발 간 등 록 번 호
11-1371029-000236-01

국가서지를 활용한 주제명 자동 분류 적용방안 연구

2022. 09

문화체육관광부
국립중앙도서관

제 출 문

국립중앙도서관장 귀하

본 보고서를 『국가서지를 활용한 주제명 자동 분류 적용방안 연구』의 최종보고서로 제출합니다.

2022년 9월 25일

연구수행기관 : **경북대학교 산학협력단**
책 임 연 구 원 : **이용구 교수**(경북대학교 문헌정보학과)
공 동 연 구 원 : **이종택 교수**(경북대학교 컴퓨터공학과)
연 구 보 조 원 : **이혜경**(경북대학교 문헌정보학과 박사수료)
나상오(경북대학교 문헌정보학과 학부과정)

이 연구는 2022년도 국립중앙도서관 연구용역비로 수행되었으며, 본 연구에서 제시한 대안이나 의견은 국립중앙도서관의 공식 의견이 아니라 수행 기관의 견해임을 밝힙니다.

<목 차>

<표 목 차>

< 그 림 목 차>

Ⅰ. 서론

1. 연구의 필요성과 목적

2002년 국립중앙도서관은 「국립중앙도서관 주제명표목표 개방을 위한 고품질화 연구」를 통해 주제명표목표 구축 방법과 수록 범위, 데이터 구조에 관한 개선안 및 주제명표목표 관리시스템의 기능과 이용자의 효율적인 접근방식을 위한 검색 절차 개선안 등을 도출하였으며, 주제명표목표의 개방과 이용 활성화를 도모하고자 하였다. 이 중 주제명표목표의 검색 효율을 증대시키고자 하는 개선 방안으로 현재 존치 중인 시스템보다 부가적으로 기능이 추가된 주제명 DB 시스템을 도입을 제안하였으며, 주제명표목표의 용어 관계 중 연관관계의 경우 업무자가 수동으로 추출하고 정의하는 것이 타 관계에 비하여 상대적으로 객관성의 결핍 우려가 있으므로, 기계적인 처리 방법을 함께 활용하여 업무의 효율과 용어 선정 등의 정확성을 높이는 방법을 고안할 필요성을 언급하였다.

2012년부터 국립중앙도서관은 서지데이터에 주제명표목을 부여하여, 이용자의 주제 정보 탐색을 원활하게 지원하기 위하여 노력하고 있으나, 사서의 전문성을 기반으로 하는 주제명 부여 과정은 개인의 경험과 역량에 따라 그 정확성과 효율성의 편차가 존재하는 것을 파악하였다.

한편, 최근 미국이나 일본, 네덜란드, 독일 등의 다양한 주요 국가도서관을 중심으로 문헌 분류 혹은 주제명 부여 업무의 자동화를 위해 인공지능과 기계학습 기술을 적용한 다양한 과업들이 추진 중이며, 이러한 시류 선상에서 국립중앙도서관은 2020년 「인공지능 기술을 활용한 사서 업무 지원 도구 개발에 관한 연구」를 진행하였다. 이 연구에서 온라인 서점 데이터를 활용해 특정 도서의 주제어를 자동으로 부여하는 방식의 주제어 추천 시제품을 고안하였으나, 문학류에 한정한 학습 데이터로 인하여 성능 결과의 신뢰성 결여와 의구성은 여전히 존재하였다.

이에 본 연구는 국립중앙도서관의 서지데이터가 지닌 다양한 자질(목차, 서지, 주제명, 원문 등)을 기반으로 하여 이전 연구의 결과를 되짚어 보고, 최근의 기계학습기법을 반영한 주제명 자동 분류 알고리즘을 설계 및 검증하고, 실제 국립중앙도서관에서의 적용 가능성을 제시하며, 구축 최적화 개선 방안 등의 개발에 관련한 사항을 제언하고자 한다.

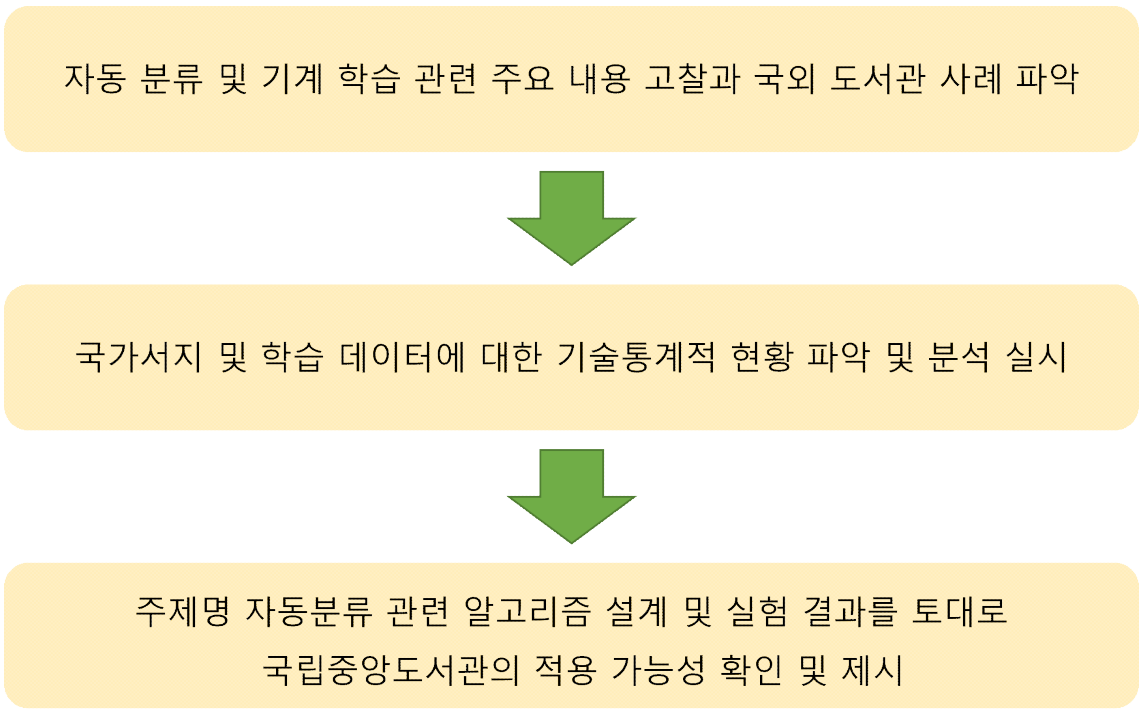
국가서지는 한 국가에서 출판된 출판물들의 존재 기록이며, 출판물을 명확히

식별 가능케 해주는 리스트로, 국가가 보유한 “지식의 총체”로서의 정치적 문화적 가치를 지는 것으로 그 중요성을 표현할 수 있다(이미화, 이지원, 2021).

본 연구는 일차적으로 주제명의 자동 분류 또는 자동 부여라는 맥락에서 실무자들의 효율적인 업무 수행을 지원하고, 나아가 앞서 설명한 국가서지를 활용하여 국민의 효율적인 정보 접근의 기틀을 마련해, 궁극적으로 국민의 정보 학습 및 향유 욕구를 충족시키고, 유연한 지식 확충을 도모하는 방안으로서 그 가치가 있음을 판단한다.

2. 연구의 범위

본 연구는 최근 추세의 기계학습을 적용하여 주제명의 자동 분류를 가능케 하는 알고리즘을 설계 및 검증하여 국립중앙도서관에 적용 가능성을 제시하고 구축 최적화 개선안 등의 개발을 지원하고자 한다. 이에 따른 연구의 범위는 <그림 1>과 같다.



<그림 1> 전체 연구의 범위

<그림 1>의 연구 범위를 수행하기 위하여, 연구 방법을 문헌 연구 및 사례조사, 데이터 분석, 그리고 자동 분류 실험으로 구성하였고, 각 연구의 세부 내용을 <표 1>과 같이 수행하였다.

<표 1> 연구 구성 및 방법

연구 방법	수행 내용
문헌 연구 및 사례조사	선행 연구조사를 통한 현재 최상의 자동 분류 성능을 나타내는 대표적인 자동 분류 모형 및 동향 조사 분석 자동 분류 및 기계학습 적용 연구와 해외 및 국내 도서관에서의 인공지능 및 자동 분류 관련 우수 사례 연구 분석
데이터 분석	국립중앙도서관 서지데이터 약 121만 건과 서지데이터에 부여된 주제명 표목 약 51만 건에 해당하는 데이터에 기반한 기술 통계 및 현황 분석
실험 및 검증	입수 데이터와 자동 분류 모형의 다양한 변인을 대상으로 하여 최적의 성능을 도출하기 위한 다양한 실험 시행 성능 최적화 모형 선별 이후 국립중앙도서관에 실제 적용 가능성을 파악하고 향후 자동 분류 모형 성능 개선을 위한 방안 제언

3. 연구의 기대효과

본 연구는 국가서지를 활용하여 기계학습 및 딥러닝을 접목한 자동 분류를 수행한 실험적 연구로서, 국가 지식 관리기관으로서의 지식 재창출 및 국민의 지식 접근의 효율성 증대 등을 숙고하는 정책 수립 의지를 나타내는 것으로 가치가 있다. 더불어 본 연구는 다음과 같은 부수적인 효과에 이바지할 수 있을 것으로 판단한다.

- 국가서지 기반의 학습데이터를 활용함으로써, 국립중앙도서관의 지식 재구성 및 기반 데이터 구축의 자동화 도모를 지원
- 국가서지 데이터의 현황 파악을 통하여 현재의 국가서지 데이터의 현황을 살펴보고, 미래에 구축할 서지데이터의 효과와 효율성 제고를 위한 업무 방향 개선 및 증진을 도모
- 국립중앙도서관의 현업에 있어서 효율적인 업무 프로세스 도입 가능성을 확인하여 발전 가능성을 파악함으로써, 정보화 사회에서 국립중앙도서관이 시대를 이끄는 차세대 지식정보 관리기관으로서의 정책적 기반 마련을 지원

Ⅱ. 도서관 자동 분류 해외 사례

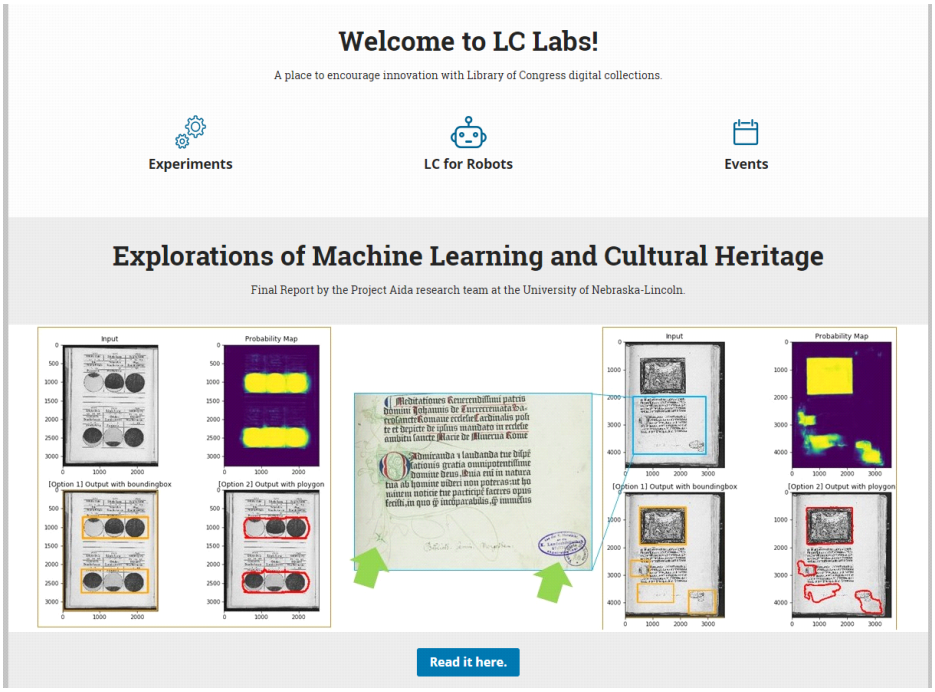
국립중앙도서관 주제명표목표의 자동 분류 기법을 고안하기 위하여, 국외 도서관에서의 자동 분류의 활용 사례를 파악하고 고찰하였다. 이는 주제명표목의 자동 분류 사례뿐만 아니라, 도서관에서 활용 중이거나 혹은 활용하고자 하는 연구 등에 등장한 자동 분류 사례를 함께 조사하여 본 연구의 선행자료로 제시하였다.

1. 미국의회도서관

미국의회도서관은 HITL(Humans in the Loop)이라는 이미지 검색을 위한 자동 메타데이터 작성 기법을 도입하고 LC library LABs에서 이용자에게 서비스를 제공 중이다(<그림 2> 참조).

미국의회도서관은 디지털 혁신 전략의 세부 목표 중 하나로, 도서관이 소장하고 있는 고문헌이나 역사적 사실을 보유하고 있는 사진이나 예술작품의 시각적 특성을 추출하는 데 인공 지능적 요소를 도입하였으며, 자동으로 메타데이터를 생성하도록 하였다.[1)]

1) https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1418&context=libraryscience



<그림 2> LC labs

2. 핀란드국립도서관

핀란드국립도서관은 도서관에서 구축한 서지 데이터베이스를 기반으로 다양하고 방대한 학습데이터로서의 활용성을 인지하여 자동색인용 오픈소스인 Annif 프레임을 구축하고 온톨로지 서비스인 Finto AI를 운영하고 있다.

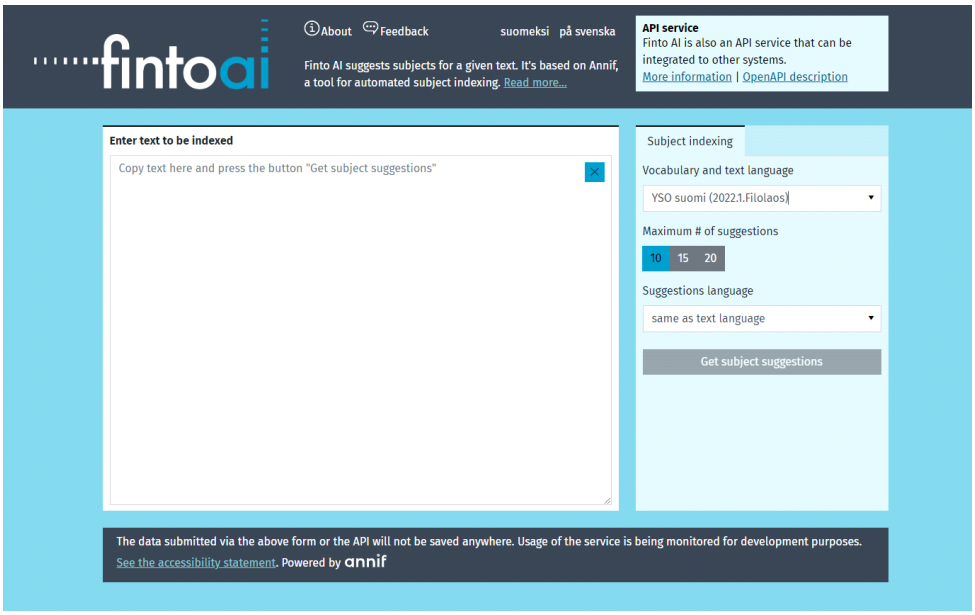
Annif[2](<그림 3> 참조)는 Tensorflow기반 신경망 모델과 Maui, fastText Omikuji 등을 활용한 다양한 텍스트 분류 알고리즘을 결합하여 구축한 자동색인용 오픈소스 도구이다. Finto AI[3](<그림 4> 참조)는 Annif를 기반으로 하여 2020년에 도서관에 설치한 자동색인 제공 서비스로, 웹 환경에서 원문에 해당하는 주제어를 추출하고 핀란드국립도서관의 온톨로지인 YSO와 연결하여 주제어에 대한 재검색 및 검토를 진행할 수 있게 한다.

Annif 알고리즘(<그림 5> 참조)은 어휘적 접근방식과 연관적 접근방식으로 나눌 수 있다. 어휘적 방식으로는 MLLM, STWFSA, Maui 등이 있고, 연관적 방식으로 TF-IDF, fastText, Omikuji 등이 있다. 이들을 통해 학습데이터의 양에 따라 알고리즘의 수행 성능을 조절한다.

2) https://annif.org/
3) https://ai.finto.fi/?locale=en

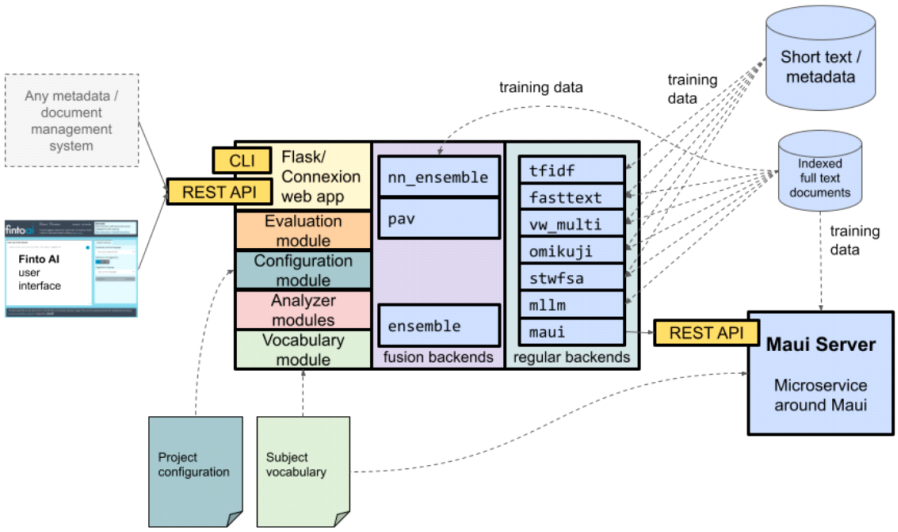
더불어 이 프레임은 오픈소스로 github[4]에 공개하고 있으므로 관련 연구자가 연구의 수행 과정 및 관련 알고리즘 등을 확인할 수 있으며, REST API로 다른 시스템에서도 Annif 서비스에 접근이 가능하다.

<그림 3> Annif 개괄



<그림 4> Finto AI 서비스 제공화면

4) https://github.com/NatLibFi/Annif



<그림 5> Annif 모듈러 (Suominen, Inkinen & Lehtinen, 2002)

3. 독일국립도서관

독일국립도서관(DNB)은 최근 자동 주제분류(automatic cataloguing) 프로젝트로 자연어 처리와 컴퓨터 언어학 및 기계학습 분야를 적용하여 도서관에서 다루는 다양한 매체의 주제분류를 수행하고자 하였다.

이 프로젝트의 기계학습 기법에서 사용하는 데이터는 대부분 비정형 데이터로서지 입수 시에 함께 입수하는 원문과 목차이며, 자동 주제분류에 있어 필요한 경우 메타데이터를 분석하여 데이터를 수집하기도 한다. 이와 더불어 주제분류에서 사용하는 DDC와 세부 설명, LCSH 및 독일국립도서관전거파일(Die Gemeinsame Normdatei, GND) 등을 활용하여 개체와 전거파일 간의 의미론적 연결을 결착시켰다.

특히, 방대한 양의 온라인 미디어 수집과 목록화에 많은 시간과 과정이 투입되기에, 이를 획기적으로 보완할 방법으로 자동으로 주제를 분류하고 목록을 작성할 수 있도록 기반을 마련하고자 하였다.

한편 2021년 4월, 독일은 인공지능 기반의 국가전략 중 하나로 핀란드국립도서관의 자동색인 시스템인 Annif를 벤치마킹하여 독자적 자동색인 시스템의 구축 연구를 수행 중이다. 이의 구현을 위해 기계 입력에 수월하도록 DNB 주제분

류와 DDC Short Numbers라는 도서관 자체적인 분류 규칙을 개발하여 적용 중이다.

DNB 주제 분류(Subject Categories)[5]는 2004년 오스트리아와 스위스 등 독일어권의 국가서지를 구성하기 위한 체계이다. 각 주제를 DDC 자체 분류번호, 주제, 전체 DDC 분류번호 구역의 총 3개 섹션으로 관리하고 있다(<그림 6> 참조).

DDC Subject Categories and included DDC classes for the New Release Service and the Series A, B, C, H and O of the Deutsche Nationalbibliografie (based on DDC 23)

Class	Discipline	Included DDC Classes
000	***Generalities, computers, information***	
000	Generalities	000-003
004	Computer science	004-006
010	Bibliography	010
020	Library and information sciences	020
030	Encyclopedic works	030
050	Magazines, journals and serials	050
060	Organizations and museology	060
070	News media, journalism, publishing	070
080	General collections	080
090	Manuscripts and rare books	090
100	***Philosophy and psychology***	
100	Philosophy	100-120, 140, 160-190
130	Parapsychology, occultism	130
150	Psychology	150

<그림 6> DDC 주제 분류(일부)

한정적 주제영역에 대해서는 DDC 대신 별도의 문자 기호를 활용하여 주제를 분류하고 메타데이터를 생성하게 한다. 예를 들어 교과서는 S, 소설은 B로 분류한다.

DDC Short Number는 서지 목록의 자동화에 적용되는 방식으로 특정 주제에 한정한 온라인 출판물이나, 의학 및 건강에 해당하는 DDC 강목 610, 상업적 정기간행물 B군, 대학 발간물 H군 등에 자체 축약된 DDC 번호를 부여한다.

<그림7>에서 Pediatrics Adipositas(소아비만)의 DDC 전체번호는 618.92398이나, DDC Short Number의 경우 Pediatrics(소아)까지 분류하는 618.92로 부여한다. 마찬가지로 Terrorism Muslims Middle East(중앙아시아의 이슬람교도 테러)의 경우, 303.6250882970956에서 303.6 Conflict and conflict resolution(갈등 및 갈등 해결)으로 부여한다.

5) https://www.dnb.de/SharedDocs/Downloads/EN/Professionell/DDC/ddcSachgruppenDNBAb2013.pdf?__blob=publicationFile&v=4

DDC Subject Category	610
Full DDC Number	618.92398
DDC Short Number	618.92

618.92398 Pediatrics Adipositas

DDC Subject Category	300
Full DDC Number	303.6250882970956
DDC Short Number	303.6

303.625 Terrorism Muslims Middle East
303.6 Conflict and conflict resolution

<그림 7> DDC short Number 예시[6)]

이와 같은 방식의 DDC Short Number는 <그림 8>과 같이 목록작성 시에 주제명표목과 함께 자동으로 부여된다.

Link to this record	https://d-nb.info/1254019936
title	Psychotherapeutic specialist literature of the GDR and FRG: A comparative citation analysis
Persons)	Storch, Monika (Other) Schneider, Nico (Other) Kirschner, Hariet (Other) Arp, Agnès (Other) Rauschenbach, Manuel (Other) Gallistl, Adrian (Other) Strauß, Bernhard (Other)
scope/format	online resource
Persistent identifiers	URN: urn:nbn:de:101:1-2022032412004609738422 DOI: 10.1055/a-1718-4071
Chronological order	Release date: 03.02.2022
DDC notation	616.89 (machine-determined DDC short notation)
Languages)	German (ger)
Relationships	Contained in: Psychotherapy, Psychosomatic Medicine, Medical Psychology (02/03/2022)
Tags	psychotherapy* ; Germany (GDR)* ; Specialist literature* ; Psychotherapist* ; Germany (Federal Republic)* ; Reception* (*determined by machine)
subject group(s)	610 medicine, health
online access	Open archive object

<그림 8> DDC 자동 부여 예시[7)]

앞서 언급한 바와 같이 Annif 알고리즘을 기반하고 DDC 자동 분류에서는 omikuji를 활용, 주제어 자동 분류에서는 omikuji-bonsai와 mllm의 앙상블을 활용하고 있다. <그림 9>는 각 자동 분류 적용 분야의 어휘 데이터 현황이다.

6) https://swib.org/swib21/slides/03-02-uhlmann.pdf 21p
7) https://swib.org/swib21/slides/03-02-uhlmann.pdf 9p

	GND descriptors	DDC subject categories	DDC short numbers (e.g., medicine)	
number of labels	1,303,547	100	121	
language	de	de	de	en
annif backend	ensemble (omikuji-bonsai + mllm)	omikuji	omikuji	omikuji
number of training docs	2,415,549 + 37,001	473,000	66,700	10,900
model size on disk	13 GB + 984 MB	21 GB	3.3 GB	716 MB

<그림 9> 독일국립도서관의 자동 분류시스템의 어휘 데이터 현황

독일국립도서관은 이 자동 목록 시스템을 활용하여 전자 자원, 학술지 논문, 대학 논문 등에 주제어를 부여하고 있다.

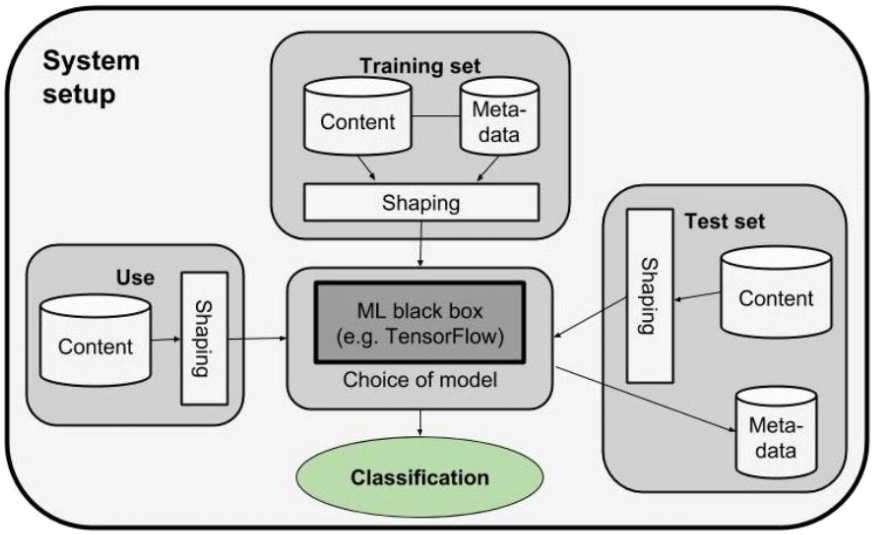
4. 노르웨이국립도서관

노르웨이국립도서관은 목록작성의 효율화를 도모하기 위하여 이미 작성된 메타데이터 및 콘텐츠와 오픈소스 소프트웨어를 기반으로 기계학습을 적용한 DDC 자동 분류 연구를 수행 중이다(Brygfjeld, Wetjen & Walsøe, 2017).

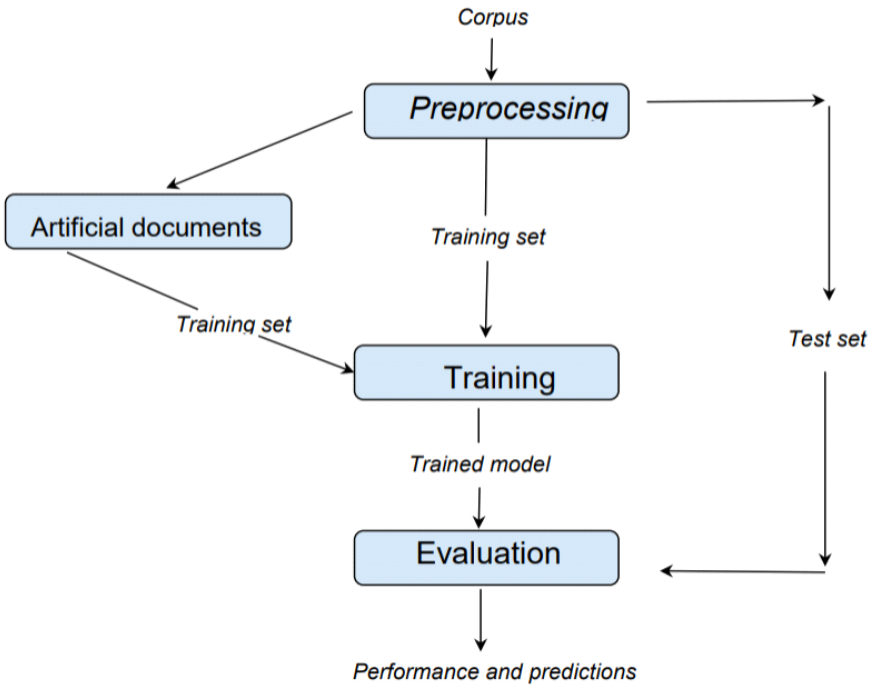
노르웨이굴립도서관의 Norart(Norske og nordiske tidsskriftartikler)는 노르웨이 및 북유럽 지역에서 출판한 논문 기사 등 600,000건을 수록하고 있는 메타데이터를 소장하는 데이터베이스이며, 특히 과학 기술 분야에 특화되어 있다. 이 중 36,000건 이상이 디지털 형태의 논문이나 기사에 대한 데이터와 해당 문헌으로 연결하는 링크를 함께 제공하고 있다. 메타데이터는 DDC 및 NORMARC을 기반으로 하고 있다.

Brygfjeld, Wetjen과 Walsøe(2017)의 연구는 DDC 자동 분류를 위한 기계학습용 콘텐츠로 노르웨이나 북유럽 출판의 과학·기술 분야의 전자 학술지를 대상으로 하며, 원문의 형태는 주로 PDF나 HTML이다. 이 연구의 실험 수행 시에 실제 학습모델에 사용할 데이터셋의 보완을 위하여 인공적으로 문헌을 생성해 데

이터셋을 구현하였으며, 모델 성능 향상을 보완하고 도모하도록 하였다(<그림 10>, <그림 11> 참조). 그 결과, 데이터셋 크기에 따라 성능의 변화가 유의미하며, 95% 이상의 높은 정확률을 보여주었다.



<그림 10> 학습 및 테스트 기본 설정 (Brygfjeld, Wetjen & Walsøe, 2017)



<그림 11> 학습모델 수행 과정(Brygfjeld, Wetjen & Walsøe, 2017)

5. 스웨덴국립도서관

스웨덴국립도서관의 KBLab(https://www.kb.se/)에서는 도서관에서 AI를 적용하고 관련 서비스를 제안하기 위해, 자연어를 처리하기 위해 기계학습의 딥러닝 기법인 Transformer 모형이 적용된 BERT를 스웨덴어로 된 자관의 입수자료에 적절히 활용하여 사용할 수 있는 방법을 강구하고 있다.

Type of named entity	AF-AI	M-BERT	KB-BERT
Person	0.913	0.945	0.961
Organization	0.780	0.834	0.884
Location	0.913	0.942	0.958
Time	0.655	0.888	0.906
Measure	0.828	0.853	0.890
Work of art	0.596	0.631	0.720
Event	0.716	0.792	0.834
Object	0.710	0.761	0.770
Total average	0.876	0.906	0.927

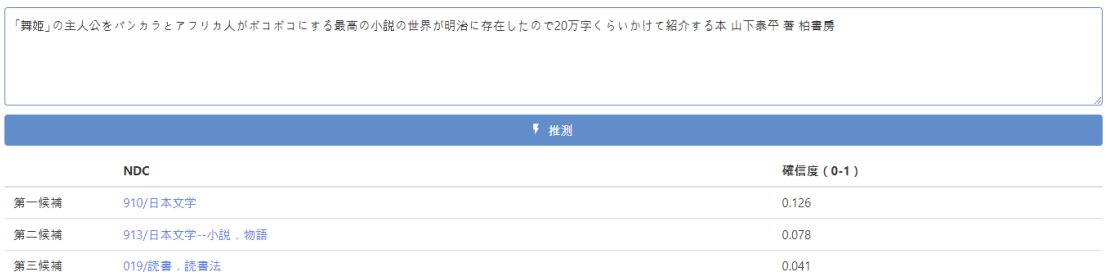
<그림 12> KB-BERT의 성능(Accuracy)

KB-BERT는 이러한 목적에서 개발된 자연어 처리 모델로 스웨덴국립중앙도서관의 서지데이터를 학습데이터로 사용하였다. 이 모델의 적용으로 예상되는 도서관의 변화는 첫째, '텍스트 자동 분류'인데, 특히 입수하는 디지털자료에 대하여 자동 분류시스템을 구축하는 데 도움을 줄 것으로 기대하였다. 둘째, '검색 접근점 향상'으로 서지데이터 작성 시 본 모델을 사용한다면 이용자의 정보자료 접근을 보다 효율적이고 효과적으로 할 수 있도록 사서가 도움을 줄 것으로 기대하였다. 셋째는 'OCR 변환의 성능 향상'으로 현재 작업 중인 디지털화된 자료

들을 온라인이나 컴퓨터상에서 재구성할 때, 각 자료의 내용에 대한 식별성을 높일 수 있을 것으로 기대하였다. <그림 12>는 KB-BERT의 성능을 나타낸 것으로, 기존의 BERT와 타 자연어 처리 모델 중 가장 높은 성능을 가져오는 것을 보여준다.

6. 일본국회도서관

일본국회도서관은 서지데이터에 기계학습을 적용해 일본십진분류를 예측하고 추천하는 서비스를 실험 중이다.



<그림 13> NDC 분류 예시

<그림 14> NDC predictor 화면

이 서비스는 일본국회도서관의 2017년 7월 기준 서지데이터 4,256,238건 중 서명과 출판사 및 책임표시사항에 해당하는 저자와 그 외의 정보를 학습 데이터로 하여 Facebook AI research가 개발한 fastText 알고리즘을 활용해 추천 3순위까지의 분류 결과를 보여준다.

예를 들어 다음 문장 "「舞姫」の主人公をバンカラとアフリカ人がボコボコにする最高の小説の世界が明治に存在したので20万字くらいかけて紹介する本[8] 山下泰平 著 柏書房"을 대상으로 서명 저자명 출판사 등과 같은 책임표시사항을 자동으로 식별하고 분류하여 910 일본문학, 913 일본문학－소설(小説、しょうせつ), 모노가타리(物語, ものがたり)[9], 019 독서, 독서법을 추천한다(<그림 13>, <그림 14> 참조).

더불어 분류 기호는 일본도서관 협회의 NDC Linked Data에서 추출하여 사용하며, 자동 분류 학습을 위한 서지데이터 4,256,238건의 서명 등의 형태소분석에는 일본어 기반 단어와 신조어 기반의 형태소 분석기인 mecab-ipadic-neologd[10]와 검색을 위한 형태소 분석기인 kuromoji[11]를 적용하였다. 한편 일본국회도서관은 앞에서 언급한 서지데이터와 형태소 분석기 및 알고리즘을 적용하여 제작한 학습모델[12]은 github에 공개하였으며, 그 내용을 확인할 수 있다.

8) 번역: '무희'의 주인공이 반카라와 아프리카인에게 몰매를 맞는 최고의 소설 세계가 메이지 시대에 존재했음을 20만 자 정도로 소개하는 책
여기서 '반카라'는 메이지 시대에 서양풍 양식, 옷차림을 하이칼라라고 불렀는데, 이에 반발하여 남루한 옷차림을 하거나 낡은 교복을 입은 사람 또는 언행이 불량한 사람을 지칭하는 말이다.

9) 일본에 있는 헤이안 시대에서 가마쿠라시대에 걸친 산문(散文)의 문학의 한 갈래(작품)로, 근대 문학의 소설(novel)에 대응하는 개념이다. 배경은 다르나 한국의 홍길동전이나 춘향전의 전(傳)에 해당하는 갈래이다.
(https://namu.wiki/w/%EB%AA%A8%EB%85%B8%EA%B0%80%ED%83%80%EB%A6%AC)

10) https://github.com/neologd/mecab-ipadic-neologd

11) https://github.com/atilika/kuromoji

12) https://github.com/ndl-lab/ndc_predictor

Ⅲ. 자동 분류의 개요

단일 레이블(single label)은 일반적인 분류 모델에서 가장 많이 사용하는 식별 방식으로, 다수의 범주에서 하나의 범주를 선택하게 된다. 범주의 수에 따라 2개의 범주 중 한 개를 고르게 되면 이진 분류(binary classification)에 해당하며, 3개 이상의 범주 중 한 개를 고르게 되면 다중 클래스 분류(multi-class classification)에 해당한다.

단일 레이블은 분류할 다수의 범주에 대해 하나의 식별 표시(label) 또는 태그(tag)를 가지게 되는데, 달리 설명하면 하나의 범주에 이 레이블이 선정되면 다른 범주는 자동으로 이 레이블에서 탈락하게 된다. 이러한 점을 통해 단일 레이블의 데이터를 표현하기 위해서는 one hot encoding 방식을 이용하는데, 단일 레이블 형식의 범주의 특징은 상호배타적(mutually exclusive)이면서, 전체를 포괄한다(collectively exhaustive)는 점이다.

