정보통신관련법 판례분석 방법론

컴퓨터공학과 15학번 최현호

지도교수: 이영구 교수님 / 한용구 멘토님

요약

법조계 종사자 및 정보통신분야 종사자 간의 간극 해소를 위하여, 정보통신 유관 판례에 대한 주제 추출 알고리즘을 제안한다. 주제 도식화 및 주제별 판례 분류를 통해 기존 판례명만으로는 파악할 수 없었던 판례 간 유사성 확인 및 판례 내용 유추를 지원한다.

1. 서론

1.1. 연구배경

2020년 12월 정식 서비스를 시작하였던 스캐터랩(SCATTER LAB)의 열린 주제 대화형 인공지능 챗봇(Open-Domain Conversational AI Chatbot)인 이루다는 여러 사회적 파장을 일으켰지만, 그 중 직접적으로 법적인 제제를 받았던 이유는 개인정보보호법 제28조의2제2항을 위시한 8개 조항을 위반하였기 때문이다.[1] 비단 이루다의 경우뿐만 아니라, 현재 제4차 산업혁명의 최전선을 담당하는 인공지능을 연구하는 데에 있어 연구 데이터는 정보통신정책과 법에서 자유로울 수 없다. 위반했을 경우의 처벌 수위 역시 마찬가지인데, 현행 징벌적 손해보상의 범위인 3배 배상제도의 수위를 사건의 경중에 따라 강화해야 한다는 목소리도 있으며[2], 한편으론 강력한 제제는 산업 전반의 발전 저해를 초래한다는 목소리가 동시에 있어 양측의 의견을 적절히 절충할 필요성이 있다.

사실 법을 직접 전공하지 않는 경우, 일반적인 사람들이 법에 대한 관심을 크게 가지는 경우는 많지 않다. 또한, 업무 상 저촉 가능한 법 역시 직접적인 관계자가 아니라면 신경 쓰지 않는 경우가 대부분이다. 물론, 직접적으로 법을 다루는 사람들에게도 필요한 판례를 바로 찾아내는 것은 쉬운 일이 아니다. 현재 법제처에서 제공하는 판례를 찾는 방법에는 키워드 검색, 판례번호, 관련 법을 통한 접근의 총 3가지가 있다. 하지만, 이러한 방식은 이미 알고 있는 판례를 찾아내는 것엔 도움이 될지 모른다. 하지만, 특정 판례와 비슷한 판례를 찾아내야 할 경우나, 특정 주제를 관통하는 판례들을 뽑아내기에는 현재의 판례 제공 시스템은 너무나도 단순하다. 따라서 본 연구에서는 정보통신 및 개인정보 관련 판례를 이용한 주제 모델링을 진행하여, 기존의 판례명으로는 알 수 없었던 판례의 개략적인 내용 및 유사 판례 분류를 진행한다.

1.2. 연구목표

현재 대한민국 법원의 종합법률정보 포털을 제외하고 양질의 판례를 제공하는 대중성 있는 플랫폼은 약 4개가 있다[3]. 저작권법 제7조에 의거하여, 법률 및 판례는 저작권법의 보호를 받지 못한다. 하지만 프로젝트 진행에 있어서의 문제를 최대한 피하기 위해, 판례 수집은 대한민국 법원 종합법률정보에서 조회 가능한 정보통신 분야의 판례에 한한다. 주제 모델링 진행 전, 법령용어가 제대로 분류되지 않는 것을 피하기 위해 한국어 전처리 시스템에 법령용어를 수록한다. 그리하여 토큰화 진행 시 의도와 전혀 다른 단어가 핵심 단어로서 분류되는 것을 방지한다. 법령용어 역시 대한민국 법원 종합법률정보에서 제공하는 데이터를 기반으로 수록한다. 이 중, 합성어는 필수 단어가 아닌 경우 제거하고 수록한다. 추가로, 분류된 주제와 추출 판례의 유사성을 검증하기 위하여 각 주제 별 대표 판례 하나씩을 선정한 후, 해당 판례와 타 판례 간의 문서 간 유사도를 도입하여, 비슷한 정도를 확인하는 방안을 고려한다. 해당 추가작업을 진행하기에 데이터가 충분하지 않다면 육안검별을 진행한다.

