https://doi.org/10.5392/JKCA.2021.21.02.049

데이터 마이닝 기반 대학입시를 위한 학교생활기록부 분석시스템

Analysis System of School Life Records Based on Data Mining for College Entrance

양진우*, 김동현*, 임종태, 유재수****

㈜바론*, 충북대학교 정보통신공학부**

Jinwoo Yang(school@ibaron.co.kr)*, Donghyun Kim(dh@ibaron.co.kr)*, Jongtae Lim(jtlim@chungbuk.ac.kr), Jaesoo Yoo(yjs@chungbuk.ac.kr)****

요약

대한민국 교육과정과 입시제도는 수많은 변화들을 통해 진화해 왔다. 현재 우리나라 대학진학률이 약 70%에 육박하고 있고, OECD회원국 중 가장 높다. 이러한 환경 속에 대학진학에 관심이 높고 또한 우리나라 교육제도에서 가장 비율이 높은 수시전형 중 학생부종합전형에 필요한 학교생활기록부의 중요성을 높아져가고 있다. 행복은 성적순이 아니지만 적극적인 학교생활을 통해 나의 미래와 행복을 동시에 찾을 수 있다. 학교생활기록부 분석시스템을 통해 자신에게 맞는 흥미와 진로를 찾을 수 있고, 가고자하는 대학, 학과에 맞는 요소들을 분석하고 보완하여 성공적인 진학에 한발 더 나아갈 수 있다. 학교생활기록부의 각 항목을 3가지로 분리시켜 필요 단어와 불필요한 단어들을 구분해 분석한다. 분석한 데이터를 시각화&수치화 하여 학교생활에서 보완할 수 있는 분석시스템을 구축한다. 기존 선행연구로 단어빈도와 유사도 분석을 이용한 다중주제 회의록 요약 시스템을 응용하여 다른 요소의 문장을 간결하게 요약하고 단어를 추출함으로써 데이터 마이닝을 통한 분석시스템을 활용할 수 있다.

■ 중심어 : | 데이터 마이닝 | 대학입시 | 학교생활기록부 | 분석 시스템 | 빅데이터 |

Abstract

The Korean curriculum and admission system have evolved through numerous changes. Currently, the nation's college entrance rate stands at nearly 70 percent, and it is the highest among OECD members. Amid this environment, the importance of school life records is increasing among students who are interested in going to college and who have the highest percentage in the nation's education system. Happiness is not the order of grades, but I can find my future and happiness at the same time through active school life. Through the analysis system of school life records, you can find interests and career paths suitable for yourself, and analyze and supplement factors suitable for the university and department you want to go to, so that you can take a step further in successful advancement. Each item in the school records is divided into three categories to analyze the necessary and unnecessary words. By visualizing and numericalizing the analyzed data, an analysis system is established that can be supplemented in school life. An analysis system through data mining can be utilized by concisely summarizing sentences of different elements and extracting words by applying the multi-topic minutes summary system using word frequency and similarity analysis as an existing prior study.

■ keyword : | Data Mining | College Entrance Examination | School Life Record Book | Analysis System | Big Data |

* 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2019R1A2C2084257), 과학기술정보통신부 및 정보통신기획평가원의 Grand ICT연구센터지원사업(IITP-2020-0-01462), 2017년 대한민국 교육부와 한국연구재단의 지원(NRF-2017S1A5B8059946), 그리고 중소벤처기업부 '산업전문인력역량강화사업'의 재원으로 한국산학연협회(AURI)의 지원(2020년 기업연계형연구개발인력양성사업, 과제번호 : S2929950)을 받아 수행된 연구임

접수일자 : 2020년 09월 21일 심사완료일 : 2020년 10월 14일

수정일자 : 2020년 10월 14일 교신저자 : 유재수, e-mail : yjs@chungbuk.ac.kr

I. 서 론

대한민국 교육과정과 입시제도는 수많은 변화들을 통해 진화해 왔다. 1940년대 대학별 단독시험제를 시작으로 시험제도의 교육과정들이 뿌리내리기 시작했다. 이후 대입예비고사·본고사, 대입학력고사, 현재 대학수학능력시험까지 우리나라 교육의 진학과 입시를 평가하는 요소로 나타내고 있다. 2000년대 중반까지 대학입시의 강력한 힘으로 작용하는 대학수학능력시험이 2008년 입학사정관제가 도입되면서 점점 약화되기 시작하였다. 이후 입학사정관제를 통해 학생 개개인의 학교생활기록부의 중요성이 증대되었고, 대학 입시제도의 평가요소들이 달라지기 시작하였다. 현재 우리나라 대학진학률이 약 70%에 육박하고 있고, OECD회원국 중 가장 높다. 우리나라 대입제도 70%를 차지하고 있는 수시전형의 학생부종합전형에 필요한 학교생활기록부를 분석하여 변별력을 높이기 위해 분석시스템을 구축한다. 학교생활기록부 분석시스템을 이용하여 현재 내가 가고자하는 대학, 학과에 맞는 키워드를 분석하고 보완하여 좀 더 질 높은 학교생활을 할 수 있도록 도와준다. 자연어로 이루어진 비구조화 자료(학교생활기록부)에서 필요한 정보를 얻어내기 위해 구조화 데이터로 변환 한 후 분석하는 텍스트 마이닝 기법인 word2vec 모델을 활용한다[1]. 또한 데이터 마이닝 기법들 중 하나의 토픽모델링을 통해 비구조화된 텍스트를 자료들의 뭉치로부터 의미있는 주제(토픽)들을 추출해주는 확률모델 알고리즘을 사용하여 기존의 전통적 텍스트 분석방법들에 비해 귀납적으로 적용된다[2]. 본 논문에서는 word2vec모델을 학교생활기록부에 적용하여 각 단어의 벡터 공간에서 거리를 계산하고 요소별 중요도를 심층적으로 학습한다. 학습한 단어의 벡터 값을 이용해 유사단어를 출력하고 한국어 데이터를 좀 더 효과적으로 모델링하기 위해 토픽모델링 적용하여 학교생활기록부에 단어의 벡터 값을 이용해 유사단어를 출력하고 자동으로 문장을 분류한다. 제안하는 방식을 통해 입증하고자 하는 바는 아래와 같다.