단일 레이블과 달리 다중 레이블(multi-label)은 다수의 범주에서 다수의 범주를 선택하게 된다. 다중 레이블 데이터의 범주 특징은 상호 비 배타적(mutually non-exclusive)인 경우가 존재한다는 점인데, 선택되는 범주의 수가 제한이 없기에, 상황에 따라 0개에서 전체 범주 개수가 될 수 있다. 예를 들어 신문 기사를 읽고 그 기사에 대해 여러 가지 태그를 부여하는 방식으로 분류를 하는 것은 다중 레이블 분류(multi-label classification)에 해당한다.

1. 문서 전처리

1.1 토큰화(Tokenization)

토큰화란 연속적인 텍스트를 단어, 구문, 기호, 또는 다른 의미 있는 요소로 쪼개는 과정을 의미한다. 이 과정에서 가장 중요한 목표는 문장에 사용된 단어의 의미를 파악하는 것이다. 텍스트 마이닝(mining)과 텍스트 분류와 같은 애플리케이션에서는 모두 텍스트에 대해 토큰화 처리를 수행하는 파서(parser)가 필요하다.

다음 영어 예를 보면, "I am going to school."의 경우 띄어쓰기를 이용하여 ['I', 'am', 'going', 'to', 'school']과 같이 토큰화를 할 수 있다.

반면, 한국어 예를 보면 "나는 학교에 간다."의 경우 마찬가지로 띄어쓰기를 이용한다면 ['나는', '학교에', '간다']와 같이 어절 단위 토큰화를 할 수도 있다. 그러나 한국어는 교착어의 특성을 가지기에 하나의 단어가 여러 개의 형태소로 이루어지는 경우가 있어서, ['나', '는', '학교', '에', '가', 'ㄴ다']와 같이 형태소 단위 토큰화를 할 수도 있다.

1.2 Bag of Words(BoW)

BoW 모델은 단어 빈도와 같은 특정 기준을 기반으로 텍스트의 선택된 부분에서 텍스트를 축소 및 단순화하는 기법이다.

BoW은 컴퓨터 비전, NLP, 베이지안 스팸 필터(Bayesian spam filter)와 같은 여러 도메인과 기계학습에 의한 텍스트 분류 및 정보 검색에 사용되는데(Sivic & Zisserman, 2008), BoW에서 문서나 문장과 같은 텍스트의 본문을 단어 꾸러미(bag)처럼 간주하고, 단어 목록은 BoW 프로세스에서 생성된다.

BoW는 텍스트의 순서(문장의 구문)와는 상관없이 단어(1-gram)를 사용하여 표현한 방법으로, 이는 얻기가 매우 쉽고 텍스트 크기를 통해 벡터로 표현할 수 있다.

이 단어 모음은 문장과 문법을 구성하는 문장이 아니며, 이들 단어 사이의 의미적 관계는 단어 모음과 구성에서 무시된다. 단어는 종종 문장의 내용을 나타내기도 한다. 문법과 단어의 출현 순서는 무시되지만, 중복도를 계산하여 문서의 초점을 결정하는 데 사용할 수 있다.

1.3 N-gram

N-gram 기법은 텍스트에서 순서대로 출현하는 n-단어의 하위 문자열 집합을 추출한다. 이는 실제 텍스트를 대표하는 것은 아니지만, 텍스트를 나타내는 기능으로 사용할 수 있다.

한편 N-gram에서 N은 숫자에 해당하는데, 1-gram을 사용하면 텍스트를 표현하기 위한 BoW의 자질(feature)과 같게 되는데, 일반적으로 2-gram이나 3-gram을 사용한다. 이러한 방식으로 추출된 텍스트의 특징은 1-gram에 비해 더 많은 정보를 감지할 수 있다는 것이다.

1.4 Word2Vec

Word2Vec 알고리즘은 지역적 문맥 정보를 사용하여 단어 벡터를 생성하는데, 이 단어 벡터는 말뭉치의 모든 단어에 대한 단어 벡터로 제공되는 고정 길이의 실숫값 벡터를 나타낸다(Mikolov et al., 2013). 여기서 Word2Vec은 두 가지 모델을 사용하는데 첫 번째 모델인 CBOW는 문맥이 이해되었다는 가정하에 현재 단어를 예측하며, Skip-gram은 현재 단어를 안다는 가정하에 문맥을 예측한다.

1.5 Global Vectors for Word Representation(GloVe)

GloVe(Pennington, Soche & Manning, 2014)는 텍스트 분류에 사용되는 강력한 단어 임베딩(word embedding) 기술이다. 접근방식은 각 단어를 고차원 벡터로 표시하고 거대한 단어 말뭉치에서 주변 단어를 기반으로 훈련하는 Word2Vec 방법과 매우 유사하다.

흔히 사용되는 사전 훈련된 단어 임베딩은 Wikipedia 2014 및 Gigaword 5를 통해 400,000개의 어휘와 50개 차원의 단어 표현을 기반으로 하며, Twitter 콘텐츠를 포함해 더 큰 말뭉치로 훈련한 100, 200, 300개 차원의 사전 훈련된 단어 벡터화를 제공한다.

2. 텍스트 기반 확률 모형

2.1 확률론적 그래픽 모델

확률론적 그래픽 모델(Probabilistic Graphical Model, PGM)은 모델의 효율성 때문에 전통적인 방법으로서 널리 적용된다. PGM은 그래프 속성 간의 조건부 관계를 나타내기 위해 확률 이론과 그래프 이론을 이용하는데(Zhang & Zhang, 2010), 매우 간단하고 가장 자주 사용되는 모델은 베이즈 정리(Bayes' theorem)를 기반으로 하는 Naive Bayes(NB)이다(Maron, 1961).

NB와 같은 모델은 구조 및 계산 절차의 단순성이 장점인데, 이 단순성은 각 속성이 다른 속성에 영향을 미치지 않는다고 여겨지는 독립성의 전제에서 비롯된다(Xu, 2017).

NB 기법은 훈련 세트에서 볼 수 있는 특성이 주어진 사전 확률을 기반으로 클래스의 사후 확률을 계산하는 것을 기반으로 하는데, NB의 독립성 전제가 적

용되기 어려운 경우, 텍스트 데이터의 순차적 특성을 이용하기 위해 HMM (Hidden Markov Model)(van den Bosch, 2017) 및 CRF(Conditional Random Field)(Sutton & McCallum, 2012)와 같은 다른 PGM 모델이 제안된다.

2.2 kNN(k-Nearest Neighbors)

kNN 기법은 k개의 가장 가까운 샘플 중에서 가장 많은 샘플이 부여된 범주를 찾아 레이블이 지정되지 않은 샘플을 분류하는 방법을 기반으로 한다(Cover & Hart, 1967).

마찬가지로 kNN 알고리즘을 활용한 텍스트 분류는 분류하고자 하는 텍스트와 가장 유사한 k개의 텍스트를 찾아 이 텍스트에 가장 빈번한 범주로 레이블을 지정하는 형태로 분류 문제에 접근한다.

단순히 데이터 포인트 간의 거리를 계산해야 하므로 이 비모수적 기술은 상당히 빠를 수 있으나, 성능은 사용된 거리 함수에 따라 크게 상이해진다. kNN 기반 접근방식이 제대로 수행되지 않을 가능성이 존재할 경우, 거대한 데이터셋을 처리하기 위해 다양한 함수 또는 근사치가 필요할 수 있다.

kNN 기법은 주로 특징 유사성(Ali, Neagu & Trundle, 2019), k 값(Baoli, Qin & Shiwen, 2004), 인덱스 최적화(Cortes & Vapnik, 1995)를 통해 개선되고 있지만, 일반적인 kNN 기법은 모델에서의 시·공간 복잡성과 데이터양 사이의 높은 연관성 때문에 대규모 데이터셋에서는 비정상적으로 긴 시간이 요구된다.

2.3 SVM(Support Vector Machine)

SVM은 패턴 인식의 이진 분류를 해결하기 위해 Cortes와 Vapnik(1995)가 제안한 분류기로, 영상과 텍스트 등의 많은 데이터 분야에서 전통적으로 지도학습 기반의 분류를 수행할 때, 높은 성능을 보인 강력한 예측 시스템이다. SVM은 1차원 입력 공간 또는 특징 공간에서 이상적인 초평면(hyperplane)을 생성하여 초평면과 두 범주의 훈련 세트 사이의 거리를 최대화하여 최상의 일반화 능력을 얻는데, 초평면에 수직인 방향을 따라 범주 경계 사이의 거리를 최대화하여 가장 낮은 분류 오류율을 얻게 된다.

그리고 SVM은 다차원, 비선형 분류에서 훈련 범주를 더 잘 구별하기 위해 입력을 고차원 공간으로 변환하는데, 이 고차원 공간으로 변환하는 함수를 커널(kernel) 함수라고 하기에 이 방법을 커널 트릭이라고 일컫는다. 대표적인 SVM 커널 함수로는 선형(linear), 다항(polynomial), 방사(radial basis), 시그모이드

(sigmoid) 함수가 있다.

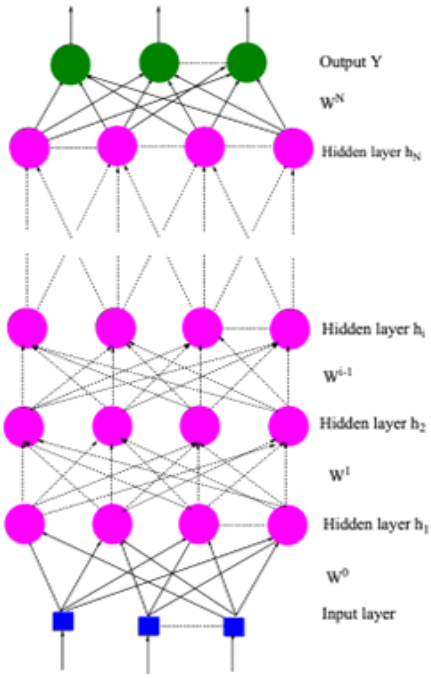
2.4 Decision Tree and Random Forest

의사 결정 트리(Decision Tree, DT)(Safavian & Landgrebe, 1991)는 데이터 공간을 계층적으로 분해하는 직관적인 트리 구조 학습 알고리즘으로, 분할 정복 원칙에 기반한 재귀 지도 트리 구조 학습 접근방식이다. DT는 트리 생성과 트리 가지치기의 두 단계로 분리될 수 있는데, 의사 결정 트리의 child 노드는 데이터 집합의 하위 집합을 나타내고 각 leaf 노드는 범주를 나타낸다. DT를 구축하는 목적은 클래스와 특성 간의 관계를 발견한 다음 알려져 있지 않은 신규 레코드의 범주를 예측하는 데 사용하는 것이나, DT 방법은 대부분 빠르게 확장되는 데이터 크기를 처리하는 데 있어 비효율적이라는 단점이 있다. 이러한 단점을 극복하기 위해 무작위 데이터 선택을 사용하여 훈련된 다수의 DT로 구성된 랜덤 포레스트(Random Forest)(Ho, 1995)는 DT보다 효과적인 결과를 보였고, 최근까지도 자주 사용되고 있다(Islam et al., 2019).

3. 딥러닝 기반 기계학습

3.1 MLP(Multi-Layer Perceptron)

다층 퍼셉트론(Multi-Layer Perceptron, MLP)은 자질을 자동으로 추출하기 위해 사용되는 가장 간단한 신경망 구조 중 하나이다(Khalil Alsmadi et al., 2009). <그림 15>는 N개의 은닉 계층으로 구성된 MLP 모델을 보여주는데, 이 모델은 입력 계층(input layer), 모든 노드에 활성화 함수(activation function)가 연결된 은닉 계층(hidden layer) 및 출력 계층(output layer)을 포함하고, 각 노드는 특정 가중치 W^k로 연결된다.



<그림 15> 다중 퍼셉트론 구조
(Ramchoun et al., 2017)

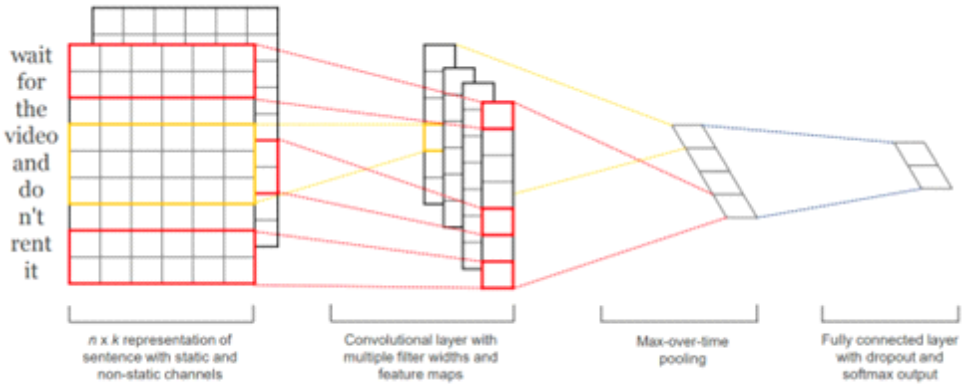
텍스트 분류를 위해서는 각 입력 텍스트를 BoW기반으로 처리하고 MLP의 입력 하는데, 기존 여러 모델에 비하여 상대적으로 다량의 텍스트 분류 작업에서 높은 성능을 보이는 것으로 나타난다.

3.2 CNN(Convolutional Neural Networks)

CNN(LeCun & Bengio, 1995)은 시각적 정보를 추출할 수 있는 콘볼루션 필터를 사용한 이미지 분류를 위해 제안된 모델이다.

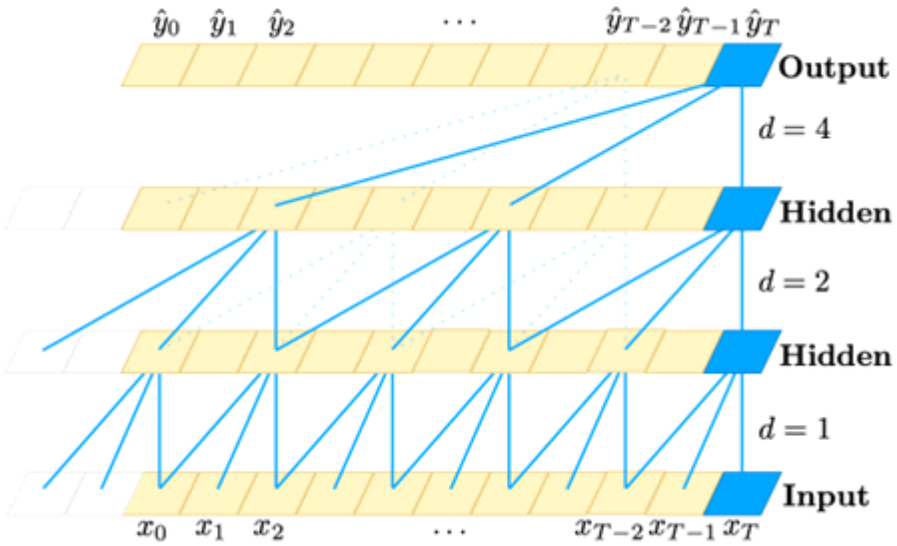
CNN은 별개의 커널에 의해 지정된 콘볼루션을 시퀀스의 수많은 청크(chunk)에 동시에 적용할 수 있어서, 텍스트 분류를 포함한 다양한 NLP 응용 프로그램에 활용되고 있다.

<그림 16>은 텍스트를 이미지 형태와 유사하게 2D 행렬로 표현한 TextCNN 방법을 표현한 것으로, 행렬은 이후 다양한 크기의 필터가 있는 콘볼루션 계층으로 전달된다. 이후 합성곱 계층 결과는 풀링 계층을 통해 전송되고 풀링 결과와 연결되어 텍스트의 최종 벡터 표현을 생성한다. 이 결과 최종 벡터에 저장된 값을 통해 범주를 예측한다.



<그림 16> TextCNN 구조(Yoon, 2014)

이후로 텍스트 분류를 위한 Sequence modelling에서 더 효율적으로 활용될 수 있는 TCN(temporal convolutional network)이 <그림 17>과 같이 제안되었다. TCN은 간단한 구조를 유지하면서 동시에 여러 작업에서 성공적이었던 CNN 구조를 결합하여 기존의 recurrent based 모델보다 높은 성능을 보여준다.

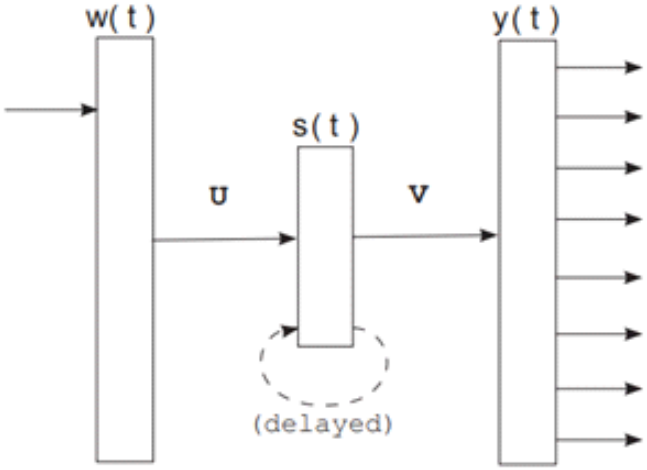


<그림 17> TCN 구조(Bai, Kolter & Koltun, 2018)

3.3 RNN(Recurrent Neural Networks)

RNN은 순환 계산 과정을 통해 장거리 종속성을 찾는 방법으로 널리 이용되고 있다. RNN 언어 모델은 텍스트 분류 작업에 적합한 모든 단어들의 위치 정보를 고려한 과거 정보를 학습한다.

<그림 18>은 텍스트 분류를 위한 RNN 모델로, 각 입력 단어는 단어 임베딩 기술을 사용하여 특정 벡터로 표현된다. 임베딩된 단어 벡터는 한 번에 하나씩 RNN 셀의 입력으로 들어가고, 출력은 입력 벡터와 동일한 차원을 가지며 다음 은닉층(hidden layer)으로 전송된다.



<그림 18> RNN 구조(Mikolov et al., 2011)

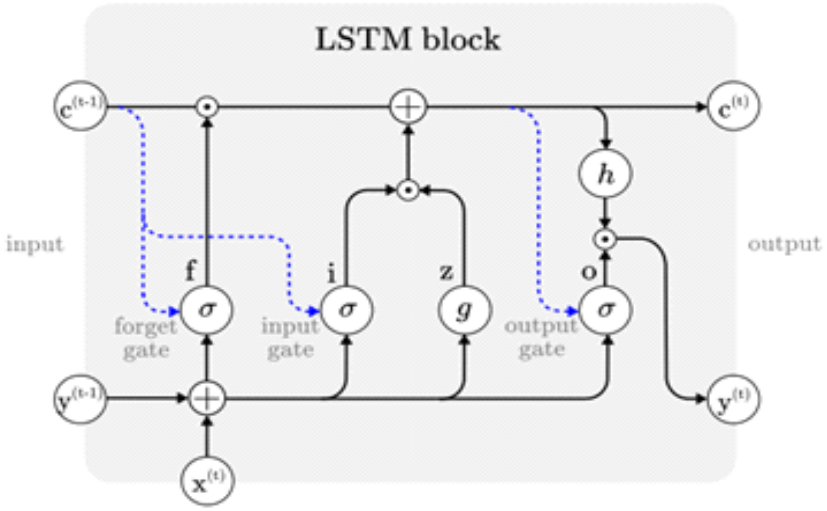
RNN은 모델 전체에서 매개변수를 공유하고 각 입력 단어에 대해 동일한 가중치를 사용하며, 은닉층의 마지막 출력은 입력 텍스트의 레이블을 예측하는 데이터로 사용된다.

3.4 LSTM(Long Short-Term Memory)

RNN 모델은 학습 시 역전파(backpropagation) 과정에서 구하는 기울기(gradient) 값이 도함수의 계속된 곱연산으로 인해 작아져 소멸하는 기울기 소실(gradient vanishing) 문제가 발생할 수 있다.

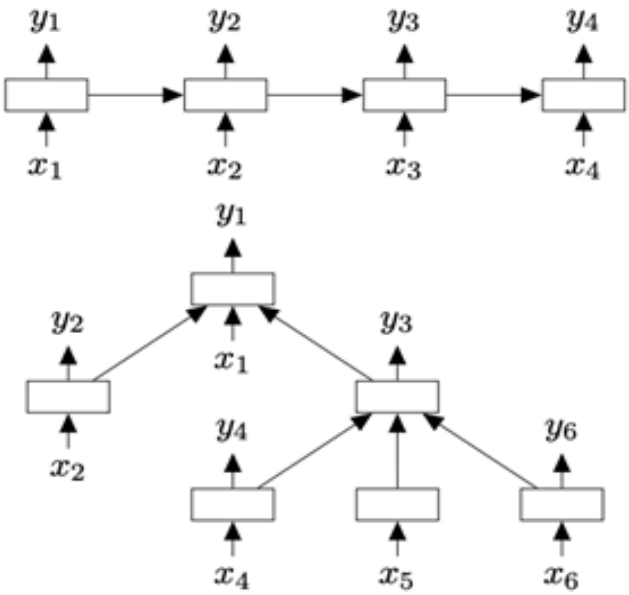
LSTM은 이러한 RNN의 문제를 개선한 모델로 기울기 소실 문제를 효율적으로 완화한다(Van Houdt, Mosquera & Nápoles, 2020). <그림 19>와 같이 LSTM block은 망각 게이트(forget gate), 입력 게이트(input gate), 출력 게이트(output gate) 구조가 정보 흐름을 제어하고, 임의의 시간 간격에 대한 값을 기억하는 셀이 추가되어 있다.

LSTM 분류 방법은 망각 게이트 구조를 사용하여 쓸모없는 정보는 필터링하고, 중요한 정보는 장기 기억하므로 분류기의 전체적인 능력을 높이는 데 도움이 된다.



<그림 19> LSTM 구조(Van et al., 2020)

한편, Tree-LSTM(Tai, Socher & Manning, 2015)은 <그림 20>과 같이 LSTM 모델 시퀀스를 트리 구조로 확장한 것으로, 게이트 및 메모리 셀 업데이트는 이전 단어의 상태가 아니라 노드의 child 상태에 따라 달라진다. 단일 망각 게이트 대신 Tree-LSTM 장치에는 각 child 정보를 선택적으로 통합하기 위해 각 child에 대해 하나의 망각 게이트가 있다.

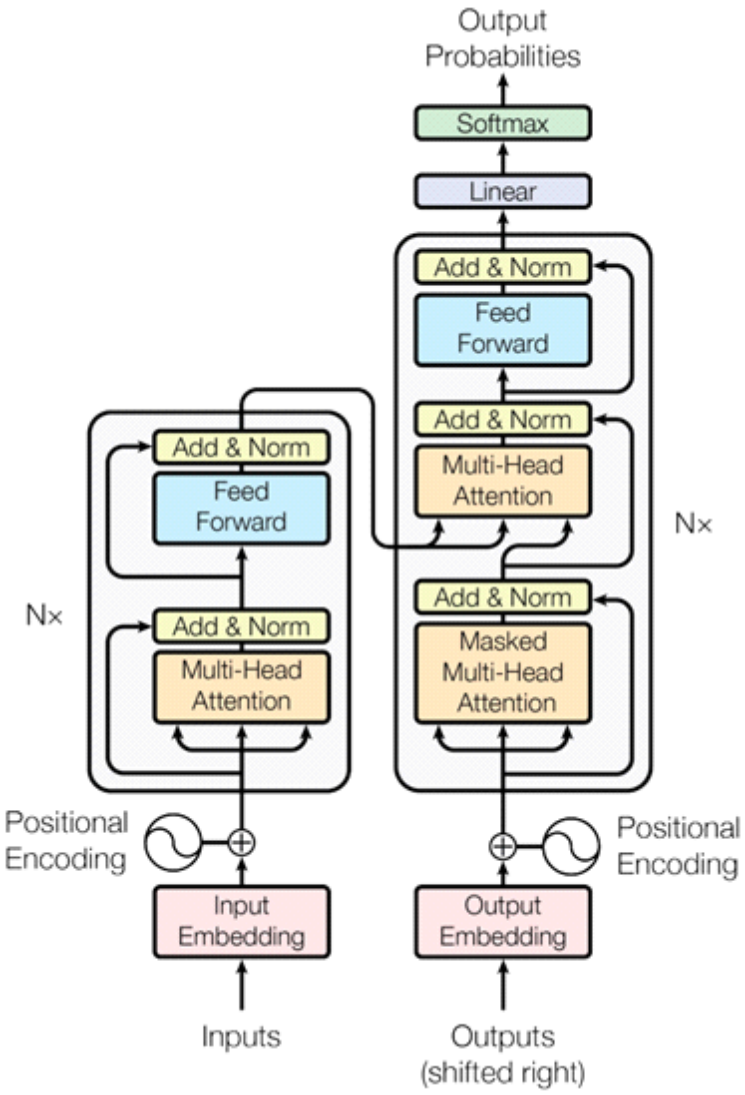


<그림 20> 기존 LSM 구조(위)와 Tree-LSTM 구조(아래)(Tai, Socher & Manning, 2015)

3.5 Transformer

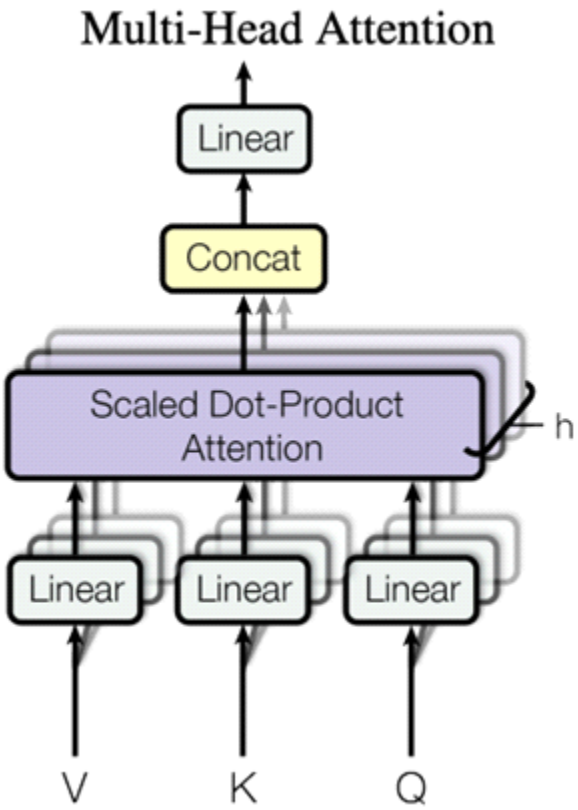
Vaswani 외(2017)는 모든 입력 토큰(예: 단어)을 순차적이 아닌 동시에 처리할 수 있는 고유한 인코더-디코더(encoder-decoder) 접근방식인 Transformer 아키텍처를 제안하였다(<그림 21> 참조).

Transformer는 순서에 대한 개념이 없는 bag-of-words 방식으로 입력 시퀀스를 제공하며, Transformer는 "self-attention" 기술을 사용하여 토큰 종속성을 학습한다. 그리고 인코더의 첫 번째 레이어에 대해 수행된 특정 인코딩 단계는 문장의 다양한 위치에 나타나는 동일한 단어에 대한 임베딩이 별개의 표현을 갖도록 보장한다. 이 단계를 "위치 인코딩"이라고 하며, 이 단계가 없으면 손실되었을 단어의 상대적 위치에 대한 정보를 삽입하는 것이 목적이다.



<그림 21> Transformer 모델 구조(Vaswani et al., 2017)

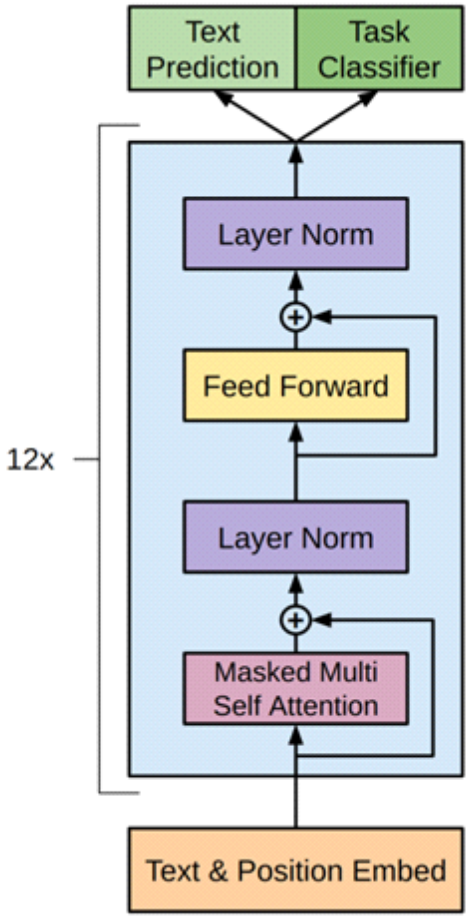
인코더가 단어 중 하나를 처리할 때마다 입력 구문의 다른 단어를 자연스럽게 볼 수 있도록 하는 'attention 레이어'는 이 시스템의 중요한 구성 요소이다. <그림 22>에서 볼 수 있듯이 이러한 유형의 레이어를 여러 개 쌓으면 다중 헤드(multi-head) attention 레이어가 생성된다(Vaswani et al., 2017). 이 상태에서 헤드 출력을 연결하고 선형 레이어를 통해 결과를 실행하면 개별 출력이 단일 행렬로 압축된다.



<그림 22> 다중 헤드(multi-head) attention 레이어

3.6 GPT(Generative Pre-trained Transformer)

Transformer 아키텍처는 언어 모델링에 매우 적합하지만, 연구자들은 일부 작업 수행 시 인코딩 및 디코딩 단계에서 중복 정보를 학습하는 경우가 존재한다고 주장하였다. 따라서 인코더나 디코더로 아키텍처를 제한하면 동등한 성능에 더 가벼운 모델이 될 수 있다는 점에 착안하여 GPT(Generative Pre-trained Transformer)를 제안하였다(Radford et al., 2018)(<그림 23> 참조).



<그림 23> GPT 모델 구조
(Radford et al., 2018)

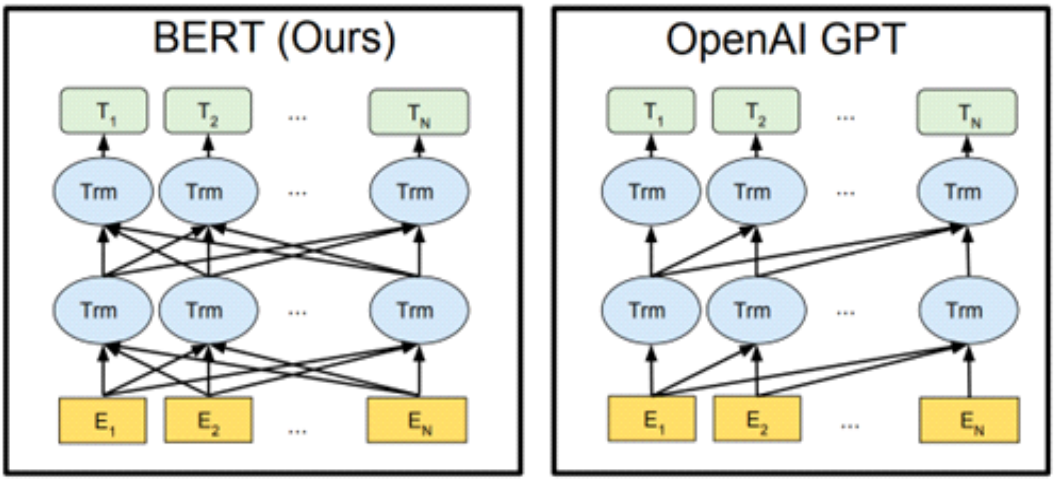
GPT는 여러 Transformer-decoder 레이어를 쌓는 디코더 전용 아키텍처를 사용하는데, 이전 블록이 주어진 대상 시퀀스의 조건부 확률 분포를 정의하기 때문에 Transformer의 디코더 블록은 자동회귀 또는 자기회귀(auto-regressive)가 된다.

GPT의 사전 훈련 작업은 다음 단어를 예측하기 위함이며, 자기회귀 모델이기 때문에 <그림 23>에서 보는 것과 같이 예측이 왼쪽 또는 오른쪽 문맥(GPT의 원래 아키텍처는 왼쪽에서 오른쪽 문맥임)에 의해서만 조절되는 단방향 언어 모델을 생성한다.

한편, 디코더 전용 아키텍처는 긴 텍스트 시퀀스를 처리할 때 기존 인코더-디코더 모델보다 더 나은 성능을 제공하는 추가 이점이 있어 생성 작업(예: 추상 텍스트 요약 및 생성 질문 응답)에 특히 적합하다. Radford 외(2018)는 분류를 포함하여 다양한 다운스트림 작업에서 사용할 수 있도록 수정 권한을 제공하였다.

3.7 BERT(Bidirectional Encoder Representations from Transformers)

GPT모델과 밀접한 연관이 있는 BERT는 양방향 컨디셔닝을 위해 자동회귀 기능을 희생한다(Devlin et al., 2019). 따라서 BERT의 아키텍처는 <그림 24>와 같이 GPT와 달리 다층 양방향 Transformer가 된다.



<그림 24> BERT(좌)와 GPT 구조(우) 비교
(Devlin et al., 2019)

BERT의 목표는 마스킹 된 언어 모델(masked language model, MLM) 절차를 따라 입력의 일부분이 마스킹 되고 모델의 네트워크는 다음 단어를 예측하는 대신 마스킹 된 토큰을 예측하려고 시도하는 것이다.

이 훈련 전략은 양방향 조건화를 달성하는 데 중요하며 모델이 마스크 단어와 왼쪽 및 오른쪽 문맥 사이의 관계를 학습할 수 있도록 하는 것으로, 두 번째 문장이 제공되고 두 번째 문장이 첫 번째 문장 다음에 오는지 아닌지를 이진 방식으로 추측해야 하는 보조 작업, 즉, 다음 문장 예측(next sentence prediction, NSP)에 대해 훈련된다. 이 작업을 통해 BERT는 모델이 문장 관계를 더 잘 학습할 수 있도록 도모한다.

또한 BERT를 다운스트림 작업에 적용하는 것은 간단하며, 인코더에서 얻은 표현을 단일 레이어 피드 포워드(single-layer feed-forward) 신경망을 통해 전달하는 모델을 단순히 미세 조정함으로써 분류에서의 뛰어난 결과를 얻을 수 있다.

4. 다중식별 분류 데이터셋

Reuters(2017)는 로이터 통신사에서 제공하는 금융 뉴스 서비스에서 파생된 텍스트 분류를 위해 널리 사용되는 데이터셋으로, 다중 클래스 및 다중 레이블이 있는 90개의 범주, 7,769개의 학습 데이터와 3,019개의 테스트 데이터로 구성된다. R8, BR52, RCV1 및 RCV1-v2를 포함하여 Reuters 데이터의 다양한 하위 집합이 존재한다.

RCV1(Lewis et al., 2004)은 1996년부터 1997년까지 103개의 범주로 수동으로 분류한 로이터 뉴스 기사를 편집한 데이터셋으로, 23,149개의 학습 데이터와 784,446개의 테스트 데이터로 구성된다. RCV1-2K는 RCV1과 동일한 특성이 있지만, RCV1-2K 레이블 세트에 몇 가지 새로운 레이블이 추가되어 2,456개의 레이블이 존재한다.

AAPD(Arxiv Academic Paper Dataset)(2018)는 웹사이트에서 컴퓨터 과학 분야 논문의 초록을 대상으로 다중 레이블 텍스트 분류를 위한 거대한 데이터셋으로, 여기에는 총 54개의 레이블이 있는 55,840건의 학술지 논문의 초록이 포함되어 있다. 이 데이터셋의 목적은 초록의 내용을 기반으로 그 초록이 수록된 논문의 주제를 예측하는 것이다.

WOS-11967(Kowsari et al., 2018)은 Web of Science에서 추출한 데이터로 각 인스턴스에 대해 두 개의 레이블이 있는 출판된 논문의 초록으로 구성되어 있다. 클래스 수는 적으나, 넓은 범위에서의 데이터를 확인할 수 있는 데이터셋이다.

5. 성능 측정 평가지표

5.1 단일평가지표

· 정확도(Accuracy)

가장 직관적인 성능 측정이며 단순히 전체 관찰에 대해 정확하게 예측된 관찰의 비율로, 정확도는 거짓 긍정(false positive)과 거짓 부정(false negative) 표본 수가 크게 다른 경우에는 의미 있는 값을 도출하기 어려운 단점이 있다.

