ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 March: 28(1): 155~174 http://dx.doi.org/10.13088/jiis.2022.28.1.155

토픽모델링을 활용한 COVID-19 학술 연구 기반 연구 주제 분류에 관한 연구

유소연
한양대학교 경영대학
(soyeonyoo@hanyang.ac.kr)

임규건
한양대학교 경영대학
(gglim@hanyang.ac.kr)

2020년 1월부터 2021년 10월 현재까지 COVID-19(치명적인 호흡기 증후군인 코로나바이러스-2)와 관련된 학술 연구가 500,000편 이상 발표되었다. COVID-19와 관련된 논문의 수가 급격하게 증가함에 따라 의료 전문가와 정책 담당자들이 중요한 연구를 신속하게 찾는 것에 시간적·기술적 제약이 따르고 있다. 따라서 본 연구에서는 LDA와 Word2vec 알고리즘을 사용하여 방대한 문헌의 텍스트 자료로부터 유용한 정보를 추출하는 방안을 제시한다. COVID-19와 관련된 논문에서 검색하고자 하는 키워드와 관련된 논문을 추출하고, 이를 대상으로 세부 주제를 파악하였다. 자료는 Kaggle에 있는 CORD-19 데이터 세트를 활용하였는데, COVID-19 전염병에 대응하기 위해 주요 연구 그룹과 백악관이 준비한 무료 학술 자료로서 매주 자료가 업데이트되고 있다. 연구 방법은 크게 두 가지로 나뉜다. 먼저, 47,110편의 학술 논문의 초록을 대상으로 LDA 토픽 모델링과 Word2vec 연관어 분석을 수행한 후, 도출된 토픽 중 'vaccine'과 관련된 논문 4,555편, 'treatment'와 관련된 논문 5,791편을 추출한다. 두 번째로 추출된 논문을 대상으로 LDA, PCA 차원 축소 후 t-SNE 기법을 사용하여 비슷한 주제를 가진 논문을 군집화하고 산점도로 시각화하였다. 전체 논문을 대상으로 찾을 수 없었던 숨겨진 주제를 키워드에 따라 문헌을 분류하여 토픽 모델링을 수행한 결과 세부 주제를 찾을 수 있었다. 본 연구의 목표는 대량의 문헌에서 키워드를 입력하여 특정 정보에 대한 문헌을 분류할 수 있는 방안을 제시하는 것이다. 본 연구의 목표는 의료 전문가와 정책 담당자들의 소중한 시간과 노력을 줄이고, 신속하게 정보를 얻을 수 있는 방법을 제안하는 것이다. 학술 논문의 초록에서 COVID-19와 관련된 토픽을 발견하고, COVID-19에 대한 새로운 연구 방향을 탐구하도록 도움을 주는 기초자료로 활용될 것으로 기대한다.

주제어 : 코로나 19, 토픽 모델링, LDA(잠재 디리클레 할당), Word2vec, 키워드 추출

논문접수일 : 2021년 12월 30일 논문수정일 : 2022년 1월 24일 게재확정일 : 2022년 2월 1일
원고유형 : 학술대회용 Fast Track 교신저자 : 임규건

1. 서론

2020년 치명적인 호흡기 증후군 코로나 바이러스-2(이하 COVID-19)에 의해 2021년 10월 현재, 전 세계 COVID-19 확진자수는 2억 4천 5백만 명, 사망자 수는 497만 명으로 보고되고 있다. 2020년 1월부터 2021년 11월 현재까지 COVID-19와 관련된 학술 연구가 500,000 편 이상 발표되었다. COVID-19 위기를 맞아 이와 관련된 학술 논문 출판은 급격한 증가세를 보이고 이에 관한 분석도 다양한 방법으로 이루어지고 있다(Shin, 2021). 프랑스 네커병원 연구팀이 실시한 메타분석 결과, COVID-19 팬데믹 상황에서 전례 없는 속도로 쏟아지는 연구 논문의 품질은 현저하게 떨어진다는 연구 결과가 발표되었다(Chu, 2020). 논문의 수가 빠른 속도로 증가함에 따라 의료 전문

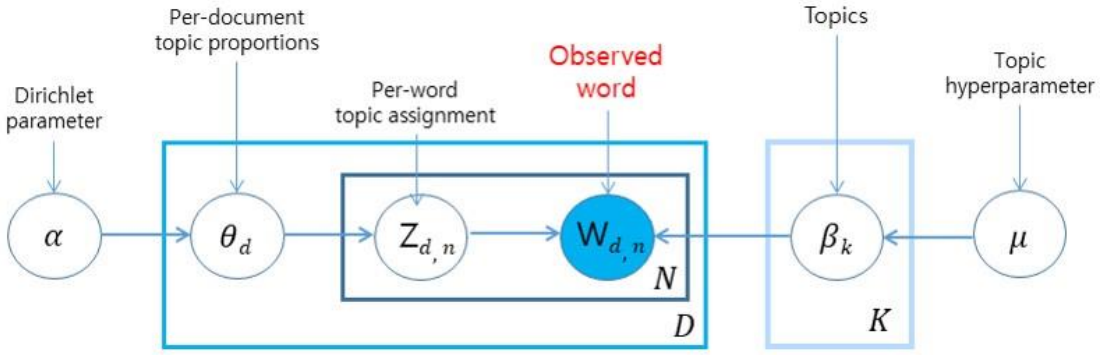
가와 정책 담당자들이 중요한 연구를 신속하게 찾는 것에 시간적 · 기술적 제약이 따르고 있다. 본 연구는 LDA와 Word2vec 알고리즘을 활용하여 대량의 문서에서 주요 토픽을 도출하고, 도출된 주제별 문서를 추출하는 방법을 제안하며, 문서를 주제별로 군집화하여 분류하고 시각화하는 방안을 제시하고자 한다. 본 연구의 목표는 (1) 대량의 문서에서 탐색하고자 하는 키워드를 입력하여 논문에 대한 특정 정보를 추출하는 방법을 제시하며, (2) LDA와 Word2vec 알고리즘을 사용하여 문헌에서 주제와 유사어를 추출하는 방식을 제안하며, (3) PCA 차원 감소를 통한 군집화 방법으로 문서의 구조화된 조직을 t-SNE알고리즘을 사용하여 시각화하는 최적의 방안을 탐색하고자 한다. 많은 과학자들이 그들이 소중한 노력과 시간을 기울임에도 불구하고, 급증하는 COVID-19와 관련된 학술 문헌을 따라잡을 수 없는 상황에 의료 전문가와 정책 담당자들에게 소중한 시간과 노력을 줄일 방법과 신속하게 원하는 정보를 얻을 방안을 제시하고자 한다.

본 연구의 순서는 다음과 같은 순서로 진행된다. 먼저 2장에서는 COVID 19 연구 논문의 분석에 대한 연구와 대량의 문서에서 키워드를 추출하는 연구에 대해 살펴보고, 3장에서는 본 연구에서 수행한 연구 방법을 제시하고, LDA, Word2vec, PCA, t-SNE 알고리즘에 대해 설명한다. 4장에서는 분석 결과와 시각화에 대해 보여주고, 5장에서는 본 연구의 결론과 향후 연구 방향에 대해 설명하고자 한다.

2. 선행 연구

급격히 증가하는 COVID-19와 관련된 학술 연구를 대상으로 머신러닝 기법을 사용하여 주제를 식별하거나, 정보를 추출하려는 연구가 활발히 진행되고 있다. Eren et al.(2020)은 치명적인 COVID-19에 관련된 과학 문헌이 급격히 증가함에 따라 COVID-19과 관련된 연구 데이터인 CORD19 데이터 세트를 가지고 학술 논문을 대상으로 머신러닝 기법을 적용하여 텍스트의 잠재적인 관계를 기반으로 논문을 매핑하여 시각화하는 방안을 제시하였다. COVID-19 관련 학술 논문을 지속적으로 문서화하는 작업은 머신러닝 알고리즘으로 분석하여 바이러스 확산 패턴을 더 잘 분석하고, 정확도와 개선 속도를 개선하며 효과적이고 새로운 치료 방법을 개발하는 등 실시간 데이터를 요구하는 새로운 연구 프로젝트를 가속화하는 데를 위해 사용할 수 있다고 한다(Alimadai et al., 2020). Ahamed et al.(2020)은 10,863개의 초록을 사용하여 ‘전염(transmission)’, ‘약물 유형(drug type)’, ‘게놈(genome)’의 세 가지 주제를 대상으로 핵심 정보를 찾는 그래프 기반 모델을 연구하였고, 머신러닝 기반 텍스트 분석은 많은 양의 과학 문헌에서 짧은 시간 안에 귀중한 정보를 얻을 수 있다고 주장하였다.