첫째, 단어 임베딩 방식의 벡터 표현을 통해 의미가 비슷한 단어끼리 벡터 공간상 비슷한 위치에 놓일 수 있는지 시각적으로 확인하고, 이와 같은 데이터 처리 방식이 학교생활기록부 분석시스템에 어떠한 영향을 미치는지 분석한다[3].

둘째, 학교생활기록부를 자연어 처리, 불용어 처리하여 명사단어를 word2vec모델과 wordcloud 방식을 이용하여 그래프로 시각화한다.

셋째, 시각화된 데이터를 기반으로 학교생활기록부에서 필요한 3가지 항목으로 나누어 토픽 모델링의 방식을 이용해 주제별, 요소별 수치화한다. 이를 통해 견고한 단어분류와 수치화된 분석시스템을 구축한다.

II. 관련연구

현대사회의 빠른 발전으로 소셜미디어와 스마트폰의 등장으로 방대한 데이터들이 쏟아지고 있다. 무분별한 데이터 확산으로 인해 산업 전반에 분석기술은 종요하게 여기지기 시작했다. 쏟아져 나오는 데이터를 분석·이용하여 부가가치를 창출하고 결과를 분석하는 기술을 빅 데이터(Big Data)라고 한다[4]. 이러한 빅 데이터는 중·고등학교의 학생들에게도 학교생활기록부의 문서를 통해 적용이 되고 있다. 다양한 대입제도를 통해 학교생활기록부의 구성요소들이 학생들의 특성에 맞게 세밀해지고 구체적으로 기록되어지고 있다. 그로 인해 학교생활기록부의 방대한 양의 데이터를 파악하기엔 오랜 시간이 걸린다. 하지만 학교생활기록부를 분석하는데 텍스트마이닝 등의 분석 기법을 사용한 연구는 미흡한 상태이다. 따라서 본 논문에서는 텍스트마이닝 기법을 활용한 학교생활기록부 분석시스템을 구축하여 활용할 수 있도록 한다. 본 논문에서는 데이터의 형태에 따라 여러 가지 분석기법 중 데이터 마이닝(Data Mining) 분석기법을 이용하여 효율적인 학교생활기록부 분석시스템을 구축한다. 데이터 마이닝의 종류에는 텍스트 마이닝, 웹 데이터 마이닝, 공간 데이터 마이닝 등이 있다[1-31]. 그 중 텍스트 마이닝의 경우, 분석대상은 문서, 웹상자의 문자와 같은 텍스트로써 단어들의 의미를 파악하여 효과적으로 분석·분류해주는 기법이다[5]. 또한 텍스트 데이터에 대한 분석을 통해 유의미한 정보를 얻기 위한 분석을 텍스트 마이닝(Text Mining)이라고 한다. 텍스트 마이닝의 word2vec 모델

에서는 문서 내에서 주위 단어의 분포가 가까운 단어일수록 결과로 산출되는 벡터 값이 유사해지며[6], 산출된 단어 표상의 값이 비슷한 단어는 의미가 유사한 것으로 간주된다[2]. 같은 의미를 지니는 단어들이 유사한 벡터 값을 지니는 것이 실험을 통해 보고되었다[7]. 사례로는 본 연구의 선행연구인 단어빈도와 유사도 분석을 이용한 다중주제 회의록 요약시스템을 바탕으로 단어빈도와 단어 간의 유사도 분석을 통해 문장을 분리하고 다양한 주제로 진행되는 토의와 토론에 대한 의사결정 결론을 도출한다. 이를 통해 효율적인 문서 요약방법과 유사한 회의록 요약 방법을 개선하였다. 또한 Joshi et al. (2015)은 E-Commerce 분야에서 word2vec을 통한 학습 자질이 개체명 인식 모델의 성능을 향상 시킬 수 있다는 것을 보고하였는데, eBay의 핸드폰, 구두, 시계, 의류 분류 내 자료를 바탕으로 개체명 인식 실험을 수행하였다. 그 결과, 모든 분류의 자료에서 word2vec로 학습된 벡터 기반 학습 자질을 사용하였을 때 개체명 인식 모델의 성능이 향상되었다[5]. 토픽모델링 기법은 기존의 텍스트 분석과는 달리 선험적 이론에 기초한 사전적 코딩범주의 투입을 요구하지 않으며, 전통적인 가까이서 읽기의 방법으로는 방대한 양의 텍스트뭉치에서 유의미한 토픽들을 자동적으로 산출해준다[8]. 또한 테스트의 분석 이전에 사전적 지식을 요구하지 않으며, 문서를 어휘의 자루(bag or words)로 가정하여 어휘들의 관계 속에서 잠재적 의미구조를 포착하고 한글 자연어처리과정의 안정성이 높이면서 문서 분석시스템에 활용도가 높다[2].

III. 학교생활기록부 분석시스템 구조 및 설계

1. 분석시스템 구조

[그림 1]은 제안하는 학교생활기록부 분석시스템의 순서도를 보여준다. 본 논문에서 제안하는 시스템은 학교생활기록부 문서를 기반으로 학생 개인마다 특성과 활동들을 알기위해 텍스트 마이닝 도구인 파이썬의 Word2vec 와 LDA 토픽모델링 패키지를 사용하였다. 모든 문장을 자연어처리, 불용어 처리 후 분류된 단어 중 최대빈도단어 순서대로 주제별로 항목을 분류시켜 그래프로 시각화하였다.