2. 관련연구

2.1. 주제 모델링 알고리즘

통상적인 LDA 알고리즘은 특정 주제와 관련된 댓글 뭉치를 기반으로 분류를 진행한다. 목적은 주로 산문 분석을 통한 목적 및 주제 파악이다. 허나, 본 연구에 활용하는 데이터는 판례라는 공통점과 비슷한 작성 양식이라는 점을 제외하고는 대체로 다루는 주제가 제각각이기 때문에 댓글 분석과는 다른 양상을 띌 가능성이 높다. 하지만, 판결의 형태 및 사건 양상의 유사성을 분류하는 것이 본 연구의 목적이며, 유사 판례명으로 명명된 판례 간의 차이점을 구별하는 것에 해당 알고리즘을 응용할 수 있다.

2.2. 판례 제공 서비스

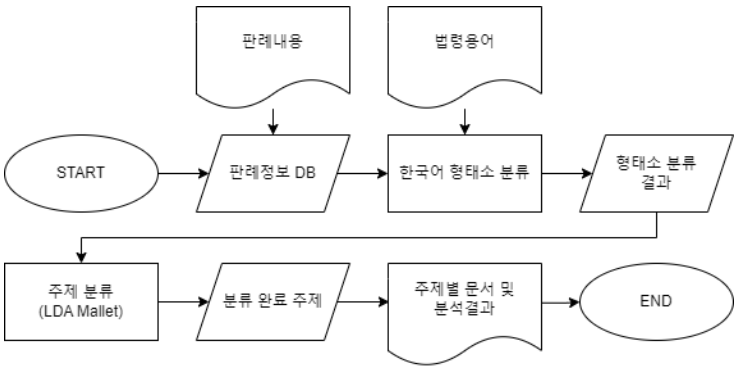
현재 판례 제공 플랫폼의 서비스 양상은 전체 공개되지 않은 판결문에 대한 서비스가 주가 된다. 독보적인 판례를 확보하여 타 플랫폼과의 차별점을 강조하는 서비스가 대부분이며, 빅케이스의 경우 단순 검색이 아닌 쟁점별 판례 검색이나 핵심요약, 유사도 높은 판례를 제공하는 등 NLP를 도입한 검색 서비스가 특징이다.

3. 프로젝트 내용

3.1. 요구사항

사용자는 분류 완료된 주제명을 확인하고, 해당 주제명을 통해서 핵심 키워드 및 바인딩된 판례를 확인하고 그 내용을 확인할 수 있도록 하는 것을 목표로 한다. 한국어 자연어처리(Natural Language Processing, 이하 NLP)에는 전처리 과정에서의 형태소 단위 분리가 필수적이며, 의미있는 단위의 단어가 분해되는 것을 방지하는 것을 1차 목표로 한다. 이를 통해 누락된 핵심 용어로 인한 프로젝트 신뢰성 저하를 방지한다.[4)]

판례 분석의 순서는 다음과 같다. 우선적으로 판례정보 DB에 접근하여 정보통신 관련 키워드("정보통신", "개인정보", "인터넷")를 포함한 판례명을 추출한 후, 해당하는 판례 내용의 텍스트 문서에 직접 접근한다. 이후 법령용어를 추가 탑재한 한국어 형태소 분류기 Okt를 이용하여 판례상세내용을 토큰화시킨 후 배열에 저장한다. 해당 배열을 기반으로 LDA Mallet을 이용, 최적화된 주제 개수를 계산한 후 그 주제에 맞게 판례를 분류한다. 마지막으로 그 분석 모델 및 분류 결과를 출력 가능한 형태로 생성한다. 위 과정은 [그림 1]에 도식화하였다.



[그림 1]

프로젝트 진행에 여유가 있다면, 판례에 명시된 법 조항을 해당하는 내용으로 치환하여 전처리 과정에서 법과 직결되는 부분이 분해되는 것을 방지하도록 한다. 또한, 분류가 완료된 판례에 대하여 키워드 분석 및 텍스트 요약 기술(TextRank), 유사도 비교 모델(Doc2Vec)을 활용하여 육안 감별 절차를 간단하게 하거나, 혹은 그 과정 자체를 생략할 수 있도록 한다. 현재 제공되는 판례의 이름은 법제처에서 임의로 명명하며, 그 이름으로는 판례 내용을 유추할 수 없어, 우선적으로는 육안 감별을 기준으로 한다.

