LDA, Top2Vec, BERTopic 모형의 토픽모델링 비교 연구

- 국외 문헌정보학 분야를 중심으로 -

A Comparative Study on Topic Modeling of LDA, Top2Vec, and BERTopic Models Using LIS Journals in WoS

이 용 구 (Yong-Gu Lee)*
김 선 욱 (SeonWook Kim)**

목 차

초 록

이 연구는 토픽모델링 모형인 LDA, Top2Vec, BERTopic을 대상으로 실험데이터에서 토픽을 추출하고, 그 결과를 비교 분석함으로써 각각의 모형 간의 특성과 차이를 파악하는데 목적이 있다. 실험데이터는 Web of Science(WoS)에 등재된 문헌정보학 분야 학술지 85종에 게재된 논문 55,442편을 대상으로 하였다. 실험 과정으로 우선 각 모형의 파라미터를 기본값 그대로 이용하여 1차 토픽모델링 결과를 얻었고, 최적의 토픽 수를 설정하여 각 모형의 2차 토픽모델링 결과를 얻었으며, 이들을 각 모형과 단계별로 비교분석하였다. 1차 토픽모델링 단계에서는 LDA, Top2Vec, BERTopic 모형이 각각 100개, 350개, 550개의 토픽을 생성하여 세 모형은 각각 매우 다른 크기의 토픽 개수를 가져왔으며, LDA 모형에 비해 Top2Vec이나 BERTopic 모형이 토픽을 3배, 5배 더 세분화하였다. 또한 세 모형은 토픽 당 문서 수의 평균이나 표준편차에서도 많은 차이가 났다. 구체적으로 LDA 모형은 비교적 적은 수의 토픽에 많은 문서를 부여하는 반면, BERTopic 모형은 반대의 경향을 보였다. 25개의 토픽 수를 생성하는 2차 토픽모델링 단계에서는 다른 모형에 비해 Top2Vec 모형이 평균적으로 토픽 당 많은 문서를 부여하고 토픽 간에 고르게 문서를 할당하여 상대적으로 편차가 작았다. 또한 모형간의 유사 토픽의 생성여부를 비교하면, LDA와 Top2Vec 모형이 전체 25개 중에 18개(72%)의 공통된 토픽을 생성하여 BERTopic 모형에 비해 두 모형이 더 유사한 결과를 보였다. 향후 토픽모델링 결과에서 각 토픽과 부여된 문서들이 주제적으로 올바르게 형성되었는지에 대한 전문가의 평가를 통해 보다 완전한 분석이 필요하다.

ABSTRACT

The purpose of this study is to extract topics from experimental data using the topic modeling methods(LDA, Top2Vec, and BERTopic) and compare the characteristics and differences between these models. The experimental data consist of 55,442 papers published in 85 academic journals in the field of library and information science, which are indexed in the Web of Science(WoS). The experimental process was as follows: The first topic modeling results were obtained using the default parameters for each model, and the second topic modeling results were obtained by setting the same optimal number of topics for each model. In the first stage of topic modeling, LDA, Top2Vec, and BERTopic models generated significantly different numbers of topics(100, 350, and 550, respectively). Top2Vec and BERTopic models seemed to divide the topics approximately three to five times more finely than the LDA model. There were substantial differences among the models in terms of the average and standard deviation of documents per topic. The LDA model assigned many documents to a relatively small number of topics, while the BERTopic model showed the opposite trend. In the second stage of topic modeling, generating the same 25 topics for all models, the Top2Vec model tended to assign more documents on average per topic and showed small deviations between topics, resulting in even distribution of the 25 topics. When comparing the creation of similar topics between models, LDA and Top2Vec models generated 18 similar topics(72%) out of 25. This high percentage suggests that the Top2Vec model is more similar to the LDA model. For a more comprehensive comparison analysis, expert evaluation is necessary to determine whether the documents assigned to each topic in the topic modeling results are thematically accurate.

키워드: 토픽모델링, LDA, Top2Vec, BERTopic, 문헌정보학
Topic Modeling, LDA, Top2Vec, BERTopic, Library and Information Science

* 경북대학교 사회과학대학 문헌정보학과 부교수(yglee@knu.ac.kr / ISNI 0000 0004 6437 6752) (제1저자)
** 대구가톨릭대학교 사회과학대학 문헌정보학과 강사(sewokim@gmail.com / ISNI 0000 0004 9360 1074) (공동저자)
논문접수일자: 2023년 12월 23일 최초심사일자: 2024년 1월 5일 게재확정일자: 2024년 1월 29일
한국문헌정보학회지, 58(1): 5-30, 2024. http://dx.doi.org/10.4275/KSLIS.2024.58.1.005

1. 서 론

일반적으로 토픽모델링은 비정형의 텍스트(unstructured text)로 구성된 다량의 문헌 집합을 대상으로 분석하는 비지도 학습 기반의 머신러닝 기법에 해당한다. 이는 문헌 클러스터링 방법과 유사한 측면이 있으며 자연언어처리 분야에서 의미 있는 정보를 발견하기 위한 텍스트 마이닝 기법 중 하나이다.

토픽모델링은 문서 안에 숨겨진 주제 또는 토픽(topic)을 발견하는 분석기법을 통칭한다. 이 기법은 특정 분야의 문헌 집단을 대상으로 어떤 주제가 분포하고 그 주제의 비중을 파악할 수 있어 해당 분야의 정보 탐색에서 연구자의 시간과 노력을 크게 줄여줄 수 있으며 동시에 주제 영역의 세분화에 따른 정확성까지 높일 수 있는 훌륭한 도구로써 이미 다양한 분야[1]에서 그 효용성을 입증하고 있다.

또한 학문 분야에 대한 토픽모델링과 그에 따른 시각화를 활용하면 장시간 누적된 많은 양의 데이터를 쉽게 이해할 수 있도록 도와주는 역할을 할 뿐만 아니라, 복잡도가 높은 데이터 안에 숨겨진 패턴을 드러내어 독자에게 통찰력, 설득력, 방향성을 제시하는 역할을 한다(Dur, 2014). 즉 연구자가 다양한 연구주제를 발견하고 새로운 연구 방향을 모색할 수 있도록 단초를 제공할 수 있다.

초기의 토픽모델링의 주요 모형은 통계 및 확률 이론에 기반을 두었다. Deerwester et al.(1990)의 LSA(잠재의미분석)는 전통적 토픽모델링 방법으로 문헌-용어 행렬에서 잠재적 의미를 도출하는 방식이며, 이후 Hofmann(1999)의 pLSA(확률잠재의미할당)로 확장되었다. 최근 폭넓게 응용되고 있는 LDA(잠재디리클레할당)은 LSA와 pLSA의 단점을 보완한 방법으로 하이퍼파라미터를 사용하여 통계 모델을 제어할 수 있어 다양한 분야에서 사용되고 있다(Jelodar et al., 2019; Vayansky & Kumar, 2020).

최근에는 토픽모델링 분야에 머신러닝 분야의 핵심적인 개념까지 도입하여, 신경망 학습을 통해 구축된 언어모델을 사용한 토픽모델링 방법도 등장하였다(Angelov, 2020; Sia, Dalmia, & Mielke, 2020; Grootendorst, 2022). 전통적인 통계 및 확률 기반의 토픽모델링이 가지는 제약으로 BoW(Bag-of-Words) 방식의 채택을 들 수 있는데, Top2Vec 모형(Angelov, 2020)은 이러한 제약을 극복하기 위해 단어와 문헌의 의미적 임베딩 표현을 결합하고 이를 통해 분산된 토픽 벡터(distributed topic vector)의 개념을 적용하였다. Grootendorst(2022)가 제안한 BERTopic 모형도 입력데이터를 BoW 방식이 아닌 자연어 그대로 처리할 수 있는데, 최신의 BERT 계열의 사전학습(pre-training) 모델이 생성한 임베딩 표현을 활용함으로써 문장이나 텍스트 전체에 대한 문맥 정보를 반영할 수 있어 보다 좋은 토픽모델링 성능을 기대할 수 있다. 다만 이들 토픽 모델링이 LDA 모형뿐만 아니라 임베딩을 사용하는 두 모형 측면에서도 다소 상이하고 토픽모델링의 결과도 다를 것으로 판단되므로 서로 다른 이들 모형들을 동일한 데이터셋을 적용하여 성능을 비교

1) 2023년 4월 14일 기준 KCI(https://www.kci.go.kr)에서 "토픽모델링"을 키워드로 검색하면 1,338편의 논문이 확인되며 그 중 분야별 1위는 사회과학(628편), 2위는 복합학(243편), 3위는 공학(216편)이다.

해볼 필요가 있다.

이러한 배경에서 이 연구는 2001년 1월부터 2021년 10월까지 약 21년간 Web of Science에 등재된 문헌정보학 분야 학술지(Library and Information Science - SSCI) 85종에 게재된 논문 55,442편에 대한 서지정보를 분석 대상인 실험집단으로 선정하였으며, 오픈 소스 유형의 주요 토픽모델링 도구인 LDA, Top2Vec, BERTopic 모형을 이용하여 실험집단으로부터 토픽을 추출하고 그 결과를 비교 분석함으로써 각각의 모형 간의 특성과 차이를 파악하는 데 목적을 두고 있다.

2. 이론적 배경

2.1 LDA

Deerwester et al.(1990)이 문헌과 그 문헌에 출현한 단어의 빈도를 표현한 문헌-용어 행렬을 차원 축소해 잠재적 의미를 추출하는 LSA를 제안한 후, Hofmann(1999)은 확률적인 분포로부터 토픽을 생성하는 pLSA를 제안하였다. 이 맥락에서 등장한 LDA 모형은 2개의 하이퍼파라미터(문서의 토픽 분포를 통제하는 α와 토픽 내에서 단어의 분포를 통제하는 β)를 사용하여 문헌의 토픽 분포와 토픽의 단어 분포를 예측한다(Blei, Ng, & Jordan, 2003). 이후 다수의 연구에서 사용되면서 LDA의 내재적 단점을 찾아내고 보완하기 위한 다양한 후속 연구들이 이루어져 왔다(Jing, Zhang, & Tang, 2004; Mehrotra et al., 2013; Chen, Sheble, & Eichler, 2013).

LDA 모형은 토픽모델링을 하기 위해서는 2개의 하이퍼파라미터인 α와 β를 추정해야 하며 이로 인해 같은 데이터셋에 대해서도 모델링 결과가 달라질 수 있을 뿐만 아니라, LDA 모형이 디리클레 확률분포에 근거하기에 토픽의 분포를 서로 독립적이라고 전제해야 한다(Vayansky & Kumar, 2020). 이를 해결하기 위해 토픽 분포에서 상관관계를 나타내기 위해 로지스틱 정규분포를 적용하는 새로운 토픽모델링 모형으로 CTM(correlated topic model)이 제안되었다(Blei & Lafferty, 2005). 다만 CTM 모형도 두 주제 쌍의 공분산 행렬(covariance matrix)을 활용하기에 그 이상의 다양한 주제 사이의 관계를 파악할 수 없으므로 이를 하기 위해 DAG(directed acyclic graph) 혼합 모형을 적용한 PAM(pachinko allocation model) 모형도 제시되었다(Li & McCallum, 2006).

