LDA와 BERTopic을 이용한 토픽모델링의 증강과 확장 기법 연구*

Topic Model Augmentation and Extension Method using LDA and BERTopic

김선욱 (SeonWook Kim)**
양기덕 (Kiduk Yang)***

초 록

본 연구의 목적은 LDA 토픽모델링 결과와 BERTopic 토픽모델링 결과를 합성하는 방법론인 Augmented and Extended Topics(AET)를 제안하고, 이를 사용해 문헌정보학 분야의 연구주제를 분석하는 데 있다. AET의 실제 적용결과를 확인하기 위해 2001년 1월부터 2021년 10월까지의 Web of Science 내 문헌정보학 학술지 85종에 게재된 학술논문 서지 데이터 55,442건을 분석하였다. AET는 서로 다른 토픽모델링 결과의 관계를 WORD2VEC 기반 코사인 유사도 매트릭스로 구축하고, 매트릭스 내 의미적 관계가 유효한 범위 내에서 매트릭스 재정렬 및 분할 과정을 반복해 증강토픽(Augmented Topics, 이하 AT)을 추출한 뒤, 나머지 영역에서 코사인 유사도 평균값 순위와 BERTopic 토픽 규모 순위에 대한 조화평균을 통해 확장토픽(Extended Topics, 이하 ET)을 결정한다. 최적 표준으로 도출된 LDA 토픽모델링 결과와 AET 결과를 비교한 결과, AT는 LDA 토픽모델링 토픽을 한층 더 구체화하고 세분화하였으며 ET는 유효한 토픽을 발견하였다. AT(Augmented Topics)의 성능은 LDA 이상이었으며 ET(Extended Topics)는 일부 경우를 제외하고 대부분 LDA와 유사한 수준의 성능을 나타내었다.

ABSTRACT

The purpose of this study is to propose AET (Augmented and Extended Topics), a novel method of synthesizing both LDA and BERTopic results, and to analyze the recently published LIS articles as an experimental approach. To achieve the purpose of this study, 55,442 abstracts from 85 LIS journals within the WoS database, which spans from January 2001 to October 2021, were analyzed. AET first constructs a WORD2VEC-based cosine similarity matrix between LDA and BERTopic results, extracts AT (Augmented Topics) by repeating the matrix reordering and segmentation procedures as long as their semantic relations are still valid, and finally determines ET (Extended Topics) by removing any LDA related residual subtopics from the matrix and ordering the rest of them by F_1 (BERTopic topic size rank, Inverse cosine similarity rank). AET, by comparing with the baseline LDA result, shows that AT has effectively concretized the original LDA topic model and ET has discovered new meaningful topics that LDA didn't. When it comes to the qualitative performance evaluation, AT performs better than LDA while ET shows similar performances except in a few cases.

키워드: 문헌정보학, 연구동향, 토픽모델링, 매트릭스 재정렬, 합성, LDA, BERT, BERTopic, WORD2VEC, AET
library and information science, research trends, topic modeling, matrix reordering, synthesis, LDA, BERT, BERTopic, WORD2VEC, AET

---

* 이 논문은 경북대학교 문헌정보학과 박사학위논문을 축약한 것임.
** 경북대학교 사회과학대학 문헌정보학과 강사(seonwook.kim@knu.ac.kr) (제1저자)
*** 영남고문헌아카이브센터 이사(yangkiduk@gmail.com) (교신저자)
■ 논문접수일자 : 2022년 8월 13일 ■ 최초심사일자 : 2022년 9월 1일 ■ 게재확정일자 : 2022년 9월 16일
■ 정보관리학회지, 39(3), 99-132, 2022. http://dx.doi.org/10.3743/KOSIM.2022.39.3.099

1. 서 론

현대사회의 통신망과 정보처리기기는 기록과 보존을 위한 인류의 오랜 도구인 종이와 펜을 빠르게 대체하고 있으며, 이러한 변화는 대중을 위한 SNS부터 연구자의 학술연구 커뮤니티까지 큰 영향을 미치고 있다. 오늘날 연구자는 통신망이 연결된 환경에서 전자화된 데이터를 수집하고 전자 저널에 논문을 투고하는 방식을 취하고 있다. 대학이나 기관은 물론, 출판사와 학회도 초기에는 단순히 논문의 목록만 제공하였으나 최근에는 오픈 액세스 정책에 부합하기 위한 리포지터리 사용을 장려하고 있으므로 전자정보원의 규모는 계속 커질 전망이다.

그러나 전자정보원의 규모가 커질수록, 그 정보의 내용을 파악하는 일은 점차 어려워지고 있다. 저자가 임의로 부여한 태그나 키워드가 주제를 온전히 반영한다고 보기는 어려우며, 그렇다고 수작업으로 분류하는 것은 천문학적인 시간과 비용이 소요될 뿐만 아니라 평가자의 주관적 의견이 반영되어 결과가 오염될 수 있다는 단점이 있다. 따라서 전자정보원의 규모가 커질수록 객관적 기준에 따른 기계적 주제분석의 필요성이 함께 커진다고 볼 수 있다(임소라, 권용진, 2017).

토픽모델링은 문서 내 숨겨진 의미(주제)를 찾는 방법을 통칭한다. 그중에서 Blei, Ng, Jordan (2003)이 제안한 잠재디리클레할당(Latent Dirichlet Allocation, 이하 LDA)은 정형 데이터뿐만 아니라 비정형 데이터에서도 그 활용도가 높아, 현재 토픽모델을 활용한 연구에서 가장 일반적으로 활용되고 있다(Jelodar et al., 2019; Vayansky & Kumar, 2020). LDA는 단어의 등장빈도가 중요하며 순서와 의미는 중요하지 않은 확률적 모델이므로, 입력데이터를 반드시 전처리하여 BoW(Bag of Words) 형태로 준비해야 한다. 이외에도 연구자가 사전에 설정해둔 하이퍼파라미터와 토픽의 개수가 통계적으로 전체 문서를 포용할 수 있는지에 따라 성능에 큰 영향을 미치므로, 하이퍼파라미터 설정이 매우 중요한데 최적의 값을 찾기가 쉽지 않다. 그럼에도 불구하고 오픈소스를 통해 구현하기 편리하고 정량적 평가법(coherence score[1]와 perplexity score[2])을 적용할 수 있으므로 많은 연구자가 LDA를 선택하고 있다.

최근에는 자연어 처리모델인 BERT를 기반으로 한 BERTopic이라는 토픽모델링 기법이 등장하여, 세상의 현상과 추세를 파악하는 연구를 진일보시키는 데 일조하고 있다. BERTopic은 입력데이터를 자연어 형태인 '온전한 문장'으로 가정하고, 문맥(context)을 고려해서 학습된 BERT 사전학습 모델을 참조해 임베딩하므로 문맥 정보를 반영할 수 있으며, 하이퍼파라미터를 설정할 필요 없고, LDA와 비교해 더 많은 토픽을 찾아냄으로써 토픽 상세도(topic granularity)를 높이는 특성이 있다. BERT 특성상 BERTopic의 수치적 성능평가법은 아직 존재하지 않으나, 대상 언어에 높은 성능을 보이는 사전학습 모델일수록 BERTopic의 결과도 더 낫다고 알려져 있다(Grootendorst, 2022).

이처럼 LDA와 BERTopic을 비교하는 행위는 사과와 오렌지를 비교하는 것(comparing apples

1) 토픽 내 상위 단어 간 의미적 유사성 점수
2) 정보 엔트로피의 값을 사용한 혼잡도 점수

and oranges) 같이 고유 특성이 전혀 다르므로 그 성능의 결과 자체를 서로 비교하는 것은 무리가 있다. 그러나 이러한 서로 다른 토픽모델링 결과를 합성하는 것이 가능하다면 각각의 장점을 보완하고 단점을 상쇄시키는 효과를 기대할 수 있을 것이다. 이에 본 연구는, LDA 토픽모델링 결과와 BERTopic 토픽모델링 결과를 합성하는 방법론인 AET(Augmented and Extended Topics, 이하 AET)를 제안하고자 한다. AET는 LDA가 발견한 토픽을 중심으로 세분된 BERTopic 토픽을 군집하여 LDA 토픽의 의미를 증강(augment)하는 한편, LDA 토픽과 의미론적 관계에서 정반대에 있지만 BERTopic이 발견해낸 비주류 주제를 발굴함으로써 토픽모델링을 확장(extend)하는 데 목적을 둔 실험적 연구이다. 그리고 AET를 사용해 국외 문헌정보학분야 학술지 논문의 연구주제와 최근의 추이를 파악함으로써, 국내 연구자가 다양한 연구 주제를 발견할 수 있는 단초를 제시하고자 한다.

2. 이론적 배경

2.1 워드 임베딩

워드 임베딩은 단어를 일정 규칙에 따라 벡터로 변화하여 공간에 투사하는 방법으로, 특히 워드 임베딩을 사용하면 어휘집합 내 각 단어에 특성 벡터(feature vector)를 부여함으로써 단어 간 관계와 유사도를 알 수 있다(Bengio, Ducharme, & Vincent, 2000). 그러나 문서에 등장하는 단어의 수만큼 차원이 증가하는 희소 벡터 표현(sparse vector representation) 문제가 발생한다.

Mikolov et al.(2013)의 WORD2VEC은 목표 단어와 그 주변의 단어가 무엇인지 softmax regression을 통해 예측하는 과정에서 학습이 이뤄진다. 단점은 문장 내 지역적인 동시출현 정보에 기반하여 단어를 학습하므로 어떤 단어가 여러 위치에서 다른 의미로 사용되었더라도 그 차이를 포착하기 힘들다는 것이다. 이를 보완하기 위해서 전역 동시출현 정보를 사용하는 Glove와 각 단어를 문자 단위 n-gram으로 분해하여 학습하는 FastText가 있다(이다빈, 최성필, 2019).

Peters et al.(2018)의 ELMO(Embedding from Language MOdel)가 등장하고 나서부터는, 맥락에 따른 표현을 반영할 수 있는 BERT 혹은 GPT와 같은 사전학습 모델 기반 임베딩 방법론이 자연어 처리의 주된 방법론으로 사용되고 있다(Ajayi, 2020).

본 연구에서의 토픽용어는 문장 구조를 갖추지 못해 문맥이 없는 BoW 형태의 단어 집합이므로, 토픽간 코사인 유사도 측정수단으로는 문맥 위주의 BERT보다 WORD2VEC이 더 유리할 것이라 가정하였다.

2.2 LDA

토픽모델링의 기원은 잠재의미분석(Latent Semantic Analysis, 이하 LSA)으로, 문서-단어 행렬에서 단어의 출현빈도만으로는 나타낼 수 없는 잠재적 의미(주제)를 도출하고자 하였다(Deerwester et al., 1990). 이후 Hoffman (1999)은 주제마다 숨겨져 있는 의미를 파악

하고, 이를 확률적인 분포로 나타내 각각의 토픽을 생성하는 확률잠재의미할당(Probabilistic Latent Semantic Allocation, 이하 pLSA)를 제안하였다. 그러나 pLSA는 변수가 대상 말뭉치 크기에 따라서 증가할수록 과적합(overfitting)이 발생한다.

Blei, Ng, Jordan(2003)이 제안한 잠재디리클레할당(Latent Dirichlet Allocation, 이하 LDA)은 pLSA와 구현방식은 유사하나, 하이퍼파라미터를 사용하여 pPLSA의 한계였던 과적합을 방지하고 토픽모델링의 결과를 유연히 도출한다.

2.3 BERT

2018년 Google이 공개한 BERT는 입력데이터를 양방향에서 접근하여 문맥을 파악할 수 있고, 미세조정과 사전학습이 가능한 비지도 학습모델이다. BERT의 사전학습은 단어와 문장의 맥락을 파악할 수 있도록 학습 과정에서 입력 텍스트의 15%를 임의로 가리고 학습하는 Masked Language Model과 2개의 문장을 짝지어 참과 거짓으로 절반 나누어 다음 문장 예측을 학습하는 Next Sentence Prediction Model을 사용하므로, 단어가 문장 내 등장 위치에 따라 맥락정보를 학습하고 문장의 연관성을 학습할 수 있다. 이런 특성 때문에 BERT는 여러 연구에서 최고 성능을 증명하고 있는데(배장성 외, 2020), 임베딩 성능을 한층 더 높여 문맥 표현 능력을 증가시키려는 노력도 진행되고 있다(Schick & Schutze, 2019).

BERT의 또 다른 장점은 전이학습(transfer learning)을 지원한다는 데 있다. 미리 특정 분야에서 학습된 데이터를 전혀 새로운 분야에서 재사용할 수 있는 전이학습을 구현하기 위해, BERT는 사전학습 모델을 미세조정(fine-tuning)하는 과정에서 추가 학습을 시행하고 원하는 다운스트림 태스크대로 업데이트하는 방안을 제공한다. 미세조정할 때는 주로 Huggingface[3]가 제공하는 트랜스포머(transformers)가 사용되는데, 특히 BERT 사전학습데이터를 기반해 텍스트를 워드 임베딩으로 전환하는 경우 Sentence Transformers가 사용된다. 우리나라도 한국전자통신연구원(ETRI)에서 한국어 언어모델인 KorBERT(코버트)[4]를 공개한 바 있다.

2.4 BERTopic

BERTopic은 BERT를 통해 사전학습된 언어모델데이터를 호출하여 자연어 형태의 입력데이터를 Sentence Transformers로 워드 임베딩한 뒤, UMAP[5]으로 차원축소하고 HDBSCAN을 통해 토픽을 군집화한 후, 단어의 빈도와 역문서 빈도(TF-IDF)를 클래스 기반으로 적용하는 c-TF-IDF를 사용해 각 토픽용어의 순위를 결정한다(Grootendorst, 2022).

BERTopic의 장점으로는 문맥을 고려해서 학습된 BERT 사전학습 모델을 참조할 수 있으므로 BoW 형태의 입력데이터를 처리하는 LDA와 달리 문맥 정보를 반영할 수 있다. 다만 임베

3) 머신러닝 모델에 사용되는 오픈소스를 제작하고 배포하며, 사용자간 공유할 수 있는 허브를 제공하는 스타트업
4) https://aiopen.etri.re.kr/service_dataset.php
5) 고차원 특성으로 표현된 데이터를 저차원으로 표현하는데 사용되는 차원축소 알고리즘

딩 특성상 기존의 평가법으로 BERTopic의 성능을 평가하는 것이 불가능하다는 단점이 있다(Bodrunova et al., 2020).

