할당되는 반면 토픽모델링은 하나의 문서에 여러 개의 토픽이 존재할 수 있으므로 현실적으로 더 적합한 모델로 평가받고 있다(Kim et al., 2017). 또한, 내용분석방법의 한계점을 극복하고 대량의 자료에서 잠재된 토픽을 찾아낼 수 있는 점에서 유용하다(Park & Song 2013). 토픽모델링의 알고리즘은 1990년 Deerwester 등이 제안한 LSA (Latent Semantic Analysis, LSI:Latent Semantic Indexing이라고도 불림) 알고리즘이 시초라고 볼 수 있다. 이후 Hofmann(1999)는 LSA 알고리즘의 확장된 알고리즘으로 pLSA (Probabilistic Latent Semantic Analysis) 알고리즘을 제안하였다. 그러나 pLSA 알고리즘은 새로운 문서에 대한 확률을 계산할 수 없으며 많은 매개변수로 인해 수식이 복잡해질 수 있다(Blei et al., 2003; Lee & Park, 2017). 또한 훈련데이터에 지나치게 맞춰지는 과적합(overfit) 현상이 나타날 수 있는 단점이 있다(Blei et al., 2003). 이러한 단점을 보완하고자 Blei 등(2003)은 LDA (Latent Dirichlet Allocation) 알고리즘을 제안하였다. 이 알고리즘은 자료 차원 축소의 유용성과 의미적으로 일관성 있는 주제를 생성할 수 있다는 점에서 장점이 있으며(Mimno et al., 2008) 높은 성능으로 학계에서 표준으로 인식되고 있다. 따라서 이 연구에서도 LDA 알고리즘을 소개하고 치매 관련 신체활동 뉴스 기사에 적용하고자 한다.

LDA 알고리즘은 문서, 단어 등 관찰된 변수(observed variable)를 통해 문맥, 문서의 구조 등 잠재된 변수(hidden variable)를 추론하는 방법으로 전체 문서 집합의 주제, 문서별 주제 비율, 각 단어가 각 주제에 포함될 확률 등을 파악할 수 있다(Park & Oh, 2017; Park & Song, 2013). 다음 〈Fig. 1〉은 LDA 알고리즘을 그래프 표현한 것이다.

α, β는 Hidden Parameter이며, θ, z는 Hidden Variable, w는 유일한 관찰변수이다. α는 k(토픽의 수) 차원 디리클레분포(Dirichlet Distribution)의 매개변수이며, 해당 문서가 어떤 토픽 비율을 가질 것인가 결정하는 Parameter이다. 즉 θ^i는 문서가 i번째 토픽에 속할 확률분포를 나타낸다. z_n^i는 단어 w_n이 i번째 토픽에 속할 확률분포를 나타낸다. β는 $k \times V$(단어집) 크기의 행렬 매개변수, β_{ij}는 i번째 토픽이 단어집 j번째 단어를 생성할 확률을 나타낸다. 각 문서에 대해 k개의 주제에 대해 가중치 θ가 존재하며 문서 내의 각 단어 w_n은 k개의 주제에 대한 가중치 z_n을 가진다. z_n은 θ에 의한 다항분포로 선택되며 마지막으로 실제 단어 w_n이 z_n에 기반을 두어 선택된다.

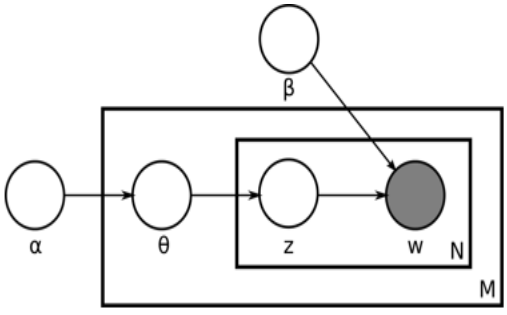

Fig. 1. Graph model of LDA algorithm

토픽모델링의 분석절차

토픽모델링을 적용하기에 앞서 텍스트 자료는 비정형 자료이기 때문에 분석에 소요되는 시간과 정확성을 높이기 위해 텍스트 전처리와 텍스트 정형화 단계를 걸쳐야 한다. 또한, 토픽모델링은 사전에 설정한 토픽의 수에 따라 산출되는 결과가 상이하기 때문에 전체 문서 집합으로부터 몇 개의 토픽이 존재하는가를 결정해야 한다. 다음 〈Fig. 2〉는 토픽모델링의 분석절차를 나타낸 것이다.

