Journal of The Korean Association
of Information Education Vol. 26, No. 5,
October 2022, pp. 439-448

http://dx.doi.org/10.14352/jkaie.2022.26.5.439

LDA기반 토픽모델링을 활용한 공공데이터 기반의 교육용 데이터마이닝 연구

신승기
서울교육대학교 컴퓨터교육과

요약

본 연구에서는 공공데이터포털에서 제공하는 교육관련 데이터를 검색하고 토픽모델링 기법을 활용한 분류를 통해 어떠한 데이터의 종류가 구축되어 있으며 활용이 가능한지를 살펴보고자 하였다. 공공데이터포털의 데이터에 대하여 분류체계를 기준으로 교육분야의 파일데이터는 3,072건이 수집되었으며, 검색어를 활용하여 '교육'을 검색하여 나타난 파일데이터 2,361건으로 나타났다. 각각의 데이터셋에 대하여 불용어처리를 실시하고 데이터전처리를 수행하여 LDA기반 토픽모델링을 활용하여 텍스트마이닝 분석을 실시하였다. 사전에 교육으로 분류된 데이터셋에서는 현재 재학중인 학교급별 학생을 대상으로 지원하는 프로그램과 정보에 대한 내용이 제공되고 있었다. 한편, 교육으로 검색하여 수집된 데이터셋에서는 장애인, 학부모, 노인, 아동 등 평생교육의 관점으로 제공되는 교육 프로그램 및 지원현황이라는 특징이 나타났다. 데이터과학기반의 의사결정 및 문제해결력을 기르기 위해 공공데이터포털이 제공하는 데이터에서 교육과정 및 내용이 충분히 제공되는 것도 좋은 기회가 될 것이다.

키워드 : 공공데이터포털, 데이터마이닝, 텍스트마이닝, 토픽모델링, 데이터과학교육

A Study on Educational Data Mining for Public Data Portal through Topic Modeling Method with Latent Dirichlet Allocation

Seungki Shin
Department of Computer Education, Seoul National University of Education

Abstract

This study aims to search for education-related datasets provided by public data portals and examine what data types are constructed through classification using topic modeling methods. Regarding the data of the public data portal, 3,072 cases of file data in the education field were collected based on the classification system. Text mining analysis was performed using the LDA-based topic modeling method with stopword processing and data pre-processing for each dataset. Program information and student-supporting notifications were usually provided in the pre-classified dataset for education from the data portal. On the other hand, the characteristics of educational programs and supporting information for the disabled, parents, the elderly, and children through the perspective of lifelong education were generally indicated in the dataset collected by searching for education. The results of data analysis through this study show that providing sufficient educational information through the public data portal would be better to help the students' data science-based decision-making and problem-solving skills.

Keywords : Public Data Portal, Data Mining, Text Mining, Topic Modeling, Data Science Education

논문투고 : 2022-10-09
논문심사 : 2022-10-25
심사완료 : 2022-10-27

1. 서론

인공지능과 빅데이터를 기반으로 사회현상을 예측하고 문제해결과정에서 활용하고자 하는 디지털 대전환의 시대가 도래함에 따라 미래인재를 길러내고자 하는 교육의 모습도 변화되고 있다. 맥킨지(McKinsey) 보고서에 따르면 기업의 70%이상이 디지털 전환의 시도에서 실패하고 있으며 이는 기술의 단순한 활용과 덧붙음으로 발생하고 있음을 설명하고 있다[6][7]. 우리나라 정부에서 사회현상 변화의 가장 큰 흐름으로 디지털 대전환을 제시하고 있으며, 교육은 사회의 모습을 담고 있으며 미래인재를 길러낸다는 관점에서 디지털 전환에 따른 인재양성 방향을 국정과제와 연계하여 발표하였다[10]. 디지털 분야 인재양성의 배경으로 청소년 및 성인의 컴퓨터기반 문제해결력이 OECD 평균과 비교하여 낮은 수준임을 제시하고 있으며, 전 국민의 디지털 교육 기회 확대의 필요성과 역량을 강화하기 위한 방향을 제시하였다[10].

디지털 인재를 길러내기 위한 세부 내용 중의 하나로 교육과정에서 데이터과학을 비롯한 최신 인공지능 관련 과목을 편성하는 것을 제시하고 있으며, 2020년 7월 발표된 정책으로 포스트 코로나 시대를 준비하기 위한 '한국판 뉴딜'의 '디지털 뉴딜' 정책에서 제시하고 있는 D.N.A.(Data, Network, AI) 정책과 연계된 부분이라고 할 수 있다[4]. 특히, 데이터를 수집하고 분석하여 활용하기 위한 데이터댐을 구축하는 내용과 교육 인프라 구축을 통한 디지털 전환을 추진하기 위해 '교육용 데이터댐' 구축에 대한 내용이 제시되었다[4]. 이는 학교 교육에서 사회현상에 대한 데이터를 수집하고 분석하여 컴퓨팅사고력을 기반으로 문제해결력을 기르는 교수학습 방법으로서의 내용을 제시하는 의미를 갖고 있다. 아울러 우리나라의 교육체제를 디지털 기반의 의사결정 및 학교단위 특성을 반영한 교육과정을 체계적으로 구성하기 위한 정책적 의사결정에서도 데이터를 활용한다는 방향성을 설명하고 있는 것이다.