· 정확률(Precision)

총 예측된 긍정적인 문서에 대하여 올바르게 예측된 긍정적인 문서의 비율로, 높은 정확률은 낮은 거짓 긍정 비율과 관련이 있다.

· 재현율(Recall)

실제 클래스의 모든 문서에 대하여 올바르게 예측된 긍정적인 문서의 비율로, 높은 재현율은 낮은 거짓 부정 비율과 관련이 있다.

· F1(F1-score)

정확율과 재현율의 조화 평균으로, 이 점수는 거짓 긍정과 거짓 부정을 모두 고려한다. F1 점수는 클래스 분포가 고르지 않은 경우에 정확도보다 더 의미 있는 값을 도출한다.

5.2 복합평가지표

· 거시적 평균(Macro-averaging)

클래스에 대한 단순 평균을 제공하는 것으로, 거시적 평균 점수는 빈도를 고려하지 않고 각 범주에 동일한 가중치를 할당하므로 범주별 평균을 의미한다.

· 미시적 평균(Micro-averaging)

클래스 전체에 걸쳐 문서별 평가 결과를 결합한 다음 분할 테이블에 효과적인 측정값을 출력하는 것으로, 미시적 평균 점수는 결과적으로 모든 문서에 동일한 가중치를 할당하기에 문서당 평균으로 간주한다.

· Hamming loss

전체 레이블 수에 대해 잘못된 레이블 수의 비율로, 다중 레이블 분류에서 Hamming loss는 정답과 예측 사이의 Hamming 거리로 계산된다.

Ⅳ. 학습 데이터 현황

1. 데이터 개요 및 범위

주제명을 자동 분류 또는 자동 부여하기 위해 수집된 학습 데이터의 전반적인 현황은 <표 2>와 같이 주제명이 부여된 서지데이터가 총 1,218,867건으로 이 중 목차를 포함한 데이터가 474,980건, 원문을 포함한 데이터가 67,038건이었으며, 이와 함께 주제명표목표의 용어 데이터는 511,980건을 수집하였다.

각 유형의 데이터별로 기술 통계적 현황 분석을 수행하였으며, 구체적인 주요 내용은 다음과 같다.

- 서지데이터 : 주제별 서지데이터 입수 현황, 주제명 데이터 사용 빈도
- 목차 데이터 : 목차 데이터 입수 현황, 목차 기입된 서지데이터의 주제별 빈도 파악
- 원문 데이터 : 원문 데이터 입수 현황, 수집 원문 데이터에 해당하는 서지데이터 파악
- 주제명표목표 용어 데이터 : 주제명 범주별 데이터 수록 현황, 활용 빈도, 주제명 데이터 간의 관계

<표 2> 데이터 입수 현황

데이터 종류	건수
오프라인 서지데이터	1,218,867
목차 데이터	474,140
원문 데이터	67,038
주제명표목표 용어 데이터	511,980

2. 주제명표목표

2.1 전체 용어 현황

국립중앙도서관은 2002년부터 「국립중앙도서관 주제명표목표」(이하 주제명표목표)를 개발하고 2013년 주제명표목표의 고품질화 연구를 통해 주제명표목표 업무지침을 제정하였다.

국립중앙도서관의 주제명표목표의 용어는 총 511,980건으로, 우선어가 257,103건으로 약 51%로 주제명표목표 중 절반 정도의 비중을 차지하고, 이어 외국어 169,466건으로 약 33%, 비우선어 61,442건으로 약 12%, 분류어 23,969건으로 약 4.7% 순으로 나타났다(<표 3> 참조).

<표 3> 주제명표목표 용어별 건수 현황

용어분류	건수 (비율)
우선어	257,103 (50.22)
비우선어	61,442 (12.00)
외국어	169,466 (33.10)
분류어	23,969 (4.68)
소계	511,980 (100.0)

이에 우선어를 대상으로 각 활용현황을 살펴보면, 1회 미만의 비활용 주제명은 199,783건으로 우선어 중 약 78%를 차지하였고, 한 번이라도 사용한 활용 주제명은 57,320건으로 나타났다.

활용 주제명 57,320건 중 100회 이상 사용된 경우는 3,506건에 불과하였는데 <그림 25>와 같이 활용 주제명 전체 중 6%에 해당하며, 특히 1회 이상 30회 이하로 사용한 주제명은 전체의 85% 정도로 나타나, 우선어 중 활용 주제명의 대부분은 30회 이하로 사용한 것을 확인할 수 있었다.

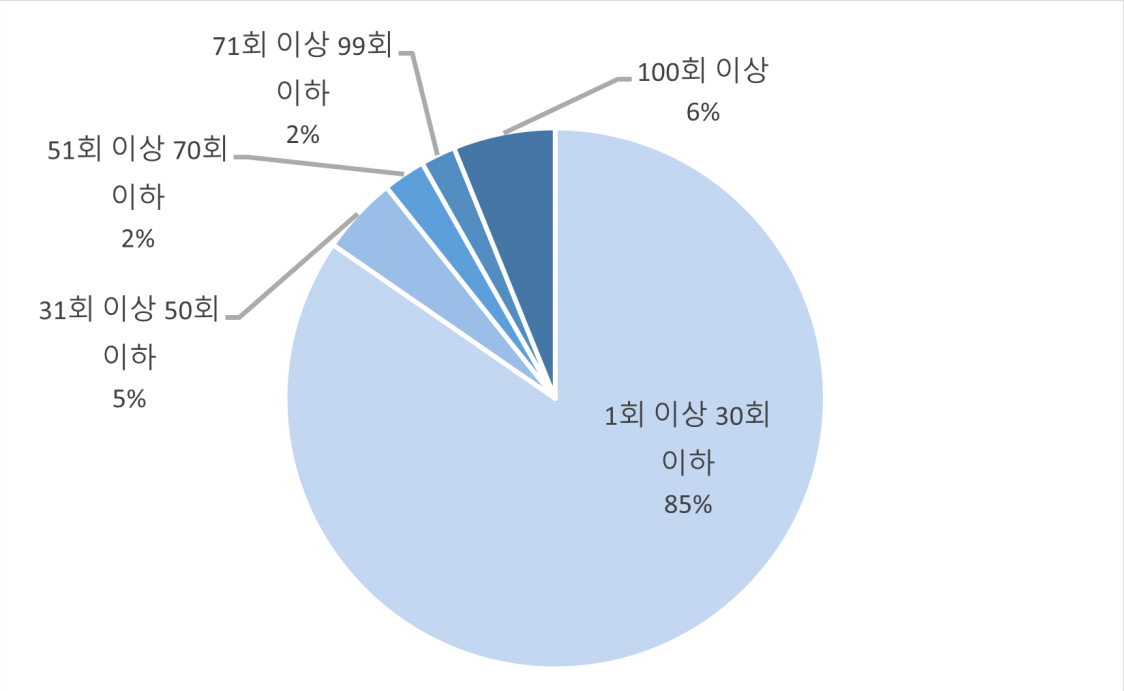
<표 4>와 <표 5>에 따르면 1회 이상 30회 이하 활용 주제명 중 10회 이하의 활용 주제명이 40,298건으로 이는 30회 이하 활용 주제명의 83%에 달한다. 결국

우선어 기준으로 활용 중인 주제명은 주로 10회 이하로 활용하는 것을 파악할 수 있었다.

또한 <표 6>은 우선어 기준 활용 주제명 중 최고빈도순으로 상위 30위를 정렬한 것으로, 최고 41,147회 활용한 주제명이 가장 많이 부여된 것으로 나타났으며, 이후 32,590회로 줄어들고 순위가 낮아질수록 급격히 활용 정도가 낮아지는 것을 알 수 있다.

<표 4> 주제명 활용 빈도 현황(우선어 기준)

활용 횟수	건수(비율, %)	소계(비율, %)
미 부여	199,783 (77.71)	199,783 (77.71)
1회 이상 30회 이하	48,506 (18.87)	57,320 (22.29)
31회 이상 50회 이하	2,657 (1.06)	
51회 이상 70회 이하	1,457 (0.57)	
71회 이상 99회 이하	1,194 (0.46)	
100회 이상	3,506 (1.36)	
합계	257,103 (100.0)	257,103 (100.0)



<그림 25> 활용 주제명 빈도 현황

<표 5> 1회 이상 30회 이하의 주제명 활용 빈도(우선어 기준)

부여 빈도	건수	비율 (%)	부여 빈도	건수	비율 (%)	부여 빈도	건수	비율 (%)
1	16,379	28.57	11	836	1.46	21	330	0.58
2	7,429	12.96	12	790	1.38	22	321	0.56
3	4,493	7.84	13	690	1.20	23	272	0.47
4	3,053	5.33	14	605	1.06	24	289	0.50
5	2,316	4.04	15	553	0.96	25	266	0.46
6	1,845	3.22	16	500	0.87	26	264	0.46
7	1,477	2.58	17	447	0.78	27	256	0.45
8	1,324	2.31	18	401	0.70	28	226	0.39
9	1,047	1.83	19	406	0.71	29	183	0.32
10	935	1.63	20	381	0.66	30	192	0.33
1-10	40,298	70.30	11-20	5,609	9.79	21-30	2,599	4.53

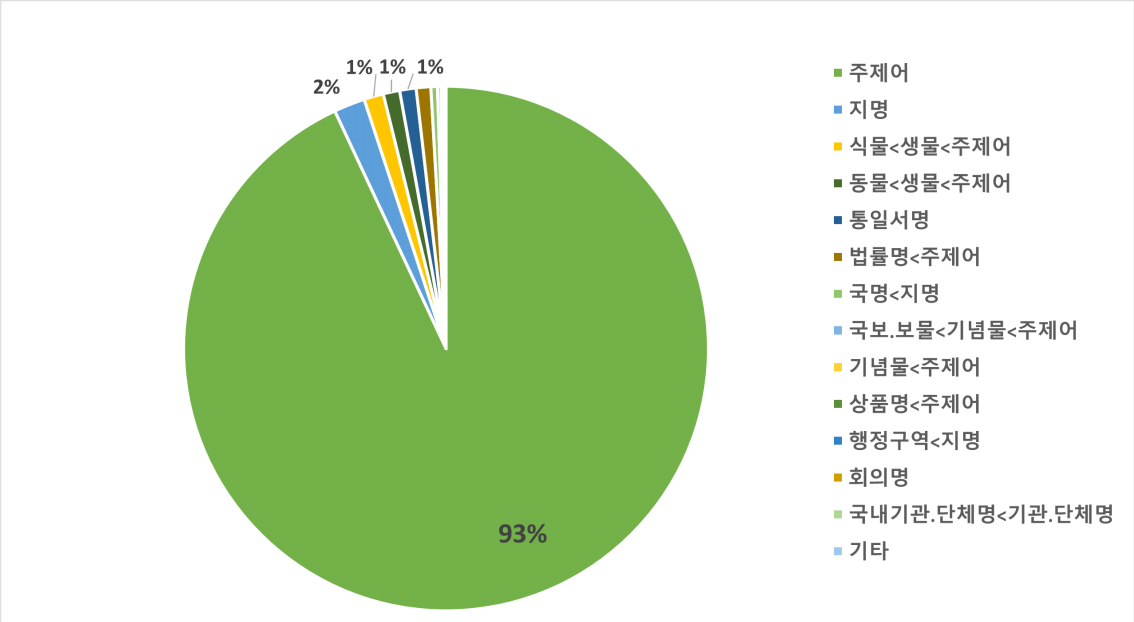
<표 6> 최고빈도순 상위 30위의 주제명 활용 빈도(우선어 기준)

부여 빈도	건수	부여 빈도	건수	부여 빈도	건수
41,177	1	8,009	1	5,442	1
32,590	1	8,007	1	5,324	1
28,681	1	7,954	1	5,170	1
24,537	1	7,264	1	5,114	1
15,877	1	6,721	1	5,011	1
12,421	1	6,613	1	4,893	1
10,693	1	6,032	1	4,892	1
9,454	1	5,931	1	4,828	1
8,823	1	5,908	1	4,696	1
8,629	1	5,871	1	4,657	1

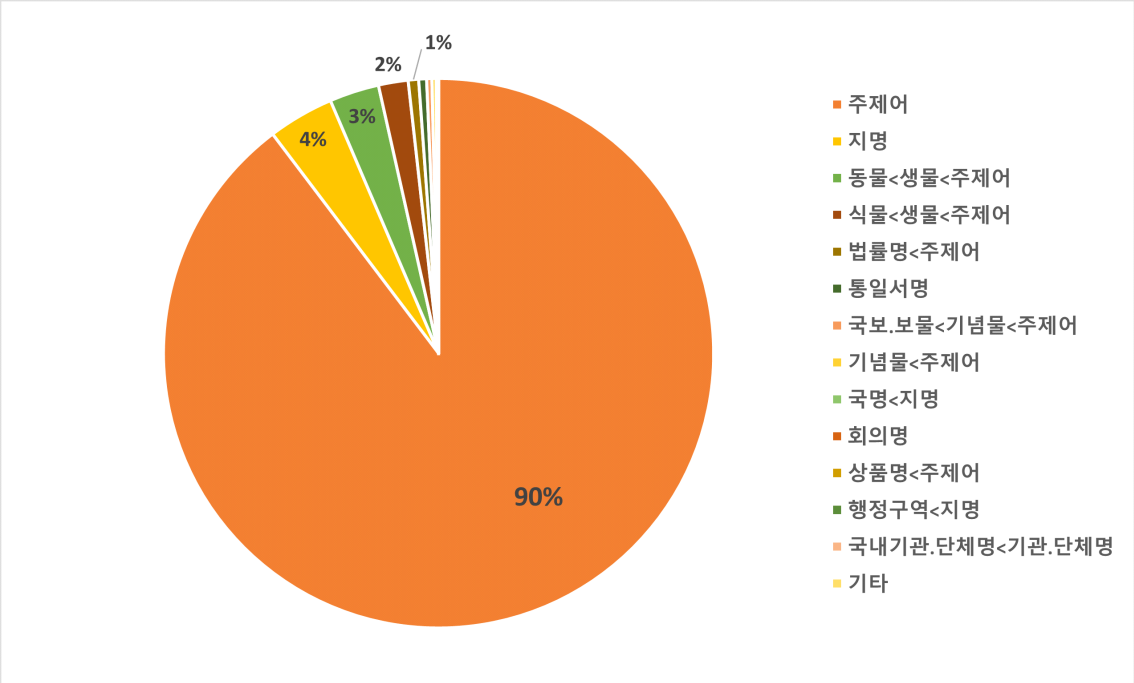
한편, 우선어 기준으로 사용된 주제어의 범주는 14개로, 가장 많이 사용되는 주제어 범주는 갈래 없는 일반 '주제어'로 약 93% 정도를 차지하였으며, 이어 '지명'이 약 2%에 이은 비율을 보여, 주제명을 부여할 때, 일반 주제어를 주로 사용하는 것으로 나타났다.

<표 7> 주제어 종류별 부여 현황(우선어 기준)

현황 횟수와 비율 / 주제어 종류	미 부여		부여		소계	
	횟수	비율(%)	횟수	비율(%)	횟수	비율(%)
주제어	179,204	89.70	53,319	93.03	232,523	90.44
지명	7,738	3.87	1,098	1.92	8,836	3.44
동물<생물<주제어	5,830	2.92	581	1.01	6,411	2.49
식물<생물<주제어	3,484	1.74	689	1.20	4,173	1.62
법률명<주제어	1,222	0.61	514	0.90	1,736	0.68
통일서명	921	0.46	580	1.01	1,501	0.58
국보.보물<기념물<주제어	617	0.31	128	0.22	745	0.29
기념물<주제어	492	0.25	95	0.17	587	0.23
국명<지명	125	0.06	231	0.40	356	0.14
상품명<주제어	57	0.03	40	0.07	97	0.04
회의명	60	0.03	11	0.02	71	0.03
행정구역<지명	38	0.02	22	0.04	60	0.02
국내기관.단체명<기관.단체명	2	0	2	0	4	0
기타	0	0	1	0	1	0
소계	199,790	100	57,311	100	257,101	100



<그림 26> 주제어 종류별 부여 현황(우선어 기준)



<그림 27> 주제어 종류별 미 부여 현황(우선어 기준)

<표 8> 부여 주제어 세부 빈도(우선어 기준)

현황 횟수 / 주제어 종류	1	2 ~ 15	16 ~ 30	31 ~ 50	51 ~ 100	101 ~ 500	501 ~ 1000	1001 ~ 10000	10001 ~	소계
주제어	14,982	25,652	4,465	2,493	2,491	2,677	326	226	7	232,523
지명	283	480	85	49	100	85	10	6	0	8,836
동물<생물<주제어	255	257	27	18	11	12	1	0	0	6,411
식물<생물<주제어	357	272	27	17	11	5	0	0	0	4,173
법률명<주제어	146	248	42	27	24	24	1	2	0	1,736
통일서명	172	281	44	27	22	29	3	2	0	1,501
국보.보물<기념물<주제어	93	27	5	1	1	1	0	0	0	745
기념물<주제어	36	50	5	2	0	1	1	0	0	587
국명<지명	30	97	25	19	19	32	4	5	0	356
상품명<주제어	6	12	8	4	3	5	1	1	0	97
회의명	7	4	0	0	0	0	0	0	0	71
행정구역<지명	6	8	1	0	0	1	1	5	0	60
국내기관.단체명<기관.단체명	1	1	0	0	0	0	0	0	0	4
기타	1	0	0	0	0	0	0	0	0	1
소계	16,375	27,389	4,734	2,657	2,682	2,872	348	247	7	257,101

2.2 관계 현황

주제명과 주제명 간의 관계를 표현하는 관계지시기호는 총 80개의 유형이 존재하였으며, 관계형의 사용횟수는 1,308,367건으로 파악되었다.

관계지시기호의 사용 빈도에 따른 상위 30개 유형을 나타내면, <표 9>와 같다. 30개 유형 사용횟수 합은 총 1,304,603건으로 관계지시기호의 전체 사용횟수의 99.7%에 해당하였으며, 이는 현재 주제명표목표에서 구축해둔 관계지시기호 80개 유형 중 실제로 사용한 관계지시기호는 30개 유형 이내임을 나타낸다.

최빈도 관계지시기호는 용어와 용어의 연관관계를 나타내는 RT였으며, 이어 상하위 관계의 BT, NT 순으로 사용 비율을 확인할 수 있다. 그리고 한국어와 영어의 관계를 서로 표현하는 KEN, ENG, 동등관계에서의 유의 관계를 표현하는 USE, UF 등이 이어서 사용 빈도가 높은 것으로 나타났다.

<표 9> 관계지시기호 상위 30개 활용 빈도

관계 지시기호	건수	비율	관계 지시기호	건수	비율	관계 지시기호	건수	비율
RT	406,666	31.08	GER	17,780	1.36	SK	1,645	0.13
BT	191,377	14.63	KJA	15,525	1.19	KDC6	1,302	0.10
NT	191,374	14.63	KFE	9,879	0.76	ORG	1,142	0.09
KEN	125,205	9.57	FRA	9,878	0.75	KES	1,069	0.08
ENG	125,158	9.57	SNN	8,619	0.66	ESP	1,067	0.08
USE	59,189	4.52	KSN	8,618	0.66	ITA	387	0.03
UF	59,186	4.52	KCH	3,009	0.23	KIT	387	0.03
UNS	21,825	1.67	CHI	2,981	0.23	KRU	384	0.03
JPN	18,039	1.38	KJP	2,786	0.21	RUS	379	0.03
KGE	17,781	1.36	NK	1,647	0.13	LT	319	0.02

<표 10> 100회 이상 활용 주제명의 관계지시기호 활용 빈도

관계형	주제명(비율)	관계 활용빈도 (비율)	관계형	주제명(비율)	관계 활용빈도 (비율)
RT	3,198 (23.3)	33,066 (48.14)	KDC6	134 (0.98)	152 (0.22)
NT	1,883 (13.72)	22,764 (33.14)	ORG	141 (1.03)	141 (0.21)
BT	2,578 (18.78)	3,527 (5.13)	NK	114 (0.83)	126 (0.18)
UF	1,492 (1.87)	3,136 (4.57)	ROM	51 (0.37)	99 (0.14)
ENG	2,102 (15.31)	3,093 (4.50)	ESP	58 (0.42)	62 (0.09)
GER	489 (3.56)	753 (1.10)	PT	16 (0.12)	27 (0.04)
JPN	583 (4.25)	719 (1.05)	SNN	14 (0.10)	14 (0.02)
FRA	386 (2.81)	453 (0.66)	LT	7 (0.05)	9 (0.01)
CHI	276 (2.01)	335 (0.49)	LAT	7 (0.05)	7 (0.01)
UNS	195 (1.42)	203 (0.30)	ITA	2 (0.01)	2 (0.00)

한편, <표 10>은 우선어 기준 100회 이상 사용한 주제명들의 관계지시기호 사용횟수를 파악한 것이다. 활용 주제명 3,506개에 사용한 관계지시기호는 총 68,688건이었으며, 사용한 관계지시기호의 유형은 20개인 것으로 나타났다. 또한 연관관계 RT를 가장 많이 사용하고(48.14%), 이어 NT(33.14%), BT(5.13)의 상하위 관계, 동의 관계 UF(4.57%), 외국어 ENG(4.5%) 순서로 관계지시기호를 사용하고 있었다.

주목할 점은 주제명으로서 주로 선정하는 주제명 간의 관계가 계층이나 연관관계 위주로 연결되어 있다는 것이다. 이외 활용하는 관계지시기호는 외국어 정도이며 이것도 영어(ENG) 정도만 활용한다는 점에서 결국 주제명표목표 관계지시기호 80개 중 실제 구축에 사용하는 관계지시기호는 매우 한정적이고 활용 주제명은 주로 광의적 용어를 사용하는 것으로 판단되며, 관계 형성이 단조로울 수 있다는 점을 시사한다.

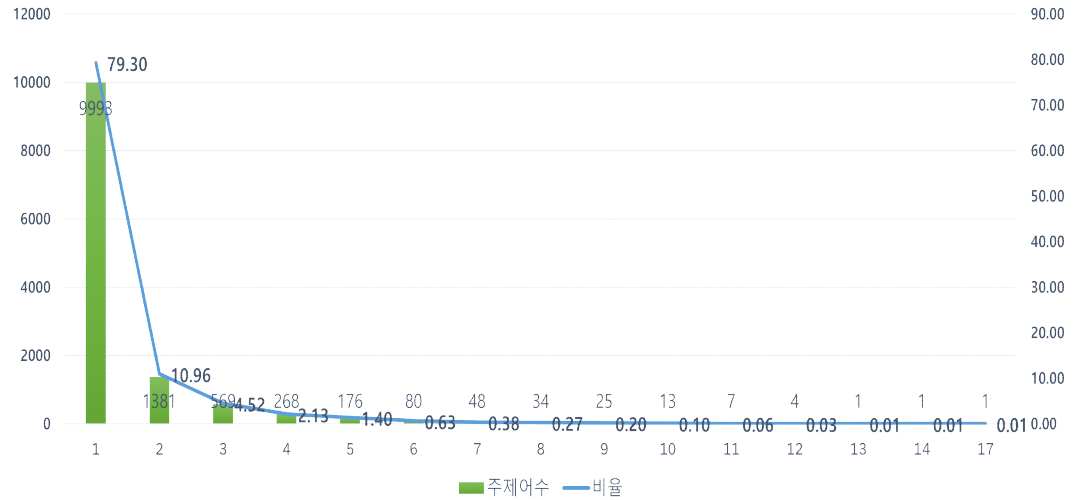
2.3 계층 현황

앞서 살펴본바, 주제명표목표는 여러 지시기호를 통하여 각각의 용어들과 관계를 이루어나가고 있다. 지시기호의 사용 빈도 중 최고빈도 RT는 계층의 동일선상에서 용어들을 연결하고 있다면, BT와 NT는 그 용어 간의 상하의 관계를 나타낸다. <표 11>은 현재 주제명표목표 자체의 계층구조에서 우선어 중에서 상하위 관계만을 확인하여 그 현황을 나타낸 것으로, 주제명 중 최상위어는 12,602건으로 나타났고, 그 하위로 143,704건의 개념어가 계층화 되어 존재하고 있다. 또한 상하 계층관계가 설정되어있지 않은 용어의 경우 73,319건으로, 68% 이상의 용어들이 상하위 관계를 지니는 것으로 확인하였다.

<표 11> 용어 갈래에 따른 건수

용어 갈래	건수(건)
최상위 개념어	12,602
최하위 개념어	143,704
계층관계 미존재 개념어	73,319
계층 연결 경로 총수	400,330

주제명의 하위 최대 심도는 17단계였으며, <그림 28>은 최상위 주제어 중 하위 심도에 대한 비율을 나타낸 것이다. 79.3%가 심도 1 즉, NT 관계를 하나만 지닌 주제명을 의미하는데 이 수가 최상위 주제어 중 대략 80%에 해당한다. 대부분의 주제어가 하위로 상세하게 혹은 세부적으로 구축되어 있지 않고 있음을 추측할 수 있다.



<그림 28> 최상위 주제어 보유 하위 심도 비율

주제어의 범주에 따른 주제어의 심도 평균을 살펴보면, <표 12>와 같다. 법률명 범주의 주제어가 0.49, 일반 주제어는 2.47로 대체로 상하관계 주제어 형성이 상대적으로 적거나 차상위까지 구성되어 있음이 나타났다. 반면 동물과 식물 등의 유강목이 정확한 주제 분야나 행정구역 등 세부지리적 요소가 명확히 드러날 때는 주제어 구성도 이와 맥을 같이 하므로 심도가 상대적으로 깊은 것으로 분석되었다.

<표 12> 주제어 범주에 따른 평균 심도

주제어 범주	평균 심도	주제어 범주	평균 심도
동물<생물<주제어	11.17	국명<지명	4.05
행정구역<지명	8.68	국내기관.단체명<기관.단체명	4.00
식물<생물<주제어	7.38	통일서명	3.90
지명	6.63	주제어	2.47
상품명<주제어	5.37	기타	1.00
기념물<주제어	4.96	회의명	0.97
국보.보물<기념물<주제어	4.36	법률명<주제어	0.49

한편, 주제명표목표 내에서 전체 우선어의 심도 현황과 부여 횟수에 따른 활용 주제명 우선어에 대한 심도 현황을 분석한 결과, <표 13>과 같다. 주제명표목표의 전체 우선어는 평균적으로 2.86 단계의 깊이에 따른 상하관계를 보유하고 있음을 확인할 수 있었다. 구체적으로 심도 현황을 살펴보면 최소 1회 부여된 모든 주제어의 경우 2.32, 100회 이상 고빈도로 부여된 주제어는 2.59, 여기에서 문학에 부여된 경우를 제외한 100회 이상 부여된 주제어는 2.52로 나타났다. 최소 한 번 이상 부여된 주제어의 경우 심도가 평균보다 얕은, 즉 매우 일반적이고 광의의 주제어들이 주로 활용되고 있음을 시사한다. 또한 100회 이상 부여되는 주제어는 최소 1회 이상 부여되는 주제어보다 심도가 상대적으로 깊은 듯하나, 전체 주제어 평균 심도보다는 얕은데 주로 서지에 주제명을 부여할 때 넓은 개념의 용어를 사용하는 것으로 판단된다.

<표 13> 주제명표목표의 전체 우선어와 활용 주제명의 심도 현황

	전체 우선어	1회 이상 부여	100회 이상 부여	100회 이상 부여 (문학제외)
개수	257,102	57,312	3,505	3,259
평균	2.86	2.32	2.59	2.52
편차	3.19	2.81	2.58	2.57
최대	17	17	15	15

<표 14> 100회 이상 부여 주제어의 심도 사례

주제어	최대 심도	해당 주제어 심도	사용횟수
수(숫자)[數]	10	0	272
예술[藝術]	11	0	631
언어학[言語學]	11	2	484
금융 상품[金融商品]	8	0	180
통계 분석[統計分析]	9	2	762
철학(사상)[哲學]	14	2	1,850

100회 이상 부여된 주제어의 심도 사례를 살펴보면, <표 14>와 같다. 실제 부여된 주제어의 상하관계에서 최대 심도는 9에서 14까지 나타나지만, 비중 있게 사용되는 주제어들은 최상위(심도 0)이거나 차상위(심도 1-2) 개념어에 해당하는 것을 볼 수 있다.

2.4 주제명 부여 현황

서지데이터의 하나의 레코드에는 한 개 이상의 주제명을 부여할 수 있다. 서지 레코드의 내용이 어떤 주제나 형식을 내포하느냐에 따라 그 주제나 형식을 표현하는 용어에 해당하는 하나 또는 그 이상의 주제명을 선정하게 된다.

국립중앙도서관의 서지데이터에서 개별 서지 레코드에 부여된 주제명 개수에 따른 서지데이터 비율을 나타내면, <표 15>와 같다. 하나의 서지 레코드에 하나의 주제명이 부여된 경우가 556,204건으로 전체 서지데이터의 45.63%에 해당하였다. 주제명이 2개 부여된 서지데이터가 39.29%, 3개 부여된 서지데이터는 약

12.77%로, 주제명이 3개 이내 부여된 서지데이터가 전체의 97.69%이었고, 이는 「국립중앙도서관 주제명표목 업무지침(2021)」에서 가급적 3개 이내로 주제명을 부여하도록 명시하고 있는바, 서지데이터에 부여된 주제명 개수는 적절하다고 판단하였다.

이 밖에 주제명 수가 10개 이상 부여된 서지데이터의 사례는 <표 16>과 같이 확인할 수 있었다.

<표 15> 주제명 부여 개수에 따른 서지데이터 비율

개수	1	2	3	4	5	6	7	8	9	10	11	12
서지	556,204	478,901	155,606	24,670	2,798	506	105	40	22	3	10	2
비율	45.63	39.29	12.77	2.02	0.23	0.04	0.01	0.00	0.00	0.00	0.00	0.00

<표 16> 주제명 10개 이상 부여 서지데이터 사례

서명	KDC	주제명 수
(2009 공인중개사 1차) 부동산학개론.n1-3	321.32077	10
수도권의 변화	539.7	11
우리 소리의 맥脈을 찾아서	679.311	12

또한 주제의 특정성에 따라 부여된 주제명의 현황을 확인하기 위한 간접적인 방법으로 한국십진분류법(KDC)의 분류 기호 길이에 해당하는 자릿수를 이용하였다. 일반적으로 KDC와 같은 십진분류법은 상위 주제를 승계하여 하위 주제로 세분하므로 분류 기호가 길어질수록 상위 개념에 종속되는 세부 주제를 표현하게 되어 있어 대상 자료의 주제가 특정성이 높아 복합적이고 구체적일수록 긴 자릿수를 갖게 된다. 이러한 특성을 이용하여 주제의 특정성에 따른 주제명의 부여 현황을 그 개수와 심도로 분석하면 각각 <표 17>, <표 18>과 같다. KDC 분류 기호의 자릿수가 길수록 주제명 입력개수도 증가하는 경향이 나타났고, 심도 역시 깊어서 특정 영역의 주제를 분류한 서지일수록 주제명 부여에서도 연관이 있음을 보여주었다.

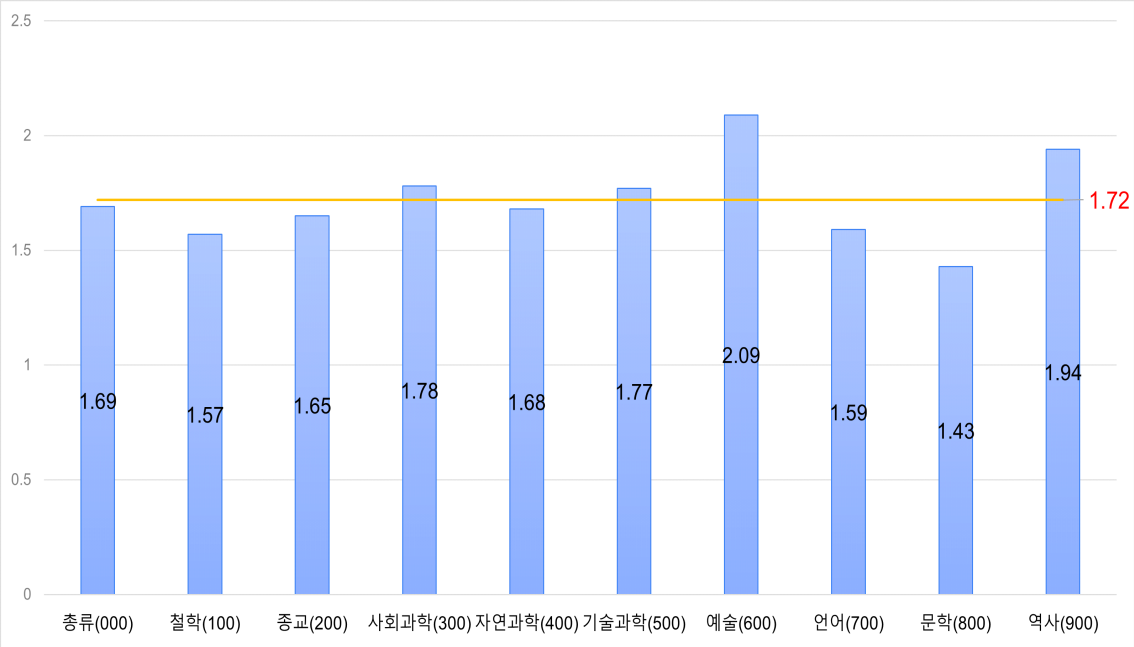
<표 17> 주제 특정성에 따른 주제명 부여 현황

	전체 서지	KDC 7자리 이상	KDC 10자리 이상
개수	1,218,867	313,388	13,733
평균	1.72	1.94	1.97
편차	0.78	0.85	0.84
최대	12	11	7

<표 18> 주제 특정성에 따른 주제명 부여 심도 현황

	전체 서지 부여 주제명	KDC 7자리 이상	KDC 10자리 이상
개수	2,097,682	609,191	27,021
평균	2.77	2.79	3.18
편차	2.56	2.47	2.90
최대	17	17	15

주류별 주제어 부여 개수 현황을 살펴보면, <그림 29>와 같다. 전체 서지데이터 기준 평균적으로 주제어가 1.72개 부여되는 것을 고려할 때 학문 분야 철학과 문학 등은 주제어 부여 개수가 상대적으로 적은 편이었다. 반면 예술이 2.09개로 가장 높았고, 역사가 차상위로 주제어를 많이 부여하고 있었는데, 역사 분야의 경우 일반 주제어보다 상대적으로 지명 중심으로 주제어 부여가 많은 것으로 확인되었다.



<그림 29> 학문 분야별 주제명 부여 현황

3. 서지데이터

3.1. 주류(main class) 분석

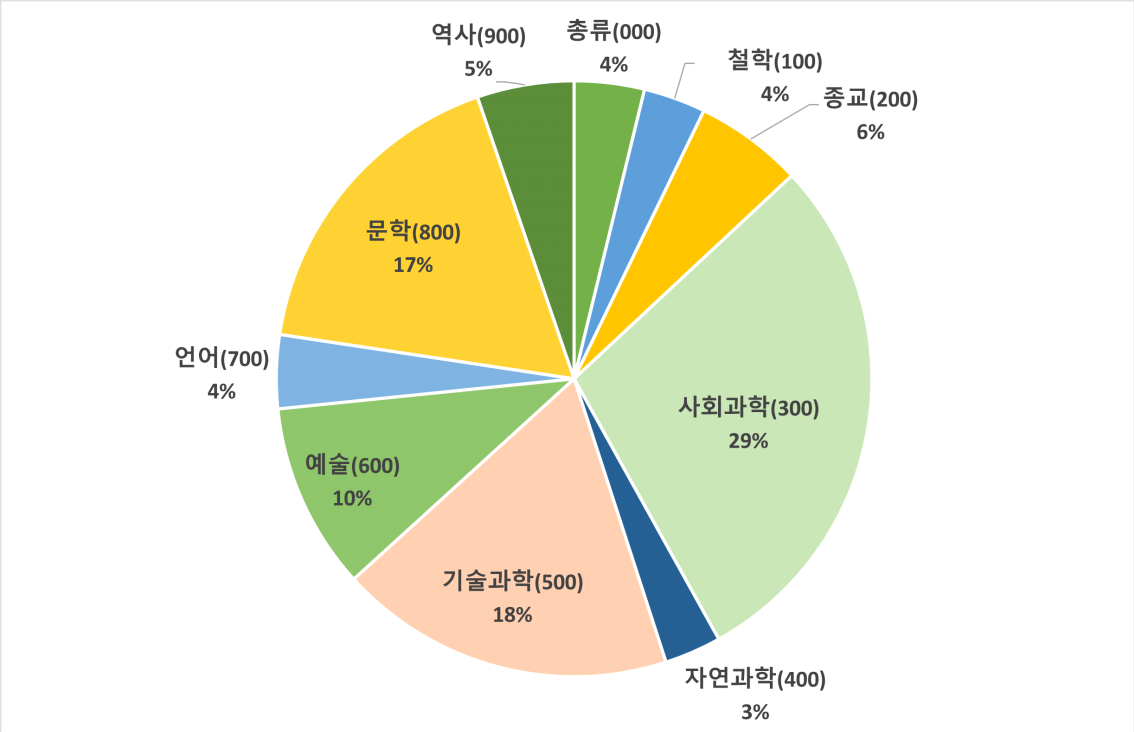
본 연구는 입수한 서지데이터에서 학습데이터로 사용이 가능한 데이터를 구분하기 위하여 주제명이 기입된 서지데이터를 추출하였고, 여기에서 한국십진분류기호를 추출하여 기계학습 대상 서지데이터에 대해 주제별 분석을 수행하였다.