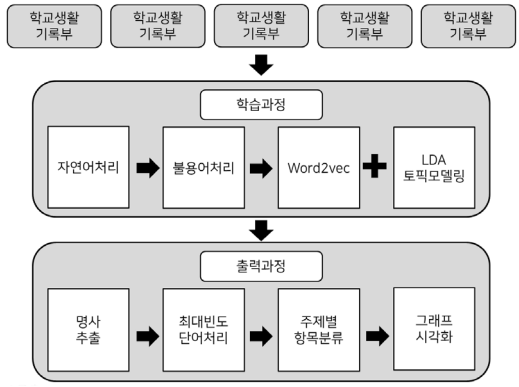


그림 1. 분석시스템 순서도

2. 자연어 처리

자연어 처리(Natural Language Processing, NLP)는 사람의 언어를 컴퓨터 프로그램이 이해할 수 있도록 하는 연구 분야이다. 컴퓨터와 사람의 언어 사이의 상호작용에 대해 자연스럽게 말하는 언어, 즉 자연어를 형태 분석과 의미 분석, 문서 분석 등의 과정을 통해 컴퓨터가 처리하여 변환시키고 이를 통해 사람의 편의성에 입각해 표, 그래프, 음성 또는 그래픽 등으로 생성하는 작업 전반을 일컫는다[9]. 자연어 처리 기술의 분석 처리 과정에는 전처리(띄어쓰기, 문장 분할, 철자 오류 관리), 형태소 분석(품사태깅, 미등록어 추정), 고정밀 구문 분석, 의미 분석과 같은 구조 분석을 바탕으로 한 담화 분석, 정보 요약, 문서 분류 기술 등이 있다[9]. 형태소 분석(Morphological Analysis)이란 실제 문장에서 하나의 단어가 여러 가지 형태로 변형되어 사용된다. 예시로 '하는', '해서', '하니' 같은 단어들이 모두 '하다'라는 하나의 단어로 볼 수 있다. 이렇게 일반적인 형태로 분석하면 단어의 수를 줄여 분석의 효율성을 높일 수 있게 된다. 형태소 분석에 있어 한국어는 조사와 복합명사 등을 분리하기 어려워 상대적으로 힘든 언어라고 할 수 있다[8].

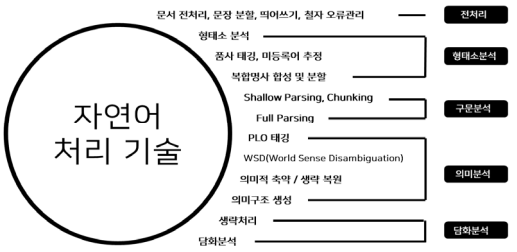


그림 2. 자연어 분석 처리과정

2.1 벡터 공간 모델

벡터 공간 모델(VSMs: Vector Space Models)은 단어를 벡터화 하여 연속적인 벡터 공간 안에 표현하고 의미론적으로 비슷한 단어들이 서로 가까운 곳에 있도록 하는 방법이다[10]. 또한 정보필터링, 문서 내에서의 정보검색, 색인과 유사도를 계산하기 위한 수학모델로서, 다차원 선형공간에서의 벡터 정보를 이용하여 자연어를 포함한 문서의 중요도를 분석하고 단어의 벡터 표현을 word-embeddings라고 한다. 벡터 공간 모델은 모두 분포 가설(Distributional Hypothesis)에 입각한다[6]. 예시로 [그림 3]의 문서집합(Corpus)의 A1, A2, A3, A4, A5의 5개의 문서를 벡터공간에 표현한다. 주요단어 '대학'과 '전공'을 2차원 벡터공간에 표현할 때 A1 문서는 '대학' 단어만 표현되어 '대학' 축과 나란한 벡터로 표현되고, A1, A2문서는 '전공' 축과 나란한 벡터로 표현되고, A5 문서는 '대학' 단어가 2번, '전공' 단어가 1번 있으므로 대각선 위치인 (1,2) 위치의 벡터로 표현된다[6].

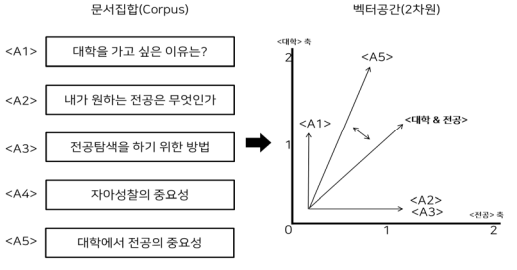


그림 3. 벡터공간모델의 개념

2.2 Word2vec

Word2vec은 Word-Embedding를 위한 하나의 방법이다. 많은 양의 텍스트 데이터, 텍스트 코퍼스를 입력 받아 주어진 텍스트에서 벡터 집합을 생성한다[15]. Word2vec 방법론은 신경 연결망 기법에 기반을 둔 방식을 통해 단어들의 의미를 특정 차원의 벡터 값으로 계산, 표현하는 비지도 학습 기법이다. 두 가지 모델로 CBOW(Contiunous Bag-of-Word)와 Skip-gram 모델이다[7]. [그림 4]을 통해 w(t)를 입력 단어로 정했을 때, 입력 단어를 기준으로 주위 지정된 범위 내의 단어들을 예측하는 기법이다. Word2vec 모델에서는 주위 단어의 분포가 가까운 단어일수록 결과로 산출되는 벡터 값이 유사해진다[11].

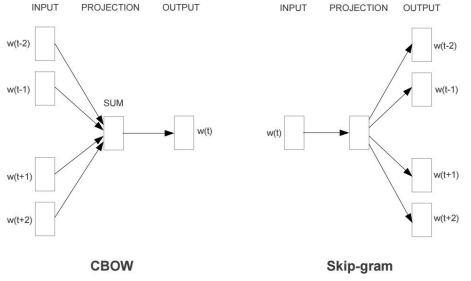


그림 4. CBOW 모델과 Skip-gram 모델의 구조

Mikolov et al.(2013)에서는 Skip-gram 기법을 〈식 1〉의 식은 평균 log 확률을 최대화시키는 계산을 수행하도록 작성되었다고 설명하고 있다[7].