우리나라 정부에서는 데이터의 중요성을 인식하고 정부 기관에서 보유하고 있는 데이터를 공공의 영역으로 확대하여 원하는 누구나 데이터에 접근하고 활용할 수 있도록 이른바 공공데이터법이라고 불리는 "공공데이터의 제공 및 이용 활성화에 관한 법률(제11956호)"을 통해 개방하고 있다. 공공데이터법에 따르면 공공데이터의 제공 및 이용 활성화에 관한 법률이 제정된 목적과 원칙으로 모든 국민이 공공데이터를 이용할 수 있는 권리를 보장하여 국민경제 발전에 기여하며 보편적 접근과 평등한 데이터 활용의 기회를 제공하는데 있음을 설명하고 있다[5].

교육분야에서도 데이터를 활용한 인공지능 교육의 교수학습방법에 대한 내용과 함께 데이터를 기반으로 교육정책을 수립하고 학습자 개별화 교육을 추진하기 위한 초개인화 학습환경을 마련하기 위해 데이터를 활용하는 방향을 구체적으로 제시하고 있다[3]. 정부에서는 인공지능시대의 핵심과제를 정리하여 대한민국의 미래교육에 대한 방향을 제시하였으며, 모두를 위한 공평한 교육 기회를 제공하기 위해 디지털 도구를 활용하여 데이터를 수집하고 가공함으로서 교육 빅데이터를 기반으로 새로운 시대에서 인공지능을 활용한 개별화 교육을 추진하고 있다[3].

교육부는 2022 개정교육과정 총론 시안을 통해 기존의 언어 소양, 수리 소양과 더불어 디지털 소양을 기초 소양으로 정의하고 총론 및 교과에 관련 내용을 반영하도록 하였다[8]. 특히, 초·중·고 학생들의 디지털 및 인공지능 소양 함양 교육을 강화하기 위해 학교급별 내용 기준을 마련하여 정보교과에서 기초 소양을 함양하고 모든 교과에서 이를 활용할 수 있는 인공지능 및 데이터 활용의 방향을 제시하고 있다[8]. 아울러 교수학습방법을 개선하고 학습자 개별 맞춤형 수업을 제시하기 위해 빅데이터 기반의 학습자 진단과 처방에 이르는 일련의 과정을 제시하고 있다[8].

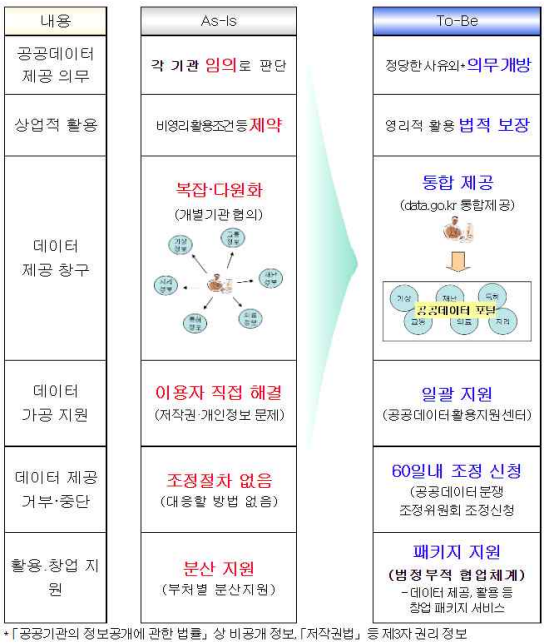
사회의 변화는 교육의 방향과 내용에 영향을 미치며 국가적으로 길러내고자 하는 인간상에도 밀접한 관련성을 갖고 있다. 인공지능 중심의 사회에서 예측을 통한 적극적인 미래의 대비는 데이터를 근간으로 수행되는 부분이라는 점에서 교육의 영역에서도 두 가지의 관점으로 고려해볼 필요가 있다. 첫 번째는 학습자의 데이터를 활용하는 역량을 길러주는 활동으로, 학생들이 문제를 해결하기 위해 데이터를 수집하고 분석하여 적절한 의사결정을 할 수 있도록 하는 것이다. 두 번째는 교사 및 연구자들의 교육데이터를 활용한 인사이트를 발굴하고 새로운 교육의 방향을 제시하는 관점이다. 그러나 교육데이터는 민감 데이터가 대다수라는 점에서 공식적으

로 공개되어 있는 데이터 이외에는 데이터 분석의 결과 및 아이디어를 공개적으로 공유하거나 활용하는데 제한이 따른다는 어려움이 있다. 따라서 본 연구에서는 정부에서 공공데이터법을 기반으로 모든 국민이 활용할 수 있도록 공유하고 있는 공공데이터포털에서 제공하는 교육관련 데이터를 검색하고 토픽모델링 기법을 활용한 분류를 통해 어떠한 데이터의 종류가 구축되어 있으며 활용이 가능한지를 살펴보고자 하였다.

2. 이론적 배경

2.1. 공공데이터 이용과 활용의 법률적 근거

공공데이터법이라고 불리는 "공공데이터의 제공 및 이용 활성화에 관한 법률"과 동법 시행령 및 시행규칙은 2013년에 제정되어 시행이 시작되었다[11]. 공공데이터법 은 2012년부터 구체화되어 2013년에 발표된 정부 3.0의 추진의 일환이며 빅데이터와 클라우드 컴퓨팅 등 최신의 기술을 활용하여 공공의 문제를 해결하기 위함이다[11].

