https://doi.org/10.32431/kace.2019.22.3.012

자유학기제 운영계획서에 대한 텍스트 빅데이터 분석 및 요약

이수안† · 박범준†† · 김민규†† · 신혜숙††† · 김진호††††

요 약

사회 각 분야에서 관련 주제에 대한 보다 직접적인 정보를 수집하고 분석하기 위하여 빅데이터 분석이 활발하게 활용되고 있다. 우리나라에서 사회적 관심과 파급 효과가 큰 교육 분야에서도 빅데이터 분석 기술을 활용하여 교육이나 정책의 효과를 파악하고 정책 수립에 활용하는 것에 관심이 높아지고 있다. 본 논문에서는 교육 분야에서 빅데이터 분석 기술을 활용하는 방안을 소개하고자 한다. 현재 핵심 교육정책 중의 하나인 자유학기제에 초점을 두고, 각 학교가 작성한 운영계획서에 대해 텍스트 분석과 시각화를 통하여 주요 관심 사항과 차이점에 대해 살펴보았다. 특히 서울과 강원도 지역의 중학교 자유학기제 운영계획서를 대상으로 지역적으로 주요 특성과 관심 사항이 서로 다르다는 것을 비교하였다. 본 연구는 빅데이터 분석 기술을 교육 분야의 필요와 요구에 따라 적용하고 활용하였다는 것에 큰 의의가 있다.

주제어 : 빅데이터, 텍스트 분석, 교육정책, 자유학기제, 시각화

Text Big Data Analysis and Summary for Free Semester Operational Plan Document

Suan Lee† · Beomjun Park†† · Minkyu Kim†† · Hye Sook Shin††† · Jinho Kim††††

ABSTRACT

Big data analysis is actively used for collecting and analyzing direct information on related topics in each field of society. Applying big data analysis technology in education field is increasingly interested in Korea, because applying this technology helps to identify the effectiveness of education methods and policies and applying them for policy formulation. In this paper, we propose our approach of utilizing big data analysis technology in education field. We focus on free semester program, one of the current core education policies, and we analyze the main points of interests and differences in the free semester through analysis and visualization of texts that are written on the operation reports prepared by each school. We compare regional differences in key characteristics and interests based on the free semester operation reports from middle schools particularly at Seoul and Gangwon-do regions. In conclusion, applying and utilizing big data analysis technology according to the needs and requirements of education field is a great significance.

Keywords : Big Data, Text Analysis, Education Policies, Free Semester, Visualization

†정 회 원: 강원대학교 IT대학 연구교수
††정 회 원: 강원대학교 IT대학 컴퓨터학과 학부생
†††정 회 원: 강원대학교 교육학과 교수
††††정 회 원: 강원대학교 IT대학 컴퓨터학과 교수(교신저자)
논문접수: 2019년 4월 18일, 심사완료: 2019년 5월 23일, 게재확정: 2019년 5월 28일
* 본 논문은 2016년도 강원대학교 대학회계 학술연구조성비(관리번호-520160150)의 지원으로 수행되었음.

1. 서론

사회 각 분야에서 빅데이터를 이용한 자료 분석이 광범위하게 활용되고 있다. 예를 들어, 새로운 이슈나 정책에 대한 시민들의 의견을 수집하기 위하여, 설문지를 실시하기보다는 관련 뉴스나 검색어 목록 등을 활용하여 직접 관심사를 분석하는 경우가 많아지고 있다[1][2]. 기존에 설문자료나 조사자료만을 활용하였을 때와 비교하여 빅데이터를 활용하면 보다 생생하고 다각적인 자료를 통해 연구대상을 보다 직접적으로 드러낼 수 있다는 장점이 있다.

교육 분야에서도 NEIS나 교육정보공시 자료 등 빅데이터의 활용이 증가하고 있다. 교육 분야에 빅테이터를 활용한 이전 연구를 살펴보면 주로 빅데이터를 통해 세부 교육 영역의 방향을 탐색하는 연구[3], 교육의 세부 정책 및 프로그램에 대한 사회적 인식을 살펴보는 연구[4][5], 빅데이터가 활용된 교육연구의 동향을 분석하는 연구[6], 빅데이터를 교육에 활용하는 방안에 대한 연구[7][8] 등이 수행되었다. 이러한 연구에서는 주로 신문기사 등을 활용하여 정책 등에 대한 일반 여론을 분석하거나, 특정 기간에 발표된 학술지 및 박사학위 논문 등의 키워드를 분석하는 방법을 적용하였다.

자유학기제 등과 같은 새로운 교육정책이 도입되었을 때, 실제로 그 정책이 어떻게 운용되고 있는지를 생생하고 현장성 있게 살펴보기 위해서는 그 정책을 시행하는 주체들에 의해 기록된 보고서를 분석하는 것이 가장 바람직할 것이다. 신문 기사의 경우 해당 정책이 실제로 교육 현장에서 어떻게 이루어지는지에 대한 정보가 부족하며, 정책 보고서나 논문 등은 정책이 일정 기간 이상 시행이 된 이후에 작성이 되기 때문에, 현재 그 정책이 어떻게 운용되고 있는지에 대한 정보를 제공하지 않는다. 따라서 교육정책이 실제로 어떻게 운용되는지를 살펴보기 위해서는 정책을 시행하며 작성한 운영계획서나 성과보고서 등을 분석할 필요가 있다.

본 연구에서는 빅데이터 기술을 교육 분야에서 실제로 어떻게 활용할 수 있는가를 보이기 위해, 현재 전국의 모든 중학교에서 운영되고 있는 자유학기제의 운영계획서를 바탕으로 빅데이터 분석을 수행하였다. 자유학기제는 지난 2015년 시범운영 이후 2016년부터 전국의 모든 중학교에서 시행되고 있으며, 현 정부의 교육부 6개 과제 중 '교실혁명을 통한 공교육 혁신'의 세부과제로서 자유학기제 확대가 포함되어 있을 만큼 현 정부의 핵심 교육정책 중의 하나이다. 자유학기제는 중학교 과정 중, 한 학기 (또는 두 학기) 동안 시험의 부담에서 벗어나 학생 참여형 수업을 하고, 학생의 소질과 적성을 키울 수 있는 다양한 체험 활동을 중심으로 교육과정을 운영하는 제도이다. 오전에는 주로 교과수업이 이루어지며 오후에는 주로 주제선택 활동, 진로 탐색 활동, 동아리 활동, 예술 체육활동 등 자유 학기 활동이 시행된다. 각 학교별로 운영하는 자유학기제에 대한 내용은 자유학기제 운영계획서에 제시돼 있다.

본 연구에서는 비정형 텍스트 형태로 작성된 자유학기제 운영계획서에 텍스트 분석 기술을 적용하여, 학교별로 주요 관심 사항과 특징을 분석하고 서로 비교하였다. 구체적으로, 서울지역과 강원도 지역에 있는 중학교의 자유학기제 운영계획서를 대상으로 텍스트 분석 기술을 적용하여 주요 키워드를 추출하고, 유사 단어 간의 거리에 따라 벡터화하여 시각화하며, 토픽 모델링을 적용하여 주제별로 요약하였다. 이를 통해 자유학기제에 대한 지역적 관심 사항의 특징과 차이점을 서로 비교하였다. 자유학기제 운영계획서에 대해 빅데이터 분석 기술을 접목한 최초 연구라는 점과 교육 분야의 필요와 요구에 따라 적용하고 활용할 수 있음을 실증적으로 보였다는 점에 큰 의의가 있다.