Verma와 Gustafsson(2020)은 COVID-19 문헌을 수집하여 비즈니스 및 관리 분야의 영역 연구 결과를 제시하고, COVID-19가 비즈니스에 미치는 영향, COVID-19 및 기술, 공급망 관리, 서비스 산업의 주요 4가지 연구주제와 18개의 하위 주제를 식별하였다. COVID-19가 비즈니스 및 관리에 미치는 영향에 관한 새로운 연구 동향을 서지(bibliometric) 및 텍스트 마이닝 기법을 사용하여 주요 연구 주제와 하위 주제에 대한 통찰력을 제시하였다.



〈Figure 1〉 LDA Model (Blei et al., 2003)

2.1. LDA(잠재 디리클레 할당, Latent Dirichlet Allocation)

토픽 모델링 기법의 하나인 LDA(Latent Dirichlet Allocation, 잠재 디리클레 할당)은 대량의 문서에서 잠재적인 주제를 찾고, 도출된 주제에 관한 확률 분포로 각각의 문서를 설명하는 확률론적 토픽 모델링 방법이다(Blei et al. 2003).

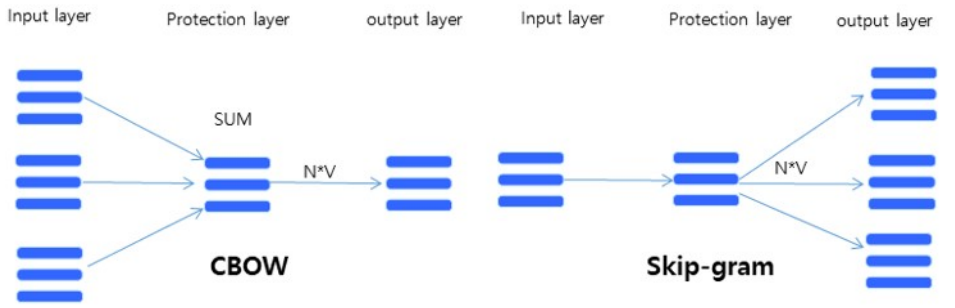
LDA 모델 요소는 <Figure 1>에 나타내었다. 여기서 K는 전체 토픽 수, D는 전체 문서의 수, N은 문서 d의 전체 단어 수를 의미한다. LDA 모델 요소인 β_k 는 k번째 토픽을 구성하는 요소로, 말뭉치의 비중 벡터이다. d번째 문서를 구성하는 토픽의 비중벡터인 Θ_d 와 k번째 토픽을 구성하는 말뭉치의 비중 벡터인 β_k 가 주어질 때, 토픽별 단어 분포에 대한 확률 모형을 선택한다(Blei et al., 2003).

LDA는 단어-주제-문서로 이어지는 계층적 관계라고 정의하고, 주변 문맥 단어에서 대상 단어를 예측하는 단어 임베딩 모델 중 하나인 Word2vec을 사용하여 보다 포괄적인 문서 표현 방법을 제안하였다(Wang et al., 2016). Heo and Yang(2021)은 DBpia에서 '코로나 19(covid 19)'를 키워드로 검색하여 총 543개의 논문을 대상으로 토픽 모델링 기법인 LDA와 감성 분석 기법을 통해 COVID-19 관련 연구 논문의 경제적 영향, 생물 의학 관련, 사회적 보호 및 복지 등 연구 주제를 탐색하였다.

본 연구에서는 COVID-19와 관련되어 연구되고 있는 주제를 살펴보기 위해 잠재 디리클레 할당(LDA) 알고리즘을 사용하여 문서 내 존재하는 단어를 대상으로 잠재적인 주제를 분석하여 문서의 내용이 어떤 토픽을 가졌는지 확인하였다.

2.2. Word2vec

본 연구에서는 COVID-19과 관련된 전문을 가진 41,062편의 연구 논문을 대상으로 토픽 모델링 기법의 하나인 LDA 알고리즘을 사용하여 연구 주제를 분석하고, Word2vec 알고리즘을 사용하여 입력한 키워드와 유사한 단어를 예측하였다. Word2vec 모델의 구조는 <Figure 2>과 같이 입력층 벡터에 여러 개의 단어를 사용하고, 하나의 단어와 비교하는 방법인 Continuous Bag-Of-Words(CBOW) 모델과 입력층 벡터에 하나의 단어를 사용하고, 주변 다른 단어와 비교하



〈Figure 2〉 Word2vec Model

는 방법인 Skip-gram 모델로 나뉜다. 본 연구에서는 단어 사이의 유사도를 측정하고, 복잡한 특징도 잘 파악할 수 있는 Word2vec 모델을 사용하여 키워드 사이의 유사도를 측정하고 예측하였다.

2.3. PCA 차원 축소 및 t-SNE

차원 축소 알고리즘은 데이터의 품질을 향상하기 위해 데이터의 복잡성을 줄여(Anowar et al., 2021) 더 직관적으로 데이터를 분석하는 방법이다. t-SNE는 데이터 탐색과 시각화에 주로 사용되는데, 단어의 위치와 거리를 시각화하고, 그룹으로 분류하여 분석할 때 많이 사용된다. 본 연구에서는 PCA 차원 축소 후, LDA 토픽 모델링 결과 도출되었던 토픽을 대상으로 레이블을 부여하고, t-SNE 알고리즘을 사용하여 유사한 문서를 그룹화하고 그룹이 형성하고 있는 것을 시각화하였다. PCA 차원 감소를 통한 군집화 방법을 적용하여 t-SNE으로 시각화하여 유사한 주제를 가진 논문의 그룹을 그룹화하였다. 대표적인 차원 축소 방법인 주성분 분석(Pricipal Component Analysis, PCA)은 선형 특성을 가진 데이터를 분석하기 위해 사용되는 비지도 학습 방법으로 여러 변수가 가지는 상관관계를 이용하여 그들을 대표하는 주성분을 추출해 차원을 축소하는 기법이다(Kwon, 2020). t-SNE(t-분포 확률적 임베딩, t-distributed Stochastic Neighbor Embedding)은 복잡하고 차원이 높은 데이터를 2차원으로 축소하는 방법으로, 주로 비선형 특성을 가진 데이터를 분석하기 위해 사용된다(Anwar et al., 2020:Buljan and Nordqvist, 2020). t-SNE 알고리즘은 기존에 가지고 있던 데이터의 중요한 구조는 유지하면서 차원이 높은 데이터에서 차원이 낮은 데이터로 매핑되는 비선형 특성을 가진 데이터를 분석하는데 사용되는 방법이다(Maaten et al., 2008). 차원 축소 방법은 연구자에게 새로운 통찰력을 제공하고, 데이터의 구조를 잘 이해하도록 도움을 준다. 본 연구에서는 PCA 차원 축소 후, LDA 토픽 모델링 결과 도출되었던 토픽을 대상으로 레이블을 부여하고, t-SNE 알고리즘을 사용하여 유사한 문서를 그룹화하고 그룹이 형성하고 있는 것을 시각화하였다.