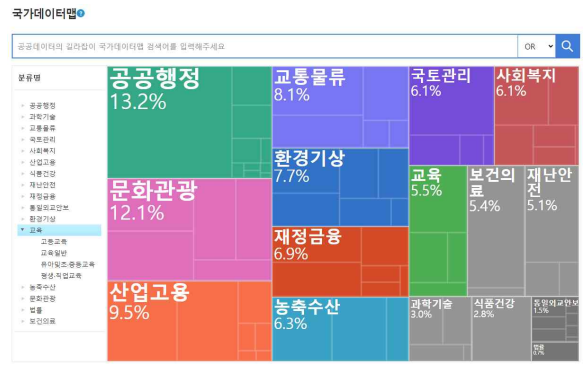


(Fig. 1) Comparison before and after the enforcement of the Public Data Act[11]

공공데이터법의 시행에 따라 변화되는 모습에 대하여 당시 안전행정부에서는 (Fig. 1)과 같이 데이터의 제공과 활용 범위에 대하여 설명하였다[11]. 공공데이터법이 시행되기 이전에는 정부 부처별로 판단하여 공개가 필요한 데이터의 범위와 종류를 설정하였고 사용자가 별도로 요청하여 데이터를 제공받아야 하였으나, 공공데이터법의 시행으로 공공데이터포털(https//www.data.go.kr)을 통해 데이터를 개방 및 제공하고 영리적 활용을 보장하며 공공데이터활용지원센터를 통해 데이터 가공 절차를 지원하도록 하였다[11].

2.2. 공공데이터포털과 교육

공공데이터포털에서는 공공기관이 보유하는 데이터를 국민들에게 개방하고 공유하는 통합플랫폼으로 운영되고 있으며, 파일데이터(CSV, XLS 등)의 형태와 오픈API의 형태 등으로 제공하여 필요한 데이터를 쉽게 활용할 수 있도록 지원하고 있다[13]. 공공데이터포털의 설치와 운영에 대한 근거로는 공공데이터법의 제21조(공공데이터포털의 운영)에 근거하여 행정안전부에서 운영의 책임을 맡고 있다[5][13]. 2022년 10월 1일 기준 보유하고 있는 검색 가능한 데이터 목록으로 총 72,997건이 제공되고 있으며, 파일 데이터 55,113건, 오픈API 9,501건, 표준데이터셋147개(8,383건)이 사용가능하다. 공공데이터포털에서 보유하고 있는 데이터는 국가데이터맵으로 16개의 분류체계를 기반으로 제공하고 있으며, 교육은 4개의 세부영역으로 구성되어 있다.



(Fig. 2) Data Classification according to the National Data Map[14]

인공지능 시대에의 교육의 내용과 방법에 대한 변화의 요구가 높아짐에 따라 2022 개정교육과정에서도 정보교육을 강화하여 데이터를 활용한 예측과 분류 등의 내용을 반영한 컴퓨팅사고력 기반의 인공지능교육이 교육과정에 반영되고 있다[9]. 특히 데이터를 수집하고 가공하여 예측하고 분류하는 데이터과학교육이 확대되고 있으며, 교과융합의 사례와 정부부처 및 공공기관에서 추진하고 있는 다양한 데이터활용대회의 수가 늘어남은 이를 설명하는 부분이라고 할 수 있다. 데이터과학교육을 위해 가장 먼저 해결이 필요한 부분은 학습용 데이터 수집과 발굴이다. 일반적으로 단위학교에서 수행되는 데이터과학교육의 학습용 데이터는 공공데이터포털을 이용하게 된다는 점에서 본 연구에서는 교육에서 활용할 수 있는 공공데이터포털의 데이터를 분류하고 특징을 살펴보고자 하였다.

3. 연구목적 및 연구방법

본 연구의 목적은 공공데이터 포털에서 제공하고 있는 데이터를 분석하여 주제별 분류를 통해 데이터과학교육을 위해 사용가능한 교육관련 데이터를 살펴보고 시사점을 제시하는데 있다. 이를 위해 토픽모델링 기반의 텍스트마이닝 기법을 활용하여 토픽을 발견하고 잠재적 의미를 살펴봄으로서 교육관련 데이터가 공공데이터포털에 공유되고 있는 현황을 살펴보았다. 토픽모델링의 기법 중에서 본 연구에서는 용어가 문서내 및 문서간 제시되는 출현 확률로 의미를 찾을 수 있게 하는 잠재 디리클레 할당(Latent Dirichlet Allocation: LDA)을 사용하였다 LDA기법은 토픽모델링을 수행하기 위한 방법 중 하나로서 단어의 출현 빈도를 기반으로 주제를 추출하는 잠재의미문석(LSI) 기법이 출현 확률을 기반으로 의미를 찾게 하는 확률적 잠재 의미 인덱싱(PLSI)로 대체 되었으며, 이를 일반화하여 제시한 모형이다[1]. 특히, 텍스트 데이터를 분석하는 자연어처리의 방법으로서 비지도학습을 통해 확률적인 출현 빈도를 기반으로 비율을 산출하여 토픽이 추출되는 방법이다[1][2].

공공데이터포털에서 교육관련 데이터의 보유 현황 및 주제를 추출하기 위하여 두 가지 방법을 활용하여 데이터를 수집하였다. 첫 번째는 공공데이터포털에서 제공하는 16개의 데이터의 분류 체계 중에서 교육에 해당하는 데이터를 수집하고 파일데이터 형태로 제공하고 있는 데이터에 대해서 제목과 내용의 텍스트 데이터를 분석하는 것이다. 두 번째는, 공공데이터포털에서 제공하고 있는 키워드 검색 기능을 활용하여 '교육'이라는 용어로 데이터를 찾고 파일데이터로 제공하는 자료들의 제목과 내용을 수집하여 텍스트마이닝을 수행하는 것이다. 공공데이터포털에서 제공하는 자료를 수집하는 기준은 연구가 수행되고 있는 2022년 10월 1일을 기준일로 정하여 자료를 수집하고 분석하도록 하였다.