2. 교육 빅데이터와 자유학기제 운영계획서

2.1 교육 빅데이터

빅데이터란 단순히 양이 많은 데이터가 아니라 다양한 형태로 존재하는 모든 종류의 문서나 자료로부터 목적에 맞는 지식을 추출하고 이를 의사 결정에 활용하는 제반 행위들을 말한다. 따라서 교육 빅데이터란 교육과 관련되어 생성되거나 활용될 수 있는 모든 종류의 데이터의 모임이라 할 수 있다. 이는 학교나 교육청 등 교육 관련 단체가 보유한 데이터에서부터 학부모 인터넷 카페나 소셜네트워크서비스(SNS)의 댓글에 이르기까지 다양한 대상

으로부터 생성될 수 있으며, 단순 통계표와 같은 정형적인 형태뿐만 아니라 문서나 강의자료, 사진과 동영상 등 다양한 매체와 형태로 존재할 수 있다. 하지만, 대표적인 교육 빅데이터의 원천은 시도교육청의 교육행정시스템(NEIS)[9], 지방교육재정시스템(EDUFINE)[10], 초중등 교육정보공시, 고등교육정보공시 등의 교육정보시스템[11]과 이들 시스템으로부터 통계 활용을 위해 만들어진 교육정보통계시스템(EDS), 그리고 한국교육개발원의 교육통계 등이라 할 수 있다.

1. 교육정보통계시스템(EduData System, EDS)

교육행정정보시스템(NEIS)를 중심으로 하는 원천 데이터를 통계 활용을 목적으로 변환해 만든 데이터웨어하우스로써, 별도의 프로그램 없이 온라인분석처리(OLAP, Online Analytical Processing) 도구를 통해 필요한 통계를 추출 및 가공할 수 있는 시스템이다. 2017년을 기준으로 16개 분야 총 1,648종 57,030항목에 대한 통계 자료를 제공하고 있으며, 교육부와 시도교육청 내부적으로 활용하고 있다.

2. 교육정보공시

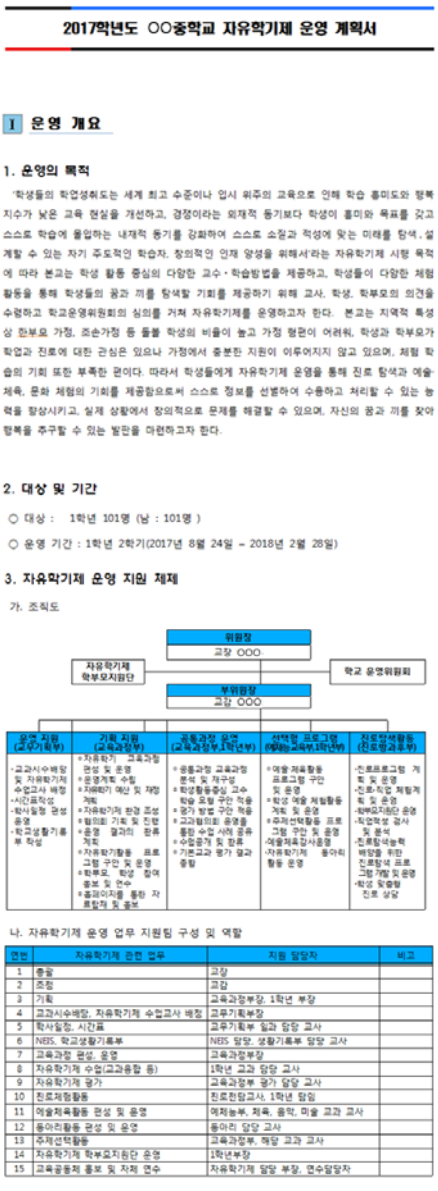
초.중등학교 정보공시제에 따라 학생, 학부모 등 교육 수요자에게 인터넷 홈페이지를 통해 학교 알리미[11]를 통해 교육 활동 전반에 관한 정보를 서비스하고 있다. 또한, 대학정보 공시제에 따라 대학알리미[12]를 통해 각 대학의 통계 정보도 제공하고 있다.

3. 한국교육개발원 교육통계

우리나라 교육통계 조사는 1963년부터 시행하여 매년 교육통계 연보를 제작하여 배포하고 있으며, 1998년부터 한국교육개발원에서 컴퓨터 시스템을 통해 교육통계 조사가 시행되고 관리되고 있다. 교육통계에서 생산된 모든 데이터는 교육통계 연부와 교육통계 서비스 홈페이지[13]를 통해 공개되고 있다.

2.2 자유학기제 운영계획서

위에서 설명한 바와 같이, 학교 알리미[11]는 교육 빅데이터를 구성하는 중요 구성 요소 중의 하나로서 각 학교에 대한 현황과 통계뿐만 아니라 학교 운영에 대한 여러 텍스트 형태의 문서들이 제공되고 있다. [그림 1]과 같이 자유로운 형식의 자유학기제 운영계획서 문서들에 텍스트 마이닝 등 빅데이터 분석 기술을 적용할 경우, 고정된 설문 조사나 통계 조사에서 발견할 수 없었던 유용한 정보를 추출할 수 있을 것이다.



2017학년도 ○○중학교 자유학기제 운영 계획서

Ⅰ 운영 개요

1. 운영의 목적

'학생들의 학업성취도는 세계 최고 수준이나 입시 위주의 교육으로 인해 학습 흥미도와 행복지수가 낮은 교육 현실을 개선하고, 경쟁이라는 외재적 동기보다 학생이 흥미와 목표를 갖고 스스로 학습에 몰입하는 내재적 동기를 강화하여 스스로 소질과 적성에 맞는 미래를 탐색, 설계할 수 있는 자기 주도적인 학습자, 창의적인 인재 양성을 위해서'라는 자유학기제 시행 목적에 따라 본교는 학생 활동 중심의 다양한 교수·학습방법을 제공하고, 학생들이 다양한 체험활동을 통해 학생들의 꿈과 끼를 탐색할 기회를 제공하기 위해 교사, 학생, 학부모의 의견을 수렴하고 학교운영위원회의 심의를 거쳐 자유학기제를 운영하고자 한다. 본교는 지역적 특성상 한부모 가정, 조손가정 등 돌볼 학생의 비율이 높고 가정 형편이 어려워, 학생과 학부모가 학업과 진로에 대한 관심은 있으나 가정에서 충분한 지원이 이루어지지 않고 있으며, 체험 학습의 기회 또한 부족한 편이다. 따라서 학생들에게 자유학기제 운영을 통해 진로 탐색과 예술·체육, 문화 체험의 기회를 제공함으로써 스스로 정보를 선별하여 수용하고 처리할 수 있는 능력을 향상시키고, 실제 상황에서 창의적으로 문제를 해결할 수 있으며, 자신의 꿈과 끼를 찾아 행복을 추구할 수 있는 발판을 마련하고자 한다.