2.4. 기존 연구와 본 연구의 차별성

기존의 선행연구를 정리하면, 대량의 문헌 전체를 대상으로 주제를 도출하고, 논문을 매핑하여 시각화(Eren et al., 2020)하거나, COVID 19

와 관련된 논문을 대상으로 비즈니스 및 관리 분야의 영역 연구 결과를 제시하였다(Verma and Gustafsson, 2020). 10,863개의 초록을 사용하여 연구자가 지정한 세 가지 주제 '전염(transmission)', '약물 유형(drug type)', '게놈(genome)'의 핵심 정보를 찾는 그래프 기반 모델을 연구(Ahamed et al., 2020)하거나, LDA와 감성 기법을 사용하여 COVID-19 관련 연구 논문의 경제적 영향, 생물 의학 관련, 사회적 보호 및 복지 등 연구주제를 탐색하는 연구가 진행되었다. 기존의 연구는 대량의 문서 전체를 대상으로 연구하거나, 특정 한 분야만을 연구하여 대량의 문헌에서 연구자가 필요로 하는 유용한 정보를 제공하는 방안은 제시하지 못하였다는 한계점을 가지고 있다.

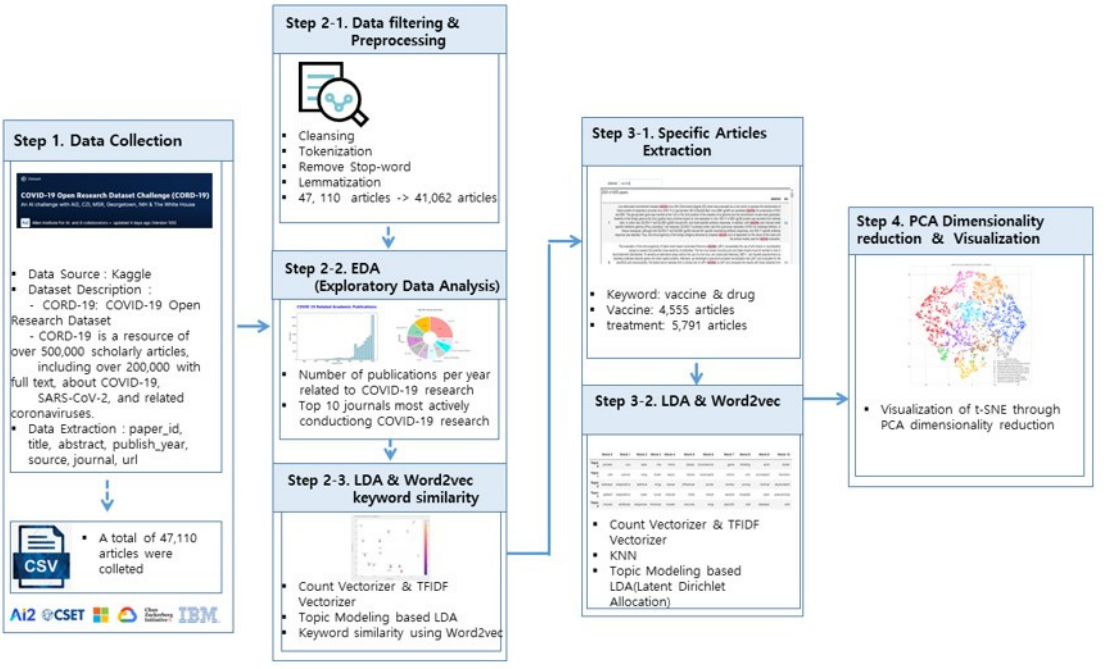
본 연구는 COVID-19와 관련된 전문(full-text)을 가진 41,062편의 연구 논문을 전체를 대상으로 토픽 모델링 기법을 활용하여 연구 주제를 분류하고 분석하였다. 대량의 문헌에서 연구자가 탐색하고자 하는 키워드를 입력하여 주제별로 문서를 분류하고, 이를 기반으로 유용한 정보를 추출하고, 군집화하여 시각화하는 방법을 제시하였다는 것이다. 본 연구의 차별성은COVID-19와 관련된 전체 논문을 대상으로 도출된 주제 중 'vaccine'과 'treatment'를 대상으로 연구되고 있는 주제가 무엇인지 탐색하여 연구 주제별 분류 및 분석 방안을 제시하였다는 것이다. 본 연구의 결과가 의료 전문가와 정책 담당자들의 소중한 시간과 노력을 줄이고, 특정 정보를 신속하게 얻을 수 있도록 도움을 줄 수 있기를 기대한다. 또한 연구자들에게 새로운 통찰력을 제공하고, 새로운 연구 방향을 탐색하는 데 기초 자료로 활용될 것으로 기대한다.

3. 연구 방법

3.1. 분석 절차

본 연구는 COVID-19 연구 논문을 대상으로 토픽 모델링을 활용하여 연구되고 있는 주요 토픽을 발견하고, 발견된 토픽 중 'vaccine' , 'treatment'를 포함하고 있는 특정 논문만을 추출하여 특정 주제에 대한 논문을 분류하고 분석하는 방법으로 진행되었다. 본 연구에 사용된 데이터 세트는 Kaggle에 COVID-19 전염병에 대응하기 위해 주요 연구 그룹과 백악관이 준비한 연구 데이터인 CORD-19 데이터 세트(COVID-19 Open Research Dataset, 이하 CORD-19)[1]를 활용하였다. CORD-19 데이터 세트는 COVID-19와 관련하여 연구된 무료 학술자료로서 매주 자료가 업데이트되고 있다. 2021년 10월 현재까지, COVID-19와 관련된 학술 논문이 500,000편 이상 발표되었다. 이 중 전문(full-text)을 가지고 있는 학술 논문 47,110편을 수집하여 데이터 필터링을 통해 총 41,062편의 논문을 수집하였다. 수집된 논문의 초록을 대상으로 파이썬을 이용하여 탐색적 데이터 분석(EDA, Explore Data Analysis)을 통해 연도별 COVID-19 관련 출판 수를 분석하였고, 활발하게 연구 중인 상위 10개의 저널을 확인하였다. 41,062편의 논문을 대상으로 텍스트에서 기호나 불필요한 문자를 제거하고, 초록에 있는 문장을 단어로 토큰화(Tokenization) 작업을 거친 후, 큰 의미가 없는 단어(Stop-words)를 제거하고, 원형 단어를 찾기

1) Data source: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge



〈Figure 3〉 Research Process

위해 Lemmatization 등의 텍스트 정규화 작업을 수행하였다. CountVetorizer와 TFIDF-Vectorizer 클래스를 이용하여 텍스트의 피처 벡터화 변환 후, LDA 알고리즘을 활용하여 현재 COVID-19과 관련되어 진행되고 있는 연구 주제를 도출하였다. Shon et al., (2020)은 COVID-19와 관련된 연구가 역학 조사, 예측 모델링, 임상 연구에서 진단, 치료제 및 백신 개발을 위한 연구가 증가하고 있다고 하였다. 본 연구에서는 연구 주제로 'vaccine'과 'treatment'를 키워드로 선정하여 대량의 문헌에서 탐색하고자 하는 키워드를 입력하여 특정 문헌을 추출하였다. 'vaccine'과 'treatment'을 포함하고 있는 논문을 추출하여 'vaccine'과 관련된 논문은 총 4,555편, 'treatment'와 관련된 논문은 총 5,971편을 수집하였다. 각각 수집된 논문을 대상으로 LDA 알고리즘을 활용하여 탐색하고자 하는 연구 주제를 추출하고 Word2vec 모델을 사용하여 유사어를 추출하였다. LDA 모델과 Word2vec 모델의 결합은 문서와 LDA 주제 사이의 관계, Word2vec 문맥 사이의 관계도 파악할 수 있어 더 좋은 성능을 나타낸다고 한다(Wang et al., 2016). PCA 차원 축소 후, LDA 토픽 모델링 결과 도출되었던 토픽을 대상으로 레이블을 부여하고, t-SNE 알고리즘을 적용하여 유사한 주제를 가진 논문의 그룹을 t-SNE으로 시각화하였다(<Figure 3> 참조).