2.5 매트릭스 재정렬

계량서지학의 네트워크 분석법에서 주로 사용되는 1-모드 매트릭스의 경우, 동일한 데이터가 2개의 축에서 접근하므로 같은 값이 열과 행에서 만나는 대각선에 큰 값이 집중하게 된다. 이는 코사인 유사도 매트릭스에서도 마찬가지인데, 단어 집합에 대해 코사인 유사도 계산을 시행한 결과에서 높은 유사도 값이 대각선을 이룬다.

그러나 서로 다른 2개의 데이터를 비교하는 경우, 그 상관관계를 예측할 수 없으므로 넓은 공간에 의미를 알 수 없는 값이 무작위 분포된 희소 행렬(sparse matrix)처럼 보이게 되며 매트릭스의 크기가 커지면 그 해석을 전문가의 눈에 맞길 수밖에 없다(Behrisch et al., 2016). 따라서 연구자가 필요한 형태로 매트릭스를 재정렬하는 작업이 필요하다.

그중에서 대각선 모델은 매트릭스 내 최댓값을 어떻게 정렬할 수 있는지 나타내는 대표적인 사례이다. Ermann, Chepelianskii, Shepelyansky (2012)는 Google matrix의 특성을 연구하여 PageRank와 CheiRank로 만들어진 2차원 검색엔진을 구현하는 연구를 진행하였는데, 이 과정에서 2차원 밀도 매트릭스의 한 축에 해당하는 Filtered CheiRank의 사잇값을 작게 할수록 대각선이 형성되었다. Behrisch et al.(2016)은 다수의 재정렬 방법론을 검토한 결과, 매트릭스 위에 그은 대각선을 기준으로 단일 대각선, 다중 대각선, 쌍곡선 형태에 따라 군집화하는 방법이 있다고 정리하였다. 벡터 공간 안에 표현된 데이터의 차원을 축소하는 방법인 선형 PCA(Principal Component Analysis)도 특성을 추출하기 위해 데이터들의 분산이 가장 커지는 축을 발견한 뒤, 그 축에 수직인 축을 다시 반복적으로 찾아 나가는 대각화 과정(diagonalization)을 거친다(송은영, 최회련, 이홍철, 2019).

3. 선행연구

3.1 LDA를 활용한 토픽모델링 개선 연구

LDA는 대표적인 토픽모델링 연구방법론으로 디리클레 파라미터가 실제 문서 데이터베이스 내 분포 상황을 정확히 반영하지 못하는 지적도 있으나, 많은 연구자에 의해 수정되고 응용 영역 또한 확장되어 왔다(Gerlach, Peixoto, & Altmann, 2018).

박준형, 오효정(2017)은 국내 기록관리학 연구동향을 LDA와 HDP로 각각 토픽모델링한 결과, LDA의 성능은 토픽용어 빈도에 크게 영향을 받는다고 판단하였다. 최원준 외(2018)는 700여 개의 논문 서지정보를 대상으로 대해 3가지 서로 다른 토픽모델링 기법(LDA, LSI, HDP) 성능을 비교하였는데, LSI와 HDP는 토픽 개수가 증가할수록 coherence score가 크게 떨어지거나 등락 폭이 컸지만 LDA는 꾸준히 높은 수치를 보인다는 사실을 확인하였다.

또한 LDA의 입력으로 사용되는 BoW 형태의 데이터가 의미나 문맥을 반영하지 않으므로 토픽과 토픽용어 사이의 관계를 이해할 수 없

다는 단점을 보완하기 위해, 키워드나 토픽에 대해 1-모드 네트워크 목록을 만들거나 문서와 토픽 사이의 2-모드 네트워크 목록을 만들어 시각화함으로써 토픽 사이의 관계를 분석하려는 시도도 있었다(강보라, 김희섭, 2017; Chen, Sheble, & Eichler, 2013).

황승연 외(2020)는 빅데이터를 분석하고 트렌드를 예측하기 위해서 빅데이터로부터 LDA 토픽모델링 결과를 자동 생성하는 파이프라인 기법을 제시하고자 하였으나, 이 과정에서 LDA 하이퍼파라미터(k, α, β)를 일괄 지정(Hard-coded)할 수밖에 없는 한계점이 있었다.

3.2 임베딩을 활용한 토픽모델링 개선 연구

Moody(2016)는 LDA2VEC을 제안하였는데, 보통의 워드 임베딩 방식으로는 주어진 단어로부터 단어 거리가 가장 가까운 단어를 예측하지만, LDA2VEC 모델은 문서 벡터에 의해서 지정된 토픽 매트릭스와 관계된 단어를 예측하게 된다. 이후 여러 연구자에 의해 토픽모델링의 수치적 분포와 벡터 공간의 의미적 분포를 서로 병렬처리한 뒤 융합하려는 노력이 계속되었다. Li et al.(2018)이 제안한 PW-LDA(Partitioned Word2Vec-LDA)는 입력 문서를 먼저 쪼개어 분석 범위를 좁힌 뒤 LDA와 WORD2VEC을 각각 수행하고 두 결과로부터 도출된 결과에 코사인 유사도를 측정한 뒤 군집화하는 방안이다. Gao et al.(2022)은 입력데이터를 WORD2VEC로 워드 임베딩 처리한 뒤 t-SNE으로 차원을 축소하고 affinity propagation으로 의미적 군집을 생성한 뒤, 같은 데이터를 동적 토픽모델링 처리한 결과와 매치시켜 토픽의 의미적 분포가 변하는 추이를 보고자 하였다.

토픽모델링과 임베딩을 순차적으로 직렬 연결하는 방법에 관한 연구도 있었는데, Choi와 Kim(2019)은 WORD2VEC을 통해 생성된 워드 임베딩 벡터를 K-평균 군집화 알고리즘으로 1차 군집화한 뒤, 각 클러스터에 2차 토픽모델링을 적용함으로써 대규모 자연어 데이터셋을 세밀하게 토픽 클러스터링하는 방법을 제안하였다. 윤상훈, 김근형(2021)은 LDA 토픽모델링을 수행한 뒤 각 토픽 내 토픽용어 중 unique token 비율이 낮을 때 WORD2VEC을 이용하여 유사한 의미가 있는 단어를 추가 추출하여 해당 토픽에 부여하는 방법을 제시하였다.

최근에는 워드 임베딩을 사용한 새로운 토픽모델링 기법도 제안되고 있다. Angelov(2020)는 의미 공간(semantic space)이란 의미적 연관성을 벡터 공간에서 거리로 표현한 것이며 이 공간이 연속적인 토픽 표현으로 가득 차 있다고 전제하고, 인근 단어를 군집화하여 토픽으로 표현하는 TOP2VEC을 제안하였다. TOP2VEC은 DOC2VEC을 사용한 문서 임베딩과 WORD2VEC을 사용해 만든 워드 임베딩을 생성하고 특정 문서 임베딩 벡터에 가까이 있는 워드 임베딩이 그 문서 집합을 가장 잘 표현하는 토픽용어라고 가정한다. 따라서 불용어를 제거하거나 표제어와 어간을 추출할 필요가 없으며, 심지어 토픽 개수를 미리 알 필요가 없다는 장점이 있다.

임베딩기법은 그동안 토픽모델링의 성능을 향상하기 위해 보조하는 역할에 그쳤지만, 최근에는 워드 임베딩을 사용한 토픽모델링 기법의 성능이 LDA보다 우수하다는 연구가 등장하기 시작했다.

Esposito, Corazza, Cutugno(2016)는 LDA를 사용한 토픽모델링 결과와 WORD2VEC CBOW 알고리즘으로 도출한 임베딩을 K-평균 군집화한 토픽모델링 결과를 비교한 뒤, 입력데이터가 전처리되어 있다는 전제하에 LDA보다 임베딩 기법으로 토픽모델링한 경우가 성능이 더 낫다는 연구결과를 보고하였다. Sia, Dalmia, Mielke(2020)도 TOP2VEC과 유사한 연구를 시행하였는데, 사전학습 모델을 이용한 워드 임베딩을 TF 가중치를 반영한 K-menas 군집화와 TF 기반 재순위 처리와 함께 조합했을 때 토픽모델링과 같은 결과를 얻을 수 있음을 확인하였다.

2021년 공개된 BERTopic은 TOP2VEC과 같은 선임베딩 후군집화 기법이지만 이미 머신러닝이 완료된 사전학습학습모델을 응용하여 임베딩할 수 있다는 장점이 있다(Grootendorst, 2022). 박순욱, 김영국, 김명호(2021)는 다양한 분류방법의 성능을 평가하는 연구에서 BERTopic과 KoBERT를 비교한 결과, 한국어 특화모델인 KoBERT가 다국어언어모델을 기본으로 사용하는 BERTopic 보다 성능이 좋음을 확인하였다.

4. 연구방법

4.1 연구설계

본 연구는 〈그림 1〉과 같이 크게 5개 절차로 구성된다.

첫 번째, 데이터 수집 과정에서는 분석에 사용할 데이터인 논문 초록을 수집하고 선정하였다.

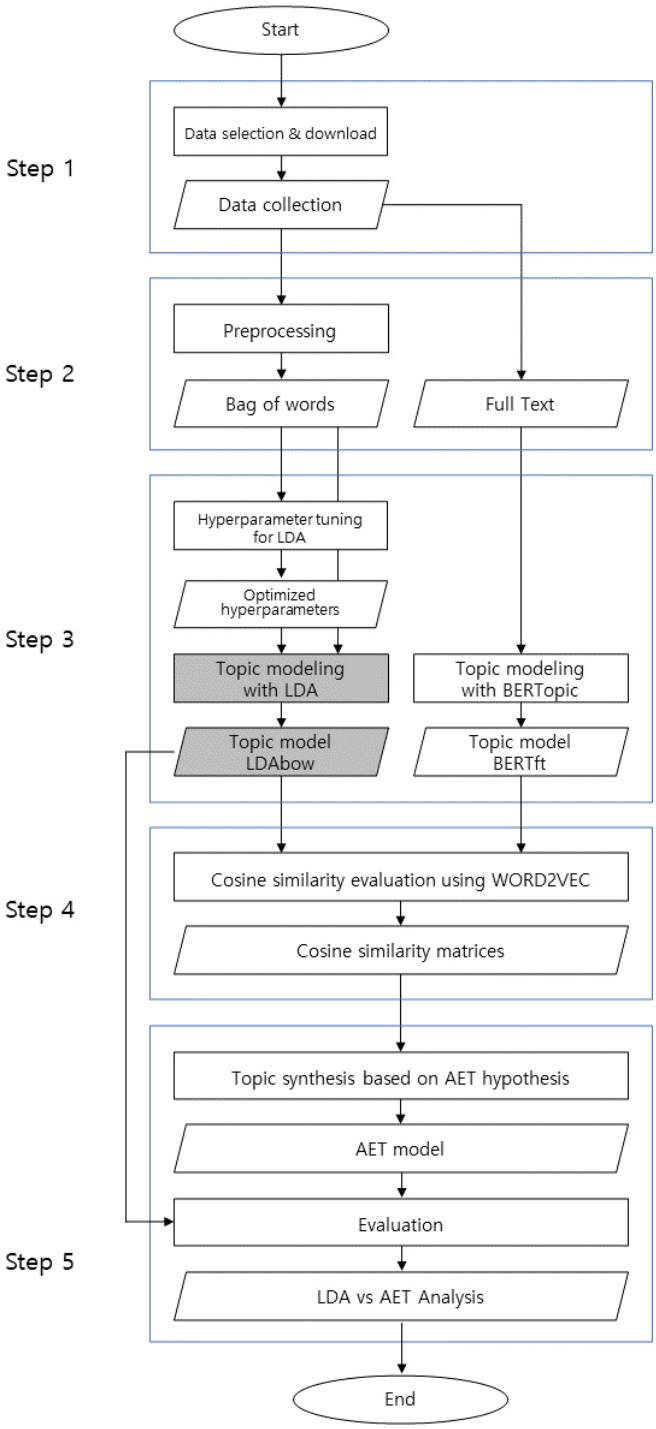
두 번째, LDA에 입력할 수 있도록 데이터로부터 노이즈 제거[6]한 뒤 전처리(명사 추출, 표제어 추출, 불용어 삭제, 연어 생성)하여 LDA에 입력할 수 있도록 BOW 형태로 변환한다. 계속해서 BERTopic에 입력할 수 있도록 데이터로부터 노이즈만을 제거하여 full text(한 문장이 주어, 동사, 목적어 등 문장을 이루는 데 필요한 단어를 모두 갖추고 있는 상태, 이하 FT) 형태의 데이터도 확보한다.

세 번째, LDA 하이퍼파라미터 튜닝 및 토픽모델링 과정은 2개의 하위절차로 구성된다. (1) LDA 하이퍼파라미터 최적화. LDA에 입력할 BoW 데이터에 대해 여러 범위의 하이퍼파라미터를 동적으로 변화시켜 LDA 토픽모델링을 실행했을 때의 coherence score 및 NAC(Normalized Absolute Coherence) 값을 확인하여 최적의 하이퍼파라미터를 도출한다. coherence score는 토픽모델링에서 자주 사용되는 c_v[7]를 사용하였다(M'sik & Casablanca, 2020). (2) 토픽모델링. 먼저 최적의 하이퍼파라미터를 사용하여 BoW 형태의 데이터를 입력한 LDA 토픽모델링 결과(이하, LDAbow)를 확보한다. 이어서 FT 형태의 데이터를 입력한 BERTopic 토픽모델링 결과(이하, BERTft)를 확보한다.

네 번째, LDAbow와 BERTft를 WORD2VEC으로 변환하여 둘 사이의 코사인 유사도 매트

6) 괄호 '( ), [ ], { }' 앞뒤에 공백문자를 추가, 연속된 빈칸을 하나의 빈칸으로 대체, 하이픈을 공백문자로 대체, Elsevier社가 초록 내 임의로 추가한 저작권 문구 삭제

7) coherence score를 계산하는 방접 중 하나로, NPMI(nomalised pointwise mutual information)와 코사인 유사도를 사용하여 계산된 점수


Start
Step 1
Data selection & download
Data collection
Step 2
Preprocessing
Bag of words
Full Text
Step 3
Hyperparameter tuning for LDA
Optimized hyperparameters
Topic modeling with LDA
Topic modeling with BERTopic
Topic model LDAbow
Topic model BERTft
Step 4
Cosine similarity evaluation using WORD2VEC
Cosine similarity matrices
Step 5
Topic synthesis based on AET hypothesis
AET model
Evaluation
LDA vs AET Analysis
End

〈그림 1〉 연구 절차

릭스를 구성한 뒤 본 논문에서 제안하는 방법으로 매트릭스를 재정렬한다.