2. 대상 및 기간

○ 대상 : 1학년 101명 (남 : 101명)

○ 운영 기간 : 1학년 2학기(2017년 8월 24일 ~ 2018년 2월 28일)

3. 자유학기제 운영 지원 체제

가. 조직도

위원장 교장 OOO

자유학기제 학부모지원단

학교 운영위원회

부위원장 교감 OOO

운영 지원 (교무기획부)	기획 지원 (교육과정부)	교육과정 운영 (교육과정부,1학년부)	선택형 프로그램 (창체,교육부,1학년부)	진로탐색활동 (진로업무부)
·교과시수배정 및 자유학기제 수업교사 배정 ·시간표작성 ·학사일정 관리 운영 ·학교생활기록부 작성	·자유학기 교육과정 편성 및 운영 ·운영계획 수립 ·자유학기 예산 및 재정계획 ·자유학기제 환경 조성 ·협의회 기획 및 진행 ·운영 결과의 환류 계획 ·자유학기활동 프로그램 구안 및 운영 ·학부모, 학생 참여 홍보 및 연수 ·홈페이지를 통한 자료탑재 및 홍보	·융통과정 교육과정 분석 및 재구성 ·학생활동중심 교수·학습 모형 구안 적용 ·평가 방법 구안 적용 ·교과협의회 운영을 통한 수업 사례 공유 ·수업공개 및 참관 ·기본교과 평가 결과 종합	·예술·체육활동 프로그램 구안 및 운영 ·학생 예술 체험활동 계획 및 운영 ·주제선택활동 프로그램 구안 및 운영 ·예술체육강사풀운영 ·자유학기제 동아리 활동 운영	·진로프로그램 계획 및 운영 ·진로·직업 체험계획 및 운영 ·학부모지원단 운영 ·직업학생 검사 및 분석 ·진로탐색능력 배양을 위한 진로탐색 프로그램 개발 및 운영 ·학생 맞춤형 진로 상담

나. 자유학기제 운영 업무 지원팀 구성 및 역할

연번	자유학기제 관련 업무	지원 담당자	비고
1	총괄	교장	
2	조정	교감	
3	기획	교육과정부장, 1학년 부장	
4	교과시수배정, 자유학기제 수업교사 배정	교무기획부장	
5	학사일정, 시간표	교무기획부 업무 담당 교사	
6	NEIS, 학교생활기록부	NEIS 담당, 생활기록부 담당 교사	
7	교육과정 편성, 운영	교육과정부장	
8	자유학기제 수업(교과융합 등)	1학년 교과 담당 교사	
9	자유학기제 평가	교육과정부 평가 담당 교사	
10	진로체험활동	진로전담교사, 1학년 담임	
11	예술체육활동 편성 및 운영	예체능부, 체육, 음악, 미술 교과 교사	
12	동아리활동 편성 및 운영	동아리 담당 교사	
13	주제선택활동	교육과정부, 해당 교과 교사	
14	자유학기제 학부모지원단 운영	1학년부장	
15	교육공동체 홍보 및 자체 연수	자유학기제 담당 부장, 연수담당자	

[그림 1] 자유학기제 운영 계획서

3. 텍스트 빅데이터 분석 기술

자유학기제 운영계획서는 자유로운 텍스트 형식의 문서이다. 빅데이터 분석 기술에는 이러한 텍스트 형태의 데이터로부터 유용한 정보를 추출하는 여러 가지 기법들이 있다. 대표적으로 텍스트 형태의 비정형 데이터를 분석하여 정보를 추출하거나 연계성을 파악하는 기법을 텍스트 마이닝(Text Mining)이라고 한다[14]. 또한, 수많은 웹 문서나 뉴스 기사 등을 분석하여 텍스트의 문맥과 의미를 파악하여 텍스트 간의 연계성을 파악하기도 한다. 텍스트 마이닝에서 사용되는 대표적인 기법은 버즈 분석(Buzz Analysis), 키워드 빈도 분석(Keyword Frequency Analysis), 토픽 모델링(Topic Modeling), 오피니언 마이닝(Opinion Mining) 등이 있다. 이 중 자유학기제 운영계획서 등과 같은 정책의 시행과 관련된 문서를 분석하는데 활용할 수 있는 기법은 키워드 빈도 분석과 Word2Vec[15], 그리고 토픽 모델링[16]이 있다.

3.1 키워드 빈도 분석 및 워드 클라우드

키워드 빈도 분석은 특정 문서에서 자주 언급되는 키워드를 추출하고 해당 키워드의 빈도에 따라서 중요도를 분석하는 방법이다. 키워드 빈도는 단순하게 문서에서 특정 단어가 얼마나 자주 등장하는지를 나타내는 단어 빈도(TF: Term Frequency)에 따라서 결정된다. 그러나 빈도가 높은 단어가 모든 문서에서 자주 등장하는 무의미한 경우일 수 있다. 즉, 문서 빈도(DF: Document Frequency) 값이 높은 단어일 수 있다. 이렇게 단순히 자주 빈번하게 등장하는 단어를 제외하기 위해서 TF-IDF(Term Frequence-Inverse Document Frequence)를 이용한다. 이는 특정 문서에서 단어의 빈도인 TF 값에서 전체 문서와 관계된 단어의 중요도를 나타내는 IDF를 곱한 값이다. TF-IDF를 사용하면 특정 문서 내에서 단어의 빈도가 높을수록 값이 커지고, 전체 문서들 중 그 단어를 포함하는 문서가 적을 적을수록 값이 커지게 된다. 이때 키워드 빈도 분석의 결과를 시각적으로 제시하는데 있어서 워드 클라우드(Word Cloud) 방법이 흔히 사용되는데, 이는 문서의 키워드, 개념 등을 직관적으로 파악 할 수 있도록 단어의 빈도수에 따라 핵심단어를 시각적으로 돋보이게 하는 시각화 기술이다.

교육 분야에서도 텍스트 분석 기법을 활용하는 다양한 연구들이 있다[17][18][19]. 그 중에서 대학 구조개혁 평가에 대해서 키워드 및 토픽을 분석한 연구[19]가 있다. 이 연구에서는 보조 자료 형태의 교육부 문서, 대학전문지 및 종합일간지에서 제공되는 신문 기사 데이터를 활용하였다. 그리고 대학 구조개혁 평가와 관련된 주요 키워드들에 대해서 분석을 수행하였다.