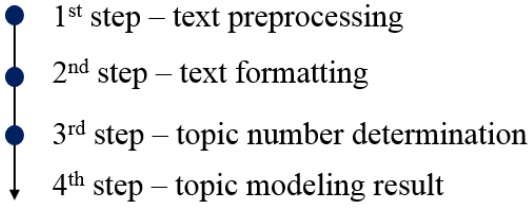

Fig. 2. Analysis procedure of topic modeling

텍스트 전처리

텍스트 자료는 정형화되지 않은 비정형 데이터로 분석에 소요되는 시간을 줄이고 정확성을 높이기 위해서는 텍스트 전처리 단계는 필수적이다(Lee & Park, 2017; Yun, 2018). 텍스트 전처리 단계의 첫 번째 단계는 토큰화(tokenization) 단계이다. 토큰화는 단어를 분리하는 단계로 사용하는 형태소에 따라 단어가 분리되는 형태가 달라질 수 있으므로 연구자는 연구의 특성에 맞게 형태소 분석기를 선택하여 사용해야 한다(Park & Oh, 2017). Python의 Konlpy 라이브러리(한국어 자연어처리 라이브러리)의 경우 5가지 형태소분석기(Kkma, Komoran, Hunnanum, Mecab, Okt(구 Twitter)를 제공하고 있다. 다음 〈Table 1〉은 형태소분석기별 토큰화 결과예시를 나타낸 표이다.

두 번째 단계는 품사추출 단계이다. 토큰화로 분리된 단어를 연구자의 판단에 따라 어떤 품사만을 추출할 것이지 결정하여야 한다. 일반적으로 주요 키워드를 도출해내는 연구(키워드분석)일 경우 명사단어만을 추출하여 적용하고 있으며 감성분석(오피니언마이닝)의 경우 명사, 동사, 형용사단어를 추출하여 사용하고 있다.

세 번째 단계는 불필요한 단어를 삭제하는 단계이다. 분석하고자 하는 자료가 치매와 관련된 자료라고 하면 '치매' 단어들 제외할 수 있으며 추출된 단어들을 전반적으로 스크린한 후 불필요한 단어들을 제외하는 단계이다.

네 번째 단계는 공통단어로 변환하는 단계이다. 이 단계는 같은 의미지만 다르게 표현된 단어들을 하나의 단어로 변환시켜주는 단계이다. 가령, '노인'과 '어르신' 단어는 같은 의미이지만 다르게 표현된 단어로 '노인'을 '어르신'으로 변환시키거나 '어르신'을 '노인'으로 변환시켜야 한다.

Table 1. Example of tokenization result by morpheme analyzer

Morpheme analyzer	Example) 중노년층의 전형적인 체형을 두고 흔히 거미형 몸매라고 한다
Kkma	〔'중', '노년층', '의', '전형적', '이', 'ㄴ', '체형', '을', '두', '고', '흔히', '거미', '형', '몸매', '라고', '하', 'ㄴ다'〕
Komoran	〔'중', '노', '년', '층', '의', '전형' '적', '이', 'ㄴ', '체형', '을', '두', '고', '흔히', '거미', '형', '몸매', 라고', '하', 'ㄴ다'〕
Hunnanum	〔'중노년층', '의', '전형적', '이', 'ㄴ', '체형', '을', '두', '고', '흔히', '거미형', '몸매', '이', '라', '고', '하', 'ㄴ다'〕
Mecab	〔'중', '노년층', '의', '전형' '적', '인', '체형', '을', '두', '고', '흔히', '거미', '형', '몸매', '라고', '한다'〕
Okt(Twitter)	〔'중', '노년', '층', '의', '전형' '적', '인', '체형', '을', '두고', '흔히', '거미', '형', '몸매', '라고', '한다'〕

Table 2. Examples of Text Formatting Methods

Example	1. The sun is shinning 2. The weather is sweet 3. The sun is shinning and the weather is sweet							
BOW	ID	and	is	shinning	sun	sweet	the	weather
	1	0	1	1	1	0	1	0
	2	0	1	0	0	1	1	1
	3	1	2	1	1	1	2	1
TF-IDF	ID	and	is	shinning	sun	sweet	the	weather
	1	0	.43	.56	.56	0	.43	0
	2	0	.43	0	0	.56	.43	.56
	3	.4	.48	.31	.31	.31	.48	.31

Raschka(2015)

다섯 번째 단계는 연속적으로 자주 사용된 단어를 하나의 단어로 변화시키는 단계이다. 가령, '신체', '활동'이라는 단어가 연속적으로 자주 등장하면 '신체_활동'으로 변환시키는 단계이다. 복합명사의 경우 형태소분석기에 따라 단어가 상이하게 분리되기 때문에 이러한 문제점을 보완할 수 있다.

텍스트 정형화

비정형 텍스트 자료를 정형화하는데 가장 보편적으로 사용되는 방법은 Harris(1954)가 제안한 BOW (Bag of Word) 방법이 있다. BOW 방법은 문서에서 단어가 몇 번 출현하였는지 빈도수에 기반으로 측정되는 방법이다. 그러나 이 방법의 문제점을 두 가지 측면으로 설명할 수 있다. 첫째, 문서의 길이는 단어 출현확률에 영향을 미친다. 예시〈Table 2〉에서 1, 2번 문서보다 3번 문서의 길이가 길기 때문에 단어가 출현할 수 있는 확률은 높다. 즉, 문서의 길이가 길어질수록 단어의 출현확률은 증가하고 문서의 길이가 짧아질수록 단어의 출현확률은 낮아질 수 있다. 둘째, 단어의 출현빈도 지나치게 클 경우 오히려 불용어 수준의 단어에 해당할 수 있다. 예시〈Table 2〉에서 'is'의 단어와 'the'의 단어는 모든 문서에 출현한 단어이며 이는 문서를 분류하는 데 있어 그 중요도는 미비하다.