주제명이 부여된 오프라인 서지데이터는 총 1,218,867건이며, 이중 분류 기호가 부여된 데이터는 1,199,643건(약 99%), 미부여된 데이터는 19,224건(0.01%)이었다(<표 19> 참조).

한국십진분류법의 주류에 따른 비율은 <그림 30>과 같다. 가장 비중이 높은 주류는 사회과학으로 346,357건, 약 29%를 차지하였으며, 다음으로 기술과학이 218,941건(약 18%), 문학 208,220건(약 17%) 순이었다. 상대적으로 비중이 낮은 주류는 자연과학, 총류, 철학, 언어 등이며, 대체로 3~4%의 비중을 차지하고 있었다.

<표 19> 전체 서지데이터 현황

분류	총류 (000)	철학 (100)	종교 (200)	사회과학 (300)	자연과학 (400)	소 계
건수	45,629	40,496	70,475	346,357	36,973	
분류	기술과학 (500)	예술 (600)	언어 (700)	문학 (800)	역사 (900)	1,199,643
건수	218,941	121,506	47,927	208,220	63,119	



<그림 30> 주류별 서지데이터 현황

전체 데이터에서 목차 콘텐츠가 있는 레코드에 주제명이 부여된 서지데이터는 <표 20>, <그림 31>과 같이 474,036건으로, 오프라인 서지데이터의 약 39% 정도인 것으로 나타났다.

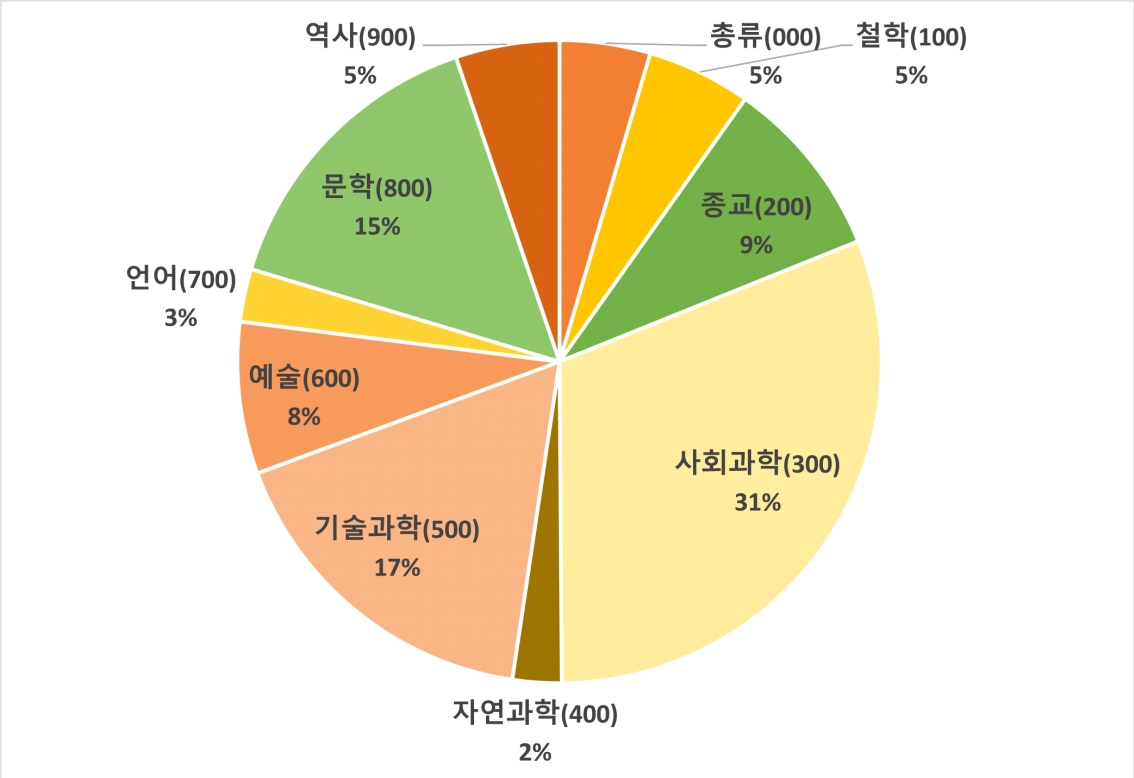
목차 기입이 가장 많은 주류는 사회과학으로 146,960건, 약 31%를 차지하였으며, 이후 기술과학이 80,626건으로 약 17%, 문학이 71,762건으로 약 15% 순이었다. 상대적으로 목차 기입이 적은 주류는 자연과학이었으며, 11,647건으로 전체

의 2%를 차지하였고, 이외 언어가 약 3% 비중을 나타내었다.

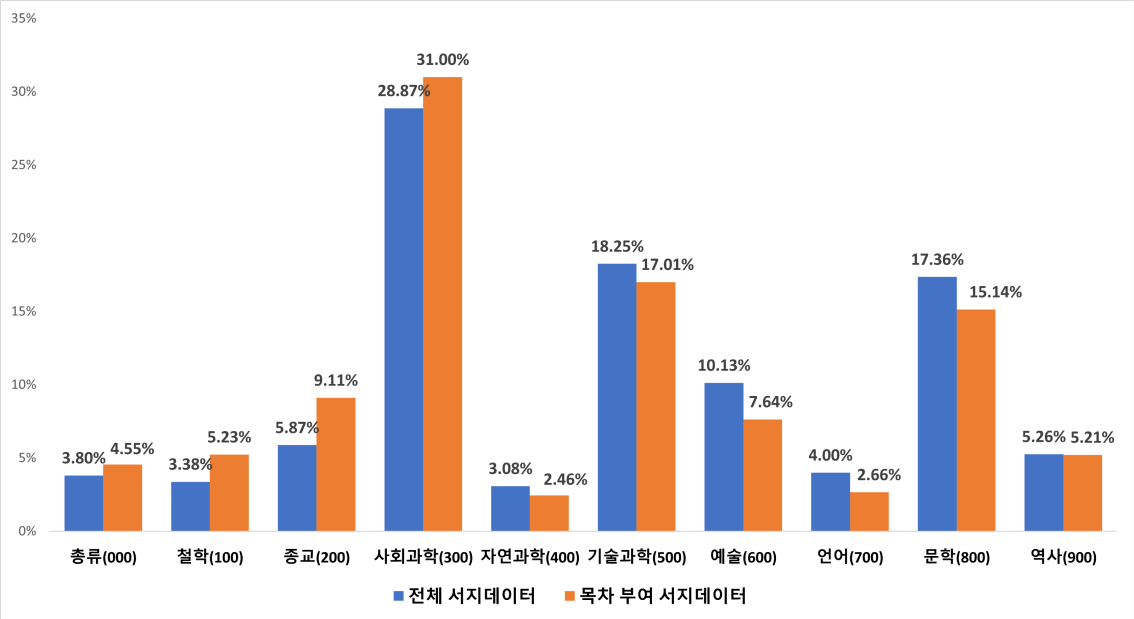
전체 서지데이터 대비 목차 기입 데이터의 비율을 확인하면 <그림 32>와 같다. 총류, 철학, 종교, 사회과학 영역이 전체 서지데이터에 대하여 상대적으로 목차 기입률이 높은 것으로 나타났으며, 예술과 언어 영역이 목차 기입률이 낮았다.

<표 20> 목차 입력 서지데이터 현황

분류	총류 (000)	철학 (100)	종교 (200)	사회과학 (300)	자연과학 (400)	소 계
건수	21,566	24,804	43,170	146,960	11,647	
분류	기술과학 (500)	예술 (600)	언어 (700)	문학 (800)	역사 (900)	474,036
건수	80,626	36,198	12,598	71,762	24,705	



<그림 31> 목차가 있는 서지데이터의 주류별 현황



<그림 32> 전체 서지데이터와 목차 기입데이터의 비율

3.2. 강목(division) 분석

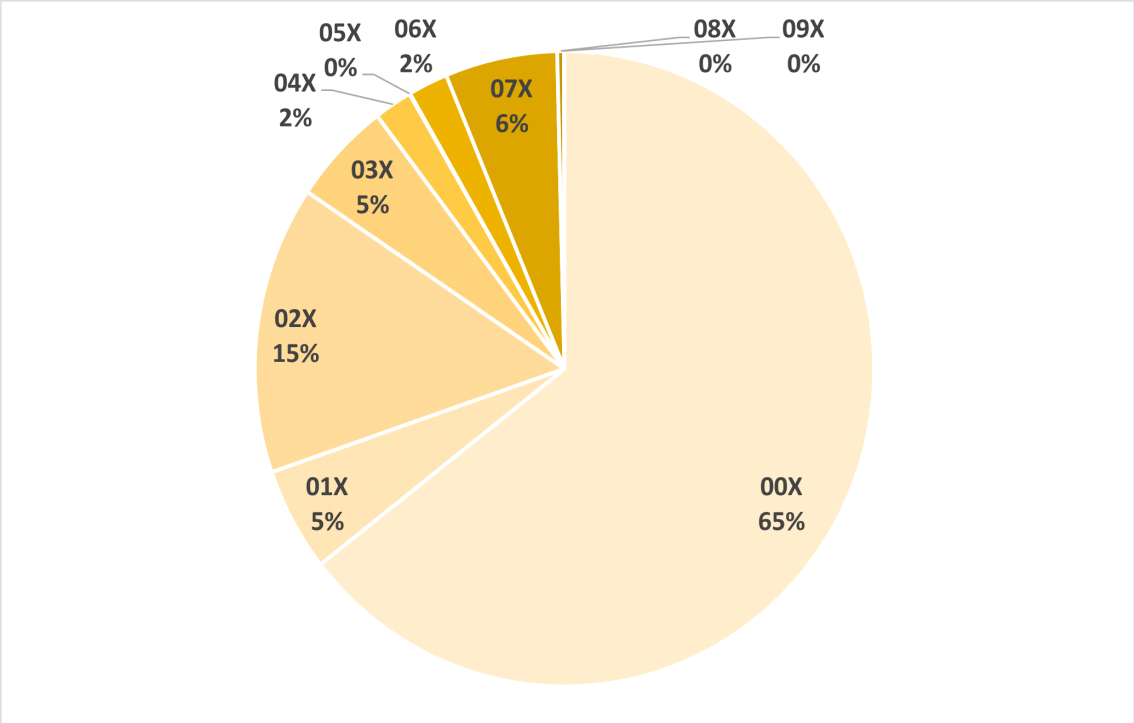
한국십진분류표 주류의 하위 세부 주제에 따른 분석을 위하여 분류 기호를 강목 수준으로 나누어 서지데이터의 분포 현황과 목차를 기입한 현황을 파악하였다.

· 총류 (<표 21>, <그림 33>, <그림 34> 참조)

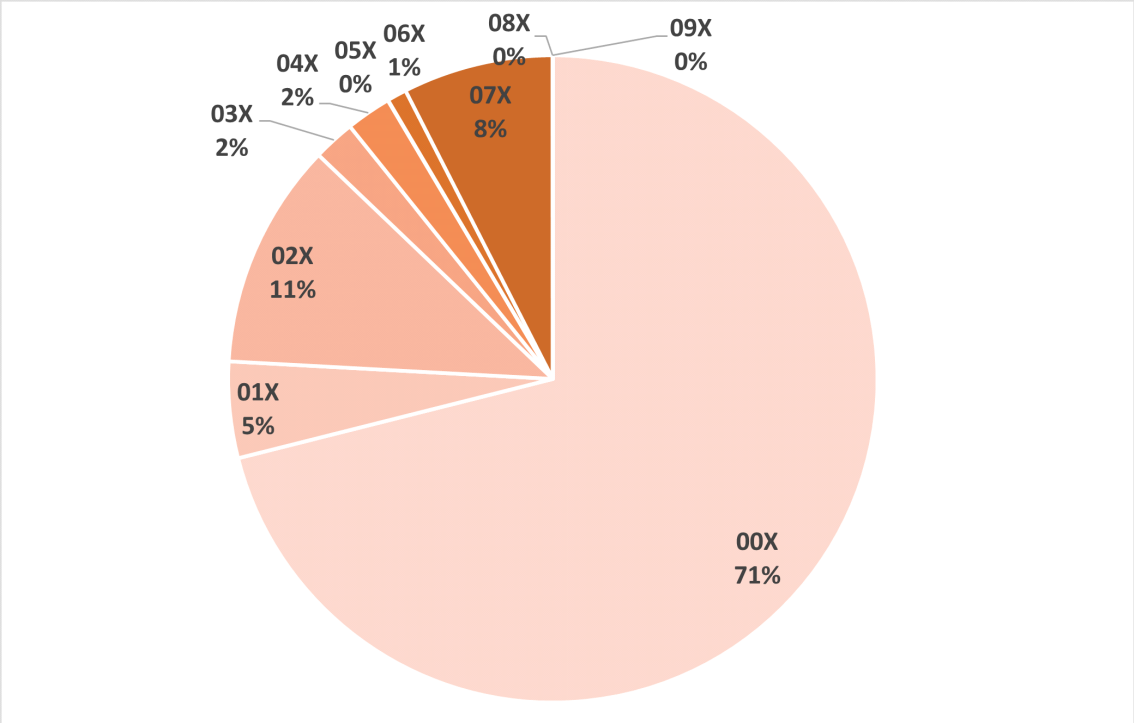
- 총류 서지데이터는 총 45,629건이며, 이 중 21,566건 약 47%의 목차가 기입되었다.
- 총류 분야에서 가장 높은 비중을 차지한 강목은 '00X-총류'로 29,402건이었고, 해당 강목의 목차가 기입된 건수는 15,325건으로 총류의 전체 목차 기입 데이터의 71%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율이 가장 높은 강목은 '07X-신문, 저널리즘'이었으며, 서지데이터 2,675건 중 1,614건으로 서지데이터 중 60% 정도는 목차가 있는 것으로 나타났다.

<표 21> 총류 서지데이터 현황

분류	00X	01X	02X	03X	04X	
	총류	도서학, 서지학	문헌정보학	백과사전	강연집,수필집,연설문집	
건수	29,402	2,413	6,697	2,400	909	(단위:건)
목차 기입	15,325	1,037	2,436	450	487	
목차 기입 비율(%)	52.10%	42.98%	36.37%	18.75%	53.58%	
분류	05X	06X	07X	08X	09X	계
	일반 연속간행물	일반학회, 단체, 협회, 기관, 연구기관	신문, 저널리즘	일반 전집, 총서	향토자료	
건수	17	945	2,675	162	9	45,629
목차 기입	4	210	1,614	2	1	21,566
목차 기입 비율(%)	23.53%	22.22%	60.34%	1.23%	11.11%	47.26%



<그림 33> 총류 서지데이터 강목 분포 현황



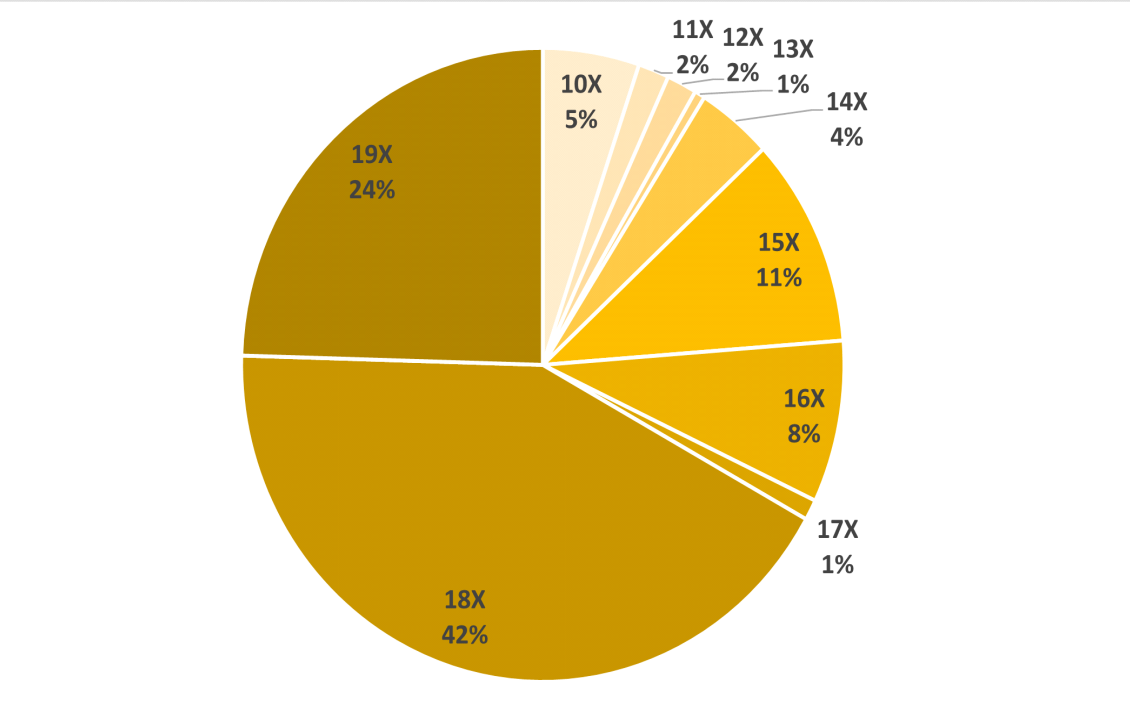
<그림 34> 총류 목차기입 서지데이터 비율 현황

· **철학 (<표 22>, <그림 35>, <그림 36> 참조)**

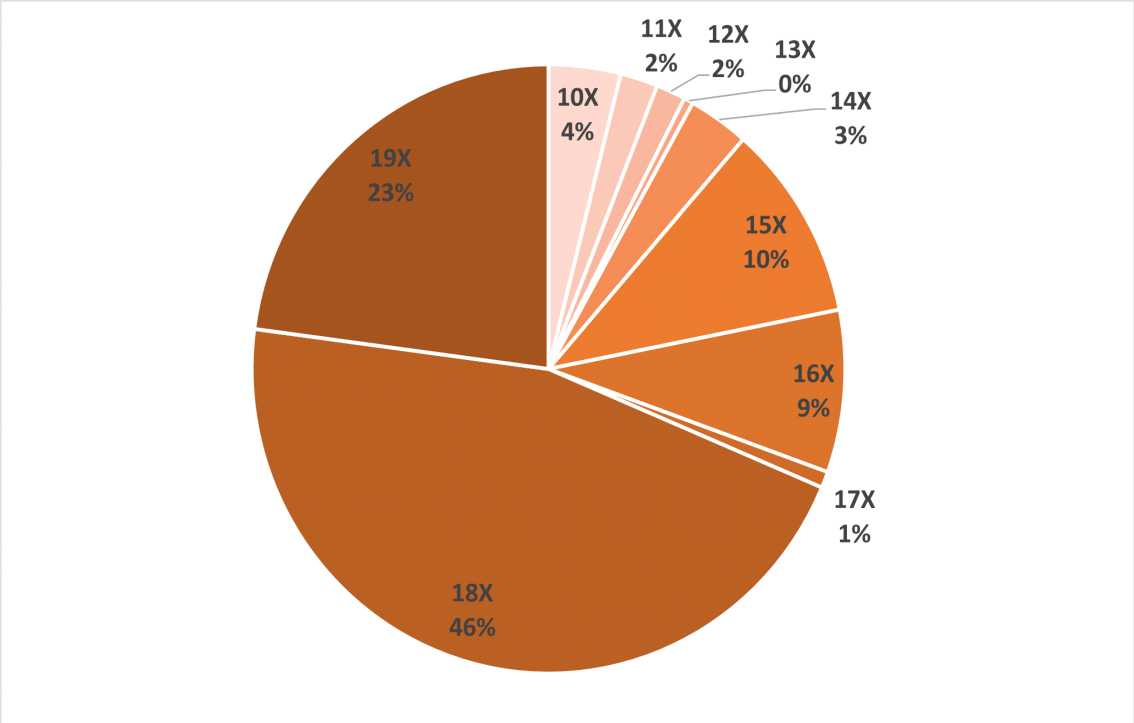
- 철학 분야 서지데이터는 총 40,496건이며, 이 중 목차 기입 건은 24,804건으로 약 61%의 목차가 기입된 것으로 나타났다.
- 철학 분야에서 가장 높은 비중을 차지한 강목은 '15X-동양철학' 으로 4,339건이었고, 해당 강목의 목차가 기입된 건수는 2,600건으로 철학의 전체 목차 기입 데이터의 46%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율이 가장 높은 강목은 '11X-형이상학'이었으며, 서지데이터 671건 중 508건으로 서지데이터 중 76% 정도 목차가 있는 것으로 나타났다.

<표 22> 철학 서지데이터 현황

분류	10X	11X	12X	13X	14X	(단위:건)
	철학	형이상학	인식론, 인과론, 인간학	철학의 세계	경학	
건수	2,096	671	652	227	1,642	
목차 기입	973	508	398	126	821	
목차 기입 비율(%)	46.42%	75.71%	61.04%	55.51%	50.00%	
분류	15X	16X	17X	18X	19X	계
	동양철학, 동양사상	서양철학	논리학	심리학	윤리학	
건수	4,339	3,351	425	17,159	9,934	40,496
목차 기입	2,600	2,139	215	11,344	5,680	24,804
목차 기입 비율(%)	59.92%	63.83%	50.59%	66.11%	57.18%	61.25%



<그림 35> 철학 서지데이터 강목 분포 현황



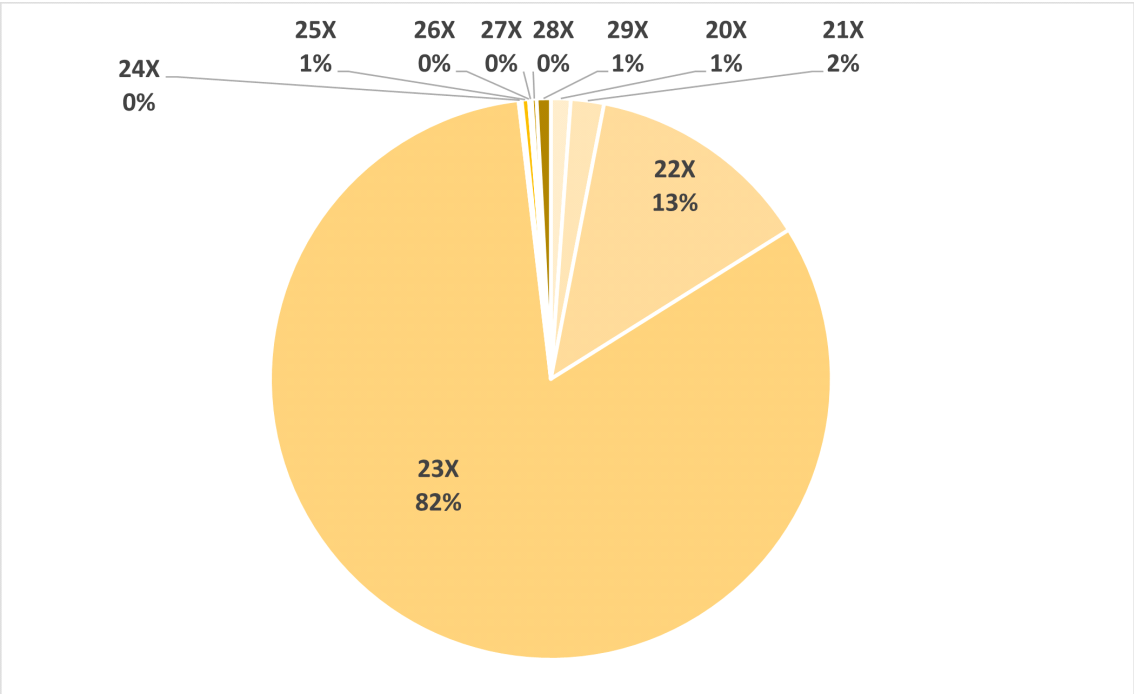
<그림 36> 철학 목차기입 서지데이터 비율 현황

· 종교 (<표 23>, <그림 37>, <그림 38> 참조)

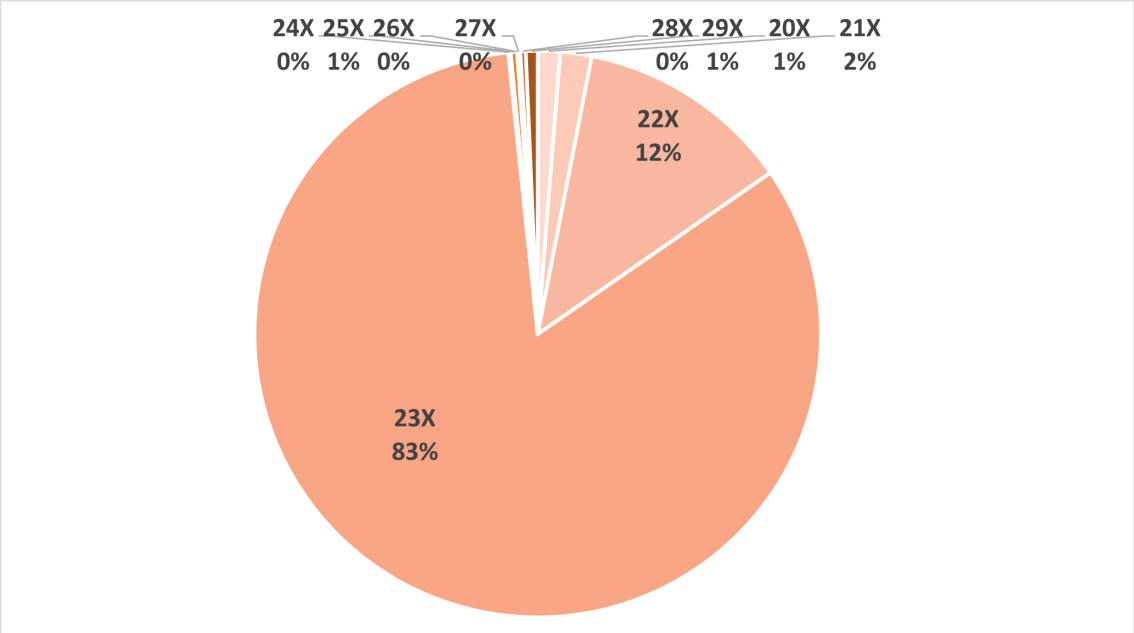
- 종교 분야 서지데이터는 총 70,475건이며, 이 중 목차 기입 건은 43,170건으로 약 61%의 목차가 기입된 것으로 나타났다.
- 종교 분야에서 가장 높은 비중을 차지한 강목은 '23X-기독교'로 57,838건이었고, 해당 강목의 목차가 기입된 건수는 35,825건으로 종교의 전체 목차 기입 데이터의 83%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율이 가장 높은 강목은 '26X-미사용'이었으며, 8건 중 8건으로 모두 목차가 기입된 것으로 나타났으나, 현재 KDC 6판에서 본 강목은 [미사용]으로, KCD 5판 기준 '신도'일 것으로 추측한다.
- 이 밖에 서지데이터 대비 목차 기입 데이터의 비율이 높은 강목은 '20X-종교'였으며, 이어 '28X-이슬람교(회교)', '27X-힌두교, 브라만교'로 나타났다.

<표 23> 종교 서지데이터 현황

분류	20X	21X	22X	23X	24X	(단위:건)
	종교	비교종교	불교	기독교	도교	
건수	802	1344	9,190	57,838	131	
목차 기입	550	771	5,305	35,825	61	
목차 기입 비율(%)	68.58%	57.37%	57.73%	61.94%	46.56%	
분류	25X	26X	27X	28X	29X	계
	천도교	[미사용]	힌두교, 브라만교	이슬람교 (회교)	기타 제종교	
건수	290	8	106	194	572	70,475
목차 기입	154	8	72	132	292	43,170
목차 기입 비율(%)	53.10%	100.00%	67.92%	68.04%	51.05%	61.26%



<그림 37> 종교 서지데이터 강목 분포 현황



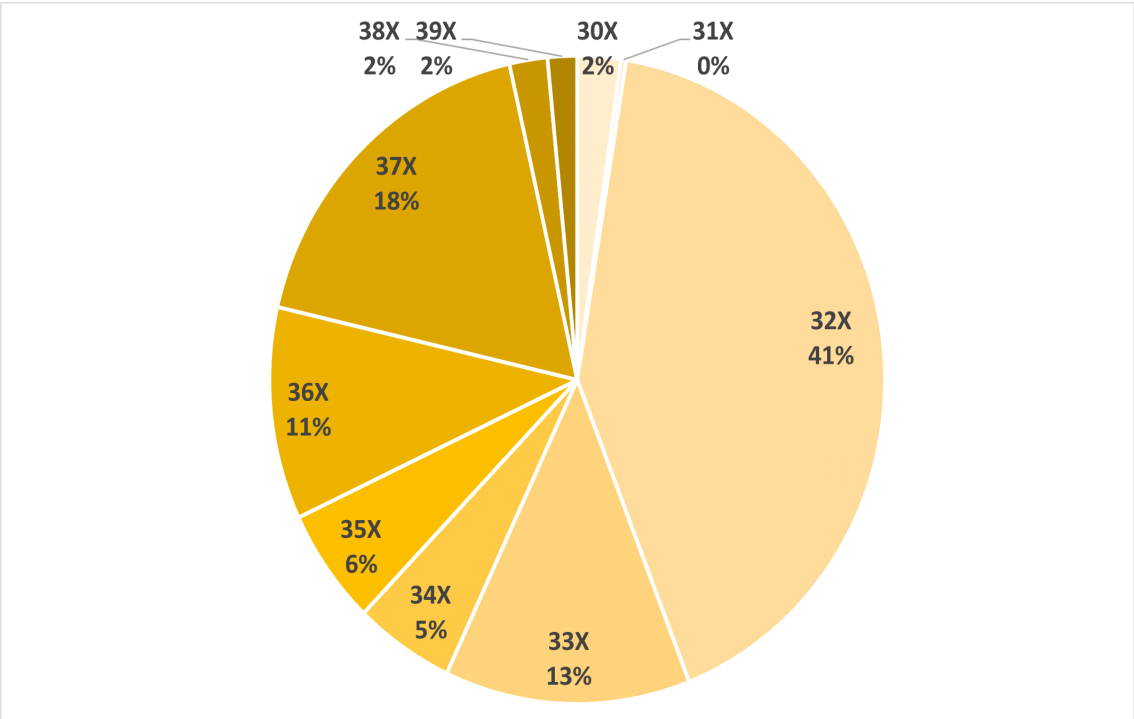
<그림 38> 종교 목차 기입 서지데이터 비율 현황

· **사회과학 (<표 24>, <그림 39>, <그림 40> 참조)**

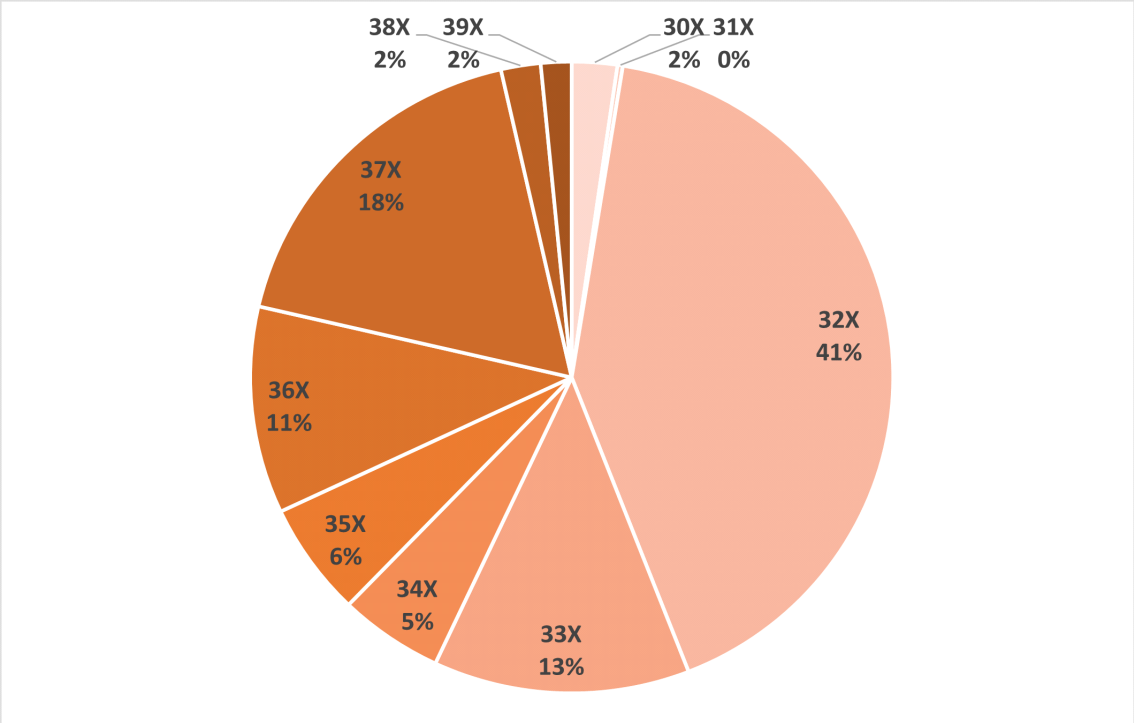
- 사회과학 분야 서지데이터는 총 346,357건이며, 이 중 목차 기입 건은 146,960건으로 약 42%의 목차가 기입된 것으로 나타났다.
- 사회과학 전체 서지데이터 중 가장 높은 비중을 차지한 강목은 '32X-경제학'으로 143.842건이었고, 해당 강목의 목차가 기입된 건수는 67,008건으로 사회과학의 전체 목차 기입 데이터의 42%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율이 가장 높은 강목은 '31X-통계학'이었으며, 913건 중 543건으로 서지데이터 중 약 59% 정도 목차가 기입되어 있었다. 이어서 '34X-정치학' 53%, '33X-사회학, 사회문제' 48% 순으로 목차 기입률을 확인할 수 있다.