3.2 Word2Vec

Word2Vec은 인공신경망(ANN)을 이용하여 Word Embedding을 수행하는 기법이다. Word2Vec은 단어 사이의 분포 관계를 기반으로 단어를 일정한 의미를 가지는 벡터 형태로 변환하는 기법이다. 문서에 나오는 단어 집합(vocabulary)의 크기가 벡터의 크기가 되고, 각 단어들을 인덱스화하여 표현할 단어의 인덱스 위치에 1을 부여하고, 다른 단어 인덱스에는 0을 부여한다. 이와 같이 원-핫 인코딩(one-hot encoding)을 수행하여 문서에 대해서 벡터화하여 원-핫 벡터(one-hot vector)를 만든다. 그리고 비슷한 위치에 있는 단어들이 서로 유사한 의미를 가진다는 가정으로 단어의 의미를 벡터화한다. 그러나 이러한 벡터는 매우 고차원의 결과를 나타내기 때문에, 결과의 가독성과 해석 가능성을 높이기 위하여 t-SNE(t-Distributed Stochastic Neighbor Embedding)[20]를 이용하여 저차원으로 변환 후 시각화한다. Word2Vec이 비교적 최신 기술이기 때문에 아직까지 교육 분야에서 Word2Vec을 이용한 사례를 찾기 어렵다. 본 논문에서는 Word2Vec을 사용하여 자유학기제 문서의 단어에 대한 의미를 벡터화하였고, 이를 시각화하기 위해 t-SNE 기술을 사용하였다.

3.3 토픽 모델링

토픽 모델링은 텍스트 데이터에서 사용되는 키워드들이 동시에 사용되는 패턴을 기반으로 대표적인

주제나 이슈에 대한 그룹들을 자동으로 추출하는 기법이다. 토픽 모델링을 사용하면 다양한 주제, 즉 토픽(topic)들이 확률적으로 혼합되어있는 문서에 대해서 토픽에 관한 키워드들을 추출해낼 수 있다. 토픽 모델링의 대표적인 기법은 LDA(Latent Dirichlet Allocation)이다. LDA는 데이터에 존재하는 잠재 변인을 유추하고 데이터를 차원 축소하여 데이터를 이해한다. 문서에 여러 토픽들이 존재할 수 있고, 토픽은 다양한 단어들로 표현된다고 전제한다. 그리고 각 문서마다 토픽의 구성 비율이 다르고, 이 비율을 결정하는 확률분포로 디리클레 분포가 존재한다고 가정한다. LDA는 결과적으로 텍스트 데이터에서 주요 토픽들과 그 토픽들을 구성하는 단어들의 집합을 찾아낸다.

4. 교육 빅데이터와 자유학기제 운영계획서

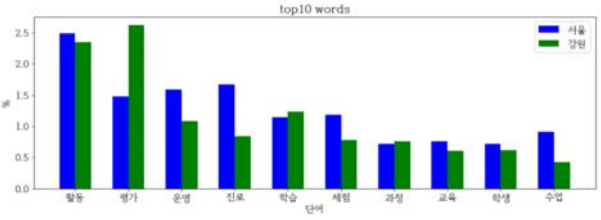
자유학기제 데이터는 텍스트로 구성된 비정형 데이터로써 동일한 문장이 반복되거나 무의미한 문장이 나열되는 부분 그리고 장황한 설명 등이 들어있을 수 있는 정제되지 않은 상태이다. 이러한 입력 데이터에 대해 빅데이터 분석을 적용하기 위해 먼저 텍스트 전처리를 통하여 중복되거나 비슷한 문장을 제거, 불용어 제거, 형태소 분석 등의 전처리 과정을 거쳐 정제된 데이터로 가공한다. 전처리 과정에서 데이터 가공을 위해 파이썬을 사용하였으며, 한국어 처리 패키지인 KoNLPy[21]를 사용한다. 먼저 KoNLPy를 사용하여 명사만 추출하고 말뭉치(corpus)를 생성한다. 이후 대소문자를 변환한다. 그리고 문장에서 불필요한 숫자나 부호, URL 등을 제거 후 언어 별로 불용어(stopword)들을 처리한다. 그리고 단어는 행으로, 문서는 열로 가지는 행렬 TDM(Term Document Matrix)을 생성하여 해당 단어가 해당 문서에서 출현하는 빈도를 계산하였다.

4.1 서울과 강원지역의 자유학기제 워드 클라우드 분석

본 논문에서는 자유학기제 문서를 이용하여 텍스트를 명사 단위로 분류한 뒤 한눈에 알아볼 수 있는 워드 클라우드(Word Cloud)를 만들고, 단어 빈도 분석을 통해 해당 데이터에 대한 시각화를 하였다. [그림 2]와 [그림 3]은 서울과 강원도 내 중학교에 대한 워드 클라우드 결과이다. 그림 4는 서울과 강원도 지역에 대해서 top 10 단어 분포를 비교한 차트이다.

[그림 2] 서울지역 전체 중학교 워드 클라우드

[그림 3] 강원도 지역 전체 중학교 워드 클라우드



[그림 4] 서울과 강원도 지역 top 10 단어 분포

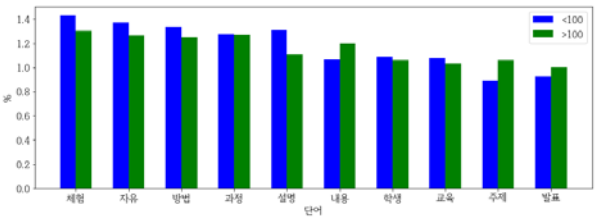
서울과 강원도 중학교의 자유학기제 활동 보고서에서 공통적으로 많이 나타난 단어는 활동, 평가, 운영, 진로, 체험 등이었다. 이러한 용어는 자유학기제의 목표와 실행전략 등에서 주로 사용된 단어로서 자유학기제를 일반 다른 학기와 구분하는 특징이라고 할 수 있다. 워드 클라우드에 높은 빈도로 나타난 결과를 지역별로 비교해보면, 서울은 진로, 운영, 체험, 수업, 선택 등이었고, 강원도는 평가, 학습, 과정 등이었다. 이를 통해 서울시내 중학

교에서 학생들의 요구나 특성에 맞는 체험 활동 등을 통하여 진로지도가 보다 많이 이루어지고 있다고 추론할 수 있다. 이에 비하여 강원도의 자유학기제 운영계획서에는 교사에 의해 주도되는 학습이나 평가가 좀더 강조되는 경향이 있었다.

다음으로 강원도내 중학교를 학생수로 구분하여 100명 이상인 학교와 100명 미만인 학교로 구분하여 자료를 분석한 결과는 다음 그림 5와 6 그리고 그림 7과 같다. 분석 결과, 평가와 활동, 학습, 운영, 수업 등이 공통적으로 자주 관찰되었다. 학교 규모에 따른 주요 단어를 비교해보면, 100명 이상 학교에서는 교사와 모둠, 주제 등이, 100명 미만 학교에서는 자유, 학생, 방법, 설명 등의 단어가 자주 관찰되었다. 이는 규모가 큰 학교에서 보다 교사 주도적인 프로그램을 운영하며 모둠 활동을 주로 활용하는 것으로 생각된다. 반면 소규모학교에서는 학생들이 자유롭게 활동에 참여하며, 그 의미를 설명하는 활동이 병행되는 것으로 추측된다.