이러한 문제점을 해결하기 위해서 TF-IDF 방법을 적용할 수 있다. 이 방법은 문서 내의 단어 간 상대적 중요도를 평가하기 위해 문서의 표현방식으로 고안된 방법(Lee & Kim, 2009)으로 공식은 다음 〈Table 3〉과 같다.

TF (Term Frequency)는 문서 내의 단어 출현빈도를 총 출현 횟수로 나눈 것을 의미한다. IDF (Inverse Document Frequency)는 전체 문서 수를 특정 단어가 나타난 문서의 수로 나눈 것을 의미한다. 즉 상대적으로 많은 문서에 출현한 단어의 IDF 값은 작게 산출되고 반대로 한쪽으로 편중하여 나타난 단어의 IDF 값은 크게 산출되는 것을 의미한다. 최종적으로 TF-IDF는 TF 값과 IDF 값을 곱한 것으로 TF-IDF 값이 큰 단어는 속한 문서의 주제를 결정짓는 가능성이 높아(Kim, 2018; Lee & Kim, 2009) 토픽모델링에서 토픽을 추출할 수 있는 척도로 활용할 수 있다.

Table 3. TF-IDF formula

TF	$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: The number of times that word t_i appeared in document d_j $\sum_k n_{k,j}$: The number of appearance of all the words in document d_j
IDF	$idf = \log \frac{\lvert D \rvert}{\lvert d_j \rvert t_j \in d_j \rvert}$ $\lvert D \rvert$: The number of total documents included in documents set $\lvert d_j \rvert t_j \in d_j \rvert$: The number of documents that word t_i appears
TF-IDF	$TFIDF_{i,j} = tf_{i,j} \times idf_i$

Lee & Kim(2009)

토픽 수 결정

토픽모델링은 사전에 설정한 토픽 수에 따라 산출되는 결과가 상이할 수 있으므로 전체 문서 집합으로부터 몇 개의 토픽이 존재하는가를 결정해야 한다. 토픽모델링을 내재적으로 평가하는 고전적인 방법은 Perplexity (곤란도 혹은 혼란도)를 통해 살펴보는 방법이 있다. 이 수치는 특정 확률모델이 실제 관측되는 값을 얼마나 잘 예측하는지를 평가할 때 사용되는 방법으로 수치가 낮을수록 성능이 높은 모델로 평가하는 방법이다(Blei et al., 2003). 그러나 Perplexity의 수치는 내재적으로 학습 성능 정도를 의미하고 있을 뿐 그 결과를 해석하기에 난해하다는 단점을 가지고 있다. Chang 등(2009)은 Perplexity는 항상 해석에 적절한 결과를 보이지 않는다고 보고하고 있으며 연구자가 해석하기에 적합한지를 확인하기 위해서는 다른 척도가 필요하다고 주장한 바 있다.

이러한 문제점을 해결하기 위해서 Coherence(일관성) Score를 적용하여 모델을 평가할 수 있다. 이 방법은 Newman 등(2010)에 의해 처음 제안된 방법으로 토픽모델링 결과로 산출된 각각의 주제들이 포함된 상위 단어들이 얼마나 높은 유사도를 가지느냐에 따라 평가되는 방법이다. 즉, Coherence Score가 높을수록 토픽모델링으로 산출되는 각각의 주제가 의미론적으로 유사한 단어들로 구성되었다고 해석한다.

반면, '토픽모델링에서 최적의 토픽 수를 찾아내는 통계적 해법은 존재하지 않는다'라는 주장도 있다. 이들 주장에 따르면 토픽모델링은 분류되지 않은 문서 집합들을 소수의 토픽에 분류시킴으로써 나머지 토픽들을 좀 더 해석 가능한 것으로 만들어주기 때문에 전체 토픽을 대상으로 평가하여 최적의 모델을 선택하는 것은 사실상 무의미하다는 것이다. 따라서 토픽 수의 결정은 추출된 토픽의 해석 가능성과 타당도, 연구문제에 비쳐 유용성 및 분석의 용이성 등이 중요한 기준이 되고 해당 연구영역에 대한 전문적 식견이 요구(DiMaggoi et al., 2013; Nam, 2016)된다고 주장한다.