4. 구현 관련 정보 및 결과

본 연구를 위해 설정한 시스템 환경 및 언어 등의 요소는 [표 1]과 같이 정리하였다.

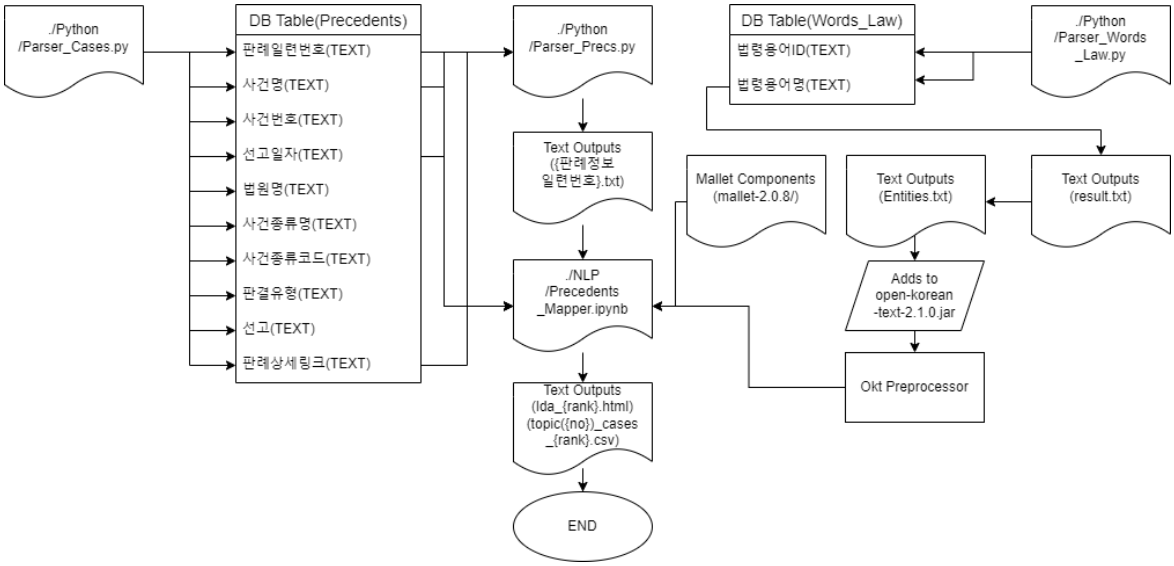
CPU	AMD Ryzen 3400G 3.7GHz
RAM	DDR4-25600 16GB x 2
OS	Windows 11 Education
Database	SQLite3
Language & Framework	Python 3.8.3 (LDA Modelling / Data Mining) Jupyter Notebook

[표 1]

LDA 모델 생성 및 분석에는 Gensim 라이브러리의 LDA Mallet을 활용하였다. 또한 데이터는 2022년 10월 기준의 대한민국 법률 종합법률정보에서 제공하는 판례, 법령용어를 활용하였다. 한국어 NLP에 활용되는 대표적인 형태소 분류기로는 Konlpy 라이브러리에 포함된 Mecab-ko(은전한닢) 및 Okt(Open Korean Text)가 있다. 비록 성능 면에서는 Mecab-ko의 처리 속도가 Okt보다 약 46배 가량 빠르지만, 본 프로젝트의 목적은 범용적인 NLP 모델을 생성하는 것이 아니기 때문에 분석하는 텍스트의 양이 많지 않다. 또한, 사용자가 임의로 분류 단어를 추가하는 것이 Mecab-ko에 비해 Okt가 더 간단하기 때문에, 본 연구에서는 Okt를 활용하였다. 한국어 전처리 결과를 기반으로 LDA 모델링 결과를 쉽게 손볼 수 있도록 Jupyter Notebook를 활용하여 전처리 코드와 LDA 코드를 분리하였다.