2.2 Top2Vec과 BERTopic

토픽모델링은 오랫동안 확률 측면에서 접근되어 왔지만, Moody(2016)가 제안한 LDA2VEC은 LDA의 개념을 벡터 공간에 구현하였다. 이후 토픽모델링을 구현하는 데 단어의 분산 표현을 이용하여 벡터 공간으로 투영하는 임베딩 기법을 적용한 다양한 접근 방법이 제안되었다(Li et al., 2018; Gao et al., 2022).

최근에는 이러한 임베딩 기법을 문장과 문헌에 응용하는 연구도 선보이고 있는데, Angelov(2020)는 단어 임베딩(word2vec)과 문헌 임베딩(doc2vec)을 이용하여 공동으로 생성한 벡터 공간에서 의미적 연관성을 거리로 표현하고 인근 단어와 문헌을 군집화하여 토픽 또는 토

픽 벡터(topic vectors)로 표현하는 Top2Vec 모형을 제안하였다. 이때 군집화 방법으로 UMAP(Uniform Manifold Approximation and Projection) 알고리즘으로 차원 축소한 뒤 HDBSCAN 알고리즘을 적용하였다.

BERTopic 모형(Grootendorst, 2022)은 최근 다양한 분야에서 높은 성능으로 인기를 끌고 있는 딥러닝 사전학습 언어 모델인 BERT 계열의 모형을 사용하여 문헌 임베딩의 성능을 높이는데 구체적으로 Sentence-BERT(Reimers & Gurevych, 2019)를 적용하였다. 이후 Top2Vec 모형과 같이 UMAP 알고리즘과 HDBSCAN 알고리즘을 적용하여 군집화함으로써 토픽을 생성하고 각 토픽에 속한 모든 문헌에 출현한 용어를 클래스 기반 TF-IDF(c-TF-IDF)로 계산하여 중요성을 나타내고 이를 토픽 표현으로 나타낸다. 이후 각 시간대의 c-TF-IDF를 해당 벡터의 크기(L1 Norm)로 정규화하여 평활화(smoothing)한 것을 동적 토픽모델링의 시각화 결과로 제공한다.

LDA 모형은 확률적인 모델이므로 주어진 하이퍼파라미터에 따라 모든 문서에 토픽을 올바르게 할당하는 문제를 해결하는 반면, Top2Vec이나 BERTopic 모형과 같은 임베딩 방식은 벡터 공간에 투영된 문서의 군집에서 토픽을 추출하는 것으로 요약할 수 있다. LDA 모형은 평가를 통해 가장 핵심이 되는 주제를 파악할 수 있다는 장점을 가지지만, 약점으로 이 모형은 기존의 전통적인 토픽모델링 모형과 마찬가지로 토픽 수의 최적화를 결정하는 과정이 필요하거나, 스테밍이나 형태소 분석과 같은 언어학적 처리가 필요하며 BoW 방식의 적용으로 문헌 내의 단어의 순서나 의미를 반영하지 못하는 점을 들 수 있다. 반대로 Top2Vec이나 BERTopic 모형은 일반적으로 토픽 수를 자동으로 결정할 수 있으며 임베딩 표현 방식을 적용으로 언어학적 처리에서 자유롭고 단어의 의미 반영이 가능하다.

2.3 국내외 주요 선행 연구

토픽모델링 기법내지 모형의 차이를 비교한 연구를 보면, 박준형, 오효정(2017)은 기록관리학 관련 논문 1,027건을 LDA와 HDP 기법을 이용해 각각 토픽모델링한 뒤 그 결과를 비교분석함으로써 두 기법의 차이를 확인하고자 하였다. 그 결과, LDA 토픽을 구성하는 토픽 단어는 주로 기록관리학에서 주요 대주제로 사용하고 있는 키워드인 경향이 있었으며 HDP 토픽을 구성하는 토픽 단어는 LDA와 비교해서 세부적인 주제를 파악할 수 있는 키워드가 주로 발견되었다. Egger와 Yu(2022)는 트위터의 단문 텍스트를 4개의 토픽모델링 기법(LDA, NMF, Top2Vec, BERTopic)으로 분석함으로써 각 방법의 특성을 파악하고자 하였다. 저자는 임베딩기반 기법인 Top2Vec와 BERTopic 모형은 연구자가 특정 주제를 깊이 탐구할 수 있도록 토픽 용어 검색이 큰 장점임을 강조하면서 특히 BERTopic의 유용성을 강조하였으며, 통계적 기법인 LDA와 NMF 중에서는 LDA 모형이 많은 연구에서 채택되고 있으나 정확한 결과를 얻기 위해 사전확률과 하이퍼파라미터 등 여러 제한 사항이 존재하므로 이에 대한 이해가 충분하지 않은 연구자라면 대안으로 NMF 모형을 사용하도록 권고하였다. 다만 어느 모형이든 특성에 따른 장·단점이 있으므로, 연구

자는 정성적 결과 해석에 노력을 기울여야 함을 강조하였다.

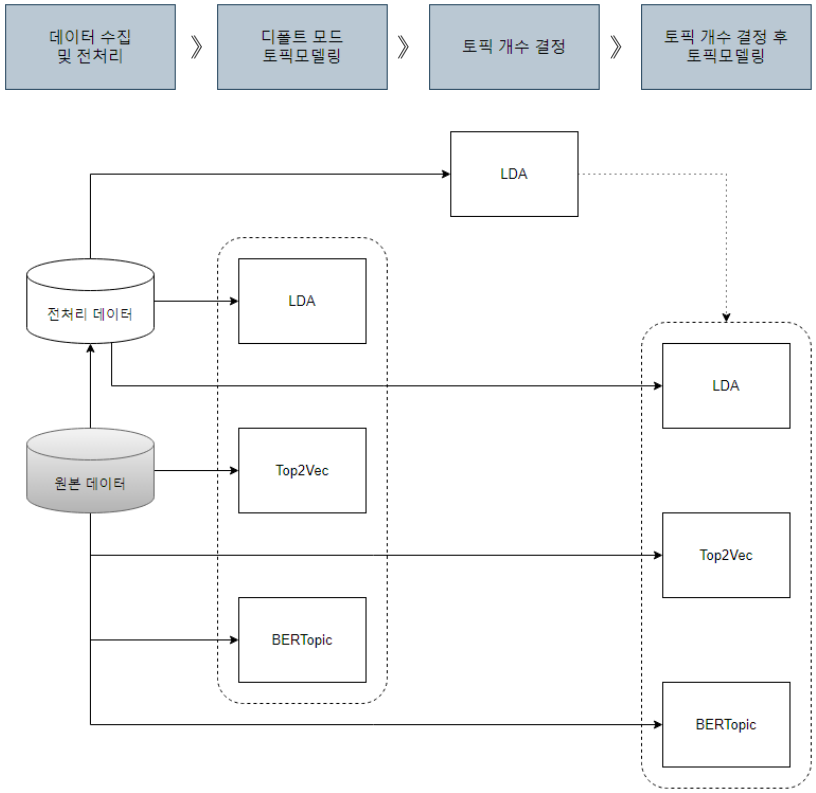
특히 박자현, 송민(2013)은 국내 문헌정보학 분야 토픽모델링의 기준이 되는 틀을 마련했는데 한국 문헌정보학 분야 4대 학술지에 1970년부터 2012년도까지 실린 총 3,834편의 논문 초록을 LDA 모형을 사용하여 토픽모델링하고, 시계열 분석을 진행하였다. 임정훈(2022)은 LDA의 확장판 중 하나인 STM을 사용하여 토픽모델링을 수행한 뒤 연도별 토픽 동향을 분석함으로써 정보 활용 교육의 방향성과 그 후속 연구를 제안하였다. 김선욱, 양기덕, 이혜경(2022)은 2001년부터 2020년까지 문헌정보학 SSCI 85종 학술지에 게재된 55,442편을 DTM으로 분석하고, 연구주제마다 시대적 변화가 반영되고 있음을 밝혀냈다.

토픽모델링에 관한 연구가 다양해지면서, 연구자의 LDA 응용법도 다양하게 변화하고 있다. 이지용 외(2022)는 최적의 LDA 토픽 개수를 찾기 위해 직접 Gensim API를 호출하였으며 김태경, 김창식(2018)은 여러 상용 분석 도구를 연구의도 대로 선형 연결하는 독창적 아이디어를 선보였다. 이외에도 김선욱, 양기덕(2022)과 Ali와 Naeem(2022)이 제안한 것처럼 이종 토픽모델링을 합성해서 쓰는 방안도 연구된 바 있다. 구체적으로 김선욱, 양기덕(2022)은 LDA 모형이 발견한 토픽에 유사도 높은 여러 BERTopic 모형의 토픽을 합성함으로써 토픽의 의미를 증강하는 한편, LDA 토픽과 의미론적 관계에서 정반대에 있지만 BERTopic 모형이 발견해낸 비주류 주제를 발굴함으로써 토픽모델링을 확장하는 AET 기법을 제안하였다. 부수적으로 각 토픽모델링의 내부 토픽 점유율 지표를 합성하여 동적 토픽모델링으로 표현할 수 있는 지수를 만들어 낼 수 있다.

3. 연구 방법

주요 3개의 토픽모델링 모형 내지 도구(LDA, Top2Vec, BERTopic)를 이용하여 국외 문헌정보학 분야의 5만 5천여 건의 서지 데이터를 대상으로 토픽 분석을 수행하고 그 결과를 비교하기 위해, 먼저 각 모형이 기본으로 설정된 상태에서 토픽을 생성하여 결과를 분석하여 도구의 특성을 파악하였으며, 모형 당 동일한 크기의 토픽을 생성하도록 설정하여 모형간의 토픽모델링 최종 결과를 비교하도록 연구를 설계하였다. 실험은 크게 4단계로 구성되는데 보다 구체적인 실험설계와 과정은 〈그림 1〉과 같다.

첫 번째, 실험을 위한 데이터는 김선욱, 양기덕(2022) 연구에서 수집하고 전처리한 결과 그대로 사용하였다. 이 데이터셋은 Web of Science가 제공하는 문헌정보학 분야 85종의 학술지에 게재된 서지데이터를 대상으로 하였으며, 수집 레코드의 대상 기간은 2001년 1월부터 2021년 10월까지 약 20년 10개월로 한정하였다. 이 기간 동안의 영문 초록을 수록한 서지레코드는 총 55,442건이며, 이들 데이터에 대해 spaCy로 토큰을 식별하여 고유 토큰(unique token) 64,976개를 얻었으며 단·복수 및 불용어 제거와 같은 전처리를 거치고 지프 법칙을 이용하여 토큰을 한정하여 앞서의 고유 토큰을 8,070개로 축소하는 과정을 거쳤다. 이후 LDA 모형을 위한 전처리 데이터로 변환하였다. LDA 이외 다른 두 모형의 경우 자연어를 그대로 이용하여 임



〈그림 1〉 실험과정 개요

베딩하므로 이러한 과정이 태생적으로 필요하지 않다.