치매 관련 신체활동 뉴스 기사 적용 예제

예제자료의 특성 및 자료수집

이 연구는 비정형 텍스트 자료에서 잠재적 정보를 추출하는 방법으로 토픽모델링을 소개하고 치매 관련 신체활동 뉴스 기사를 적용 예제로 설명하는 것이 목적이었다. 예제자료는 N 포털사이트에서 검색되는 뉴스 기사의 본문을 예제자료로 선정하였다. 검색조건은 다음과 같다. 첫째, 신체활동, 치매를 키워드로 사용하였다. 이때, 신체활동의 키워드가 정확히 일치하는 기사와 치매 키워드가 포함된 기사만을 선정하였다. 둘째, 2000년 1월 1일부터 2018년 12월 31일까지 기간을 한정하였다. 셋째, 언론사를 일간지(경향신문, 국민일보, 내일신문, 동아일보, 매일일보, 문화일보, 서울신문, 세계일보, 아시아투데이, 조선일보, 중앙일보, 한겨레, 한국일보)만을 한정하였다. 이와 같은 조건으로 검색하였을 때 총 335건의 기사가 검색되었으며, 치매, 신체활동과 관련이 없는 기사와 중복되는 기사를 제외하고 최종적으로 274건의 기사를 최종 연구자료로 선정하였다. 자료수집을 위해 python 3을 활용하여 웹 크롤링 자동화 프로그램을 개발하였다.

Table 4. Text preprocessing procedure

Preprocessing		Specific method
1st step	Tokenization	Separate words (example 특히/ 전/ 세계/적/으로/ 고령화/ 사회/로 .....)
2nd step	Only extract nouns	Only extract nouns (example 전/ 세계/ 고령화/ 사회 .....)
3rd step	Common word conversion	'가능', '가능성' → '가능', '뇌졸증', '뇌졸중' → '뇌졸중'
4th step	Stopword	Remove unnecessary words such as Local name, person name, etc.
5th step	Continuous word conversion	'신체', '활동' → '신체_활동', '건강', '검진' → '건강_검진'

텍스트 전처리 및 정형화

이 연구에서 수행한 텍스트 전처리는 5단계의 절차로 실시하였으며 구체적인 방법은 다음 〈Table 4〉와 같다.

텍스트 전처리의 1단계로 연구자료로 선정된 뉴스 기사 본문을 토큰화를 통해 단어들을 분리하였으며 이때 형태소분석기는 Okt분석기를 사용하였다. 2단계로 분리된 단어들에서 명사단어만을 추출하였으며 3단계로 연구자 판단에 따라 같은 의미지만 다른 형태로 표현된 단어들을 공통단어로 변환하였다. 가령, '알츠하이머'를 '치매'로 변환하였으며, '가능성'을 '가능'으로 변환하였다. 4단계로는 지역명, 사람명 등 분석에 불필요한 단어를 제거하였으며, '치매', '노인' 단어도 불용어로 판단하여 제거하였다. 5단계로는 전체 자료에서 5회 이상 연속적으로 관찰되는 단어를 한 단어로 변환하였다. 가령, '신체'와 '활동'이라는 단어가 연속적으로 5회 이상 나타났다면 '신체_활동'의 단어로 변환하였으며, '건강', '증진' 단어가 연속적으로 5회 이상 나타나면 '건강_증진'단어로 변환하였다.

텍스트 정형화는 앞서 설명한 TF-IDF 방법을 적용하여 텍스트 정형화를 실시하였다.

자료처리

이 연구의 모든 자료처리는 python 3을 활용하였으며 python에 사용된 연구방법별 라이브러리 Package는 다음 〈Table 5〉와 같다.

Table 5. Python package by research method

Research method	Package
Collecting data	Requests, BeatifulSoup
Text preprocessing	konlpy, gensim
TF-IDF, Coherence, LDA	gensim
Keyword frequency analysis	nltk

토픽모델링 적용

토픽모델링을 적용하기에 앞서 텍스트 전처리 과정을 걸친 최종자료를 기반으로 키워드빈도분석을 실시하였다.

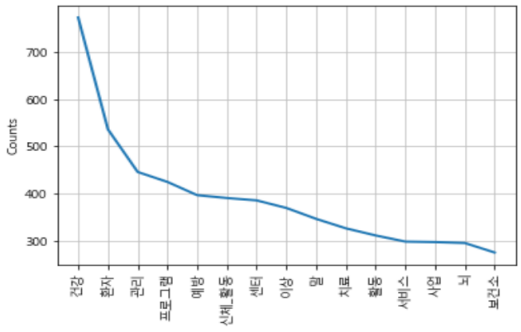

Fig. 4. Top 15 keyword(frequency)

그 결과 총 4132개의 단어가 관찰되었으며 '건강'단어가 772회로 가장 많이 나타났으며, 두 번째로 많이 나타난 단어는 '환자'로 나타났다. 이외에도 '관리' 446회, '프로그램' 425회, '예방' 397회, '신체_활동' 391회, '센터' 386회, '이상' 370회, '말' 347회, '치료' 327회 등으로 나타났다. 다음 〈Fig. 4〉는 빈도수가 높은 상위 15개의 단어를 그래프로 나타낸 것이다.

이 연구에서는 토픽모델링의 LDA 알고리즘을 적용하였으며 토픽 수는 사전에 연구자가 설정하여야 한다. 따라서 최적의 토픽 수를 탐색하기 위해 토픽 수를 2개부터 9개까지 설정하여 모델의 Coherence Score를 산출하였으며 그 결과 다음 〈Fig. 3〉과 같다.