마지막으로, 본 논문에서 제안하는 합성 방법인 AET를 사용하여 코사인 유사도 매트릭스로부터 증강토픽(Augmented Topic, 이하 AT)과 확장토픽(Extended Topic, 이하 ET)을 도출한다.

4.2 데이터 수집·선정 및 전처리

본 연구를 위해 먼저 2001년 1월부터 2021년 10월까지(약 20.8년), Web of Science가 제공하는 문헌정보학 학술지(Library and Information Science - SSCI) 86종에 게재된 학술논문 서지데이터를 수집하였다. 이렇게 수집된 188,674건의 서지정보 중에서 Library Journal -학술지에 수록된 학술논문이 97,726건으로, 이는 전체 86종 학술지 중 나머지 85종의 학술지에 게재된 논문 편수를 모두 합한 90,948건보다 많은 것이다. 따라서 본 연구에서는 Library Journal을 이상치(outlier)으로 판단하고 분석 대상에서 제외하였다(양기덕, 김선욱, 이혜경, 2021; Yang et al., 2021). 계속해서 Language 칼럼의 값이 English가 아닌 레코드를 배제하고, Publication Year 칼럼이 비어있는(null) 레코드를 제외하고, Abstract 칼럼에 포함된 문자열의 단어 수가 25개 미만인 레코드를 제외[8]하였다. 그 결과 분석에 사용될 서지 레코드의 수는 90,948개에서 55,442개로 줄어들었다.

전처리에는, 자연어 처리에 특화된 Python 기반 오픈 소스 라이브러리인 spaCy를 사용하였다. 우선 몇 차례 시험적 시도에서 확인된 문제를 제거하기 위해 아래와 같이 데이터를 일부 수정하였다. 첫째, spaCy의 버그[9]를 회피하기 위해 괄호 특수문자 '( ), [ ], { }' 앞뒤에 공백문자를 추가하였다. 둘째, 연속된 빈칸을 하나의 빈칸으로 대체하였다. 셋째, 하이픈을 공백문자로 대체하였다. 넷째, 예비 전처리에서 spaCy가 '%'를 명사로 인식되는 경우가 확인되어 이 특수문자를 수동으로 미리 제거하였다. 다섯째, Elsevier社가 초록 내 임의로 추가한 두 가지 패턴의 저작권 문구(예. (C) 2020 Elsevier Ltd. All rights reserved 및 Published by Elsevier Ltd.)가 확인되어, 이를 찾아 삭제하였다.

그다음 spaCy 학습모델 중 가장 정확도가 높은 영문 언어모델인 en_core_web_lg를 사용하여, 인적 개입을 최대한 배제한 자동화 전처리를 시행하였다. 문헌 내 모든 단어를 출현빈도 순으로 정렬하면 그 출현빈도가 해당 단위의 순위에 반비례한다는 Zipf의 법칙에 근거한, Luhn의 모델을 응용하여(Losee, 2001) 상위 1~6위에 해당하는 과빈도 단어 6개와 하위 1~6위(공동순위 허용)에 해당하는 저빈도 단어 13,470개를 spaCy의 기본 불용어 목록에 추가하였다. 그 결과, 고유토큰(unique token)의 개수를 본래의 64,976개에서 8,070개까지 줄일 수 있었다.

---

8) Elsevier社가 제공하는 논문 작성규칙에 따르면 초록의 최대 길이는 제한하고 있으나 최소 길이에 대한 언급은 찾아볼 수 없다. 따라서 본 연구에서는 하나의 초록이 최소한 하나의 문장을 포함하고 있다고 가정하고, 단어 개수가 25개 미만인 초록은 분석에서 제외하였다(Deveci, 2019).

9) https://github.com/explosion/spaCy/issues/3454

4.3 제안된 증강/확장토픽(AET)의 구성

본 논문에서 제안하는 AET 모델은 〈그림 1〉의 연구 절차 내 Step 3부터 Step 5에 해당하는 부분으로, 다음과 같이 요약할 수 있다. 첫째, 최적의 LDA 토픽모델을 생성할 수 있는 하이퍼파라미터를 도출한 뒤 이를 사용하여 LDA 기반 토픽모델링 결과인 LDAbow를 생성한다. 이와 동시에 full text 형태의 데이터를 BERTopic에 입력하여 워드 임베딩 기반 토픽모델링 결과에 해당하는 BERTft를 생성한다. 둘째, 베이스라인인 LDAbow를 기준으로, 'LDAbow vs BERTft'에 대해 WORD2VEC을 사용한 코사인 유사도 매트릭스를 구축한다. 셋째, 매트릭스 재정렬을 통해 증강토픽 후보군(이하 AT 후보군)과 확장토픽 후보군(이하 ET 후보군)을 결정하고 이를 합성하여 AET 결과를 도출한다.

4.3.1 LDA 토픽모델링

먼저 최적의 LDA 토픽모델링 연산에 필요한 하이퍼파라미터를 결정하기 위해, 개별 하이퍼파라미터를 독립변수로 하고 coherence score와 perplexity score를 종속변수로 하는 연산 시뮬레이션을 실행하였다. 그동안 연구자 사이에서 coherence score가 높을수록 perplexity score가 낮을수록 그 결과가 정확하다는 것이 일반적인 접근법이었다(박종도, 2019; 이유빈 외, 2020). 그러나 최근 연구에 따르면 실제로 perplexity score는 신뢰도가 매우 낮아 분석에 사용하기 어려우며, coherence score도 그대로 사용하는 것보다는 정규화(NAC)해서 보는 편이 신뢰도가 더 높다(Hasan et al., 2021). 본 연구에서도, 실험이 진행될수록 perplexity score는 분산되는 경향이 확인되어, coherence score와 NAC 모두 가장 큰 값을 기록한 하이퍼파라미터 조합만을 LDA 토픽모델링에 투영하였다.

전처리가 완료되어 BoW 형태로 전환된 입력데이터에 Gensim[10]이 제공하는corpora.dictionary API를 사용하면 num_topics에 설정한 만큼 토픽용어 집합을 추출할 수 있다. 본 연구에서는 연산속도를 가속화하기 위해 병렬연산을 지원하는 ldamulticore 클래스를 사용하였다.

4.3.2 BERTopic 토픽모델링

BERTopic은 Sentence Transformers를 이용하여 입력 데이터에 대한 워드 임베딩을 생성하며, 다양한 언어 모델을 반영할 수 있다. 본 연구에서는 BERTopic의 임베딩 매핑을 위해 현재까지 가장 성능이 좋다[11]고 알려진 all-mpnet-base-v2 사전학습 모델을 채택하였다. BERTopic과 같은 임베딩 기반 '선 분류 후 군집' 토픽모델링 방법은 LDA처럼 클러스터 개수(k)를 미리 지정하지 않더라도 차원축소 과정에서 자동으로 토픽 개수가 결정된다(Wang et al., 2018).

4.3.3 AET(증강/확장토픽) 방법

1) 매트릭스 재정렬

LDAbow와 BERTft의 코사인 유사도 측정 결과를 측정할 때, LDAbow의 전체 토픽용어 집합(m)을 BERTft 전체 토픽용어 집합(n)에 대해 코사인 유사도 측정한 결과를 $m \times n$ 크기의 매트릭스(m 행: LDAbow 토픽번호, n 열:

10) 확률기반 비지도학습을 위한 Python 기반 오픈소스 라이브러리로, LDA 구현을 위한 모듈도 포함하고 있다.
11) https://www.sbert.net/docs/pretrained_models.html

BERTft 토픽번호)에 저장한다.

이때 비교되는 두 토픽은 온전한 문장 형태가 아니므로, 코사인 유사도 측정에는 GoogleNews-vectors-negative300 사전학습 모델에 기반한 WORD2VEC을 사용하였다.

작성된 매트릭스 안에서 행과 열 사이에 숨겨진 의미를 발견하기 위해서는 재정렬(reordering) 작업을 거칠 필요가 있다(Behrisch1 et al., 2016). 본 연구에서는 〈그림 2〉와 같이 최댓값을 순차적으로 대각 정렬하는 방법을 제안한다.

	x1	x2	x3
y1	0.2	1.2	0.9
y2	0.1	0.6	0.5
y3	0.4	0	0.7

a. 원본 데이터

	x2	x1	x3
y1	**1.2**	0.2	0.9
y2	0.6	0.1	0.5
y3	0	0.4	**0.7**

b. 1차 정렬

	x2	x3	x1
y1	**1.2**	0.9	0.2
y3	0	**0.7**	0.4
y2	0.6	0.5	0.1

c. 2차 정렬

	x2	x3	x1
y1	**1.2**	0.9	0.2
y3	0	**0.7**	0.4
y2	0.6	0.5	**0.1**

d. 3차 정렬 후 정렬 완료

〈그림 2〉 제안된 매트릭스 재정렬 방법

① 〈그림 2a〉의 회색 영역 안에서 1차 시도 ($j=1$). 최댓값인 1.2가 위치한 셀이 $(x2, y1)$이다. 최댓값을 j 기준 위치인 (1, 1)로 이동시키기 위해서 $x2$ 열을 $x1$ 열의 자리를 서로 바꾼다. 그 결과 1차 시도 정렬이 〈그림 2b〉처럼 완료되었다.

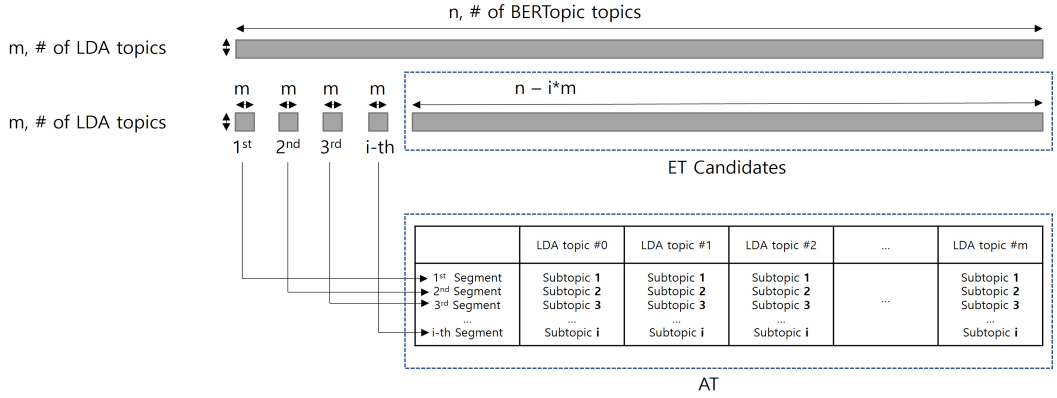
② 〈그림 2b〉의 회색 영역 안에서 2차 시도 ($j=2$). 최댓값인 0.7이 위치한 셀이 $(x3, y3)$이다. 최댓값을 j 기준 위치인 좌표(2, 2)로 이동하기 위해 $y3$ 행을 $y2$ 행과 서로 바꾸고, $x3$ 열을 ,$x1$ 열과 서로 바꾼다. 그 결과 2차 정렬이 〈그림 2c〉처럼 완료되었다.

③ 〈그림 2c〉의 회색 영역 안에서 3차 시도 ($j=3$). 남은 회색 영역은 $(x1, y2)$가 유일하므로 3차 시도 시 최댓값은 0.1이다. 이렇게 완료된 재정렬 결과는 〈그림 2d〉와 같다.

2) 토픽 증강(AT)

LDAbow와 BERTft를 확보한 뒤 코사인 유사도 매트릭스에 〈그림 2〉의 재정렬을 적용하면, y축은 LDAbow에 x축은 BERTft에 대응한다. BERTopic이 LDA보다 많은 토픽을 발견하므로 실제 토픽모델링을 수행한 뒤 코사인 유사도 매트릭스를 만들면 〈그림 3〉처럼 y보다 x가 긴 매트릭스가 만들어진다. 코사인 유사도 매트릭스 재정렬은 한 번에 짧은 축의 길이만큼만 수행되므로 LDAbow의 토픽 개수를 m, BERTft의 토픽 개수를 n, 재정렬 수행횟수를 i라고 할 때 $m\times(n-im)$ 크기의 매트릭스가 남을 때까지 $m\times n$ 크기의 재정렬 세그먼트를 i개 생성할 수 있다.

각 세그먼트는 LDA 결과에 대한 코사인 유사도 최댓값을 기준으로 정렬된 상태이므로, '의미적 연결관계'가 무너지지 않는 i 범위에서 생성된 각 재정렬 세그먼트를 구성하는 BERTft 토픽은 모두 LDAbow의 서브토픽이라 정의할 수 있다. 이때, 메인토픽과 서브토픽 관계에 있



〈그림 3〉 AET 합성 개념

는 LDAbow 토픽과 BERTft 토픽의 집합은 기존의 LDA 토픽을 지지하고 세분화하는 역할을 한다. 이 관계를 이루는 집합을 AT(증강토픽)로 정의한다. i 영역의 임계치는 연구자의 주관적 판단으로 결정되므로 재정렬을 반복 진행할 때, '의미적 연결관계'가 손실되는 임계치를 포함하는 i를 놓치지 않도록 주의를 기울여야 한다. 코사인 유사도에서 유무를 가리키는 일반적인 임계치가 0.5이므로, 임계치는 0.5 인근에서 발견될 가능성이 크다고 예상할 수 있다.

3) 토픽 확장(ET)

전체 코사인 유사도 매트릭스에서 AT 영역으로 선정하고 남은 $m\times(n-im)$ 크기의 세그먼트를 ET 후보군으로 정의한다. 그러나 ET 후보군 안에 여전히 임계치 이상의 수치를 최댓값으로 포함하는 열이 존재할 수 있으므로(잠재적 LDA 서브토픽), 코사인 유사도 수치가 하나라도 임계치 이상인 셀을 포함하는 BERTft 토픽은 ET 후보군에서 배제해야 한다.