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}|w_t) \qquad (1)$$

Skip-gram 모델은 단어 주변의 단어들을 예측하는 신경망 모델로, 간단한 신경망 훈련을 통해 은닉층에서의 가중치를 학습시켜 단어벡터를 얻는다. 윈도우 사이즈(window-size)를 지정할 수 있는데, 해당 단어를 중심으로 주변 단어를 몇 개까지 학습하는지 나타내는 변수이다. [그림 5]는 Skip-gram 과정을 통한 문장 분석의 예시이다. [그림 5]를 통해 "나는 충북대학교 전자정보공학과를 다니면서 파이썬을 배웠다" 예시문장을 사용해 학습 샘플로 적용해본다. 윈도우 사이즈 3로 설정되었을 때, 해당 문장의 단어 앞뒤로 3개의 단어와 함

께 학습한다. 이 때 조사나 어미를 제거한다. 학습 후 특정단어를 입력하는 모든 단어들이 입력 값과 얼마나 가까운 지 확률로 출력하게 되어 많이 학습될수록 더 높은 확률을 출력 값을 갖게 된다. 예를 들어 입력 값이 '전자정보공학과'일 때, 출력단어 '충북대학교'와 '배웠다' 단어 중에서는 '충북대학교' 단어가 윈도우 사이즈에 포함되어 더 높은 확률 값을 출력할 것이다. [그림 6]은 그에 대한 word2vec 모델 신경망 학습구조 그림이다[12].

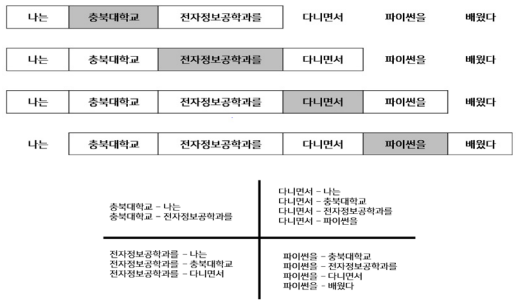


그림 5. Skip-gram 과정을 통한 문장 분석의 예

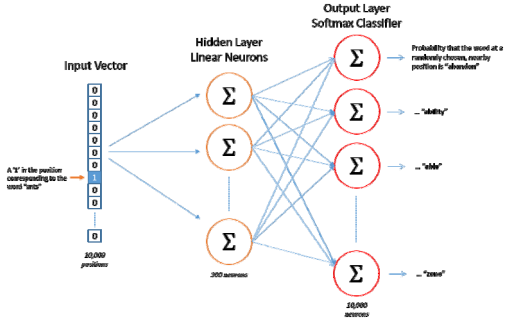


그림 6. Word2vec 모델 신경망 학습구조

2.3 LDA 토픽모델링

텍스트 마이닝 분석에서 가장 많이 활용되고 있는 LDA(Latent Dirichlet Allocation)는 기존의 LSA(Latent Semantic Analysis) Deerwester(1990)와 PLSA(Probabilistic Latent Semantic Analysis) Huang(2008)등의 약점을 보완한 방법으로 문서 집단 내에 잠재되어 있는 주제(Topic)들을 추출할 수 있게 하는 생성확률모델(Generative probabilistic model)이다[13]. 또한 구조화되지 않은 방대한 문헌집단에서 주제를 찾아내기 위한 알고리즘으로, 맥락과 관련된 단서들을 이용하여 의미를 가진 단어들을 클러스터링하여 주제를 추론하는 모델이다[13]. LDA는 문서 내 여러 주제 중 각 주제에 속할 확률분포와 특정 단어가 주제에 속할 확률분포를 깁스 샘플링(Gibbs Sampling)을 통해 구하고, 이미 관찰된 변수 (observed variable)를 통해 확률을 계산하고 토픽을 생성하는 사후 추론방법이다. 단어 벡터는 집합을 문서벡터로 표현 될 수 있고 벡터간의 코사인(Cosin) 각으로 유사도를 계산한다. 토픽 모델은 사서(narative), 구문법(syntax) 등 한 문서 내에서 단어의 동시발생(co-occurence)하는가를 측정한다[10]. [그림 7]은 LDA 알고리즘을 통한 문서생성 과정을 보여준다. [그림 7]에서 α와 β은 전체 문서에서 동일하게 적용되는 파라미터(매개변수)이다. α에 의해 문서별 토픽의 비율(θ)이 결정되고 β에 의해서 토픽별 단어의 분포(ϕ_k)가 결정된다. θ에 의해 단어별 토픽 할당(Z)이 결정된다. 단어의 토픽들을 나타내는 $Z_{d,i}$ 값과 토픽별 단어 분포를 나타내는 ϕ_k값에 의해 단어 $W_{d,i}$가 결정된다.

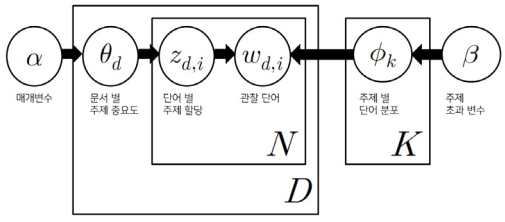


그림 7. LDA 알고리즘을 통한 문서생성 과정

$$p(z_{d,i}=j|z_{-i},w)=\frac{n_{d,k}+\alpha_j}{\sum K_{i=1}(n_{d,i}+\alpha_i)}\times\frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum V_{j=1}v_{k,j}+\beta_j}=AB \quad (2)$$

[표 1] LDA 모델 수식 정리표를 보면 A는 d번째 문서가 j번째 토픽과 맺고 있는 연관성 강도를 나타낸다. B는 d번째 문서의 n번째 단어 ($w_{d,n}$)가 j번째 토픽을 맺고 있는 연관성의 강도를 나타낸다.

