토픽모델링을 활용한
최근 5년 간 국내 문헌정보학의 연구동향 분석

데이터사이언스와 R 1조

2017314643 김현우

2018312990 조명재

2021313897 허지원

2021314373 안세연

2021314642 신민서

초록

본 연구는 최근 5년 간 국내 문헌정보학의 연구동향을 파악하고, 어떤 주제의 연구가 진행되었는지 확인하기 위해서 문헌정보학 분야의 주요 학술지인 한국문헌정보학회지, 한국비블리아학회지, 정보관리학회지의 최근 5년간 발간된 논문들을 수집하여 분석을 진행하였다. 이때, 논문 전체 내용에 기반한 정성적인 평가가 아닌, 논문 초록데이터를 바탕으로 토픽모델링을 진행하여 연구동향을 파악하는데 최대한 연구자의 주관을 배제하고자 하였다. 국회도서관에서 제공하는 국가학술정보 Open API를 통해 학술지에 게재된 논문 838건에 대한 한글 초록 데이터를 수집하였고, 이 중 결측치를 제외한 후 한글 초록데이터에 대해 형태소 분해를 실시하여 총 611건의 한글 초록데이터에 대하여 분석을 실시하였다. 분석 결과, 최적의 연구 주제 수는 6이 나왔고, 이를 바탕으로 LDA를 진행한 결과 도서관 체계와 도서관 경영에 해당하는 주제, 정보서비스 이용에 대한 주제, 기록관리학에 대한 주제, 도서관 서비스에 대한 주제, 자료조직과 디지털 도서관, 서지학에 대한 복합적인 주제, 주제분석 연구와 메타데이터에 대한 주제가 진행되고 있는 것을 확인하였다. 이 중 몇몇 논문들을 살펴본 결과, 대체로 논문들이 6개 중 하나의 주제에 대해서 가중치가 높게 나타나고 있지만, 일부에서는 주제에 대한 가중치가 고르게 나타났고, 이를 문헌정보학 분야가 가지는 융복합적인 성격으로 인한 것으로 파악하였다. 이후 연도별 문헌정보학 내 주제 변화를 분석하여 2020년을 기점으로 도서관과 도서관의 디지털화에 대한 연구가 활발히 진행되었고, 이는 코로나-19의 확산으로 인한 것으로 분석하였다. 이후 계층적 클러스터링을 통해 단어의 문서별 출현 빈도를 바탕으로 단어들의 계층적인 구조를 확인해보았고, 이를 LDA를 통한 토픽모델링 결과와 비교하여 토픽모델링 결과와 단어의 계층적인 구조와의 연관성을 확인하였다. 본 연구에서는 데이터셋과 분석 프로그램인 R로 인한 분석의 한계가 존재하였지만, 기존 선행연구보다 더욱 최신의 트렌드를 반영한 연구주제를 압축하여 제공하는 것으로 향후 문헌정보학 연구설계의 기초자료를 생성하는 연구목표를 달성하였고, 기존에 진행된 토픽모델링을 통한 연구현황분석과 달리 한글 텍스트데이터를 바탕으로 토픽모델링을 진행하여 국내 연구동향 분석에 새로운 방법론은 제시하였다는 의의를 가진다.

성균관대학교
SUNG KYUN KWAN UNIVERSITY

0. 목차

1. 서론

1.1. 연구배경

우리나라의 문헌정보학계는 정보환경의 급격한 변화와 이로 인한 혼란 속에서도 거듭한 성장을 해왔다. 이와 같은 발전이 가능했던 것은 1970년에 창립한 한국도서관학회 (현 한국문헌정보학회)를 시작으로 한국정보관리학회, 한국도서관·정보학회, 한국비블리아학회 등 문헌정보학 분야 주요 학회들이 설립되며 활발히 연구를 진행했기 때문이다. 국내 문헌정보학의 연구 활동이 활발해짐에 따라 연구자들은 문헌정보학이 어떤 주제를 중심으로 연구를 수행하고 있고, 문헌정보학자의 관심분야가 어떻게 변화하고 있으며 학문적 유형은 어떤 양상을 보이는지 등 그 현상을 밝히기 위해 다각적인 측면에서의 연구동향 분석을 시도하고 있다. (박자현,2013)

기존의 문헌 연구 방식은 인간이 갖는 정보 획득과 처리의 인지적 한계로 인해 논문을 탐색하고 수집하는 과정에서 오류와 편향성이 발생할 수 있으며(Lau et al. 2007), 이는 문헌 연구의 편향적 논리로 이어질 수 있었다. 또한 Logan et al(2010)의 연구에 따르면 저자 자신의 연구를 관련 연구로 추가할 때 과대평가하거나 과소평가하는 편향성 또한 무시할 수 없었다. 이러한 기존 문헌 연구 방식의 한계와 학술 데이터의 지속적인 축적으로 인해 인간이 전체 데이터의 내용을 분석하는 것은 어려워 정량적 분석으로 보완하고 있으며, 내용 분석, 양적 분석, 계량 서지학, 텍스트 마이닝 등의 기법이 활용되고 있다. 본 연구는 대표적인 텍스트 마이닝 기법인 LDA를 활용한 토픽 모델링을 통해 도출되는 토픽들을 활용해 근 5년간 문헌정보학 분야의 연구 현황을 분석하여 확인하고, 향후 문헌정보학 연구 설계의 기초 자료로 확인될 수 있는 자료를 만드는 것을 목표로 한다.

또한, 이전의 연구들에서는 영문 초록을 활용하여 토픽 모델링을 진행하였는데, 본 연구는 한글 초록을 한글 초록의 형태소를 분석해 키워드를 추출하고, 이를 바탕으로 토픽 모델링을 진행하여 연구동향을 분석하고자 한다. 이전에는 영문 초록에 대해서만 이러한 분석이 진행되었지만, 한글로만 작성된 논문에 대해서도 연구동향의 분석이 가능하게 되어, 앞서 언급한 문헌정보학의 학술적 가치를 더욱 높일 수 있을 것이다.

1.2. 선행연구

선행연구를 통해, 이미 문헌정보학의 연구동향분석에 텍스트마이닝 기법이 활발하게 적용되고 있는 것을 확인하였다. 박자현, 송민(2013)의 연구에서는 기존의 내용분석이 갖는 단점을 텍스트 마이닝 기법을 적용하여 보완해, 문헌정보학 분야 주요 학술지인 정보관리학회지, 한국문헌정보학회지, 한국도서관·정보학회지, 한국비블리아학회지의 1970년대부터 2012년도까지의 발표 논문 초록을 수집하여 LDA 기반의 토픽 모델링 실험을 수행한다. 연구 결과, 텍스트 마이닝 기법을 적용하여 연구동향을 분석한 결과가 기존의 연구자가 직접

내용을 판단하여 분석한 결과와 유사하게 도출되어, 연구동향 분석에 새로운 방법론을 제시하였다는 점에서 의의를 가진다.

이기헌, 정효정, 송민(2015)의 연구에서는 모델링, 개념/문헌 연구, 연구 협업 분석, 웹 데이터 분석 및 LDA 기법을 활용하여 1990년부터 2014년까지 문헌정보학 학술지에 게재된 논문을 대상으로 연구 주제와 연구 방법을 구분하여 학술지의 연구 누적 현황을 분석하였다. 위 연구에선 25년간 빈번히 짝을 이루는 연구 주제들과 연구 방법들의 군집과 2014년부터 5년간의 군집을 비교하여 개량 서지연구와 네트워크분석 방법이 저변을 넓히고 있으며 의료정보시스템, 이용자 인터페이스와 같은 연구 주제에 텍스트마이닝 연구 방법이 특화되고 있음을 확인했다. 이러한 연구 결과를 통해 선도 연구자들의 관심 연구 영역과 관점을 분석하여 확인하는 것이 향후 문헌정보학 발전을 위한 연구 설계의 기초 자료로 확인될 수 있음을 강조하였다.

진설아, 송민(2016)의 연구에서는 LDA 기반의 토픽 모델링을 통해 정보학 분야 학술지의 세부토픽을 추출하여 분석을 진행하였다. 이를 통해 산출된 학술지 별 토픽 분포와 토픽 간 유사도를 바탕으로 개별 학술지의 다양성과 응집성을 지수 산출 방실을 활용해 측정하였다. 측정한 다양성과 응집성을 모두 고려한 학제적 유형에 따라 학술지를 분류하고 각 유형의 대표 학술지를 선정하여 유형과 특성에 따라 다르게 나타나는 학술지 별 토픽 네트워크를 분석하였다. 본 연구는 지수 산출을 통한 학제성의 측정과 유형 분류, 시각화를 통한 세부주제의 표현이 학제적 현위치를 파악하고 앞으로의 발전 가능성을 예측하는 단서를 제공한다는 점에서 학술지의 학제성을 나타내는 지표로서 다양하게 활용될 수 있음을 보여주었다.

박준형, 오효정(2017)의 연구에서는 LDA 기법과 이를 변형한 HDP 토픽 모델링 기법을 활용하여 국내 기록 관리학 연구 동향을 분석하고 국내 기록관리학 관련 학술지 2종과 문헌정보학 관련 학술지 4종에서 1997년부터 2016년까지의 기록관리학 관련 논문 1,027건을 수집하고 전처리 과정을 거친 뒤, 두 방식을 모두 활용하여 토픽 모델링을 진행하였다. 연구 결과, LDA 기법은 전반적으로 해당 도메인을 대표하는 주요 키워드의 빈도수에 영향을 많이 받는다는 사실을 알게 되었고, 따라서 기록 관리학 내에서 공통으로 많이 다루고 있는 주제 키워드 분석에 유리하다는 것을 알 수 있었다.

