[DOI] http://dx.doi.org/10.21487/jrm.2022.3.7.1.61

【연구경향】

빅데이터 시대에 맞춘 국회 사용법: 자동화된 국회 데이터베이스

엄기홍*·김대식**·서대원***·강원식****·육태훈*****

논문요약

국회 자료의 풍부함과 필요성에도 불구하고, 국회 자료를 사용하는 데에는 상당한 비용과 시간이 요구된다. 이러한 한계로 말미암아 국회 연구는 시간적으로 단발적인 그리고 공간적으로 제한되어 이루어져 왔다. 본 연구의 목적은 학계 등과 공유할 수 있는 자동화된 국회 데이터베이스를 구축함으로써 국회 자료 사용의 시간적·공간적 한계를 극복하는 데 있다. 또한 본 연구에서 소개한 자료 수집 및 정제 방식, 데이터베이스 구축 방식, 그리고 자동화 방식은 선거관리위원회, 통계청 등과 같이 다량의 데이터를 생성하는 기관에 적용할 수 있는 논리로써 데이터베이스의 확장성을 기할 수 있다. 따라서 본 연구의 시도는 국회를 시작으로 사회과학 전반으로 확장되어 적용될 수 있을 것으로 기대한다. 궁극적으로 이와 같은 시도는 언론 및 시민단체 등에 의해 활용되어 한국 민주주의 운용의 투명성과 건전성 확보에 이바지하리라 기대한다.

주제어: 국회 데이터베이스, 회의록, 한글 문서 구문분석기, 스키마, 자동화

* 경북대학교 정치외교학과, 교수, 제1저자
** 경북대학교 정치외교학과, 강사, 교신저자
*** 경북대학교 대학원 정치학과, 석사 졸업생
**** 경북대학교 대학원 정치학과, 석사 과정생
***** 경북대학교 대학원 정치학과, 석사 과정생

Ⅰ. 서론

국회는 다양하고 방대한 자료를 공개하고 있다. 국회는 국회의원에 대한 인적 사항뿐만 아니라 의원실 구성원에 대한 자료를 공개하고 있다. 누가 어떤 법률안을 대표발의했는지, 공동발의자는 누구인지 등을 공개하고 있다. 법률안뿐만 아니라 본회의, 상임위원회, 특별위원회, 소위원회 등에서의 국회의원 발언을 공개하고 있다. 또한 국회의원의 의정활동, 예를 들어 본회의 참석 여부, 법률안 표결 자료 등을 공개하고 있다.

학계, 언론, 시민단체 등은 이와 같은 자료를 활용하여 한국 민주주의가 제대로 작동하고 있는지를 연구하고 감시하고 있다. 예를 들어, 국회의원의 사회경제적 배경에 대한 변화(가상준·유성진·김준석 2008; 전용주 2012), 사회경제적 배경과 의정활동 간의 연관성(문우진 2010; 정회옥·장혜영 2013), 정치 대표성(조진만 2019), 이념이 의정활동에 미치는 영향(강우창 외 2020; 이갑윤·이현우 2011; 이현우 2005a 그리고 2005b; 전진영 2006a 그리고 2006b), 그리고 이러한 요소가 선거 결과에 미치는 영향력 등을 분석하고 있다(유승익·김현우 2003). 또한 국회의원의 상임위원회 활동(손병권 외 2004), 국회의원의 표결 결정요인 등 다양한 차원의 국회 연구가 진행되고 있다(전진영 2006a; Jun and Hix 2010). 이처럼 국회에서 제공하는 자료는 학문적 차원뿐만 아니라 한국의 민주주의 건전성 평가에도 상당히 기여하고 있다(한국 국회와 정치과정에 대한 전반적 논의는 임성호 외 2010).