[그림 5] 강원도 내 학생 수 100명 이상 중학교 워드 클라우드

[그림 6] 강원도 내 학생 수 100명 이하 중학교 워드 클라우드



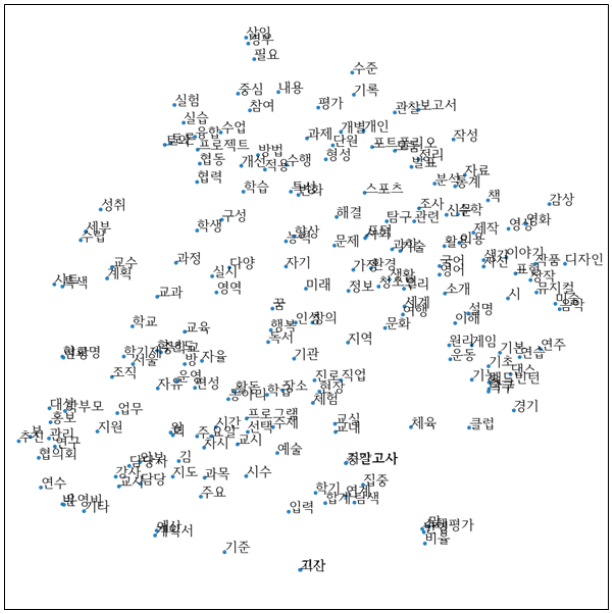
[그림 7] 강원도내 100명 이상과 100명 미만인 학교의 단어의 분포 비교

4.2 서울과 강원지역의 자유학기제 Word2Vec

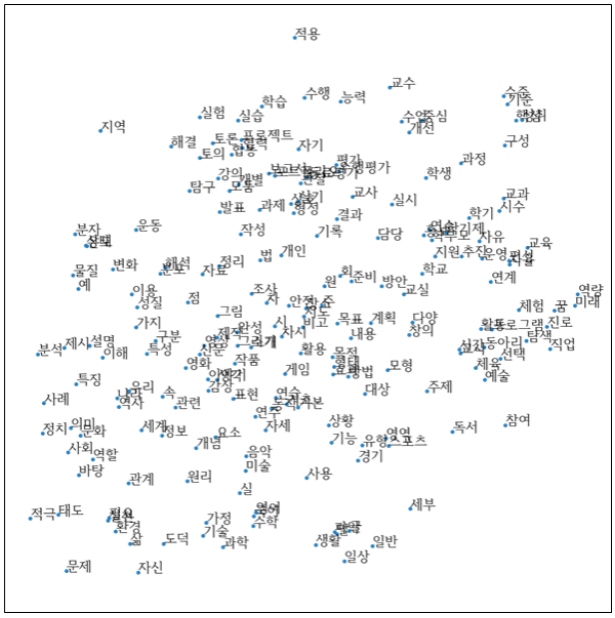
본 논문에서는 Word2Vec을 활용하여 서울과 강원지역의 모든 중학교의 자유학기제 운영계획서를 벡터 형태로 변환하였다. 이후 고차원 데이터를 시각화하기 위해 차원 감소기술인 t-SNE를 활용하여 상호 연관이 높은 키워드들을 2차원으로 제시하였다. 그림 8과 9는 자유학기제 문서에 대해서 t-SNE를 이용하여 2차원으로 시각화한 결과이다.

표 1과 2는 자유학기제 문서에 대해 Word2Vec을 이용하여 산출된 2차원 벡터 중에서 중요도를 고려하여 높은 순으로 유사 단어들을 10개씩 뽑아낸 결과이다. 이렇게 추출된 유사단어 결과를 보면 두 지역에서 공통적으로 추출된 유사단어도 일부 있지만, 지역에 따라 동일한 키워드에도 추출된 유사단어에 일부 차이가 있었다. 예를 들어 평가를 기준으로 산출한 유사단어 결과를 보면, 두 지역 모두에서 수행평가, 동료평가, 서술, 결과(물) 등이 추출되었다. 이는 중간고사와 기말고사 등의 지필검사를 치르지 않는 자유학기제의 평가가 공통적으로는 수행평가의 형식으로 이루어지고 있으며, 동료평가 등을 활용하며 그 결과는 서술의 형태로 제시되고 있다는 것을 보여준다. 이는 학생의 성장과 발달의 과정에 초점을 두고 그 과정을 서술하고 기록한다는 점에서 자유학기제가 보다 바람직하게 운영되고 있다는 것을 나타낸다. 지역에 따른 차이를 구체적으로 살펴보면, 서울지역에서는 형성평가와 피드백, 기록, 종합 등이 추출되었고, 강원지역은 성찰, 포트폴리오, 논술, 성적표, 동료 등이 추출되었다. 이는 서울지역에서는 수행평가 외에도 형성평가가 활용되고 있으며, 그 특징으로 피드백 등이 함께 추출되었다. 강원지역은 수행평가 유형 중 포트폴리오와 논술 등이 추출된 것으로 보아, 자유학기제의 평가에 있어서 수행평가를 주로 활용하는

것으로 해석할 수 있다.



[그림 8] 서울지역 2차원 t-SNE 결과



[그림 9] 강원지역 2차원 t-SNE 결과

<표 1> Word2Vec에서 서울지역 단어별 유사 단어 10개

평가	*활동*	*진로*	*과학*	*체험학습*	*미술*
형성평가	프로그램	제공	생연	견학	음악
결과	학생	기회	과학자	탐색	예술음악
수행평가	동아리활동	직업	공상	소규모	예술
동료평가	동아리	미래	다빈치	장기	연극
확인	기회	체계	해저	청진기	앨범커버
피드백	제공	제공	탐구	특강	르네상스
서술	활성화	적성	과학지식	캠프	수학과
기록	진로	꿈	수학과	현장	향유
종합	연계	창체	고도	일제	실용
결과물	자율	독서	기술	보릿고개	접목
탁구	*연습*	*목적*	*과정*	*수업*	*교육*
배드민턴	패스	공교육	시행	학습	강화
농구	안무	기여	행동	통합	역량
풋살	백핸드	방향	누적	활성	활성화
축구	슛	취지	수시	참여	의식
배구	파트	논의	과제	개선	인성
배트	셈	지속	서술	사용료	교육활동
피구	드리블	계기	확대	준비	혁신
리그	복습	합리	수행평가	효과	자치
여학생	주법	정착	성장	적응력	함양
핸드볼	스텝	수렴	학습자	준비물	민주