또한 문헌정보학 분야의 학술정보 분석이 아닌 다른 분야에서도 텍스트마이닝 기법을 활용한 분석을 통해 유의미한 결과를 도출하고 있는 것을 확인할 수 있었다. 김태경, 최회련, 이홍철(2016)의 연구에서는 핀테크 관련 특허 데이터에 토픽 모델링 기법을 적용해 핀테크 세부 기술들을 추출해 정의하고, 그 세부 기술에 대해 유망 기술(Hot topic)과 쇠퇴 기술(Coldtopic)을 도출하였다. 추출된 핀테크 세부 기술에 할당된 상위 키워드의 동시발생매트릭스를 구축하고 시각화하여 각 토픽별 관계를 살펴보았다. 본 연구는 기존의 단순한 핀테크 기술 및 산업 동향 분석과 달리 핀테크 산업 기술에 대한 특허 데이터의 초록을 이용하여 텍스트마이닝 기법으로 분석을 수행하였다는 점과, 일반적으로 특허 분석에 주로 사용하는

전통적인 계량서지학적 방법론이 아닌 토픽모델링과 네트워크 분석을 수행하여 분석한 점에서 의의를 가진다

양명석, 이성희, 박근희, 최광남, 김태현(2021)의 연구에서는 국가과학기술지식정보서비스에서 제공하는 국가연구개발사업에 대한 과제정보를 대상으로 LDA 토픽모델링 기법을 적용하여 연구주제와 관련된 토픽들을 추출하고, 이러한 토픽을 활용하여 인공지능이라는 특정 분야와 관련한 국가연구개발사업에 대한 연구주제와 투자방향에 대하여 분석한다. 연구 결과, 인공지능 분야의 경우에는 인공지능 관련 기초 기술(딥러닝, 기계학습 등)에 관한 연구뿐만 아니라 인공 지능 기술을 활용한 의료, 소재, 스마트시티, 로봇 등 다양한 영역에 걸쳐 연구가 폭넓게 진행되고 있을 알 수 있었으며, 정부 차원에서 인공지능 기술을 활용한 다양한 분야에 투자를 지속적으로 확대해가고 있음을 알 수 있었다.

정혜리(2022)의 연구에서는 온라인 커뮤니티에 게시된 대학생들의 텍스트를 분석하여 대학의 온라인 교육에 관한 의견을 LDA 토픽모델링을 통해 주제화하여 분석하였다. 연구 결과 토픽은 각각 개강, 학점관리, 카메라 등 몇 개의 단어들로 추려졌으며, 공통적으로 "등록금"이라는 키워드가 발견되었다. 따라서 본 연구에서 도출된 주요 토픽들을 기초로 하여 앞으로 진행되어야 할 온라인 교육과 관련된 연구에 학술적 기여를 기대할 수 있었다.

1.3. 연구문제

본 연구를 통해 알아보고자 하는 문제들은 다음과 같다.

연구 문제 1: 최근 5개년 간 문헌정보학 분야에서 연구되고 있는 최적의 연구 주제는 몇개이며, 어떤 주제를 담고 있는지를 확인한다.

연구 문제 2 : 2017~2020년 문헌정보학 분야에서 해당 주제들의 비율은 어떻게 변화하였는지를 시각화한 플롯을 통해 확인한다.

2. 연구 설계

토픽모델링을 통하여 최근 5년간의 문헌정보학의 연구 동향을 파악하기 위해서 국가학술정보 Open API를 통해 초록데이터를 수집하고 이를 전처리를 하는 과정을 거쳤고, 이후 대표적인 토픽 모델링 기법인 LDA를 통하여 초록데이터에서 최적의 토픽을 추출하고, 이후 결과를 분석하는 방향으로 진행하고자 한다.

전체적인 데이터 분석에는 R을 사용하였다. API 스크래핑을 위하여 httr 패키지와 jsonlite 패키지를 사용하였고, 텍스트마이닝을 위한 Corpus(말뭉치)생성, Document-Term Maxtrix(문서-단어 행렬)로의 변환을 위해 tm 패키지를 사용하였다. 이 때 한국어로 되어있

는 초록을 형태소 단어로 분리하여 이 중 명사만을 추출하기 위해서 KoNLP와 형태소 분해를 위해 한국지능정보사회진흥원 빅데이터 센터에서 만든 NIADic를 사용하였다. 이후 진행하는 토픽 모델링에서 LDA(잠재 디레클레 할당)을 사용하기 위해서 topicmodels 패키지와 lda 패키지를 사용하였고, 이 과정에서 최적의 Topic 수 K를 찾기 위해 ldatuning 패키지를 추가로 사용하였다.

2.1. 논문 초록 데이터 수집 및 전처리

2.1.1. 논문 초록 데이터 수집 및 결측치처리

선행연구에서는 영문 초록 데이터를 각 문헌정보학 관련 학회지의 홈페이지에서 java를 통해 크롤링하는 방식으로 데이터를 수집하였지만, 본 연구에서는 선행연구와는 달리 국회도서관에서 제공하는 국가학술정보 Open API를 활용하였다.

현재 국가학술정보 Open API는 연구자 정보와 주제어 정보, 논문 통합검색과 논문 상세정보를 제공하고 있다. 모든 API는 POST방식으로 데이터를 호출받고 있으며, POST방식으로 호출받은 데이터를 바탕으로 검색 결과값을 Python의 dictionary와 유사하면서도 계층적으로 이루어져있는 JSON 형식으로 반환하게 된다.

논문의 초록 데이터를 수집하기 위해서, 본 연구에서는 API중 '통합검색' 과 '상세보기'를 사용하였다. 먼저 통합검색 API를 통해 문헌정보학 관련 학술지 이름(한국문헌정보학회, 정보관리학회, 한국도서관·정보학회, 한국비블리아학회)를 검색 쿼리로 하여 작성기간 기준 최근 5년인 2017년부터 2021년까지의 논문을 수집하였다. API를 호출하여 반환 받은 값 중에서 논문의 고유 번호인 LOD 아이디와 논문명, 발행년, 제목만을 수집하였고, 이 과정에서 검색이 되지 않은 한국 도서관·정보학회의 논문은 제외하였다.

다음으로 수집한 데이터를 바탕으로 상세보기를 통해 다시 논문의 제목, 논문명, 발행년, 제목, 그리고 논문 초록정보를 수집하였다. 이후 논문의 초록이 정상적으로 수집되지 않거나, 수집에는 성공하였으나 초록이 미등재되어 결측치로 나타난 값들을 제거해주는 과정을 진행하였다. 이 과정을 통해 전체 수집한 논문 838건 중 739건을 본 연구의 분석 대상으로 삼았다. 이를 정리하면 다음과 같이 표로 나타낼 수 있다.

학술지명	수집한 논문수	초록 결측치 처리 후 논문수	비고
한국문헌정보학회지	320	273	총 47건 제외 (API 오류로 인한 누락 1건, 초록 미등재 46건)

			2021년 발행된 논문의 초록데이터가 누락됨
정보관리학회지	225	224	총 1건 제외 (API 오류로 인한 누락 1건)
한국비블리아학회지	293	242	총 51건 제외 (초록 미등재 51건)
한국도서관·정보학회지	-	-	API에서 검색이 되지 않아 분석 대상에서 제외

<표 1> 학술지 별 수집한 논문 수와 학술지 별 논문 제외 사유

이를 통해 초록데이터를 수집하였고, 이후 한글 데이터 전처리 과정을 거쳤다.

2.1.2. 한글 초록데이터 추출 및 토큰화, 문서-단어 행렬 생성

2.1.1.에서 수집한 데이터를 바탕으로 초록데이터에서 한글로만 이루어진 초록 데이터만 수집하고, 이를 토큰화하여 말뭉치(Corpus)의 형태로 변환하는 과정을 거쳤다.

초록데이터에서 한글 초록데이터만을 얻어내기 위해서는 정규표현식(Regex)를 통해서 초록데이터에서 한글 초록을 제외한 모든 부분을 없애 주어야 한다. 여기서 정규표현식이란, 텍스트에서 특정한 규칙을 가진 문자열의 집합을 표현하는 형식 언어이다. 이를 통해서 텍스트에서 세세하게 한글 초록을 골라내지 않고, 정규표현식을 통해 단어, 혹은 언어가 나타나는 패턴을 텍스트에서 찾아내어 추출할 수 있게 한다. 본 연구에서는 정규표현식과 stringr 패키지의 str_replace_all을 통해 정규표현식으로 찾은 단어 집합들을 제거하는 것으로, 초록데이터에서 한글 초록데이터 부분만을 얻어내었다. 사용한 정규표현식은 다음과 같다.

정규표현식	의미	사용처
[^[가-힣]]+$	텍스트 맨 뒤에서 한글이 아닌 모든 단어들의 모음을 찾기	한글 초록 뒤에 있는 영문 초록 제거
^[^[가-힣]]+	텍스트 맨 앞에서 한글이 아닌 모든 단어들의 모음을 찾기	한글 초록 앞에 있는 영문 초록 제거
[一-龥]+	한자로 되어있는 모든 단어를 텍스트 내에서 찾기	초록 데이터에 존재하는 한문단어 제외

<표 2> 초록 데이터 전처리에서 사용한 정규표현식

이러한 방법을 통해 초록에서 한글 초록데이터만 남긴 후, 초록 데이터 중 일부에서만 나타나는 한문 텍스트를 제거해주는 것으로 한글 초록데이터에서 최대한 유의미한 단어만을 남겼다. 이후 제거된 영문 초록데이터로 인해 생기는 결측치들을 제거하였고, 이 과정을 통해 한글 초록데이터 611건을 추출할 수 있었다.