그 다음 ET 후보군이 갖는 특성을 파악한 뒤, 연구자가 설정한 기준에 따라 정렬한 뒤 top-n을 ET로 선정하는 것이 타당하다 할 수 있다. ET 후보군은 LDA 토픽에 대한 BERTtopic 토픽의 코사인 유사도 매트릭스의 일부분이므로, ET 후보군에 랭킹을 부여하는 기준으로 topic size(토픽 크기 순위, 이하 TS)와 similarity average(코사인 유사도 평균값 순위, 이하 SA)를 고려할 수 있다. TS가 클수록 BERTopic이 주류(major)로 판단한 토픽을 의미한다. SA가 낮으면 LDA 토픽으로부터의 벡터 공간 내 거리가 멀어짐을 의미하므로, 이 값이 낮을수록 LDA가 발견하지 못한 토픽을 의미한다. 둘 다 ET 측면에서 중요한 의미가 있는 특성이므로 둘 다 고려하기 위해, 수식 (1)과 같이 F_1 score(조화평균)의 확장인 F_β score를 사용한다.

$$F_\beta = (1+\beta^2)\cdot\frac{TS\cdot SA}{(\beta^2\cdot TS)+SA} \qquad \text{수식 (1)}$$

TS와 SA 중에서 어느 쪽에 더 가중치를 둬야 하는지 또는 동등하게 F_β score를 계산해야 하는지 판단하기 위해서는, 여러 β값에 대한 F_β(TS, SA)를 기준으로 ET 후보군을 정렬한 뒤

TS와 SA를 모두 포용하는 β를 실제로 확인해야 한다.

ET 후보군이 정렬되면, 연구자는 필요에 따라 순위에 따라 ET의 수를 유연하게 설정할 수 있다. 본 연구에서 제안하는 AET는 ET 개수를 AT 개수(LDAbow의 토픽 개수)와 같게 설정한다.

4.3.4 AET 평가

앞서 LDA 하이퍼파라미터 튜닝 과정에서 coherence score와 NAC와 같이 수학적 접근법으로 토픽모델의 특성을 평가한 것이 정량적 평가라면, 실제 생성된 토픽모델의 토픽 목록과 토픽용어 목록을 인간이 이해할 수 있는 수준인지 평가하는 것은 정성적 평가이다. 본 연구에서는 AET 정성적 평가를 위해 다음 2가지 방법을 채택하였다.

먼저 의학 분야에서 다수의 진단 방법이나 측정방법의 일치도를 판단할 때 사용하는 카파통계량(Cohen's Kappa Statistic)을 사용하였다(김동기, 한무영, 한혜리, 2004; 박창언, 김현정, 2015). 그리고 각 토픽의 토픽용어를 WORD2VEC으로 워드 임베딩하여 시각화한 뒤, 그중에서 단어거리가 먼 것이 실제 토픽의 의미와 분리된 단어인지 평가하였다.

5. 연구결과

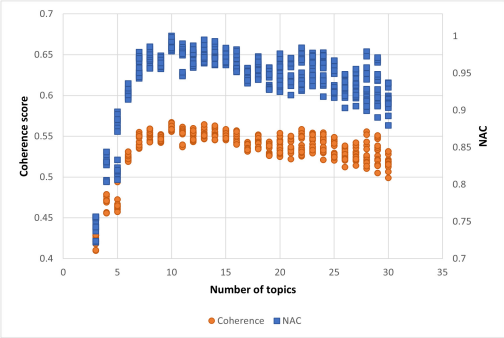
5.1 LDA 토픽모델링 하이퍼파라미터 최적화

최적의 LDA 하이퍼파라미터를 구하기 위한 연산 결과를 시각화하면 〈그림 4〉와 같다.

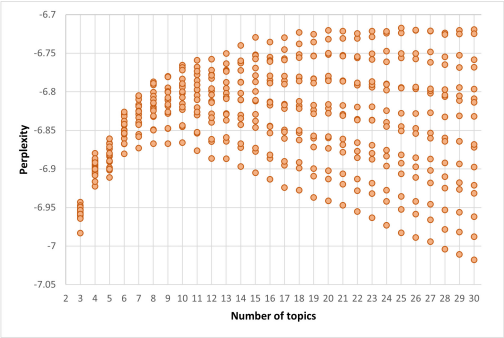
Gensim 환경에서 LDA 기반 토픽모델링을 수행하기 위해서는 먼저 연구자가 보통 k로 불리는 토픽 개수(num_topics), 그리고 문서-토픽 사전확률 α(alpha)와 토픽-단어 사전확률 η(eta)를 결정해야 한다. 최고 coherence score는 〈num_topics = 10, alpha = 0.31, eta = 0.61〉에서 0.5672였고, 최저 coherence score는 〈num_topics = 3, alpha = 0.01, eta = 0.61〉에서 0.4095였다. 이외에도, perplexity score도 계산해보았지만, 토픽의 개수가 커질수록 분산하는 경향이 확인되어 LDA 하이퍼파라미터 성능을 판별하는데 적절하지 않았다.

iterations는 토픽 확률을 구하기 위해 코퍼스 내 문서마다 수행하는 최대 계산 횟수의 범위를 말한다. num_topics = 10, alpha = 0.31, eta = 0.61을 고정한 채, iterations 변화에 따른 최고 coherence score는 iterations = 40에서 0.5515였고, 최저 coherence score는 iterations = 10에서 0.5423였다. NAC도 iterations가 40일 때 가장 높은 수치를 나타내었다. 따라서, 본 연구에 사용된 입력데이터는 iterations = 40일 때 가장 의미 있는 분류가 가능하다고 판단하였다.

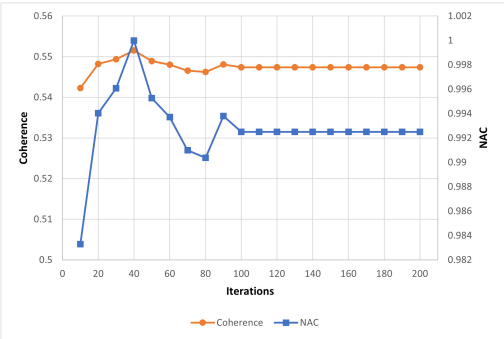
passes는 전체 코퍼스 수준에서 모델을 학습하는 횟수로써, 머신러닝에서 epoch와 유사한 개념이다. 마찬가지로 num_topics = 10, alpha = 0.31, eta = 0.61, iterations = 40을 고정한 채, passes 변화에 따른 최고 coherence score는 passes = 4500에서 0.5589였고, 최저 coherence score는 passes = 100에서 0.5489였다. NAC도 iterations가 40일 때 가장 높은 수치를 나타내었다. 따라서, 본 연구에 사용된 입력데이터



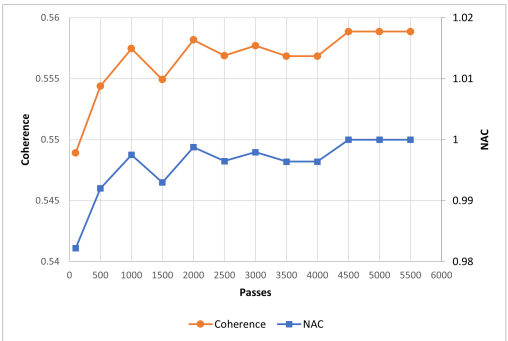
(num_topics, alpha, eta 조합에 따른 coherence 및 NAC 분포)



(num_topics, alpha, eta 조합에 따른 perplexity 분포)



(iterations에 따른 coherence 및 NAC 비교)



(passes에 따른 coherence 및 NAC 비교)

〈그림 4〉 최적의 LDA 하이퍼파라미터 탐색

는 passes = 4500일 때 가장 의미 있는 분류가 가능하다고 판단하였다.

5.2 LDA 토픽모델링 결과

최적의 토픽모델 LDAbow을 도출하기 위해 앞서 확보해둔 하이퍼파라미터를 Gensim의 models.ldamulticore API에 적용한 결과, 〈표 1〉과 같은 토픽모델링 결과가 도출되었다. 10개의 토픽에 각각 0부터 9까지 일련번호를 부여함으로써, 각 토픽을 〈T0〉에서 〈T9〉까지 명명하였다.

〈T0〉을 구성하는 토픽용어 중에서 확률 기준 상위 10개는 community, medium, article, communication, people, analysis, group, practice, way, work이고, 온라인 커뮤니티에서의 소통에 대한 토픽이다. 〈T1〉을 구성하는 토픽용어 중에서 확률 기준 상위 10개는 journal, article, citation, publication, author, field, science, analysis, impact, researcher이고, 연구자/저자 영향도와 인용분석에 대한 토픽이다. 〈T2〉를 구성하는 토픽용어 중에서 확률 기준 상위 10개는 health, patient, woman, care, intervention, risk, participant, child, treatment, disease이고, 건강 취약계층 진료에 대한 토픽이다. 〈T3〉를 구성하는 토픽용어 중에서 확률 기준 상위

〈표 1〉 LDA 토픽모델 결과 - 토픽과 토픽용어

토픽번호	토픽의미	토픽용어
T0	온라인 커뮤니티에서의 소통	community, medium, article, communication, people, analysis, group, practice, way, work
T1	연구자/저자 영향도와 인용분석	journal, article, citation, publication, author, field, science, analysis, impact, researcher
T2	건강취약계층 진료	health, patient, woman, care, intervention, risk, participant, child, treatment, disease
T3	이용자 행동의도	user, factor, effect, behavior, model, use, consumer, finding, trust, intention
T4	웹 정보원 관리와 검색	user, web, search, content, document, review, database, website, finding, source
T5	대학도서관에서의 학습활동 지원	library, student, librarian, service, resource, university, learning, collection, education, program
T6	지식공유가 기업의 혁신역량/혁신성과에 미치는 영향	knowledge, organization, business, performance, firm, innovation, relationship, finding, capability, company
T7	신경망 모델과 머신러닝을 사용한 분류	method, model, approach, network, algorithm, classification, analysis, performance, feature, technique
T8	정보체계 관련 이론	system, process, technology, development, approach, model, framework, design, implementation, theory
T9	인터넷 접근규제정책이 창업/기업에 미치는 영향	service, government, policy, market, country, access, cost, technology, internet, network

10개는 user, factor, effect, behavior, model, use, consumer, finding, trust, intention이고, 이용자 행동의도에 대한 토픽이다. 〈T4〉를 구성하는 토픽용어 중에서 확률 기준 상위 10개는 user, web, search, content, document, review, database, website, finding, source이고, 웹 정보원 관리와 검색에 대한 토픽이다. 〈T5〉를 구성하는 토픽용어 중에서 확률 기준 상위 10개는 library, student, librarian, service, resource, university, learning, collection, education, program이고, 대학도서관에서의 학습활동 지원에 대한 토픽이다. 〈T6〉을 구성하는 토픽용어 중에서 확률 기준 상위 10개는 knowledge, organization, business, performance, firm, innovation, relationship, finding, capability, company이고, 지식공유가 기업의 혁신역량/혁신성과에 미치는 영향에 대한 토픽이다. 〈T7〉을 구성하는 토픽용어 중에서 확률 기준 상위 10개는 method, model, approach, network, algorithm, classification, analysis, performance, feature, technique이고, 신경망 모델과 머신러닝을 사용한 분류에 대한 토픽이다. 〈T8〉을 구성하는 토픽용어 중에서 확률 기준 상위 10개는 system, process, technology, development, approach, model, framework, design, implementation, theory이고, 정보체계 관련 이론에 대한 토픽이다. 마지막으로, 〈T9〉을 구성하는 토픽용어 중에서 확률 기준 상위 10개는 service, government, policy, market, country, access, cost, technology, internet, network이고, 인터넷 접근규제정책이 창업/기업에 미치는 영향에 대한 토픽이다.

5.3 BERTopic 토픽모델링 결과

LDAbow의 증강/확장 대상으로 평가할 FT 입력 BERTopic 토픽모델링 결과인 토픽모델 BERTft를 확보하기 위해 1개의 GPU를 동원한 처리에 걸린 총 연산시간은 23분 49초였다. 총 377개 토픽이 도출[12]되었으며, LDAbow와 마찬가지로 〈T0〉부터 〈T376〉까지 토픽번호를 명명하였다. 〈표 2〉는 377개 BERTft 토픽 중에서 top 10을 나타낸다. BERTft는 전처리를 거치지 않은 초록을 그대로 입력하여 얻은 결과이므로, 명사 이외의 품사를 지닌 단어나 복수 표현 등 LDAbow와 다른 형태의 토픽용어가 관찰된다.

5.4 AET 수행

5.4.1 코사인 유사도 매트릭스 재정렬

WORD2VEC으로 LDAbow 토픽 10개와 BERTft 토픽 377개에 대한 코사인 유사도를 계산하면 10×377 크기의 매트릭스가 만들어지지만, LDAbow와 BERTft 사이의 관계를 가늠하기 어려운 상태의 전형적인 희소 행렬이다. 본 논문에서 제안하는 매트릭스 재정렬 방안에 따라 재정렬을 반복 수행하면, 매번 10×10 크기의 매트릭스 세그먼트가 분리 생성되므로 377개의 BERTft 토픽으로부터 최대 37개의 세그먼트를 만들어 낼 수 있다.

이렇게 매트릭스 재정렬을 반복해서 수행하는 과정에서, 〈그림 5〉와 〈표 3〉처럼 15번째 세그먼

〈표 2〉 BERTopic 토픽모델 결과 - 상위 10위 토픽과 토픽용어

토픽번호	문헌빈도	토픽용어
T0	2047	spatial, land, urban, gis, data, geographic, map, algorithm, model, method
T1	643	information literacy, literacy, il, students, instruction, information, skills, learning, librarians, teaching
T2	590	care, caregivers, illness, life, family, experiences, cancer, chronic, living, women
T3	492	health, health sciences, librarians, medical, library, sciences, skills, libraries, health information, students
T4	452	information systems, systems, research, discipline, theory, information, field, critical, paper, information systems research
T5	452	km, knowledge, knowledge management, management, organizational, culture, management km, knowledge management km, organizations, knowledge creation
T6	428	government, citizens, public, social media, governments, citizen, government services, media, services, service
T7	422	information seeking, seeking, information, behaviour, information behaviour, librarians, faculty, students, academic, information behavior
T8	393	privacy, disclosure, privacy concerns, concerns, personal, information privacy, personal information, protection, users, self disclosure
T9	392	innovation, firm, firms, knowledge, ict, investments, growth, smes, open innovation, productivity

12) BERTopic에서 생성된 토픽번호 -1은 이상치(outlier)를 의미하므로 토픽 해석과정에서 무시하였다.