표 1. LDA 모델 수식 정리표

표기	내용
$n_{d,k}$	k번째 토픽에 할당된 d번째 문서의 단어빈도
$v_{k,w_{d,n}}$	전체 말뭉치에서 k번째 토픽에 할당된 단어 $w_{d,n}$의 빈도
$w_{d,n}$	d번째 문서에 n번째로 등장한 단어
α	문서의 토픽 분포 생성을 위한 디리클레 분포 파라메터
β	토픽의 단어 분포 생성을 위한 디리클레 분포 파라메터
K	사용자 지정하는 토픽 수
V	말뭉치에 등장하는 전체 단어 수
A	d번째 문서가 k번째 토픽과 맺고 있는 연관성 정도
B	d번째 문서의 n번째 단어 ($w_{d,n}$)가 k번째 토픽과 맺고 있는 연관성 정도

IV. 학교생활기록부 분석시스템 구현

본 논문에서는 학교생활기록부 문서를 명사(nouns) 형태로 분석하고 이를 토대로 word2vec모델 skip-gram 알고리즘, LDA 토픽 모델링을 이용하여 학습한 데이터를 활용해 학교생활기록부 문서에서 학생의 특성과 희망하는 진학목표에 대해 분석한다.

1. 시스템 환경 및 데이터

본 연구를 위하여 사용된 시스템의 구현 환경은 다음과 같다. 윈도우 10과 리눅스 환경에서 C 언어 기반으로 개발된 파이썬 3.7.5 언어를 사용하였다. 원활한 연구를 위해 한국어 분석 도구 라이브러리 KoNLPy패키지 사용해 문장분리, 토큰화, 정규화, 품사 태깅까지 한꺼번에 수행하였다. 또한 Gensim 패키지의 word2vec 모델을 사용하였다.[9]

시스템에 사용된 데이터는 충청북도 청주시 소재 일반계 고등학교 졸업생의 학교생활기록부 4개의 데이터를 샘플 데이터로 사용하였다. 데이터는 2015년 기준으로 학교생활기록부 작성 지침에 따라 작성되었다. 시스템 분석의 신뢰도를 높이기 위해 학교생활기록부의 총 10가지 항목 중 3가지 항목만 분석한다. 다음 항목들은 [표 2]를 참고한다

표 2. 학교생활기록부 분석시스템 데이터 활용 항목

학교생활기록부 전체 항목	데이터 분석에 사용할 항목
1. 인적사항 2. 학적사항 3. 출결상황 4. 수상경력 5. 자격증 몇 인증 취득상황 6. 진로 희망사항 7. 창의적 체험활동 상황 (자율활동, 동아리활동, 봉사활동, 진로활동) 8. 교과학습발달상황 (세부능력 및 특기사항) 9. 독서활동상황 10. 행동특성 및 종합의견	1. 창의적 체험활동 상황 (자율활동, 동아리활동, 봉사활동, 진로활동) 2. 교과학습발달상황 (세부능력 및 특기사항) 3. 행동특성 및 종합의견

학교생활기록부의 3가지 항목(창의적 체험활동 상황, 교과학습발달상황, 행동특성 및 종합의견)에서 학생의 특성과 진학목표를 가장 뚜렷이 확인할 수 있다[6]. 3가지 항목의 모든 데이터를 명사(nouns)형태의 단어로 추출하여 빈도수를 확인한다. 명사(nouns)형태로 추출하기 위해서 KoNLPy의 Okt, Counter, matplotlib, pytagcloud, wordcloud 모듈을 활용해 데이터를 분석하였다[14]. [그림 8]에서는 모듈 이용해 명사(nouns)형태로 불용어처리, 단어 추출, 빈도수 분석을 위한 소스코드과정이다. 처리 과정에서 중요한 한국어 자연어 처리를 위해 형태소 분석기로 Okt(Open Korea Text), Counter 모듈을 사용해 학교생활기록부를 분석한다. 문서 전체의 단어, 명사(nouns)로 추출도 중요하지만 불필요한 단어를 제거하는 부분도 중요하다. '은', '임', '나', '등', '알', '함' 등 조사 및 한 글자 단어와 '학년', '번호', '학기', '관련', '고등학교', '대해', '내용', '위해' 등 불필요 중복단어, 불용어(명사(nouns)는 아니지만 단어로 인식하는 단어) 제거도구로는 파이썬 프로그램 NLTK 패키지의 stopword를 사용하였다[15]. 또한 if문과 for문을 사용하여 데이터 분석을 원만히 진행하였다.

```
from konlpy.tag import Okt
nlpy = Okt()
nouns = nlpy.nouns(lines)
print(nouns)

from collections import Counter

count = Counter(nouns)
```

Konlpy 패키지 Okt 모듈

```
for n, in count.most_common(100):
    dics={'tag':n, 'count':c}

if len(dics['tag'])>2 and len(tags)<=15:
    tag_count.append(dics)
    tags.append(dics['tag'])

for tag in tag_count:
    print("{:<14}".format(tag['tag']),end='\t')
    print("{}".format(tag['count']))
```

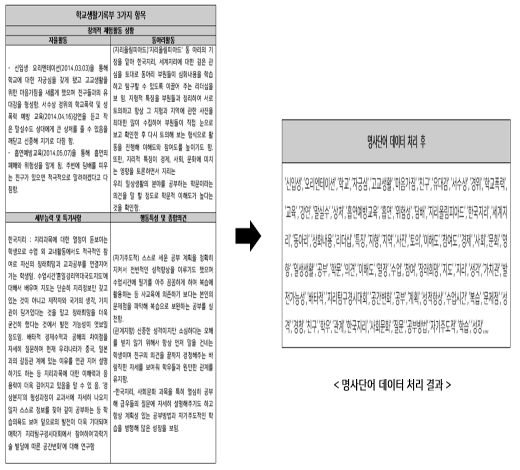
Counter 모듈 처리과정

```
myfile=file_stopword.readlines()

stopwords = []

for oneword in myfile:
    stopwords.append(oneword.strip('\r\n'))

stopwords = set(stopwords)

print(stopwords)
print('-'*40)

for word, count in sortedData[:500]:
    if not(word in stopwords):
        if(count >= 15 and len(word)>=2):
            wordinfo.append((word, count))

print(wordinfo)

saveWordCloud(wordinfo)
showGraph(wordinfo)

print('finished')
```

stopword 모듈 처리과정

그림 8. Konlpy 패키지 모듈 처리 과정 소스코드

[그림 9]는 Okt 모듈을 활용한 처리과정 및 결과를 보여준다. [그림 9] 소스코드를 사용해 [표 2] 학교생활기록부 4가지 항목의 부분을 데이터 분석한 예시이다.