두 번째 단계는 세 모형에 대해 기본 설정(default, 디폴트 모드)에 의한 1차 토픽모델링 실험으로, 기본 설정 상태에서 각 도구 또는 모형마다 토픽모델링 결과를 생성하고 이를 비교·분석하였다. 모형을 구현하는 패키지는 일반적인 환경에서 그 모형의 목적에 부합하는 결과를 가져오도록 핵심적인 파라미터를 기본 값으로 설정해 놓는다. 이러한 이유로 패키지의 기본 설정만 사용해도 보편적으로 작동하며 이는 해당 패키지의 특성을 반영한다 할 수 있다. 연구자가 자신의 필요에 의해 최적화하지 않고 패키지의 기본 설정에 따른, 즉 이 연구에서는 토픽모델링의 성능향상을 도모하기 위해 추가 설정을 하지 않고 토픽모델링 모형이 갖는 자체의 기본 설정 값으로 실행함으로써 각 토픽모델링 모형의 기본 특성을 파악하고자 하였다. 따라서 이 단계에서는 LDA, Top2Vec, BERTopic 모형 모두 토픽 개수를 따로 설정하지 않았다. 다만 모형마다 최소한으로 설정해야 하는 파라미터가 있는데, LDA의 경우 Gensim API의 파라미터인 alpha와 eta의 디폴트값이 모두 'symmetric'인데 이를 기반으로 학습을 진행하는 값인 'auto'로 설정하였다. Top2Vec의 경우 파라미터 speed의 디폴트값인 'learning'을 명시

적으로 선언해 주었으며, 임베딩 표현을 생성하기 위한 기본 설정인 Doc2Vec을 사용하도록 하였다. BERTopic은 학습에 관련된 파라미터는 존재하지 않으므로 이를 설정하지 않았으나, 2023년 12월 현재 최신 BERTopic의 임베딩 기본 설정대로 'all- MiniLM-L6-v2' 모형을 사용하도록 하였다.

세 번째, 동일한 토픽 수를 생성하여 세 모형의 토픽모델링 성능을 비교하기 위해 모형이 최종으로 산출하는 토픽 개수를 결정하였다. 이 단계에서는 주어진 실험 데이터셋에 대한 최적의 토픽 개수를 결정하기 위해 LDA를 중심으로 일관성 점수를 측정하기 위한 다수의 파라미터(c_v, c_npmi, c_uci, u_mass 등)를 사용하였다. Top2Vec과 BERTopic 모형은 벡터 공간 상에 임베딩한 단어나 문헌 표현을 차원 축소하고 이를 HDBSCAN 알고리즘으로 군집화하는데, 여기서 정식으로 최적의 군집 수에 대한 성능을 평가하는 방안은 현재까지 아직 존재하지 않는다. 다만 기존의 전통적 방법 중 하나인 elbow method와 실루엣 분석(silhouette analysis)은 K-means 클러스터링에서 군집의 개수인 k를 결정하기 위한 방법이기에(Yuan & Yang, 2019), 임베딩 기반 토픽모델링 기법 자체적으로 최적의 토픽 개수를 결정할 수 없으므로 LDA를 통해 결정된 것을 기준으로 삼았다.

네 번째 단계는 성능 비교를 위한 각 모형의 2차 토픽모델링 실험으로, 기본 설정에 따른 토픽모델링 단계에서 수행한 것처럼 토픽모델링 작업을 다시 수행하되, 다만 토픽 개수를 설정하였다. 이때 앞서 세 번째 단계에서 확인한 최적의 토픽 개수가 사용하였다. 이렇게 확보한 LDA, Top2Vec, BERTopic의 토픽 결과를 비교・검토한 뒤 세 모형간의 토픽 형성에서 유사한 정도를 평가하였다. 이때 세 모형이 최종적으로 생성한 토픽 결과에 대해 문헌정보학 박사학위 소지자 3명이 각 토픽에 맞는 토픽명을 부여하였으며 이후 일치하지 않는 부분은 합의를 통해 통일하였다.

4. 기본 설정에 의한 토픽 분석

4.1 LDA 모형

gensim 패키지를 이용하여 LDA를 학습할 때 고속으로 병렬 처리할 수 있는 gensim.models.ldamulticore API는 하이퍼파라미터를 지정한 경우에만 사용할 수 있으나 이 연구에서는 토픽모델링 모형을 비교하는 것이 목적으로 하이퍼파라미터 지정 없이 1개의 CPU만을 이용하는 models.ldamodel API를 사용하였다. 기본 설정의 변화를 최소화하기 위해 토픽의 개수(파라미터 num_topics)는 지정하지 않고, 하이퍼파라미터 α(파라미터 alpha)와 하이퍼파라미터 η(파라미터 eta)를 모두 'auto'로 설정하여 LDA gensim API가 디폴트 모드에서 토픽모델링을 실시하도록 하였다. 토픽모델링 학습결과 총 100개의 토픽이 생성되었으며, 그중에서 토픽 내 점유율(documents per topic, 이하 점유율)이 가장 높은 상위 10개를 정리하면 〈표 1〉과 같다.

100개의 토픽 중에 가장 높은 비율을 차지한 1위 토픽(토픽 8)은 전체 5만 5천여 건의 데이터 중에서 10,505건이 포함되어 18.95%의 가장

〈표 1〉 LDA 모형의 토픽 결과(기본 설정, 상위 10개)

순위	토픽 번호	토픽 단어	문서 (개수)	비율 (%)
1	8	approach, framework, theory, process, context, work, development, case, analysis, design	10,505	18.95
2	91	library, collection, librarian, need, material, access, patron, development, institution, acquisition	2,877	5.19
3	33	citation, publication, author, index, impact, researcher, number, journal, analysis, year	2,704	4.88
4	85	student, literacy, education, course, school, faculty, teaching, instruction, teacher, skill	2,015	3.63
5	63	effect, influence, relationship, intention, variable, finding, attitude, hypothesis, survey, theory	1,952	3.52
6	45	feature, method, approach, word, dataset, performance, text, algorithm, accuracy, set	1,831	3.30
7	55	management, organization, business, strategy, process, implementation, stage, case, enterprise, phase	1,631	2.94
8	1	journal, access, publishing, publisher, publication, author, list, database, subscription, impact_factor	1,224	2.21
9	65	country, institution, production, database, level, funding, output, period, trend, productivity	1,174	2.12
10	26	patient, care, hospital, disease, treatment, provider, health_care, clinician, nurse, diagnosis	1,169	2.11

높은 비율을 보였다. 2위(토픽 91)와 3위(토픽 33)는 각각 5.19%(2,877건)와 4.88%(2,704건)로 나타났다. 이들은 1위 토픽(토픽 8)과의 비율에서 큰 차이를 보여 토픽 8이 상대적으로 다수의 논문이 포함된 것을 알 수 있다. 2위(토픽 91)부터 7위(토픽 55)까지는 완만한 감소를 보이다가 8위(토픽 1)부터는 미미하게 감소하고 있음을 알 수 있다.

상위 10개 토픽 중 1위인 토픽 8은 학술지 논문에서 '지식 체계 및 담론'과 관련되어 주로 사용되는 단어들로 구성된 주제로 나타났다. 다음으로 토픽 91은 '도서관 장서 및 개발', 토픽 33은 '인용분석 및 지표', 토픽 85는 '리터러시 교육', 토픽 63은 '연구 방법', 토픽 45는 '텍스트 분석 및 성능', 토픽 55는 '기업 관리 및 조직', 토픽 1은 '학술지 출판 및 접근', 토픽 65는 '국가 및 기관의 연구 성과', 토픽 26은 '의료서비스'와 관련된 주제로 볼 수 있다.

4.2 Top2Vec 모형

Top2Vec 모형에서는 파라미터 설정은 학습 속도를 결정하는 speed 파라미터를 기본 값인 'learning'으로 명시적으로 지정하는 것 외에는 다른 파라미터는 따로 설정하지 않고 토픽모델링을 실시하였으며 이때 Top2Vec 모형은 임베딩 과정에서 암묵적으로 Doc2Vec이 사용되었다. 토픽모델링 결과 총 350개의 토픽이 생성되었으며, 그중에서 토픽 내 점유율이 높은 상위 10개를 제시하면, 〈표 2〉와 같다.

〈표 2〉 Top2Vec 모형의 토픽 결과(기본 설정, 상위 10개)

토픽 번호	토픽 단어	문서 (개수)	비율 (%)
0	spatial, elevation, raster, land, polygon, dem, dems, areal, lidar, terrain	1,391	2.51
1	philosophical, epistemological, epistemology, philosophy, hermeneutics, ontological, critique, positivist, habermas, realist	1,362	2.46
2	instruction, instructional, course, student, teaching, courses, students, instructor, teach, classroom	1,003	1.81
3	usa, countries, productive, japan, publications, authored, domestic, output, prolific, internationally	956	1.72
4	illness, semistructured, traumatic, survivors, suffering, caring, schizophrenia, illnesses, trauma, distress	830	1.50
5	hirsch, egghe, exponent, zipf, lotka, distributions, index, informetrics, mathematical, lognormal	636	1.15
6	search, engines, queries, searchers, engine, searcher, query, searching, logs, reformulation	634	1.14
7	km, codification, kmrp, tacit, knowledge, organisational, organisations, organisation, organizational, ic	615	1.11
8	interlibrary, consortial, acquisitions, loan, serials, purchased, print, subscriptions, monograph, loans	607	1.09
9	ambidexterity, firm, capabilities, capability, absorptive, firms, agility, ambidextrous, turbulence, innovation	596	1.07

350개의 토픽 중에서 가장 높은 비율을 차지한 토픽 0은 '지리 정보'를 나타내는데 LDA의 상위 10개의 토픽 결과와는 다소 차이가 있다. 토픽 1은 '철학 및 인식론'을 나타내는 단어들로 구성된 주제를 보였으며, 토픽 2는 '도서관 이용교육', 토픽 3은 '국가의 연구성과', 토픽 4는 '건강 정보', 토픽 5는 '계량정보학', 토픽 6은 '검색엔진', 토픽 7은 '지식 경영', 토픽 8은 '상호대차', 마지막으로 토픽 9는 '기업 혁신'에 대한 주제로 나타났다. Top2Vec 모형의 토픽 모델링 결과를 LDA 모형과 비교하면 Top2Vec 모형의 토픽 1은 LDA의 1 순위인 토픽 8번과 유사한데, 이들 토픽은 두 모형에서 토픽 점유율이 상위 그룹에 속한다. 토픽 2는 LDA의 4위인 토픽 85와 유사하며, 토픽 3과 5는 각각 65와 33, 토픽 8이 91 등이다.