	BERTft T115	BERTft T72	BERTft T192	BERTft T193	BERTft T28	BERTft T372	BERTft T242	BERTft T220	BERTft T344	BERTft T277
LDAbow T5	**0.6404**	0.2925	0.4421	0.3311	0.3748	0.3708	0.4087	0.2988	0.3199	0.3043
LDAbow T4	0.4118	**0.6015**	0.5240	0.4293	0.3961	0.3463	0.4063	0.2876	0.4826	0.3114
LDAbow T9	0.4734	0.3761	**0.5748**	0.2118	0.4465	0.3612	0.4047	0.2703	0.2902	0.3387
LDAbow T1	0.3926	0.3189	0.4065	**0.5641**	0.4912	0.2524	0.3669	0.2377	0.3284	0.2990
LDAbow T6	0.4607	0.3315	0.4535	0.2521	**0.5256**	0.3062	0.5203	0.4295	0.3511	0.4562
LDAbow T2	0.3931	0.2285	0.3121	0.2240	0.2560	**0.5147**	0.4669	0.2330	0.1970	0.4166
LDAbow T0	0.5076	0.4166	0.4623	0.4057	0.4664	0.3366	**0.5114**	0.4373	0.3838	0.4544
LDAbow T8	0.4090	0.3721	0.4625	0.2756	0.4575	0.1534	0.3767	**0.5068**	0.4041	0.2822
LDAbow T7	0.3643	0.4931	0.4353	0.3414	0.4032	0.1777	0.3128	0.3956	**0.4906**	0.2612
LDAbow T3	0.3731	0.4126	0.4069	0.2445	0.3662	0.3462	0.3854	0.4098	0.3154	**0.4867**

〈그림 5〉 최초 의미적 관계 손실 위치 - 15번째 세그먼트

〈표 3〉 15번째 세그먼트 내 최초 의미손실 관계

토픽번호	토픽용어
LDAbow T3	user, factor, effect, behavior, model, use, consumer, finding, trust, intention
BERTft T277	hiding, knowledge hiding, knowledge, territoriality, employees, psychological ownership, abusive supervision, hiding behavior, psychological, knowledge hiding behavior

트에서 코사인 유사도 0.4867를 갖는 LDAbow 〈T3〉와 BERTft 〈T277〉의 의미적 관계가 유실된 것이 최초 확인되었다. 이후 16번째 세그먼트와 17번째 세그먼트에서도 0.48* 대의 코사인 유사도를 갖는 토픽 관계에서 의미적 관계가 손실되는 것이 확인되어, 의미상 유사도를 기준 짓는 임계치는 0.49로 설정하였다.

5.4.2 토픽 증강(AT)

〈그림 3〉에서 서술한 AET 방법론에 따라, 1번째 매트릭스 세그먼트부터 임계치를 포함하지 않는 14번째 매트릭스 세그먼트에 해당하는 LDAbow 토픽과 BERTft 토픽을 합성하여 AT를 구성한다. 즉, AT는 1개의 LDAbow 토픽과, 이를 뒷받침하는 14개의 서브토픽(14개의 BERTft 토픽)을 합성해서 만들어진다.

AET의 AT 관점에서는, 각 LDAbow 토픽을 뒷받침하는 구체적 연구주제를 BERTft로부터 발굴해 낼 수 있다. 즉 연구자가 LDA를 통해 얻은 토픽으로는 추상적인 이해밖에 할 수 없었던 반면, AET는 한층 더 실제적인 토픽(AT)을 확보할 수 있으므로 연구자는 토픽의 이해도를 높이고 세분된 사례를 확인할 수 있다.

각 LDAbow 토픽이 증강된 내역을 정리하면 다음과 같다.[13] AT 중에서 지식과 정보에 관련된 연구주제가 높은 비중을 차지하고 있다. 전처리 데이터에 기반한 LDAbow의 토픽용어는 마치 전처리된 것처럼 보이나, BERTft는 자연어 기반 모델이므로 다양한 형태의 명사와 품사에 해당하는 단어를 포함하고 있다.

- 〈AT0〉 토픽: LDAbow 〈T0〉 토픽은 온라인 커뮤니티에서의 소통을 의미하는 것으로 이해되었으나 AET를 거쳐 지식/학습 공동체와 관련된 다양한 연구 주제가 발견되었다.
- 〈AT1〉 토픽: LDAbow 〈T1〉 토픽은 연구자/저자 영향도와 인용분석을 의미하는 것으

13) 자세한 AT 내역은 [부록 1] ~ [부록 10]에 걸쳐 수록하였다.

로 이해되었으나 AET를 거쳐 계량서지학과 인용분석과 관련된 다양한 연구 주제가 발견되었다.

- 〈AT2〉 토픽: LDAbow 〈T2〉 토픽은 건강 취약계층 진료를 의미하는 것으로 이해되었으나 AET를 거쳐 보건의료와 관련된 다양한 연구 주제가 발견되었다.
- 〈AT3〉 토픽: LDAbow 〈T3〉 토픽은 이용자 행동의도를 의미하는 것으로 이해되었으나 AET를 거쳐 이용자(소비자) 심리와 관련된 다양한 연구 주제가 발견되었다.
- 〈AT4〉 토픽: LDAbow 〈T4〉 토픽은 웹 정보원 관리와 검색을 의미하는 것으로 이해되었으나 AET를 거쳐 온라인 정보원과 관련된 다양한 연구 주제가 발견되었다.
- 〈AT5〉 토픽: LDAbow 〈T5〉 토픽은 대학도서관에서의 학습활동 지원을 의미하는 것으로 이해되었으나 AET를 거쳐 도서관의 교육적 기능과 리터러시 교육과 관련된 다양한 연구 주제가 발견되었다.
- 〈AT6〉 토픽: LDAbow 〈T6〉 토픽은 지식공유가 기업의 혁신역량/혁신성과에 미치는 영향을 의미하는 것으로 이해되었으나 AET를 거쳐 지식과 기업활동과 관련된 다양한 연구 주제가 발견되었다.
- 〈AT7〉 토픽: LDAbow 〈T7〉 토픽은 신경망 모델과 머신러닝을 사용한 분류를 의미하는 것으로 이해되었으나 AET를 거쳐 데이터 사이언스와 관련된 다양한 연구 주제가 발견되었다.
- 〈AT8〉 토픽: LDAbow 〈T8〉 토픽은 정보체계 관련 이론을 의미하는 것으로 이해되었으나 AET를 거쳐 정보체계 개발 및 응용과 관련된 다양한 연구 주제가 발견되었다.
- 〈AT9〉 토픽: LDAbow 〈T9〉 토픽은 인터넷 접근규제정책이 창업/기업에 미치는 영향을 의미하는 것으로 이해되었으나 AET를 거쳐 기술관련 정책과 규제와 관련된 다양한 연구 주제가 발견되었다.

5.4.3 토픽 확장(ET)

최초 10×377 크기의 매트릭스로부터 AT영역을 제외하고 남은 10×237 크기의 매트릭스는 ET 후보군이 되지만, 코사인 유사도 임계치에 해당하는 0.49 이상인 값을 포함하는 BERTft 토픽은 ET 후보군에서 배제해야 한다. 이런 식으로 LDA와 연관성 높은 BERTft 토픽을 제거한 뒤, 최종 ET 후보군의 크기는 10×145으로 줄어들었다.

ET 후보군은 LDA 토픽에 대한 BERTtopic 토픽의 코사인 유사도 값의 집합이므로, ET 후보군에 랭킹을 부여하는 기준으로 topic size(토픽 크기 순위, 이하 TS)와 similarity average(코사인 유사도 평균값 순위, 이하 SA)를 고려할 수 있다. TS가 클수록 BERTopic이 주류(major)로 판단한 토픽을 의미한다. SA가 낮으면 LDA 토픽으로부터의 벡터 공간 내 거리가 멀어짐을 의미하므로, 이 값이 낮을수록 LDA가 발견하지 못한 토픽을 의미한다. TS 순위와 SA 순위 둘 다, ET 측면에서 중요한 의미를 갖는 특성이므로 다양한 β를 고려하여 F_β score를 구한 결과, F_1 score(조화평균)일 때 가장 우수한 것으로 판별되었다. TS 순위와 SA 순위의 F_1 score에 따라 정렬한 뒤 그 중 top 10을 선정한 결과는 아래와 같다.[14]

- 〈ET0〉 토픽: 국제임상표준용어체계(SNOMED CT)와 국제질병분류체계(ICD)
- 〈ET1〉 토픽: 개인정보

14) 자세한 ET 내역은 [부록 11]에 수록하였다.

- 〈ET2〉 토픽: 성소수자 포용
- 〈ET3〉 토픽: 특허
- 〈ET4〉 토픽: IT 아웃소싱
- 〈ET5〉 토픽: 잠자는 미녀의 문제
- 〈ET6〉 토픽: 약물상호작용 문제와 전자처방
- 〈ET7〉 토픽: 노벨상 수상자 대상 계량서지학적 분석
- 〈ET8〉 토픽: 교정시설 도서관
- 〈ET9〉 토픽: 보존처리

5.5 AET 평가

〈표 4〉는 AET 토픽별 카파통계량(K) 계산 결과를 담고 있다.

먼저 LDAbow의 각 토픽에 대한 평가를 확인하면, 두 명의 평가자 중 한 명이 모두 yes를 부여하였고, 그 결과 평균 0.94로 매우 큰 값을 나타내었다.

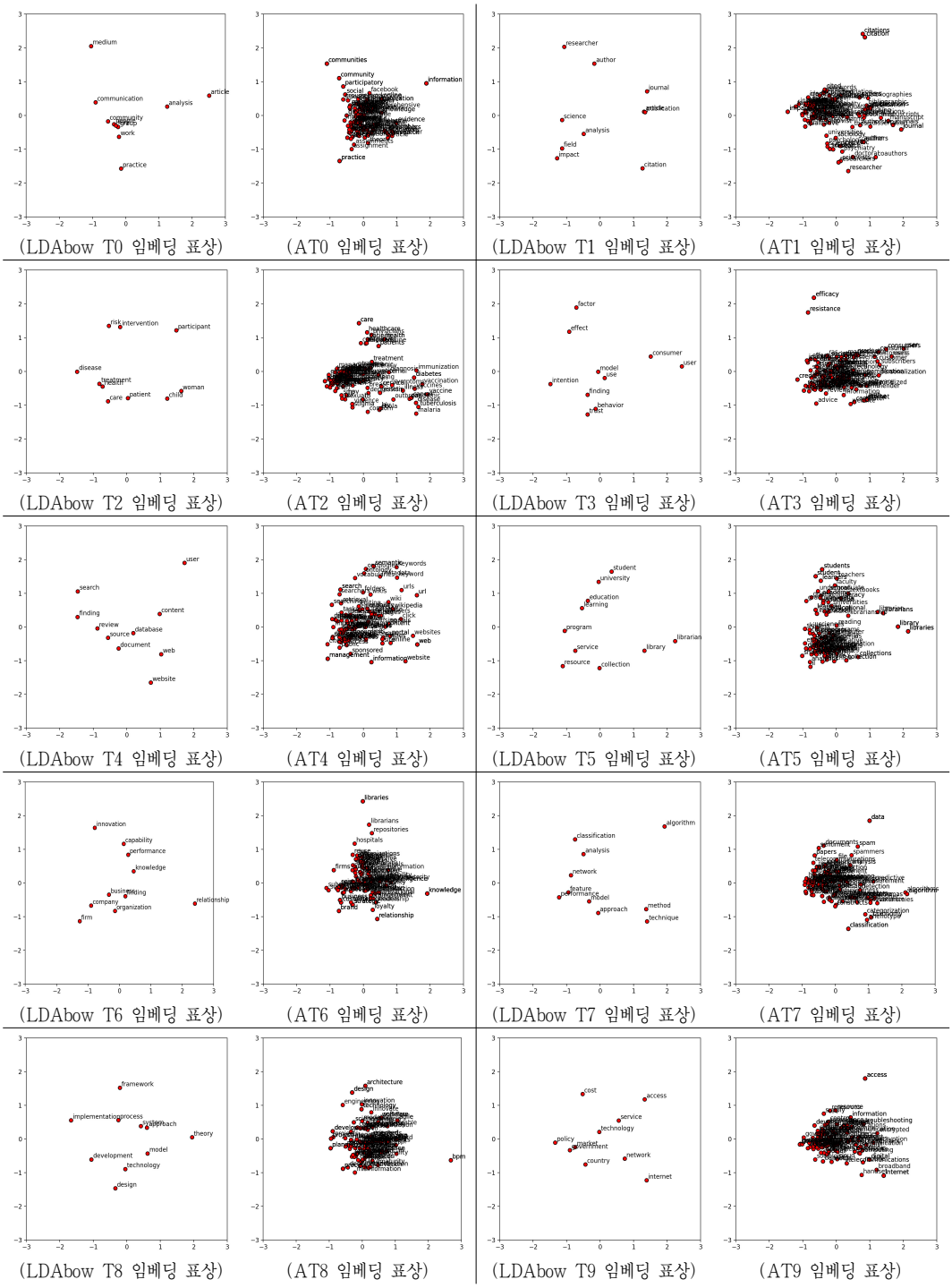
AT 토픽에 대한 평가를 확인하면, 먼저 10개 중 5개 토픽에 대한 카파통계량 계산이 가능했으며 〈AT0, AT3, AT5, AT8〉 토픽은 범위 0.61~0.80 (일치도 매우 높음) 또는 0.801~1.000 (일치도가 완벽함)이었으나, 〈AT1〉토픽은 0.201~0.400 (일치도가 어느 정도 높음)에 해당하였다. 나머지 5개 토픽은 두 명의 평가자 중 한 명이 모두 yes를 부여하여 카파통계량을 계산할 수 없어 일치율을 계산한 결과 평균 0.998로 매우 큰 값을 나타내었다.

ET 토픽에 대한 평가를 확인하면, 두 명의 평가자 모두 yes를 부여하여 카파통계량을 계산할 수 없었으며 일치율은 1.0으로 매우 큰 값을 나타내었다. 각 AT 토픽은 1개의 LDAbow 토픽(토픽용어 10개)을 중심으로 14개의 BERTft 토픽(토픽용어 140개)이 서브 토픽으로서 합성된 150개의 단어 집합임에도 불구하고, 본래의 LDAbow 토픽 못지않은 높은 수치의 카파통계량 및 일치율을 보였다.