<표 24> 사회과학 서지데이터 현황

분류	30X	31X	32X	33X	34X	
	사회과학	통계자료	경제학	사회학, 사회문제	정치학	
건수	7,906	913	143,842	44,668	18,145	(단위:건)
목차 기입	3,545	543	67,008	21,586	9,608	
목차 기입 비율(%)	44.84%	59.47%	46.58%	48.33%	52.95%	
분류	35X	36X	37X	38X	39X	계
	행정학	법률, 법학	교육학	풍습, 예절, 민속학	국방, 군사학	
건수	20,127	36,697	61,802	6,906	5,351	346,357
목차 기입	5,498	13,832	20,658	3,030	1,652	146,960
목차 기입 비율(%)	27.32%	37.69%	33.43%	43.87%	30.87%	42.43%



<그림 39> 사회과학 서지데이터 강목 분포 현황



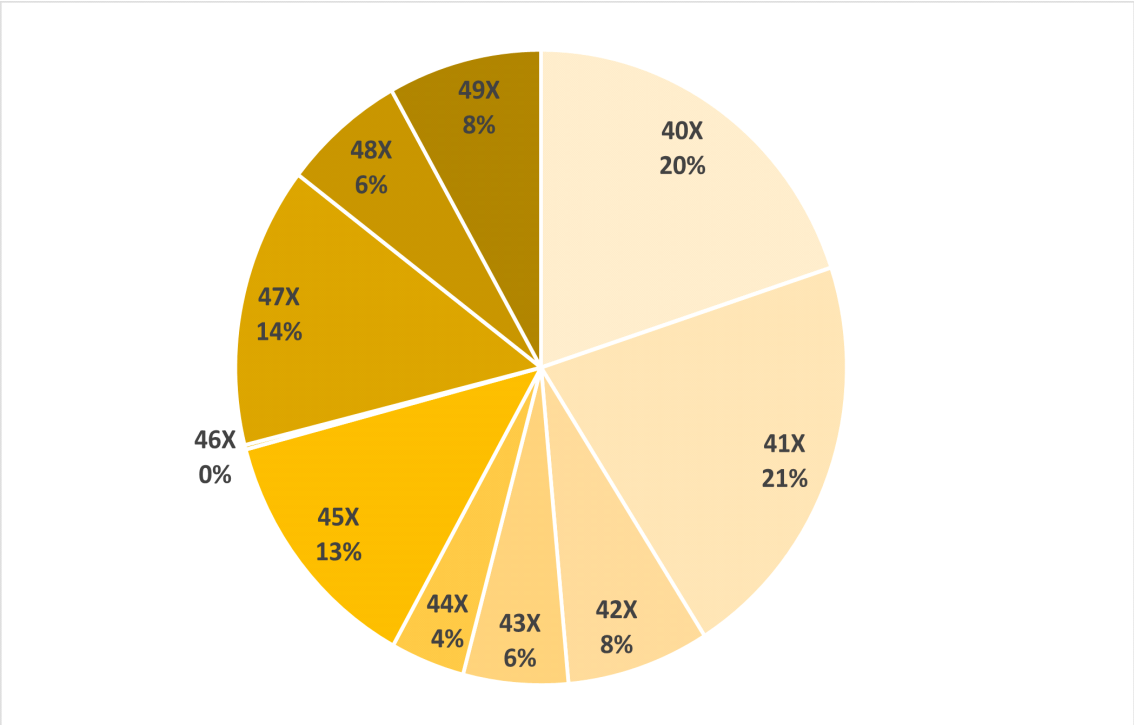
<그림 40> 사회과학 목차 기입 서지데이터 비율 현황

· 자연과학 (<표 25>, <그림 41>, <그림 42> 참조)

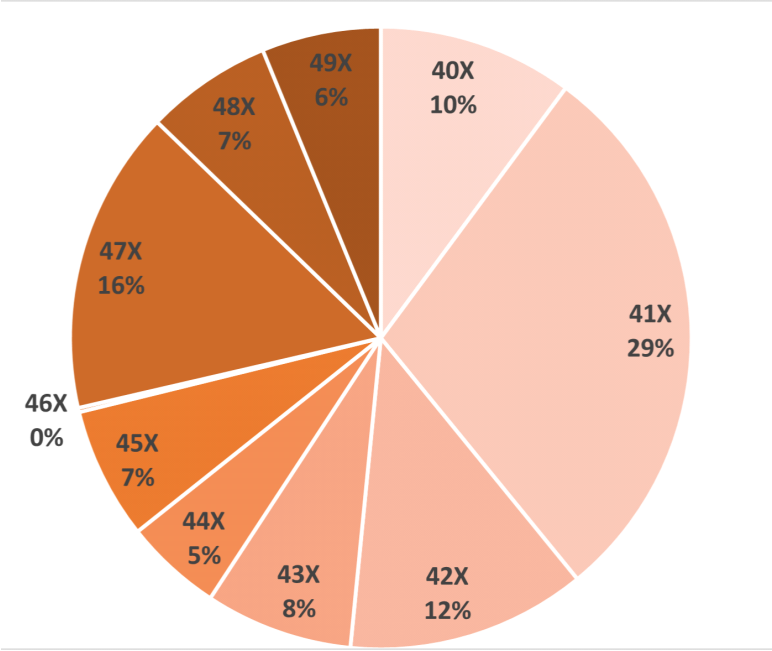
- 자연과학 분야 서지데이터는 총 36,973건이며, 이 중 목차 기입 건은 11,647건으로 약 32%의 목차가 기입된 것으로 나타났다.
- 자연과학 분야에서 가장 높은 비중을 차지한 강목은 '41X-수학'으로 7,805건이었고, 해당 강목의 목차가 기입된 건수는 3,372건으로 자연과학의 전체 목차 기입 데이터의 29%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율이 가장 높은 강목은 '42X-물리학'이었으며, 2,792건 중 1,452건으로 서지데이터 중 약 52% 정도 목차가 기입되어 있었다.

<표 25> 자연과학 서지데이터 현황

분류	40X	41X	42X	43X	44X	(단위:건)
	자연과학	수학	물리학	화학	천문학	
건수	7,360	7,805	2,792	2,050	1,453	
목차 기입	1,182	3,372	1,452	895	590	
목차 기입 비율(%)	16.06%	43.20%	52.01%	43.66%	40.61%	
분류	45X	46X	47X	48X	49X	계
	지학	광물학	생명과학	식물학	동물학	
건수	4,733	84	5,290	2,399	3,007	36,973
목차 기입	798	26	1,842	767	723	11,647
목차 기입 비율(%)	16.86%	30.95%	34.82%	31.97%	24.04%	31.50%



<그림 41> 자연과학 서지데이터 강목 분포 현황



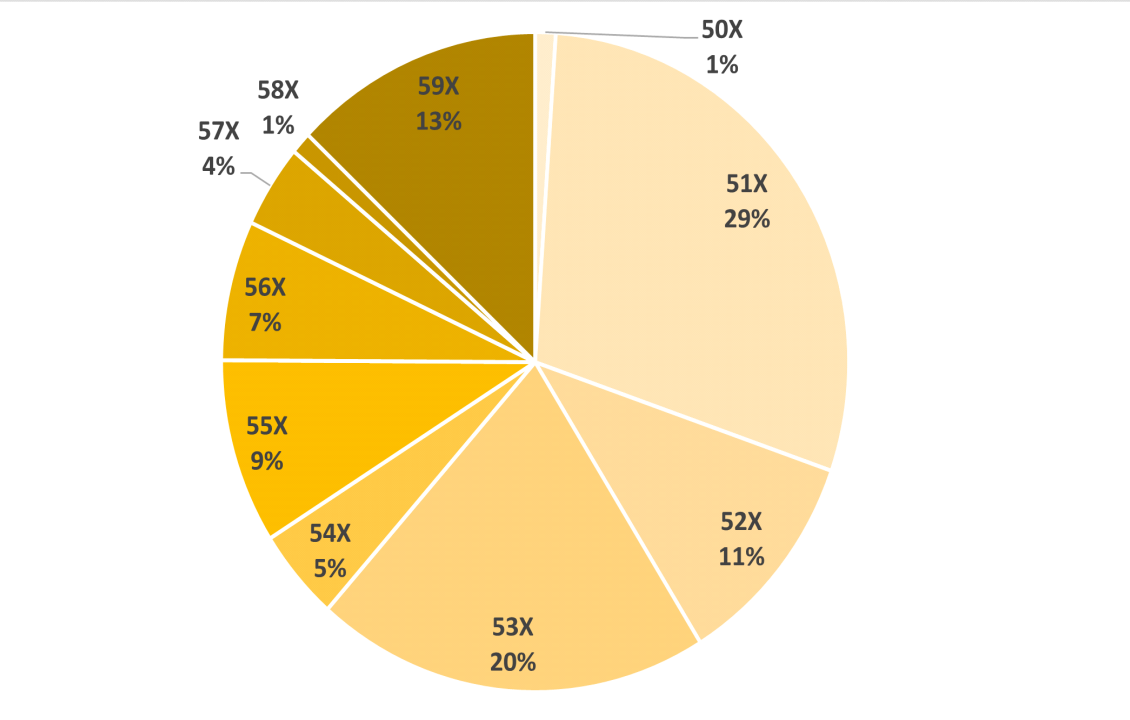
<그림 42> 자연과학 목차 기입 서지데이터 비율 현황

· 기술과학 (<표 26>, <그림 43>, <그림 44> 참조)

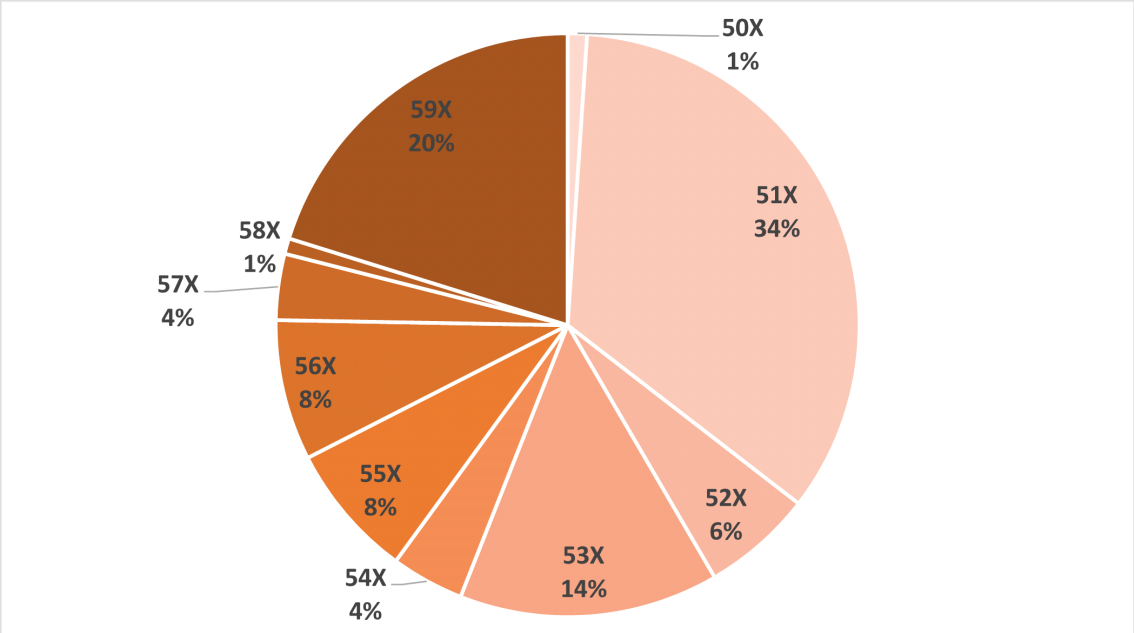
- 기술과학 분야 서지데이터는 총 218,941건이며, 이 중 목차 기입 건은 80,626건으로 약 37%의 목차가 기입된 것으로 나타났다.
- 기술과학 분야에서 가장 높은 비중을 차지한 강목은 '51X-화학'으로 64,310건이었고, 해당 강목의 목차가 기입된 건수는 27,735건으로 기술과학의 전체 목차 기입 데이터의 29%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율이 가장 높은 강목은 '59X-생활과학'이었으며, 28,229건 중 16,282건으로 서지데이터 중 약 58% 정도 목차가 기입되어 있었다.

<표 26> 기술과학 서지데이터 현황

분류	50X	51X	52X	53X	54X	(단위:건)
	기술과학	의학	농업, 농학	공학, 공업일반, 토목공학, 환경공학	건축, 건축학	
건수	2,308	64,130	23,709	44,537	9,871	
목차 기입	861	27,735	4,945	11,592	3,261	
목차 기입 비율(%)	37.31%	43.25%	20.86%	26.03%	33.04%	
분류	55X	56X	57X	58X	59X	계
	기계공학	전기공학, 통신공학, 전자공학	화학공학	제조업	생활과학	
건수	19,857	15,027	8,932	2,341	28,229	218,941
목차 기입	6,019	6,282	2,952	697	16,282	80,626
목차 기입 비율(%)	30.31%	41.80%	33.05%	29.77%	57.68%	36.83%



<그림 43> 기술과학 서지데이터 강목 분포 현황



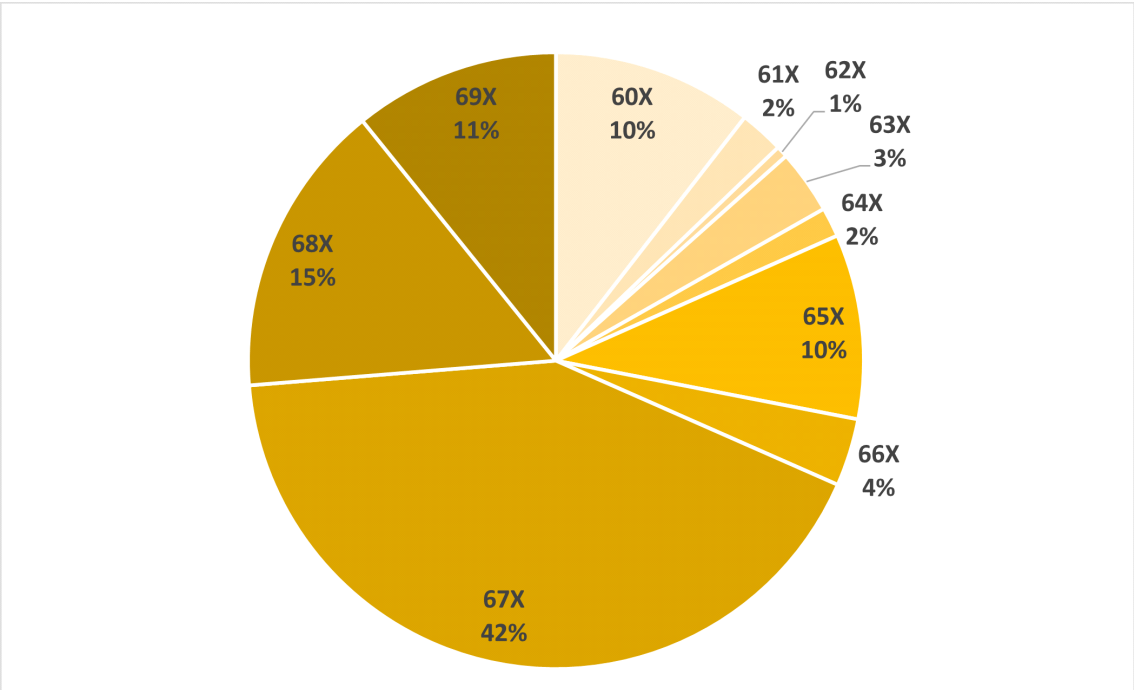
<그림 44> 기술과학 목차 기입 서지데이터 비율 현황

· **예술 (<표 27>, <그림 45>, <그림 46> 참조)**

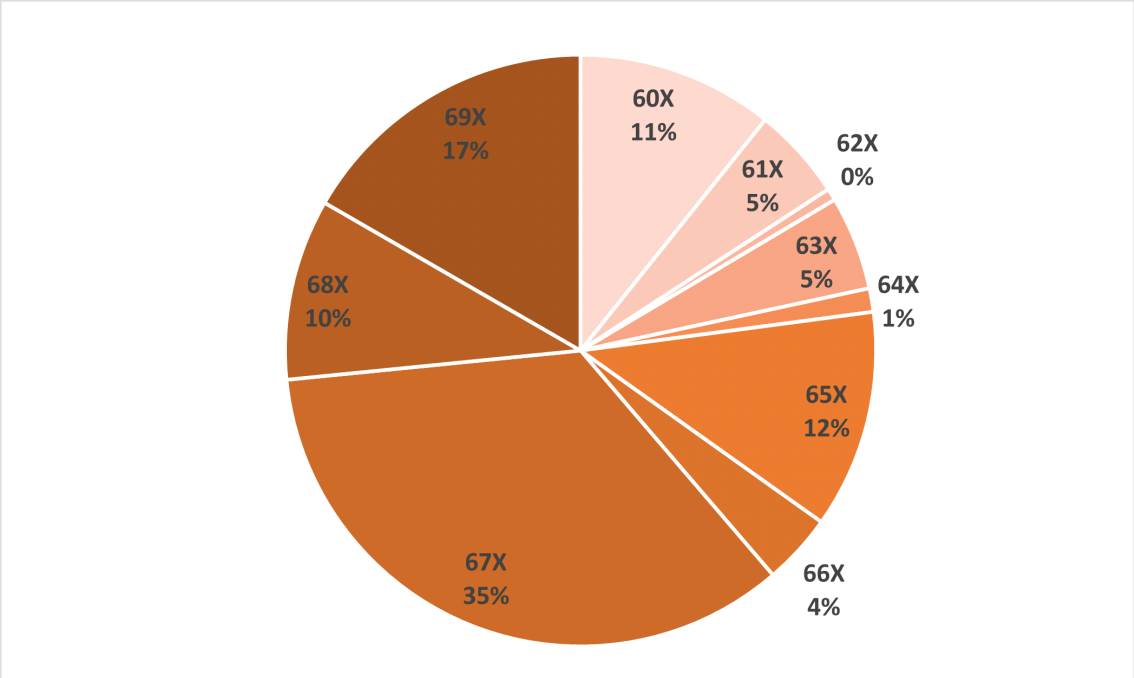
- 예술 분야 서지데이터는 총 121,506건이며, 이 중 목차 기입 건은 36,198건으로 약 30%의 목차가 기입된 것으로 나타났다.
- 예술 분야에서 가장 높은 비중을 차지한 강목은 '67X-음악'으로 51,182건이었고, 해당 강목의 목차가 기입된 건수는 12,565건으로 예술의 전체 목차 기입 데이터의 58%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율이 가장 높은 강목은 '61X-[미사용]'이었으며, 2,821건 중 1,836건으로 서지데이터 중 약 65% 정도 목차가 기입되어 있었으나, 현재 KDC 6판에서는 미사용 강목으로 기존 서지데이터는 KDC 4판의 '건축술'에 해당하는 것으로 확인하였다.

<표 27> 예술 서지데이터 현황

분류	60X	61X	62X	63X	64X	(단위:건)
	예술	[미사용]	조각, 조형미술	공예	서예	
건수	12,774	2,821	741	4,081	1,881	
목차 기입	3,891	1,836	238	1,861	466	
목차 기입 비율(%)	30.46%	65.08%	32.12%	45.60%	24.77%	
분류	65X	66X	67X	68X	69X	계
	회화, 도화, 디자인	사진예술	음악	공연예술, 매체예술	오락, 스포츠	
건수	11,787	4,306	51,182	18,792	13,141	121,506
목차 기입	4,319	1,403	12,565	3,584	6,035	36,198
목차 기입 비율(%)	36.64%	32.58%	24.55%	19.07%	45.92%	29.79%



<그림 45> 예술 서지데이터 강목 분포 현황



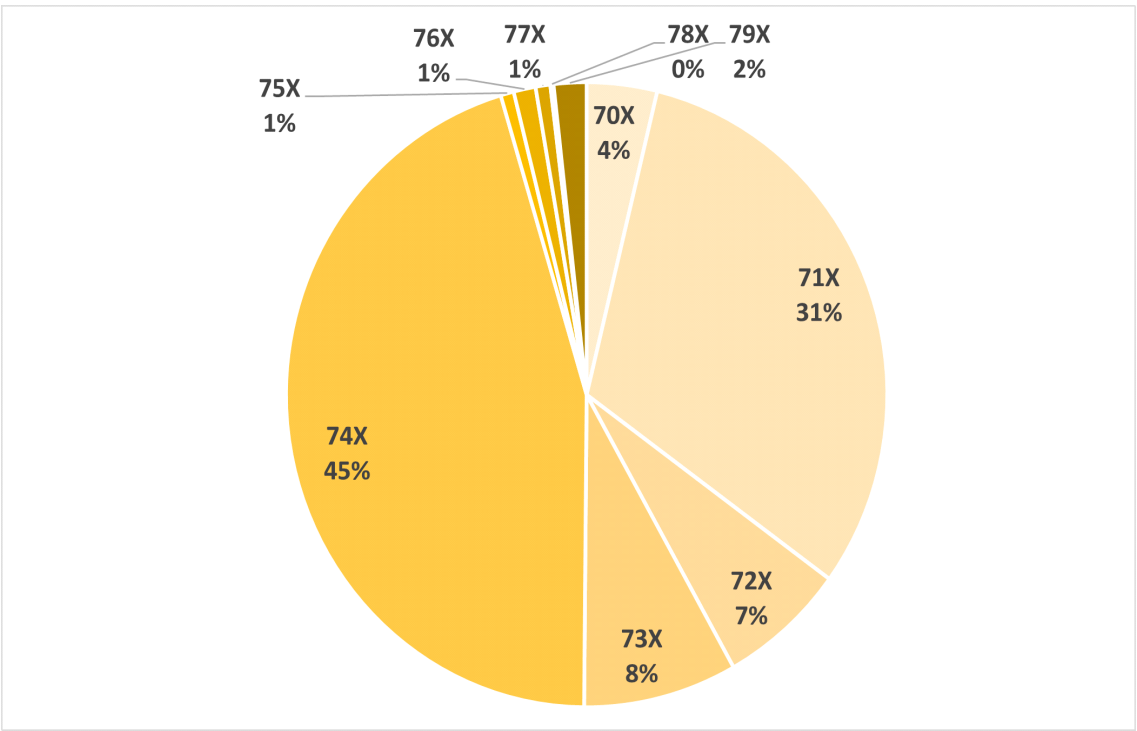
<그림 46> 예술 목차 기입 서지데이터 비율 현황

· **언어 (<표 28>, <그림 47>, <그림 48> 참조)**

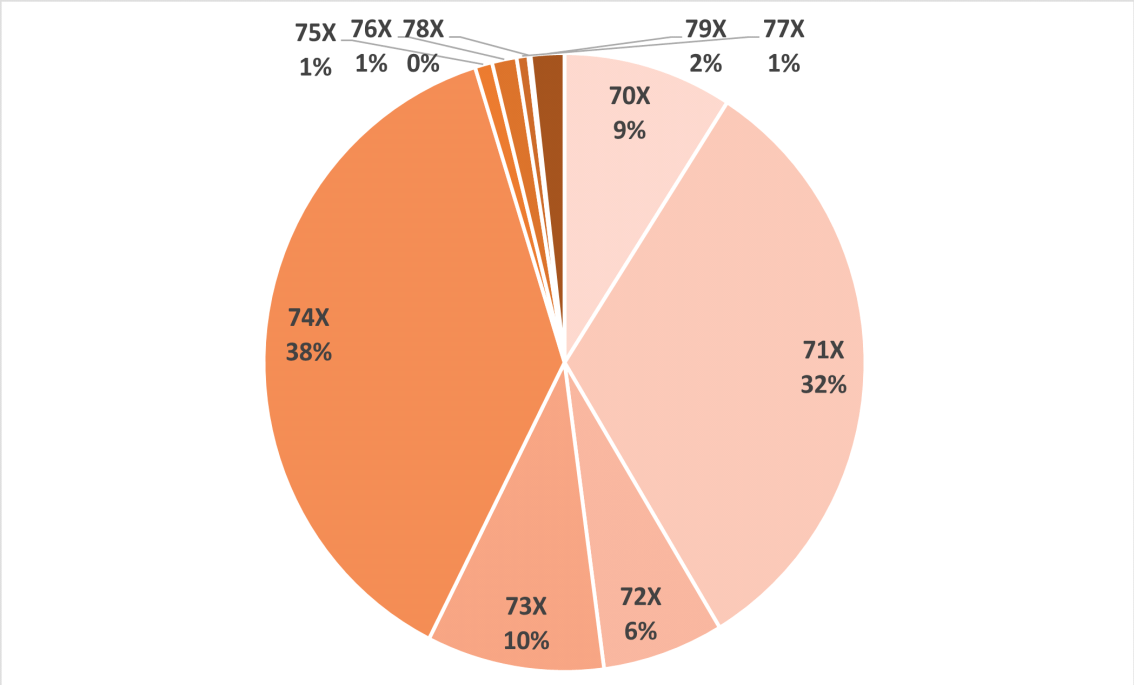
- 언어 분야 서지데이터는 총 47,927건이며, 이 중 목차 기입 건은 12,598건으로 약 26%의 목차가 기입된 것으로 나타났다.
- 언어 분야에서 가장 높은 비중을 차지한 강목은 '74X-영어'로 21,688건이었고, 해당 강목의 목차가 기입된 건수는 4,756건으로 언어의 전체 목차 기입 데이터의 38%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율이 가장 높은 강목은 '70X-언어'였으며, 1,812건 중 1,146건으로 서지데이터 중 약 63% 정도 목차가 기입되어 있었다.

<표 28> 언어 서지데이터 현황

분류	70X	71X	72X	73X	74X	(단위:건)
	언어	한국어	중국어	일본어 및 기타 아시아제어	영어	
건수	1,812	14,980	3,275	3,960	21,688	
목차 기입	1,146	4,066	822	1,202	4,756	
목차 기입 비율(%)	63.25%	27.14%	25.10%	30.35%	21.93%	
분류	75X	76X	77X	78X	79X	계
	독일어	프랑스어	스페인어 및 포르투칼어	이탈리아어	기타 제어	
건수	340	567	379	83	843	47,927
목차 기입	117	167	79	16	227	12,598
목차 기입 비율(%)	34.41%	29.45%	20.84%	19.28%	26.93%	26.29%



<그림 47> 언어 서지데이터 강목 분포 현황



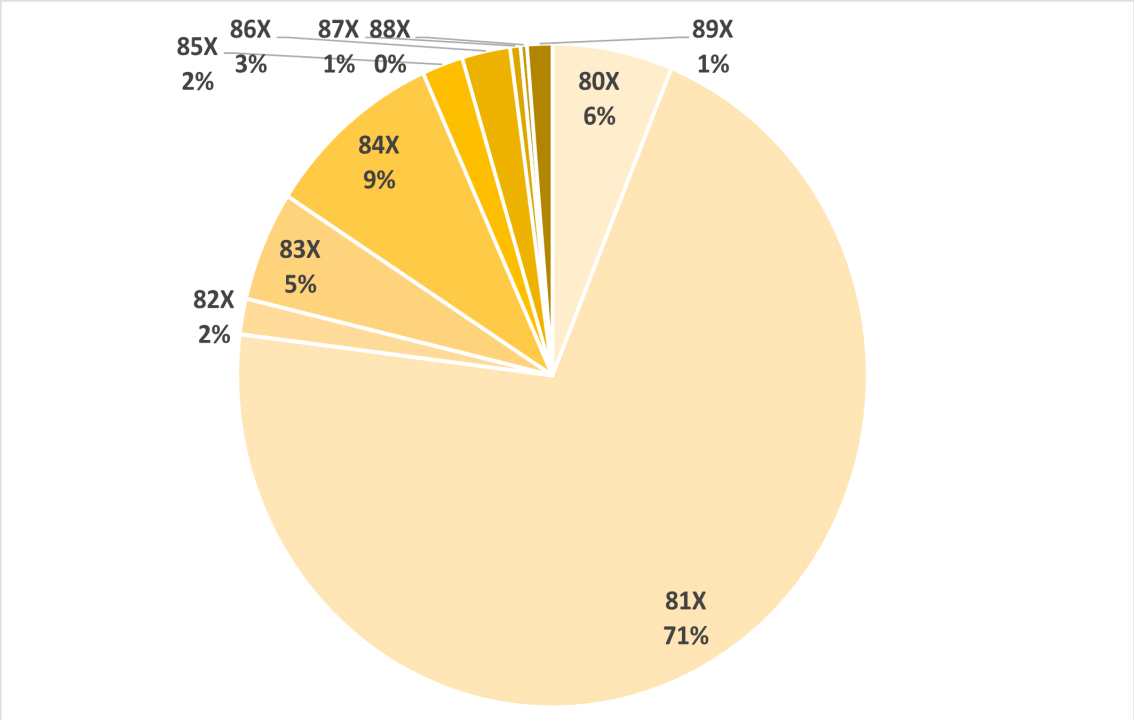
<그림 48> 언어 목차 기입 서지데이터 비율 현황

· 문학 (<표 29>, <그림 49>, <그림 50> 참조)

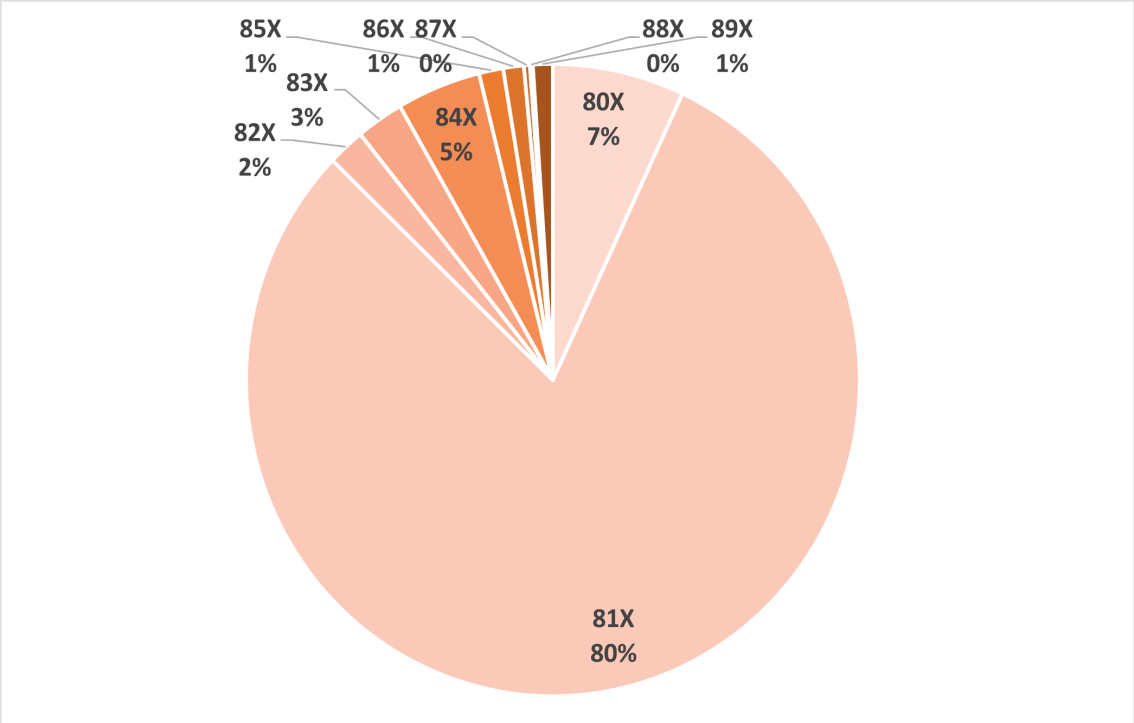
- 문학 분야 서지데이터는 총 208,220건이며, 이 중 목차 기입 건은 71,762건으로 약 34%의 목차가 기입된 것으로 나타났다.
- 문학 분야에서 가장 높은 비중을 차지한 강목은 '81X-한국문학'으로 147,464건이었고, 해당 강목의 목차가 기입된 건수는 57,618건으로 문학의 전체 목차 기입 데이터의 80%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율에서 '80X-문학', '81X-한국문학', '82X-중국문학'등이 모두 약 39%의 정도 목차가 기입되어 있었다.

<표 29> 문학 서지데이터 현황

분류	80X	81X	82X	83X	84X	(단위:건)
	문학	한국문학	중국문학	일본문학 및 기타아시아 제문학	영미문학	
건수	12,838	147,464	3,624	11,173	19,090	
목차 기입	4,982	57,618	1,405	1,793	3,184	
목차 기입 비율(%)	38.81%	39.07%	38.77%	16.05%	16.68%	
분류	85X	86X	87X	88X	89X	계
	독일문학	프랑스 문학	스페인 및 포르투갈 문학	이탈리아 문학	기타 제문학	
건수	4,389	5,141	1,113	686	2,702	208,220
목차 기입	916	779	225	110	750	71,762
목차 기입 비율(%)	20.87%	15.15%	20.22%	16.03%	27.76%	34.46%



<그림 49> 문학 서지데이터 강목 분포 현황



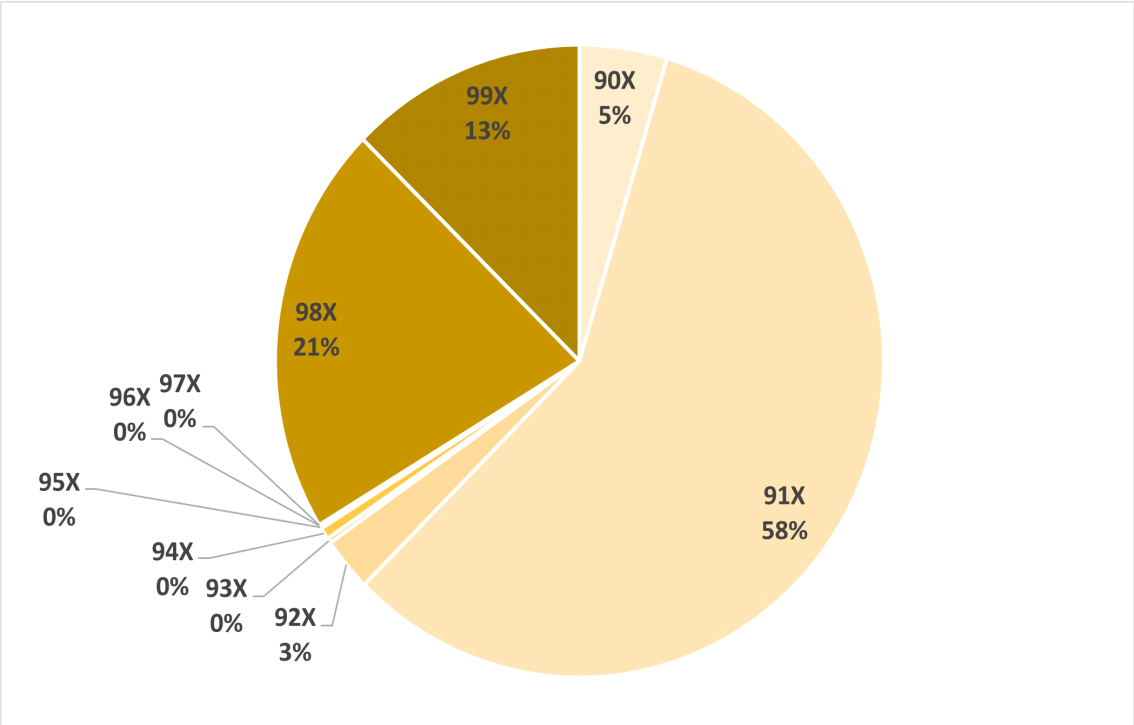
<그림 50> 문학 목차 기입 서지데이터 비율 현황

· **역사 (<표 30>, <그림 51>, <그림 52> 참조)**

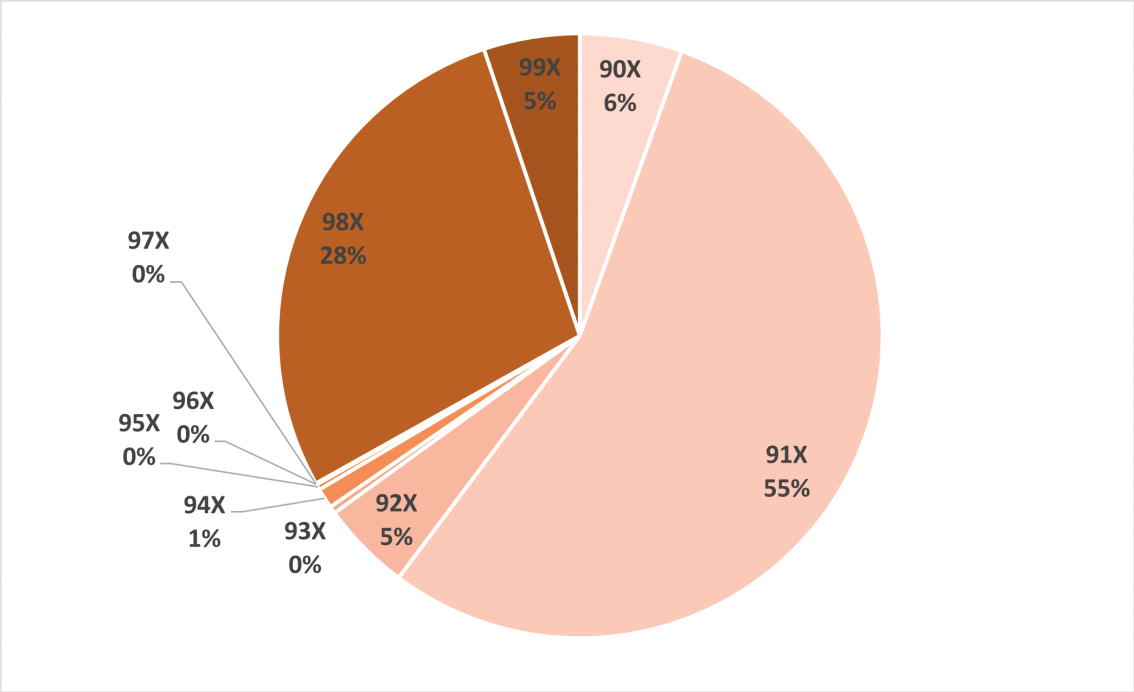
- 역사 분야 서지데이터는 총 63,119건이며, 이 중 목차 기입 건은 24,705건으로 약 26%의 목차가 기입된 것으로 나타났다.
- 역사 분야에서 가장 높은 비중을 차지한 강목은 '91X-아시아'로 36,518건이었고, 해당 강목의 목차가 기입된 건수는 13,528건으로 역사의 전체 목차 기입 데이터의 55%를 차지하였다.
- 각 강목 서지데이터 대비 목차 기입 데이터의 비율이 가장 높은 강목은 '95X-남아메리카'였으며, 93건 중 72건으로 서지데이터 중 약 77% 정도 목차가 기입되어 있었다. 이어서 '94X-북아메리카' 70%, '92X-유럽'이 68% 순으로 나타났다.