3.2. Keyword Extraction Approach

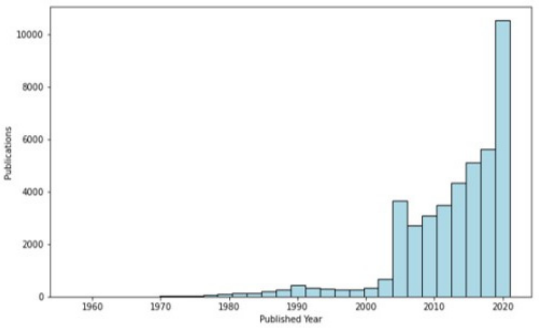
비지도 학습을 기반으로 하는 키워드 추출 접근법은 토픽 모델링, 통계, 그래프 기반으로 나눌 수 있다(Vasta et al., 2021). 본 연구에서는 파

이썬 대화형 인터페이스인 'ipywidgets'을 이용하여 논문의 초록에서 검색하고자 하는 키워드를 입력하여 방대한 양의 문서에서 특정 정보를 추출하는 방식을 제안한다. 토픽 모델링 결과 도출된 토픽 중 'vaccine'과 'treatment'와 관련된 논문을 검색하여 수집하였다.

4. 분석 결과

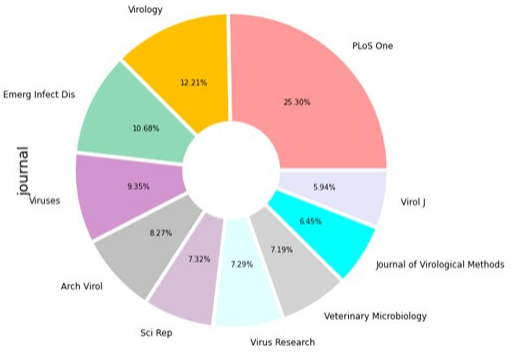
4.1. EDA(Explorer Data Analysis, 탐색적 데이터 분석)

데이터를 분석하여 인사이트를 얻기 위해 탐색적 데이터 분석(EDA) 과정을 수행하였다. 연도별 출판물 수와 학술지별 출판 현황을 살펴보았다. <Figure 4>는 COVID-19과 관련하여 연도별 출판된 학술 논문의 수를 그래프로 나타낸 것이다. COVID-19가 세계적으로 발생한 2020년에 출판물 수가 급격히 증가한 것을 확인할 수 있었다.



〈Figure 4〉 Number of Publications related COVID-19 by year

<Figure 5>은 COVID-19과 관련하여 가장 활발한 연구를 진행하고 있는 상위 10개의 학술지를 그래프로 나타낸 것이다. 1위는 과학과 의학의 연구를 주로 수행하는 PLos One이 25.3%이고, 세계 바이러스학 저널인 Virology가 12.21%로 2위를 차지했고, 의학 분야 저널인 Emerge Infect Dis가 10.86%로 그 뒤를 따르고 있다.



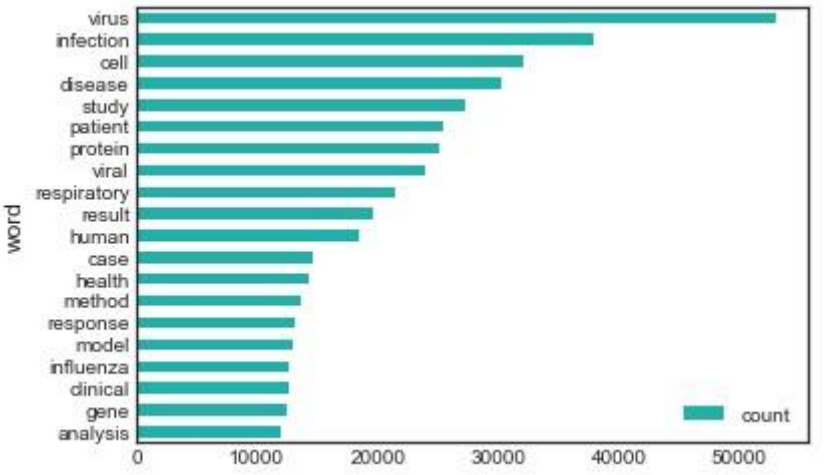
〈Figure 5〉 Publication status of research papers related to COVID-19 by academic journal

4.2. 빈도 분석

COVID-19와 관련된 연구 동향을 살펴보기 위해 초록을 대상으로 빈도 분석을 실시하였다. 전문을 가진 총 47,110건의 논문을 대상으로 파이썬 프로그램을 이용하여 텍스트 클렌징, 토큰화, 불용어 제거, 어근 추출 과정을 통해 수집된 41,062편의 논문의 초록을 대상으로 키워드 빈도 분석을 한 결과를 <Table 1>과 <Figure 6>에 나타내었다. 빈도 분석을 위해 파이썬의 사이킷런(sklearn)의 TFIDF-Vectorizer를 사용하여 문장에서 단순히 빈도수를 기준으로 키워드를 찾는 것이 아니라, 문장에서 단어의 빈도수와 자주 등장하는 단어에 가중치를 높게 부여하여 해당 키워

〈Table 1〉 Top 30 Keyword Frequency of COVID-19 articles

No.	Keyword	Frequency	No.	Keyword	Frequency	No.	Keyword	Frequency
1	virus	53117	11	human	18406	21	rna	11642
2	infection	37931	12	case	14678	22	vaccine	11289
3	cell	32128	13	health	13648	23	covid	11279
4	disease	30332	14	method	13109	24	control	10472
5	study	27304	15	respond	12865	25	abstract	10408
6	patient	25441	16	model	12583	26	group	10361
7	protein	25107	17	influenza	12529	27	coronavirus	10332
8	viral	23935	18	clinical	12365	28	sample	9645
9	respiratory	21497	19	gene	11984	29	pathogen	9638
10	result	19667	20	analysis	11984	30	infectious	9372



〈Figure 6〉 Top 30 Keyword Frequency of COVID-19 articles graph

드의 중요도를 고려해 상위 30개의 단어를 확인할 수 있었다.

4.3. LDA 토픽 모델링

토픽 모델링은 방대한 양의 문서를 구성하고 있는 주요 토픽을 도출하는 방법으로, 잠재 데이터를 발견하고, 데이터와 텍스트 문서 사이의 관계를 찾기 위해 가장 많이 사용되는 기법이다(Jelodar, H. et al., 2018). LDA 기반 토픽 모델링은 문서별 토픽 분포와 토픽별 단어 분포를 확률적으로 나타내는 확률 모형이다(Blei et al., 2003). 본 연구에서는 LDA 기반 토픽 모델링 기

〈Table 2〉 LDA Topic modeling analysis results

	Topic 01 (diagnostic test)	Topic 02 (aerosol transmission)	Topic 03 (rna virus)	Topic 04 (pandemic)	Topic 05 (healthcare)
1	pcr	air	rna	epidemic	health
2	detection	airway	gene	outbreak	care
3	positive	exposure	replication	spread	healthcare
4	diagnostic	aerosol	translation	global	hospital
5	test	ariborn	genome	pandemic	medical
	Topic 06 (zoonotic disease)	Topic 07 (pneumonia)	Topic 08 (confirmed cases and deaths)	Topic 09 (antiviral agent)	Topic 10 (mutant)
1	animal	lung	country	antiviral	hiv
2	disease	pulmonary	rate	drug	mutant
3	pathogen	pneumonia	number	resistance	mutation
4	zoonotic	inflammation	death	antimicrobial	protein
5	cause	inflammatory	infected	agent	amino
	Topic 11 (COVID-19 envelop protein)	Topic 12 (interferon effect)	Topic 13 (patient care and clinical outcome)	Topic 14 (vaccine)	Topic 15 (protein interaction)
1	protein	ifn	patient	vaccine	response
2	particle	effect	care	antibody	mechanism
3	microscopy	induced	clinical	immune	function
4	virion	infected	outcome	antigen	pathway
5	enveloped	replication	conclusion	vaccination	interaction
	Topic 16 (respiratory syndrom)	Topic 17 (gene sequence)	Topic 18 (treatment)	Topic 19 (alpha coronavirus)	Topic 20 (lamp extraction)
1	sars	gene	research	pedv	lamp
2	mers	sequence	article	pig	extraction
3	ibv	genome	treatment	prrsv	efficiency
4	respiratory	genetic	drug	Tgev	method
5	syndrome	genotype	vaccine	rotavirus	test
	Topic 21 (antibody)	Topic 22 (analysis)	Topic 23 (zika virus)	Topic 24 (symptom)	Topic 25 (children respiratory infections)
1	protein	analysis	denv	exacerbation	respiratory
2	binding	approach	zika	asthma	influenza
3	peptide	application	flavivirus	symptom	child
4	receptor	technique	pregnancy	result	rev
5	protease	technology	pregnant	effect	detected