공공데이터포털을 통해 수집된 교육관련 데이터들은 KoNLpy를 활용하여 형태소 분석을 수행하여 텍스트데이터 전처리를 실시하였다. LDA기법에서 필요한 디리클레 파라미터(Dirichlet Parameter)로서 $\alpha = 0.1$, $\beta = 0.01$로 값을 설정하였으며, 토픽의 개수(K)는 15개로 설정하여 주제를 추출하고 상위 10개의 유의미한 토픽에 대한 세부 토픽의 내용과 의미를 분석하였다. 토픽 추출을 위한 반복의 횟수는 1000번으로 설정하였으며, perplexity가 감소하여 안정적으로 수렴할 수 있도록 제시되었다.

4. 연구결과

본 연구에서는 공공데이터포털에서 교육관련 데이터를 분류체계 및 검색어 활용 등 두 가지 방법으로 수집하였으며 아래의 <Table 1>과 같이 데이터가 추출되었다. 분류체계를 기준으로 교육분야의 데이터는 총 4,155건으로 파일데이터 3,072건, 오픈API 466건, 표준데이터셋 147개(617건)가 수집되었다. 검색어를 활용하여 '교육'을 검색하여 나타난 전체 데이터는 2,953건으로 파일데이터 2,361건, 오픈API 273건, 표준데이터셋 2개(319건)가 수집되었다. 각각의 데이터 수집방법에 대하여 파일데이터를 기준으로 토픽모델링을 수행하였으며, 인덱스 데이터인 제공기관, 수정일, 조회수, 다운로드, 키워드 등의 용어는 불용어처리하여 데이터를 전처리하였다.

<Table 1> Data Search and Collection from Data Portal

Source	Total	File Data	Open API	Data Set
Category	4,155	3,072	466	147
Search	2,953	2,361	273	2

4.1. 분류체계 활용 토픽모델링 분석 결과

공공데이터포털에서 제시하고 있는 분류체계를 활용하여 교육분야의 데이터를 수집하여 3,072개의 파일데이터를 대상으로 LDA기반 토픽모델링을 수행한 결과는 아래의 <Table 2>와 같이 나타났다. 추출된 10개의 토픽은 모두 교육과 관련된 분야의 내용을 담고 있으며, 현황과 정보제공의 내용이 구성된 것을 살펴볼 수 있다.

Topic 1에서는 연구와 강좌 등의 운영에 대한 내용들과 함께 역량 강화를 위한 교과별 강사 및 학업과 도서에 대한 현황으로 제시된 것을 살펴볼 수 있다. Topic 2에서는 홈페이지를 통해 제공되고 있는 학습 프로그램과 평가에 대한 내용과 함께 프로그램의 개발과 등록 및 직업 교육과 관련된 내용과 시설의 내용들이 나타난 것을 살펴볼 수 있다. Topic 3에서는 플랫폼을 활용하여 정책과 관련된 정보를 제공하거나 교육 현황을 관리하기 위해 프로그램의 시작과 위치 및 종합 정보 등이 제시되어 있다. Topic 4에서는 지역별 책에 대한 정보와 어린이를 위한 행사와 프로그램 입학 및 선발의 내용과 더불어 관련된 분야의 재원과 제공 형태 및 성취에 대한 내용이 제시되어 있다. Topic 5에서는 수치데이터에 대한 내용들이 제시되어 있으며, 시스템에서 제공하고 있는 코드와 인원 및 공시되어 활용되고 있는 현황 및 지역별 문화 프로그램에 대한 내용과 결과를 비슷한 검색 현황이 제시되어 있다. Topic 6에서는 어린이집과 급식 및 연수원 등 지원체제에 대한 내용을 나타났으며, Topic 7에서는 학생과 학급을 대상으로 동아리와 콘텐츠의 내용과 대상 등 정보가 제공되고 있다. Topic 8에서는 학교와 도서관 및 대학 등 기관별 과정과 사업에 대한 내용이 제시되었으며, Topic 9는 온라인 기반의 강의에 대한 내용들이 학사, 설문, 참여, 컴퓨터 등의 키워드와 함께 나타났음을 살펴볼 수 있다. Topic 10은 청소년을 지원하기 위한 제반 환경에 대한 내용들이 키워드로 도출되었다.

<Table 2> Result of Topic Modeling with LDA Method about Categorical Educational Data from Public Data Portal in South Korea

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
연구	제공	관리	책	수	자격	학생	학교	항목	지원
0.0941078	0.10768	0.0891019	0.0435528	0.104224	0.0449381	0.093145	0.106203	0.0808267	0.0655799
운영	학습	시작	지역	시스템	정원	내용	도서관	온라인	과목
0.0799918	0.0571456	0.0628078	0.0321398	0.0513269	0.0290378	0.0791735	0.0797521	0.0421194	0.0364832
도서	홈페이지	위치	재원	코드	지원청	분류	과정	훈련	일자
0.0723026	0.0268103	0.031449	0.0315391	0.0357459	0.0285769	0.0420887	0.0689168	0.0410963	0.0320522
강좌	프로그램	여부	일정	인원	전문	대상	대학	강의	청소년
0.0406277	0.0251944	0.0250431	0.0285356	0.0331272	0.0221246	0.0336368	0.0444576	0.0230215	0.0295413
자치	평가	개선	어린이	공시	어린이집	기간	목록	소장	계획
0.0329385	0.0248272	0.0217431	0.0237301	0.0301158	0.0214332	0.0265648	0.0376855	0.0199522	0.0236333
강사	개발	정책	입학	활용	시민	학급	기관	저자	환경
0.0183634	0.0227705	0.0207725	0.0231295	0.0216052	0.0182071	0.0207002	0.0372871	0.017565	0.0177254
역량	등록	플랫폼	선발	결과	보유	콘텐츠	교육부	학사	스마트
0.0168715	0.0219626	0.0201902	0.0165219	0.0195103	0.0177462	0.017078	0.0291606	0.0143251	0.0172823
교과	직업	종합	행사	지역	급식	동아리	사업	설문	장소
0.0128547	0.0182166	0.0190255	0.0138188	0.0154514	0.0138287	0.0169055	0.018644	0.0131315	0.0155099
공공	시설	소재	성취	문화	체제	수강	조사	참여	구성
0.0123957	0.0179962	0.0151431	0.0117164	0.0136183	0.0117548	0.016733	0.0184846	0.0127905	0.0149191
학업	교육원	진로	형태	검색	연수원	영문	센터	컴퓨터	특성
0.0096413	0.0178493	0.0139784	0.0102147	0.012309	0.0110634	0.016388	0.0156961	0.0117674	0.0112266