이후 수집한 한글 초록데이터 611건을 바탕으로 말뭉치를 만들고 토큰화하기 이전에, 단어들이 띄어쓰기 단위로 되어있는 영어와 달리 형태소단위로 단어들이 결합하여 의미 단위로 띄어쓰기된 한글의 특성을 고려하여 형태소분석기를 통해 초록데이터를 형태소 단위로 분해하고, 명사만을 추출한 후 띄어쓰기로 연결해주어야 한다. 이때, 명사 이외의 품사는 토픽 모델링을 진행하는 과정에서 유의하지 않은 것으로 판단하여 제외하였고, 명사를 추출한 후에 다시 띄어쓰기로 연결하는 이유는 본 연구에서 사용하는 패키지 tm이 영문 텍스트마이닝을 위한 패키지이기 때문이다. 앞에서도 언급한 것처럼, 영어의 경우에는 단어들이 띄어쓰기 형태로 되어있기 때문이기 때문에 이후 말뭉치로 문서-단어행렬을 형성할 때 띄어쓰기 단위로 말뭉치 내의 단어들을 띄어쓰기로 분해하여 단어로 인식하기 때문이다.

한글 초록데이터의 형태소 분해를 위해서 KoNLP패키지를 사용하였고, 이 중에서 명사들만을 추출하기 위해서 extractNoun 함수를 사용하였다. 이때, 텍스트에서 최대한 많은 명사들을 추출하기 위해서 KoNLP에서 지원하는 형태소 분해의 기준이 되는 사전 중 가장 많은 단어 수를 가진 NIADic을 사용하였다. NIADic을 만든 한국지능정보사회진흥원 빅데이터 센터의 소개에 따르면, 기존 형태소 사전이 37만개의 단어를 수록하고 있는 반면 NIADic은 총 93만개의 단어와 함께 전문용어, 인물사전, 신조어 등을 추가하여 기존 Sejong Dictionary보다 2.56~5.36배의 형태소를 발생시킬 수 있다고 소개하고 있다. 원래 계획 상 NIADic에 초록데이터 수집 과정에서 함께 수집 가능한 키워드를 NIADic에 추가하여 문헌정보학 분야에 대한 형태소 분석 정확도를 향상시키려고 하였으나, KoNLP가 기반을 두고 있는 Java에서 오류가 발생하면서 이 과정은 진행할 수 없었다. 이처럼 각 초록데이터에서 명사만 추출한 후 다시 띄어쓰기로 연결해주고, 이를 tm 패키지의 Corpus 함수를 통해 말뭉치 형태로 변환한다.

이후 말뭉치에서 다시 전처리를 진행해준다. KoNLP의 형태소 분해 성능 향상을 위해 사전 데이터에 단어를 추가하지 못하였고, 또한 KoNLP에서 문장 끝 부분에 있는 단어들의 경우에는 제대로 형태소 분해를 해내지 못하는 모습을 보여주었기 때문에, 제대로 분류가 되지 않은 단어들에 대해서 일일히 제거하는 과정을 동시에 진행하였다.

말뭉치 전처리 과정에서 사용한 함수과 그 기능들은 다음과 같다.

함수명	기능	비고
removePunctuation	특수문자 제거	!,"' 등의 특수문자 제거
removeNumbers	숫자 제거	-
tolower	영문 단어 소문자 변환	영문 단어를 통일시키는 것으로 불용어인 영단어를 삭제하는데 도움을 줌
removeWords	불용어 사전에 저장된 단어들을 말뭉치 내에서 제거	불용어 사전에 저장된 단어들만 제거

<표 3>말뭉치 전처리 과정에서 사용한 R 함수

전처리 과정 중 삭제할 불용어를 선별하는 과정에서 영문 단어의 경우에는 토픽 내에서 주요한 의미를 가지는 단어보다는 어떠한 분석 기법의 이름이나, 이해를 돕기 위해 한글과 병기되어 표시되는 경우에 나타난 것을 확인하였기 때문에, 불용어 제거 과정에서 영문 단어들을 다시 전부 제외하는 과정을 진행하였다.

이 과정에서 'lda'와 같은 영문 단어, '연구', '대표', '주요', '현황' 등 많이 나타나지만 토픽에서 주요한 의미를 가지지 않는 단어, '정보''등 미처 전처리 과정에서 특수문자가 제거되지 않은 단어, '제안하였' 등 불완전한 형태소 분해로 인해 명사로 처리된 단어들을 모두 불용어 처리하여 불용어 사전으로 저장시켰고, 이 과정에서 불용어 1202개를 제거하였다.

이제 전처리를 모두 거친 한글 초록데이터 말뭉치를 DocumentTermMatrix 함수와 TermDocumentMatrix 함수를 통해 각각 문서-단어 행렬과 단어-문서 행렬로 변환해준다. 이 중 문서-단어 행렬은 이후 진행될 잠재 디레클레 할당을 통한 토픽 모델링에 사용할 예정이며, 단어-문서 행렬은 단어의 출현 빈도에 따른 계층적 클러스터링에 사용할 예정이다.

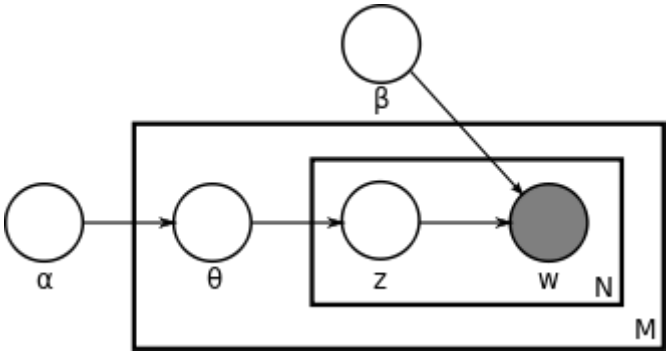
2.2. 분석방법

2.2.1. LDA((Latent Dirichlet allocation, 잠재 디레클레 할당)

위의 과정에서 전처리 된 텍스트 데이터를 분석하는데 대표적인 토픽 모델링 기법인 LDA를 사용하고자 한다. LDA는 베이즈 기반의 머신러닝 모델로, 주제별 단어 분포를 바탕으로 각 문서 별 단어 분포를 분석하는 것으로 역으로 해당 문서가 어떤 주제를 다루고 있을지 예측한다. (David M. Blei 외, 2003)

LDA에서 사용되는 모수 중 중요한 모수는 다음과 같다. 문서의 개수를 M, 토픽의 개수를 K, 단어의 개수를 V라고 하면

- θ는 k차원 벡터이고, 문서가 각 주제에 속할 확률 분포이다.
- Z_n는 k차원 벡터이고, 특정 단어가 각 주제에 속할 확률분포이다.
- β는 K × V 행렬로, 각 주제가 문헌에서 단어를 생성할 확률을 나타낸다.



<그림 1> LDA에서의 문서 생성 과정

LDA 에서의 문서 생성을 그래프로 도식화하면 다음과 같고, 도식화된 과정을 간단히 풀면 다음과 같다.

1. $N \sim Poisson(\xi)$을 선택한다.
2. $\theta \sim Dir(\alpha)$를 선택한다(α는 디레클레 분포의 매개변수).
3. 문서 내의 단어 w_n에 대해서
    - $Z_n \sim Multinomial(\theta)$를 선택한다.
    - Z_n이 주어졌을 때 w_n은 $P(w_n|Z_n, \beta)$로부터 선택한다.

본 연구에서는 실제 관측 가능한 w_n을 통해 θ와 Z_n을 추론하는 것에 초점을 두는 것으로, 현재 문헌정보학 논문들에서 활발히 연구중인 주제를 확인하고자 한다.

또한, LDA의 특성 상, 한 문서내에서 주제들은 정확히 하나의 주제로 분류되는 것이 아닌, 여러 개의 주제들의 가중치의 형태로 혼합되어 나타나는 형태로 나타난다. 따라서, 여러 주제들이 복합적으로 나타나고 융복합적인 성격을 띄는 문헌정보학의 논문들에서 나타나는 주제들에 대한 토픽모델링에 효과적일 것으로 예상된다.

2.2.2. 계층적 클러스터링(Hierarchical Clustering)

계층적 클러스터링이란, 트리 모형을 이용해서 개별 개체들을 순차적이고 계층적으로 유

사한 개체 혹은 그룹과 함께 클러스터를 만들어주는 알고리즘이다. 계층적 클러스터링은 트리 형태의 구조인 덴드로그램을 사용하고, 사전에 클러스터의 개수를 정하지 않고도 학습이 가능하다는 장점이 있다.