법을 사용하여 COVID-19와 관련된 연구가 어떠한 주제를 가지고 있는지 분석하였다. <Table 2>에 LDA 분석 결과를 나타낸 것으로, 토픽에 포함된 상위 키워드 5개와 각 토픽에 따른 주제를 나타내었다. 토픽 모델링 결과를 분석할 때, 연구자는 도출된 키워드의 유사도가 높은 키워드를 파악하여 토픽 명을 선정해야 한다. 예를 들어, Topic 01은 pcr(polymerase chain reaction, 중합 효소 연쇄반응), detection(검출), positive(양성), diagnostic(진단), test(검사)라는 키워드를 바탕으로 '진단 검사(diagnostic test)'라는 토픽 명을, Topic 14는 vaccine(백신), antibody(항체), immune(면역체계), antigen(항원), vaccination(예방접종)이라는 키워드를 바탕으로 '백신(vaccine)'이라는 토픽 명을 선정해주었다. 같은 방법으로 도출된 25개의 토픽에 토픽명을 선정해주었다. LDA 기반 토픽 모델링 결과, COVID-19와 관련하여 연구되고 있는 토픽들은 Topic 01은 '진단 검사(diagnostic test)', Topic 02는 '에어로졸 전파(aerosol transmission)', Topic 03은 'rna 바이러스(rna virus)', Topic 04는 '팬데믹(pandemic)', Topic 05는 '헬스케어(healthcare)', Topic 06은 '인수공통감염병(zoonotic disease)', Topic 07은 '폐렴(pneumonia)', Topic 08은 '확진자 수와 사망자 수(confirmed cases & death)', Topic 09는 '항바이러스제(antiviral agent)', Topic 10은 '돌연변이(mutant)', Topic 11은 '외피 단백질(envelope protein)', Topic 12는 '인터페론 효과(interferon effect)', Topic 13은 '환자 치료와 임상 결과(patien care & clinical outcome)', Topic 14는 '백신(vaccine)', Topic 15는 '단백질 상호작용(protein interaction)', Topic 16은 '호흡기 증후군(respiratory syndrom)', Topic 17은 '유전자 염기서열(gene sequence)', Topic 18은 '치료(treatment)', Topic 19는 '알파 코로나바이러스(alpha coronavirus)', Topic 20은 '램프 추출(lamp extraction)', Topic 21은 '항체(antibody)', Topic 22는 '분석(analysis)', Topic 23은 '지카 바이러스 (zika virus)', Topic 24는 '증상(symptom)', Topic 25는 '소아 호흡기 감염증(children respiratory infections)'으로 도출되었다. COVID-19로부터 벗어나기 위해 진단 검사, 백신, 치료제 연구, 돌연변이에 대한 다양한 연구가 진행되고 있다는 사실을 확인할 수 있었다. 실제로 항바이러스제로 개발 중인 약물을 COVID-19 치료제로 사용하기 위해 에볼라 바이러스 출혈열 치료제, C형 간염 치료제, B형 간염 치료제, 인간 면역 결핍 바이러스(AIDS) 치료제를 활용하는 연구도 보고되고 있다**(Lee, 2020)**.

COVID-19를 치료하기 위한 치료제가 없는 상황에서 새롭고 효과적인 치료제를 개발하는 일이 시급한 이 시점에, 도출된 토픽 중 '백신(vaccine)'과 '치료(treatment)'을 포함하고 있는 논문을 각각 추출하여 LDA 알고리즘과 Word2vec 알고리즘을 활용하여 문헌을 분류하는 연구를 진행하였다.

4.4. 논문 전체를 대상으로 한 Word2vec 연관어 분석

본 연구는 COVID-19와 관련된 학술 논문 47,110편을 대상으로 토픽 모델링 기법을 활용하여 도출된 토픽들을 가지고, 파이썬의 genism 패키지의 Word2vec 라이브러리를 이용하여 중심 단어로 주변 단어를 예측하는 Skip-gram 방식을 적용하여 토픽별 단어 사이의 유사도를 측정하였다. Wang et al.(2016)은 LDA기반 토픽모델링과 Word2vec을 결합하면 LDA의 주제와 문서 사이의 관계 뿐만 아니라, Word2vec의 문맥 사이

〈Table 3〉 Word2vec keyword similarity

No.	word	similarity	No.	word	similarity
1	cov	0.7789	6	syndrome-coronavirus	0.7012
2	coronaviruses	0.7637	7	syndrome coronavirus	0.6896
3	corona	0.7294	8	middle east	0.6840
4	coronavirus2	0.7288	9	syndrome cov	0.6812
5	sars coronavirus	0.7041	10	mers coronavirus	0.6807

의 관계도 파악할 수 있어 더 나은 성능을 보인다고 하였다. Xia et al.(2016)은 말뭉치 기반 접근 방식은 자연어 처리의 기본 문제를 극복하여 정확도를 향상시켰지만, 텍스트의 형식과 길이에 따라 예측 오류를 발생시키기도 한다고 주장하였다. 단어 임베딩(word embedding)은 단어를 실수 벡터로 표현하는 방법으로, 단어나 문장에 숨겨진 의미와 정보를 파악하고, 단어와 문장 사이의 유사도를 확인하는데 사용되기도 한다(Kim et al., 2021). 단어 임베딩(word embedding) 모델 중 하나인 Word2vec은 예측 기반 신경망 언어 모델로 주변 단어들을 분석하여 핵심 단어들을 예측한다. 문맥적 의미를 반영하여 생성된 어휘들은 비슷한 의미를 가지는 단어들이 같은 공간에 위치하게 되고(Jeong et al., 2018), 각 단어의 의미를 고려한 특성벡터를 만들어 유사한 단어를 사용한 문서를 보다 정확하게 찾을 수 있고, 검색 성능이 향상되는 결과를 가져올 수도 있다고 한다(Kim et al., 2016).

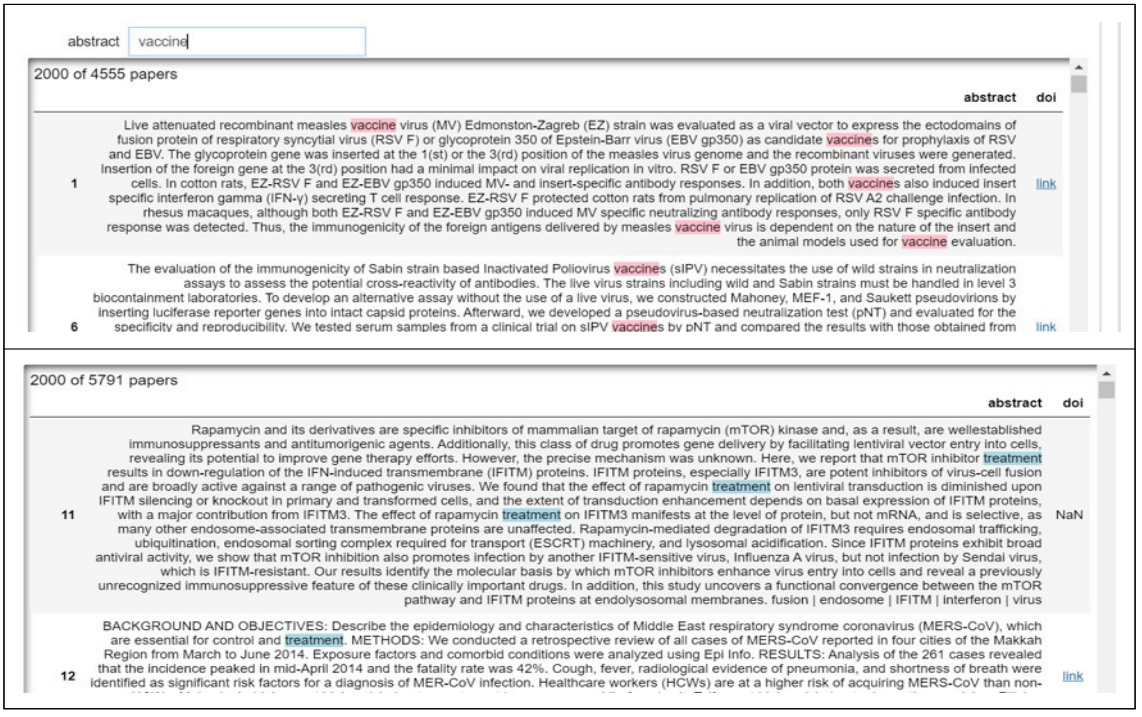
본 연구에서는 Word2vec과 LDA 기반 토픽 모델링 기법을 활용하여 주제와 문서의 관계와 단어 사이의 관계를 문서 벡터에 적용하였다. <Table 3> 은 각 토픽별 중심단어에 대한 유사도를 나타낸 것으로, 토픽 중 ‘coronavirus’를 선정하여 이와 유사한 단어 10개를 추출하고, 유사도 결과를 표로 나타낸 것이다. 본 연구에서는 LDA 모델과 Word2vec 모델을 결합하여 숨겨진 의미를 찾아 주요 키워드는 무엇이고, 어떠한 주제에 관심이 있는지를 파악하였고, 핵심 단어의 유사도가 0.6807이상으로 더 나은 성능을 보이는 것을 확인할 수 있었다.