그러나 국회 자료의 이와 같은 중요성에도 불구하고, 기존 연구는 '지속성의 한계'라는 난관에 부딪히고 있다. 예를 들어, 대부분의 기존 연구는 국회 대수 1회를 분석했으며(유승익·김현우 2003; 윤종빈 2004; 전용주 2012; 전진영 2006a 그리고 2006a; 정회옥·장혜영 2013), 길어도 2회의 대수를 넘기지 못하고 있다(가상준·유성진·김준석 2008; 문우진 2010). 국회의원 표결행태에 관한 연구가 아닌 다른 차원의 연구, 예를 들어 국회의원 발언에 관한 연구는 특정 위원회에 관한 단발성 연구에 머무르고 있었다(최준영·조진만 2013; 최준영·조진만·가상준·손병권. 2008).

국회 연구가 이처럼 제한되는 이유는 국회에서 제공하는 공공자료를 정제·분석하기 위해 상당한 비용과 노력이 필요하기 때문이다. 물론 국회는 "열린국회정보" 사이트를 통해 국회 자료를 공개하고 있으며, 사용자 편의를 위한 Application Programming Interface(이하 API)를 제공하고 있다.[1] 또한 Team Popong은 웹 스크랩핑 기법을 활용하여 국회 자료를 수집·정제하였고, 참여연대의 "열려라 국회"는 국회의원 표결 자료를 정제하여 제공하고 있다. 오마이뉴스의 "국회의원 정치자금"은 정치자금의 수입·지출 보고서를 보기 쉽게 제공하고 있다.[2]

그러나 이와 같은 국회 자료를 제공하는 기관은 저마다 한계에 부딪혀 있다. 예를 들어, Team Popong은 더 이상 정보를 업데이트하지 않고 있으며, "열려라 국회"는 국회 표결마다 개별적으로 그리고 수작업으로 표결 자료를 입력하고 있다. "국회의원 정치자금"은 PDF 버전으로 제공하는 "국회의원의 정치자금 수입·지출 보고서"를 수작업으로 입력하고 있다. 궁극적으로 국회의 "열린국회정보" 사이트에서 제공하는 API의 반환물은 PDF 또는 HWP 포맷으로 제공하고 있어, 학계의 연구나 언론 및 시민단체의 감시를 위해서는 상당한 노력과 비용이 요구된다. 이러한 이유로 국회 자료의 중요성에도 불구하고 한국 민주주의의 건전성을 측정하고 견제하는 데에 상당한 한계가 있다.

본 연구의 목적은 학계를 비롯한 다양한 영역에서 국회 자료를 손쉽게 사용할 수 있는 국회 데이터베이스를 소개하는 데 목적이 있다. 특히, 본 연구에서 소개하는 국회 데이터베이스는 컴퓨터에 의해 자동으로 수집·정제되기 때문에 한국 정치학계가 안고 있는 연구 지속성의 한계를 해소하는 데 이바지하고자 한다. 또한 이와 같은 시도는 궁극적으로 국회 투명성을 강화하기 때문에 한국 국회의 고질적인 문제점인 국회 불신을 해소하는 데 다소나마 기여하리라 본다. 나아가 본 연구가 소개하는 데이터베이스 구성 논리를 선거관리위원회, 통계청 등과 같이 방대한 양의 데이터를 생산하는 기관에게 제공함으로써 기관의 데이터 활용성을 높이는

1) 국회의 "열린국회정보" 사이트는 https://open.assembly.go.kr/이며, 방문일은 2021년 11월 23일이다.

2) Team Popong 사이트는 https://www.popong.com/, 참여연대의 "열려라 국회" 사이트는 http://watch.peoplepower21.org/, 오마이뉴스의 "국회의원 정치자금" 사이트는 http://www.ohmynews.com/NWS_Web/Event/Special/20spf.aspx이다. 방문일은 2021년 11월 23일이다.

데 이바지하고자 한다.