계층적 클러스터링을 진행하기 위해서는 모든 개체들 간의 유사도, 혹은 거리를 구할 수 있어야 한다. 개체 간 거리를 측정하는 방법으로는 유클리드 거리, 맨하튼 거리, 민코우스키 거리, 마할라노비스 거리 등이 있지만, 본 연구에서는 유클리드 거리를 통하여 개체들 간의 거리를 측정할 계획이다. 유클리드 거리를 수식으로 나타내면 다음과 같다.

$$Euclidean\ distance = \left||P - Q|\right| = \sqrt{(P - Q) \cdot (P - Q)}$$
$$where\ P = (p_1, \cdots, p_{n)},\ Q = (q_1, \cdots, q_{n)}$$

유클리드 거리를 통해 모든 개체들의 거리를 구했다면, 이를 바탕으로 거리 행렬을 만들어 서로 가까운 거리들끼리 묶어주는 것으로 트리형의 계층적인 덴드로그램을 형성한다. 이때, 묶어주는 기준은 최소연결법, 최대연결법, 평균연결법 등이 존재하지만, 본 연구에서는 ward 연결법을 통하여 각 개체들끼리 거리를 기준으로 묶어주었다. 이 때 와드 연결법은 군집 간 거리에 따라 데이터를 연결하는 것이 아닌, 군집내의 편차들의 제곱합을 기준으로 군집간 결합 시 오차제곱합의 증가분이 최소가 되도록 묶는 방법이다. 이 과정을 통해 완성된 덴드로그램에서 기준선을 어디에 두느냐에 따라 임의로 클러스터의 수를 조절할 수 있다.

본 연구에서는 단어-문서 행렬을 통해 얻을 수 있는 각 단어당 문서별 단어 출현 빈도를 바탕으로 이를 표준화하여 유클리드 거리로 거리 행렬을 생성하여, 이를 덴드로그램을 통해 계층적인 구조로 표현하고, 데이터 분석과정에서 얻어낸 LDA의 최적의 토픽 개수만큼 클러스터를 만들 수 있도록 cut-off하여 단어들의 계층적인 구조 형태를 살펴볼 예정이다. 이를 통해 단순히 단어의 출현 빈도만을 확인하는 것이 아닌, 각 문서에서의 단어의 출현 빈도에 따른 계층적 구조를 통해 LDA를 통한 토픽 모델링 결과와 비교하는 것으로 단어의 구조와 잠재적인 주제들에 대한 관계를 확인할 수 있을 것으로 기대된다.

3. 데이터 분석 및 결과 해석

3.1. LDA를 통한 토픽 모델링

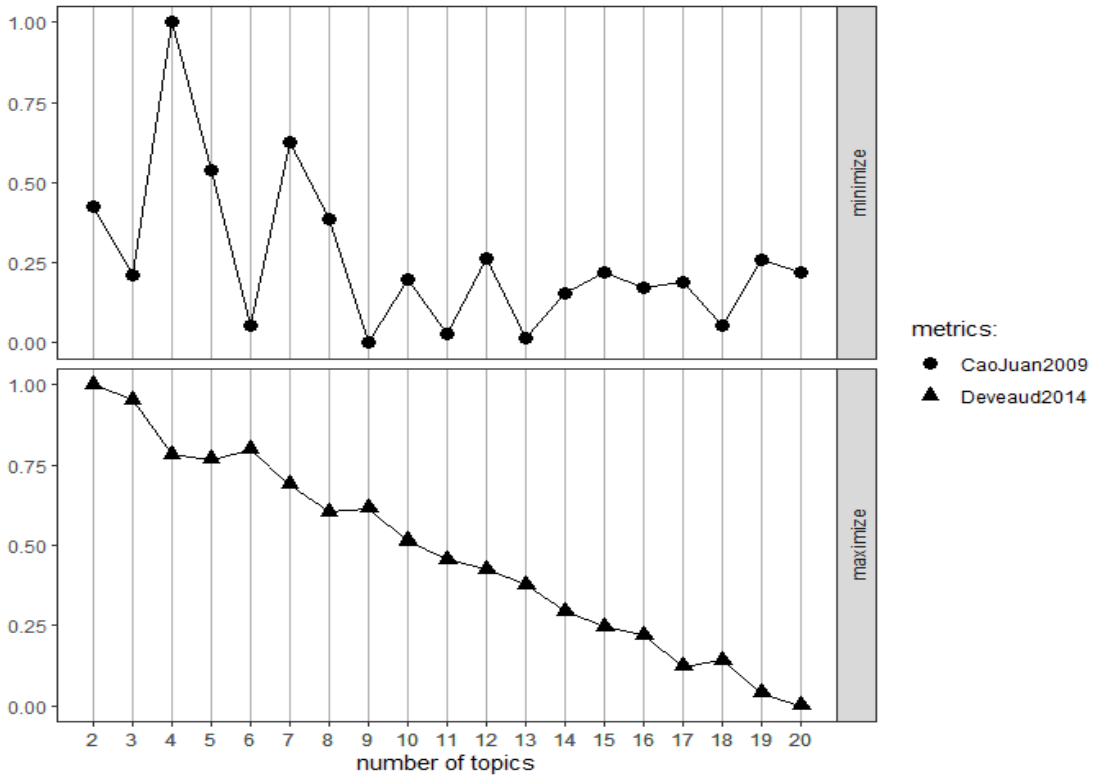
3.1.1. 토픽 모델링을 위한 최적의 주제 수 K 파라미터 튜닝

LDA를 통해 토픽모델링을 진행하기 앞서, LDA를 진행하기 위해 구해야 하는 주제 수 K에 대한 파라미터 튜닝을 실시하였다. LDA의 특성상 사전에 설정한 주제 수 K를 바탕으로 분석

을 진행하기 때문에, 최적의 결과를 얻기 위해서는 사전에 설정해야 하는 하이퍼 파라미터인 K에 대한 튜닝이 필수적이다. 최적의 주제수 K에 대한 파라미터 튜닝을 위해서 앞에서도 소개한 ldatuning 패키지를 사용하였다.

LDA의 최적의 주제수 K에 대하여 파라미터 튜닝을 진행할 때, 모델의 성능 지표로 사용되는 값들은 다양하다. 대표적으로 사용되는 지표로는 Perplexity나 Coherence가 있지만, 본 연구에서는 ldatuning 패키지에서 제공하는 'CaoJuan2009'과 'Deveaud2014'를 사용하였다. 이 중 'CaoJuan2009' 지표는 토픽 분포 간의 코사인 유사도를 최소화하는 값을 구해주고, 값이 최소화 될수록 모델의 성능이 높다고 판단할 수 있게 되고, 'Deveaud2014'는 토픽 분포 간의 젠슨-샤논 거리가 최대화하는 값을 구해주어 값이 최대화될수록 모델의 성능이 높다고 판단 할 수 있게 된다(shawnk123.log, 2022).

ldatuning 패키지에서 최적의 토픽 수 K를 2개와 20개 사이로 설정하고, K값 별 모델의 성능을 'CaoJuan2009'과 'Deveaud2014' 지표를 평가한 후, 이를 시각화하여 확인해 본 결과 다음과 같은 결과를 얻을 수 있었다.



<그림 2>최적의 주제 수 K 파라미터 튜닝 결과 시각화

전체적으로 'CaoJuan2009' 지표는 최적의 주제 수 K가 커짐에 따라 진동하다가 K값이 작아질수록 그 값들이 최소화되는 경향이 나타났고, 'Deveaud2014' 지표는 K가 2일 때 제일 최대

화되다가 그 이후로는 6에서 살짝 증가하고, 그 이후로는 계속 감소하는 추세를 보였다.

이를 바탕으로 최적의 주제 수 K를 선정한다면, K가 6이 되는 지점에서 값들이 최적화되는 것을 확인할 수 있다. K=6 에서 모든 지표가 최대화 혹은 최소화되지는 않았지만, 'CaoJuan2009' 지표에서는 K=6 일때와 그 외 최소화되는 지점과 그 값의 차이가 크게 나타나지 않았고, 'Deveaud2014' 지표에서는 전체적으로 감소하는 추세이지만 K=6 인 지점에서 다시 그 값이 상승하는 경향이 있기 때문에 최적의 주제 수 K를 6으로 선정하여 이후 분석을 진행하였다.

3.1.2. LDA 토픽 모델링 및 결과 해석

3.1.1.에서 얻어낸 최적의 주제 수 K=6으로 설정한 후, 반복 수를 1000으로 설정하여 LDA를 실시한 결과, 5년간 발간된 논문에서 얻어낸 한글 초록 데이터 611건에서 나타나는 6개의 토픽에서 주요하게 나타나는 상위 10개 단어들은 다음과 같았다.

주제 1	주제 2	주제 3	주제 4	주제 5	주제6
도서관	이용자	기록	교육	정보	데이터
서비스	이용	관리	인식	자료	논문
공공도서관	평가	기관	프로그램	주제	국내
공간	만족	지원	학교도서관	유형	학술지
운영	요인	구축	독서	디지털	모델
정책	관계	국가	개발	기술	인용
특화	환경	업무	학생	선정	분류
문화	모형	대학	사서	사회	주제
지역	측정	체계	사서교사	콘텐츠	문헌
역할	설문조사	국내	운영	한국	네트워크

<표 4> LDA 토픽모델링 주제 별 주요 상위 10개 단어

<그림 3.1> 주제 1의 주요 단어 Word Cloud

해당 주제들에서 나타나는 단어들을 바탕으로 문헌정보학 주제분류표와 비교한 결과, 다음과 같은 결과를 얻어낼 수 있었다. 주제 1의 경우 “도서관”, ”서비스”, ”공공도서관”, ”공간”, “운영”, ”정책”등의 단어가 주를 이뤘으며 해당 논문들을 분석한 결과 “도서관 체계”, “도서관 경영”에 해당하는 주제임을 확인할 수 있었다. 전반적으로 도서관의 지속과 발전을 위한 연구가 주를 이뤘으며 특정 지역이나 국가, 집단, 연구 방법에 국한되지 않고 모든 형태의 도서관을 다룬 연구들이 주제 1에 포함되었다. 또한 지역 공공도서관이나 대학, 박물관, 병영, 병원 등 기관 내 도서관, 특화 도서관 등에 대한 활성화 방안에 대한 연구 또는 도서관 서비스 개발 연구, 관리 방안 및 관리 기준 연구가 주로 토픽 1에 해당하였으며, 4차 산업 혁명과 연결 지어 도서관의 발전 방향과 미래 동향 예측 연구 또한 포함되었다. 대표적으로 수집기간 중 발간된 논문 중 “대전지역 공공도서관 운영현황과 발전과제”(윤혜영, 2019)의 경우 대전지역을 중심으로 24개의 공공도서관의 일반 현황과 행정 체계를 분석하는 지역 특화 적인 단기 분석 연구를 진행하였으나, “빅데이터 분석 기법을 활용한 도서관발전종합계획 동향 분석 연구”(노영희, 2018)의 경우 텍스트 마이닝 기법을 통해 2009년부터 2017년까지의 도서관발전종합계획에 대한 언론보도를 수집하여, 빅데이터 분석 기법을 통해 도서관발전종합계획이 시행되는 시기 동안에 발생한 시기별 동향과 시사점을 도출하였다.