['지리학', '기타', '도시', '개발', '도시', '개발', '지리학', '지도', '사진', '문화', '변성', '수종', '하늘', '우주', '존재', '수도', '핵심', '이론', '해리스', '도시', '계획', '독서', '개발', '도상', '도시', '건설', '다짐', '계기', '인식', '냉정', '생활', '예방', '흡연', '교내', '인생', '자율', '강연', '대처', '공부', '계획', '공간변화', '열정', '학문', '의견', '수업', '참여', '친구', '학습', '성장', '관계', '학우', '생각', '관광', '의견', '리더십', '지역', '경제', '사회', '문화', '영향', '경청', '질문', '자기주도', '성장', '교육', '강연', '친구', '신입생', '참여', '적극', '문제해결', '태도', '관심', '이해', '노력', '공부', '관심', '전략', '계획', '절차', '생각', '다짐', '발전', '영어', '상황', '문화', '계기', '환경', '성적', '특기', '평가', '스스

그림 9. Okt 모듈을 활용한 처리과정 및 결과

분석한 핵심단어를 시각화하기 위해 wordcloud 패키지를 이용한다. 해당 문서에서 중요단어들을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하고 단어의 크기에 따라 빈도수를 표현하는 기법이다. 주로 방대한 양의 데이터를 분석할 때 특정 정보를 도출하기 위해 활용한다. 학교에 다니는 학생이라면 모두 학교생활기록부를 가지고 있다. 학생마다 데이터의 양이 다르기 때문에 진학과정에서 학교생활기록부의 중요성이 나타난다. 이러한 학교생활기록부는 평균적으로 6500자의 데이터 중 명사단어 추출 단어는 1800자의 데이터로 집계되었다. 수많은 단어 중 빈도수를 나타내기 위해 Counter 모듈을 통해 [그림 10]은 15번 중복 이상 단어 중 가장 많이 나타나는 최대 빈도 단어를 그래프로 시각화한 결과이다.

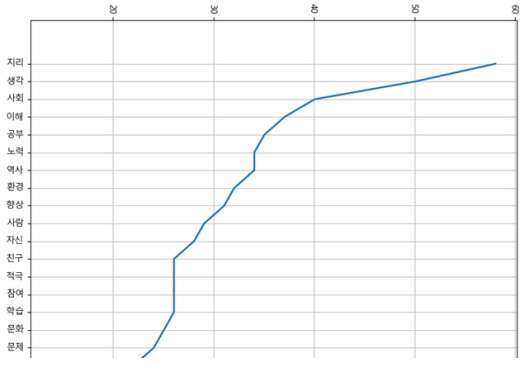


그림 10. 그래프 시각화

[그림 11]은 학교생활기록부를 wordcloud를 통해 한눈에 보기 간편하게 시각화한 결과를 보여준다. '지리'가 가장 많은 빈도로 나타났고 그 다음으로는 '생각', '사회', '이해', '역사', '공부' 등 큰 글자 순으로 빈도수가 표현되었다. 본 논문에서 제안하는 분석시스템은 학교생활기록부를 단어 분석과정을 거쳐 최대빈도 단어와 wordcloud 시각화를 표현하고 LDA 토픽모델링으로 4개의 학교생활기록부의 주제를 출력하기 위해 LDA 모델의 변수를 선정한다. 정확한 토픽 데이터를 얻기 위해서는 처리된 텍스트의 데이터 품질, 토픽의 다양성, 토픽 모델링 알고리즘, 알고리즘에 공급된 주제 수, 알고리즘 튜닝 파라미터로 정해진다. 토픽 모델에 필요한 사전과 코퍼스를 생성한다. 사전을 만들고 bag-of-words 코퍼스로 변환하여 필요한 사전과 코퍼스를 생성 및 저장한다. LDA에서 데이터를 3가지 항목으로 찾도록 요청한다.

그림 11. 최대빈도 단어 수와 wordcloud 시각화

[그림 12]는 학교생활기록부의 3가지 항목(창의적 체험활동 상황, 교과학습발달상황, 행동특성 및 종합의견)에 대한 데이터의 모델링된 가장 최적화된 단어선정이다.

```
from gensim import corpora
dictionary = corpora.Dictionary(text_data)corpus = [dictionary.doc2bow(text) for text in text_data]
import pickle
pickle.dump(corpus, open('corpus.pkl', 'wb'))
dictionary.save('dictionary.gensim')

import gensim
NUM_TOPICS = 3
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = NUM_TOPICS, id2word=dictionary, passes=15)
ldamodel.save('model5.gensim')
topics = ldamodel.print_topics(num_words=4)

for topic in topics:
    print(topic)

ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = 3, id2word=dictionary, passes=15)
ldamodel.save('model3.gensim')
topics = ldamodel.print_topics(num_words=4)
for topic in topics:
```

```
(0, '0.032 * "공부" + 0.027 * "적극" + 0.009 * "aid" + 0.002 * "태도"')
(1, '0.024 * "참여" + 0.015 * "문제해결" + 0.037 * "관심" + 0.001 * "이해"')
(2, '0.019 * "질문" + 0.017 * "열정" + 0.022 * "노력" + 0.010 * "인성"')
```

그림 12. 토픽 모델 생성 소스코드 및 결과 값

[표 3] 학교생활기록부 3가지 항목을 Topic1 (창의적체험활동상황), Topic2 (교과학습발달상황), Topic3 (행동특성 및 종합의견)으로 선택하여 전체의 단어 개수만큼 길이를 표현한다. 각 Topic에서 [그림 5] LDA 알고리즘을 통한 문서생성 과정을 살펴보면 ϕ_k는 k번째 토픽에 해당하는 벡터이다. ϕ_k의 각 요소는 확률이므로 [표 3] 열의 모든 요소의 합은 1이 된다. 학생1의 Topic1의 가장 많이 언급된 핵심단어는 '관심', Topic2는 '열정', Topic3는 '노력'으로 나타난다. 이외의 단어 빈도수를 통해 각 Topic에 보완해야 할 핵심 단어들을 생각해볼 수 있다.