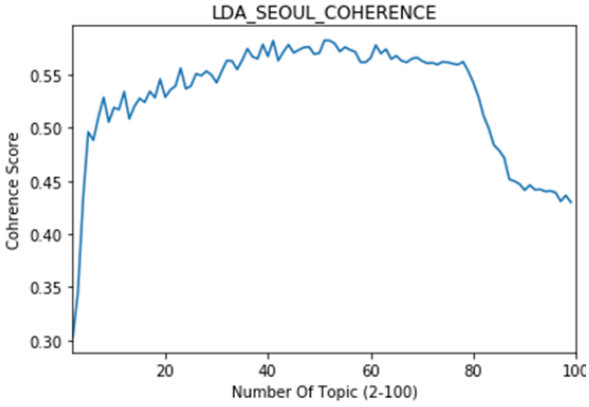
<표 2> Word2Vec에서 강원지역 단어별 유사 단어 10개

평가	*활동*	*진로*	*과학*	*체험학습*	*미술*
수행평가	프로그램	적성	수학	현장	음악
동료평가	자율	직업	세모	특강	예술
성찰	동아리	눈높이	호기심	봉사	시대
서술	확대	프로그램	과학자	탐색	문학
결과	기회	요강	실생활	연계	세계어
포트폴리오	영역	제공	만물	멘토링	가치
논술	학생	가진	발명	캠프	고전
성적표	봉사	기회	인지도	벼룩시장	동시대
동료	진로	실질	한계	발굴	동서양
성취	다양	미래	도덕	일터	배경
탁구	*연습*	*목적*	*과정*	*수업*	*교육*
배드민턴	변형	설정	경직	학습	자율
농구	구사	다음	교과	중심	연계
축구	운지	어려움	학기	융합	조성
테니스	비트	독자	익명성	교수	기반
풋살	연습곡	구체	효율	개선	학교
플로어	실전	부합	평가	교과	활성
디스크	텅잉	고려	확대	확대	학기제
볼	칼립소	다양	가능	프로젝트	운영
핸드볼	드럼	효과	내용	모색	마련
프리	파트	효율	중심	위주	개발

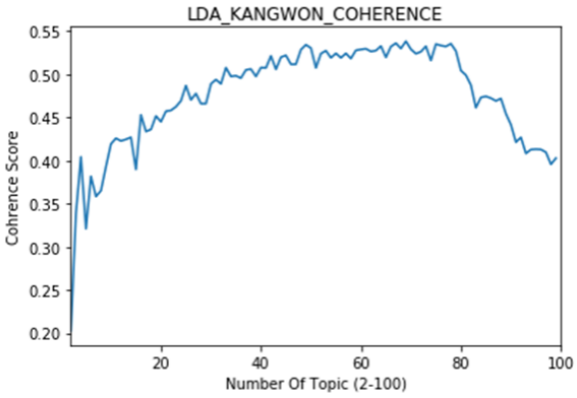
4.3 서울과 강원지역의 자유학기제 토픽 모델링

마지막으로 자유학기제 운영계획서에 포함된 단어들의 빈도수를 바탕으로 각 문서에 어떤 토픽들이 존재하는지를 확률 모형으로 만들기 위해서 LDA(Latent Dirichlet Allocation)를 이용하였다. LDA는 각 문서에 어떤 주제들이 존재하는지에 대한 확률 모형으로 문서에 포함된 단어들의 빈도수를 가지고 표현한다. 본 논문에서는 자유학기제 보고서에 대해 토픽의 의미론적 일관성을 살펴보기 위해서 coherence를 통해 적절한 토픽의 수를 찾아본다. 토픽이 얼마만큼 일관성을 가지는지에 따라서 coherence score가 정해진다. 그림 10와 11은 자유학기제 보고서에 대해서 서울지역과 강원지역에 대해서 토픽의 수에 따라 coherence score를 분석한 결과이다. 토픽의 수를 정함에 있어서 coherence score가 크게 오르는 지점으로 토픽의 수를 정하였다. 토픽의 수가 변할 때마다 coherence score를 분석해보면, 서울은 토픽의 수가 5개일 때 coherence score가 크게 높아졌으며, 강원도는 토픽의 수가 4개일 때 coherence score가 크게 높아졌다. 즉, 서울의 경우에는 토픽의 수를 5개로하고, 강원지역은 토픽의 수를 4개로 정하여 토픽 모델링을 수행하였다.

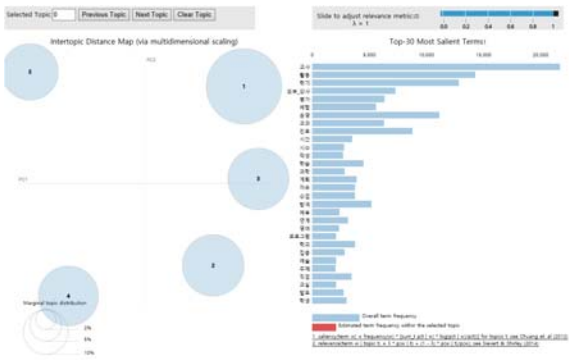
그림 12는 서울지역의 자유학기제에 대한 토픽들을 시각화한 결과이다. 결과를 보면, 총 5개의 토픽들이 적절하게 구분되어 있음을 알 수 있다. 그리고 표 3은 토픽 별로 연관된 주요 단어들을 정리한 결과이다. 연관된 단어들을 통해 각 토픽이 의미하는 것을 살펴보면, 토픽 1에는 활동, 학기, 운영, 교과, 탐색, 계획, 자유, 수업, 연계, 학생, 집중, 중심, 진로, 교육, 과정 등이 포함되며, 이러한 단어는 자유학기제의 '교육과정 구성'을 보여준다고 할 수 있다. 토픽 2는 외부강사, 학습, 체육, 영어, 직업, 음악, 국어, 표현, 중간고사, 미술, 자료 등이 포함되며, 이는 자유학기제의 '운영교과 및 내용'을 보여준다고 할 수 있다. 토픽 3은 체험, 시간, 학교, 예술, 교실, 선택, 제작, 사회, 문화, 현황, 창의 등이 포함되며, 이는 자유학기제 '운영 방식'을 나타낸다고 할 수 있다. 토픽 4는 평가, 과학, 발표, 주제, 모듈, 방법, 탐구, 포트폴리오 등으로, 자유학기제에서의 '평가방법'이라고 할 수 있다. 마지막으로 토픽 5는 교사, 시수, 작성, 프로그램, 지도, 예산_계획서 등이 포함되었으며, '교사 활동'이라고 할 수 있다.



[그림 10] 서울지역의 토픽 coherence



[그림 11] 강원지역의 토픽 coherence

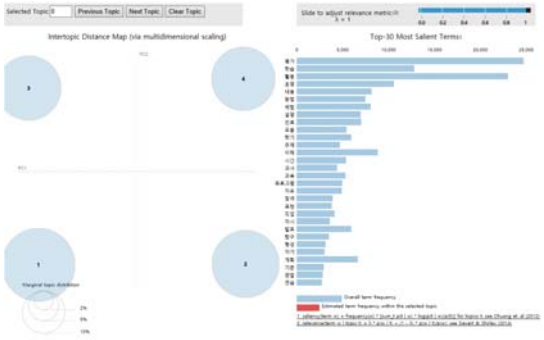


[그림 12] 서울지역 자유학기제 토픽 모델링 시각화

<표 3> 서울 자유학기제 토픽에 대한 주요 단어

토픽명	주요 단어
토픽 1 교육과정 구성	활동, 학기, 운영, 교과, 탐색, 계획, 자유, 수업, 연계, 학생, 집중, 중심, 진로, 교육, 과정
토픽 2 운영교과 및 내용	외부강사, 학습, 체육, 영어, 직업, 음악, 국어, 표현, 중간고사, 미술, 자료
토픽 3 운영 방식	체험, 시간, 학교, 예술, 교실, 선택, 제작, 사회, 문화, 현황, 창의
토픽 4 평가방법	평가, 과학, 발표, 주제, 모듈, 방법, 탐구, 포트폴리오
토픽 5 교사 활동	교사, 시수, 작성, 프로그램, 지도, 예산_계획서