<그림 3.2> 주제 2의 주요 단어 Word Cloud

주제 2에서는 ‘이용자’, ‘평가’, ‘만족’, ‘설문조사’, ‘품질’, ‘정보서비스’ 등의 단어들이 주를 이루었으며, 해당 논문들을 분석한 결과 ‘정보 이용과 관련한 연구’에 해당하는 주제임을 확인할 수 있었다. 전반적으로 정보서비스를 이용하는 이용자의 만족도 평가 또는 품질에 대한 평가에 대한 연구들이 주제 2에 속해있는데, 도서관 서비스뿐 아니라 학술정보 서비스 등 다양한 정보서비스 이용에 대한 연구들을 다루고 있다. 대표적으로 수집기간 중 발간된 논문 “의과대학 대학원생의 학술정보 검색엔진 이용 동기 및 이용자 만족도에 관한 연구”(심새봄, 이용정, 2019)에서는 의과대학 대학원생들이 학술정보 검색엔진을 이용하는데 있어 편리성과 전문성이라는 두 가지 이용 동기 요인이 학술정보 검색엔진의 이용자 만족도에는 유의미

한 영향을 미치고, 특히 편리성은 이용자들이 학술정보 검색엔진을 선택하는데 있어서도 가장 큰 영향을 미치는 것을 도출해냈다. "학술정보포털에 대한 이용자만족 관련 인식에 관한 연구 - NAVER 전문정보의 학술자료 검색 기능을 중심으로"(김양우, 2017)에서는 다양한 전공영역의 학부 학생들이 본인 전공에 관한 학술적 탐색주제에 대한 검색을 수행하는 과정에서, 학술정보 전문포털에 대한 만족이나 불만족 등의 인식과 그 이유를 조사했다. 인터페이스, 검색메커니즘 및 검색결과 등 세 가지 범주에 속하는 다양한 평가 항목을 통해 만족도를 평가하고, 이에 기반하여 이용자 만족도를 높일 수 있는 개선 방향을 제시하였다.

<그림 3.3> 주제 3의 주요 단어 Word Cloud

주제 3은 '기록관리학'에 해당하는 주제로 "기록", "관리", "장서", "아카이브", " 자원", "체계" 등의 단어가 주를 이뤘다. 기록관리학은 인간활동의 증거와 기억으로서의 기록을 연구하는 학문으로서, 해당 주제에서는 도서관의 장서 구성 및 관련에 관련된 연구부터 기록의 디지털화에 따른 패싯 기반 연구까지 포괄적인 범위의 기록관리학에 대한 연구를 포함한다. 이처럼 주제 3은 기록에 관한 광범위한 연구를 포함하며, 기록을 적절하게 관리·생산해 효율적으로 사용하고 증거적 가치나 영구보존 가치가 있는 기록을 보존하여 쉽게 검색하고 활용할 수 있도록 체계를 구축하는 것에 그 목적을 둔다. 일례로 데이터 수집 처리된 논문 중, "교육지원청 기록관리 기관평가 개선방안에 관한 연구"(윤영조·정연경, 2020)는 교육지원청의 기록관리 기관평가에 대한 문제점을 개선하고자 문헌연구, 설문조사, 심층면담을 수행하였고 이를 기반으로 해 교육지원청의 특수성을 반영한 개선방안을 제안하였다. 또한 "방송영상기록물의 저작자 식별을 위한 패싯 기반 식별체계 구축 연구"(정연주·이승민, 2020)에서는 방송영상기록물을 대상으로 방송영상기록물의 제작에 참여한 저작자를 식별하기 위한 패싯 기술항목을 제시하여 기록물에 대한 저자식별체계 구축 서비스를 도출하였다.

<그림 3.4> 주제 4의 주요 단어 Word Cloud

주제 4는 '도서관 서비스'에 해당하는 주제로 '교육', '인식', '프로그램', '학교도서관', '독서', '개발' 등의 단어가 주를 이뤘다. 이 주제의 연구는 주로 도서관 유형별 서비스에 관한 연구와 서비스 평가 연구가 활발하게 연구되고 있다. 또한, 주제 4의 도서관 서비스 연구는 도서관 이용자의 교육과 정보 리터러시에 대한 연구도 포함한다. 일례로 "이용자 인식조사를 기반으로 한 공공도서관 비대면 서비스 운영 방향에 대한 연구" (윤다영,2021)에서는 팬데믹 상황에서도 도서관 서비스를 원활하게 진행하기 위해 도서관 비대면 서비스의 범위를 파악하고, 이용자의 인식조사를 통해 도서관 비대면 서비스 운영 방향을 제안하고자 하였다. 온라인 설문조사를 통해 이용자의 의견을 수렴하였으며, spss 통계 프로그램을 활용하여 분석한 결과 유의미한 결과를 얻었다. 또한 "국가 수준의 도서관 계획을 기반한 학교도서관의 교육공동체 협력 방안"(이승길,2019) 에서는 학교도서관이 교육공동체에 협력할 수 있는 방안을 도출하기 위해 학교도서관진흥기본계획과 도서관발전종합계획의 연계성을 분석하였고, 이를 통해 관종별 도서관 전문성 강화, 학교도서관 교육과정 설정 등을 제시하였다.

<그림 3.5> 주제 5의 주요 단어 Word Cloud

주제 5는 '정보', '자료', '디지털', '콘텐츠', '분류체계', '작품' 등의 단어들이 주요하게 나타나고 있다. 문헌정보학 분류체계표를 통해 분석해본 결과, 자료조직 분야의 분류에 대한 연구와 디지털 도서관, 서지학에 대한 연구들이 복합적으로 나타나는 주제인 것으로 확인되었다. 주제 5가 이렇게 복합적으로 나타나게 된 이유는 최근 발간된 서지학 논문들에서 종이책과 디지털 책과의 비교 연구를 진행하거나, 고문헌에 대해서 작품, 주제별로 디지털 아카이브를 생성하거나, 분류를 진행하는 연구들이 많이나타나게 되면서 자료분류에 대한 연구와 도서관의 디지털화에 대한 연구에 대한 중간 고리 역할이 되었기 때문으로 보인다. 이 주제에 해당하는 연구로는 "스마트디지털 환경 기반 도서관 구축에 관한 사서 인식 연구"(강필수, 노영희 and 김윤정, 2021)가 있다. 해당 연구에서는 디지털서비스 스마트화 추진 방안을 모색하기 위해서 디지털 자료 서비스 이용현황과 운영상의 애로사항을 살펴보았고, 이를 바탕으로 스마트도서관을 도입해야되는 우선 지역과 스마트 도서관 도입을 위해 필요한 장비 및 기술에 대해서 제언하고 있다. 또 다른 연구로는 "작품·주제형 디지털 인물 아카이브의 콘텐츠와 구성에 관한 연구 - 조선 중기 여류 문인을 중심으로 -"(김희숙 and 장우권, 2020)가 있었다. 해당 연구에서는 신사임당, 송덕봉, 허난설현을 대상으로 작품·주제형 인물 아카이브 콘텐츠와 그 구성방안을 수립하였고, 컨텐츠들을 분류하여 효과적으로 인물 아카이브를 구축하고자 하였다.