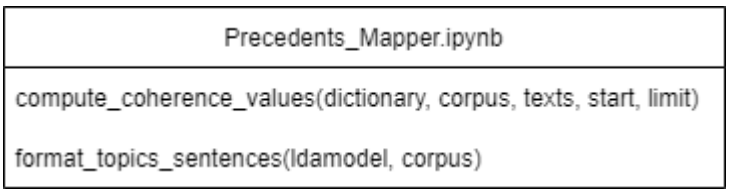
판례 분류 및 추출 작업을 용이하게 하기 위하여 SQLite3를 사용하여 판례 정보 DB를 구축하였고, 이를 기반으로 판례내용을 비롯한 상세 내용을 일반 텍스트로 추출하였다.

Github 게시 코드 및 프로젝트 구동에 필수적인 요소를 [그림 2]와 같이 정리하였다. 이 중, 판례 분석에 직접적으로 이용되는 DB, 파일 및 코드는 Precedents.db, Precedents_Mapper.ipynb이며, 해당 그림에 언급되지 않은 파일 및 코드는 프로젝트 재구상 후 용도폐기된 코드이다. 주제 분류의 정확도가 검증 완료된 후, 실제 서비스에 이용될 파일은 Precedents_Mapper 실행 후 생성되는 모델을 추출한 파일과 그 결과로써 추출된 판례번호만을 이용할 것으로 전망한다.



[그림 2]

[그림 2]에 언급된 Precedents_Mapper.ipynb 코드에서 주제 모델링 및 분류를 진행하는 코드는 [그림 3]과 같다.



[그림 3]

compute_coherence_values() 함수는 분류된 모델의 coherence 값을 통하여 최적의 주제 개수를 계산하는 함수이다. 이 중 limit 파라미터를 통하여 주제 개수 제한을 설정하는데, 최대 주제 개수를 20개로 설정하고 10회 분류를 진행하였을 때, 약 17개의 주제가 선정되었다. 허나, 주제를 육안으로 검토하였을 때 핵심 키워드와 분류된 판례가 일치하지 않는 점이 확인되었다. 추가로, 일부 판례의 경우 같이 분류된 판례 간의 유사성이 희미한 경우도 적지 않았다. 다만, 최대 개수 제한이 20개였을 때 12개의 주제가 분류되었을 때의 결과가 가장 유의미함을 육안으로 검증하였고, 이후 최대 개수를 16개 및 10개로 제한한 후 실험을 진행하였을 때, Coherence값이 0.59 이상 검출되었을 때 상대적으로 분류 결과가 유의미했다.

format_topics_sentences() 함수는 compute_coherence_values()함수를 구동하였을 때 출력되는 최적 주제를 기반으로 판례를 분류하는 함수이다. ldamodel 파라미터에는 compute_coherence _values() 함수를 통해 얻은 최적 주제 변수(optimal_model) 및 형태소 분류된 결과(corpus)를 대입하여 판례별 적합 주제의 확률을 계산하고, 그 확률이 가장 높은 주제를 'Dominant_Topic' 열에

저장한다. 이런 과정을 모든 입력 판례 및 그 corpus에 대해 진행하여, 지배적 주제(Dominant Topic)를 기반으로 유사 판례를 묶어낸다.

다만, format_topics_sentences() 함수를 구동하였을 때, 각 주제에 맞는 주제어 모음을 생성하는 부분에서 문제가 발견되었다. 가장 대표적으로 눈에 띄는 단어는 '국가정보원' 및 '국정원'이었는데, 그 결과를 이하 [그림4, 5, 6]에 정리하였다.