본 연구의 구성은 다음과 같다. 이어지는 제Ⅱ장에서는 국회 자료의 수집·정제에 대해 설명한다. 자료 수집과 관련하여 “열린국회정보”에서 제공하는 Application programming interface(이하 API)에 대한 설명과 API를 통해 국회 자료를 가져오는 방식을 설명한다. 자료 정제와 관련하여 HWP 형식으로 구성된 한글 문서에서 원하는 정보를 추출하는 구문분석(parser) 프로그램과 자료 추출 논리를 소개한다. 제Ⅲ장에서는 국회 데이터베이스 스키마(schema)를 설명한다. 데이터베이스 스키마 구축을 위한 선결과제인 개념적 모델 설계, 논리적 구조 설계 그리고 물리적 구조 설계 과정을 제시하고 있다. 이와 같은 스키마는 제Ⅳ장에서 구현된 자동화한 스케줄러를 통해 일일 단위로 업데이트되어 국회 자료를 수집하고 데이터베이스에 쌓이게 된다. 제Ⅴ장은 학계 등과 공유하는 실제 사이트를 예로 들어 입력 키워드 및 조건, 그리고 이에 따른 반환물을 소개한다. 끝으로 제Ⅵ장에서는 본 연구가 제시하는 국회 데이터베이스의 함의와 추후 과제를 논의해 보고자 한다.

Ⅱ. 국회 자료의 수집과 정제: 회의록을 중심으로

국회 자료의 수집과 정제 단계는 국회 자료, 예를 들어 회의록 자료를 자동으로 수집하고, 이를 정제하여 저장하는 단계이다. 정제된 자료의 저장 형태는 JSON 형식(.json)이다. 이 과정은 수집, 정제, 저장의 단계로 진행되며, 가장 주요한 단계는 정제 단계이다. 작성 언어는 파이썬(python)이다.

국회 자료 수집은 “공공데이터포털”(data.go.kr)에서 제공하는 API를 활용한다. 예를 들어, 공공데이터포털은 국회의원 정보, 의안 정보, 회의록 정보 등을 제공하며, 각각의 API는 공공데이터포털 검색창에 “국회 국회사무처_의원 정보”, “국회 국회사무처_의안 정보”, “국회 국회사무처_회의록 정보” 키워드를 입력하여 획득할 수 있다. API를 통해 반환된 자료는 XML 형식(.xml)이다. 정제 단계란 수집 단계에서 반환된 XML 형식의 문서를 JSON 형식으로 변환하는 작업, 변환된 JSON 문서에서 한글 문서를 수집하는 작업, 그리고 한글 문서를 텍스트 또는

JSON 문서로 변환하는 작업을 일컫는다. 끝으로 저장 단계는 데이터베이스로 파일을 넘기기 이전 임시로 저장하는 형태이며, JSON으로 저장한다.

공공데이터포털 API를 통해 국회 자료를 요청할 때는 두 가지 정보가 필요하다. 하나는 공공데이터포털 회원가입 이후 활용신청을 통해 받는 API 키값, 다른 한 가지는 요청하는 정보이다. 요청하는 정보는 API 주소와 API가 요구하는 파라미터의 키-값 쌍으로 구성되어 있다. "국회 국회사무처_의원 정보" API를 통해 국회의원 정보를 가져오는 방식들을 설명하면 다음과 같다.

예제)

http://apis.data.go.kr/9710000/NationAssemblyInfoService/getMemberCurrStateList?numOfRows=100&ServiceKey=****

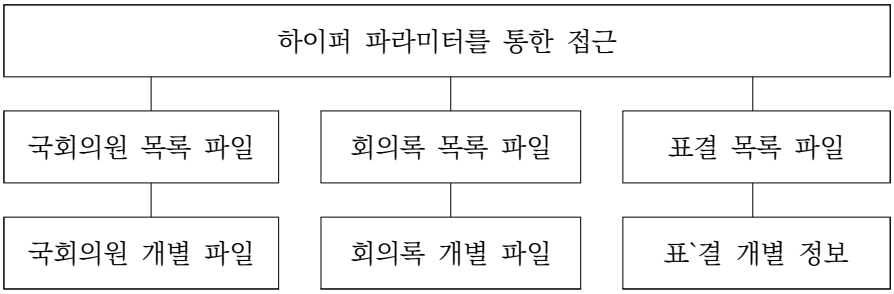
- http://apis.data.go.kr/: 공공데이터 포털 API 기본 URL
- 9710000: API 제공자에 할당된 부호, 9710000은 국회사무처를 의미
- NationalAssemblyInfoService: 국회사무처가 지정한 대분류
- getMemberCurrStateList: 국회사무처가 지정한 소분류(국회의원 정보)
- ?: 주소가 끝나고 파라미터가 이어진다는 부호
- numOfRows=100: 키=값, 예제에서는 100개 행(즉 100명)을 요청
- &: 파라미터 간의 구별자, &은 그리고를 의미함
- ServiceKey=****: 키=값, API 키값은 가림