Top2Vec 모형은 LDA와 달리 토픽 번호는 점유율대로 생성되었으며, 점유율이 제일 높은 토픽을 낮은 토픽과 비교하였을 때 점유율의 편차내지 비율이 그리 크지 않았다. 이는 다른 모형과 달리 Top2Vec 모형이 비교적 균등하게 토픽을 생성하는 것을 의미한다. 결과적으로 Top2Vec은 앞서 기술한 복잡한 과정의 전처리를 거치지 않았음에도 불구하고 LDA에 비해 더 많은 토픽을 추출하고 세부 주제가 효과적으로 분산된 토픽모델링을 수행하는 것으로 보인다.

4.3 BERTopic 모형

BERTopic 모형에서도 API에서 사전학습 모델을 기본 값으로 지정하는 것 외에는 파라미터

를 설정하지 않고, 토픽모델링을 실시하였다. 결과적으로 BERTopic 임베딩에는 Top2Vec 모형에 비해 좋은 임베딩 성능을 가져오는 센텐스-트랜스포머(sentence-transformers) 모형의 'all-MiniLM-L6-v2'가 사용되었다. 토픽 모델링 학습 결과 총 550개의 토픽이 생성되었으며, 그중에서 토픽 내 점유율이 높은 상위 10개를 정리하면, 〈표 3〉과 같다.

BERTopic 모형은 다른 모형과 달리 이상치(outliers)를 따로 토픽으로 분류하는데, 디폴트 모드에서 발생한 이상치(표 하단에 나와 있듯이 토픽 번호를 -1로 부여함)는 모두 24,388개로, 43.99%에 이르렀다. 이는 BERTopic이 사용하는 HDBSCAN 알고리즘이 군집화하지 못한 영역을 나타내며(Grootendorst, 2022), 최적화 내지 설정을 통해 이 비율을 낮추거나 개선할 수 있으나, 되도록 동일한 상황에서 세 모형의 결과를 비교하는 이 연구의 목적에 부합하지 않아 더 이상의 개선을 위한 실험을 수행하지 않았다.

BERTopic 모형에서도 Top2Vec처럼 토픽 번호는 점유율의 크기에 비례하여 순서대로 생성되었으며, 토픽 0은 다른 토픽에 비해 점유율이 높게 확인되었는데 Top2Vec과 동일하게 '지리 정보'에 대한 주제를 나타내었다. 토픽 1은 '기술 특허'에 대한 주제를 보였으며, 토픽 2는 '정보 리터러시', 토픽 3은 '건강 정보', 토픽 4는 '정치 및 소셜미디어', 토픽 5는 '인용 분석', 토픽 6은 '정보 추구 행태', 토픽 7은 'HIV 바이러스', 토픽 8은 '기업 공급망', 마지막으로 토픽 9는 '협력 학습'에 대한 주제로 나타났다.

BERTopic 모형의 상위 10개 토픽을 다른 모형과 비교하면, BERTopic 모형에서 3위에 해당하는 토픽 2의 경우 LDA 모형의 4위인 토

〈표 3〉 BERTopic 모형의 토픽 결과(기본 설정, 상위 10개)

토픽 번호	토픽 단어	문서 (개수)	비율 (%)
0	spatial, land, urban, gis, geographic, data, algorithm, map, landuse, method	1,957	3.53
1	patent, patents, technological, innovation, technology, nanotechnology, patenting, patentcitation, industry, citation	780	1.41
2	informationliteracy, literacy, il, instruction, students, skills, information, learning, teaching, librarians	624	1.13
3	health, healthsciences, librarians, medical, library, sciences, skills, training, librarian, healthinformation	529	0.95
4	political, news, media, twitter, socialmedia, election, social, candidates, online, campaign	498	0.90
5	citation, topic, network, topics, scientific, clustering, science, networks, keywords, clusters	407	0.73
6	information, informationseeking, informationbehaviour, seeking, behaviour, informationscience, science, informationneed, informationbehavior, informationstudies	348	0.63
7	hiv, aids, hivaids, sexual, sex, prevention, condom, men, stigma, hivprevention	310	0.56
8	commerce, supply, supplychain, chain, business, adoption, b2b, firms, smes, edi	302	0.54
9	cscl, learning, collaborative, students, collaborativelearning, group, learners, computersupported, teachers, script	290	0.52
-1	research, information, study, knowledge, paper, use, data, social, library, analysis	24,388	43.99

픽 85와 유사하다 볼 수 있다. 이외는 두 모형의 상위 10위 안에 공통된 토픽은 없다. 또한 Top2Vec 모형과 비교하면 BERTopic 모형은 토픽 0과 3이 거의 동일하며 토픽 5가 유사하지만, 토픽 1과 4를 비롯하여 상위 10위에서 나머지 토픽들이 전혀 다른 주제를 추출하였는데, 이는 두 모형의 알고리즘 차이에 기인한 결과로 생각된다. 또한 BERTopic 모형이 Top2Vec 모형보다 200여개 더 많은 토픽을 추출한 부분과, 토픽 7('HIV 바이러스')와 토픽 9('협력 학습')와 같이 문헌정보학 주제와 다른 복합학 주제를 명확히 분류해 냈다는 사실에 주목할 필요가 있다.

5. 최적화에 의한 토픽 분석

5.1 LDA 모형을 통한 토픽 최적화

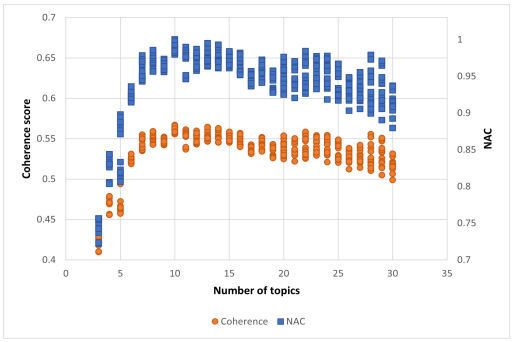
LDA 하이퍼파라미터를 결정하는데 척도로써 사용되는 일관성 점수(예. c_v, c_uci, c_npmi)는 연구자의 의도, 연구 목적, 연구 환경에 따라 달라질 수 있다. 즉, 동일한 데이터셋을 사용하더라도 연구자의 접근 방법에 따라 토픽모델링의 결과에 영향을 미치는 하이퍼파라미터 값은 달라질 수 있으므로 신중한 접근이 필요하다.

LDA 모형의 최적 토픽 개수를 결정하는 방법은, 연구에 따라 접근법이 다소 상이한 형태를 보이는데 그 이유는 같은 데이터를 사용하더라도 토픽모델링을 수행하는 장비의 성능이나 연구 주제에 따라 연구자가 하이퍼파라미터의 설정 범위를 결정할 수 있기 때문이다. 예를 들어 〈그림 2〉는 본 연구에 사용된 데이터로부터 그리드서치(grid search) 기법으로 최적으로 토픽 개수를 찾고자 시도한 결과인데(김선욱, 양기덕, 2022), 최적의 토픽 개수로 10개를 제시하고 있다. 그러나 앞서 수행한 디폴트 모드의 연장선상에서, LDA 모형의 디폴트 모드에서 토픽 개수만 3~50개 범위로 한정하고 토픽모델링을 학습하여 일관성 점수(coherence score)를 측정한 결과 〈그림 3〉, 〈그림 4〉, 〈그림 5〉에서 확인되는 바와 같이 값이 25개이다. 따라서 본 연구에서 전제했던 디폴트 모드를 고려했을 때, 최적의 LDA 토픽 개수도 25개로 결정하는 것이 타당하다 할 수 있다.

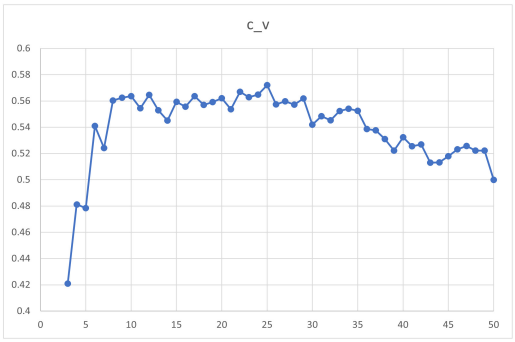
5.2 LDA 모형

앞서 진행했던 디폴트 모드의 LDA 토픽모델링에 토픽 개수 파라미터만을 추가 설정한 뒤 학습하여 얻은 토픽 최적화 결과는 〈표 4〉와 같다. 먼저 토픽 개수를 지정하지 않았을 때 점유율이 가장 높았던 토픽 8('지식 체계 및 담론') 관련 주제가 최적화된 LDA 토픽모델링에서도 마찬가지로 가장 높은 점유율(12.40%)로 나타났다. 그 외에 〈표 1〉에서 일부 '인용분석 및 지표'나 '도서관 장서 및 개발'과 같은 토픽들은 〈표 4〉의 상위 10위에서도 공통으로 나타나지만 토픽 15('취약 계층 지원'), 토픽 17('이용자 행태'), 토픽 11('텍스트 자동분류') 등은 새로이 등장하였다. 대체로 100개의 토픽을 25개로 압축하였기에 1위의 토픽 8을 제외하면 토픽 당 문서의 수나 비율이 상대적으로 많아지고 커진 것을 알 수 있다.

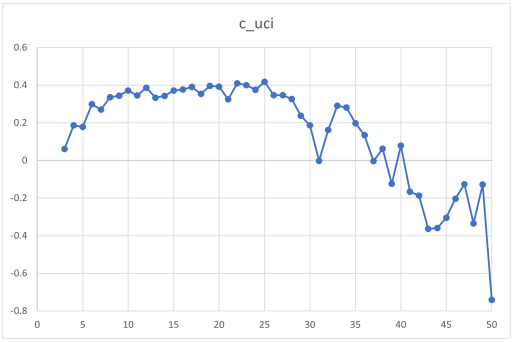
전체 25개 토픽을 점유율 기준으로 상위권 위주로 정리하면 토픽 8('지식 체계 및 담론'), 토



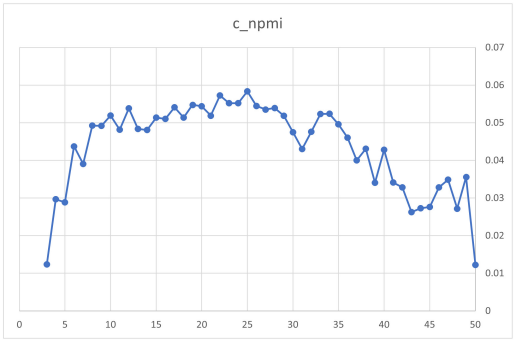
〈그림 2〉 그리드서치를 통한 토픽 개수 결정
(김선욱, 양기덕, 2022)