4.2. 검색어 활용 토픽모델링 분석 결과

공공데이터 포털에서 기존의 분류체계 이외에도 키워드 검색이 가능하며, 교육분야 이외의 사회영역에서 교육과 관련된 데이터를 제공하고 있음을 고려하여 키워들 “교육”을 활용하여 검색하여 도출된 2,301개의 파일 데이터에 대해 LDA기반 토픽모델링을 수행하여 아래의 <Table 3>과 같이 결과를 얻을 수 있었다.

Topic 1에서는 평생교육의 관점으로 도서관과 보건 및 장애인을 지원하는 교육에 대한 내용으로 도출되었으며, Topic 2에서는 기관에서 제공하고 있는 교육훈련의 유형과 장소 및 기간 등에 대한 내용이 제시되었다. Topic 3에서는 연수원과 교육부 및 공단에서 제시하고 있는 연수의 인원과 정책 등의 내용으로 도출되었으며, Topic 4에서는 학부모와 어린이를 대상으로 제공되고 있는 맞춤형 프로그램의 내용과 성취도에 대한 내용과 함께 취업과 급여지원에 대한 내용도 나타났으며 산업 분야별 지원되고 있는 현황이 나타났다. Topic 5에서는 안전 관리를 위한 프로그램과 평가 및 인재양성의 내용과 서비스 등의 내용이 제시되었으며, Topic 6에서는 지역별 학생과 여성을 지원하기 위한 프로그램과 행정적 지원의 내용으로 도출되었다. Topic 7에서는 교육관련 연구와 체험의 내용을 홈페이지를 통해 제공하는 보고서와 주기적 현황의 내용이 나타났으며, Topic 8에서는 데이터와 통계의 내용으로 교육관련 시설과 대상 및 온라인으로 제공하고 있는 재정 현황과 기술 목록 등에 대한 자료 제공의 내용이 제시되었다. Topic 9는 센터에서 운영하는 교육과정 및 시험 등 정보 제공의 내용들이 제시되었으며, Topic 10에서는 교육관련 사업의 내용으로 아동과 노인을 지원하고 일자리와 연계된 재원과 자원의 내용에 대한 정보가 도출되었다. 검색어를 활용하여 도출된 데이터를 토픽모델링으로 분석하여 도출된 데이터는 대체로 정보전달 및 현황에 대한 내용과 각종 프로그램 내용으로 구성됨을 살펴볼 수 있다.

<Table 3> Result of Topic Modeling with LDA Method about Educational Search Data from Public Data Portal in South Korea

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
운영	정보	인원	산업	과정	학습	연구	데이터	센터	사업
0.098505	0.207521	0.0724254	0.0476019	0.0864359	0.0462129	0.0616676	0.0701217	0.0686299	0.0655435
평생	기관	공단	신청	자료	직업	체험	통계	기본	강좌
0.0923935	0.0596384	0.0507534	0.0350223	0.0711546	0.034776	0.0393406	0.0596739	0.0364537	0.0628229
지원	내용	정책	취업	안전	지역	홈페이지	시설	문화	재원
0.0722615	0.0391725	0.0397803	0.0258426	0.0626783	0.0338487	0.03748	0.0380751	0.0351036	0.030918
보건	훈련	교육부	급여	관리	구분	일반	대상	직종	아동
0.0605777	0.0354314	0.0299045	0.0217627	0.0414278	0.0324577	0.024456	0.0293352	0.0245282	0.0187991
시스템	청소년	진행	과제	프로그램	학생	생활	온라인	접수	일자리
0.0393671	0.0218975	0.0268869	0.0204028	0.0385625	0.0296757	0.0241902	0.0261205	0.020028	0.0187991
과학	국가	검사	참여	평가	계획	초등학교	재정	보호	건강
0.0388279	0.0211272	0.0260639	0.0190428	0.0230425	0.0264301	0.0199374	0.0220016	0.019803	0.0183045
분류	조사	연수	성취도	대학	여성	시행	연구	공개	농업
0.0275036	0.0185965	0.0246923	0.0180228	0.0187446	0.0239573	0.0170137	0.019892	0.0162029	0.0175625
장애인	유형	목적	어린이	기초	능력	보고서	목록	과정	노인
0.0170781	0.0156256	0.0167368	0.012243	0.0145662	0.0160751	0.0146215	0.019691	0.0162029	0.0153366
사이버	장소	창업	맞춤	서비스	행정	점검	기술	교육장	자원
0.0133033	0.0139752	0.0164624	0.011223	0.0136111	0.0132931	0.0143557	0.0184855	0.0119277	0.0148419
도서관	기간	연수원	학부모	양성	실적	주기	등록	시험	수준
0.0095285	0.0138651	0.0139935	0.0108831	0.0108652	0.0119021	0.0138241	0.0153713	0.0112527	0.0148419