4.5. 특정 논문 추출

파이썬의 대화형 인터페이스인 ipywidgets을 활용하여 COVID-19와 관련된 연구논문 41,062편을 대상으로, 입력창에 ‘vaccine’이라는 키워드를 입력하여 초록에서 ‘vaccine’ 단어를 포함하고 있는 논문을 검색하여 총 4,555편의 논문을 수집하였다. 같은 방식으로 입력창에 ‘treatment’라는 키워드를 입력하여 초록에서 ‘treatment’를 포함하고 있는 논문을 검색하여 총 5,791편의 논문을 수집하였다(<Figure 7> 참조). 대량의 문서에서 특정 정보를 포함하고 있는 논문을 추출하여 문헌을 분류하는 것은 연구자들에게 새로운 통찰력을 제공할 것이라고 기대한다.

4.6. 연구 주제별 토픽 모델링 결과

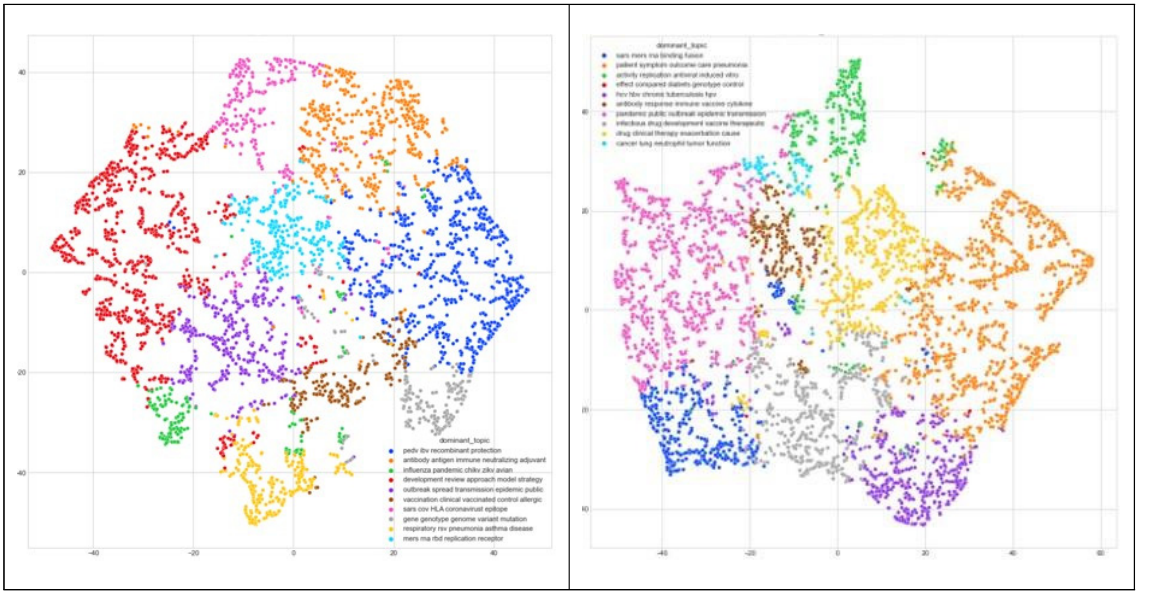
수집된 논문을 대상으로 vaccine 관련 4,555편, treatment 관련 5,791편을 대상으로 불용어(stop



〈Figure 7〉 Top 30 Keyword Frequency of COVID-19 articles graph

〈Table 4〉 Topic modeling analysis results by research topic

Research Topic 01: vaccine					
	Topic 01 (influenza)	Topic 02 (respiratory disease)	Topic 03 (recombinant dna vaccine)	Topic 04 (COVID-19 variants)	Topic 05 (neutralizing antibodies)
1	influenza	respiratory	pedv	gene	antibody
2	pandemic	rsv	rbv	genome	antigen
3	avian	pneumonia	recombinant	genotype	immune
4	zivk	asthma	protection	variants	neutralizing
5	chikv	disease	response	mutant	adjuvant
Research Topic 02: treatment					
	Topic 01 (respiratory virus)	Topic 02 (cancer)	Topic 03 (drug)	Topic 04 (patient symptom)	Topic 05 (cytokine)
1	sars	cancer	drug	patient	antibody
2	mers	lung	clinical	symptom	response
3	rna	neutrophil	therapy	outcome	immune
4	binding	tumor	exacerbation	care	vaccine
5	fusion	function	cause	pneumonia	cytokine



word) 제거, 토큰화(tokenization), lemmatization과 같은 전처리 과정을 통해 'vaccine' 관련 4,448편, 'treatment' 관련 5,721편을 수집하였다. <Table 4>는 연구 주제별 LDA 토픽 모델링 결과를 표로 나타낸 것이다. vaccine과 관련된 논문을 대상으로 토픽을 추출한 결과, Topic 05 '중화 항체(neutralizing antibodies)'라는 새로운 토픽이 도출된 것을 확인할 수 있었다. 중화항체는 바이러스가 몸에 침투했을 때, 세포가 감염되는 것을 방어해주는 항체로, 치료제 생산과 백신 개발에 중요한 역할을 한다(Shin, 2020). 'treatment'와 관련된 논문을 대상으로 토픽을 추출한 결과, Topic 05 '사이토카인(cytokine)'이라는 새로운 토픽을 발견할 수 있었다. 사이토카인 폭풍이란 우리 몸의 면역 세포가 공격을 방어하지 않고, 정상세포를 공격하는 것이다(Jo, 2021). COVID-19에 치명적인 사이토카인에 대한 연구도 진행되고 있음을 파악할 수 있었다. 전체 논문을 대상으로 찾을 수 없었던 숨겨진 주제를 키워드에 따라 문헌을 분류하여 토픽 모델링을 수행한 결과 세부 주제를 찾을 수 있었다.

4.7. 연구 주제별 PCA 차원 축소 후 t-SNE 문서 군집화

연구 주제별로 PCA 차원 축소를 수행하였다. 토픽 모델링 결과 도출된 토픽을 주요 주제로 선정하여 레이블을 정해주고, t-SNE 알고리즘을 사용하여 단어의 위치와 거리를 시각화하였다. <Figure 8>은 비슷한 주제를 가지고 있는 문헌을 분류하여 그룹으로 형성하고 있는 것을 산점도로 나타내었다. 선정한 주제에 맞게 주제별로 문헌이 군집화되어 있는 것을 파악할 수 있었다.