〈그림 3〉 토픽 개수 변화에 따른 c_v



〈그림 4〉 토픽 개수 변화에 따른 c_uci



〈그림 5〉 토픽 개수 변화에 따른 c_npmi

〈표 4〉 LDA 모형의 토픽 최적화 결과

토픽 번호	연구 주제	토픽 단어	문서 (개수)	비율 (%)
8	지식 체계 및 담론	system, process, technology, framework, approach, development, case, practice, design, management	6,873	12.40
1	학술지 인용 분석	journal, citation, article, publication, author, field, number, science, database, analysis	4,873	8.79
0	문헌정보학 일반	article, science, researcher, field, work, concept, issue, practice, discipline, way	3,914	7.06
10	이용자 요구 기반 도서관 서비스	library, service, resource, user, librarian, use, access, finding, need, university	3,663	6.61
15	취약 계층 지원	community, experience, participant, people, interview, child, life, analysis, finding, support	3,315	5.98
17	이용자 행태	factor, model, user, behavior, effect, finding, trust, relationship, use, intention	2,981	5.38
18	건강 관리	health, patient, care, hospital, use, conclusion, system, health_care, physician, healthcare	2,701	4.87

토픽 번호	연구 주제	토픽 단어	문서 (개수)	비율 (%)
11	텍스트 자동분류	text, classification, method, feature, approach, model, language, domain, image, representation	2,673	4.82
19	지식 경영 및 조직 혁신	knowledge, innovation, organization, finding, performance, capability, firm, knowledge_management, relationship, process	2,549	4.60
21	장서 개발	librarian, program, collection, book, record, development, professional, institution, article, library	2,401	4.33
6	마케팅 및 촉진	business, market, customer, firm, value, service, cost, product, company, industry	2,260	4.08
9	통신망/인터넷 접근 개선 정책	country, policy, internet, technology, development, access, growth, level, infrastructure, region	2,064	3.72
7	정보 검색	search, user, document, system, method, retrieval, query, task, term, performance	2,022	3.65
14	정보 리터러시 교육	student, learning, literacy, education, skill, course, school, university, training, group	1,934	3.49
12	지리 정보	model, method, algorithm, time, approach, location, scale, app, area, distribution	1,925	3.47
5	대학 성과지표	indicator, university, performance, impact, evaluation, measure, patent, security, collaboration, institution	1,741	3.14
22	콘텐츠 플랫폼 및 공유	user, medium, content, platform, privacy, use, device, sharing, video, finding	1,483	2.67
24	건강정보 추구 행태	risk, message, health, woman, gender, behavior, communication, cancer, campaign, effect	1,358	2.45
3	인사 및 조직 관리	group, communication, team, work, employee, member, control, organization, task, project	1,125	2.03
16	웹 사용성 평가	quality, web, website, site, content, web_site, tool, evaluation, usability, question	1,105	1.99
20	네트워크 분석	network, analysis, structure, topic, pattern, method, cluster, map, community, similarity	1,080	1.95
13	공공기관의 서비스와 청렴도	service, government, citizen, participation, governance, sector, city, agency, policy, transparency	613	1.11
4	이용자 리뷰 평가	review, source, software, product, rating, opinion, report, disaster, developer, type	347	0.63
2	경매	event, uncertainty, agent, time, advertising, auction, newspaper, flow, treatment, effect	300	0.54
23	소비자 선호 및 추천	consumer, recommendation, preference, news, channel, item, attribute, choice, shopping, signal	142	0.26

픽 1('학술지 인용 분석'), 토픽 0('문헌정보학 일반'), 토픽 10('이용자 요구 기반 도서관 서비스'), 토픽 15('취약 계층 지원'), 토픽 17('이용자 행태'), 토픽 18('건강 관리'), 토픽 11('텍스트 자동분류'), 토픽 19('지식 경영 및 조직 혁신'), 토픽 21('장서 개발') 등의 순으로 주제가 나타났다.

종합적으로 살펴보면, 상위 토픽부터 점유율

이 비교적 차이를 두며 점차 감소하며 하위 토픽에서는 크기가 작아져 상위 토픽과 하위 토픽에서 점유율에서 편차가 다소 보이는데, 토픽의 규모 측면에서는 군집이 잘 형성된 것으로 보인다. 또한, 최종 도출된 토픽 목록은 상호 중첩되거나 간섭하지 않으므로 이해할 수 있는 수준의 토픽모델링 결과로 판단된다. 참고로 토픽 2('경매')의 경우 『Information Systems Research』를 비롯한 여러 학술지에서 온라인 경매시스템에 관련된 논문이 활발히 게재되고 있기 때문으로 판단된다.

5.3 Top2Vec 모형

LDA 모형과 달리 Top2Vec 모형은 디폴트 모드로 생성되었던 모든 토픽을 토픽 개수에 해당하는 파라미터를 새롭게 지정하면 토픽의 유사도에 따라 재군집하여 새로운 토픽을 병합하는 특성을 갖는다. 물론 처음부터 지정한 개수의 토픽을 생성할 수도 있으나 이렇게 자체적으로 모든 토픽을 생성한 후 이미 생성된 결과에서 병합함으로써 토픽모델링을 처음부터 다시 학습하지 않고 결과를 도출할 수 있어 컴퓨터 처리 측면에서 이득을 가져올 수 있다. 이러한 Top2Vec 모형의 토픽 병합 과정을 사례로 제시하면 〈표 5〉와 같다.

〈표 5〉에서 병합 전의 350개의 토픽 중에서 토픽 37, 토픽 118, 토픽 346, 토픽 2 등은 각각 '교육 요소', '교과과정', '온라인 공개수업', '지도 및 수업' 등의 주제를 나타내는데, 전체 25개의 토픽으로 모델링할 때 이들 세부 주제들이 병합되어 토픽 12를 생성하게 되며, 이는 HDBSCAN 기법으로 구현된다. 구체적인 토픽 단어를 살펴보면 토픽 12는 '학교도서관'과 관련된 전반적인 주제를 의미하는 것으로 보인다. 다만 Top2Vec 패키지는 이 과정에서 기존의 토픽 점유율 같은 특성은 모두 손실되므로, 병합 이전의 토픽 특성을 고려해야 하는 단점을 갖기에 개선이 필요하다.

병합 과정에서 중복된 토픽 단어는 추가 나

〈표 5〉 Top2Vec 토픽 병합 사례

병합	토픽 번호	토픽 단어
전	37	cscl, classroom, argumentation, teacher, classrooms, learners, pedagogical, script, teachers, metacognitive
	118	teacher, teachers, school, classroom, schools, curriculum, classrooms, students, elementary, teaching
	345	moocs, mooc, courses, massive, course, learners, learner, instruction, taught, instructional
	2	instruction, instructional, course, student, teaching, courses, students, instructor, teach, classroom
후	12	instruction, instructional, student, students, classroom, course, teaching, courses, curriculum, pedagogical, instructor, instructors, teach, pedagogy, teacher, undergraduate, teachers, assignments, taught, il, acrl, literacy, skills, semester, graduate, tutorials, learners, school, librarians, educators, curricular, classrooms, pbl, librarian, faculty, writing, educational, sessions, lecture, introductory, curricula, learning, lesson, rubric, tutorial, learner, cscl, shot, skill, education

열되지 않으므로, 병합 대상의 수가 많다고 하더라도 병합된 토픽의 토픽 단어 수가 이에 비례한다고 볼 수는 없다. Top2Vec 모형을 디폴트 모드에서 학습하고 추가로 토픽 개수 파라미터를 25개로 설정하여 최적화한 토픽모델링 결과를 정리하고 토픽 단어의 경우 10개 내외로 제시하면, 〈표 6〉과 같다.

Top2Vec 모형의 전체 25개 토픽을 점유율이 높은 상위권 중심으로 정리하면 토픽 0('철학 및 인식론'), 토픽 1('기업 비즈니스'), 토픽 2('기록학'), 토픽 3('건강 정보'), 토픽 4('대학도서관 이용교육'), 토픽 5('정보검색 및 자동분류'), 토픽 6('소셜미디어 분석'), 토픽 7('의료 기록'), 토픽 8('계량정보학'), 토픽 9('인용색인 데이터베이스') 등의 순으로 주요 주제가 나타났다. Top2Vec 모형의 상위권의 우세 토픽은 LDA 모형보다 좀더 세분화되어 보다 구체적인 주제를 잘 드러내는 것으로 보인다.

Top2Vec 모형은 전체적으로 상위 토픽부터 점유율이 급격한 감소 없이 비교적 점차 감소하며 토픽 군집이 잘 형성된 것으로 보인다. LDA 모형과 비교하면 Top2Vec 모형의 토픽 최적화는 토픽간의 점유율 측면에서 편차가 많이 작아진 것을 알 수 있다.

〈표 6〉 Top2Vec 모형의 토픽 최적화 결과

토픽 번호	연구 주제	토픽 단어	문서 (개수)	비율 (%)
0	철학 및 인식론	epistemology, philosophical, epistemological, foucault, habermas, hermeneutics, critique, ontological, realist, sociomaterial, ...	2,559	4.62
1	기업 비즈니스	firm, business, capabilities, strategic, capability, agility, outsourcing, rbv, executives, investments, ...	4,432	7.99
2	기록학	nineteenth, century, twentieth, historians, preservation, archives, historical, archivists, founding, essay, ...	3,185	5.74
3	건강 정보	illness, semistructured, caring, experiences, caregiving, narratives, living, schizophrenia, caregivers, distress, ...	3,966	7.15
4	대학도서관 이용교육	librarians, faculty, library, colleges, instruction, graduate, students, liaison, campuses, academic, skills, university, curriculum, instructional, ...	3,580	6.46
5	정보검색 및 자동분류	retrieval, supervised, sentence, classifiers, corpus, vector, query, embeddings, trec, baselines, wsd, extraction, svm, unsupervised, precision, ...	2,521	4.55
6	소셜미디어 분석	snss, sns, facebook, prosocial, gratifications, microblogs, media, instagram, technostress, microblogging, ...	3,613	6.52
7	의료 기록	ambulatory, ehr, patient, medication, clinical, prescribing, outpatient, clinician, alerts, cpoe, dosing, ...	2,575	4.64
8	계량정보학	hirsch, citation, indicator, index, rankings, bibliometric, egghe, scientometric, jif, percentile, normalized, ...	2,468	4.45
9	인용색인 데이터베이스	journals, publication, articles, scopus, scholarly, editorial, citations, publisher, indexed, mendeley, wos, citation, ...	3,355	6.05
10	서지 네트워크	serials, oclc, print, catalog, consortial, acquisitions, interlibrary, cataloging, holdings, purchased, aggregators, ...	2,228	4.02