<그림 3.6> 주제 6의 주요 단어 Word Cloud

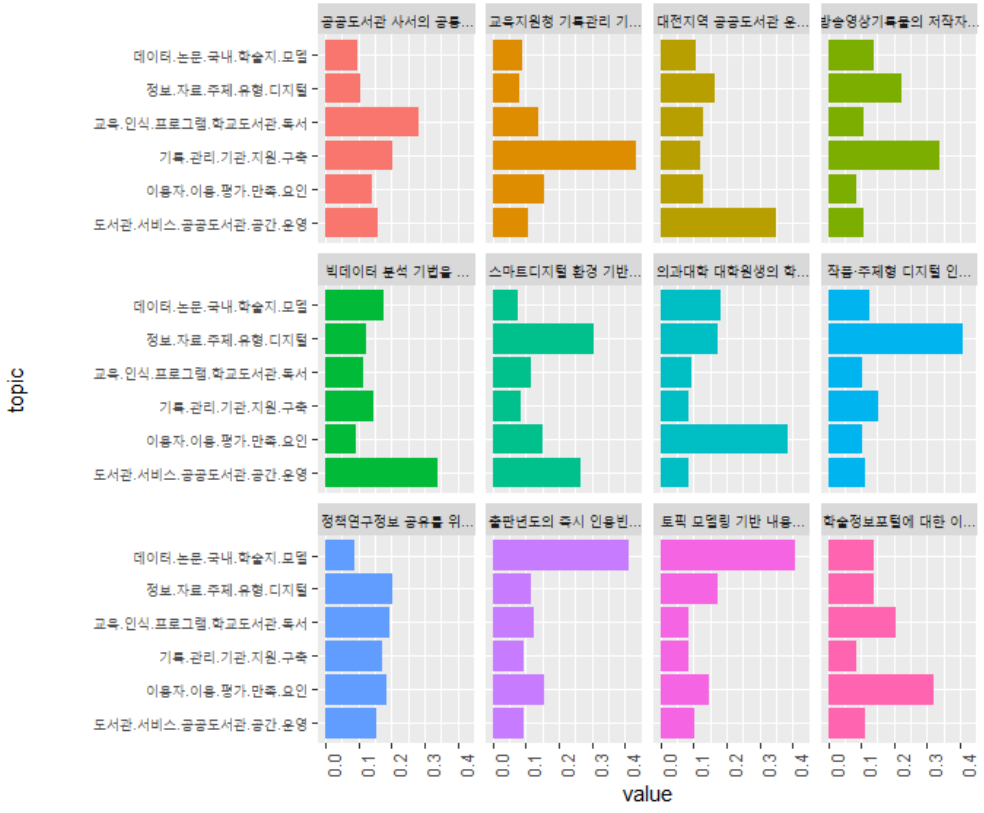
마지막으로 주제 6은 '데이터', '논문', '학술지', '인용', '네트워크', '추출' 등의 단어들이 주요하게 나타나고 있다. 문헌정보학 주제분류표를 통해 분석해본 결과, 자료조직분야의 주제분석 연구와 메타데이터의 연구인 것을 확인할 수 있었고, 추가로 본 연구와 같이 메타 연구들이 나타나는 주제인 것을 확인할 수 있었다. 특히, 이 주제에 해당되는 연구들에서는 텍스트마이닝을 위시로 한 전통적인 머신러닝과 딥러닝을 결합하려는 시도가 많이 나타나고 있는 것 또한 확인하였다. 이 주제에 해당하는 연구로는 "출판년도의 즉시 인용빈도를 포함하는 학술지 인용지수 개발"(이재윤,2018)이 있다. 해당 연구에서는 기존 학술지 영향력 측정에서 당해년도의 인용을 배제하는 문제를 해결하기 위해 즉시 인용을 포함하는 인용지수 TIF와

MIF를 개발하고, 연도별 변동성이 훨씬 적다는 것을 실험을 통해 도출해내어 3년마다 재평가를 시행하는 국내 상황에 더욱 맞는 장점을 가진다는 것을 확인하였다. 또 다른 연구로는 "토픽 모델링 기반 내용 분석을 통한 학제 간 융합기술 도출 방법"(정도헌 and 주황수, 2018)가 있었다. 해당연구에서는 본 연구와 마찬가지로 생명공학-정보문화기술 간의 융합기술 도출 방법을 구하기 위하여 생명공학 분야의 메타데이터로 지적 구조를 분석하고 이를 토픽 모델링한 후, 토픽모델링을 통해 얻어낸 단어들을 통해 그 용어의 상위어를 도출해내어 정보문화기술 분야와 매칭하는 방식으로 유망한 생명공학-정보문화기술간의 융합 기술 아이템을 얻어내었다.

3.1.3. θ를 통한 문서 별 주제 가중치 확인

2.2.1.에서 설명한 것처럼, θ는 문서가 각 주제에 속할 확률 분포이다. 각 문서 내에서 문서의 내용은 θ를 각 주제에 대한 가중치로 하는 선형 결합의 형태로 나타나게 된다. 따라서, 각 문서에서 θ의 원소들의 값을 바탕으로 해당 문서가 어떤 주제를 가질 확률이 높은지를 확인할 수 있게 된다.

본 연구에서 활용한 논문들과 몇몇 논문들을 뽑아, 해당 문서들에서 나타난 θ값을 시각화 해본 결과, 다음과 같은 결과를 얻을 수 있었다.



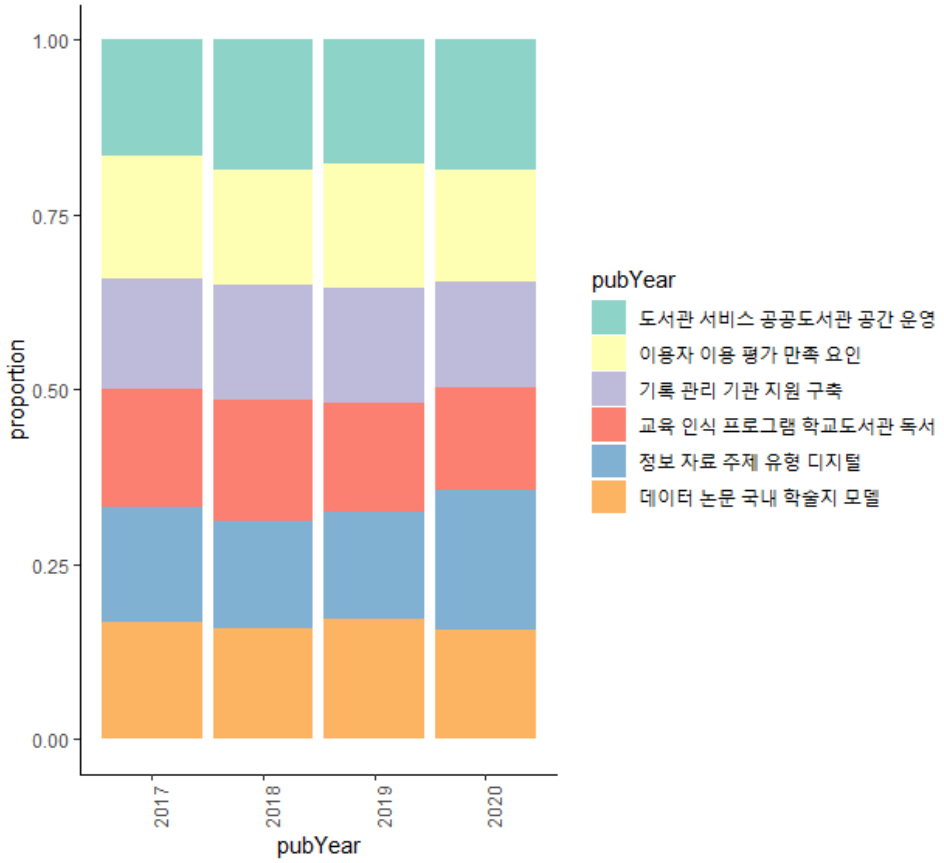
<그림 4> 논문 별 주제 분포 θ 시각화

전체적으로 살펴보았을 때, 각 논문들은 6개 중 하나의 주제에 대해서 가중치가 높게 나타나고 있었다. 그러나 이와 동시에, 주제 가중치가 가장 높은 값과 비슷하게 나타나거나, 혹은 전체적으로 주제 가중치가 고르게 나타나는 논문 또한 존재하는 것이 확인되었다.

이는, 문헌정보학 분야 연구에서 주로 각 주제별 연구로 진행되지만, 이 주제들이 서로 개별적으로 진행되는 것이 아닌 서로 유기적으로 연관되어 연구가 진행되기 때문인 것으로 해석할 수 있다. 즉, 논문들 내에서의 주제의 분포는 문헌정보학이 갖는 융복합적인 특성으로 인해 나타나는 것임을 알 수 있다.

3.1.4. 연도별 문헌정보학 내 주제 변화 분석

연도별 문헌정보학 분야에서의 주제 변화를 분석하기 위해서 데이터 수집 과정에서 논문이 게재된 년도를 수집하였고, 이 중 게제된 월을 포함하여 수집된 경우에 이를 년 단위로 변경해주는 전처리 과정을 거쳤다. 전처리 이후, 해당 년도에 게제된 논문들에서 나타나는 주제 분포 θ의 평균을 구하여 이를 시각화하였다. 시각화한 결과는 다음과 같았다.



<그림 5> 게재 연도별 평균 주제 가중치 θ 시각화

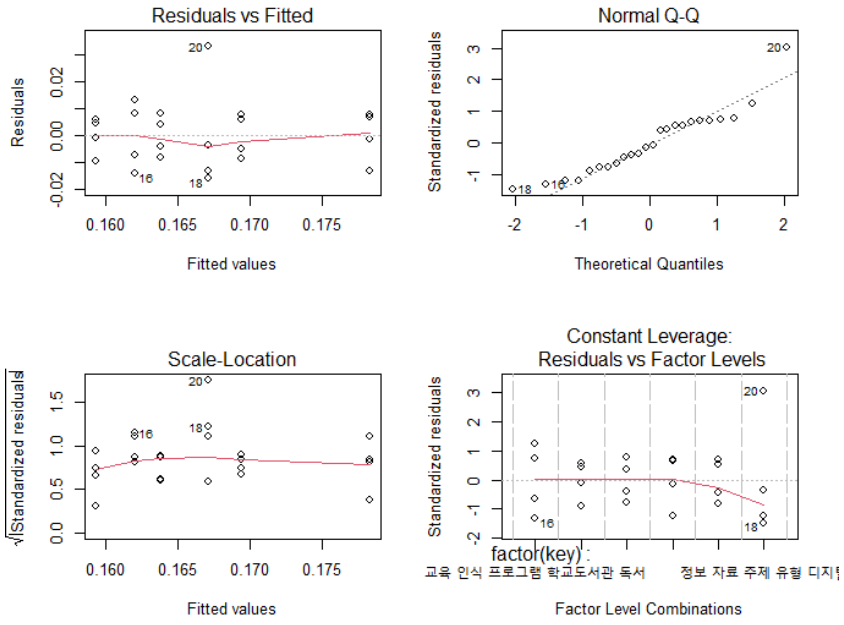
시각화 결과를 바탕으호 확인해본 결과, 연도별 평균 주제 가중치 값들에 큰 차이가 없어 보인다는 생각이 들었고, 연도별 평균 주제 가중치 값들이 주제에 따라 차이가 있는지 확인해보기 위하여 One-Way ANOVA를 실시하였고, ANOVA 결과는 다음과 같았다.