5. 데이터과학교육을 위한 교육용 공공데이터 분석

분류체계를 활용하여 교육분야 파일데이터 3,072건에 대해 LDA기반 토픽모델링을 수행하여 10개의 토픽을 추출하여 살펴본 결과, 교육분야 연구 및 프로그램의 내용과 일정 및 지원체계를 비롯하여 학생과 학교의 현황에 대한 내용들이 주요 토픽으로 추출되었다. 추출된 10개의 토픽과 관련 키워드를 중심으로 살펴볼 수 있는 내용은 아래의 <Table 4>와 같다.

<Table 4> Data Index from Categorical Educational Data

Topic	Abstracted Data Index from Public Data Portal
1	연구와 강의 및 학업지원을 위한 공공의 데이터와 교과별 강사 및 도서 관련 내용
2	홈페이지를 통해 제공되고 있는 직업교육 관련 프로그램과 평가에 대한 내용
3	플랫폼을 활용하여 정책관련 정보 제공 및 교육 현황 관리에 대한 정보
4	지역별 책에 대한 정보와 어린이를 위한 행사와 프로그램 참여 정보 제공
5	교육관련 시스템에서 제공하고 있는 코드와 인원 및 공시정보 등 수치데이터 중심 자료 제공
6	교육지원체제로서 어린이집과 급식 및 연수원에 대한 현황과 정보 제공
7	학생과 학급을 대상으로 운영되고 있는 콘텐츠와 내용에 대한 현황 제공
8	교육관련 기관에서 제공하고 있는 각종 사업 및 프로그램에 대한 내용
9	온라인 기반 강의 운영 및 편성을 위한 학사일정 및 참여방법 등의 내용 제공
10	청소년을 지원하기 위한 환경제공과 구성에 대한 일정과 운영 방법에 대한 내용

<Table 4>에서 제시된 내용과 같이 분류체계를 통해 제공되고 있는 교육분야 파일데이터에서는 다양한 교육프로그램에 대한 운영 실태와 현황에 대한 내용들이 주로 구성되어 있음을 살펴볼 수 있다. 1번 주제와 4번 주제를 통해 도서관 및 책에 대한 데이터가 상당히 제공되고 있음도 알 수 있었다. 공공데이터로서 제공되는 데이터들은 주로 온라인을 통해 수집되거나 제공 및 정리된 내용임을 고려해볼 때 2번, 3번, 9번을 비롯한 많은 주제에서 온라인 환경에서 제공되는 정보들이 공공데이터로서 구성되고 있음을 알 수 있다.

모든 분류체계에 대하여 “교육”을 키워드로 검색하여 수집된 파일데이터 2,361건에 대해 LDA기반 토픽모델링을 수행하여 10개의 토픽을 추출하여 살펴본 결과, 다양한 분야에서 실시되는 교육 현황과 지원체계에 대한 내용 및 프로그램으로 제시되고 있음을 살펴볼 수 있었다. 10개의 토픽을 추출하였으며 키워드를 분석하여 살펴볼 수 있는 내용은 아래의 <Table 5>와 같다.

<Table 5> Data Index from Educational Search Data

Topic	Abstracted Data Index from Public Data Portal
1	평생교육의 관점으로 도서관과 보건 및 장애인을 지원하는 교육 현황 내용
2	국가의 기관별로 제공하고 있는 교육훈련의 유형과 장소 및 기간에 대한 현황 내용 제공
3	교육부와 연수원 및 공단에서 제시하고 있는 연수 현황 및 정책에 대한 정보 제공
4	학부모와 어린이 대상의 맞춤형 프로그램 및 취업과 급여지원 등 산업분야별 지원 현황
5	안전관리 프로그램의 현황과 평가 및 인재양성 지원을 위한 내용 제공
6	지역별 학생과 여성을 지원하는 프로그램과 행정적 지원 체제의 내용
7	분야별 교육관련 연구와 체험의 내용의 보고서와 주기적 현황 자료 제공
8	교육관련 시설과 대상 및 온라인으로 제공하고 있는 재정 현황의 데이터와 통계 자료 제공
9	센터에서 운영하는 교육과정 및 시험 등 정보 제공의 내용과 현황
10	교육관련 사업의 내용으로 아동과 노인을 지원하고 일자리 관련 재원과 지원의 내용 제공

교육으로 검색하여 공공데이터포털에서 수집된 데이터를 분석한 <Table 5>의 결과와 같이 대부분의 데이터가 분야별 교육관련 현황에 대한 내용으로 구성되어 있음을 살펴볼 수 있다. 주제 1번부터 4번까지는 모두 기관의 특성에 따라 제공하고 있는 교육 현황 등 프로그램의 자료로 구성되어 있는 내용이며, 주제 5번부터 10번까지는 대상에 따른 교육지원을 분야별로 구성하고 있는 현황과 자료를 제시하고 있음을 살펴볼 수 있다.

<Table 4>는 현재 학교급별 학생을 대상으로 지원하는 프로그램과 정보라고 한다면 <Table 5>는 장애인, 학부모, 노인, 아동 등 평생교육의 관점으로 학생을 제외한 대상의 현황이라는 특징이 있으며, 공통점으로는 대체로 프로그램 및 지원 현황에 초점을 두고 있다.