그림 13은 강원도의 자유학기제에 대해서 시각화한 토픽들이다. 강원도는 총 4개의 토픽에 대해서 살펴볼 수 있다. 표 4는 강원도의 주요 토픽에 대해서 연관된 주요 단어들을 정리한 결과이다. 연관 단어에 따라서 토픽들을 분류하면 토픽 1은 활동, 운영, 체험, 진로, 학기, 시간 교육, 프로그램, 직업, 계획, 과정 등의 단어를 포함하며, 자유학기제의 '교육과정 구성'을 나타낸다. 토픽 2는 학습, 설명, 이해, 차시, 탐구, 개별, 글, 작품, 과학, 협동, 제작, 발표 등으로 자유학기제 '수업방법'을 나타낸다. 강원지역 자유학기제의 토픽 3은 평가, 내용, 방법, 활동, 모둠, 주제, 교사, 참여, 형성, 자기 등으로 자유학기제의 '평가방법'이라고 할 수 있다. 마지막으로 토픽 4는 발표, 표현, 이해, 연습, 토론, 음악, 문화, 활용, 기본, 토의 시수 등으로서 자유학기제의 '수업내용'이라고 할 수 있다.



[그림 13] 강원지역 자유학기제 토픽 모델링 시각화

<표 4> 강원 자유학기제 토픽에 대한 주요 단어

토픽명	주요 단어
토픽 1 교육과정 구성	활동, 운영, 체험, 진로, 학기, 시간 교육, 프로그램, 직업, 계획, 과정
토픽 2 수업방법	학습, 설명, 이해, 차시, 탐구, 개별, 글, 작품, 과학, 협동, 제작, 발표
토픽 3 평가방법	평가, 내용, 방법, 활동, 모둠, 주제, 교사, 참여, 형성, 자기
토픽 4 수업내용	발표, 표현, 이해, 연습, 토론, 음악, 문화, 활용, 기본, 토의 시수

토픽 모델링을 통해서 살펴본 서울과 강원지역 중학교의 자유학기제 운영계획서의 토픽들을 살펴보면, 서로 공통적으로 교육과정 구성과 평가방법이 주요 토픽임을 알 수 있다. 교육과정 구성과 관련하여 활동, 운영, 교과, 체험, 진로, 직업, 계획, 과정 등이 주요 단어로 언급되고 있으며, 평가방법의 경우에도 평가 발표, 주제, 방법, 탐구, 포트폴리오, 참여 등이 공통적으로 언급되고 있는데, 이러한 단어들로 자유학기제가 일반 다른 학기와 구별되는 특징을 알 수 있다. 그러나 서울지역의 경우, 운영교과의 내용이나 교사의 활동에 대해 보다 구체적으로 기술하고 있었고, 강원지역의 경우 자유학기제에서 다루는 수업내용에 대하여 보다 구체적으로 기술하고 있었다. 다른 방법과는 달리 토픽 모델링의 경우 자유학기제 운영계획서가 어떠한 주제로 어떤 내용을 기술하고 있는지를 보여주며, 지역간 비교를 통하여 각 지역별 특색을 나타내준다는 장점이 있다.

5. 결론

최근 교육과 사회 전반에서 빅데이터를 활용하여 사회현상을 심층적으로 분석하고자 하는 사회적·학문적 요구가 높아지고 있다. 다른 분야와 마찬가지로 교육의 영역에서도 다양한 교육정책이 학교현장에서 실제로 어떻게 운영되고자 하는지를 분석하기 위하여 빅데이터 분석을 활용할 수 있다. 정책 운영 현황 분석을 위해 기존에 주로 사용되었던 정형 자료(조사 자료)에 비하여 운영계획서나 성과보고서 등과 같은 비정형 빅데이터는 특정 정책이 실제로 어떻게 운용되고 있는가에 대한 생생한 정보를

가지고 있다는 점에서 분석의 가치가 높다. 특히 정책의 시행 주체가 작성한 운영계획서 등은 관련 정책의 구성과 운영상의 특징을 매우 직접적으로 보여준다.

본 연구에서는 현 정부의 대표적인 교육정책인 자유학기제의 운용 현황과 시도별 운영 특징을 살펴보기 위하여 서울과 강원지역 전체 중학교의 자유학기제 운영계획서를 분석하였다. 구체적으로 서울과 강원지역 중학교의 자유학기제 운영계획서에 포함된 주요 단어와 빈도를 분석하여 지역별 주요 관심 사항을 비교하고 시각화하였다. 또한, 지역별 자유학기제 운영계획서의 단어분포를 비교하였고, 토픽 모델링을 통해서 지역별로 강조되는 주제의 차이를 분석하였다. 본 연구를 통해 도출된 결과를 중심으로 그 시사점을 논의하면 다음과 같다.

첫째, 자유학기제 운영계획서를 바탕으로 워드 클라우드 분석을 수행한 결과, 서울은 진로와 운영, 체험 등이, 강원도는 평가, 학습, 과정 등이 빈도수가 높은 것으로 나타났다. 이는 서울에서 학생 특성에 맞는 체험 활동 등의 진로지도가 더 많이 이루어지고 있으며, 강원도에서는 교사에 의해 주도되는 학습이나 평가가 조금 더 강조되는 경향이 있었다는 것을 시사한다. 이러한 점을 기반을 보았을 때, 이후 강원지역에서는 학생들의 진로 발달에 초점을 두고 진로를 점차 발전시킬 수 있는 다양한 프로그램을 개발하여 제공할 필요가 있다.

둘째, Word2Vec 분석결과를 요약해보면, 두 지역 모두에서 공통으로 수행평가와 동료평가 등이 추출되었다. 이는 지필 검사가 없는 자유학기제에서 수행평가가 활용된다는 것을 보여준다. 지역에 따른 차이를 보면 서울에서는 형성평가와 피드백 등이, 강원지역은 성찰, 포트폴리오, 논술 등이 추출되었는데, 이는 서울에서 형성평가도 함께 활용되고 있으며, 강원지역은 수행평가를 주로 활용하는 것으로 해석할 수 있다. 따라서 이후 강원지역에서도 학생들의 피드백을 중심으로 학생들의 발달에 초점을 둔 학생 맞춤형 교육이 제공될 필요가 있다.

마지막으로 토픽 모델링 분석결과를 보면 서울과 강원지역 모두에서 활동, 운영, 교과, 체험, 진로, 직업 등의 교육과정 구성과 평가, 발표, 주제, 방법, 탐구, 포트폴리오, 참여 등의 평가방법이 주요 토픽이었다. 지역에 따른 강조점을 살펴보면, 서울은 운영 교과의 내용이나 교사의 활동에 대해, 강원도는 수업내용에 대하여 더욱 구체적으로 기술하고 있었다.