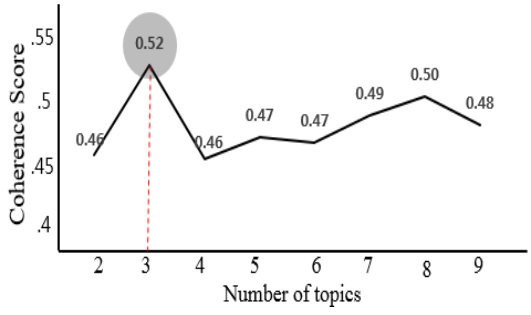

Fig. 3. Coherence score by number of topics models

토픽의 수를 3개로 설정하였을 때 Coherence Score가 가장 높게 산출되었으며, 이는 토픽 수를 3개로 설정하였을 때 각각의 토픽에 포함된 상위 단어 간의 유사도가 가장 높다고 해석할 수 있다. 따라서 이 연구에서는

토픽 수를 3개로 설정하여 분석을 실시하였다. 토픽모델링 분석결과 〈Table 5〉와 같다.

토픽별 주요 단어를 살펴보면 토픽1의 경우 '센터', '서비스', '환자', '이용', '시설', '지원', '사회', '말', '의료', '병원'의 단어가 주요 단어로 나타났으며, 이러한 결과를 비추어볼 때 토픽1은 치매 환자를 위한 복지서비스와 관련된 주제라고 판단해 볼 수 있다. 토픽2의 경우 '건강', '관리', '프로그램', '사업', '보건소', '예방', '대상', '신체_활동', '교육', '통해'의 단어가 주요 단어로 나타났으며, 이러한 결과를 비추어볼 때 토픽2는 치매 예방과 관련된 주제라고 판단해 볼 수 있다. 토픽3은 '환자', '뇌', '이상', '치료', '위험', '검사', '사람', '인지', '교수', '활동'의 단어가 주요 단어로 나타났으며, 이러한 결과를 비추어볼 때 토픽3은 치매 연구와 관련된 주제라고 판단된다. 토픽별로 분류된 문서의 수를 살펴보면 토픽2가 126건(46.0%)으로 가장 높게 나타났으며, 토픽3이 91건(33.2%)으로 두 번째로 높게 나타났다. 반면 토픽1은 57건(20.8%)으로 가장 낮게 나타났다. 즉, 치매 관련 신체활동 뉴스 기사들의 이슈는 치매 예방과 관련된 이슈가 중점적이며, 그다음으로는 치매 연구, 치매 환자를 위한 복지서비스 순으로 파악된다.

Table 6. Topic modeling result

Category	Topic 1	Topic 2	Topic 3
Keyword	센터	건강	환자
	서비스	관리	뇌
	환자	프로그램	이상
	이용	사업	치료
	시설	보건소	위험
	지원	예방	검사
	사회	대상	사람
	말	신체_활동	인지
	의료	교육	교수
	병원	통해	활동
	기관	개	경우
	대한	교실	예방
	생활	활동	연구
	경우	참여	질환
	요양보호사	진행	기억
n	57	126	91
(%)	(20.8)	(46.0)	(33.2)

n: The number of classified documents

논의 및 결론

이 연구는 최근 산업공학 및 문헌정보학 분야에서 텍스트 자료를 분석하는 데 많이 활용하고 있는 토픽모델링의 기본개념과 절차를 소개하고 체육학 분야에서 토픽모델링의 적용 가능성을 살펴보고자 하였다. 이를 위해 토픽모델링의 개념을 소개하고 토픽모델링을 위한 텍스트 전처리 과정, 텍스트 정형화 과정, 토픽 수 결정 과정을 제시하였다. 또한, 체육학 분야의 적용 가능성을 확인하기 위하여 치매 관련 신체활동 뉴스 기사의 본문을 자료로 토픽모델링 방법을 예제로 정리하였다.

텍스트 내의 잠재적 정보를 추출하고 범주화하는 방법은 내용 분석적 방법과 통계적 방법으로 구분할 수 있다. 통계적인 방법은 네트워크 기반의 의미연결망 분석과 토픽모델링이 가장 대표적으로 활용되고 있으며 이 연구에서는 최근 많이 언급되고 있는 토픽모델링을 소개하였다. 사실 텍스트 자료는 자료의 특성상 문맥의 흐름에 따라 잠재적 의미는 변화할 수 있으므로 통계적으로 잠재적 정보를 추출한다는 것이 무의미한 일일 수도 있다. 오히려 내용 분석적 방법이 보다 더 타당한 방법일 수도 있다. 그러나 다량의 자료가 쏟아지는 최근의 데이터분석 환경에서는 통계적 방법으로 활용하지 않고서는 해결하기 힘든 것이 현실이다.