6. 공공데이터 활용 데이터과학교육 방안

본 연구에서 살펴보고자 한 것은 공공데이터 포털에서 제공하고 있는 데이터를 활용하여 데이터과학교육에서 활용할 수 있는 시사점을 제공하는데 있다. 공공데이터포털은 이른바 공공데이터법을 통해 개설 및 운영이 시작된 것으로 손쉽게 다양한 데이터를 검색하고 활용할 수 있는 환경을 정부에서 공식적으로 제공하는 시스템이라고 할 수 있다. 그러나, 공공데이터포털에서 제공되는 데이터는 정부기관 및 관련부처에서 공급자 중심으로 제공되는 환경이라는 점에서 데이터를 사용하는 관점에서는 기대하는 데이터와 실제 제공되는 데이터가 다를 수 있다는 점이 고려되어야 한다. 교육현장에서 교육관련 데이터를 검색하고 수집하여 활용하는 상황에서 공공데이터포털의 공공데이터를 활용하는 방안을 제시하면 다음과 같다.

첫째, 공공데이터포털에서 제공하는 교육분야 데이터에 대한 활용 방안이다. 공공데이터 포털에서 제공하고 있는 16개의 분류체계 중 하나로서 교육관련 데이터는 학생과 학교의 현황 및 교육분야 연구와 프로그램의 내용과 일정 등 지원체계에 대한 내용으로 제시되어 있음을 살펴볼 수 있다. 따라서, 교육기관에서 제공되는 프로그램에 대한 현황 및 교육관련 현황을 살펴보게 되는 경우에는 교육분야의 데이터에서 키워드 검색을 통해 교차분석의 형태로 활용할 수 있다. 다만, 교육의 현황과 세부 운영 실태에 관련된 자료는 제한적이라는 점을 고려하여 별도로 데이터를 수집하거나 데이터를 요청하여 수집할 수 있는 방안이 고려되어야 한다.

둘째, 공공데이터포털이 제공하는 모든 분류체계에 대한 교육데이터의 활용 방안이다. 교육은 모든 영역에서 필수적인 부분이라는 점에서 검색 결과를 통해 분야별 제공되는 교육현황과 지원체계의 내용이 제시되어 임음을 살펴볼 수 있다. 특히, 교육을 통해 사회구성원들에 대한 지원을 지자체 및 관련기관별로 운영하고 있으며 학교교육에서 소외될 수 있는 대상에 대한 현황이 대부분이라는 점을 고려하였을 때, 교육의 현황이라는 관점보다는 보편적 기회제공과 여건을 마련하기 위한 사회구성원 단위별 교육프로그램의 종류에 대한 정보라는 점을 고려하여 데이터를 수집하고 활용하는데 참고해야할 것이다.

7. 결론 및 제언

디지털 인재양성을 위한 국가정책 및 교육과정에서는 정보교육을 통한 정보소양 함양 및 컴퓨팅사고력을 기반으로 문제해결력을 기르기 위한 활동으로 구성되어 있으며, 소프트웨어(SW)교육과 인공지능(AI)교육을 주요 내용으로 제시하고 있다[8][9]. 인공지능교육을 위해 최근 데이터과학교육의 중요성이 확대되고 있으며 빅데이터를 기반으로 다양한 문제상황에서 문제해결을 통해 인사이트를 제시하고 사회공공의 영역에서 눈에 보이지 않는 문제를 발굴하여 해결할 수 있는 역량을 기르는 활동들이 안내되고 있다. 이는 교육부에서 추진하고 있는 교육공공데이터대회가 올해들어 4회를 맞이하며 더욱 확대되고 있는 모습과 일치한다고 할 수 있으며, 많은 정부부처와 공공기관에서도 공공데이터를 활용하는 대회가 열리고 있는 취지이기도 하다.

교육에서 데이터를 활용하는 목적은 크게 2가지로 분류할 수 있다. 첫 번째는 교육에서 정책을 입안하고 제언을 마련하기 위해 교육데이터를 활용하여 분석하고 방향을 제시하는 부분이다. 두 번째는 학교에서 인공지능교육을 위해 데이터과학교육을 실시하는 경우 교육데이터를 활용하여 학교와 교육에 대한 실제적인 데이터를 분석하여 문제해결력을 기르고 예측하는 활동이 수행되는 것이다. 본 연구에서는 두 번째 목적인 데이터과학교육을 위해 학습용 데이터를 수집하기 위하여 교육데이터를 발굴하는 과정에서 공공데이터포털을 이용하여 관련 데이터를 검색하고 수집하게 되는 상황을 중점적으로 검토하였다. 특히, 공공데이터포털에서 분류체계를 통해 제시하고 있는 교육데이터를 살펴보고 어떤 주제의 데이터가 제공되고 있는지를 분석함으로서 데이터과학교육을 하게 되는 경우 검색어 사용에 대해 예측가능한 주제를 사전에 파악하고 데이터의 종류와 현황을 알아보기 위해 본 연구가 수행되었다. 분류체계에서 제시되는 교육데이터 뿐만 아니라 검색어로 “교육”을 활용하여 도출되는 데이터의 현황에 대해서도 특징을 분석하고자 하였다. 분류체계를 기준으로 수집된 교육분야의 파일데이터는 3.072건이었으며, 검색을 통해 수지된 파일데이터는 2,361건이었다. 수집된 데이터는 텍스트데이터 전처리를 통해 텍스트 마이닝이 실시되었으며, 이를 위해 LDA기반 토픽모델링이 수행되었다.