토픽 번호	연구 주제	토픽 단어	문서 (개수)	비율 (%)
11	국가의 연구성과	countries, publications, bibliometric, authored, productive, usa, scientific, domestic, output, international, ...	2,066	3.73
12	학교도서관	instruction, student, classroom, course, teaching, curriculum, pedagogical, instructor, teacher, assignments, taught, il, ...	1,501	2.71
13	지리 정보	spatial, raster, elevation, polygon, land, areal, spatially, dem, lidar, terrain, topographic, gis, ...	1,441	2.60
14	지식 경영	km, tacit, knowledge, organizational, kmrp, cops, codification, subsidiaries, seci, npd, ...	1,063	1.92
15	통신망	broadband, penetration, telephony, competition, fixed, telecom, cable, unbundling, subscribers, nga, wireless, ...	1,421	2.56
16	건강정보 추구 행태	whites, hispanics, antismoking, adolescents, smoking, colorectal, hispanic, messages, adults, smokers, cancer, ...	1,413	2.55
17	민주주의와 정치	citizen, democratic, political, government, civic, election, deliberation, presidential, public, trump, ...	1,560	2.81
18	기술 수용 행태	tam, equation, intention, ease, usefulness, continuance, sem, antecedents, utaut, acceptance, ttf, ...	1,594	2.88
19	특허 정보	patent, citespace, bibliometric, assignees, uspto, scientometric, patenting, trademark, inventors, cocitation, ...	1,797	3.24
20	검색엔진	search, engine, searcher, query, interface, navigational, reformulation, retrieval, bing, browsing, yahoo, session, ...	1,545	2.79
21	시맨틱 웹	ontology, semantic, thesaurus, metadata, skos, semantics, xml, vocabulary, rdf, schema, folksonomies, dublin, ...	1,396	2.52
22	의학도서관	amia, informatics, librarian, workshop, nhs, outreach, partnerships, director, liaison, nlm, librarianship, staff, coordinator, bioinformatics, ...	918	1.66
23	개인정보 및 보안	security, threat, unauthorized, privacy, breaches, attacks, protection, deterrence, violation, malware, cybersecurity, ...	1,509	2.72
24	소비자 행태	purchase, consumer, product, seller, shopping, retailer, buyers, sales, price, commerce, ebay, loyalty, ...	1,736	3.13

5.4 BERTopic 모형

BERTopic 모형의 디폴트 모드 토픽모델링에 토픽 개수 파라미터만을 추가 설정한 뒤 학습한 결과를 정리하면 〈표 7〉과 같다. LDA와 Top2VEC과 달리 BERTopic은 이상치를 구분하여 배제하는 특성을 가지는데, 토픽 개수를 명시적으로 설정하더라도 이상치 비율은 디폴트 모드와 다를 바 없는 43.99%를 유지했다.

25개의 토픽으로 최적화하여 BERTopic 모형이 생성한 토픽을 점유율이 높은 상위권 중심으로 정리하면 토픽 0(‘문헌정보학 일반’), 토픽 1(‘대학도서관과 정보 리터러시’), 토픽 2(‘인용 분석’), 토픽 3(‘건강 정보’), 토픽 4(‘정보 검색’), 토픽 5(‘지리 정보’), 토픽 6(‘건강 정보관리’), 토픽 7(‘통신망’), 토픽 8(‘개인정보 및 보안’), 토픽 9(‘온라인 마켓 플랫폼’) 등의 순으로 주요 주제가 나타났다. 앞서 두 모

〈표 7〉 BERTopic 모형의 토픽 최적화 결과

토픽 번호	연구 주제	토픽 단어	문서 (개수)	비율 (%)
0	문헌정보학 일반	knowledge, information, study, research, social, paper, use, findings, model, data	8,407	15.16
1	대학도서관과 정보 리터러시	library, libraries, information, librarians, data, research, study, students, literacy, academic	6,407	11.56
2	인용 분석	research, journals, citation, journal, science, scientific, citations, impact, patent, articles	4,346	7.84
3	건강 정보	health, information, care, study, cancer, patients, use, hiv, research, literacy	3,216	5.80
4	정보 검색	based, retrieval, search, method, information, results, proposed, query, documents, approach	2,031	3.66
5	지리 정보	spatial, data, model, based, land, urban, gis, method, geographic, time	1,961	3.54
6	건강 정보관리	clinical, data, objective, methods, drug, medical, results, medication, patient, health	1,235	2.23
7	통신망	broadband, telecommunications, mobile, market, spectrum, competition, policy, paper, network, access	864	1.56
8	개인 정보 및 보안	privacy, security, information, data, informationsecurity, disclosure, users, personal, concerns, study	790	1.42
9	온라인 마켓 플랫폼	blockchain, market, price, online, crowdsourcing, platform, information, crowdfunding, product, auction	673	1.21
10	코로나19	disaster, emergency, covid, covid19, crisis, 19, information, media, social, response	350	0.63
11	인공지능과 윤리	robots, ai, moral, human, robot, ethical, social, article, artificial, ethics	322	0.58
12	보존 처리	paper, ink, ageing, cellulose, iron, treatment, samples, degradation, conservation, deacidification	254	0.46
13	클라우드 컴퓨팅	cloud, cloudcomputing, computing, adoption, services, service, security, cloudservices, model, factors	158	0.28
14	음악 정보	music, musical, musicinformation, information, classification, mood, informationseeking, seeking, retrieval, user	86	0.16
15	의사 결정	fuzzy, criteria, decision, selection, supplier, approach, method, proposed, supply, supplychain	67	0.12
16	무선 주파수 기술	rfid, radiofrequency, radiofrequencyidentification, frequencyidentification, identification, radio, technology, identificationrfid, frequencyidentificationrfid, frequency	65	0.12
17	감정 및 신경과학	emotion, emotions, neurois, fmri, neurols, neuroscience, research, brain, emotional, neurophysiological	36	0.06
18	IPP	distribution, distributions, ipp, informetric, function, concentration, lotkaian, ipps, informetrics, leimkuhler	25	0.05
19	인도학	indian, treaties, indianaffairs, tribes, treaty, affairs, federal, american, states, americanindian	20	0.04
20	스프레드시트	spreadsheet, spreadsheets, errors, spreadsheetapplications, applications, error, mt, spreadsheeterror, simulations, errorcorrection	16	0.03

토픽 번호	연구 주제	토픽 단어	문서 (개수)	비율 (%)
21	-	works, initiated late constance, new editions standard, standard works provided, set semiannual, sheehy, set semiannual series, purpose list, brief roundup new, comprehensive brief	15	0.03
22	정신 건강	schizophrenia, neurocognitive, cognitive, impairment, brain, nmda, dopamine, daily, dailytasks, symptoms	14	0.03
23	동영상 검색	video, image, features, semantic, question, neural, caption, retrieval, level, deep	14	0.03
24	공학 인증	engineering, competencies, graduates, education, skills, engineers, competences, curriculum, engineeringeducation, professional	11	0.02
-1	-	information, research, study, data, knowledge, paper, use, social, based, results	24,059	43.39

형과 비교하면 BERTopic 모형의 전체적인 결과는 상위 20%의 토픽(1위부터 5위까지 해당)은 유사하지만, 하위 50%의 토픽들(대략 15위 이후에 해당)은 다른 모형에서 찾아보기 힘든 주제들에 해당하였다. 이에 대한 방증으로 토픽의 점유율 내지 논문 수를 보면 이들 토픽들이 현격히 적은 것을 알 수 있다. 즉 최적화에 따른 BERTopic 모형의 결과는 토픽간의 점유율 측면에서 매우 큰 편차를 보인다.

그밖에 토픽 21은 주제를 판단할 수 없었는데, 이 토픽에 부여된 원본 데이터를 확인한 결과 그 이유는 『Association of College & Research Libraries』가 발행하는 『Selected reference books』 시리즈[2]에 있었다. BERTopic은 동일한 문장이 반복되는 15건의 논문 초록과 그 어휘를 중요하다고 판단하여 하나의 토픽을 생성한 것으로 보인다.

5.5 모형에 따른 비교

앞서 세 모형 LDA, Top2Vec, BERTopic의 기본 설정을 이용한 디폴트 모드에서 토픽모델링을 수행한 결과와, 이 연구의 실험 데이터에 대해 최적의 토픽 개수를 찾아내고 그 값을 명시적으로 설정하여 각 모형에서 토픽모델링을 하여 얻을 결과를 제시하였다. 이들 결과를 통해 이 세 모형이 각각의 단계에서 생성한 토픽 모델링 결과가 얼마나 차이가 나는지를 파악하기 위해 세 모형에 대해 각 토픽 별로 부여된 문서 개수에 대한 기술통계를 계산하면 〈표 8〉과 같다.

기본 설정 단계에서 각각의 모형은 매우 다른 토픽 결과를 가져온 것을 알 수 있다. LDA, Top2Vec, BERTopic 모형 각각이 100개, 350, 550개의 토픽을 생성했으며, 각 토픽에 부여된 문서 개수의 평균도 554개, 158개, 56개로 매우 차이가 큼을 알 수 있다. 즉 LDA 모형의 경우

2) 이 서지레코드는 전체 데이터에서 15건에 해당하는데, 다음과 같은 문구로 시작하는 초록을 공통으로 포함한다. *This article follows the pattern set by the semiannual series initiated by the late Constance M. Winchell more than ** years ago and continued by Eugene Sheehy.*

〈표 8〉 토픽 내의 문서 수에 대한 기술통계

모형	기본 설정 단계			최적화 단계		
	LDA	Top2Vec	BERTopic*	LDA	Top2Vec	BERTopic*
topic 수	100	350	550	25**		
mean	554.42	158.41	56.46	2,217.68	2,217.68	1,255.32
std	1,137.29	172.26	109.99	1,521.83	972.45	2,171.95
min	20	23	10	142	918	11
25%	144.5	58.25	16	1,125	1,501	25
50%	262.5	104.5	27.5	2,022	1,797	254
75%	543	196.25	56.75	2,701	2,575	1,235
max	10,507	1,391	1,957	6,873	4,432	8,407