토픽모델링은 비교적 최근에 소개된 방법으로 산업공학, 문헌정보학, 정보시스템학 등 다양한 학계에서 활용되고 있는 실정이다(Lee & Park, 2017). 토픽모델링의 적용사례는 증가하고 있지만, 토픽모델링 분석방법에 관한 정보를 체계적으로 소개하고자 하는 연구는 현저히 부족한 실정이다. 특히, 텍스트 자료는 비정형화된 자료로 텍스트 전처리를 어떻게 하느냐에 따라 혹은 텍스트 정형화를 어떠한 방법으로 하느냐에 따라 그 결과는 상이할 수 있으므로 체계적인 분석방법을 설명하는 것은 유용한 정보가 될 것으로 판단된다.

이 연구에서는 토픽모델링을 분석절차를 4단계로 설명하였다. 첫 번째로 소개한 텍스트 전처리 단계는 텍스트 자료를 분석하는데 가장 중요한 단계이면서 가장 논란의 여지가 많은 단계이다. 왜냐하면, 텍스트의 경우 띄어쓰기나 문맥에 따라 그 의미가 달라질 수 있으나 컴퓨터가 이를 정확하게 판단하기는 아직 기술적으로 보완되어

야 할 부분이 많기 때문이다. 대표적인 사례로 '아빠가 방에 들어가셨다'와 '아빠 가방에 들어가셨다'의 문장은 띄어쓰기 때문에 그 내용은 완전히 다른 의미를 지니지만 아직 컴퓨터는 이를 분류하는데 한계점을 가진다. 따라서 불용어 제거 및 공통단어로 변환, 연속단어로 변환 등의 단계를 걸쳐 오류를 최소화해야 하지만 대량의 자료로부터 관찰되는 모든 오류를 통제할 수 없는 것이 현실이다. 이와 같은 이유로 이 단계에서 오류들을 얼마나 통제하느냐가 모델의 성능을 높이는 관건이라 할 수 있다.

토픽모델링을 적용한 선행연구에서는 사전에 설정한 토픽 수에 대한 근거를 제시한 연구는 찾아보기 쉽지 않다. 통계적으로 최적의 토픽 수를 탐색하는 것은 사실상 무의미하다는 견해(DiMaggoi et al., 2013; Nam, 2016)도 있지만 토픽모델링을 제안하고 평가하는 연구(Blei et al., 2003; Chang et al., 2009; Newman et al., 2010)에서 적용하는 인덱스(Perplexity, Coherence Score)가 존재하는 것 역시 사실이다. 토픽모델링은 사전에 설정한 토픽 수에 따라 결과는 상이하게 도출되기 때문에 토픽 수는 토픽모델링에서 중요한 역할을 한다. 그러나 모델을 평가하는 방법이 존재함에도 불구하고 해석의 용이성 때문에 연구자 주관에 따라 토픽 수를 결정한다면 객관적인 통계적 방법으로 잠재적 정보를 추출해낸다는 토픽모델링의 장점에 모순이 생길 것이다. 따라서 토픽모델링을 평가하는 Coherence Score를 활용하여 토픽 수를 결정하는 것이 보다 더 적절하다고 판단된다.

토픽모델링은 연구자의 주관성을 배제하여 잠재적 정보(주제)를 추출할 수 있으며 각 문서에 할당된 주제의 확률 역시 산출해 낼 수 있다. 이는 각각의 문서를 주제별로 분류할 수 있으며 각각의 주제가 전체에서 차지하는 비중을 산출해낼 수 있다. 구체적으로 이 연구의 적용 예제 결과를 살펴보면 치매 관련 신체활동 뉴스 기사에서 '치매 환자를 위한 복지서비스', '치매 예방', '치매 연구'라는 3개의 잠재적 토픽을 추출하였으며, 총 274개의 뉴스 기사가 어떠한 토픽에 해당하는가를 분류하여 비율로 제시하였다. 그 결과 '치매 예방' 관련 뉴스가 46.0%로 가장 많았으며, '치매 환자를 위한 복지서비스'와 관련한 뉴스가 20.8%로 가장 적은 것으로 나타났다. 이러한 결과는 뉴스 기사에서 주요한 이슈가 무엇인지 파악할 수 있으며 또한 자료를 연도별로 범주화하여 주제별 비중 변화를 비교한다면 시기에 따라 이슈가 어떻게 변화하는지 파악할 수 있을 것이다.