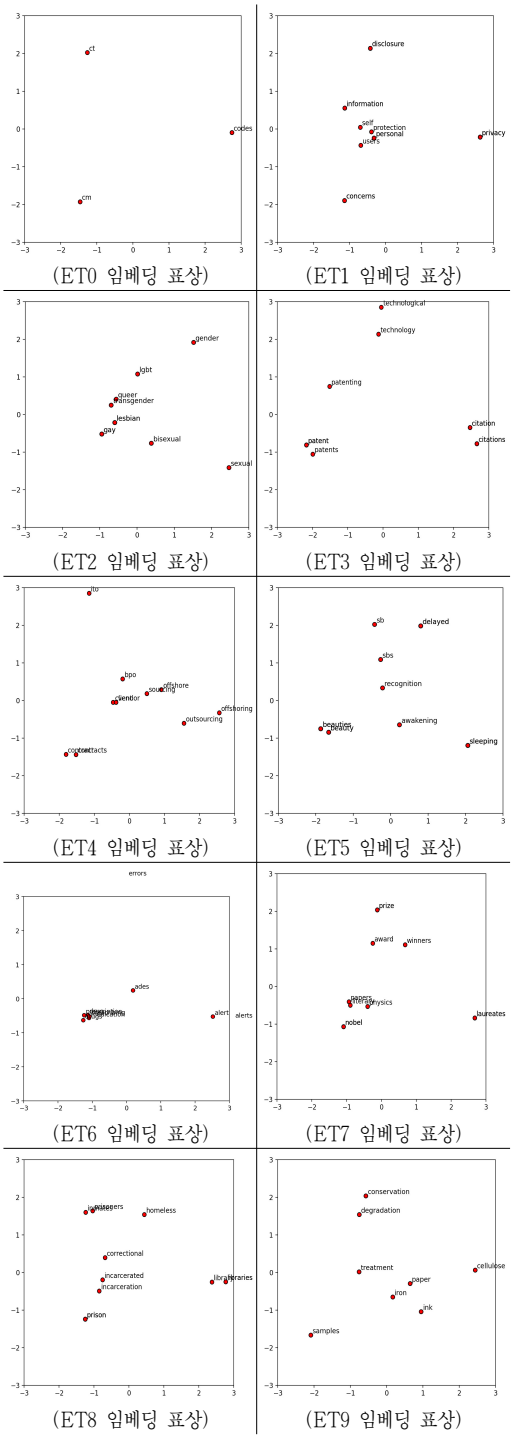
〈그림 6〉과 〈그림 7〉은 각각 LDA 대비 AT의 임베딩 표상과 ET의 임베딩 표상을 담고 있다. 단어의 임베딩을 2차원으로 시각화하면 상대적 거리를 통해 유사한 의미의 단어끼리 모이고 다소 의미가 멀어지는 단어의 거리는 멀어지는 양상을 확인할 수 있다. 단 여기서 주의할 점은 단어 사이의 거리가 상대적이라는 것이다. 이렇게 시각화하는 과정에서는, WORD2VEC 사전학습 모델에 포함되지 않아 코사인 유사도를 계산할 때는 조용히 무시되었던 예외값이 'not in vocabulary'는 오류를 반환하였는데 이를 정리한 것이 〈표 5〉이다.

〈표 4〉 AET 카파통계량(K)

토픽번호	LDAbow T 0	LDAbow T1	LDAbow T2	LDAbow T3	LDAbow T4	LDAbow T5	LDAbow T6	LDAbow T7	LDAbow T8	LDAbow T9
K	-	-	-	-	-	-	-	-	-	-
일치율	1	1	0.8	1	1	1	1	1	1	0.6
토픽번호	AT0	AT1	AT2	AT3	AT4	AT5	AT6	AT7	AT8	AT9
K	0.797	0.32	-	1	-	0.664	-	-	0.664	-
일치율	-	-	1	-	1	-	0.99	1	-	1
토픽번호	ET0	ET1	ET2	ET3	ET4	ET5	ET6	ET7	ET8	ET9
K	-	-	-	-	-	-	-	-	-	-
일치율	1	1	1	1	1<	1	1	1	1	1



〈그림 6〉 LDAbow와 AT 임베딩 표상 비교



〈그림 7〉 ET 임베딩 표상 비교

〈표 5〉 WORD2VEC 임베딩 불가 토픽용어

토픽번호	WORD2VEC 비학습 단어
AT0	behaviour, ebp, ebl, cbpr, ecrs, pubpeer, ohcs, sheehy
AT1	jif, orcid, peirce, disciplinarity, bourdieu
AT2	t1dm, ehrs, rcm, sociotechnical, covid 19, ipv, ptsd
AT3	delone, mclean, hcgs
AT4	opac, opacs, catalogue, web2p, sensemaking, skos
AT5	unima, cscl
AT6	smes, acap, scrm, cios, ekr, ekrs, mncs, mnc, mnes
AT7	temsets, huffman, diachronous, formatively, dbn
AT8	dsr, ethnomethodology, vsd, sisp, fsc, bpr, standardisation, scdma
AT9	ogd, blockchain, bitcoin, ict4d, icts, libguides
ET0	snomed, icd 10, icd, loinc, snomed ct, cm, icd 10 cm, 10 cm
ET1	—
ET2	lgbtq
ET3	uspto
ET4	—
ET5	—
ET6	cpoe
ET7	—
ET8	—
ET9	ageing, deacidification

AET의 각 AT 토픽에 대한 평가를 확인하면, 한정된 공간 내에서 일정한 간격을 두고 분포되어 있던 LDAbow 토픽 임베딩이 AET를 거쳐 높은 밀도로 밀집된 임베딩 분포로 변모한 것을 확인할 수 있다. AET의 각 ET 토픽에 대한 평가를 확인하면, 대부분 LDAbow 토픽처럼 일정한 간격을 두고 분포된 모습을 보이지만 〈ET0, ET4, ET6〉은 조금 더 멀리 분산된 형태를 나타낸다. 그중 〈ET0〉와 〈ET4〉 토픽은 예

외어를 여럿 포함하고 있으며, 〈ET6〉의 경우 토픽용어 ades가 ade의 복수로 처리된 것으로 보이지만 실제로는 adverse drug events를 의미하는 두문자어이다. 이처럼 미학습 단어가 등장하거나 일부 두문자어가 동음이의어로 잘못 해석된 경우, 단어 관계가 충실히 반영되지 못한 경우를 확인할 수 있었다.

6. 결 론

본 연구는 2001년 1월부터 2021년 10월까지 약 21년간 Web of Science에 등재된 문헌정보학 분야 55,442편의 논문을 대상으로 LDA와 BERTopic 토픽모델링 결과를 의미론적 유사도를 중심으로 합성함으로써, LDA가 이미 발견한 토픽은 BERTopic으로 증강하고 LDA가 발견하지 못한 토픽은 대신 BERTopic을 통해 발견하는 AET(Augmented and Extended Topics)를 제안하였다. 그 세부내용 및 결과는 다음과 같다.

첫째, LDA의 경우 하이퍼파라미터들을 시뮬레이션한 결과 토픽 개수(k)는 10개, 디리클레 분포는 alpha = 0.31과 eta = 0.61, Gensim 파라미터는 iterations = 40, passes = 4500으로 설정할 때 LDA 토픽모델링의 성능이 가장 높았다. BERTopic의 경우 연구자가 토픽의 개수를 직접 지정하지 않고 기본 알고리즘을 그대로 이용한 결과, 토픽 개수는 377로 결정되었다.

둘째, 확률기반 LDA 토픽모델링 결과와 임베딩기반 BERTopic 토픽모델링 결과의 의미론적 유사성을 파악하기 위해 WORD2VEC으로 LDA 토픽 10개와 BERTopic 토픽 377개에 대한 코사인 유사도를 계산하여 전형적인 희소 행렬인 10×377 크기의 매트릭스를 구축하였다. 본 논문에서 제안하는 코사인 유사도 재정렬 방안을 적용하여 의미적 관계가 손실되는 위치를 발견할 수 있었다.

셋째, 앞서 언급한 매트릭스 재정렬 알고리즘을 사용하여 LDA 토픽모델링 결과와 BERTopic 토픽모델링 결과로부터 AT(Augmented Topics)과 ET(Extended Topics)을 도출하였다. 그 결과 AT는 지식/학습 공동체, 계량서지학과 인용분석, 보건의료, 이용자(소비자) 심리, 온라인 정보원, 대학도서관에서의 학습활동 지원, 지식과 기업활동, 데이터 사이언스, 정보체계 개발 및 응용, 기술관련 정책과 규제에 대한 주제로 구분할 수 있었다. ET는 국제임상표준용어체계(SNOMED CT)와 국제질병분류체계(ICD), 개인정보, 성소수자 포용, 특허, IT 아웃소싱, 잠자는 미녀의 문제, 약물상호작용 문제와 전자처방, 노벨상 수상자 대상 계량서지학적 분석, 교정시설 도서관, 보존처리에 대한 주제로 확인되었다.

정성적 평가 과정에서, AT와 ET의 카파통계량은 LDA 토픽의 그것과 유사하거나 더 높은 수치를 나타냈다. 임베딩 표상으로 평가하였을 때 AT는 LDA보다 높은 응집력을 보였으며, ET는 〈ET0, ET4, ET6〉를 제외하고 대부분 LDA 토픽과 유사한 수준으로 분포되었다. WORD2VEC 사전학습 모델이 학습하지 못한 단어를 포함하고 있을 때, 단어 관계가 충실히 시각화되지 못한 것이 원인으로 추측된다.

본 연구의 한계를 바탕으로 한 후속 연구 제언은 다음과 같다. 첫째, 타 인용색인이나 표본데이터를 적용하면 토픽모델링의 결과가 상이

할 수 있다. 특히, 국내외 문헌정보학 연구주제 변화와 특성을 파악할 수 있는 연구가 진행되어야 할 것이다. 둘째, 타 학문에 적용하기 위해서는 적절한 보완이 필요할 수 있다. 따라서 타 학문 분야 학술지 데이터를 AET 분석함으로써 AET 범용성을 확인하거나 개선하기 위한 연구의 수행이 필요할 것이다. 셋째, 다른 언어 모델을 사용하면 결과가 상이할 수 있다. 따라서 추후 특정 주제영역(subject domain) 문헌만을 대상으로 머신러닝을 실시하여 범용 언어 모델 대비 AET 성능변화 여부를 확인할 수 있는 추가 연구 수행이 필요할 것이다.

본 연구는 기존에 찾아볼 수 없었던 토픽모델링 방법론을 설계함으로써, 대규모의 문헌정보학 연구논문 집합 속에 숨겨진 연구주제와 경향을 새로운 관점에서 발견했다는 것에 의의가 있다. 더불어 미래의 문헌정보학이 나아갈 수 있는 다양한 방향을 제시하였다는 점에서도 의의를 더할 수 있을 것이다. 이러한 시도는 더욱 향상된 형태로 연구주제 추이를 분석하거나 문헌을 구체적으로 주제분류하는 연구에 기여할 것으로 사료되며, 도출된 결과는 학계 연구자들이 더욱 새롭고 창의적인 연구과제를 모색하고, 깊이 있는 연구주제를 발굴하는 데에 비계가 될 것으로 기대한다.

참 고 문 헌

강보라, 김희섭 (2017). 국내 디지털 도서관 연구 동향 분석. 정보관리학회지, 34(3), 49-66. https://doi.org/10.3743/KOSIM.2017.34.3.049

김동기, 한무영, 한혜리 (2004). 의학통계학적 방법의 사용과 오류. 신경정신의학, 43(2), 141-147.

박순욱, 김영국, 김명호 (2021). 코로나19 재난문자 데이터를 활용한 의도 분류 모델 설계. 한국정보과학회 학술발표논문집, 1810-1812.

박종도 (2019). 토픽 모델링을 활용한 다문화 연구의 이슈 추적 연구. 한국문헌정보학회지, 53(3), 273-289. https://doi.org/10.4275/KSLIS.2019.53.3.273

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.48.201712.235

박창언, 김현정 (2015). 체계적 문헌고찰에서 평가자 간의 신뢰도 측정. Hanyang Medical Reviews, 35(1), 44-49. https://doi.org/10.7599/hmr.2015.35.1.44

배장성, 이창기, 임수종, 김현기. (2020). BERT를 이용한 한국어 의미역 결정. 정보과학회논문지, 47(11), 1021-1026. https://doi.org/10.5626/JOK.2020.47.11.1021

송은영, 최회련, 이홍철 (2019). Word Embedding에 PCA를 적용한 개체명 인식 모델을 위한 효율적

인 학습방법 연구. 대한산업공학회지, 45(1), 30-39. https://doi.org/10.7232/JKIIE.2019.45.1.030

양기덕, 김선욱, 이혜경 (2021). 국제 및 국내 문헌정보학 분야의 연구성과 비교 분석. 한국문헌정보학회지, 55(1), 365-392. http://dx.doi.org/10.4275/KSLIS.2021.55.1.365

윤상훈, 김근형 (2021). Word2Vec를 이용한 토픽모델링의 확장 및 분석사례. 정보시스템연구, 30(1), 45-64. http://dx.doi.org/10.5859/KAIS.2021.30.1.45

이다빈, 최성필 (2019). 대용량 텍스트 자원을 활용한 한국어 형태소 임베딩의 모델별 성능 비교 분석. 정보과학회논문지, 46(5), 413-418. http://doi.org/10.5626/JOK.2019.46.5.413

이유빈, 이영호, 성정창, 애나 스타네스쿠, 지상훈, 황철수 (2020). 계량적 모델을 통한 지리학 연구의 최신동향 및 토픽 분석. 대한지리학회지, 55(6), 589-599. http://doi.org/10.22776/kgs.2020.55.6.589

임소라, 권용진. (2017). 특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류. 인터넷정보학회논문지, 18(1), 77-88. https://doi.org/10.7472/jksii.2017.18.1.77

최원준, 설재욱, 정희석, 윤화묵. (2018). 국내 학술논문 주제 분류 알고리즘 비교 및 분석. 한국콘텐츠학회논문지, 18(8), 178-186. http://doi.org/10.5392/JKCA.2018.18.08.178

황승연, 안윤빈, 신동진, 오재곤, 문진용, 김정준 (2020). 빅데이터 기반 문서 토픽 추출 시스템 연구. 한국인터넷방송통신학회 논문지, 20(5), 207-214. http://doi.org/10.7236/JIIBC.2020.20.5.207

Ajayi, D. (2020). How BERT and GPT models change the game for NLP. Available: https://www.ibm.com/blogs/watson/2020/12/how-bert-and-gpt-models-change-the-game-for-nlp/

Angelov, D. (2020). Top2vec: Distributed representations of topics. https://doi.org/10.48550/arXiv.2008.09470

Behrisch, M., Bach, B., Riche, H. N., Schreck, T., & Fekete, J. D. (2016). Matrix reordering methods for table and network visualization. In Computer Graphics Forum, 35(3), 693-716. https://doi.org/10.1111/cgf.12935

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. Advances in Neural Information Processing Systems, 13.

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

Bodrunova, S. S., Orekhov, A. V., Blekanov, I. S., Lyudkevich, N. S., & Tarasov, N. A. (2020). Topic detection based on sentence embeddings and agglomerative clustering with markov moment. Future Internet, 12(9), 144. https://doi.org/10.3390/fi12090144

Chen, A. T., Sheble, L., & Eichler, G. (2013). Topic modeling and network visualization to explore

patient experiences. Visual Analytics in Healthcare Workshop 2013.