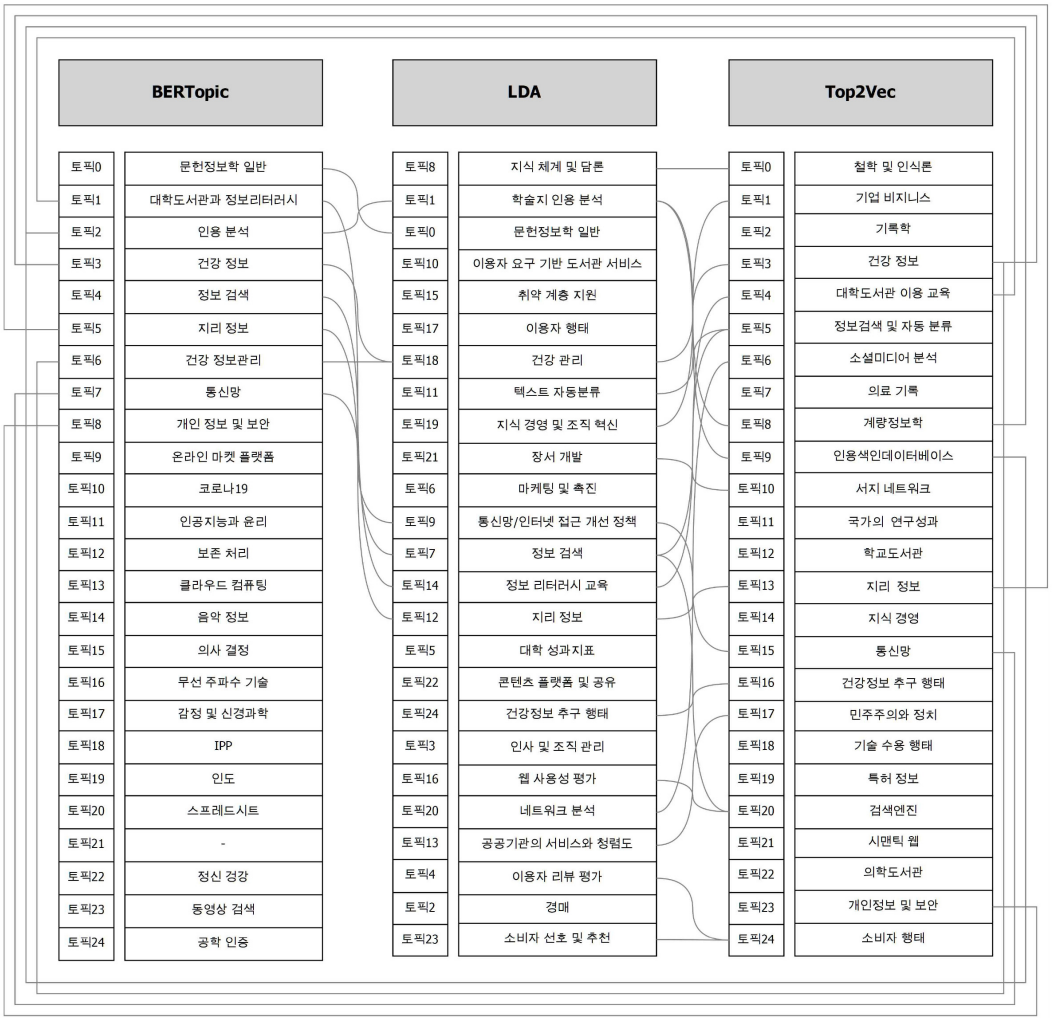
* 이상치를 제외하고 계산한 통계치임
** 최적의 토픽 개수를 동일하게 적용

비교적 적은 수의 토픽을 생성하고 한 토픽 당 평균적으로 554개의 많은 문서가 부여되었으며, 이와 반대로 BERTopic은 많은 수의 토픽을 생성하고 한 토픽에 대해 비교적 작은 수의 문서를 가지는 경향을 유추해 볼 수 있다. 또한 두 모형 중에 LDA는 상대적으로 토픽 당 부여된 문서 수에서 토픽 사이에 매우 큰 편차를 보인 반면, BERTopic은 그렇지 않은 모습을 보였다. BERTopic이 전체 문서에 44%에 해당하는 문서를 이상치로 부여하여 이 기술통계에서 제외하였기에 LDA보다 상대적으로 작은 통계치를 보일 수는 있으나 기술통계량에서 보이는 매우 큰 차이는 다수 토픽의 생성에 따른 것으로 보인다.

한편 최적화된 토픽 수를 25개로 설정하여 얻은 토픽모델링의 결과에 대해서는 상대적으로 Top2Vec 모형이 토픽 당 문서 수의 평균에서 큰 값을 보이고 토픽 간 차이를 나타내는 표준편차에서 다소 작은 값을 보여 25개의 토픽에 대해 고르게 문서가 분포되어 있는 것을 알 수 있다. 다만 이러한 모습에 대해 각 토픽에 부여된 문서들이 올바른지를 평가하지 않았기에 Top2Vec 모형의 토픽모델링 결과가 더 우수하다는 것을 의미하지는 않는다. 세 모형 사이에 동일한 토픽 개수를 설정하였기에 Top2Vec 모형이 편차가 적은 모델링 결과를 가져온다는 것만은 확실하다.

최적화 토픽의 개수를 25개로 설정한 후 세 모형을 이용해 토픽모델링하여 생성한 결과를 대상으로 서로 공통된 또는 유사한 토픽을 식별하고 이를 연결하여 분석한 결과는 〈그림 6〉과 같다.

세 모형 서로 간에 공통된 토픽을 산출하면 LDA와 Top2Vec 모형이 토픽간의 유사 정도가 18개(72%)로 상당히 높았으며 LDA와 BERTopic 모형이 8개(32%), Top2Vec과 BERTopic 모형이 8개(32%) 순이었다. 세 모형이 모두 공통으로 생성한 토픽은 전체 25개 중에서 모두 7개(28%)인데, 구체적인 토픽의 연구 주제로 계량정보학과 관련된 인용 분석, 건강 정보, 정보검색과 자동분류, 지리 정보, 정보 리터러시, 통신망 등이 해당하였다. 이들 주제 중의 일부는 각 모형별 유사한 토픽을 구성하는 문서의 수도 큰 차이를 보이지 않았다. 예를 들어 LDA



〈그림 6〉 BERTopic, LDA, Top2Vec 모형간 유사 토픽 관계

와 BERTopic 모형에서 '지리 정보' 토픽의 경우 약 1천 9백여 건, '정보검색'은 2천여 건이 식별되었다. 따라서 주요 7개 영역의 주제는 문헌정보학 내에서 최근 20년 동안 주로 많이 논의된 비교적 핵심 주제 영역으로 볼 수 있다. 모형 측면에서는 LDA와 Top2Vec 모형이 공통되거나 유사한 토픽이 다수이면서 토픽의 점유율 순위와 무관하게 골고루 분포되어 있어 이 두 모형이 토픽모델링에서 비슷한 측면이 있음을 알 수 있다.

LDA와 달리, BERTopic과 Top2Vec 모형은 작위적으로 토픽 개수를 설정하지 않아야 각 모형에서 설계된 알고리즘에 따라 의도하는 최적의 결과를 만들어 내도록 구현되었다. 두 모형은 임베딩 기반의 토픽모델링이라는 이론적 공통점에도 불구하고 세부적인 알고리즘의

차이가 있다. 이는 모형별 최적화 결과에서 두 모형의 공통된 토픽은 세 모형의 공통된 7개 주제와 함께 두 모형만의 공통된 1개 토픽('개인정보 및 보안')을 더해 전체가 8개 밖에 안 되는 것에서 알 수 있다.

한편 최적화를 통해 토픽 개수를 모두 같게 설정했을 때는, 세 모형에서 LDA와 유사한 결과를 보인 모형은 Top2Vec이었다. 사실 모형 구조 측면에서 보면 Top2Vec 모형과 BERTopic 모형은 임베딩의 사용하고 UMAP를 적용하여 차원 축소하고 토픽 형성을 위해 HDBSCAN의 사용하는 등 유사한 접근 방식을 취하는데 이로 인해 토픽 수를 자동으로 찾고 대부분의 경우 전처리가 필요하지 않는다. 하지만 실제 동일한 실험 데이터셋을 적용하여 얻은 두 모형의 토픽모델링 결과는 다소 상이한 모습을 보였으며, 이는 두 모형의 세부적인 차이로 기인한다고 볼 수 있다. 구체적으로 Top2Vec 모형이 차원 축소 전의 임베딩 공간에서 문헌 벡터의 밀집 영역의 대한 센트로이드(centroid) 기법을 적용하여 주위 단어로 토픽 벡터를 생성하는 반면, BERTopic 모형은 토픽을 정확히 표현하기 위해 군집화 이후에 클러스터 수준에서 변형된 TF-IDF 기법을 적용하여 토픽 표현을 생성한다. 따라서 두 모형이 토픽을 표현하는 방식이나 찾아내는 시점이 다르다 할 수 있다.

마지막으로 BERTopic 모형이 추출한 토픽 결과는 다른 모형과 상당한 차이를 보였는데, 이는 BERTopic 모형의 기능적 불리함을 의미하는 것이 아니라, 토픽모델링을 수행함에 있어 연구자가 자신의 의도에 부합하게 결과를 가져오기 위해 적합한 모형을 선정해야 할 필요성이 있음을 제시하는 것이라고 보는 것이 타당하다. 특히 BERTopic 모형에서 주목해야 할 부분으로 토픽에 할당된 문서에 대한 기술통계(〈표 8〉)과 토픽모델링의 결과(〈표 3〉과 〈표 7〉)에서 보이듯이, 이 모형은 LDA 확률모델에서 고려되지 못한 다양하고 세부적인 토픽을 식별 가능한 것으로 보인다. 즉 BERTopic 모형은 주제를 보다 명료하게 구분되도록 하여 더 많은 토픽을 찾아내는 토픽 세분화 또는 토픽 세분성(topic granularity)을 높이는 기능을 가졌다고 할 수 있다.

6. 결 론

본 연구는 최근 토픽모델링 관련 연구에서 사용되는 주된 기법인 LDA, Top2Vec, BERTopic 모형의 특성을 파악하기 위해 2001년 1월부터 2021년 10월까지 Web of Science에 등재된 문헌정보학 분야 85종의 학술지에 게재된 논문 55,442건의 초록을 대상으로 토픽모델링을 수행하고 결과를 분석한 뒤 세 모형을 비교・분석하였다. 실험 과정은 크게 두 단계로 나누어지는데, 첫 단계는 각 모형에 대해 파라미터를 변경하지 않고 기본 설정을 그대로 적용하여 1차 토픽모델링을 수행하였으며, 두 번째 단계는 최적의 토픽 수에 따라 각 모형을 2차 토픽모델링하고 그 결과를 직접 비교・분석하였다. 최적의 토픽 수를 선정하기 위해 실험데이터에 대한 일관성 점수를 활용하였다. 이에 연구 결과는 다음과 같다.

첫째, 기본 설정을 그대로 적용하여 1차 토픽모델링을 수행한 결과, 개별 모형들의 특성은 다음과 같다. ① LDA 모형은 전처리된 데이터

를 입력 받아 100개의 토픽을 생성하였으며 전처리 후 토픽모델링을 실시했음에도 불구하고 점유율 1위인 토픽 8('지식 체계 및 담론')은 전체 55,442건에서 10,505건(19%)의 문서가 부여되어 세 모형의 모든 토픽에서 가장 높은 비율을 보였다. 이 토픽은 논문에서 지식 체계를 나타내기 위해 주로 사용되는 단어들로 구성되었다. ② Top2Vec 모형은 Doc2Vec을 사용하여 논문의 초록을 임베딩하였고, 토픽모델링 학습 결과 총 350개의 토픽을 생성하였다. 초록 텍스트에 대한 전처리를 거치지 않았음에도 불구하고 세부 주제가 효과적으로 분산된 토픽모델링을 수행하였다. ③ BERTopic 모형은 초록 텍스트를 임베딩하는 과정에서 사전학습모델인 센텐스-트래스포머의 'all-MiniLM-L6-v2'가 사용되었고, 토픽모델링 학습 결과 총 550개의 토픽이 생성되었다. 기본 설정으로 인해 BERTopic의 특성에 따라 전체 데이터의 44%가 이상치(outliers)로 분류되었음에도 불구하고, LDA와 Top2Vec 모형에 비해 다양한 주제를 분류해 냈다. 세 모형에 대해 토픽을 많이 찾아내는 또는 작게 분할하는 측면을 고려한다면, LDA 모형보다 Top2Vec이나 BERTopic 모형은 3배나 5배 정도 토픽을 잘게 세분한다고 할 수 있다. 만약 이렇게 세분된 토픽이 주제 적합성에 잘 부합하거나 더 명료하다면 토픽 활용에 많은 도움이 될 것이다.