Source	DF	SS	MS	F	Pr(>F)
주제	5	0.000910	0.000182	1.173	0.36
Error	18	0.002792	0.000155	3.218380	0.000393
Total	23	0.003702			

<표 5> 주제에 대한 One-Way ANOVA 결과

ANOVA 결과 주제에 대한 P-Value는 0.36으로 유의수준 0.05보다 크게 나타났다. 즉, 주제별로 주제 가중치들의 평균 값에는 큰 차이가 없음을 ANOVA를 통해 확인할 수 있었다.

이후 ANOVA 또한 이상치 및 데이터 구조에 영향을 받기 때문에, 모델의 기본 가정을 확인해야 한다. 본 연구에서는 잔차 Plot을 통해 모델이 기본 가정을 위반하고 있는지를 확인하고, 문제가 있을 것으로 보이는 부분에서 추가 검정을 진행하는 방식으로 모델적합성 검사를 진행하였다.



<그림 5> 모델 적합성 평가를 위한 잔차 plot 시각화

R을 통해 잔차 plot을 확인해본 결과, 잔차들이 전체적으로 무작위로 분포해있는 것을 확인하였고, 또한 잔차 plot에서 예측 값이 커질수록 잔차가 마치 확성기 모양처럼 커지거나 작아지는 경향도 없는 것을 확인하였다. 그러나, 모델의 정규성을 확인해보았을 때, 전체적으로는 정규성을 만족하는 것으로 보이나 가장 큰 값으로 인하여 정규성이 위배될 수 있을 것으로 판단되었다. 만약 모델의 정규성이 위배되었을 때 모델의 기본 가정인 오차가 정규분포를 따르지 않게 되어 ANOVA 결과를 신뢰할 수 없게 되기 때문에, Shapiro-Wilks 검정을 통해 정규성을 추가로 검정해보았다. 검정결과, P-Value 0.051로 유의수준 0.05에서 잔차가 정규분포를 따른다는 귀무가설을 기각하지 못하였기 때문에, ANOVA 모델이 정규성을 만족한다고 판단하였다.

모델 적합성을 확인해 본 결과 ANOVA 결과를 신뢰할 수 있게 되었고, 이를 바탕으로 연도별 주제 가중치 평균이 주제별로는 차이가 없음을 확인할 수 있었다. 따라서, 연도별로 주제별 평균 주제 가중치에 대한 분석은 진행하지 않고, 연도에 따라 평균 주제 가중치가 어떻게 변화하였는지 분석하는 것이 유의미할 것이라고 결론 내렸다.

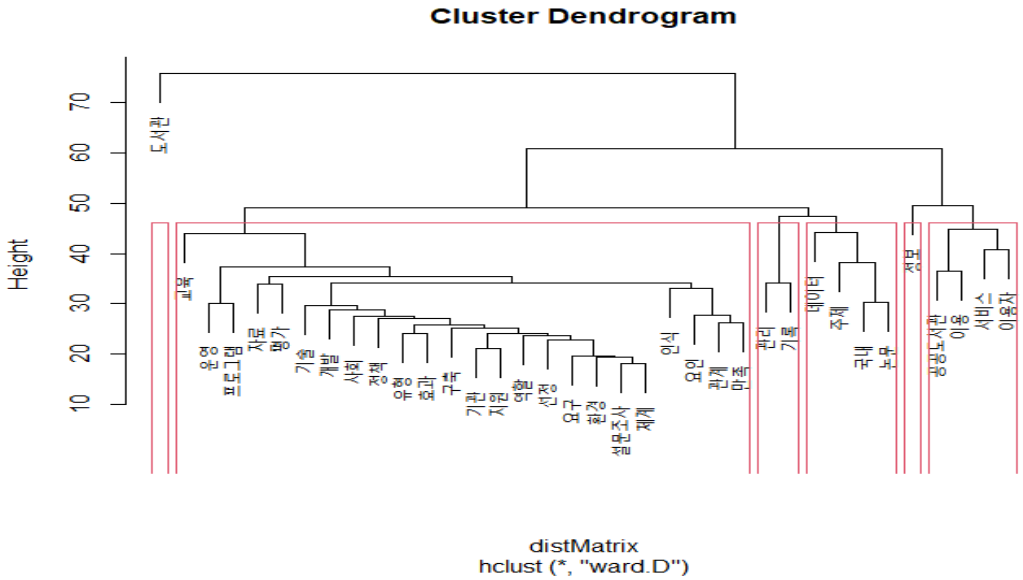
	주제1	**주제2**	**주제3**	**주제4**	**주제5**	**주제6**
2017	0.1651	0.1752	0.1582	0.1700	0.1633	0.1678
2018	0.1853	0.1645	0.1639	0.1751	0.1513	0.1596
2019	0.1768	0.1770	0.1652	0.1548	0.1539	0.1721
2020	0.1859	0.1608	0.1498	0.1479	0.1999	0.1555

<표 6> 게재 연도별/주제 별 평균 주제 가중치 θ

이를 바탕으로 연도별 평균 주제 가중치를 확인해본 결과, 2020년에는 주제 1에 속하는 도서관 체계, 경영 연구와 주제 5에 속하는 자료조직과 디지털 도서관, 서지학에 대한 복합적인 주제의 연구가 많이 나타난 것으로 확인되었다. 이는, 2020년 코로나-19의 확산으로 인하여 사람과의 접촉을 최소화하는 '언택트'시대로 변화하면서 시대의 변화에 발맞추어 포스트 코로나 시대의 도서관의 위기를 타파하기 위해 다양한 해결책을 모색하고, 기존에 진행하던 도서관의 디지털화에 대한 연구를 더욱 활발하게 진행하였기 때문으로 해석할 수 있다.

3.2. 계층적 클러스터링을 통한 단어의 계층적 구조 확인

마지막으로 단어의 문서 별 출현 빈도를 바탕으로 계층적 클러스터링을 통하여, 문서 전체에서 주요하게 나타나는 단어들끼리 어떤 관계를 가지고 있는지 확인하고자 하였다. 전처리 과정에서 생성한 단어-문서 행렬을 바탕으로 거리 행렬을 만들고, 이를 바탕으로 계층적 클러스터링을 진행하였다. 이후 생성된 덴드로그램을 6개의 군집이 생성되도록 cut-off를 진행해주었고, 이 과정을 통해 생성된 6개의 군집은 다음과 같이 나타났다.



<그림 6> 출현 빈도 기반 덴드로그램과 계층적 클러스터링 결과

클러스터링 결과를 확인해보면, 전반적으로 단어의 출현 빈도에 따라 단어들이 계층적 관계를 이루고 있는 것을 확인할 수 있다. 문서 전체적으로 많이 등장할 것으로 예측되는 '도서관' 이라는 단어가 하나의 클러스터에 배정되었고, '기술', '개발', '설문조사', '구축' 등 주요하지만 방법론이나 분석 기법에 해당되는 단어들끼리 같은 클러스터에 배정되었음을 알 수 있었다. 해당 단어들을 제외하고 클러스터를 살펴본다면, 우리가 LDA 토픽모델링을 통해 얻어낸 주제들에서 나타난 단어들과 비슷하게 나타나고 있는 것을 확인할 수 있었다. 토픽모델링에서 주제 3에서 주요하게 나타난 단어인 '기록'과 '관리가 하나의 클러스터로 묶였고, 주제 6에서 주요하게 나타난 데이터, 주제, 국내, 논문 또한 하나의 클러스터로 묶였으며, 주제 5에서 주요하게 나타난 '정보' 또한 하나의 클러스터에 배정되었다. 또한 주제 1과 주제 2에서 주요하게 나타난 '공공도서관'과 '이용', '서비스'와 '이용자'가 하나의 클러스터에 배정된 모습을 확인할 수 있었다. 또한 주제 4에서 주요하게 나타난 '교육'이란 단어도 동시에 주요하게 나타난 '프로그램'이라는 단어와 묶인 모습을 확인할 수 있었다.

정리하면, 문서 전반적으로 도서관이라는 단어가 주요하게 나타나고 있으며, 그 외 단어들

은 LDA 토픽 모델링 결과와 비슷하게 클러스터링이 진행되었으며, 이러한 유사함을 바탕으로 비교하자면 주제 1과 주제 2는 단어들의 출연 빈도적으로 유사하게 나타나는 것 또한 확인할 수 있었다. 이를 바탕으로, LDA를 통해 얻어낸 주제들에 대해서 단어들이 가진 출현 빈도에 따른 계층적인 구조에 큰 영향을 받았음을 알 수 있었다.