국내 체육학 분야에서 토픽모델링이 소개된 사례는 많지 않다. 아직까진 텍스트 자료는 내용분석 방법과 네트워크기반 분석방법들이 주를 이루고 있다. 특히 내용분석 방법은 연구자의 주관성을 완전히 배제할 수 없으며 빅데이터의 경우 분석에 소요되는 시간이 많기 때문에 현실적으로 불가능한 것이 사실이다. 따라서 체육학 분야에서 토픽모델링을 적용한다면 보다 객관적으로 텍스트 내의 주제를 추출하고 추출된 주제로부터 구체적이고 다양한 정보를 제공할 수 있을 것이다. 구체적으로 체육학 분야에서 적용 가능성을 3가지 측면으로 제언하고자 한다. 첫째, 체육학 분야의 연구 동향을 파악하는데 적용 가능할 것이다. 연구의 초록을 대상으로 잠재적 주제를 파악하고 연도별로 주제가 자치하는 비중을 산출한다면 상향하고 있는 주제와 하향하고 있는 주제를 파악할 수 있을 것이다. 둘째, 체육 관련 주제에 대해 대중들의 의견을 종합적으로 파악하는데 적용 가능할 것이다. 가령 체육 정책이 시행되었을 때 대중들이 의견을 댓글이나 SNS를 통해 파악할 수 있을 것이며 스포츠이벤트(올림픽, 월드컵 등)가 개최되었을 때 대중들의 의견을 반영하여 이벤트를 개최 성공 여부를 평가할 수 있을 것이다. 셋째, 체육 분야의 사회적 이슈가 발생하였을 때 이슈의 쟁점을 파악하는데 적용 가능할 것이다. 사회적 이슈의 발전은 언론매체에서 보도하는 뉴스 기사의 역할이 중요하기 때문에 뉴스 기사를 토대로 토픽모델링을 적용한다면 이슈에 대한 주요한 쟁점이 무엇인지 파악할 수 있을 것이다.

참고문헌

Bae, J. H. (2019). *Analysis on the change process of specific news topics using Dynamic Topic Models.* Unpublished master's thesis, Dongguk University, Seoul, Korea.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993-1022.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M.

(2009). Reading tea leaves: How humans interpret topic models. *In Advances in neural information processing systems* (pp. 288-296).

Chang, M., Jonsson, P. V., Snaedal, J., Bjornsson, S., Saczynski, J. S., Aspelund, T., ... & Gudnason, V. (2010). The effect of midlife physical activity on cognitive function among older adults: AGES—Reykjavik Study. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, 65*(12), 1369-1374.

Choi, Y. J., & Kweon, S. H. (2014). A semantic network analysis of the newspaper articles on big data. *Journal of Cybercommunication, 31*(1), 242-284.

Chung, D. H., Lee, J. K., Kim, S. E., & Park, K. J. (2013). An analysis on congruency between educational objectives of curriculum and learning objectives of textbooks using semantic network analysis: Focus on earth science I in the 2009 revised curriculum. *Journal of the Korean Earth Science Society, 34*(7), 711-726.

Crossman, J., Vincent, J., & Speed, H. (2007). The Times They are A-Changin' Gender Comparisons in Three National Newspapers of the 2004 Wimbledon Championships. *International Review for the Sociology of Sport, 42*(1), 27-41.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics, 41*(6), 570-606.

Hahm, Y. K., & Lee, S. J. (2016). The Distinctiveness of Big Data Business Model in Its Components: A Comparative Analysis of Korea-US Cosmetic Big Data Business Cases. *Journal of Information Technology and Architecture, 13*(1), 63-75.

Hamer, M., & Chida, Y. (2009). Physical activity and risk of neurodegenerative disease: a systematic review of prospective evidence. *Psychological medicine, 39*(1), 3-11.

Han, M. K., Kim, W. K., & Yoon, J. W. (2016). Perceptions of Disabled Sports in Newspapers Using Semantic Networks Analysis. *Journal of Rehabilitation Research, 20*(4), 157-175.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2-3), 146-162.

Hofmann, T. (1999). Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 50-57.

Iyengar, S. (1994). *Is anyone responsible?: How television frames political issues.* Chicago: University of Chicago Press.

Kang, B. I., Song, M., & Jho, W. S. (2013). A study on opinion mining of newspaper texts based on topic modeling. *Journal of the Korean Library and Information Science Society, 47*(4), 315-334.

Kang, A. T., & Kang, Y. O. (2018). Analyzing spatial characteristics of stress topics using tweet data. *Journal of the Korean Cartographic Association, 18*(2), 53-69.

Kim, N. G., Lee, D. H., & Choi, H. C. (2017). Investigations on techniques and applications of text analytics. *The Journal of Korean Institute of Communications and Information Sciences, 42*(2), 471-492.

Kim, B. S. (2018). Analysis of research trends in personal assistance service in korea using topic modeling. *Journal of Disability and Welfare, 42*, 163-190.

Kim, D. S., & Lee, H. O. (2014). A new approach to public segmentation theory suitable for social media public. *Journal of Public Relations, 18*(3), 394-429.

Kim, J. Y., & Kim, D. S. (2016). A study on the method for extracting the purpose-specific customized information from online product reviews based on text mining. *The Journal of Society for e-Business Studies, 21*(2), 151-161.

Kim, S. K., & Jang, S. Y. (2016). A study on the research trends in domestic industrial and management enginering using topic modeling. *Journal of the Korea Management Engineers Society, 21*(3), 71-95.

Lee, S. J., & Kim, H. J. (2009). Keyword extraction from news corpus using modified TF-IDF. *The Journal of Society for e-Business Studies, 14*(4), 59-73.