5. 결론

본 연구는 급속하게 증가하고 있는 COVID-19와 관련된 대량의 문서에서 LDA와 Word2vec 알

고리즘을 활용하여 주요 토픽을 도출하고, 도출된 주제별 문서를 추출하는 방법을 제안하며, 문서를 주제별로 군집화하여 분류하고 시각화하는 방법을 제시하였다. 연구에 사용된 데이터는 Kaggle에 있는 COVID-19 전염병에 대응하기 위해 주요 연구 그룹과 백악관이 준비한 연구데이터인 CORD-19 데이터 세트(COVID-19 Open Research Dataset, 이하 CORD-19)를 활용하였다. 전문(full-text)을 포함하고 있는 학술 논문 47,110편을 분석에 사용하였다. 본 연구 절차는 크게 두 가지로 나눌 수 있다. 먼저, 학술 논문 47,110편을 대상으로 데이터 필터링(filtering)과 전처리 과정을 통해 41,062편을 수집하였다. 41,062편의 논문을 대상으로 파이썬 프로그램을 이용하여 탐색적 데이터 분석(EDA, Explore Data Analysis)을 통해 연도별 COVID-19 관련 출판 수를 분석하였고, 활발하게 연구 중인 상위 10개의 저널을 확인하였다. Count-Vetorizer와 TFIDF-Vectorizer 클래스를 이용하여 텍스트의 피처 벡터화 변환 후, LDA 알고리즘을 활용하여 현재 COVID-19과 관련되어 진행되고 있는 연구 주제를 도출하였다. 파이썬의 gensim패키지의 Word2vec 라이브러리를 이용하여 중심 단어로 주변 단어를 예측하는 Skip-gram 방식을 적용하여 유사도를 측정하였다. 두 번째로, 도출된 주제 중 ‘vaccine’과 ‘treatment’을 포함하고 있는 논문을 추출하여 ‘vaccine’과 관련된 논문은 총 4,555건, ‘treatment’와 관련된 논문은 총 5,971건을 수집하였다. 각각 수집된 논문을 대상으로 LDA와 Word2vec 알고리즘을 활용하여 세부 주제를 분석하고, PCA 차원 감소를 통한 군집화 방법을 적용하여 유사한 주제를 가진 논문의 그룹을 t-SNE으로 시각화하였다.

본 연구의 결과에서 주목할 만한 점은 COVID-19와 관련하여 연구되고 있는 전체 논문을 대상으로 도출된 토픽(<Table 2> 참조)에서는 도출되지 않았던 토픽들이 연구 주제별 토픽 모델링 결과(<Table 4> 참조)에서 도출된 것을 확인할 수 있었다. 예를 들면, ‘vaccine’과 관련된 논문을 대상으로 한 토픽 모델링 결과, Topic 05 ‘중화 항체(neutralizing antibodies)’라는 새로운 토픽이 분석되었는데, 중화항체는 바이러스가 몸에 침투했을 때, 세포가 감염되는 것을 방어해주는 항체로, 치료제 생산과 백신 개발에 중요한 역할을 한다고 한다(Shin, 2020). 또한, ‘treatment’와 관련된 논문을 대상으로 토픽을 추출한 결과, Topic 05 ‘사이토카인(cytokine)’이라는 새로운 토픽을 발견할 수 있었다. 사이토카인 폭풍이란 우리 몸의 면역 세포가 공격을 방어하지 않고, 정상세포를 공격하는 것이다(Jo, 2021). 전체 논문을 대상으로 찾을 수 없었던 숨겨진 주제를 키워드에 따라 문헌을 분류하여 토픽 모델링을 수행한 결과 세부 주제를 찾을 수 있었다.

본 연구의 실무적 시사점은 대량의 문서에서 탐색하고자 하는 키워드를 입력하여 논문에 대한 특정 정보를 추출하는 방법을 제안하였다는 점이다. COVID-19를 극복하기 위해 많은 연구자들의 노력이 급증하는 COVID-19와 관련된 학술 논문 출판 속도를 따라잡을 수 없는 상황에 의료 전문가와 정책 담당자들의 소중한 시간과 노력을 줄이고, 신속하게 새로운 통찰력을 얻을 수 있도록 도움을 줄 것으로 기대한다. 또한 연구자들이 새로운 연구방향을 탐색하는 데 기초자료로 활용될 것으로 기대한다.

본 연구의 학술적 의의는 LDA 알고리즘을 사용하여 대량의 문헌에서 주제를 추출하고, Word2vec의 모델 중 중심 단어로 주변 단어를 예측하는 Skip-gram 방식을 활용하여 유사어를 추출하는

방식을 제안하였다. LDA 모델과 Word2vec 모델의 결합은 문서와 LDA 주제 사이의 관계와 Word2vec 문서 사이의 관계도 파악하여 더 좋은 성능을 나타내기 위해 노력하였다. 또한, PCA 차원 축소를 통한 군집화 방법으로 문서의 구조화된 조직을 t-SNE 기법을 사용하여 비슷한 주제를 가지고 있는 문헌을 분류하고 그룹으로 형성하고 있는 것을 시각화하여 문헌을 직관적으로 분류하는 방안을 제시하였다는 점이다.

본 연구에서는 Kaggle에서 제공하는 데이터 세트를 사용하였지만, 향후 연구에서는 다른 학술 연구 검색사이트에서 COVID-19와 관련된 연구 논문을 수집하여 국내 연구도 분석해야 할 것이다. 또한, LDA 모델을 사용하여 토픽을 분석하였는데, 향후 연구에서는 토픽 모델링의 다른 모델인 CTM, STM 모델을 사용하여 성능을 비교하고, 최적의 모델을 무엇인지 비교 분석해 볼 필요가 있을 것이다.

참고문헌(References)

Ahamed, S. and M. D. Samad, *"Information mining for COVID-19 research from a large volume of scientific literature,"* Cornell University, 2020. Available at https://arxiv.org/abs/2004.02085/ (Downloaded 07 November, 2021).

Alimadadi, A., S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight COVID," *Physiol Genomics,* Vol. 52, No. 4(2020), 200~202.

Anowar, F., S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms(PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Computer Science Review,* Vol. 40(2021), 100378.

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

Buljan, M., J. Nordqvist, and R. M. Martins, "An Investigation on the Impact of Non-Uniform Random Sampling Techniques for t-SNE", 2020 Swedish Workshop on Data Science (SweDS), (2020), 1~8.

Chu, Y. J., *Research papers pandemic brought by COVID-19,* Medical Observer, 2021, Available at http://www.monews.co.kr/news/articleView.html?idxno=302210 (Downloaded 12 Nov, 2021)

Eren, M. E., N. Solovyev, E. Raff, C. Nicholas, and B. Johnson, "COVID-19 Kaggle Literature Organization," *Proceedings of the ACM Symposium on Document Engineering,* 2020.

Heo, S. M. and J. Y. Yang, "A Convergence Study on the Topic and Sentiment of COVID 19 Research in Korea Using Text Analysis," *Journal of the Korea Convergence Society,* Vol.12, No. 4(2021), 31~42.

Jelodar, H., Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, *"Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey,"* Cornell Univerity, 2018. Available at https://arxiv.org/abs/1711.04305/(Downloaded 11 November, 2021)

Jo, S. W., *Corona 19 fact check ㊻ fatal 'cytokine storm',* hidoc, 2021. Available at https://www.hidoc.co.kr/healthstory/news/C0000595237/ (Downloaded 11 November, 2021).

Jeong, J. Y., K. H. Mo, S. W. Seo, C. Y. Kim, H. D. Kim, and P. S. Kang, "Unsupervised

Document Milti-Category Weight Extraction based on Word Embedding and Word Network Analysis : A Case Study on Mobile Phone Reviews," Journal of the Korean Institute of Industrial Engineers, Vol. 44, No. 6(2018), 442~451.

Kim, W. J., D. H. Kim, and H. W. Jang, "Semantic extention search for documents using the Word2vec," Journal of the Korea Contents Association, Vol. 16, No. 10(2016), 687~692

Kim, T. K., W. S. Shon, and S. M. Jeon, "Mining Loot Box News: Analysis of Keyword Similarities Using Word2Vec," Journal of Information Technology Service, Vol. 20, No. 2(2021), 77~90.

Kwon, C. M., *Python MachineLearning Perfect Guide,* Wikibooks, Seoul, Korea, 2020.

Lee, Z. A., *New Coronavirus without a cure, hepatitis C, Ebola, and AIDS treatments are on the rise,* Dong-A Science, 2020. Available at https://www.dongascience.com/news.php?idx=34026/ (Downloaded 11 November, 2021).