<표 30> 역사 서지데이터 현황

분류	90X	91X	92X	93X	94X	(단위:건)
	역사	아시아	유럽	아프리카	북아메리카	
건수	2,938	36,518	1,738	164	383	
목차 기입	1,351	13,528	1,189	95	268	
목차 기입 비율(%)	45.98%	37.04%	68.41%	57.93%	69.97%	
분류	95X	96X	97X	98X	99X	계
	남아메리카	오세아니아, 양극지방	[미사용]	지리	전기	
건수	93	32	3	13,285	7,965	63,119
목차 기입	72	18	2	6,914	1,268	24,705
목차 기입 비율(%)	77.42%	56.25%	66.67%	52.04%	15.92%	39.14%



<그림 51> 역사 서지데이터 강목 분포 현황



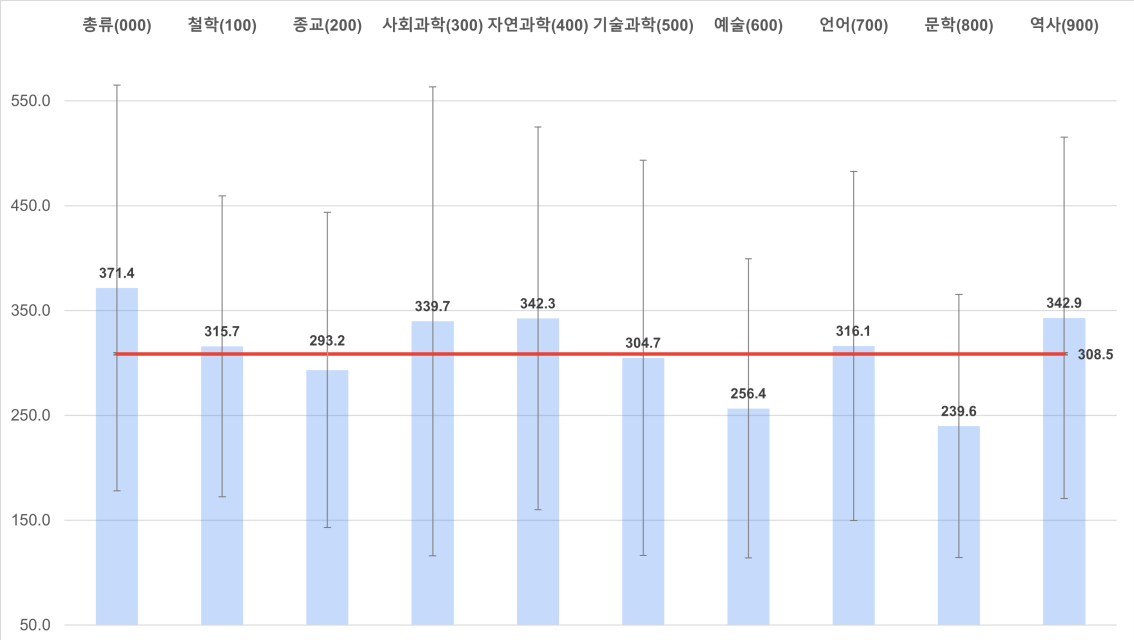
<그림 52> 역사 목차 기입 서지데이터 비율 현황

4. 목차 데이터

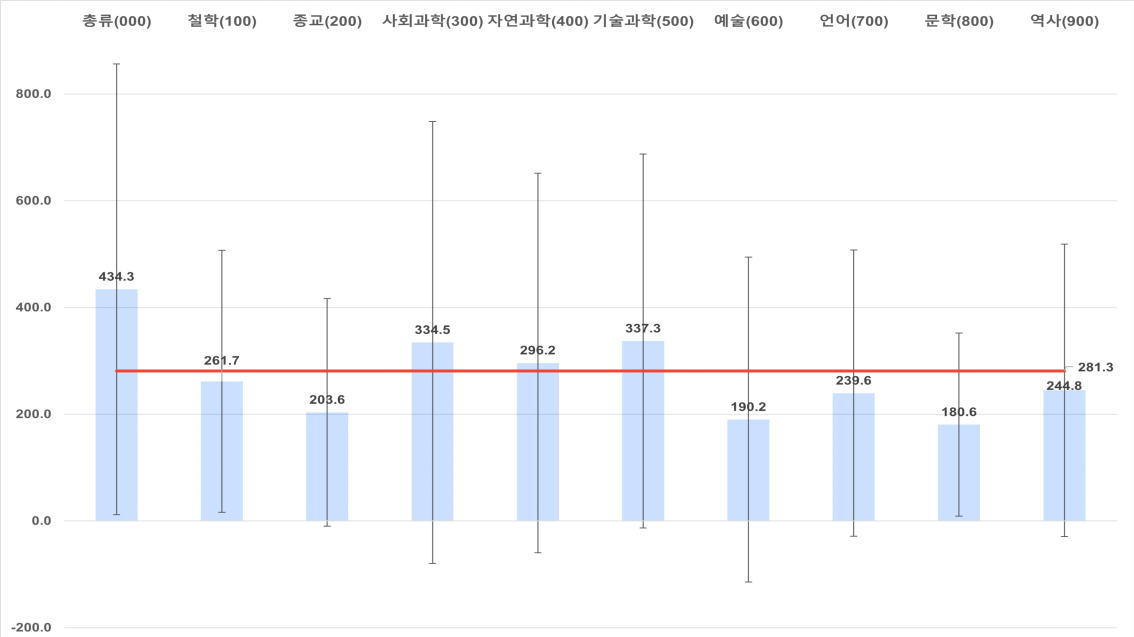
주제명이 부여된 전체 서지데이터 중 목차가 기입된 서지데이터는 총 474,140건으로 약 38.9%에 해당한다. 목차에 대한 수치적 정보로 각 목차가 어느 정도의 길이를 갖는지 파악하고자 목차를 구성하는 어절과 음절의 수를 계산하였으며, 또한 해당 목차가 수록된 대상 자료의 전체 쪽수를 파악하였다. 이를 KDC 주류에 따른 기술 통계 현황을 살펴보면 <표 31>과 같다. 평균적으로 총류가 쪽수, 목차의 어절 수와 음절 수 모두 많은 것으로 보였으며, 사회과학의 경우 쪽수, 어절 그리고 음절의 편차가 상대적으로 큰 것을 확인할 수 있다. 그리고 문학의 경우, 목차 어절과 음절의 평균과 편차 모두가 다른 주류에 비해 낮은데, 문학 분야의 목차가 다른 주제 분야보다 그 길이나 수록 내용이 적음을 의미한다. 한편, 자연과학은 쪽수는 많으나 어절과 음절이 상대적으로 적은 것으로 나타났다(<그림 53>, <그림 54>, <그림 55> 참조).

<표 31> 주류별 목차의 길이 관련 통계 현황

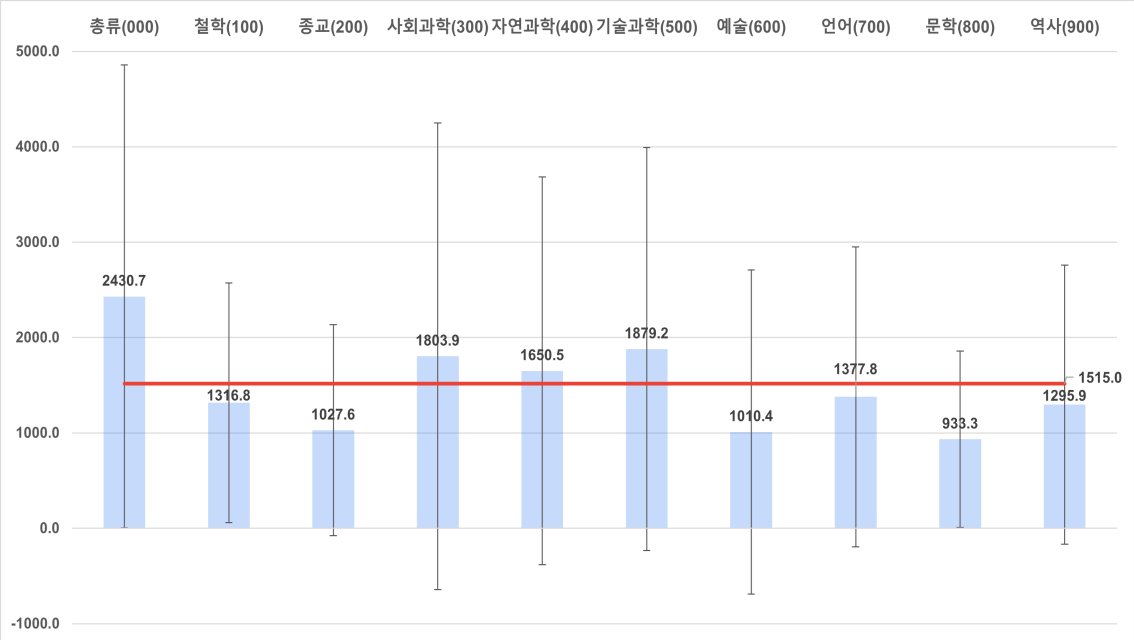
주류	쪽수				어절				음절			
	평균	편차	최소	최대	평균	편차	최소	최대	평균	편차	최소	최대
총류 (000)	371.4	193.5	3	2,918	434.3	422.3	3	13,857	2430.7	2427.6	26	73,337
철학 (100)	315.7	143.4	3	4,477	261.7	245.3	1	5,840	1316.8	1256.3	11	24,182
종교 (200)	293.2	150.4	2	4,952	203.6	213.5	1	4,308	1027.6	1105.8	3	24,722
사회 과학 (300)	339.7	223.7	1	6,834	334.5	414.1	1	40,380	1803.9	2444.7	3	210,893
자연 과학 (400)	342.3	182.5	3	4,010	296.2	355.5	3	7,052	1650.5	2033.0	24	38,804
기술 과학 (500)	304.7	188.4	1	4,455	337.3	350.1	3	17,718	1879.2	2111.0	16	176,512
예술 (600)	256.4	142.7	2	3,274	190.2	304.4	1	40,651	1010.4	1698.9	8	227,774
언어 (700)	316.1	166.4	2	2,110	239.6	268.3	4	5,436	1377.8	1572.6	16	33,857
문학 (800)	239.6	125.5	1	1,784	180.6	171.4	1	9,005	933.3	924.6	10	47,633
역사 (900)	342.9	172.3	1	4,018	244.8	274.0	1	13,839	1295.9	1462.3	17	70,831



<그림 53> 주류별 쪽수 평균과 표준편차



<그림 54> 주류별 어절 평균과 표준편차

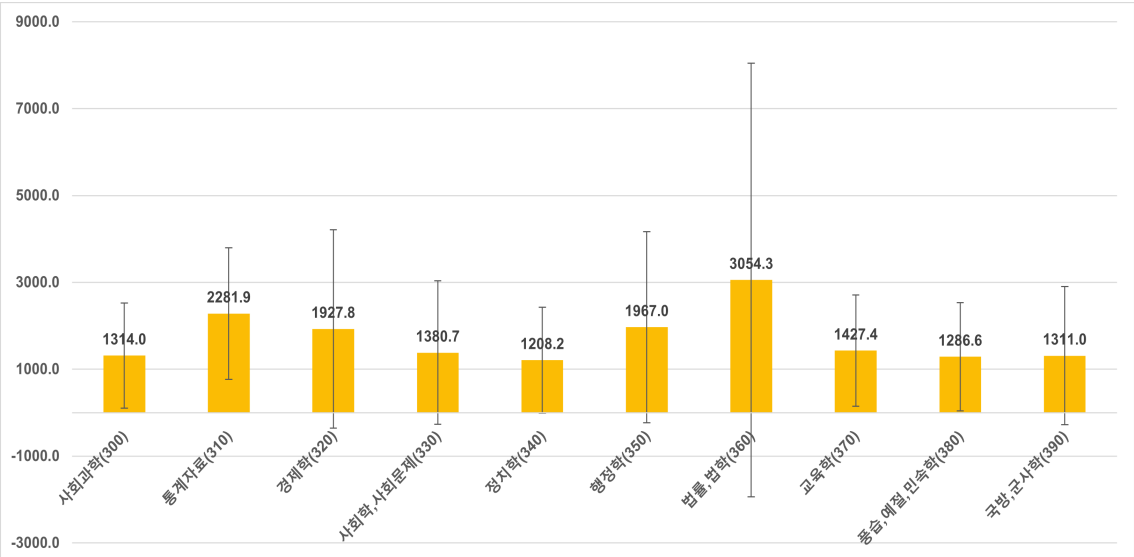


<그림 55> 주류별 음절 평균과 표준편차

성격이 다른 두 학문 분야인 사회과학과 문학에 대해 좀 더 깊이 있는 분석을 위해 이들 분야의 강목에서 목차 길이의 차이를 분석하면, 각각 <표 32>와 <표 33>과 같다. 사회과학의 하위 분야에서 '법률, 법학'이 가장 음절이 많고 편차가 큰 것으로 나타났으며, 최소 음절은 3개, 최대 음절은 210,893개였다. 문학의 경우, '중국문학'이 음절이 가장 많고 편차도 큰 것으로 나타났으며, 최소 음절은 22개, 최대 음절은 12,735개이었다. '한국문학'은 세 번째로 음절이 많은 강목이었다. '한국문학'의 하위 요목을 살펴보면, <표 34>와 같다. 주요 문학 갈래인 시, 희곡, 소설, 수필에서는 목차의 음절 편차가 크게 나타나지 않았으며, 연설과 웅변(815)에서 편차가 2457.9로 가장 컸다(<그림 56>, <그림 57>, <그림 58> 참조).

<표 32> 사회과학 강목별 목차 음절 통계 현황

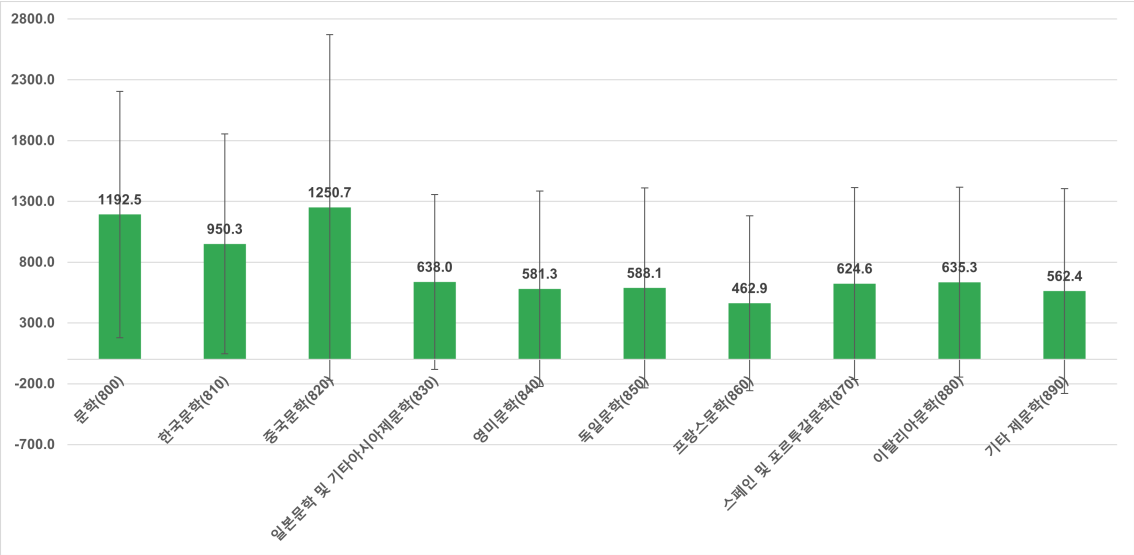
사회과학 강목	평균	편차	최소	최대
사회과학 (300)	1,314.0	1,210.0	33	27,234
통계자료 (310)	2,281.9	1,517.9	72	8,481
경제학 (320)	1,927.8	2,284.0	3	124,565
사회학, 사회문제 (330)	1,380.7	1,651.7	37	118,974
정치학 (340)	1,208.2	1,221.8	14	48,013
행정학 (350)	1,967.0	2,203.0	22	46,720
법률, 법학 (360)	3,054.3	4,990.0	3	210,893
교육학 (370)	1,427.4	1,282.2	19	27,635
풍습, 예절, 민속학 (380)	1,286.6	1,241.4	30	15,605
국방, 군사학 (390)	1,311.0	1,590.5	44	43,803



<그림 56> 사회과학 강목별 음절 평균과 표준편차

<표 33> 문학 강목별 목차 음절 통계 현황

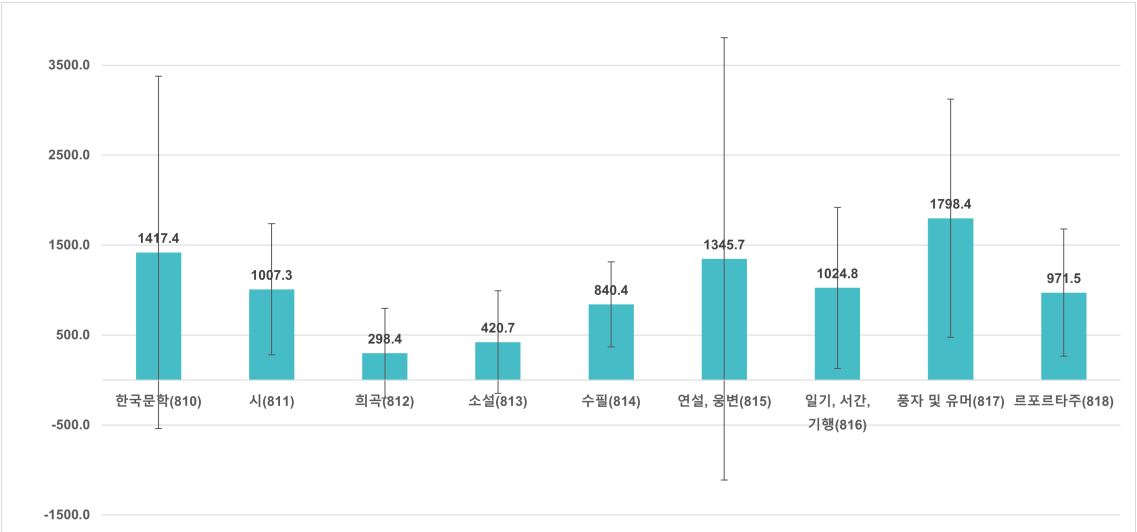
문학 강목	평균	편차	최소	최대
문학 (800)	1,192.5	1,013.5	41	18,986
한국문학 (810)	950.3	905.0	11	47,633
중국문학 (820)	1,250.7	1,421.5	22	12,735
일본문학 및 기타아시아제문학 (830)	638.0	719.3	18	6,847
영미문학 (840)	581.3	803.5	11	7,938
독일문학 (850)	588.1	822.5	11	7,221
프랑스문학 (860)	462.9	719.0	16	6,600
스페인 및 포르투갈문학 (870)	624.6	790.1	23	6,051
이탈리아문학 (880)	635.3	781.8	35	4,029
기타 제문학 (890)	562.4	843.8	10	7,067



<그림 57> 문학 강목별 음절 평균과 표준편차

<표 34> 한국문학의 요목별 목차 음절 통계 현황

한국문학 요목	평균	편차	최소	최대
한국문학 (810)	1,417.4	1,957.7	34	47,633
시 (811)	1,007.3	728.4	20	25,745
희곡 (812)	298.4	497.3	17	3,640
소설 (813)	420.7	570.7	13	10,566
수필 (814)	840.4	473.5	11	8,126
연설, 웅변 (815)	1,345.7	2,457.9	98	9,075
일기, 서간, 기행 (816)	1,024.8	895.2	18	9,637
풍자 및 유머 (817)	1,798.4	1,322.6	30	7,797
르포르타주 (818)	971.5	708.0	17	15,559



<그림 58> 한국문학 요목별 음절 평균과 표준편차

5. 원문 데이터

입수한 원문 데이터는 총 65,825건으로 전체 서지데이터의 약 5.4%에 해당한다. 주제명 자동 분류를 위한 학습데이터로 원문을 활용하기 위하여 목차가 입력된 서지데이터에 해당하는 원문 데이터를 확인한 결과, 26,229건으로 급격히 줄어드는 모습을 보였다.

해당 데이터에서의 출현 횟수가 300건 이상의 주제명이 2개이며, 200건 이상의 주제명 2개, 100건 이상의 주제명 10개로 부여된 것으로 나타났다. 즉 100회 이상 사용된 주제명이 전체 14개에 불과해, 서명과 같은 다른 분류 자질과 비교해서 학습데이터로서 활용도가 떨어지는 것으로 판단하였다.

한편, 원문 데이터가 존재하는 서지데이터의 주제명 부여 현황을 파악한 결과는 <표 35>와 같다. 문학류 관련 주제명 이외 '유아 교육[幼兒敎育]', '관악곡[管樂曲]', '악보(음악)[樂譜]', '전시 도록[展示圖錄]' 등의 주제명이 상위권에 위치하고 있다. 이러한 결과는 상대적으로 사회과학이나 자연과학 자료는 그 양이 적거나, 원문 추출 및 수집이 수월하고, 저작권 문제가 없어 OCR(광학식 문자판독장치) 처리가 가능한 자료에 한해 원문 데이터를 수집했기 때문으로 보인다. 이러한 경우 서명과 목차와 같은 다른 분류 자질과의 주제명 자체의 차이와 통계적 분포의 차이로 인해 직접적인 성능을 비교하는 것은 다소 무리가 있다. 다만 원문 데이터가 자체적으로 자동 분류에 기여하는 부분이나 성능 등을 파악할 필요는 있었다. 따라서 이 연구에서는 별도로 원문 데이터 65,825건만을 대상으로 자동 분류 알고리즘을 적용해 목차나 저자명 등과의 조합 없이 원문 데이터 단독으로 성능을 평가하는 추가 실험을 진행하였고, 이후 서명 자질만 조합하여 성능을 확인하였다.

<표 35> 원문 데이터 자체 주제명 부여 횟수 순위

주제명	부여 횟수
한국 현대시[韓國現代詩]	2,517
한국 현대 소설[韓國現代小說	1,860
유아 교육[幼兒敎育]	733
기독교[基督敎]	573
악보(음악)[樂譜]	460
전시 도록[展示圖錄]	444
일본 현대 소설[日本現代小說	411
한국 현대 수필[韓國現代隨筆	391
창작 그림책[創作--冊]	390
한국 현대 문학[韓國現代文學	369
관악곡[管樂曲]	365
임상 실습[臨床實習]	361
한국 문학[韓國文學]	357
영어 학습[英語學習]	333
애정 소설[愛情小說]	315

6. 요약 및 시사점

6.1 주제명데이터 측면

국립중앙도서관 주제명표목표의 전체 주제명 중 실제 사용되는 주제명은 57,320건, 이 중에서도 100회 이상 사용된 경우는 3,506건으로 구축된 전체 주제명에 비해 상대적으로 주로 사용하는 주제명만을 자주 사용하는 행태를 확인할 수 있다.

관계지시기호 측면에서 최빈도 RT, BT, NT 등의 계층구조와 연관성에 따른 지시기호 사용이 가장 높았으며, 이를 중심으로 개별 주제명 간의 계층관계를 파악하였다. 주제명표목표에서 구축된 최상위어 기준 최하위 계층은 심도 17이었으며, 주로 주제어 형성은 심도 1~2단계, 활용 주제명들도 역시 심도 0~2단계 사이의 주제어로 구성된 점을 확인하였다. 이는 넓은 의미의 개념이나 주제

에 대항하는 주제명을 서지에 부여하고 있음을 나타낸다. 따라서 다양한 주제어의 사용이 가능해지려면 주제어 간의 연관관계를 재정립할 필요가 있다.

한편, 주제명 부여 개수는 평균 1.72개로 업무지침 3개 이내 부여라는 권고사항에 부합하였으며, 다만 대상 자료나 콘텐츠의 주제가 특정적일수록 그 자료의 주제어 부여 개수가 증가하고, 심도 역시 깊어지는 경향을 보였다.

주제영역에 따른 주제어 부여 개수를 보면, 예술이 2.09개로 가장 많았고 문학이 1.43개로 가장 적었는데, 1.72개가 평균이라는 점을 감안할 때 주제에 따라 주제어 부여 개수의 편차가 다소 존재하는 것으로 나타났다.

6.2 서지 및 목차 데이터 측면

주제명이 부여된 전체 서지데이터 중 KDC 주류 측면에서 가장 비중이 높은 주제는 사회과학 분야로 전체에서 약 29%를 차지하였으며, 이어 기술과학이 18%, 문학이 17%인 것으로 나타났다. 하위 비중의 주류는 자연과학, 총류, 철학, 언어 등의 분야이며, 대체로 주제별로 3~4%의 비중을 차지하고 있었다.

주제명이 부여된 전체 서지데이터 중 대략 38.9%가 목차를 포함하고 있는데, KDC 주류 측면에서 보유 서지데이터에 비하여 목차 기입의 양이 다소 높은 주류는 총류, 종교, 사회과학 등으로 나타났다. 데이터 보유 현황에 따른 비율에서 서지데이터 전반의 양에 비하여 목차 데이터는 약 1/3 수준으로, 절반에도 미치지 못하므로 목차 데이터의 추가 수집의 필요성을 고려해야 하며, 주제명 자동분류 측면에서 보면 이는 학습데이터의 다양한 자질 중 하나로 사용되는 목차 데이터를 보완하여 자동 분류의 성능을 높이면 좀 더 적절한 주제명을 부여할 수 있다.

Ⅴ. 자동 분류 알고리즘 설계 및 검증

1. 알고리즘 설계

인공지능의 딥러닝 기법은 비전(vision) 분야에서 획기적인 성과를 가져오며 연구뿐만 아니라 실무 영역에서까지 화두가 되고 있다. 대표적으로 이미지 처리에 탁월한 성능을 보이는 합성곱 신경망(CNN) 방법을 필두로 하여 새롭고 혁신적인 다른 모형들이 등장하였다. 이 방법은 또한 분야를 달리하여 텍스트 처리에서도 다양한 시도를 통해 상당한 성과를 가져왔다.

텍스트나 자연언어 처리에서는 비전 영역의 이미지와 달리 출현 단어의 순서 또는 순차(sequential)적인 특성을 반영하는 딥러닝 모형이 필요하다. 이러한 배경에서 출현한 모형으로 순환 신경망(RNN)이 있으며 이를 개선한 다양한 모형으로 LSTM과 GRU 등이 등장하였다. 또한 텍스트 처리 분야에서 딥러닝의 지속적인 변화 발전에 따라 트랜스포머(transformer) 모형이 등장하였다.

트랜스포머 모형은 2017년 구글이 “Attention is all you need” 논문(Vaswani et al., 2017)에서 발표한 모형으로 시퀀스-투-시퀀스(sequence-to-sequence) 모형에 해당하며 입력 시퀀스 데이터를 압축 변환하여 출력 시퀀스로 보내 이를 통해 최종 결과물을 생성하는 구조를 갖는다. 이때 전자의 압축 변환하는 과정을 인코딩(encoding)이라고 하며, 후자의 과정을 디코딩(decoding)이라고 한다. 따라서 트랜스포머 모형은 크게 앞부분의 인코더(encoder)와 뒷부분의 디코더(decoder)로 구성된다.

이후 2018년에 구글은 트랜스포머 모형의 인코더를 활용한 언어 모형인 BERT(Bidirectional Encoder Representations from Transformers)를 발표하였다. BERT는 위키피디아와 도서의 많은 텍스트를 레이블 되지 않은 상태로 이용하여 사전 학습(pretrain)한 모형으로 다양한 자연언어 처리 과업 영역에서 미세조정(fine-tuned)을 통해 최고의 성능을 가져오는 것으로 알려졌다.

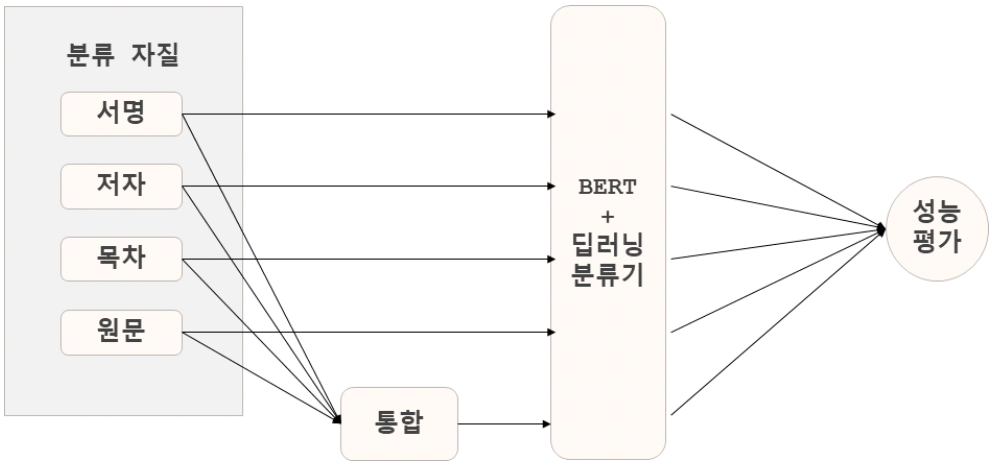
구글은 원래 영어판으로 인코더 레이어와 self-attention 헤드에 따라 두 종류의 모형을 공개하였다. 이후 70여 개의 언어의 텍스트를 이용하여 학습한 다국어 버전의 BERT를 추가하였는데, 이 안에는 한국어도 포함한다. 최근 딥러닝 기술의 발전으로 T5(Text-to-Text Transfer Transformer)나 BIGBIRD,

Longformer와 같이 보다 개선된 모형들이 등장하고 있다.

BERT는 입력된 문장이나 텍스트에 대해 단어 임베딩(word embedding) 방법을 제공하는 마스크 언어 모형인데, 여기서 임베딩은 자연언어 텍스트에 나타난 특정 단어를 밀집된 벡터화로 표현하여 그 단어가 가지는 특징이나 문맥적 의미를 추출하거나 표현하는 방법을 말한다. 가장 대표적인 예로 Word2Vec나 Doc2Vec가 해당한다. 이들 방법이 각 단어에 대해 하나의 벡터 표현이 생성하는 문맥 독립적이라면, BERT에서 사용한 단어 임베딩은 같은 단어에 대해 그 단어가 사용된 문맥에 따라 서로 다른 벡터 표현을 가능하게 하여 문맥에 의존적인 특징을 갖는다. 또한 임베딩의 수준을 단어 차원이나 문장(문헌) 차원으로 나누어 볼 수 있는데, BERT는 문장 수준의 임베딩까지 가능하기에 여러 과업에서 뛰어난 성능을 보여 널리 쓰이고 있다.

본 연구는 주제명표목표의 자동 분류를 위하여 딥러닝 사전학습 모형인 구글의 다국어 BERT를 적용하였다. 한국어 텍스트를 사전 학습한 BERT 유형의 모형들도 다수 개발되어 있는데, 이들은 대체로 다국어 BERT보다 더 좋은 성능을 보인다. 다만 구글의 다국어 BERT는 다양한 딥러닝 패키지에서 사용할 수 있을 뿐 아니라 BERT 유형 모형들의 비교 대상, 즉 베이스라인 모형이 되기에 이 연구에서는 기준 모형인 구글의 다국어 BERT를 이용하여 실험하였다.

이 연구의 국가서지 기반 자동 분류 실험을 위한 전체적인 개요는 <그림 59>와 같다. 주제명 자동 분류를 위한 딥러닝 실험은 허깅 페이스(Hugging Face)사의 Tranformers 패키지를 사용하였으며, 각각의 데이터셋에 대해 30 에포크(epoch), 배치 사이즈(batch size) 8과 학습률(learning rate) 2e-5를 적용하였다.



<그림 59> 자동 분류 실험 개요

전체적인 실험과정은 먼저 구글의 다국어 BERT를 이용하여 국가서지에서 추출된 텍스트 데이터에 대해 문장 수준 임베딩(sentence embedding)을 처리하고, 이를 딥러닝 분류기를 통해 주제명을 부여한다. 여기서 딥러닝 분류기는 BERT의 임베딩 결과물 벡터인 pooler_output에 대해 선형 변환을 적용하는 레이어에 해당한다.

분류 실험에 사용된 데이터는 국립중앙도서관에서 생성한 국가서지 데이터로, 추출 가능한 주요 메타데이터와 자연언어 형식의 텍스트를 대상으로 하였다. 구체적인 분류 자질로 서명, 저자명, 목차, 도서의 원문(full-text)을 선정하였다.

국립중앙도서관이 주제명을 부여한 서지 레코드를 대상으로 분류 실험을 하는데 이 데이터에서 분류 자질의 출현 또는 제공 현황을 보면, 서명과 저자명은 대부분의 레코드에 출현하며, 목차는 약 39%, 원문은 대략 5% 정도의 레코드에 포함된다.

실험에 사용된 4가지 자질에 대해 주제명의 자동 분류 성능은 단일 자질과 복합 자질로 나누어 평가하였다. 단일 자질은 4가지 데이터 중에 서명, 목차, 원문 데이터만을 사용하여 분류하고 그 성능을 측정하였다. 복합 자질은 서지 레코드에 가장 많이 나타나는 자질인 서명을 중심으로 저자명을 통합한 방식, 목차의 분류 성능을 파악하기 위해 목차 중심으로 서명과 저자명을 통합한 방식, 그리고 원문의 특성을 파악하기 위해 원문을 중심으로 통합한 방식으로 종류를 나누어 분류 자질을 생성하고 그에 따른 성능을 측정하였다. 다만 분류 자질로써 서명과 목차, 원문의 성능을 비교하기 위해 이들 단일 자질을 복합 자질의 성능과 비교하였다.

자동 분류 모형의 성능은 다양한 방법으로 평가하는데 해당 모형이 수행하는 과업이 분류인지 회귀인지에 따라 성능 평가 척도는 여러 종류로 나누어진다. 회귀 과업의 경우 RMSE, MAE 등에 기반하여 평가하며, 분류 과업은 정확도(Accuracy), 재현율(Precision), 정확률(Recall), F1 척도, ROC(Receiver Operating Characteristic), AUC(Area Under the ROC Curve)를 활용하여 분류 모형의 성능을 평가한다.

정확도는 실제 데이터에서 예측 데이터가 얼마나 같은지를 판단하는 척도로 직관적으로 모델 예측 성능을 나타내는 척도지만 데이터 구성에 따라 왜곡될 수 있어 정확률, 재현율, F1 점수 등을 같이 사용한다. 정확률은 분류기가 한 문헌에 대해 그 문헌에 부여된 옳은(True) 범주로 예측했을 때 그 비율을 계산한 값으로, Positive 예측 성능을 더욱 정밀하게 측정하기 위한 평가로 양성 예측도라

고도 불린다. 재현율은 실제 값이 참인 대상 중 예측과 실제 값이 참으로 일치한 데이터의 비율을 의미한다.

일반적으로 재현율과 정확률은 상호 보완적인 척도여서 이 중 하나만으로 분류 성능을 완전하게 평가하기 어려워 정확률과 재현율의 조화 평균값을 의미하는 F1 척도를 많이 사용하며 이 연구에서도 이 척도를 중심으로 분류 성능을 평가하였다. F1 척도의 경우 최종 성능을 판단하기 위해 그 관점을 분류 범주 중심으로 보느냐 분류 문헌 중심으로 보느냐에 따라 마이크로 평균 F1(micro-average F1: microF1) 척도나 매크로 평균 F1(macro-average F1: macroF1) 척도의 적용이 가능하다.

자동 분류 모형의 평가는 분류 성능을 통해 서로 다른 모형이나 사용된 자질을 비교하는 데 초점을 두고 있어 실제 올바르게 분류한 범주나 주제명을 보여주거나 제시하지 않는 경향이 있다. 이 연구는 실제 주제명과 자동 분류 모형이 올바르게 분류한 주제명을 비교하여 추가적인 분석을 하였다. 즉 이 연구에서 수행하는 분류 대상이 문헌정보학 분야나 도서관 영역에 해당하므로 분류 일치성 분석을 KDC(또는 DDC) 분류체계의 주제 분야에 따른 분석, 주제명의 유형과 범주에 따른 분석 그리고 학습 자질의 종류에 따른 분석을 수행하였다.