Topic_Num	Keywords	Num_Documents	Perc_Documents
1	광고, 프로그램, 카페, 주식회사, 게시, 글, 광고주, 전화, 항의, 중단	18	0.057
2	개인정보, 제공, 진술, 누설, 원, 처리, 당원, 자, 업무, 명부	42	0.1329
3	국가정보원, 활동, 보고, 직원, 지시, 법인, 팀, 진술, 국가정보원장, 순번	4	0.0127
4	범행, 차, 콜센터, 양형, 상담원, 조직, 사이트, 번호, 범죄수익, 금액	18	0.057
5	개인정보, 회사, 제공, 동의, 게임, 피고, 자, 원고, 수집, 이용자	46	0.1456
6	상고이유, 파기, 촬영, 법, 음란물, 청소년, 상고, 타인, 보호, 동영상	124	0.3924
7	주식회사, 적시, 표현, 비방, 작성, 후보자, 이익, 글, 기사, 허위	62	0.1962
8	국익, 국정원, 지시, 지원, 차장, 단체, 직원, 자금, 전략, 직무	2	0.0063

[그림 4]

Dominant_Topic	Topic_Perc_Contrib	판례일련번호	사건명
3	0.7142	210683	대부업등의등록및금융이용자보호에관한법률위반·정보통신망이용촉진및정보보호등에
3	0.5837	226163	정보통신망이용촉진및정보보호등에관한법률위반(명예훼손)
3	0.3449	168411	마약류관리에 관한 법률 위반(향정)·마약류관리에 관한 법률 위반(대마)·정보통신망 이·
3	0.3016	218561	무고·모욕·명예훼손·정보통신망이용촉진및정보보호등에관한법률위반(명예훼손)

[그림 5]

Dominant_Topic	Topic_Perc_Contrib	판례일련번호	사건명
8	0.7472	214359	아동·청소년의성보호에관한법률위반(음란물제작·배포등)·도박공간개설·정보통신망이용
8	0.6568	216985	아동·청소년의성보호에관한법률위반(음란물제작·배포등)·정보통신망이용촉진및정

[그림 6]

나열한 그림에서 주목한 부분은 주제3 및 주제 8이다. '국가정보원' 및 '국정원'이라는 단어가 지배적으로 드러난다고 표기된 [그림 4]와는 달리 [그림 5, 6]의 사건명을 확인하였을 때 위 두 단어는 자연적으로도 연상이 불가능하며, 명시된 판례에서도 해당 키워드를 찾을 수 없었다.

5. 결론 및 기대효과

비록 정보통신과 직접적으로 관계가 있는 것으로 명시된 판례는 현재 약 317건이며, 관련 키워드를 추가하면 최대 600여개까지 증가한다. 현재 법제처에서 제공하고 있는 판례는 [그림 7]과 같이 제공되어, 제목만으로는 그 내용을 한 눈에 파악하기는 매우 어려우며, 현재 법조계 역시 관련 법 조항을 직접 검색하여 읽는 방식으로 판례를 파악하고 있는 실정이다.