Liu, M. N. and G.G.Lim, "Word-of Mouth Effect for Online Sales of K-Beauty Products: Centered on China SINA Weibo and Meipai," *Journal of Intelligence and Information System,* Vol.25, No. 1(2019), 197-218.

Maaten, L.v.d., G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol.9(2008), 2579~2625.

Shin, E. J., "Recent Academic Publishing Trends through Bibliometric Analysis of COVID-19 Articles: Focused on Medicine and Life Science," *Korean Biblia Society for Library and Information science,* Vol. 32, No. 1(2021), 115~132.

Shin, Y. S., New technology to quickly identify neutralizing antibodies against COVID-19, PharmNews, 2020, Available at https://www.pharmnews.com/news/articleView.html?idxno=100977/ (Downloaded 11 November, 2021).

Shin, Y. S., COVID-19 fact check fata 'cytokine storm', Hidoc, Available at https://www.hidoc.co.kr/healthstory/news/C0000595237 (Downloaded 12 Nov, 2021)

Shon, E.S., S. J. Ahn, T. H. Ha, and B. Y. Coh, *COVID-19 research trends seen through archive data,* Korea Institute of science Technology Information, Available at http://mirian.kisti.re.kr/insight/insight.jsp (Downloaded 11 Nov, 2021).

Vatsa, S., S. Marthur, M. Garg, and R. Jimdal, "COVID-19 Tweet Analysis using Hybrid Keyword Extraction Approach," 2021 10th IEEE International Conference on Communication Systems and Network Technologies(CSNT), 2021, 136~140.

Verma, S. and A. Gustafsson, "Investigating the emerging COVID-19 research trends in the field of business and management: *A bibliometric analysis approach, Journal of Business Research,"* Vol. 118(2020), 253~261.

Wang, Z., L. Ma, and Y. Zhang, "A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2vec," *2016 IEEE First International Conference on Dat Science in Cyberspace(DSC),* (2016), 98~103.

Xia, C., T. He. W., Li, Z. Qin, and Z., Zou, "Similarity Analysis of Law Documents Based on Word2vec," 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion(QRS-C), (2019),

345~357.

Yoo, S.Y., and G.G.Lim, "Ananlysis of News Agenda Using Text mining and Semantic Network Analysis: Focused on COVID-19 Emotions," *Journal of Intelligence and Information System,* Vol.27, No. 1(2021), 47-64.

Yoon, J.E., and C.J.Suh, "Research Trend Analysis by using Text-Mining Techniques on the Convergence Studies of AI and Healthcare Technologies," *Journal of Information Technology Services,* Vol.18, No. 2(2019), 123-141.

Yu, D.S., and G.G.Lim, "A Study on the eWOM and Selecting Movie According to Online Media and Replies," *Journal of Information Technology Services,* Vol.14, No. 2(2015), 177-193.

Abstract

A study on the classification of research topics based on COVID-19 academic research using Topic modeling

So-yeon Yoo* · Gyoo-gun Lim**

From January 2020 to October 2021, more than 500,000 academic studies related to COVID-19 (Coronavirus-2, a fatal respiratory syndrome) have been published. The rapid increase in the number of papers related to COVID-19 is putting time and technical constraints on healthcare professionals and policy makers to quickly find important research. Therefore, in this study, we propose a method of extracting useful information from text data of extensive literature using LDA and Word2vec algorithm. Papers related to keywords to be searched were extracted from papers related to COVID-19, and detailed topics were identified. The data used the CORD-19 data set on Kaggle, a free academic resource prepared by major research groups and the White House to respond to the COVID-19 pandemic, updated weekly. The research methods are divided into two main categories. First, 41,062 articles were collected through data filtering and pre-processing of the abstracts of 47,110 academic papers including full text. For this purpose, the number of publications related to COVID-19 by year was analyzed through exploratory data analysis using a Python program, and the top 10 journals under active research were identified. LDA and Word2vec algorithm were used to derive research topics related to COVID-19, and after analyzing related words, similarity was measured. Second, papers containing 'vaccine' and 'treatment' were extracted from among the topics derived from all papers, and a total of 4,555 papers related to 'vaccine' and 5,971 papers related to 'treatment' were extracted. did For each collected paper, detailed topics were analyzed using LDA and Word2vec algorithms, and a clustering method through PCA dimension reduction was applied to visualize groups of papers with similar themes using the t-SNE algorithm. A noteworthy point from the results of this study is that the topics that were not derived from the topics derived for all papers being researched in relation to COVID-19 (<Table 2>) were the topic modeling results for each research topic (<Table 4>)

* School of Business, Hanyang University
** Corresponding author: Gyoo-gun Lim
School of Business, Hanyang University
220 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Korea
Tel: +82-2-2220-2593, E-mail: gglim@hanyang.ac.kr

was found to be derived from For example, as a result of topic modeling for papers related to ‘vaccine’, a new topic titled Topic 05 ‘neutralizing antibodies’ was extracted. A neutralizing antibody is an antibody that protects cells from infection when a virus enters the body, and is said to play an important role in the production of therapeutic agents and vaccine development. In addition, as a result of extracting topics from papers related to ‘treatment’, a new topic called Topic 05 ‘cytokine’ was discovered. A cytokine storm is when the immune cells of our body do not defend against attacks, but attack normal cells. Hidden topics that could not be found for the entire thesis were classified according to keywords, and topic modeling was performed to find detailed topics. In this study, we proposed a method of extracting topics from a large amount of literature using the LDA algorithm and extracting similar words using the Skip-gram method that predicts the similar words as the central word among the Word2vec models. The combination of the LDA model and the Word2vec model tried to show better performance by identifying the relationship between the document and the LDA subject and the relationship between the Word2vec document. In addition, as a clustering method through PCA dimension reduction, a method for intuitively classifying documents by using the t-SNE technique to classify documents with similar themes and forming groups into a structured organization of documents was presented. In a situation where the efforts of many researchers to overcome COVID-19 cannot keep up with the rapid publication of academic papers related to COVID-19, it will reduce the precious time and effort of healthcare professionals and policy makers, and rapidly gain new insights. We hope to help you get It is also expected to be used as basic data for researchers to explore new research directions.

Key Words : COVID-19, Topic Modeling, LDA(Latent Dirichlet Allocation), Word2vec, Keyword Extraction

Received : December 30, 2021 Revised : January 24, 2022 Accepted : February 1, 2022
Corresponding Author : Gyoo-gun Lim

저 자 소 개

유 소 연

한양대학교 경영대학 비즈니스 인포매틱스학과 박사과정에 재학중이다. 관심분야는 빅데이터 분석, 인공지능, 텍스트 마이닝 등이다.

임 규 건

한양대학교 경영대학 임규건 교수는 KAIST 전산학 학사, POSTECH 컴퓨터 석사, KAIST 경영공학 박사학위를 취득하였고, 삼성전자, KT, 국제전자상거래연구센터(ICEC) 연구위원, 세종대학교 경영학과 교수를 역임하였다. 관심분야는 혁신 비즈니스 모델, IT서비스 혁신, 인공지능과 경영, e-Business 등이며, 2018년 IT서비스 우수연구인상을, 2009년 IT Innovation 유공자 지식경제부 장관 표창과 2007년 SW산업발전 유공자 정통부 장관 표창을 수여하였다. 주요 저서로는 '경영을 위한 정보기술', 'e-비즈니스 경영', '디지털경제시대의 경영정보시스템' 등 전문서적과 다수의 논문과 특허가 있다. 또한, 아시아최초 상용인터넷인 KORNET 상용화, 중국 Shanghai Telecom SI사업전략, 한국영화기술 로드맵, KTI 사업전략, 나라장터 (G2B) 효과평가, 행정정보화(G4C) 성과분석, 국가정보보호지수개발, 국방정보화 수준평가모형, IT혁신인력양성종합대책, 국가디지털식별체계(UCI), 저작권정품 인증제도, SW사업자신고제도 개선, SW기술자신고제도개선 등 다양한 IT혁신 분야의 프로젝트를 수행하였다.