Choi, W. J. & Kim, E. (2019). A large-scale text analysis with word embeddings and topic modeling. Journal of Cognitive Science, 20(1), 147-187. http://doi.org/10.17791/jcs.2019.20.1.147

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%3C391::AID-ASI1%3E3.0.CO;2-9

Deveci, T. (2019). Sentence length in education research articles: a comparison between anglophone and turkish authors. The Linguistics Journal, 14(1), 73-100.

Ermann, L., Chepelianskii, A. D., & Shepelyansky, D. L. (2012). Toward two-dimensional search engines. Journal of Physics A: Mathematical and Theoretical, 45(27), 275101.

Esposito, F., Corazza, A., & Cutugno, F. (2016, December). Topic Modelling with Word Embeddings. CLiC-it/EVALITA. https://doi.org/10.4000/books.aaccademia.1666

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, 1-21. https://doi.org/10.1007/s11192-022-04275-z

Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. Science Advances, 4(7), eaaq1360. https://doi.org/10.1126/sciadv.aaq1360

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

Hasan, M., Rahman, A., Karim, M., Khan, M., Islam, S., & Islam, M. (2021). Normalized approach to find optimal number of topics in Latent Dirichlet Allocation(LDA). In Proceedings of International Conference on Trends in Computational and Cognitive Engineering, 341-354. Springer, Singapore. http://doi.org/10.1007/978-981-33-4673-4_27

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.

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

Li, C., Lu, Y., Wu, J., Zhang, Y., Xia, Z., Wang, T., Yu, D., Chen, X., Liu, P., & Guo, J. (2018, April). LDA meets Word2Vec: a novel model for academic abstract clustering. Companion

Proceedings of the Web Conference 2018, 1699-1706. https://doi.org/10.1145/3184558.3191629

Losee, R. M. (2001). Term dependence: A basis for Luhn and Zipf models. Journal of the American Society for Information Science and Technology, 52(12), 1019-1025. https://doi.org/10.1002/asi.1155

M'sik, B. & Casablanca, B. M. (2020). Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. International Journal, 9(4). https://doi.org/10.30534/ijatcse/2020/231942020

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ICLR. https://doi.org/10.48550/arXiv.1301.3781

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

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. https://doi.org/10.48550/arXiv.1802.05365

Schick, T. & Schütze, H. (2019). BERTRAM: Improved word embeddings have big impact on contextualized model performance. https://doi.org/10.48550/arXiv.1910.07181

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too. https://doi.org/10.48550/arXiv.2004.14914

Vayansky, I. & Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 101582. https://doi.org/10.1016/j.is.2020.101582

Wang, Y., Shi, Z., Guo, X., Liu, X., Zhu, E., & Yin, J. (2018). Deep embedding for determining the number of clusters. Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.12150

Yang, K., Lee, H., Kim, S., Lee, J., & Oh, D.-G. (2021). KCI vs. WoS: comparative analysis of Korean and international journal publications in library and information science. Journal of Information Science Theory and Practice, 9(3), 76-106. https://doi.org/10.1633/JISTAP.2021.9.3.6

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Bae, Jangseong, Lee, Changki, Lim, Soojong, & Kim, Hyunki (2020). Korean semantic role labeling

with BERT. Journal of Korean Institute of Information Scientists and Engineers, 47(11), 1021-1026. https://doi.org/10.5626/JOK.2020.47.11.1021

Choi, Won-Jun, Seol, Jae-Wook, Jeong, Hee-Seok, & Yoon, Hwamook (2018). Comparison and analysis of subject classification for domestic research data. The Journal of the Korea Contents Association, 18(8), 178-186. http://doi.org/10.5392/JKCA.2018.18.08.178

Hwang, Seung-Yeon, An, YoonBin, Shin, Dong-Jin, Oh, Jae-Kon, & Moon Jin-Yong (2020). A study on the document topic extraction system based on big data. The Journal of The Institute of Internet, Broadcasting and Communication, 20(5), 207-214. http://doi.org/10.7236/JIIBC.2020.20.5.207

Kang, Bora & Kim, Heesop (2017). An analysis of the digital library research trends in Korea. Journal of the Korean Society for Information Management, 34(3), 49-66. https://doi.org/10.3743/KOSIM.2017.34.3.049

Kim, Dong-Kee, Han, Mooyoung, & Han, Hae-Ree (2004). Use and misuse of biostatistical analysis. Korean Neuropsychiatric Association, 43(2), 141-147.

Lee, Da-Bin & Choi, Sung-Pil (2019). Comparative analysis of various Korean morpheme embedding models using massive textual resources. Journal of Korean Institute of Information Scientists and Engineers, 46(5), 413-418. http://doi.org/10.5626/JOK.2019.46.5.413

Lee, Yubin, Lee, Youngho, Seong, Jeongchang, Ana, Stanescu, Ji, Sanghoon, & Hwang, Chul-Sue (2020). An analysis of the latest trends and topics in geography research using topic modeling. Journal of the Korean Geographical Society, 55(6), 589-599. http://doi.org/10.22776/kgs.2020.55.6.589

Lim, Sora & Kwon, YongJin (2017). IPC multi-label classification based on functional characteristics of fields in patent documents. Journal of Internet Computing and Services, 18(1), 77-88. https://doi.org/10.7472/jksii.2017.18.1.77

Park, ChangUn & Kim, HyungJung (2015). Measurement of inter-rater reliability in systematic review. Hanyang Medical Reviews, 35(1), 44-49. https://doi.org/10.7599/hmr.2015.35.1.44

Park, Jong-Do (2019). A study on issue tracking on multi-cultural studies using topic modeling. Journal of the Korean Library and Information Science, 53(3), 273-289. https://doi.org/10.4275/KSLIS.2019.53.3.273

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.48.201712.235

Park, Soonwook, Kim, Youngkook, & Kim, Myungho (2021). A design for intent classification modelswith covid-19 disaster alerts data. Proceedings of the 2021 Korea Computer Congress, 1810-1812.

Song, Eun-Young, Choi, Hoe-Ryeon, & Lee, Hong-Chul (2019). A study on efficient training method for named entity recognition model with word embedding applied to PCA. Journal of the Korean Institute of Industrial Engineers, 45(1), 30-39. https://doi.org/10.7232/JKIIE.2019.45.1.030

Yang, Kiduk, Kim, SeonWook, & Lee, HyeKyung (2021). Comparison of research performance between domestic and international library and information science scholars. Journal of the Korean Library and Information Science, 55(1), 365-392. http://dx.doi.org/10.4275/KSLIS.2021.55.1.365

Yoon, Sang-Hun & Kim, Keun-Hyung (2021). Expansion of topic modeling with Word2Vec and case analysis. The Journal of Information Systems, 30(1), 45-64. http://dx.doi.org/10.5859/KAIS.2021.30.1.45

[부록 1] AT0 토픽 – 지식/학습 공동체

순번	토픽번호	코사인 유사도	토픽 용어
0	**LDAbow T0**		**community, medium, article, communication, people, analysis, group, practice, way, work**
1	BERTft T327	0.654504	group, group members, group work, students, collaborative, isp model, assignment, behaviour, assignments, isp
2	BERTft T152	0.61139	cops, cop, communities practice, communities, practice, police, knowledge, community, members, sharing
3	BERTft T352	0.600508	journals, network, journals structural, editorial, interlocking, ebi, structural influence, network analysis, communication journals, communication
4	BERTft T341	0.586517	students, online assessment, radar, assessment, tasks, information tools, academic search tasks, online, search, university students
5	BERTft T292	0.5804	evidence, evidence based, based practice, ebp, evidence based practice, ebl, practice, information practice, librarians, critical appraisal
6	BERTft T285	0.580263	cbpr, community, translation, researcher, participatory research cbpr, research cbpr, participatory research, participatory, based participatory research, language
7	BERTft T4	0.570772	information systems, systems, research, discipline, theory, information, field, critical, paper, information systems research
8	BERTft T298	0.564369	ecrs, scholarly, pubpeer, scholarly communication, communication, researchers ecrs, career researchers ecrs, ethics trust, trust, early career researchers
9	BERTft T218	0.563018	focus groups, focus, groups, focus group, group, face, participants, qualitative, saturation, interaction
10	BERTft T58	0.554346	community, communities, virtual, online, knowledge, virtual communities, members, online communities, virtual community, knowledge sharing
11	BERTft T180	0.539734	information, visual, information behaviour, work, information activities, practice, behaviour, information work, activities, information science
12	BERTft T90	0.529095	support, social support, online, online health, health communities, online health communities, health, communities, emotional, ohcs
13	BERTft T287	0.52621	job, job search, recruitment, social media, hiring, linkedin, networking, social media screening, media screening, facebook
14	BERTft T246	0.515807	selection recent scholarly, continued eugene sheehy, semiannual series, editions standard works, eugene sheehy, roundup new, roundup new editions, balanced comprehensive brief, scholarly general, article follows pattern

[부록 2] AT1 토픽 – 계량서지학과 인용분석

순번	토픽번호	코사인 유사도	토픽 용어
0	**LDAbow T1**		**journal, article, citation, publication, author, field, science, analysis, impact, researcher**
1	BERTft T13	0.719976	journal, journals, impact factor, jif, impact, factor, journal impact, citation, citations, journal impact factor
2	BERTft T172	0.705201	psychology, journals, articles, psychiatric, publications, cited, published, papers, negative results, psychiatry
3	BERTft T111	0.677806	recommendation, papers, citation, scientific, text, task, reviewers, based, text spans, researchers
4	BERTft T262	0.65009	submission, journal, editorial, delay, journals, manuscripts, manuscript, acceptance, submitted, online posting
5	BERTft T63	0.648781	journals, english, international, chinese, language, publishing, journal, publication, internationality, editorial
6	BERTft T27	0.645352	qualitative, qualitative research, research, researchers, qualitative health, health research, qualitative researchers, health, researcher, article
7	BERTft T210	0.619613	citation, citations, papers, articles, distributions, indicator, impact, mean, publications, indicators
8	BERTft T14	0.619093	clustering, topic, clusters, cluster, topics, citation, analysis, scientific, science, keywords
9	BERTft T70	0.616198	countries, china, science, papers, usa, water, research, publications, publication, output
10	BERTft T173	0.604137	dissertations, theses, citations, students, bibliographies, cited, citation, doctoral, thesis, dissertation
11	BERTft T176	0.602928	author, credit, authorship, harmonic, coauthors, authors, counting, attribution, authorship attribution, index
12	BERTft T137	0.575879	disambiguation, author, names, author disambiguation, orcid, author names, clustering, authors, bibliographic, identifiers
13	BERTft T370	0.573835	peirce, disciplinarity, knowledge constraints, sciences, bourdieu, science, peirce classification, reputational work, relativist sociology science, relativist sociology
14	BERTft T78	0.566076	mobility, scientists, usage statistics, scientific, researchers, productivity, statistics, international, universities, research

[부록 3] AT2 토픽 – 보건의료

순번	토픽번호	코사인 유사도	토픽 용어
0	LDAbow T2		health, patient, woman, care, intervention, risk, participant, child, treatment, disease
1	BERTft T25	0.759875	cancer, risk, screening, cancer information, health, patients, genetic, patient, breast, information
2	BERTft T206	0.732265	diabetes, self management, type diabetes, t1dm, self, care, patients, diabetes care, patient, shared decision making
3	BERTft T2	0.727966	care, caregivers, illness, life, family, experiences, cancer, chronic, living, women
4	BERTft T18	0.712432	ehr, hie, health, healthcare, care, electronic health, physicians, ehrs, adoption, patient
5	BERTft T30	0.676112	mental, mental health, depression, women, mental illness, illness, care, health, recovery, mothers
6	BERTft T306	0.668908	hit, healthcare, health, health research, care, health care, rcm, sociotechnical, ista, contingencies
7	BERTft T338	0.652699	tb, tuberculosis, treatment, tuberculosis tb, adherence, malaria, diagnosis, symptoms, stigma, care
8	BERTft T134	0.628065	portal, patient, patients, patient portal, portal use, patient portals, care, telemedicine, portals, intervention
9	BERTft T199	0.606976	mhealth, mobile health, health, mobile, healthcare, sms, mch, maternal, midwives, mch information
10	BERTft T44	0.602786	covid 19, covid, 19, pandemic, disease, surveillance, outbreak, public health, health, ebola
11	BERTft T50	0.591128	health literacy, literacy, health, numeracy, low health, patients, low health literacy, low, skills, adults
12	BERTft T112	0.585876	vaccine, hpv, vaccination, hpv vaccine, cervical, parents, vaccines, cervical cancer, immunization, cancer
13	BERTft T45	0.527527	hiv, aids, hiv aids, sexual, sex, stigma, men, prevention, living hiv, condom
14	BERTft T128	0.519483	survivors, violence, trauma, ipv, abuse, women, ptsd, intimate, intimate partner, victims

[부록 4] AT3 토픽 – 이용자(소비자) 심리

순번	토픽번호	코사인 유사도	토픽 용어
0	**LDAbow T3**		**user, factor, effect, behavior, model, use, consumer, finding, trust, intention**
1	BERTft T35	0.645424	trust, shopping, commerce, online, consumers, purchase, consumer, perceived, website, intention
2	BERTft T47	0.592284	acceptance, tam, technology acceptance, technology, intention, adoption, model, use, perceived, acceptance model
3	BERTft T87	0.587196	recommendation, recommender, collaborative filtering, recommender systems, filtering, items, recommendations, preferences, users, user
4	BERTft T74	0.550969	price, pricing, product, market, prices, consumers, software, consumer, optimal, products
5	BERTft T254	0.543804	ras, personalization, ra, recommendation, web personalization, consumers, recommendation agents, product, trade, personalized
6	BERTft T161	0.53015	success, adoption, organizational, user satisfaction, assimilation, leadership, satisfaction, delone mclean, delone, mclean
7	BERTft T280	0.53007	resistance, user resistance, user, work routines, habits, resistance behaviors, implementation, resistance information, routines, acceptance resistance
8	BERTft T188	0.526359	gamification, gamified, games, game, engagement, gamification elements, elements, hcgs, citizen science, hedonic
9	BERTft T320	0.519808	internet use, ties, internet, adolescents, social, online communication, online, relationships, sns, social relationships
10	BERTft T228	0.517064	churn, customer, switching, subscribers, satisfaction, mobile, customer satisfaction, service, customers, switching costs
11	BERTft T331	0.513857	cse, self efficacy, efficacy, computer self, computer self efficacy, computer, self, anxiety, self efficacy cse, efficacy cse
12	BERTft T125	0.497165	impact, impacts, societal, research, impact assessment, societal impact, research impact, assessment, impact research, evaluation
13	BERTft T212	0.495273	credibility, credibility assessment, cues, ratings, advice, web, web credibility, decision aid, user ratings, credibility judgments
14	BERTft T54	0.490141	reviews, product, review, online reviews, helpfulness, online, consumers, ratings, review helpfulness, products