표 3. 높은 빈도 단어의 Topic 수치표

학생1	Topic1	Topic2	Topic3
공부	0.003	0.007	0.041
참여	0.117	0.179	0.123
질문	0.000	0.277	0.000
적극	0.004	0.004	0.053
문제해결	0.184	0.011	0.074
열정	0.009	0.339	0.099
태도	0.044	0.021	0.183
관심	0.467	0.017	0.048
이해	0.145	0.099	0.073
노력	0.027	0.046	0.306

[표 4]는 4명의 학생의 학교생활기록부 3가지 항목을 [표 3] 추출된 단어를 항목마다 수치화한 것이다. 학생1의 경우 Topic1,2,3 중 Topic3가 높은 수치로 핵심단어들이 표현된다. [표 3]과 동일하게 ϕ_d는 d번째 문서가 가진 토픽 비중의 나타내는 벡터이다. 전체 토픽 개수 K만큼의 길이를 가진다. 또한 ϕ_d는 확률이므로 [표 4]의 행의 모든 요소의 합은 1이 된다. LDA 기법을 사용해 쉽고 빠르게 구현이 가능하고 Topic별 단어의 잠재적 의미를 이끌어낼 수 있다.

표 4. 학교생활기록부 전체 Topic 수치표

	Topic1	Topic2	Topic3
학생1	0.246	0.339	0.415
학생2	0.423	0.299	0.278
학생3	0.179	0.365	0.456
학생4	0.314	0.284	0.402

V. 결 론

본 논문에서는 현재 교육분야에서 관심도가 높은 학교생활기록부를 데이터 마이닝 기반 기법들을 이용해 연구하였다. 방대한 양의 데이터를 word2vec모델과 LDA 토픽모델링 과정을 통해 특정단어 추출, wordcloud

시각화, 토픽 수치화로 분석하였다. 학교생활기록부 분석시스템을 기반으로 학교생활기록부의 가장 중요한 3가지 항목(창의적 체험활동 상황, 교과학습발달상황, 행동특성 및 종합의견)의 데이터를 간편하게 볼 수 있도록 구현하였다. 실제 학생부종합전형의 정성적 평가기준에 도움이 될 수 있는 학교생활기록부를 정량적인 평가를 실시한 것이다. 학년별로 분석시스템을 통해 자신의 학교생활기록부를 확인한다면 자신에게 부족한 활동요소와 나에게 필요한 문장 또는 단어들을 보완하고 학교생활 충실도를 올릴 수 있을 것이다. 또한 학생의 적성(흥미, 성격, 능력)을 파악하는 데 있어 도움이 될 것이라 생각한다. 결과적으로 학교생활기록부 분석시스템을 통해 자기이해와 진로탐색이 필요한 학생, 전략적인 학습계획이 필요한 학생, 진학을 체계적으로 준비하는 학생 등 다양한 학생들에게 도움을 줄 수 있다. 데이터 마이닝 기법 이외에도 더 세심하고 견고한 분석시스템을 구축하기 위해서는 수많은 데이터자료(학교생활기록부)가 필요할 것으로 보인다. 데이터자료의 접근권한은 개인정보 보안 문제로 본인 이외의 부모님, 해당 선생님에게만 접근이 가능하여 데이터 수집의 어려움을 겪을 수 있다. 향후 연구 과제로서 추가적인 데이터 자료수집과 불용어처리 과정의 고급화, LDA 토픽 수 증가가 필요하다고 생각된다.

참 고 문 헌

[1] 강민영, *텍스트 마이닝 기반 개인역량평가 시스템 설계 및 구현,* 가천대학교 대학원 모바일소프트웨어학과, 석사학위논문, 2014.

[2] 구주나, *한국어 텍스트 마이닝의 특징 고찰 및 실제 빅데이터에의 적용,* 숙명여자대학교 대학원 통계학과 통계학전공, 석사학위논문, 2013.

[3] 허강호, *단어빈도와 유사도 분석을 이용한 다중주제 회의록 요약시스템,* 충북대학교 대학원 전자정보공학과 전자정보공학전공, 석사학위논문, 2018.

[4] 오석원, *텍스트 마이닝의 실제적 활용에 관한 고찰,* 고려대학교 대학원 경제통계학과, 석사학위논문, 2011.

[5] 황명하, *텍스트 마이닝 기반 국제표준 트렌드 분석시스템 개발에 관한 연구,* 과학기술연합대학원대학교 정보통신네트워크공학전공, 석사학위논문, 2018.

[6] 이현주, *텍스트 마이닝 기법을 활용한 학교생활기록부 문서 내용의 시각화,* 이화여자대학교 교육대학원 컴퓨터교육전공, 석사학위논문, 2019.

[7] 김윤덕, *Word2vec을 이용한 위키피디아 텍스트 데이터 분석시스템 구현,* 숭실대학교 소프트웨어특성화대학원, 석사학위논문, 2016.

[8] 배진아, *자연어처리 기반 교통 연구 통합검색에 관한 연구,* 충남대학교 대학원 컴퓨터 공학과 데이터 및 소프트웨어공학전공, 석사학위논문, 2019.

[9] 한남기, *word2vec 학습 자질을 사용한 새로운 한글 개체명 인식 모델 제안,* 연세대학교 대학원 문헌정보학과, 석사학위논문, 2015.