KDC 분류체계에 따른 분석은 분석 단위를 KDC의 주류 및 강목 수준의 일치 정도를 분석함으로써 이를 통해 주제 분야별 자동 분류의 용이함과 성능 차이를 파악하고자 하였다. 주제명의 유형과 범주에 따른 분석은 주제명의 유형별, 범주별 종류에 따른 분류 일치도 분석을 하여 해당 주제명이 자동 분류의 용이함과 난해함 정도를 파악하고자 하였다. 학습 자질의 종류에 따른 분석은 서지 레코드로부터 추출된 주요 자질(서명, 저자, 목차, 원문)에 따라 일치 정도의 분류 성능 파악하였다.

기계학습에 의해 분류된 주제명이 사람에 의해 수작업으로 부여한 정답과 다른 경우 그 결과(비일치)를 분석하여 그 원인이나 이유를 밝히는 오분류 분석이 필요하다. 이는 자동 분류에 실패한 오류 결과를 주제명 부여 실무 전문가가 평가하여 기계학습에 의한 분류 모형이 가지는 한계점과 개선 방향을 도출하고, 이를 자동 분류 모형에 반영하여 성능을 높이기 위한 최적화 방안을 마련하기 위한 실마리를 제공할 수 있다. 또한 실무 전문가가 부여한 결과를 기준으로 기존의 부여된 결과와 자동 분류에 의해 부여된 결과를 주제명표목 업무지침에 따라 비교 분석하여 현재 업무 상황에 대한 시사점을 제시하고자 하였다.

2. 데이터 전처리

이 연구의 실험 데이터는 국립중앙도서관이 수서하여 소장한 장서에 대해 분류와 편목 업무를 통해 생산된 국가서지 데이터의 일부이다. 이 데이터는 서지 레코드를 기반으로 하기에 도서의 서명과 저자명이 기본적으로 포함되어 있으며, 도서관 이용자에게 더욱 풍부한 정보를 제공하기 위해 일차적으로 목차와 필요에 의해 원문을 추가적으로 구축하고 있다.

서명을 비롯하여 서지데이터에서 주요한 분류 자질들은 자연언어 형식의 텍스트 유형에 해당한다. 특히 서명이나 목차의 경우 비교적 짧게 요약된 또는 압축된 형식의 자연언어에 해당하며, 반대로 원문의 경우 긴 텍스트 형식에 해당한다. 일반적으로 이러한 유형의 한국어 텍스트에 대해 대표적으로 형태소 분석과 같은 자연 언어 처리를 적용한다. 딥러닝을 이용한 자동 분류에서도 어떤 모형을 선택하느냐에 따라 그 모형이 필요로 하는 자연 언어 처리 방법을 적용해야 한다.

자연언어로 된 문장이나 텍스트를 기계가 처리하기 위해서는 토큰화가 필요하다. 텍스트를 구성하는 요소를 분석에 필요한 정도의 작은 단위로 쪼개고 이를 애플리케이션이나 과업에 맞게 컴퓨터가 처리해야 한다. 한국어의 문장은 구, 어절, 단어, 형태소 순으로 더 작게 쪼갤 수 있다. 텍스트를 이렇게 필요한 수준의 작은 단위(token)로 나누는 작업을 토큰화(tokenization)라고 하며 적용 영역이나 애플리케이션에 따라 다양한 방법이 적용된다. 이를 수행하는 것을 토크나이저라고 하는데, 자동 분류를 위한 기계학습에서는 학습과 예측에서 동일한 토큰화 또는 토크나이저를 사용한다.

BERT 모형은 토크나이저로 워드피스(wordpiece) 방법을 사용한다. 워드피스는 바이트 쌍 인코딩(byte pair encoding: BPE) 알고리즘에 기초하는데, 이 BPE는 데이터에서 자주 등장하는 문자열 쌍을 데이터 내에 등장하지 않는 바이트로 대체하여 데이터를 압축한다. BPE 기법을 이용한 토큰화는 텍스트나 말뭉치에서 자주 나타나는 문자열을 서브워드(subword) 또는 부분 문자열로 대체하고 분리하여 어휘집합으로 구축한다. 따라서 워드피스의 토큰화 기법은 분석 대상 텍스트에 쓰인 언어에 대한 지식을 요구하지 않는다. 즉 한국어나 영어에 대해 그 언어가 필요로 하는 사전 처리 등을 하지 않아도 된다.

분류 자질로 서명, 저자명, 목차, 원문이 있는데 이 중에서 저자명은 국립중앙도서관에서 제공하는 전거데이터 중에서 개인명과 단체명에 대한 식별된 제어번

호를 이용하여 인코딩하였다. 이는 동명이인이나 동일인에 대한 다양한 이명(異名)으로 인해 발생하는 문제를 해결할 수 있다. 나머지 서명, 목차, 원문 텍스트에 대하여 특수 기호를 제거하는 등 비교적 간단한 전처리를 수행하였다.

기계학습 방법 중의 지도학습에 의한 자동 분류 실험의 데이터셋은 학습 세트(train set), 검증 세트(validation set), 테스트 세트(test set)로 구성된다. 일반적으로 이들의 비율은 학습 세트와 테스트 세트를 80:20 비율로 나누며, 학습하는 과정에서 분류기의 성능을 검증을 위해서 전체의 80%에 해당하는 학습 세트를 다시 80:20 비율로 나누어 20%에 해당하는 데이터를 검증 세트를 구성한다. 이에 본 연구에서도 같은 비율로 각각의 데이터셋을 구성하였다.

자동 분류 실험을 데이터 측면에서 구체적으로 보면 실험 대상 서지 레코드의 전체 양이 약 122만여 건에 해당하며, 이 중 목차를 포함하는 수는 39%로 약 47만여 건에 해당한다. 각 자질 측면에서 보면 전체 122만여 건은 대부분 서명을 포함하고 있지만 목차는 그렇지 않기 때문에 최대한의 데이터를 활용하여 각 자질의 자동 분류 성능을 계산하기 위해서 서명, 목차, 원문의 데이터셋을 별도로 구성하였다. 즉 서명의 분류 성능을 보기 위해서는 전체 데이터를 활용하였으나 이 중 목차를 포함하는 레코드가 1/3밖에 안 되므로 직접적으로 서명과 목차의 분류 성능을 비교할 수 없다. 따라서 서명과 목차의 직접적인 분류 성능 비교를 위해서는 목차를 가진 레코드를 대상으로 서명 자질의 분류 성능을 계산하고 이를 목차의 성능과 비교하고자 하였다. 원문의 경우 원문이 포함된 데이터에 대해 서명과의 성능을 비교하였다.

또한 주제명은 각각에 대해 균등하게 부여되지 않는다. 어떤 주제명은 흔하게 사용되며 다른 주제명은 소수의 서지에 부여되거나 심지어 전혀 부여되지 않는 사례도 있다. 이는 지식 세계에서는 자연스러운 모습이지만, 자동 분류의 데이터셋 측면에서는 분류하고자 하는 범주의 분포가 균등한 것이 유리하다. 그렇다고 주제명 간의 분포의 차이를 인위적으로 가공할 수는 없고 실제 데이터의 모습을 그대로 적용하는 것이 바람직하다.

앞서 기술한 부분과 함께 주제명의 분포를 그대로 가져와 기계학습용 실험 데이터를 구축하는 데 추가로 고려해야 하는 점은 서지데이터에 거의 부여되지 않은 주제명의 경우 학습이 어렵거나 불가능하다는 점이다. 예를 들어 전체 국가서지 데이터에서 한번 내지 두 번 부여된 주제명은 학습 세트, 검증 세트, 테스트 세트의 3개 부분으로 나누어 실험 데이터를 구축하는 경우 학습이 되지 않거나 검증이 되지 않는 상황이 발생하게 된다. 따라서 이러한 어려움을 해결하기

위해 일정 횟수 이상 부여된 주제명을 대상으로 자동 분류 실험을 적용하였다. 실제 서명과 목차의 데이터 규모가 다르기에 각 자질에 대해 실험 데이터를 생성하기 위해 적용한 최소 주제명의 부여 횟수도 달리하였다.

3. 분류 성능 측정 및 평가

이 연구에서는 주제명이 부여된 전체 서지 레코드를 실험 데이터로 사용하였다. 샘플을 일부 추출하여 데이터로 사용하기보다 전체 데이터를 학습 및 테스트 데이터셋으로 나누어 사용하였다. 서명의 경우 대부분 짧은 문장이므로 자동 분류에서 어려움이 있을 수 있으므로 보다 많은 사례를 학습할 수 있도록 환경을 제공하였다.

국가서지의 주제명 자동 부여를 위한 기계학습에서 서지데이터의 분류 자질에 따른 특성을 반영하여 성능 측정과 평가를 수행하였다. 국립중앙도서관의 서지 레코드는 서명, 저자명이 목록 규칙 등에 의해 기술되어 있으며, 추가로 일부 자료에 대해 목차와 원문을 구축하고 있다. 즉 대부분의 레코드는 서명과 저자명이 있지만, 목차와 원문을 그 대상이 일부 서지 레코드에 한정되어 존재한다. 이 연구에서도 이를 분리하여 저자명을 포함한 서명 중심의 분류 실험과 목차 중심의 분류 실험을 진행하였다. 추가적으로 간략하게 원문에 대한 자동 분류 실험도 포함하였다.

3.1 서명 중심 분류

서명과 저자명을 분류 자질로 사용한 자동 분류 성능을 알아보기 국가서지 데이터 전체를 대상으로 BERT 기반의 딥러닝 기계학습을 적용하였다.

기계학습은 레이블 된 사례가 너무 적으면 학습이 불가능하여 일정 횟수 이상 출현한 학습데이터를 사용해야 한다. 그 기준은 분야나 데이터의 품질 등에 따라 다르므로 획일해서 한 가지로 설정할 수 없고 주제명의 부여는 장서의 주제에 따라 다르므로 균형적이지 않아 주제명의 부여 횟수에 따라 그 기준을 다양하게 설정하고 학습 데이터로 추출하였다. 국가서지 데이터와 주제명 부여 데이터의 통계를 나타내는 <표 2>와 <표 8>을 참조하여 서명 중심의 분류를 위한

데이터셋을 추출한 현황은 <표 36>과 같다.

<표 36> 서명 중심 데이터 현황

데이터 현황				저자 전거 현황	
출현 횟수	주제명 수	서지 레코드수	비율	출현 횟수 기준	총 저자 수
5,000	25	287,253	13.70%	4	8,903
3,000	46	372,971	17.80%	5	8,486
1,000	254	709,406	33.80%	7	9,913
500	603	947,581	45.20%	9	9,225
300	1,106	1,139,906	54.30%	10	9,809
100	3,506	1,539,076	73.40%	13	8,636

주제명 부여 횟수에 해당하는 출현 횟수를 기준으로 5,000회 이상 부여된 주제명은 25개이며, 100회 이상 출현한 횟수를 기준으로 3,506개의 주제명을 분류 범주로 사용하였다. 이들이 부여된 서지데이터의 비율을 보면 5,000회 이상의 25개 주제명이 부여된 비율이 전체 실험 대상 데이터인 122만여 건의 13.7%에 해당하며, 3,506개의 주제명의 경우 대상 서지데이터의 73.4%에 해당한다.

이 표를 주제명 자동 부여의 실무 측면에서 보면, 출현 횟수 상위 100회 이상의 3,506개 주제명을 학습하면 현재 국가서지 전체에 부여된 주제명의 73.4%를 부여할 수 있다는 의미이며 앞으로 추가된 서지데이터의 73.4%에 대해 3,506개 주제명을 추천할 수 있다는 의미로 볼 수 있다.

도서나 문헌의 저작에 있어 저자가 예외적으로 없는 경우도 있지만, 대부분 한 명 또는 다수가 참여한다. 또한 동일한 단어가 여러 문헌에 자주 출현하는 것과 달리 동일한 저자는 수천 건의 문헌에서 소수의 자료를 집필한다. 이를 분류 자질 측면에서 보면 저자 정보는 매우 희소한(sparse) 자질에 해당하며 이러한 자질을 행렬로 처리하려면 고사양의 컴퓨팅 자원을 요구한다. 데이터셋에 출현한 모든 저자를 처리하는 것이 일반적인 상황에서는 불필요하거나 쉽지 않으므로 최소 출현 횟수를 기준으로 제한하였다. 이때 각각의 데이터셋에서 총 저자 수가 10,000명이 넘지 않는 범위에서 저자의 최소 출현 횟수를 설정하였다.

출현 횟수 5,000회 이상의 상위 25개의 주제명에 대한 데이터셋은 최소 저자 출현 횟수로 4회를 기준으로 설정하였으며 이때 고유한 총 저자 수는 8,903명이

되며, 이들 저자를 저자 전거 제어번호를 사용하여 인코딩하고 분류 자질로 사용하였다. 주제명의 출현 횟수 기준을 낮게 설정하면 분류 대상 주제명의 수는 커지며 해당하는 서지 레코드가 증가하므로 데이터셋의 크기도 커진다. 출현 횟수 100에 해당하는 데이터셋의 경우 주제명 수는 3,506개이며 서지 레코드는 153만여 건이 되는데, 이 세트에 포함되는 고유 저자의 수도 따라서 증가하기에 10,000명 이하에 해당하는 최소 저자 출현 횟수도 13으로 증가함을 알 수 있다.

출현 횟수를 기준으로 주제명 수에 따른 각각의 데이터셋에서 서명만 이용한 단일 분류 자질과 저자명을 포함한 복합 분류 자질에 대한 분류 성능을 보면, <표 37>과 같다.

<표 37> 서명 중심 자질의 분류 성능(microF1 기준)

세트명	주제명 수	분류 자질	
		서명	서명+저자
Title25	25	0.8076	0.8149
Title46	46	0.7637	0.7730
Title254	254	0.6976	0.6990
Title603	603	0.6566	0.6607
Title1106	1,106	0.6367	0.6386
Title3506	3,506	0.5434	0.5504

서명과 저자명의 분류 성능을 보면, 고빈도에 해당하는 소수의 주제명에서 자동 분류의 성능이 매우 높고, 저빈도 다수의 주제명으로 갈수록 성능이 차츰 떨어지는 것을 알 수 있다. 이는 소수의 범주로 분류할 때 성능이 높은 자동 분류의 일반적인 경향과 일치하는 것으로 판단되며, 고빈도로 갈수록 학습에 사용되는 데이터가 증가하는 출현 횟수를 기준으로 설정한 이 연구 데이터셋의 특성에서 기인한다고도 볼 수 있다. 즉 세트명 Title25는 5,000번 이상 출현하여 학습할 수 있는 데이터가 그만큼 큰 데 반해, 세트명 Title3506은 100회 출현하여 학습할 수 있는 데이터가 그만큼 작아 전체적인 분류 성능에서 차이를 보이고 있음을 알 수 있다.

분류 성능은 서명의 단일 자질에서는 주제명 상위 25개의 데이터셋(세트명

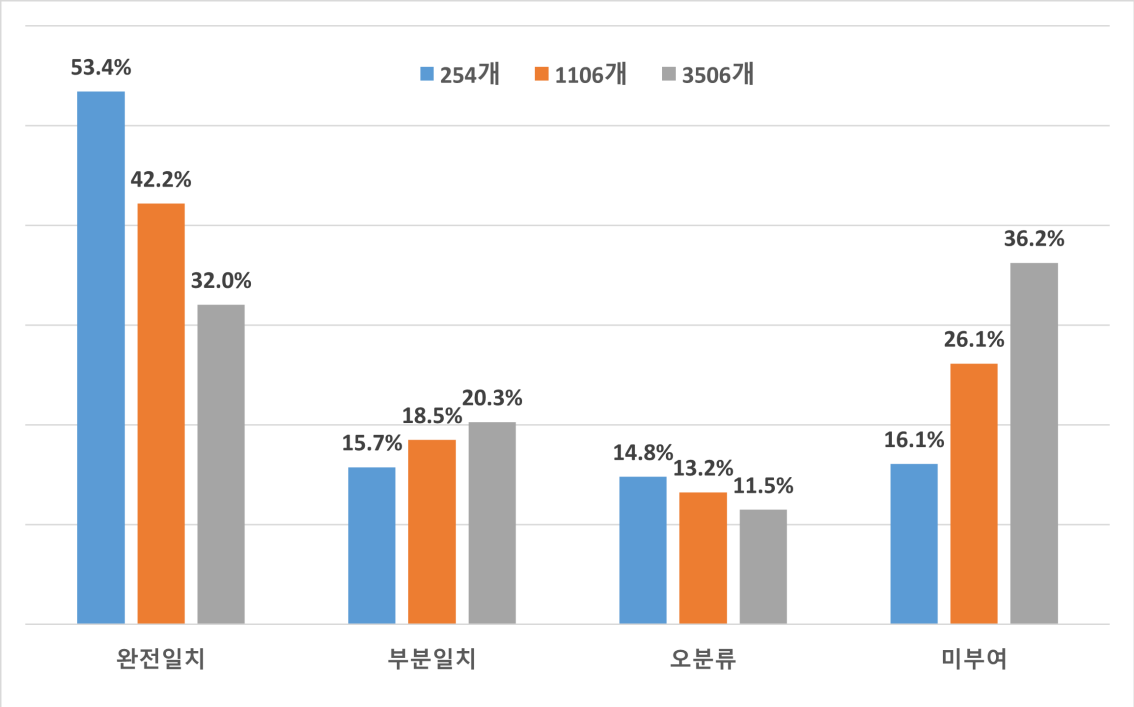
Title25)에서 0.8076으로 높게 나타났으며, 상위 3,506개(세트명 Title3506)에서는 0.5434로 나타났다. 서명과 저자가 결합된 복합 자질에서도 상위 25개의 주제명에 대한 데이터셋에서 0.8149로 가장 높게 나타났으며, 상위 3,506개에서는 0.5504로 나타났다. 단일 자질과 복합 자질의 성능을 비교하면 대체로 서명과 저자가 결합된 복합 자질이 근소하게 높은 성능을 보여 저자명 자질을 추가하는 것이 필요함을 알 수 있다. 다만 이 연구에서는 모든 저자를 사용하지 않았기에 저자 수를 늘려서 학습한다면 추가적인 성능 향상을 기대할 수 있다.

주제명 254개의 데이터셋(세트명 Title254)에 대해 학습의 횟수인 에포크(epoch)에 따른 성능을 다양한 평가 척도로 나타내면, <표 38>과 같다. 에포크 초반에는 정확률이 높지만, 더 많은 학습이 이루어질수록 낮아지며 재현율은 반대로 낮은 값에서 점차 높아지는 것을 알 수 있다. 정확도의 경우 주제명의 분포가 불균형적이고 부여해야 할 범주가 많아 분자 분모에 들어가는 true negative 값이 큰 값을 가지므로 매 에포크마다 차이가 없는 것을 알 수 있다. 대체로 실험 데이터가 규모가 크고 복합 자질을 사용할수록 많은 수의 에포크를 수행해야 최고 성능을 가져오며 작은 규모의 서명 데이터셋은 약 5회 정도의 에포크에서 가장 높은 성능을 보여주었다.

「국립중앙도서관 주제명표목 업무지침(2021)」에 따르면, 하나의 서지 레코드에 최소 1개에서 3개까지 주제명을 부여하도록 원칙으로 규정하고 있다. 이 연구에 사용된 데이터의 실제 현황인 <표 15>를 보면 주제명이 서지데이터에서 최소 1개부터 최대 12개까지 부여된 것을 알 수 있다. 하나의 서지 레코드에 다양한 수의 주제명이 부여되기에 기계에 의해 자동으로 부여된 주제명과 어떤 차이가 나는지 파악하기 위해 동일 서지 레코드에 기계와 인간이 부여한 주제명의 일치 정도를 파악해 볼 필요가 있다. 예를 들어 주제명 2개가 부여된 서지 레코드에 기계가 2개 모두를 정확하게 부여하거나 하나 또는 2개 이상, 더 나아가 하나도 부여하지 못할 수도 있다. 서명과 저자명을 이용한 분류에서 서지 레코드를 기준으로 자동 분류 결과의 일치 정도에 따른 비율을 그림으로 나타내면, <그림 60>과 같다. 그림에서 완전일치는 기계가 정답과 동일하게 주제명을 부여한 경우를 뜻한다. 부분일치는 정답과 기계가 부여한 주제명이 서로 부분집합인 경우를 말하며 이때는 기계가 더 적게 또는 더 많이 주제명을 부여한 경우이다. 오분류는 기계가 주제명을 자동 부여했지만, 정답과 하나도 일치하지 않는 경우를 뜻하며, 미부여는 기계가 주제명을 전혀 부여하지 않은 경우를 뜻한다. 다만 여기서 일치정도는 단순 일치 비율로 마이크로 평균 척도와 같은 평가지표와는 그 의미가 다르다.

<표 38> 에포크에 따른 주제명 254개 데이터셋의 성능

Epoch	Training Loss	Validation Loss	F1	Precision	Recall	Roc Auc	Accuracy
1	0.0120	0.0115	0.6126	0.8341	0.4841	0.7418	0.9968
2	0.0106	0.0104	0.6593	0.8164	0.5528	0.7761	0.9970
3	0.0093	0.0099	0.6695	0.8236	0.5639	0.7817	0.9971
4	0.0090	0.0096	0.6827	0.8129	0.5885	0.7939	0.9972
5	0.0083	0.0096	0.6850	0.8198	0.5883	0.7938	0.9972
6	0.0077	0.0096	0.6893	0.8005	0.6052	0.8022	0.9972
7	0.0073	0.0097	0.6932	0.7920	0.6164	0.8078	0.9972
8	0.0068	0.0099	0.6895	0.7857	0.6143	0.8067	0.9971
9	0.0061	0.0102	0.6928	0.7790	0.6237	0.8114	0.9971
10	0.0052	0.0104	0.6913	0.7618	0.6327	0.8158	0.9971
11	0.0052	0.0108	0.6888	0.7520	0.6354	0.8171	0.9970
12	0.0046	0.0111	0.6879	0.7507	0.6349	0.8169	0.9970
13	0.0042	0.0115	0.6847	0.7370	0.6392	0.8190	0.9970
14	0.0039	0.0119	0.6842	0.7335	0.6411	0.8200	0.9969
15	0.0033	0.0123	0.6831	0.7246	0.6460	0.8224	0.9969
16	0.0032	0.0127	0.6840	0.7253	0.6471	0.8229	0.9969
17	0.0028	0.0132	0.6852	0.7249	0.6497	0.8242	0.9969
18	0.0024	0.0136	0.6793	0.7175	0.6450	0.8218	0.9969
19	0.0022	0.0139	0.6801	0.7140	0.6492	0.8239	0.9968
20	0.0020	0.0143	0.6794	0.7106	0.6509	0.8247	0.9968
21	0.0018	0.0146	0.6791	0.7165	0.6453	0.8220	0.9968
22	0.0016	0.0150	0.6787	0.7073	0.6523	0.8254	0.9968
23	0.0014	0.0153	0.6788	0.7096	0.6506	0.8246	0.9968
24	0.0012	0.0156	0.6812	0.7138	0.6514	0.8250	0.9968
25	0.0012	0.0159	0.6824	0.7138	0.6536	0.8261	0.9969
26	0.0010	0.0162	0.6826	0.7134	0.6543	0.8265	0.9969
27	0.0009	0.0164	0.6828	0.7104	0.6573	0.8280	0.9968
28	0.0008	0.0166	0.6845	0.7170	0.6548	0.8267	0.9969
29	0.0006	0.0167	0.6832	0.7144	0.6545	0.8266	0.9969
30	0.0006	0.0168	0.6841	0.7153	0.6555	0.8271	0.9969

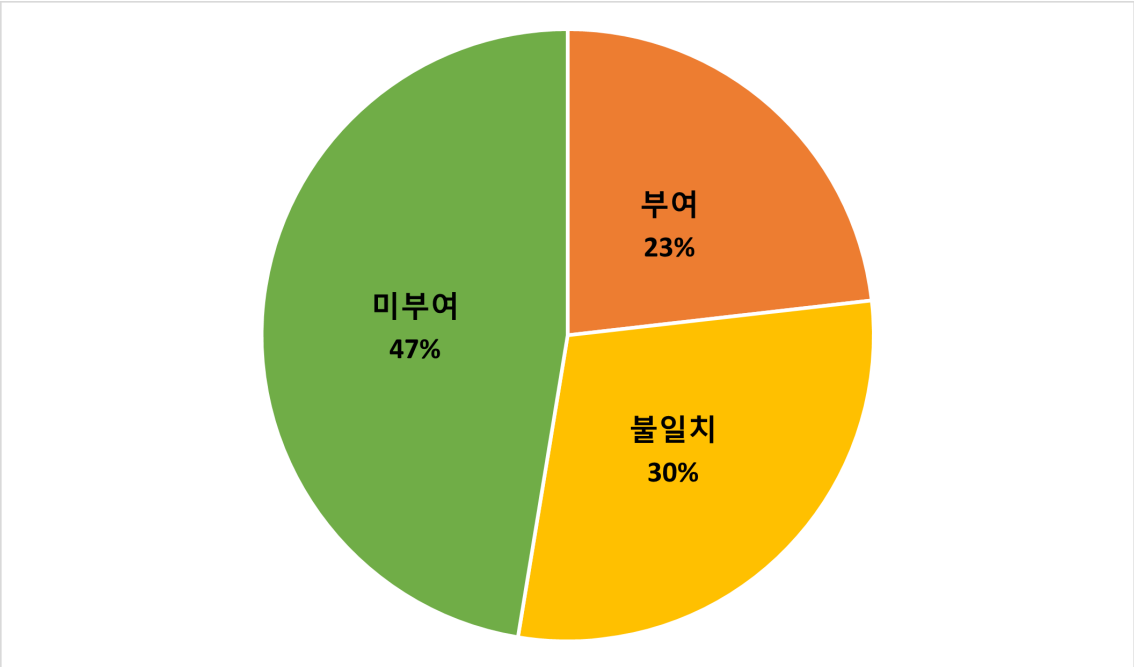


<그림 60> 서명 자질 데이터셋에 따른 자동 분류 일치정도

하나의 서지 레코드에 실무자가 부여한 주제명과 기계가 완전히 동일하게 부여한 경우인 완전 일치가 주제명 상위 254개 기준의 데이터셋에서 53.4%로 가장 높았으며, 1,106개의 데이터셋에서는 42.2%, 3,506개 세트에서는 32.0%로 나타나서 자동 분류 데이터의 주제명 수가 커질수록 낮아지는 것을 알 수 있다. 서지 레코드에 부여된 주제명 중 일부만 부여한 경우인 부분 일치는 각각의 데이터셋에 대해 15.7%, 18.5%, 20.3%로 나타나서 주제명 수가 커질수록 반대로 높아지는 것을 알 수 있다. 기계가 주제명을 부여하긴 했지만 서지 레코드의 정답인 주제명과 하나도 맞지 않는 경우인 불일치는 각각의 데이터셋에 대해 14.8%, 13.2%, 11.5%로 나타나서 주제명 수가 커질수록 낮아지는 것을 알 수 있다. 더 나아가 불일치긴 하지만, 기계가 아예 주제명을 부여하지 않은 미부여 불일치가 254개 세트에서 16.07%로 나타났고, 나머지 두 세트에서는 각각 26.1%와 36.2%로 나타났다. 출현 횟수가 낮은 1,106개의 데이터셋과 3,506개의 세트에서는 이러한 비율이 큰 폭으로 변하는 것을 알 수 있다.

주제명의 부여 분포가 장서에 따라 다르고, 되도록 많은 주제명을 자동 분류해야 하는 관점에서 보면 3,506개의 데이터셋에 대해 분류 정도를 좀 더 자세히 분석해 볼 필요가 있다. 특히 100회에서 200회 사이의 저빈도 주제명을 제대로

학습할 수 없는 상황이 될 수 있어 이들 주제명의 오분류를 파악해야 한다. 왜냐하면 되도록 많은 주제명을 학습하여 자동 분류시스템 또는 추천 시스템이 분류할 수 있는 주제명의 수를 늘리는, 또 다른 의미로 자동 분류의 범위를 넓히는 것이 또 다른 중요한 목표가 될 수 있기 때문이다. 3,506개 데이터셋에서 100회에서 200회 사이의 저빈도 주제명의 개수는 1,771개로 대략 51%에 해당하므로 이들 주제명이 이 세트의 전체 성능을 크게 좌우하여 영향을 미치게 된다. 이들 저빈도 주제명의 분류 성능을 분석하면, <그림 61>과 같다. 앞서 설명하였듯이 기계의 자동 분류에서 정확히 일치하여 부여한 비율이 23%에 해당하며, 부여는 하였지만 잘못 분류한 불일치가 30%, 미부여는 47%로 가장 높게 나타났다. 이는 저빈도 주제명이 학습데이터의 부족으로 적절한 학습이 이루어지지 않음을 의미한다.



<그림 61> 저빈도(100~200회) 부여 주제명의 미부여 정도

주제명 출현 횟수에 의한 생성된 데이터셋에 대해 KDC 주류에 따른 분류 일치 정도 성능을 계산하면 <표 39>와 같다. 주제명 254개의 데이터셋(세트명 Title254)에서는 총류(0XX), 사회과학(3XX), 기술과학(5XX), 역사(9XX) 등이 좋은 성능을 보이며 종교(2XX), 예술(6XX), 문학(8XX)이 낮은 성능을 보였다. 특히 문학과 사회과학의 경우 테스트 데이터셋에서 각각 24.4%와 31.2%의 비율로 절반 이상이 두 분야에 해당함에도 불구하고 문학의 경우는 일치도가 상대적으

로 낮고 사회과학은 높다. 결국 문학의 경우 정답과 동일하게 부여되는 주제명 분류가 잘 이루어지지 않음을 알 수 있다.

<표 39> KDC 주류에 의한 서명 중심의 분류 일치 현황(%)

세트명	유형	0XX	1XX	2XX	3XX	4XX	5XX	6XX	7XX	8XX	9XX
Title 254	완전일치	74.31	42.27	38.02	63.59	57.53	69.86	35.39	58.81	47.57	63.83
	부분일치	4.64	4.96	11.77	13.85	5.64	11.15	44.89	18.98	10.53	10.70
	불일치	21.05	52.77	50.20	22.56	36.83	19.00	19.72	22.21	41.90	25.47
Title 1106	완전일치	51.25	29.61	24.11	47.38	47.10	55.51	28.75	50.06	39.16	37.33
	부분일치	19.20	7.41	15.00	19.73	13.18	14.70	36.86	21.86	11.84	18.99
	불일치	29.54	62.98	60.89	32.90	39.72	29.79	34.39	28.08	49.00	43.68
Title 3506	완전일치	36.87	19.61	17.55	33.71	34.22	35.95	22.47	42.09	36.91	24.67
	부분일치	21.63	10.17	15.07	22.28	14.68	18.34	34.83	24.71	12.17	24.02
	불일치	41.51	70.23	67.38	44.01	51.10	45.71	42.70	33.20	50.93	51.32

여기서 데이터셋 중심으로 살펴보면 3,506개의 데이터셋(세트명 Title3506)은 254개나 1,106개 세트에 비해 일치도가 낮아지고 특히 불일치가 40~50% 넘는다. 이 불일치는 앞서 제시한 것과 같이 학습데이터의 부족으로 분류기에 의해 미부여가 많이 발생하는 것에서 기인한다. 그리고 일치 여부를 중심으로 살펴보면 완전일치의 경우 주제명 출현 횟수가 높은, 즉 소수의 주제명이 포함되고 학습데이터가 풍부한 데이터셋에서 좋은 성능을 보인 총류(0XX), 사회과학(3XX), 기술과학(5XX) 등의 분야가 부분 일치나 불일치에서 다수의 주제명을 포함하는 데이터셋에서는 급격히 성능이 저하되는 것을 알 수 있다. 반면 문학(8XX)의 경우 상대적으로 그 성능이 47.57%에서 42.08%로 상대적으로 덜 저하되는 것으로 나타났다.

출현 횟수 기준으로 생성된 3개의 데이터셋에 대해 주제명의 범주 유형에 따른 일치도로 성능을 분석하면 <표 40>과 같다. 주제명 유형 중 “지명”의 속성을 포함한 주제명의 경우에서 서명 중심 주제명 수가 증가할수록 불일치도가 높아진다. 특히, “지명”의 경우 254개 데이터셋에서 3,506개 데이터셋으로 확대하면 일치도가 79.36%에서 44.57%로 급감하며 이 밖에 “국명<지명” 44.69%, “행정구역<지명” 56.71%로 낮아진다. 한편, “법률명”의 경우 254개 데이터셋에서는 약 97.45%의 일치도를 나타났으나, 3,506개 데이터셋에서는 78.13%로 일치도가 감소한 듯 보이나, 타 주제명 유형 중 상대적으로 데이터셋 변화에 비해서 다소 변동이 적어서 데이터셋에 따라 주제명 수가 증가하더라도 편차가 적은 일치 성능이 나타나는 것을 확인할 수 있다.

<표 40> 주제명 범주에 따른 일치도(%)

주제명 수 / 주제명 유형	254개		1,106개		3,506개	
	일치	불일치	일치	불일치	일치	불일치
국명 < 지명	68.52	31.48	55.68	44.32	44.69	55.31
국보.보물 < 기념물 < 주제어	-	-	-	-	20.69	79.31
기념물 < 주제어	-	-	55.32	44.68	39.20	60.80
동물 < 생물 < 주제어	-	-	64.94	35.06	40.56	59.44
법률명 < 주제어	97.45	2.55	81.56	18.44	78.13	21.88
상품명 < 주제어	93.58	6.42	80.63	19.37	66.98	33.02
식물 < 생물 < 주제어	-	-	-	-	36.57	63.43
주제어	61.94	38.06	51.15	48.85	40.99	59.01
지명	79.36	20.64	68.08	31.92	44.57	55.43
통일서명	67.76	32.24	63.40	36.60	58.48	41.52
행정구역 < 지명	77.73	22.27	65.98	34.02	56.71	43.29

3.2 목차 중심 분류

목차는 도서나 문헌의 주요 내용을 압축하여 제시하는 형식으로 짧은 문장 형태의 서명보다 더 많은 정보를 제공한다. 사회과학 분야 도서의 서명과 목차에 대한 통계적 특성에 관한 연구(이용구, 2019)에 의하면, 두 유형의 텍스트 모두 명사 중심의 형식으로 이루어져 있으나 목차가 서명보다 50배 정도 더 많은 명사를 제공하며, 목차만이 고유하게 가지는 명사의 비율도 95% 정도 되는 것으로 파악되었다.

실험 데이터에서 목차 구축 서지 레코드는 전체의 약 39%에 해당하는 47만여 건에 해당하며, 주제명은 이 데이터에 대해 75만여 회 부여되었다. 주제명의 출현 횟수에 따른 데이터셋 현황과 학습에 사용된 저자전거 데이터의 현황은 <표 41>과 같다. 전체 서지데이터에서는 1,000회 이상 출현한 주제명의 수가 254개였으나 목차 구축 데이터셋은 그 수가 상대적으로 작아 56개에 해당한다. 해당 기준의 서지 레코드수나 전체에서 차지하는 비율 또한 낮은 것을 알 수 있다.

<표 41> 목차 중심 데이터 현황

데이터셋 현황				저자 전거 현황	
출현 횟수	주제명 수	서지 레코드수	비율	출현 횟수 기준	총 저자 수
1,000	56	128,977	17.2%	3	8,252
500	152	194,223	25.9%	4	7,265
300	348	268,304	35.7%	5	7,196
100	1,351	431,604	57.4%	6	9,102

목차가 기입된 서지데이터의 경우 자체에 서명과 저자명이 존재하므로 이 3가지 정보를 조합하여 분류 자질을 생성하고 각각에 대해 분류 성능을 산출하는 것이 가능하다. 목차를 중심으로 서명과 결합하거나 추가적으로 저자까지 결합하여 실험 데이터셋을 만들 수 있다. 여기에 목차와 서명 단일 자질로 하면 동일한 조건에 해당하므로 두 자질 간의 직접적인 비교가 가능하다. 따라서 이 연구에서는 이러한 조합을 사용하여 서명, 목차, 목차+서명, 목차+서명+저자 등으로 분류 자질을 구분하고 출현 횟수 기준 1,000, 500, 300, 100에 따라 56개, 152개, 348개, 1,351개의 주제명 수를 포함하는 데이터셋을 생성하여 각각에 대해 분류 성능을 측정하여 <표 42>의 결과를 얻었다. 동일한 주제명 수 기준에 따

라 서로 다른 분류 자질을 조합하여 생성한 데이터셋 간 분류 성능을 비교하면 대체로 분류 자질을 많이 조합할수록 성능이 향상되었다. 또한 부여한 주제명 수가 적은 데이터셋이 더 좋은 성능을 보였다. 다만, 주제명 348개의 데이터셋(세트명 Toc348)은 다소 일관되지 않은 모습을 보였다.