[부록 5] AT4 토픽 – 온라인 정보원

순번	토픽번호	코사인 유사도	토픽 용어
0	**LDAbow T4**		**user, web, search, content, document, review, database, website, finding, source**
1	BERTft T32	0.745192	web, usability, links, websites, link, sites, site, pages, website, web sites
2	BERTft T302	0.71401	ecm, portal, enterprise content, enterprise content management, content management, enterprise, portals, management, portal definition document, portal definition
3	BERTft T202	0.692352	web, web tools, web technologies, web applications, libraries, tools, applications, technologies, use web, librarians
4	BERTft T168	0.671954	opac, opacs, library, catalogue, online public access, access, online public, portal, catalog, portals
5	BERTft T283	0.666128	cms, content management, content, site, management cms, content management cms, web, management, web2py, database driven
6	BERTft T51	0.658076	search, task, relevance, tasks, searching, retrieval, user, information retrieval, ir, searchers
7	BERTft T186	0.65483	urls, web, web citations, citations, url, references, url citations, cited, articles, web resources
8	BERTft T160	0.647796	linked data, linked, ld, semantic web, vocabularies, data, semantic, rdf, metadata, web
9	BERTft T195	0.645776	discovery, usability, interfaces, search, interface, searching, catalog, discovery layers, box, library
10	BERTft T179	0.635526	collaborative, cis, collaborative information, search, information seeking, collaborative information seeking, sensemaking, seeking, collaborative search, tasks
11	BERTft T86	0.626913	wikipedia, wikis, wiki, editing, contributors, quality, content, articles, knowledge, wikipedia articles
12	BERTft T236	0.607626	pim, personal information, personal, folders, pim reference, pim reference management, reference management, information management, personal information management, files
13	BERTft T185	0.605776	advertising, ad, sponsored, advertisers, sponsored search, search, keywords, click, revenue, keyword
14	BERTft T130	0.604795	ontology, ontologies, semantic, semantic web, domain, matching, web, knowledge, skos, core ontology

[부록 6] AT5 토픽 – 도서관의 교육적 기능과 리터러시 교육

순번	토픽번호	코사인 유사도	토픽 용어
0	**LDAbow T5**		**library, student, librarian, service, resource, university, learning, collection, education, program**
1	BERTft T301	0.788776	libraries, academic programs, academic, faculty staff, academic libraries, research libraries, universities, university research libraries, collections, staff
2	BERTft T73	0.761466	children, young, reading, public libraries, public, public library, library, libraries, school, literacy
3	BERTft T122	0.723099	student, library, student success, students, retention, graduate, academic, gpa, success, undergraduate
4	BERTft T304	0.717676	learning analytics, analytics, learning, student, privacy, ethical, attention engineering, student data, education, library
5	BERTft T42	0.717028	libraries, library, public, public library, public libraries, public sphere, sphere, librarians, librarianship, digital library
6	BERTft T3	0.716387	health, health sciences, librarians, medical, library, sciences, skills, libraries, health information, students
7	BERTft T1	0.706193	information literacy, literacy, il, students, instruction, information, skills, learning, librarians, teaching
8	BERTft T279	0.694268	school, library, school libraries, libraries, shanghai library, china, chinese, shanghai, librarians, unima
9	BERTft T314	0.677466	transfer students, transfer, students, transfer student, student, instruction, library, il instruction, campus, transfer shock
10	BERTft T20	0.659107	learning, cscl, collaborative, students, collaborative learning, group, learners, teachers, computer supported, collaboration
11	BERTft T201	0.651881	oer, textbook, open educational, affordability, course, textbooks, open educational resources, educational resources, course materials, open
12	BERTft T97	0.649238	collection, collections, collection development, special collections, circulation, library, libraries, weeding, collection management, print
13	BERTft T100	0.646415	space, library, spaces, noise, library space, wayfinding, learning, place, physical, libraries
14	BERTft T300	0.64206	distance, distance education, distance learning, education, distance learners, library, learners, distance students, services distance, librarians

[부록 7] AT6 토픽 – 지식과 기업활동

순번	토픽번호	코사인 유사도	토픽 용어
0	**LDAbow T6**		**knowledge, organization, business, performance, firm, innovation, relationship, finding, capability, company**
1	BERTft T9	0.736342	innovation, firm, firms, knowledge, ict, investments, growth, smes, open innovation, productivity
2	BERTft T271	0.705966	absorptive capacity, innovation, absorptive, ambidexterity, acap, capacity, firm, capability, knowledge, collaborative innovation
3	BERTft T107	0.671128	alignment, strategic alignment, strategic, business, business alignment, strategy, business strategy, alignment business, performance, firm
4	BERTft T162	0.663331	agility, organizational agility, capabilities, capability, organizational, dynamic, dynamic capabilities, flexibility, firm, business
5	BERTft T233	0.647524	crm, customer, crm systems, customer relationship, relationship management, customer relationship management, relationship management crm, management crm, crm software, scrm
6	BERTft T147	0.64207	cio, cios, business, leadership, strategic, hospitals, organizations, executives, society information management, reporting structure
7	BERTft T367	0.607185	innovations, public libraries, libraries, strategic planning, innovation, strategic, planning, vision, strategic plans, innovation research libraries
8	BERTft T324	0.606446	ekr, knowledge, knowledge reuse, reuse, knowledge repositories, knowledge seeking, ekr usage, electronic knowledge, knowledge contribution, ekrs
9	BERTft T138	0.589026	transfer, knowledge transfer, knowledge, subsidiaries, subsidiary, multinational, mncs, firms, mnc, mnes
10	BERTft T5	0.580796	km, knowledge, knowledge management, management, organizational, culture, management km, knowledge management km, organizations, knowledge creation
11	BERTft T169	0.548596	brand, brand community, community, brand communities, brand loyalty, loyalty, engagement, online brand, customer, virtual brand
12	BERTft T183	0.546509	entrepreneurship, entrepreneurial, entrepreneurs, business, public libraries, business librarians, libraries, public, spin, small business
13	BERTft T142	0.541777	turnover, job, job satisfaction, satisfaction, career, turnover intention, professionals, personnel, commitment, intention
14	BERTft T113	0.525986	bi, business intelligence, intelligence, analytics, business, ba, bi systems, business analytics, business intelligence bi, intelligence bi

[부록 8] AT7 토픽 – 데이터 사이언스

순번	토픽번호	코사인 유사도	토픽 용어
0	**LDAbow T7**		**method, model, approach, network, algorithm, classification, analysis, performance, feature, technique**
1	BERTft T0	0.687117	spatial, land, urban, gis, data, geographic, map, algorithm, model, method
2	BERTft T96	0.625532	classification, text, categorization, text classification, text categorization, documents, feature, classifier, feature selection, label
3	BERTft T55	0.612546	phenotype, clinical, patients, predictive, ehr, prediction, algorithms, data, objective, models
4	BERTft T170	0.559514	efficiency, dea, envelopment, data envelopment analysis, envelopment analysis, data envelopment, technical efficiency, productivity, heis, telecommunications
5	BERTft T354	0.544216	taxonomy, taxonomies, thesauri, taxonomy development, organizational taxonomy, classification schemes, classification schemes thesauri, schemes thesauri, domain thesauri, navigation
6	BERTft T174	0.52618	mining, data mining, itemsets, data, rules, association rules, sensitive, frequent, text mining, algorithm
7	BERTft T190	0.516886	compression, inverted, compressed, file, inverted file, double array, space, tree, algorithm, huffman
8	BERTft T59	0.509885	sentiment, sentiment analysis, polarity, opinion, lexicon, sentiment classification, classification, aspect, reviews, words
9	BERTft T325	0.502277	distribution, gamma, gamma distribution, obsolescence, citation, diachronous, gamma distribution parameters, parameters, distribution parameters, papers
10	BERTft T339	0.50194	features, prediction, citation, feature space, papers, feature, citations, ppi model, external features, scholarly paper
11	BERTft T155	0.497134	pls, formative, measurement, constructs, formative measurement, construct, formative constructs, formatively, covariance, covariance based
12	BERTft T355	0.496942	data quality, dq, ehr, ehr data, data quality assessment, data, quality, completeness, quality assessment, dq dimensions
13	BERTft T263	0.495765	spam, detection, spammers, spam detection, features, dbn, ewa, feature, classification, emotion
14	BERTft T165	0.493805	conceptual, modeling, conceptual modeling, conceptual models, grammars, domain, models, schemas, conceptual schemas, diagrams

[부록 9] AT8 토픽 – 정보체계 개발과 응용

순번	토픽번호	코사인 유사도	토픽 용어
0	**LDAbow T8**		**system, process, technology, development, approach, model, framework, design, implementation, theory**
1	BERTft T120	0.685	dss, dsr, design, design science, decision, design science research, science research, decision support, design theory, support systems
2	BERTft T159	0.681816	workflow, process, business process, business, workflow management, process modeling, business processes, modeling, process management, processes
3	BERTft T290	0.650525	technology, trying innovate, inv, technology adoption, task technology fit, adoption, technology fit, task technology, influence, motivation
4	BERTft T362	0.644634	design, artefact, design science, grounded design, ethnomethodology, social impacts, tailorable, artifacts, vsd framework, ethnography
5	BERTft T88	0.644541	agile, software, software development, development, project, teams, spi, agile methods, team, agile development
6	BERTft T208	0.629056	evaluation, policy, innovation, innovation policy, additionality, tax incentives, firms, grants, program, evaluations
7	BERTft T322	0.611493	npd, product development, product, cpc, product development npd, development npd, new product development, collaboration, supplier, new product
8	BERTft T215	0.60442	requirements, analysts, elicitation, requirements elicitation, requirements determination, systems development, determination, requirements engineering, misinformation, misinformation effect
9	BERTft T351	0.596471	soa, soa implementation, soa governance, service, service oriented, oriented architecture soa, architecture soa, service oriented architecture, oriented architecture, shared services
10	BERTft T267	0.576156	planning, sisp, planning sisp, information systems planning, systems planning, strategic, fsc, strategic information, planning process, strategy
11	BERTft T282	0.573447	bpm, bpr, business process, maturity, maturity models, business, business process management, process, process management bpm, management bpm
12	BERTft T61	0.552778	erp, implementation, erp implementation, erp systems, enterprise, resource planning, enterprise resource, enterprise resource planning, planning erp, resource planning erp
13	BERTft T141	0.540901	standards, standardization, standard, china, standardisation, technology, firms, td, competition, scdma
14	BERTft T68	0.536401	project, projects, control, risk, project management, project managers, escalation, project risk, management, managers

[부록 10] AT9 토픽 – 기술관련 정책과 규제

순번	토픽번호	코사인 유사도	토픽 용어
0	**LDAbow T9**		**service, government, policy, market, country, access, cost, technology, internet, network**
1	BERTft T12	0.756698	broadband, telecommunications, competition, regulatory, policy, investment, regulation, market, access, infrastructure
2	BERTft T6	0.700607	government, citizens, public, social media, governments, citizen, government services, media, services, service
3	BERTft T272	0.674755	public, public libraries, internet access, access, internet, libraries, public access, public library, public internet, public internet access
4	BERTft T71	0.643335	mobile, market, operators, price, prices, diffusion, telecommunications, competition, handset, roaming
5	BERTft T19	0.623192	mobile, banking, adoption, perceived, intention, mobile banking, payment, brand, internet banking, satisfaction
6	BERTft T81	0.621178	open data, open, ogd, open government, government, data, government data, open government data, portals, public
7	BERTft T22	0.614824	security, information security, compliance, employees, isp, security policies, security policy, policies, threat, behavior
8	BERTft T102	0.608186	blockchain, blockchain technology, bitcoin, supply, technology, blockchain based, chain, supply chain, smart contracts, blockchain applications
9	BERTft T116	0.604782	divide, digital divide, digital, countries, ict, inequality, internet, digital development, access, economic
10	BERTft T67	0.601789	ict, ict4d, development, information communication, ict development, information communication technology, rural, communication technology, development ict4d, icts
11	BERTft T38	0.601618	network, networks, nodes, neutrality, social networks, social, node, net neutrality, content, diffusion
12	BERTft T175	0.597341	internet, history, sociological, arpanet, section, computers, technologies, computing, sociology, internet research
13	BERTft T288	0.592022	electronic resource, troubleshooting, libguides, electronic, er, electronic resources, resource, staff, access problems, resource troubleshooting
14	BERTft T326	0.584431	cloud, access control, security, encrypted, encryption, deduplication, scheme, secure, key derivation, access

[부록 11] ET 전체 토픽

토픽번호	원 토픽번호	토픽 용어	토픽 의미
ET0	BERTft T209	snomed, icd, loinc, codes, icd 10, ct, snomed ct, cm, icd 10 cm, 10 cm	국제임상표준용어체계(SNOMED CT)와 국제질병분류체계(ICD)
ET1	BERTft T8	privacy, disclosure, privacy concerns, concerns, personal, information privacy, personal information, protection, users, self disclosure	개인정보
ET2	BERTft T99	lgbtq, transgender, lesbian, gay, queer, lgbt, sexual, gender, lesbian gay, bisexual	성소수자 포용
ET3	BERTft T5	patent, patents, technological, technology, patent citation, patenting, citations, citation, patent citations, uspto	특허
ET4	BERTft T16	outsourcing, client, vendor, ito, offshore, contract, sourcing, offshoring, contracts, bpo	IT 아웃소싱
ET5	BERTft T216	sleeping, sbs, beauties, sleeping beauties, delayed recognition, delayed, sb, awakening, beauty, sleeping beauty	잠자는 미녀의 문제
ET6	BERTft T17	drug, medication, alerts, prescribing, alert, errors, cpoe, drugs, prescription, ades	약물상호작용 문제와 전자처방
ET7	BERTft T197	nobel, laureates, prize, nobel laureates, nobel prize, physics, papers, award, winners, literary	노벨상 수상자 대상 계량서지학적 분석
ET8	BERTft T163	prison, prison libraries, prisoners, homeless, correctional, prison library, incarcerated, libraries, inmates, incarceration	교정시설 도서관
ET9	BERTft T23	ink, ageing, iron, cellulose, paper, conservation, treatment, samples, deacidification, degradation	보존처리