판례일련번호	사건명	사건번호
필터	필터	필터
193337	개인정보보호법위반	2015도16508
204636	정보통신망이용촉진및정보보호등에관한법률위반(명예훼손) (인정된죄명:성폭력범죄의처벌등에관한특례법위반(카메라등이용촬영))	2017도17529
198572	성폭력범죄의처벌등에관한특례법위반(카메라등이용촬영)·정보통신망이용촉진및정보보호등에관한법률위반(음란물유포)	2017노1867
210693	대부업등의등록및금융이용자보호에관한법률위반·정보통신망이용촉진및정보보호등에관한법률위반	2017고단487
209257	정보통신망이용촉진및정보보호등에관한법률위반	2017고단2372
186110	사기[피고인1에대하여인정된죄명:특정경제범죄가중처벌등에관한법률위반(사기)]·개인정보보호법위반·국민체육진흥법위반·사기미수·범죄단체조직(피고인2에대하여인정된…	2017도8600
195388	정보통신망이용촉진및정보보호등에관한법률위반(음란물유포)	2012도13352
185809	공직선거법위반·개인정보보호법위반	2015도12400
210981	정보통신망이용촉진및정보보호등에관한법률위반(정보통신망침해등)	2017노309
206214	정보통신망이용촉진및정보보호등에관한법률위반(정보통신망침해등)	2017노262
185270	특정범죄가중처벌등에관한법률위반(뇌물)·개인정보보호법위반·정보통신망이용촉진및정보보호등에관한법률위반·부정경쟁방지및영업비밀보호에관한법률위반(영업비밀누설…	2017도4240
195468	정보통신망이용촉진및정보보호등에관한법률위반(명예훼손)(인정된죄명:명예훼손)	2017도5122
197622	사기[피고인1에대하여인정된죄명특정경제범죄가중처벌등에관한법률위반(사기)]·개인정보보호법위반·국민체육진흥법위반·사기미수·범죄단체조직(피고인2에대하여인정된죄…	2017노209
185129	정보통신망이용촉진및정보보호등에관한법률위반(명예훼손)	2016고정3950
184700	개인정보보호법위반 · 정보통신망이용촉진및정보보호등에관한법률위반(개인정보누설등)(이른바 '경품 응모권 1mm 글씨 고지' 등 관련 형사사건)	2016도13263
193129	특정범죄가중처벌등에관한법률위반(뇌물)·개인정보보호법위반·정보통신망이용촉진및정보보호등에관한법률위반·부정경쟁방지및영업비밀보호에관한법률위반(영업비밀누설…	2016노2667
204528	성폭력범죄의처벌등에관한특례법위반(카메라등이용촬영)·정보통신망이용촉진및정보보호등에관한법률위반	2016노4877
183967	정보통신망이용촉진및정보보호등에관한법률위반(정보통신망침해등)·개인정보보호법위반	2016도11138
183975	모욕·정보통신망이용촉진및정보보호등에관한법률위반(명예훼손)	2014도15290
197767	배상명령신청·사기[피고인1에대하여인정된죄명특정경제범죄가중처벌등에관한법률위반(사기)]·개인정보보호법위반·국민체육진흥법위반·사기미수·범죄단체조직(피고인2에대…	2016고합203
183259	부당이득금반환(공개된 개인정보를 수집하여 제3자에게 제공한 행위에 대하여 개인정보자기결정권의 침해를 이유로 위자료를 구하는 사건)	2014다235080
203797	개인정보보호법위반	2016노223
182950	시정조치등취소(인터넷 사이트에서 개인정보를 수집하면서 적법한 동의를 받았는지 문제 된 사건)	2014두2638
188224	정보통신망이용촉진및정보보호등에관한법률위반(정보통신망침해등)·개인정보보호법위반	2016노336

[그림 7]

당초의 프로젝트는 정보통신 기술과 관계된 판례를 추출하고, 관련 법 조항 및 관련 기술을 조합하여 정보통신기술분야 종사자로 하여금 법과 판례에 대한 접근성을 끌어올리는 것에 목적을 두었다. 하지만, 정보통신 유관 판례를 분석하여 나온 결론은 정보통신 기술과 관계 있는 판례는 극히 드물며, 사기나 개인정보 침해, 음란물 유통과 같은 범죄행위를 주로 다룬다는 것이다. 따라서 해당 데이터로는 도저히 프로젝트를 진행할 수 없을 것이라고 판단, 결국 막바지에 주제를 선회할 수밖에 없었다.

따라서, 당초의 프로젝트에 장애가 생긴 것이 곧 해당 프로젝트를 진행하게 된 이유다. 이 프로젝트를 통해 유의미한 결과를 거두지는 못했지만, 특정 분야에 특화된 판례를 분석하는 모델을 개발한다면, 별도의 법무팀을 구성할 수 없는 소규모 사업자에게 법적 분쟁을 대비하거나[5] 이를 예방하는 수단으로서의 활용도를 크게 높일 수 있을 것이라 전망한다.

6. 참고문헌

1) 「개인정보위, '이루다' 개발사 ㈜스캐터랩에 과징금·과태료 등 제재 처분」, 『개인정보보호위원회』, 2021년 4월 28일.

2) 「개인정보 유출로 인한 손해배상책임의 최근 동향」, 『보안뉴스』, 2016년 7월 7일.

3) 「온라인 기업 '판례 검색 사이트' 본격 경쟁 돌입」, 『법률신문』, 2022년 5월 12일.

4) 최현호·김재웅·이영구·한용구, 「단어의 의미적 유사도를 응용한 암호 생성기」, 『2022년 한

국컴퓨터종합학술대회 논문집』, 한국정보과학회, 2022.

5) 심준식·김형중, 「LDA 토픽 모델링을 활용한 판례 검색 및 분류 방법」, 『전자공학회논문지』 제54호, 대한전자공학회, 2017.