Lee, Y. K., & Park, J. H. (2017). An analysis of the research trend in the korean journal of measurement and evaluation in physical education and sport science using topic models. *The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science, 19*(2), 11-22.

Mimno, D., Wallach, H., & McCallum, A. (2008, December). Gibbs sampling for logistic normal topic models with graph-based priors. *In NIPS Workshop on Analyzing Graphs* (Vol. 61).

Myung, J. S., Lee, D. J., & Lee, S. G. (2008). A korean product

review analysis system using a semi-automatically constructed semantic dictionary. *Journal of KIISE, 35*(6), 392-403.

Nam, C. H. (2016). An illustrative application of topic modeling method to a farmer's diary. *Cross-Cultural Studies, 22*(1), 89-135.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006, May). Analyzing entities and topics in news articles using statistical topic models. *In International conference on intelligence and security informatics* (pp. 93-104). Springer, Berlin, Heidelberg.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (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* (pp. 100-108). Association for Computational Linguistics.

Paranyushkin, D. (2010). Text network analysis. *In Conférence du Performing Arts Forum, retreived form http://noduslabs com/ research/pathways-meaning-circulation/, (14.09. 2011).*

Park, C. S., & Chung, C. W. (2013). Text network analysis: detecting shared meaning through socio-cognitive networks of policy stakeholders. *Journal of Governmental Studies, 19*(2), 73-108.

Park, K. J. (2016). A design on informal big data topic extraction system based on spark framework. *KIPS transactions on software and data engineering, 5*(11), 521-526.

Park, J. H., & Song, M. (2013). A study on the research trends in library & information science in korea using topic modeling. *Korea Society for Information Management, 30*(1), 7-32.

Park, J. H., & Oh, H. J. (2017). Comparison of topic modeling methods for analyzing research trends of archives management in korea: Focused on LDA and HDP. *Journal of Korean Library and Information Science Society, 48*(4), 235-258.

Park, S. G., Park, K. W., & Kang, H. W. (2018). The research features analysis of leisure and recreation based on co-authors network and topic model. *The Korean Journal of Physical Education, 57*(2), 279-289.

Raschka, S. (2015). *Python machine learning.* Birmingham: Packt Publishing Ltd.

Rapp, A., Beitelspacher, L. S., Grewal, D., & Hughes, D. E. (2013). Understanding social media effects across seller, retailer, and consumer interactions. *Journal of the Academy of Marketing Science, 41*(5), 547-566.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424-440.

Um, H. S., Kang, E. B., Koo, J. H., Kim, H. T., Kim, E. J., Yang, C. H., ... & Cho, J. Y. (2011). Treadmill exercise represses neuronal cell death in an aged transgenic mouse model of Alzheimer's disease. *Neuroscience research, 69*(2), 161-173.

Yoon, J. E., & Suh, C. J. (2018). Research trend analysis on smart healthcare by using topic modeling and ego network analysis. *Journal of Digital Contents Society, 19*(5), 981-993.

Yoon, J. W., & Park, J. H. (2015). Semantic network analysis for content analyzing of qualitative research in adapted physical activity. *Korean journal of physical education, 54*(5), 877-889.

Yun, H. J. (2018). *A real-time players evaluation model development based on social big data in korea professional baseball: Sentiment analysis using machine learning.* Ph.D. Dissertation, Korea National Sport University, Seoul.

비정형 텍스트 자료에서 잠재정보 추출을 위한 토픽모델링 소개: 치매관련 신체활동 뉴스 기사의 이슈 분석

윤효준 · 박재현 · 윤지운(한국체육대학교)

〔목적〕 이 연구는 토픽모델링에 대한 기본개념 및 절차에 대해 소개하고 치매관련 신체활동에 대한 뉴스기사에 토픽모델링을 적용사례로 설명하는 것이 목적이다. 아울러 체육학 분야에서 토픽모델링의 활용가능성을 논의하고자 하였다. **〔방법〕** 이 연구에서는 토픽모델링의 LDA 알고리즘을 설명하고 분석절차를 텍스트전처리, 텍스트정형화, 토픽수결정으로 단계별로 요약하였다. 적용사례는 치매관련 신체활동에 대한 뉴스기사로 2000년부터 2018년까지 13개 주요일간지에 보도된 274건의 뉴스본문을 대상으로 선정하였다. **〔결과〕** 토픽의 수는 3개 일 때 Coherence Score값이 가장 높게 나타났다. 토픽1은 치매환자를 위한 복지서비스 주제, 토픽2는 치매예방 주제, 토픽3은 치매연구 주제이다. 주제별 비율은 토픽2(46.0%), 토픽3(33.2%), 토픽1(20.8%)순으로 높게 나타났다. **〔결론〕** 토픽모델링은 연구자의 주관성을 배제하여 잠재적 정보를 추출해낼 수 있는 효과적인 방법론으로 체육학분야에서도 방대한 텍스트자료에서 정보를 탐색하고자 할 때 활용되길 기대한다.

주요어: 텍스트마이닝, 토픽모델링, LDA 알고리즘