4. 결론 및 연구의 의의·한계

본 연구는 최근 5년간 문헌정보학에서 연구되고 있는 주제들을 토픽모델링을 통해 확인하고, 연도 별 주제에 대한 비율이 어떻게 변화하였는지를 LDA를 통한 토픽모델링을 통해 확인하고자 하였다.

연구 결과 최적의 연구 주제는 6개가 나왔고, 이 6개의 주제는 각각 도서관학, 정보이용연구, 기록관리학, 도서서비스, 디지털화 및 자료분류, 메타 연구 및 색인 연구인 것으로 확인되었다.

LDA를 통해 추론한 문서가 각 주제에 속할 확률 분포인 θ를 통하여 문서에서 나타나는 θ의 가중치를 확인하는 것으로 전체적으로 논문의 주제에 맞는 θ값이 크게 나타나는 것을 확인하였고, θ가 하나의 주제에만 편향되어 크게 나타나는 것이 아닌 비슷하게 나타나는 것 또한 확인하여 이를 문헌정보학이 가지는 융복합적인 특성으로 인한 것으로 해석하였다.

또한 문서들에서 나타나는 θ의 평균을 연도별로 구하여 주제별 θ값의 평균의 차이가 있는지 확인하기 위해서 일원분산분석을 사용하였고, 이를 통해 통계적으로 문서에서 나타나는 각 주제에 속할 확률들에 차이가 없는 것을 확인하였고, 모델 적합성 검사를 통해 분석 내용에 대해서 이상이 없는 것도 확인하였다. 이를 바탕으로 각 논문들이 어떠한 주제를 가질 확률은 동일하다고 판단하였고, 이 또한 문헌정보학이 가지는 융복합적인 특성으로 인한 것으로 해석하였다. 이를 바탕으로 연도별 평균 주제 가중치의 변화를 확인해본 결과 2020년에 도서관과 도서관의 디지털화에 대한 주제가 활발히 연구된 것을 확인하였고, 이를 코로나-19 확산 이후 언택트 시대로의 변화에 대응하기 위한 것임으로 해석하였다.

마지막으로 각 단어들의 문서별 출현 빈도를 통해 계층적 클러스터링을 진행하였고, 대체로 LDA 결과와 비슷하게 나타나는 것을 확인하였다. 이를 통해 LDA를 통해 각 단어들이 해당 주제들에 배정된 것이 출현 빈도에 따른 각 단어들의 계층적인 구조로 인한 것이라고 해석할 수 있었다.

본 연구는 다음과 같은 한계를 가진다. API에서 검색되지 않는 논문이나 API 내에서 초록이 미등재된 논문, 초록이 영문으로만 작성된 논문이 분석에서 제외되면서 전체 문헌정보학 분야에 대한 조사가 이루어지지 않게 되었다. 또한 위의 경우에 해당되는 논문들을 분석에서 제외하는 과정에서, 학술지별 투고한 논문의 비율이 변화하였고, 특히 한국문헌정보학회

지의 2021년도 논문을 모두 분석에서 제외하게 되면서 최근 5년간 문헌정보학의 연구 현황 정리라는 주요 연구 문제를 불완전하게 해결할 수밖에 없었다.

또한 R의 KoNLP 성능상 형태소 분해에 한계가 있고, NIADic이 많은 형태소를 분해할 수 있음에도 문헌정보학에서 주로 사용되는 학술용어를 분류하는데 한계가 있었기 때문에, 토픽모델링의 성능이 온전하게 나오지 않았다. 이를 NIADic에 논문 키워드를 추가하는 것으로 해결하고자 하였으나, Java 문제로 인하여 이 과정이 이루어지지 않아 불용어처리만으로 모델링 성능을 보완할 수 없었다.

반면, 본 연구가 가지는 의의는 다음과 같다. 기존 선행연구에서는 문헌정보학의 특정 분야에 대한 현황 분석만을 진행하거나 문헌정보학 논문 전체를 대상으로 어떤 주제를 가지고 있는지를 분석했다면, 본 연구에서는 최근 5개년동안 진행된 문헌정보학 분야의 연구를 대상으로 토픽 모델링을 진행하여 해당 기간동안 연구된 논문들의 주제를 분석하였다. 이 과정에서 기존 연구들보다 조금 더 최신의, 핵심적으로 연구가 진행되고 있는 주제들을 확인할 수 있었다. 따라서, 기존 선행연구보다 좀 더 최신 연구 트렌드에 맞는 주제들을 압축하여 제공하는 것으로 본 연구가 추후 문헌정보학 연구를 진행하는데 활용될 수 있을 것으로 기대된다.

또한 기존 연구들과는 달리 영문 텍스트 데이터를 사용하지 않고 한글 초록데이터를 바탕으로 토픽모델링을 진행하였다. 대부분의 논문들이 한글 초록과 함께 영문 초록도 함께 등록하고 있지만, 대부분의 논문들이 한글 초록데이터를 사용하였고, 대부분의 국내 연구들은 한글로 논문을 작성하고 있기 때문에 국내 연구동향 분석 연구에 새로운 방법론을 제시했다는 점에서 그 의미가 있을 것이다.

참고문헌

1. David M. Blei, Andrew Y. Ng, & Michael I. Jordan. (2003). Latent Dirichlet Allocation (pp. 993-1022). n.p.: Journal of Machine Learning Research vol.3
2. Lau, A. Y. and Coiera, E. W. (2007). "Do people experience cognitive biases while searching for information?" Journal of the American Medical Informatics Association, 14(5): 599-608.
3. Logan, D. W. et al. (2010). "Ten simple rules for editing Wikipedia." PLoS Comput Biol, 6(9): e1000941.
4. Pautasso, M. (2013). "Ten simple rules for writing a literature review." PLoS computational biology, 9(7): 1-4.
5. Webster, J. and Watson, R. T. 2002. "Analyzing the past to prepare for the future: Writing a literature review." Management Information Systems Quarterly, 26(2): 3.
6. 강필수, 노영희 and 김윤정. (2021). 스마트디지털 환경 기반 도서관 구축에 관한 사서 인식 연구. 한국비블리아학회지, 32(1), 5-33.
7. 국민청원 데이터 수집 및 토픽 모델링 . (2022). https://velog.io/@shawnk123/tautnfco.
8. 김동석, & 노영희. (2018). 빅데이터 분석 기법을 활용한 도서관발전종합계획 동향 분석 연구. *한국비블리아학회지*, *29*(2), 85-108.
9. 김양우. (2017). 학술정보포털에 대한 이용자만족 관련 인식에 관한 연구 - NAVER 전문정보의 학술자료 검색 기능을 중심으로 -. 한국문헌정보학회지 51.2, 255-279.
10. 김태경, 최회련, 이홍철. (2016). 토픽 모델링을 이용한 핀테크 기술 동향 분석. 한국산학기술학회 논문지, 17(11), 670-681.
11. 김희숙 and 장우권. (2020). 작품.주제형 디지털 인물 아카이브의 콘텐츠와 구성에 관한 연구 - 조선 중기 여류 문인을 중심으로 -. 한국문헌정보학회지, 54(1), 145-174.
12. 박자현, 송민. (2013). 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석. 정보관리학회지, 30(1), 7-32.
13. 박준형, 오효정. (2017). 국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법 비교 - LSA 와 HDP 를 중심으로 -. 한국도서관정보학회지, 48(4), 235-258.
14. 심새봄, 이용정. (2019). 의과대학 대학원생의 학술정보 검색엔진 이용 동기 및 이용자 만족도에 관한 연구. 한국비블리아학회지 30.4, 197-216.
15. 양명석, 이성희, 박근희, 최광남 and 김태현. (2021). LDA 토픽모델링을 활용한 인공지능 관련 국가 R&D 연구동향 분석. 인터넷정보학회논문지, 22(5), 47-55.
16. 윤다영 and 노영희. (2021). 이용자 인식조사를 기반으로 한 공공도서관 비대면 서비스 운영 방향에 관한 연구. 한국비블리아학회지, 32(4), 161-188.

17. 윤영조, 정연경. (2020). 교육지원청 기록관리 기관평가 개선방안에 관한 연구. 한국비블리아학회지, 31(3), 125-149.
18. 윤혜영. (2019). 대전지역 공공도서관 운영현황과 발전과제. (2019). 한국문헌정보학회지. 53(2), 69-90.
19. 이기헌, 정효정, 송민. (2015). 문헌정보학 분야 핵심 학술지들의 가중 주제-방법 네트워크 분석. 한국문헌정보학회지, 49(3), 457-488.
20. 이승길. (2019). 국가 수준의 도서관 계획을 기반한 학교도서관의 교육공동체 협력 방안. 한국문헌정보학회지, 53(2), 139-157.
21. 이재윤. (2018). 출판년도의 즉시 인용빈도를 포함하는 학술지 인용지수 개발. 한국문헌정보학회지, 52(4), 71-90.
22. 정도헌 and 주황수. (2018). 토픽 모델링 기반 내용 분석을 통한 학제 간 융합기술 도출 방법. 정보관리학회지, 35(3), 77-100.
23. 정연주, 이승민. (2020). 방송영상기록물의 저작자 식별을 위한 패싯 기반 식별체계 구축 연구. 한국비블리아학회지, 31(3), 213-234.
24. 정혜리. (2022). 대학 온라인 교육에 대한 온라인 커뮤니티 토픽 분석 : LDA 기반 토픽모델링을 활용하여. 서울: 성균관대학교 일반대학원.
25. 진설아, 송민. (2016). 토픽 모델링 기반 정보학 분야 학술지의 학제성 측정 연구. 정보관리학회지, 33(1), 7-32.