학교현장에서 데이터과학교육을 위해서 수집하게 되는 데이터의 상당수는 공공데이터포털을 이용하게 된다. Github와 같이 세계적으로 널리 사용되는 다양한 플랫폼들이 있지만, 구성주의 관점에서 우리 주변의 실제적인 데이터를 수집하고 분석을 통해 문제해결의 아이디어를 제시하는 과정에서 평소 생활과 관련이 있는 데이터를 수집하게 된다[15]. 이와 같은 관점에서 실제 속해 있는 사회의 구성원으로서 데이터를 수집하고 우리말로 적혀진 데이터를 수집하여 평소 생활과 관련성을 높일 수 있으며, 실제적 문제해결이 가능하다는 점에서 정부에서 제공하고 있는 공공데이터포털에서 데이터를 주로 수집하는 경향이 있다. 특히, 디지털인재양성전략 및 정보교육과 연계하는 과정에서 교육현장에서 데이터역량을 기르는 활동으로 적용된다는 점에서 교육관련 데이터를 찾고 활용하게 된다. 본 연구에서는 공공데이터포털에서 제공하고 있는 교육관련 데이터를 분석하고 구성된 형태를 살펴봄으로서 앞으로 데이터과학교육에서 교육관련 데이터를 수집하게 될 때 고려할 수 있는 데이터의 유형을 미리 살펴볼 수 있다는 점에서 연구의 의미가 있다고 할 수 있다.

공공데이터 포털에서 제공하고 있는 분류기준에서 교육분야의 데이터를 살펴보았을 때 현재 재학중인 학교급별 학생들을 지원하는 프로그램의 구성과 현황에 대한 내용들이 대부분이었음을 살펴볼 수 있었다. 특히 교육정책 및 운영되고 있는 교육현황과 제공되는 콘텐츠에 대한 정보가 대부분의 자료임을 알 수 있다. 한편, 검색어를 활용하여 “교육”을 키워드로 제시하여 수집된 데이터를 살펴보았을 때 수집되는 데이터는 평생교육의 관점으로 아동, 노인, 학부모, 장애인을 지원하는 프로그램에 대한 현황이 나타났음을 알 수 있었다. 두 가지 방법으로 검색되어 수집된 데이터 모두 프로그램 및 교육지원의 환경적 측면에서 현황을 제공하고 있다는 공통점이 있지만, 교육의 실제적인 내용이라고 할 수 있는 교육과정 및 학교운영에 대해서는 소수의 사례만 나타남을 살펴볼 수 있었다. 따라서, 공공데이터포털에서 제공되는 현황을 사전에 숙지하고 필요한 데이터를 찾는 것도 방법이지만, 공공데이터포털이 갖는 의미를 높이고 데이터과학기반의 의사결정 및 문제해결력을 기르기 위해 교육과정 및 내용이 충분히 제공되는 것도 좋은 기회가 될 것이다.

참고문헌

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022

[2] Kim, K., Chun, S., Koo, D., Shin, S. (2021). A Trend Analysis of Computer Education based on SNS Data through Data Mining Analysis. Journal of The Korean Association of Information Education, 25(2), 289-300.

[3] Korean Government(2020). Educational policy direction and key tasks in the era of artificial intelligence. Retrieved from https://www.korea.kr/archive/expDocView.do?docId=39237

[4] Korean Government(2020). Korean version of the New Deal Comprehensive Plan. Retrieved from https://www.korea.kr/news/pressReleaseView.do?newsId=156401053

[5] Legislative Assembly of South Korea (2020). Act on the Provision and Use of Public Data. Retrieved from https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EA%B3%B5%EA%B3%B5%EB%8D%B0%EC%9D%B4%ED%84%B0%EB%B2%95

[6] McKinsey & Companay(2018). Unlocking success in digital transformations. Survey result. Retrieved from https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/unlocking-success-in-digital-transformations

[7] McKinsey & Companay(2019). Why do most trasnformations fail? A conversation with Harry Robinson. Retrieved from https://www.mckinsey.com/capabilities/transformation/our-insights/why-do-most-transformations-fail-a-conversation-with-harry-robinson

[8] Ministry of Education, South Korea(2021). 2022 Revised National Curriculum. Retrieved from https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&boardSeq=89671&lev=0&searchType=null&statusYN=W&page=1&s=moe&m=020402&opType=N

[9] Ministry of Education, South Korea(2022). 2022 Information Education Curriculum, Draft Version.
[10] Ministry of Education, South Korea(2022). 2022 Master Plan for Comprehensive Digital Talent Cultivation. Retrieved from https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=72769&boardSeq=92573&lev=0&searchType=null&statusYN=W&page=1&s=moe&m=0315&opType=N
[11] Ministry of Safety and Public Administration, South Korea (2013). Create jobs by loosening the shackles of public data!. Retrieved from https://mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=30053
[12] Ministry of Safety and Public Administration, South Korea (2013). Major work plan of the Ministry of Safety and Public Administration in 2013. Retrieved from https://www.korea.kr/archive/expDocView.do?docId=33763
[13] Public Data Portal, South Korea(2022). Guide to using Public Data Portals.
[14] Public Data Portal, South Korea(2022). National Data Map.
[15] Shin, S. (2021). A Study on Instructional Methods based on Computational Thinking Using Modular Data Analysis Tools for AI Education in Elementary School. *Journal of The Korean Association of Information Education, 25*(6), 917-925.

저자소개

신 승 기

2017 University of Georgia(Ph.D.)
2017 미국 칼빈슨 정부연구소 연구원
2020 애리조나주립대학교
컴퓨터교육전공 교수
2020～현재 서울교육대학교
컴퓨터교육과 교수
관심분야: Computational Thinking, 인공지능교육, 보편적정보교육
e-mail: skshin@snue.ac.kr