[10] 백시온, *한국어 토픽모델링을 위한 단어 임베딩 활용 가능성 탐색,* 서울대학교 대학원 협동과정 인지과학전공, 석사학위논문, 2018.

[11] 김성민, *단어 벡터화를 통한 특징 단어 기반 문서 관련성 분석 방법 영화 스크립트 비교를 중심으로,* 건국대학교 컴퓨터공학과, 석사학위논문, 2016.

[12] 김정미, *word2vec모델을 활용한 RNN기반의 문서 분류에 관한 연구,* 인하대학교 컴퓨터공학과, 석사학위논문, 2018.

[13] 이앞길, *LDA 토픽 모델링과 Word2vec을 활용한 유사 특허문서 추천 연구,* 한밭대학교 창업경영대학원 빅데이터비즈니스학과, 석사학위논문, 2018.

[14] 안드레아스 뮐러, 세라 가이도, *파이썬 라이브러리를 활용한 머신러닝,* 한빛미디어(주), 2017.

[15] 민형기, *파이선으로 데이터 주무르기,* 비제이퍼블릭, 2017.

[16] 남춘호, "일기자료 연구에서 토픽모델링 기법의 활용 가능성 검토," 서울대학교 비교문화연구소 학술지, 2016.

[17] 정성원, *벡터공간모델을 활용한 상품추천 알고리즘에 관한 실증연구,* 숭실대학교 대학원 IT정책경영학과, 박사학위논문, 2019.

[18] 강전학, *특허 토픽모델링과 동시인용분석을 활용한 클라우드 컴퓨팅 유망기술 도출,* 서울과학기술대학교 일반대학원 데이터사이언스, 석사학위논문, 2017.

[19] 김정미, *word2vec모델을 활용한 RNN기반의 문서 분류에 관한 연구,* 인하대학교 컴퓨터공학과, 석사학위논문, 2018.

[20] Harris, Zelling, "Distributional Structure," Word, Vol.10, No.2/3, pp.146-62, 1954.

[21] Mikow, Tomas et al. "Efficient estimation of word rpresentations in vector space," arXiv preprint arXiv, 1301.3781, 2013.

[22] 박은정, 조성준, “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지,” 제26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.
[23] https://xop6.com/list-of-english-stop-words/
[24] Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents, Icml, No.14. pp.1188-1196, 2014.
[25] 김상윤, *효율적인 오피니언 마이닝을 위한 Word Embedding 기반 대체어 자동 추출*, 숭실대학교 컴퓨터학과 대학원, 석사학위논문, 2016.
[26] 잘라지 트하나키, *파이썬 자연어 퍼리의 이론과 실제*, 에이콘출판주식회사, 2018.
[27] 제이크 밴더플래스, *파이썬 데이터 사이언스 핸드북*, 위키북스, 2017.
[28] 바라가브 스리니바사 디지칸, *자연어 처리와 컴퓨터 언어학*, 에이콘출판주식회사, 2019.
[29] 서대호, *잡아라! 텍스트 마이닝 with 파이썬*, 비제이퍼블릭, 2019.
[30] A. Karami, C. N. White, K. Ford, S. Swan, and M. Y. Spinel, “Unwanted advances in higher education: Uncovering sexual harassment experiences in academia with text mining,” Information Processing & Management, Vol.57, No.2, pp.102-167, 2020.
[31] D. Shin and J. Shim, “A Systematic Review on Data Mining for Mathematics and Science Education,” INTERNATIONAL JOURNAL OF SCIENCE AND MATHEMATICS EDUCATION, 2020).

저 자 소 개

양 진 우(Jinwoo Yang) 정회원

- 2015년 2월 : 서원대학교 정보통신공학과(공학사)
- 2015년 7월 : 와이즈멘토리더십센터 강사
- 2020년 2월 : 충북대학교 전자정보공학과(공학석사)
- 2016년 2월 ~ 현재 : ㈜바론 대리

〈관심분야〉 : 데이터베이스 시스템, 소셜 네트워크 서비스,빅데이터 처리 등

김 동 현(Donghyun Kim) 정회원

- 2002년 2월 : 충북대학교 구조시스템공학과(공학사)
- 2006년 2월 : 충북대학교 구조시스템공학과(공학석사)
- 2009년 8월 : 충북대학교 구조시스템공학과(공학박사수료)
- 2017년 2월 ~ 현재 : 충북대학교 빅데이터학과 박사과정
- 2003년 ~ 2005년 : 서울대학교 선진화 연구단 실무담당
- 2006년 ~ 2007년 : 교차로, 내일신문 교육 칼럼리스트
- 2006년 ~ 2011년 : ㈜바른교육 창업콘텐츠 연구소장
- 2016년 ~ 현재 : ㈜바론 대표이사

〈관심분야〉 : 데이터베이스 시스템, 이동 P2P 네트워크, 소셜 네트워크 서비스, 빅데이터 처리 등

임 종 태(Jongtae Lim) 정회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2011년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2015년 8월 : 충북대학교 정보통신공학과(공학 박사)
- 2015년 9월 ~ 2019년 8월 : 충북대학교 정보통신공학과 Postdoc.
- 2019년 10월 ~ 현재 : 충북대학교 전자정보대학 정보통신 공학부 초빙조교수

〈관심분야〉 : 소셜 미디어, 빅데이터, 인텔리전트 데이터베이스, 시공간 데이터베이스, 위치기반 서비스 등

유 재 수(Jaesoo Yoo) 종신회원

- 1989년 2월 : 전북대학교 컴퓨터공학과(공학사)
- 1991년 2월 : 한국과학기술원 전산학과(공학석사)
- 1995년 2월 : 한국과학기술원 전산학과(공학박사)
- 1995년 2월 ~ 1996년 8월 : 목포대학교 전산통계학과 전임강사
- 1996년 8월 ~ 현재 : 충북대학교 전자정보대학 정교수

〈관심분야〉 : 데이터베이스 시스템, 멀티미디어 데이터베이스, 센서 네트워크, 바이오인포매틱스, 빅데이터 등