이처럼 본 연구에서 도입한 세 가지 텍스트 분석 방법은 서울과 강원지역 중학교의 자유학기제 운영계획서를 각각 다양한 관점에서 분석하고 있다. 즉 워드 클라우드는 각 지역의 운영계획서에서 전체적으로 어떤 단어가 자주 언급되는지를 분석함으로써 각 지역에서 초점을 두고 있는 자유학기제의 특징을 드러내 주고 있다. 다음으로 Word2Vec은 각 운영계획서에 기술된 단어 간 관계를 분석함으로써 어떤 단어 간 유사성이 높은지, 그러한 단어들을 조합하면 어떤 정보를 얻을 수 있는지를 보여준다. 마지막으로 토픽 모델링은 각 지역의 자유학기제 운영계획서가 전체적으로 어떤 토픽을 중심으로 기술되었는지를 보여줌으로써 두 지역 계획서의 차이를 더욱 생생하게 보여주고 있다.

본 연구는 현재 전국의 모든 중학교에서 시행되고 있는 자유학기제의 운용 현황을 분석하기 위하여 실제로 자유학기제의 수업을 구성하고 시행하고 있는 교사들이 작성한 수업계획서를 다양한 빅데이터 분석 방법을 활용하여 분석하였다는 데 그 의미가 있다. 즉 본 연구의 결과는 사전에 일정한 내용으로 질문을 한정하는 설문지 조사를 통해 수행된 연구에 비하여 더욱 광범위한 주제를 분석할 수 있었다는 장점이 있으며, 신문기사나 보도자료가 아닌 현장 전문가의 실제 계획서를 분석했다는 점에서 자유학기제의 실제 운영 현황에 근접한 결과를 도출하였다고 생각된다.

참 고 문 헌

[1] 박수정 · 김영태 (2018). 자유학기제 관련 신문기사 분석. **학습자중심교과교육학회. 18**(18), 683-707.

[2] 유예림 · 백순근 (2016). 자동화된 텍스트 분석을 활용한 2015 개정 교육과정 정책에 대한 언론 보도의 쟁점 분석. **교육과정평가연구, 19**(3), 127-156.

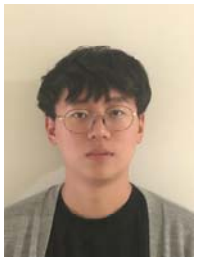
[3] 김경철 · 김은혜 (2017). 빅데이터 활용을 통한 유아부모교육 방향 탐색. **유아교육학회 정기 학술대회 논문집**, 124-138.
[4] 강승지 · 이연선 (2017). 빅데이터를 통해 바라본 유아 스마트미디어 교육에 대한 사회적 인식. **열린유아교육연구, 22**(4), 45-72.
[5] 강승지 · 이연선 (2018). 우리나라 유아 영어교육에 대한 사회적 인식 연구: 빅데이터와 사회연결망 분석을 중심으로. **미래유아교육학회지, 25**(2), 141-168.
[6] 김선아 · 박진희 · 이현정 · 정유진 (2016). 텍스트마이닝 기법을 활용한 다문화 미술교육 연구 동향 분석 연구. **다문화교육연구, 9**(2), 203-227.
[7] 이영석 · 조정원 (2016). 빅데이터의 교육적 활용방안 연구. **한국산학기술학회논문지, 17**(12), 716-722.
[8] 최제영 · 박충식 · 최광선 · 정의석 · 김성진 · 유인식 (2012). 스마트교육에서 발생하는 교육 빅데이터 활용방안. **한국지능정보시스템학회 학술대회논문집, 2012**(5), 144-148.
[9] 교육행정시스템(NEIS), 2019년 05월 02일 접속, http://neis.moe.go.kr.
[10] 지방교육재정시스템(EDUFINE), 2019년 05월 02일 접속, http://www.keris.or.kr.
[11] 학교알리미, 2019년 05월 02일 접속, https://www.schoolinfo.go.kr.
[12] 대학알리미, 2019년 05월 02일 접속, http://www.academyinfo.go.kr.
[13] 교육통계서비스, 2019년 05월 02일 접속, http://kess.kedi.re.kr.
[14] Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). *Fundamentals of predictive text mining.* Springer.
[15] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* In Advances in neural information processing systems (pp. 3111-3119).
[16] Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. *In Proceedings of the 23rd international conference on Machine learning (pp. 977-984).* ACM.
[17] 신철균 · 황은희 · 김은영 (2015). 자유학기제 운영 실태 분석 연구. **아시아교육연구, 16**(3), 27-55.
[18] 김동심 (2017). 자유학기제 운영에 따른 교육성과 변화 분석 – 진로성숙도, 인지적정의적 사회적 핵심역량, 학교만족도를 중심으로-. **교육과정평가연구, 20**(3), 101-121.
[19] 김지은 (2017). 대학전문지 기사에 나타난 대학구조개혁 평가 토픽 분석. **교육종합연구소, 15**(4), 203-231.
[20] Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. J*ournal of machine learning research, 9(Nov)*, 2579-2605.
[21] KoNLPy(Korean NLP in Python): http://konlpy.org.

이 수 안

2008 강원대학교 컴퓨터과학과(학사)
2010 강원대학교 컴퓨터과학과(전산학석사)
2012 강원대학교 컴퓨터과학과(전산학박사수료)
2015 ㈜알티베이스 연구개발본부 연구원
2017 강원대학교 컴퓨터과학과(전산학박사)
2018 강원대학교 정보통신연구소 박사후연구원
2019 ~ 현재 강원대학교 SW중심대학 연구교수
관심분야: 빅데이터, 분산병렬컴퓨팅, 머신러닝, 딥러닝, 그래프, 추천시스템, 데이터마이닝, 데이터베이스, 클라우드컴퓨팅
E-Mail: suanlab@gmail.com

박 범 준

2012 ~ 현재 강원대학교 컴퓨터과학과(학사과정)
관심분야: 텍스트마이닝, 딥러닝
E-Mail: pjh8723@gmail.com



김 민 규

2016~현재 강원대학교
컴퓨터과학과(학사과정)
관심분야: 인공지능, 신경과학,
인지심리학
E-Mail: a4855744@kangwon.ac.kr

신 혜 숙

1998 서울대학교 교육학과(학사)
2001 서울대학교
교육학과(교육학석사)
2009 미국 UCLA
교육통계및평가 전공(박사)
2013 ~ 현재 강원대학교 교육학과 조교수
관심분야: 다층모형, 구조방정식, 인과이론,
책무성 평가 등
E-Mail: hyesook@kangwon.ac.kr

김 진 호

1982 경북대학교
전자공학과(학사)
1985 KAIST
전산학과(전산학석사)
1990 KAIST 전산학과(전산학박사)
1990 ~ 현재 강원대학교 컴퓨터과학과 교수
관심분야: 대용량 빅데이터 저장 및 처리,
하둡/맵리듀스 분산/병렬처리 기술,
빅데이터 분석 기법, 데이터 마이닝,
클라우드 컴퓨팅,
데이터 웨어하우스/OLAP 다차원 분석,
데이터베이스 시스템 개발
E-Mail: jhkim@kangwon.ac.kr