<표 42> 목차 중심 자질의 분류 성능(microF1 기준)

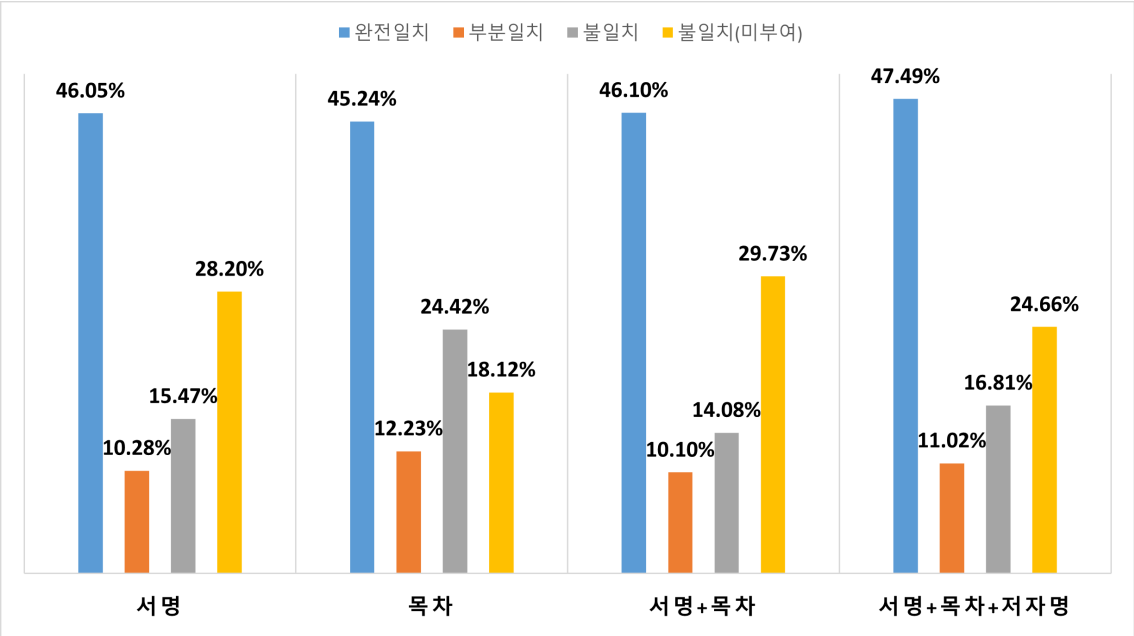
세트명	주제명 수	분류 자질			
		서명	목차	목차+서명	목차+서명+저자
Toc56	56	0.6809	0.6909	0.7016	0.7051
Toc152	152	0.6391	0.6262	0.6401	0.6449
Toc348	348	0.6025	0.5787	0.6128	0.6165
Toc1351	1,351	0.4676	0.5007	0.5431	0.5714

목차와 서명에 대한 통계적 특성에 비하면 두 자질 사이의 성능이 크게 차이가 나지 않아 기존의 다른 연구와 다른 결과를 가져오는 것을 알 수 있다. 사회과학 분야에 도서에 대한 DDC 강목에 대한 자동 분류 연구(이용구, 2020)에서는 목차의 풍부한 자질로 인해 분류 재현율과 분류 정확률 모두를 향상시키는 것으로 나타났으며, 특히 목차는 분류 정확률보다 분류 재현율을 높이는데 더 효과적인 것으로 제시하였다.

기존의 연구들과 달리, 이 연구에서는 BERT의 문장 수준 임베딩을 추출하고 이를 신경망 분류기에 적용하여 주제명 자동 분류를 수행하였다. 이는 기존의 기계학습에 의한 자동 분류 방법과는 다른 점으로 이러한 환경에서 목차 자질이나 복합 자질에서 분류 성능을 개선하기 위한 방법을 연구할 필요가 있다.

출현 횟수 300회 기준으로 주제명 수 348개의 데이터셋(세트명 Toc348)에 대해 단일 및 복합의 분류 자질에 따른 성능을 서지 레코드에 부여된 주제명과의 일치 정도에 따라 분석하면 <그림 62>와 같다. 이 데이터셋에서 분류 자질에 따라 성능 차이가 크게 나지 않지만, 목차 자질이 다른 분류 자질에 비해 가장 낮은 성능을 보였는데 일치도 비율을 보면 수작업으로 부여된 주제명을 기준으로 기계가 주제명을 완전히 잘못 분류한 불일치가 다른 자질이 비해 높았으며, 오히려 주제명 미부여로 인한 불일치의 경우 상대적으로 낮았다. 서명의 경우

목차와 반대로 미부여된 비율이 높았으며 주제명을 잘못 부여한 비율은 상대적으로 낮았다. 서명의 경우 짧은 문장이나 적은 자질로 인해 미부여될 가능성이 있지만, 서명에 좋은 자질이 출현하는 경우 짧은 문장으로도 주제명을 오분류하는 경우가 줄어드는 것을 의미한다. 목차는 풍부한 자질로 주제명의 미부여를 줄여 주지만 다소 오분류를 초래하는 경향이 있음을 알 수 있다.



<그림 62> 분류 자질에 따른 일치 정도 비율(348개 데이터셋)

출현 횟수 300회 이상의 주제명 348개 데이터셋(세트명 Toc348)에 대해 주제명의 범주 유형에 따른 분류 성능을 파악하기 위해 자질을 다양하게 조합하여 분석하였다. <표 43>에 따르면 “법률명”과 “상품명”의 경우 대체로 높은 일치도를 보이는데, “법률명”의 경우 상대적으로 목차 자질에서 좋은 성능을 보이며 “상품명”과 “행정구역”의 경우 서명에서 좋은 성능을 보였다. “지명”의 경우 서명과 목차 자질을 조합한 복합 자질이 단일 자질보다 더 좋은 성능을 보였으며, “국명”의 경우 다른 범주보다 일치도가 낮았다.

<표 43> 분류 자질과 주제명 범주에 따른 일치도(%)

자질 / 범주	서명		목차		서명+목차		서명+목차 +저자명	
	일치	불일치	일치	불일치	일치	불일치	일치	불일치
국명 < 지명	41.74	58.26	41.87	58.13	41.61	58.39	46.53	53.47
법률명 < 주제어	92.54	7.46	94.53	5.47	93.53	6.47	93.03	6.97
상품명 < 주제어	87.26	12.74	77.07	22.93	85.99	14.01	86.62	13.38
주제어	49.47	50.53	50.96	49.04	49.50	50.50	51.45	48.55
지명	71.11	28.89	65.93	34.07	74.44	25.56	75.93	24.07
통일서명	56.40	43.60	37.44	62.56	52.71	47.29	55.91	44.09
행정구역 < 지명	78.38	21.62	63.51	36.49	75.68	24.32	75.68	24.32

어떤 분류 자질을 사용하느냐 또는 어떤 분류 자질을 조합하느냐에 따라 KDC 주류에서 분류 일치 현황을 나타내면, <표 44>와 같다. KDC 주류에서 단일 자질인 서명의 경우 자연과학(4XX), 총류(0XX), 언어(7XX) 분야에서 완전 일치도가 가장 높았으며 특히 총류의 경우 목차 자질에 비해 큰 차이를 보였다. 다만 서명은 짧은 단문이라는 특성 때문에 미부여에 의한 불일치가 큰 값을 가지고 부분 일치에서 낮은 성능을 보였다. 목차의 경우 상대적으로 미부여 불일치도가 낮으며 부분 일치도가 높거나 주제명을 잘못 부여한 불일치가 다소 높게 나타났다. 텍스트 특성으로 인한 단일 자질의 성능을 그대로 보여주는 모습이지만 앞서 제시된 것처럼 목차가 더 많은 분류 자질을 가짐에도 상대적으로 낮은 성능을 보여주어 개선이 필요해 보인다.

<표 44> 분류 자질에 의한 KDC 주류의 분류 일치 현황(%)

자질	일치 정도	0XX	1XX	2XX	3XX	4XX	5XX	6XX	7XX	8XX	9XX
서명	완전 일치	57.18	35.49	25.41	51.39	68.70	52.30	40.91	50.32	49.03	39.73
	부분 일치	11.64	6.81	9.39	8.71	4.12	10.83	28.03	15.10	8.30	10.20
	불일치	16.82	17.90	20.30	14.42	11.37	15.06	12.36	14.09	14.52	17.65
	미부여 불일치	14.36	39.81	44.90	25.48	15.82	21.81	18.71	20.49	28.15	32.41
목차	완전 일치	50.03	36.36	29.06	48.63	64.09	52.64	40.91	45.22	48.86	38.52
	부분 일치	12.14	8.86	13.51	11.15	6.75	12.98	29.36	13.87	9.23	12.80
	불일치	25.87	27.11	33.13	23.40	18.45	21.93	15.37	27.89	23.18	26.89
	미부여 불일치	11.95	27.67	24.30	16.82	10.71	12.46	14.37	13.01	18.73	21.79
서명 + 목차	완전 일치	56.48	33.39	26.07	51.63	67.71	53.07	40.42	52.70	48.84	39.52
	부분 일치	11.70	6.94	10.03	8.76	4.94	9.83	28.47	13.37	7.60	10.58
	불일치	16.13	16.37	19.05	13.09	9.88	13.19	10.88	14.88	12.72	16.90
	미부여 불일치	15.69	43.30	44.84	26.52	17.46	23.91	20.23	19.05	30.84	33.00
서명 + 목차 + 저자명	완전 일치	58.44	35.92	27.97	52.44	68.20	54.62	41.61	54.64	50.46	40.07
	부분 일치	12.33	7.81	11.56	9.63	5.11	11.36	28.88	13.59	8.22	12.34
	불일치	16.19	18.73	24.35	15.96	12.69	15.84	13.40	16.82	14.99	18.78
	미부여 불일치	13.03	37.54	36.12	21.97	14.00	18.18	16.11	14.95	26.33	28.82

원문의 경우 전체 서지 레코드의 약 5.4%에 해당하는 65,825건의 데이터만 구축되어 있어 비교적 그 규모가 작은 상황이며, 주제 및 연도별 분포가 국가서지 전체 데이터와 상이하여 서명이나 목차 자질의 분류 성능과의 직접적인 비교는 어렵다. 다만, 원문이 구축된 데이터 내에서는 서명과 원문의 일부를 자질로 얼

마나 반영하느냐에 따라 분류 성능을 파악해보면, <표 45>와 같다. 출현 횟수 기준을 100회와 50회 이상인 주제명 74개와 221개의 데이터셋에 대해 원문이 서명보다 월등히 좋은 성능을 보여주었으며, 원문에서 분류 자질로 사용한 길이를 기준으로 앞부분에서 1,024개의 어절을 사용한 자질보다 2,048개의 어절을 사용한 자질이 상대적으로 더 좋은 성능을 가져왔다. 향후 원문의 구축이 더 많이 이루어진다면 추가적인 실험을 통해 보다 다양하게 분석하여 일반화된 결과를 가져오는 것이 바람직할 것으로 보인다.

<표 45> 분류 자질에 따른 원문 데이터의 분류 성능(microF1 기준)

학습 데이터 현황			분류 성능		
출현 횟수	주제명 수	비율	서명	원문 (1,024어절)	원문 (2,048어절)
100	74	18.2%	0.4706	0.7270	0.7338
50	221	28.8%	0.2464	0.5415	0.5630

4. 오류 검증

딥러닝 모형과 같이 기계학습에 의해 분류한 결과는 일반적으로 정확률과 재현율 같은 분류 성능 척도에 의해 제시되고 설명된다. 이들 척도에 의한 수치 또는 통계의 경우 전체적이고 전반적인 분류 결과를 나타내기에, 보다 세부적이거나 구체적인 오류 분석이 추가되면 보다 나은 제안점이나 발전 방안을 도출할 수 있다.

이 연구에서는 앞서 제시된 실무자의 수작업 분류와 자동 분류의 일치성 분석과 함께 자동 분류 결과에 대한 주제명 부여 실무 전문가의 평가와 분석을 수행하였다. 이를 통해 기존에 사서가 수작업으로 부여한 주제명표목과 기계학습 기반의 자동 분류 결과를 주제명표목 업무지침에 따라 비교 분석하여 현재 업무 상황에 대한 부분과 자동 분류 결과에 대한 시사점을 제시하고자 하였다. 아울러 자동 분류의 최적화 방안으로 제시하기 위해 기계가 부여한 빈도수가 많은 주제명 중에서 오류가 많은 사례에 대한 분석을 수행하였다.

주제명 부여의 일관성을 확인하기 위한 실무 전문가의 평가 대상으로 기계가 자동으로 부여한 주제명과 실제 서지에 부여된 주제명이 다른 데이터를 무작위로 추출하여 수백 건의 사례를 분석하였다. 먼저 대표적인 유형의 주제명을 사례로 제시하면 <표 46>과 같다. 첫 번째 사례와 같이 일부에서는 실무자가 잘못 부여한 주제명을 기계가 올바르게 부여한 경우가 있다. 두 번째 사례처럼 기계가 다수의 주제명을 완전하게 부여하지 못하는 사례도 있으며, 동일한 주제의 문헌에 대해 반대로 기계가 정확하게 다수의 주제명을 부여한 사례도 있다. 또한 더 이상 부여되지 않는 주제명을 기계가 잘못 학습하여 부여하는 사례도 찾아볼 수 있다. 이와 유사하게 마지막처럼 주제명이 서명에 출현하는 경우를 학습하여 기계가 완전히 잘못 분류하는 사례도 있다. 디자인이라는 단어가 서명에 들어가면 주제명 '디자인[design]'으로 부여하는 경우를 기계가 너무 일반화하여 학습한 것으로 보인다. 이와 유사하게 박노균 저자의 「김영랑: 최고의 순수 서정 시인」이라는 도서에 대해서도 정답인 '한국 근대 문학[韓國近代文學]' 주제명 대신에 '한국 현대시[韓國現代詩]' 주제명을 자동 분류하는 유사한 사례들도 다수 있다.

<표 46> 오류 검증용 주제명 부여 사례

서명	분류번호	수작업 분류 주제명	자동 분류 주제명
(예제로 배우는) 한글 2005	005.3(4)	컴퓨터 프로그래밍 [computer programming]	워드 프로세서 [word processor]
기본간호학 실습서	512.8(5)	임상 실습[臨床實習]; 간호학[看護學]	간호학[看護學]
기본간호학 실습지침서	512.8(6)	간호학[看護學]	간호학[看護學]; 임상 실습[臨床實習]
(건축 인테리어 제작실무를 위한) 한글 AutoCAD 2007	542.2027(4)	건축 설계[建築設計]; 오토캐드[AutoCAD]; 인테리어 디자인[interior design]	오토캐드[AutoCAD]
(요점) 평생교육프로그램 개발:교육	378.1(4)	평생 교육[平生敎育]	평생 교육[平生敎育]; 방송 통신 교재[放送通信敎材]
라떼아트 테크닉: 카페라떼의 디자인 원리	573.93(5)	커피[coffee]	디자인[design]

오분류를 포함한 기계의 자동 부여에 대해 실무자의 몇 가지 주요한 검증 결과를 긍정적인 측면과 부정적인 측면에서 나누어 제시하면 다음과 같다.

먼저 긍정적인 측면에서 보면, 기계가 실무자가 부여한 정답보다 더 많은 개수의 주제어를 부여한 경우 일정 부분에 있어 적합도가 매우 높은 사례를 다수 보인다는 점이다. 즉 실무자가 부여한 것이 충분하지 않은 경우는 이런 점들을 기계가 보완하는 패턴을 발견할 수 있었다. 예를 들어 홍완표 저자의 「민법 및 민사특별법」이라는 도서는 실무자의 정답으로 주제명 '민법[民法]'이 부여되어 있지만, 기계는 '민법[民法];공인 중개사 시험[公認仲介士試驗]'으로 부여하여 주제명 '공인 중개사 시험[公認仲介士試驗]'을 추가하고 있는 것을 알 수 있다. 이 경우는 서명 정보만으로는 추가된 주제명을 추측할 수 없으며, 목차에서 기인하는 주요 정보(자질)가 기여하는 것으로 판단한다. 이렇듯 실무자의 정답에 기계가 주제명을 추가하는 사례에서 높은 적합도를 보이는 것은 기계학습이나 인공지능이 가지는 장점의 단면을 잘 보여준다고 볼 수 있다. 이러한 부분을 고려한다면 실제 기계의 자동 분류 성능은 앞서 이미 제시된 자동 분류 성능에서 다소 더 많이 향상될 것으로 예측할 수 있다.

기계가 단일 주제어를 추천한 결과 사서가 부여한 것보다 적절한 경우가 있는데, 그 경향은 서명에 해당 텍스트가 들어 있는 경우가 많은 것으로 보인다. 예를 들어 「글쓰기의 지평」이라는 서명의 도서에 대해 실무자는 주제명 '한국 현대 수필[韓國現代隨筆]'로 부여했다면, 기계는 서명 자질로부터 학습에 의해 주제명 '글쓰기'를 부여하였다.

이외에 앞서 예를 든 사례처럼 서명이 짧거나 의미를 식별하기 어려운 경우 부족한 정보를 다른 분류 자질에서 활용하는 것을 알 수 있다. 구체적으로 목차에만 특정하게 출현하는 명사는 주제명으로 추천될 가능성이 높음을 알 수 있다.

특히 논픽션(과학) 분야의 경우 기계가 정답인 수작업보다 더 정확한 주제명을 부여하기도 한다. 예를 들어 「해부생리학」이라는 도서에 대해 실무자는 주제명으로 '피부 미용[皮膚美容];해부 생리학[解剖生理學]'을 부여했지만, 기계는 '인체 해부학[人體解剖學]'을 부여하여 도서의 실제 내용에 보다 근접한 것을 알 수 있다.

반대로 부정적인 측면에서 보면, 실무자가 다수의 주제명을 부여한 사례에 대해 기계가 적게 부여한 경우 적절한 것도 많지만, 자료를 나타내는 데 꼭 필요한 다양한 측면을 전반적으로 다 반영하지 못한 부분이 있어 아쉬운 점을 보인다. 이는 <표 46>에서 두 번째나 네 번째 사례에 해당한다.

기계가 부여한 주제명의 주된 특징을 살펴보면, 기계는 서명에 포함된 키워드(명사)에 반응하고 학습하여 비교적 정확하게 주제명을 부여하는 반면, 주제명이 서명이나 목차에 나타나지 않는 경우는 왜곡된 주제명을 부여하는 문제점을 노출하였다. 예를 들어 도서명「와인 영화 로맨스 그리고 여행」에 대해 실무자는 '수기(글)'라는 주제명을 부여했다면 기계는 '포도주'를 부여하였다. 다른 예로「주식회사 헌법제1조」에 대해 실무자는 '자기관리'를 부여했다면 기계는 제목에 나온 대로 '헌법'을 부여하였다.

현재는 국립중앙도서관에서 부여하지 않은 주제명인데 이번 연구에서 학습 데이터에 일부 관련 사례가 포함되어 기계가 이들 주제명에 대해 학습하여 자동 부여하였다. 향후 부여된 주제명 수정 등 서지데이터 정비 작업을 통해 재발 방지가 필요하다. <표 46>에서 제시된 주제명 '방송 통신 교재[放送通信教材]'가 여기에 해당한다.

기계가 논픽션(과학) 분야를 잘 분류한다면, 반대로 픽션(문학류)은 잘못된 주제명을 추천하는 경우 빈번히 발생하여 이에 대한 보완이 필요하며, 저작이 다룬 주제(소재)를 추천할 가능성은 희박한 것으로 판단한다. 예로 도서명「서울시민」에 대해 실무자는 '일본문학'을 부여했다면, 기계는 '서울(특별시)'를 부여하였다.

또한 기계의 자동 부여에서 문학 장르에 오류가 많이 나타나는 경향을 보였다. 이를 해결하기 위해서는 주제명표목표의 조정과 주제명 분포의 불균형에 대한 합당한 학습 데이터 구축이 선행되어야 할 것으로 보인다. 이는 문학 형식의 자동 부여에서도 나타나므로 향후 기계학습에서 적절한 조치가 필요할 것으로 판단된다. 구체적으로 대체로 명백하게 시가 아닌 자료에도 '한국 현대시'를 자주 부여하는 모습을 보였다.

오류 검증에 대해 실무 전문가의 의견을 종합적으로 정리하면, 기계가 자동 부여한 주제명은 전반적으로 실무 전문가의 수작업을 보완 또는 보조하는 용도로 충분히 참고 가능한 것으로 판단하여 제시하고 있음을 알 수 있다.

5. 요약 및 시사점

딥러닝을 포함하여 일반적으로 기계학습은 분류해야 할 범주의 수가 적을수록 좋은 성능을 보이며 범주 수가 많으면 성능이 떨어진다. 예로 이메일의 스팸 분

류와 같이 이진 분류가 다수의 범주를 갖는 신문 기사 분류보다 성능이 대체로 더 높다. 국가서지 데이터에 주제명 자동 부여 또는 분류도 주제명의 출현 횟수가 많은 소수의 주제명일수록 성능이 매우 높음을 알 수 있다. 수천 개가 넘는 주제명을 분류 범주로 하면 성능이 낮아지는 것을 알 수 있다. 이는 향후 실제적인 주제명 자동 부여 또는 자동 분류시스템을 구축한다면 분류 대상이 되는 주제명을 어느 정도 범위까지 해야 할지 결정할 때 중요한 지표가 된다.

분류 자질 측면에서 이 연구를 보면, 주제명 자동 분류에서 출현 횟수와 관계없이 단일 자질로서 서명이 기여하는 바가 타 자질에 비하여 비중이 높음을 알 수 있다. 실무자가 주제명을 부여함에 있어 서명에 출현한 단어를 주제명에 포함된 용어를 위주로 고려하는 부분이 있을 수 있으며, 또한 이 연구에서 사용한 BERT 모형의 워드피스 토크나이저와 임베딩 성능에서 기인한 결과로 볼 수 있다. 이미 다른 연구에서 보이듯이 다른 기계학습 영역뿐만 아니라 딥러닝에서도 전이 학습을 적용한 BERT 모형이 압도적으로 좋은 성능을 가져오고 있다.

서명 이외에 저자, 목차, 원문 등 다수 자질을 활용할수록 자동 분류 알고리즘의 성능이 상승하였다. 구체적으로 KDC 주제영역별로 확인한 결과, 대체로 서명의 기여가 높았으며, 함께 목차가 활용될 경우, 성능이 소폭 상승함을 보였다. 저자명을 사용하는 것 또한 분류 성능을 향상시켰다

오류 검증 결과, 기계가 자동 부여한 주제명은 실무자의 수작업을 보완할 수 있을 것으로 판단되며, 이러한 부분을 방증하는 것으로 다수의 주제명을 부여해야 하는 상황에서 기계가 나름대로 좋은 성능을 보여주는 것을 보아 알 수 있다. 다만 자연과학 등 일부 분야는 좋은 결과를 보이지만, 문학, 종교 분야 등은 그렇지 못해 주제 영역별로 성능의 편차를 보여 현재 모형의 이러한 장단점을 파악할 수 있었다.

Ⅵ. 결론 및 제언

최근 딥러닝 기법을 많은 분야에서 적용하려는 노력이 계속되고 있다. 이는 딥러닝이 최신 기법이기 때문이기도 하지만 다른 기계학습 기법에 비해 압도적으로 높은 성능을 보이기 때문이다. 문헌정보학 및 도서관 분야에서도 인공지능과 딥러닝 기법을 적용하고자 하는 사례들이 해외의 국가대표도서관을 중심으로 보고되고 있다. 예를 들어 미국, 독일, 핀란드, 일본 등의 국가대표도서관이 주제명이나 분류표에 대해 자동 부여나 자동 분류를 진행하거나 계획 중에 있다. 우리나라 국립중앙도서관도 인공지능 기반 요약 및 검색 서비스를 시범적으로 개시하였다.

앞서 이 연구에서는 주요국의 국가대표도서관이 그들의 업무나 연구에 인공지능과 기계학습을 적용하는 사례를 조사하였으며, 적용 분야 중 하나로 주제명의 자동 부여 사례가 등장하고 있음을 알 수 있다. 한편 국립중앙도서관도 이미 이십여 년 전부터 주제명표목표를 개발하고 외부에 개방하고 있으며, 국립도서관 장서에 대해 주제명을 부여하여 국가서지 데이터를 구축하고 있다. 이에 국립중앙도서관이 수행하고 있는 주제명 부여 업무에 대해 딥러닝을 활용하여 자동 분류 또는 자동 부여의 적용 가능성을 제시하며, 구축 시의 최적화 방안을 연구를 통해 제언하고자 하였다.

먼저 도서관의 업무 및 서비스에서 인공지능을 적용하기 위한 영역을 선정하고 그에 맞는 최적의 인공지능 기법을 적용하기 위해, 이 연구에서는 해외 국가대표도서관을 중심으로 문헌정보학 및 도서관 분야의 인공지능과 기계학습 적용 사례를 살펴보았으며, 최근 각광받고 있는 딥러닝 기법을 전반적으로 개괄하였다. 그 결과 주제명 자동 부여를 위해 딥러닝 Transformer 아키텍처를 채택하고 적용하였다.

이 연구는 국립중앙도서관으로부터 수집된 국가서지 데이터와 목차 데이터의 개요를 살펴보고, 이들 데이터에 부여된 주제명의 현황 분석을 통해 주제명표목표와 주제명 활용의 장단점을 파악하고 이를 통해 시사점을 제언하였다. 국립중앙도서관이 구축한 국가서지 데이터의 활용현황 분석을 통해 살펴본 바에 의하면, 개념을 표현하는 주제어의 개발과 이들 주제어 간의 의미적 관계를 재정립하고 더욱 다양한 주제명을 서지 레코드에 부여할 수 있는 방안을 강구해야 한다. 또한 목록 규칙에서 제시하는 서지 요소뿐만 아니라 목차와 원문의 구축을 통해 서지데이터의 접근 및 활용을 높일 필요가 있다. 이는 기계학습에 의한 자동 분

류 측면에서도 매우 중요하다.

주제명 자동 분류 실험을 살펴보면, 분류 자질(특성)은 서지데이터의 서명과 목차를 중심으로 사용하였다. 서명의 경우 대부분은 서지데이터에 존재하지만, 목차는 모든 데이터에 구축되어 있지 않아 따로 자질 집합을 생성하였다. 서명과 목차 각각의 자질에 대해 오직 해당 자질만 사용한 단순 자질과 다른 자질을 조합한 복합 자질을 적용하여 분류 실험을 하였다. 복합 자질의 경우 서명에 저자명을 조합하거나, 목차에 서명과 저자명을 조합하는 방식으로 진행하였다. 주제명에 따라 수만 번 부여된 경우도 있으며 소수나 미부여된 사례도 있어 주제명 부여가 불균형 분포에 가까우므로 학습 데이터 관점에서 고빈도 중심으로 출현 횟수 기준을 달리 하여 학습 및 테스트 데이터셋을 구성하였다.

실험 결과를 요약하면 일반적으로 자질의 풍부한 목차가 더 좋은 성능을 보이는 다른 연구와 달리, 다수의 데이터셋에 대해 서명과 목차가 유사한 분류 성능을 보였다. 이는 주제명 부여 규칙, 분류 모형인 BERT의 문장 임베딩 효과, 비교적 큰 규모의 학습 데이터 등에서 기인하는 것으로 보여 추가적인 연구가 필요한 것으로 나타났다. 또한 단순 자질보다 복합 자질이 더 좋은 분류 성능을 가져와 주제명 자동 분류를 위한 딥러닝도 학습 자질을 더 많이 사용하면 더 좋은 성능을 가져오는 경향을 일관되게 보였다. 다수의 데이터셋으로 실험하였기에 분류 성능이 더 좋은 복합 자질을 평균으로 살펴보면, 6개의 서명 데이터셋이 최소 0.5504에서 최대 0.8149로 평균적으로 0.6894의 마이크로 F1 성능을 보였으며, 4개의 목차 데이터셋은 최소 0.5714에서 최대 0.7051로 평균 0.6345의 마이크로 F1 성능을 보였다. 다만 목차가 구축된 서지레코드가 서명이 포함된 전체 데이터에 비해 1/3정도 수준이라 이 두 복합 자질의 성능을 직접 비교할 수는 없다. 추가적으로 서명과 목차 각각의 대표 데이터셋에 대해 실무자가 부여한 정답 주제명과 기계에 의해 부여된 주제명에 대한 KDC 주류와 주제명의 범주 유형에 따른 일치 정도 여부를 분석하였다. 마지막으로 주제명 자동 분류에 대한 실무자의 오류 검증을 실시하여 제안점이나 발전 방안 등을 도출하였다.

1. 도입 가능성 및 활용 가능성

이 연구는 전이 학습이 가능한 딥러닝 모형인 BERT를 이용하였으며 주제명 추천을 위한 자동 분류 알고리즘의 설계와 검증을 통해 다음과 같은 결과를 도

출하였다.

- 국가서지 데이터를 분석하고 이를 통해 기계학습에 필요한 다양한 분류 자질을 식별하고 이들 자질의 조합을 통해 분류 성능을 파악함으로써 주제명의 자동 부여 측면에서 국립중앙도서관에서 필요한 딥러닝 모형의 기본 방향을 제시하였다.
- 이 연구에서 제시된 자동 분류 알고리즘은 결과적으로 실무자 입장에서 주제명 선정을 위한 추천 시스템으로서의 가능성을 보인 것으로 판단되며, 추후 국립중앙도서관의 경영 관리 측면에서 해당 시스템의 구축이 필요할 것으로 예상된다.
- 향후 주제명 자동 분류 모형 시스템을 도입하기 위한 국립중앙도서관의 서지데이터 입력 및 업무 등 관리 방식과 절차를 개선할 필요가 있다. 일례로 새로 입수되는 도서에 대해 주제명을 추천하기 위해 편목 단계에서 목차를 구축하거나 반입하면 최종 추천 성능이 향상될 것으로 예상된다.

2. 자동 분류 알고리즘 최적화를 위한 제언

이 연구의 결과를 바탕으로 실용적인 주제명 추천 도구로써 자동 분류 알고리즘의 최적화를 위한 방안을 제시하면 다음과 같다.

- 전문(full-text) 성격의 특징을 분류 방법에 잘 반영하는 딥러닝 모형의 개발 및 적용이 필요하다. 즉 최근 딥러닝에서 전이 학습 모형이 대세를 이루고 있으며 BERT 모형의 한계를 넘어서는 개선된 다양한 모형이 개발되고 있는데, 도서관 환경에 맞게 이러한 모형을 개선하거나 새로 개발하는 것이 필요하다. 특히 이러한 모형은 자동 분류에 목차나 원문을 보다 효율적으로 활용할 수 있는 방안이 전제되어야 한다.
- 분류 성능을 높이기 위해 시스템 자원의 한계를 넘어 국가서지 데이터에 포함된 모든 저자 또는 되도록 많은 저자를 활용할 수 있는 방안을 모색할 필요가 있다.
- 국가서지 데이터를 주제명 부여 측면에서 살펴보면 장서의 구성에 따라 주제명의 부여가 불균형을 가진다. 이는 자연스러운 현상이지만 성능을 높이

기 위해 기계학습 측면에서 이를 보완하기 위한 데이터 증강기법이 강구되어야 한다.

- 대용량 텍스트나 말뭉치를 활용한 사전학습 기반의 T5나 Longformer 등과 같이 아키텍처 정교화를 통한 딥러닝 모형의 최적화가 필요하다. 트랜스포머 모형의 인코딩뿐만 아니라 디코딩을 활용하여 생성 모형(generative model)의 적용도 검토할 필요가 있다. 딥러닝 생성 모형은 이 연구에서 적용한 모형과 다른 특징을 보이기에 참고할 정보를 제공할 것으로 예상된다.
- 분류 성능을 주제영역별로 보면, 문학이나 예술 분야가 사회과학이나 자연과학이 비해 저조한 것으로 나타났다. 분류 모형 측면에서 이러한 부분을 학습데이터나 기법을 통해 보완할 수 있어야 한다. 그러나 이것으로 불충분할 수 있으므로 문학류나 예술 분야의 장르나 형식이 제시된 주제명표목표의 패싯화와 이들 주제영역에 특성을 반영한 자동 분류 모형의 개발이 필요하다.
- 주제명의 자동 분류나 자동 부여에서 더 많은 주제명이 활용될 수 있도록 하기 위해 부여 횟수 저빈도 주제어에 기반한 특수 딥러닝 모형을 개발하고 제안할 필요가 있다. 저빈도 주제명을 확대 포함하여 활용 범위를 높여 주제명 추천 시스템의 범용성을 높일 필요가 있다.

참고문헌

<국내문헌>

이미화, 이지원. (2021). 주요국 국가서지 현황조사를 통한 국가서지의 최신 경향 분석. 한국비블리아학회지, 32(1), 35-57.

이용구. (2019). 사회과학 분야 도서의 목차 텍스트에 대한 통계적 특성에 관한 연구. 정보관리학회지, 36(2), 255-273.

이용구. (2020). 목차 정보와 kNN 분류기를 이용한 사회과학 분야 도서 자동 분류에 관한 연구. 정보관리학회지, 37(1), 1-21.

<국외문헌>

Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences, 1*(12), 1-15.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Baoli, L., Qin, L., & Shiwen, Y. (2004). An adaptive k-nearest neighbor text categorization strategy. *ACM Transactions on Asian Language Information Processing (TALIP), 3*(4), 215-226.

Brygfjeld, S. A., Wetjen, F., & Walsøe, A. (2017). Machine learning for production of Dewey Decimal. *IFLA WLIC 2018*. https://library.ifla.org/id/eprint/2216

Chen, S. (2018). K-nearest neighbor algorithm optimization in text categorization. In *IOP conference series: earth and environmental science* (Vol. 108, No. 5, p. 052074). IOP Publishing.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory, 13*(1), 21-27.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

Islam, M. Z., Liu, J., Li, J., Liu, L., & Kang, W. (2019, November). A semantics aware random forest for text classification. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 1061-1070).

Khalil Alsmadi, M., Omar, K. B., Noah, S. A., & Almarashdah, I. (2009, March). Performance comparison of multi-layer perceptron (Back Propagation, Delta Rule and Perceptron) algorithms in neural networks. In *2009 IEEE International Advance Computing Conference* (pp. 296-299). IEEE.

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., & Barnes, L. E. (2017, December). Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)* (pp. 364-371). IEEE.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks, 3361*(10), 1995.

Lewis, D. D., Yang, Y., Russell-Rose, T., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research, 5*(Apr), 361-397.

Maron, M. E. (1961). Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM), 8*(3), 404-417. https://doi.org/10.1145/321075.321084

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010, September). Recurrent neural network based language model. In

Interspeech (Vol. 2, No. 3, pp. 1045-1048).

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011, May). Extensions of recurrent neural network language model. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5528-5531). IEEE.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Technical Report. *OpenAI*. URL:https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Ramchoun, H., Idrissi, M. J., Ghanou, Y., & Ettaouil, M. (2017, March). Multilayer Perceptron: Architecture Optimization and training with mixed activation functions. In *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications* (pp. 1-6).

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics, 21*(3), 660-674.

Sivic, J., & Zisserman, A. (2008). Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 591-606.

Suominen, O., Inkinen, J., & Lehtinen, M. (2022). Annif and Finto AI: developing and implementing automated subject indexing. *Italian Journal of Library, Archives and Information Science, 13*(1), 265-282. https://doi.org/10.4403/jlis.it-12740

Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning, 4*(4), 267-373.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

van den Bosch, A. (2017). Hidden Markov Models. In: Sammut, C., Webb, G.I. (eds) *Encyclopedia of Machine Learning and Data Mining* (pp. 609-611). Springer, Boston, MA. https://doi.org/10.1007

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review, 53*(8), 5929-5955.

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30* (pp. 6000-6010).

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science, 44*(1), 48-59.

Yoon Kim. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746-1751).

Zhang, M. L., & Zhang, K. (2010). Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 999-1008).

<웹사이트>

Reuters Corpus. (2017). https://martin-thoma.com/nlp-reuters.

Arxiv Academic Paper Dataset. (2018). https://github.com/lancopku/SGM

Library Congress labs. https://labs.loc.gov/

NDC predictor https://lab.ndl.go.jp/ndc/

국가서지를 활용한 주제명 자동 분류 적용방안 연구

연 구 주 관	‖	국립중앙도서관
연 구 수 행	‖	경북대학교 산학협력단 (책임연구원: 이용구, 공동연구원: 이종택)
발 행 처	‖	국립중앙도서관 국가서지과
발 행 인	‖	국립중앙도서관
발 행 일	‖	2022년 9월 25일
I S B N	‖	979-11-6513-294-1 (종이책) 979-11-6513-296-5 (PDF)