둘째, 실험데이터에 대한 최적의 토픽 수를 찾아 세 모형에 설정하고 토픽모델링을 실행하여 그 결과를 통해 모형을 비교하고자, LDA 모형의 파라미터(c_v, c_npmi, c_uci, u_mass)를 중심으로 최고의 일관성 점수를 가져오는 값으로 25를 얻었으며 이를 최적의 토픽 수로 설정하였다.

셋째, 최적의 토픽 수를 25로 설정하되 나머지 설정은 1차 토픽모델링과 동일하게 2차 토픽모델링을 수행한 결과, 개별 모형들의 특성은 다음과 같다. ① LDA 모형에서는 여전히 토픽 8('지식 체계 및 담론')이 가장 높은 점유율을 보이지만 그 비율이 12.4%로 많이 낮아졌으며, 나머지 24개 토픽은 상위 토픽부터 점유율에서 비교적 차이를 두며 감소하면서 편차를 보여 토픽의 규모 측면에서는 군집이 잘 형성된 것으로 보인다. ② Top2Vec 모형은 재학습하지 않고 토픽 개수를 명시적으로 지정하여 토픽간의 유사도에 따라 재군집하여 새로운 토픽을 병합하였는데, 이 모형의 상위권의 우세 토픽은 LDA 모형보다 좀더 세분화되어 보다 구체적인 주제를 잘 드러내는 것으로 나타났다. ③ BERTopic 모형은 최적화된 토픽 수를 설정하더라도 이상치 비율은 기본 설정과 다를 바 없이 동일하게 유지하였다. 이는 성능상의 문제라기보다는 BERTopic 모형이 가지는 특성이며, 이상치로 배제된 데이터를 제외하고 남은 데이터를 대상으로 토픽별 변별력을 높이는 것으로 판단된다.

넷째, 앞서 제시된 두 단계의 토픽모델링 결과를 세 모형 관점에서 비교하면, ① 기본 설정에서는 세 모형은 각기 매우 다른 크기의 토픽 개수를 가져왔으며 토픽 당 문서 수의 평균이나 표준편차에서 많은 차이가 났다. LDA 모형은 비교적 적은 수의 토픽에 많은 문서를 부여하는 반면, BERTopic 모형은 반대의 경향을 보였다. 최적화된 토픽모델링에서는 상대적으로 다른 모형에 비해 Top2Vec 모형이 평균적으로 토픽 당 많은 문서를 부여하며, 토픽간의 편차가 다

소 작아 25개의 토픽에 대해 고르게 문서가 분포됨을 알 수 있다. 다만 세 모형의 비교에서 각 토픽에 부여된 문서들이 주제적으로 올바르게 형성되었는지 추가적인 평가가 필요하기에 이러한 결과를 가지고 우열을 가리는 것은 유의미하지 않다. ② 각 모형별로 주제적으로 공통되거나 유사한 토픽을 생성했는지를 비교해보면, LDA와 Top2Vec 모형이 25개 중에 18개(72%)로 상당히 높았으며, LDA와 BERTopic 모형 8개(32%), Top2Vec과 BERTopic 모형 8개(32%) 순이었다. Top2Vec과 BERTopic 모형이 거의 유사한 방식의 알고리즘으로 설계되었음에도 토픽모델링 결과에서는 차이가 남을 알 수 있다. 오히려 Top2Vec 모형은 LDA와 더 유사한 결과를 보인 것을 알 수 있다. ③ BERTopic 모형은 주제를 보다 명료하게 구분하고 더 많은 토픽을 찾아내어 높은 토픽 세분성을 보인다고 할 수 있다.

본 연구는 토픽모델링 분야에서 자주 언급되는 LDA, Top2Vec, BERTopic 세 모형을 사용하여 대용량 서지데이터를 대상으로 토픽모델링을 수행하고 그 결과를 비교하여 세 모형이 갖는 기본적인 특성을 파악하고 동일하게 토픽을 생성하도록 설정하여 각 모형간의 차이를 분석하였다. 이 결과를 통해 토픽모델링 기법에 익숙하지 않거나 간단하게 활용하고자 하는 연구자에게 기본 정보와 안내를 제공할 것으로 본다. 아울러 Top2Vec와 BERTopic 모형처럼 인공 신경망을 이용하는 비교적 최근에 등장한 모형에 대해, 해당 분야 전문가가 직접 토픽모델링 결과를 평가하여 모형을 최적화를 할 수 있는 보다 다양한 후속 연구를 기대해 본다.

참 고 문 헌

김선욱, 양기덕 (2022). LDA와 BERTopic을 이용한 토픽모델링의 증강과 확장 기법 연구. 정보관리학회지, 39(3), 99-132. http://doi.org/10.3743/KOSIM.2022.39.3.099

김선욱, 양기덕, 이혜경 (2022). 다이나믹토픽모델링을 활용한 문헌정보학 분야의 토픽 변화 분석. 한국도서관·정보학회지, 53(2), 265-284. http://doi.org/10.16981/kliss.53.2.202206.265

김태경, 김창식 (2018). 텍스트마이닝을 이용한 정보보호 연구동향 분석. 디지털산업정보학회논문지, 14(2), 19-25. http://dx.doi.org/10.17662/ksdim.2018.14.2.019

박자현, 송민 (2013). 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석. 정보관리학회지, 30(1), 7-32. https://doi.org/10.3743/KOSIM.2013.30.1.007

박준형, 오효정 (2017). 국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법 비교: LDA와 HDP를 중심으로. 한국도서관·정보학회지, 48(4), 235-258. https://doi.org/10.16981/kliss.48.4.201712.235

이지용, 최유리, 김대건, 이승박 (2022). 스포츠 영역에서 나타나는 폭력의 유형: 텍스트 마이닝을 적용한 판례분석. 한국체육학회지, 61(5), 43-54. http://dx.doi.org/10.23949/kjpe.2022.09.61.5.4

임정훈 (2022). 키워드 네트워크 분석과 토픽모델링을 활용한 정보활용교육 연구 동향 분석. 정보관리

학회지, 39(4), 23-48. http://dx.doi.org/10.3743/KOSIM.2022.39.4.023

Ali, I. & Naeem, M. A. (2022). Identifying and profiling user interest over time using social data. In 2022 24th International Multitopic Conference (INMIC), 1-6. https://doi.org/10.1109/INMIC56986.2022.9972955

Angelov, D. (2020). Top2Vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470. https://doi.org/10.48550/arXiv.2008.09470

Blei, D. & Lafferty, J. (2005). Correlated topic models. Advances in Neural Information Processing Systems, 18, 147-154.

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

Chen, A. T., Sheble, L., & Eichler, G. (2013). Topic modeling and network visualization to explore patient experiences. In Visual Analytics in Healthcare Workshop 2013.

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. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6〈391::AID-ASI1〉3.0.CO;2-9

Dur, B. I. U. (2014). Data visualization and infographics in visual communication design education at the age of information. Journal of Arts and Humanities, 3(5), 39-50. https://doi.org/10.18533/journal.v3i5.460

Egger, R. & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. Frontiers in Sociology, 7, 886498. https://doi.org/10.3389/fsoc.2022.886498

Gao, Q., Huang, X., Dong, K., Liang, Z., & Wu, J. (2022). Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec. Scientometrics, 127, 1543-1563. https://doi.org/10.1007/s11192-022-04275-z

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 50-57. https://doi.org/10.1145/3130348.3130370

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

Jing, X. Y., Zhang, D., & Tang, Y. Y. (2004). An improved LDA approach. IEEE Transactions

on Systems, Man, and Cybernetics, Part B (Cybernetics), 34(5), 1942-1951. https://doi.org/10.1109/tsmcb.2004.831770

Li, C., Lu, Y., Wu, J., Zhang, Y., Xia, Z., Wang, T., Yu, D., Chen, X., Liu, P., & Guo, J. (2018). LDA meets Word2Vec: A novel model for academic abstract clustering. In Proceedings of the 2018 Web Conference Companion (WWW '18 Companion), 1699-1706. https://doi.org/10.1145/3184558.3191629

Li, W. & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd International Conference on Machine Learning, 577-584. https://doi.org/10.1145/1143844.1143917

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 889-892. https://doi.org/10.1145/2484028.2484166

Moody, C. E. (2016). Mixing Dirichlet topic models and word embeddings to make lda2vec. arXiv preprint arXiv:1605.02019. https://doi.org/10.48550/arXiv.1605.02019

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982-3992. https://doi.org/10.18653/v1/D19-1410

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too!. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP), 1728-1736. https://doi.org/10.18653/v1/2020.emnlp-main.135

Vayansky, I. & Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 1-15. https://doi.org/10.1016/j.is.2020.101582

Yuan, C. & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. J, 2(2), 226-235. https://doi.org/10.3390/j2020016

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Kim, SeonWook & Yang, Kiduk (2022). Topic model augmentation and extension method using LDA and BERTopic. Journal of the Korean Society for Information Management, 39(3),

99-132. http://doi.org/10.3743/KOSIM.2022.39.3.099

Kim, SeonWook, Yang, Kiduk, & Lee, HyeKyung (2022). Analysis of research topic trend in library and information science using dynamic topic modeling. Journal of Korean Library and Information Science Society, 53(2), 265-284. http://doi.org/10.16981/kliss.53.2.202206.265

Kim, Tae Kyung & Kim, Changsik (2018). Research trends analysis of information security using text mining. Journal of the Korea Society of Digital Industry and Information Management, 14(2), 19-25. http://dx.doi.org/10.17662/ksdim.2018.14.2.019

Lee, Ji-Yong, Choi, You Lee, Kim, Dae Geon, & Lee, Seungbak (2022). Types of violence appearing in the sports field: case law analysis using text mining. The Korean Journal of Physical Education, 61(5), 43-54. http://dx.doi.org/10.23949/kjpe.2022.09.61.5.4

Lim, Jeonghoon (2022). Analysis of research trends in information literacy education using keyword network analysis and topic modeling. Journal of the Korean Society for Information Management, 39(4), 23-48. http://dx.doi.org/10.3743/KOSIM.2022.39.4.023

Park, Jahyun & Song, Min (2013). A study on the research trends in Library & Information Science in Korea using topic modeling. Journal of the Korean Society for Information Management, 30(1), 7-32. https://doi.org/10.3743/KOSIM.2013.30.1.007

Park, JunHyeong & Oh, Hyo-Jung (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. https://doi.org/10.16981/kliss.48.4.201712.235