예제)

http://apis.data.go.kr/9710000/NationAssemblyInfoService/getMemberDetailInfoList?num=***&dept_cd=****&ServiceKey=**********

- getMemberDetailInfoList: 국회사무처가 지정한 소분류(국회의원 세부정보)
- num: 국회의원 개인 고유값
- dept_cd: 국회의원 개인 고유값

API 키값은 비밀번호와 같아서 외부에 공개하기 어렵다는 점, 그리고 API를 통해 자료를 요청할 때마다 고유값(unique value, 예) num, dep_cd)을 입력해야 하는 번거로움이 있다. 이와 같은 어려움을 해소하기 위해 본 연구는 API 키값과 고유값 등과 같은 하이퍼 파라미터(Hyper parameter)를 별도의 파일(즉 "목록 파일")에 담고, 이를 헤더 파일(header.py)을 통해 사용하였다. 즉 [그림 1]에 제시된 바와 같이, 국회의원에 대한 고유값은 "국회의원 목록 파일"에 담았고, 회의록 각각에 대한 고유값은 "회의록 목록 파일"에 담았고, 국회의원 표결 자료는 "표결 목록 파일"에 담았다. 이와 같은 방식은 API를 기반으로 한 국회 자료의 수집 자동화을 가능하게 하는 장점이 있다.



[그림 1] 자료 수집 구조

[그림 1]에 제시된 목록 파일에는 각 파일의 요소에 맞는 기본 정보들이 저장되어 있다. 예를 들어, "국회의원 목록 파일"에는 국회의원은 성명, 정당, 소속 상임위원회와 기간 등이 저장되어 있고, "회의록 목록 파일"에는 회기, 회의 형태, 대수 등이 저장되어 있다. 그리고 "표결 목록 파일"에는 전체 표결의 목록에 법안명과 법안 코드가 저장되어 있다. 요컨대 본 연구의 자료 요청 방식은 각 목록 파일에 담긴 고유값을 통해 개별 AP에 접근하는 구조로 작성하였다.

예를 들어, 회의록을 가져오는 목록 파일인 "회의록 목록 파일"의 일부분은 [그림 2]와 같다. [그림 2]에서 API 키값은 생략되었으며, 고유값은 "body" 하단의 "051327"이다. 이 고유값에 대한 부연 설명은 "051327"에 딕셔너리

형태로 저장되었으며, 예를 들어 “daenum”은 대수(21), “comName”은 위원회명(과학기술방송통신위원회) 등이다.

```
{
    "head": {
        "Rupdate": "2021-10-11_12:31:56",
        "Type": "LIST",
        "Path": "내부 경로"
    },
    "body": {
        "051327": {
            "daenum": "21",
            "comName": "과학기술정보방송통신위원회",
            "conferType": "(정기회)",
            "period": "제391회(2021.09.01-2021.12.09)",
            "order": "제4차(2021년10월05일)",
            "hwplink": "http://likms.assembly.go.kr/record/mhs-10-040-0040.do?conferNum=051327&fileType=hwp",
            "pdflink": "http://likms.assembly.go.kr/record/mhs-10-040-0040.do?conferNum=051327&fileType=PDF",
            "isPdfHere": true,
            "isHwpHere": true,
            "isTxtHere": false,
            "isJsonHere": false
        },
```


[그림 2] 회의록 목록 파일 일부

[그림 2]에서 주목할 점은 “회의록 목록 파일”에 실제 회의록 문서를 다운로드할 수 있는 링크가 있다는 것이다. 즉 “hwplink”에는 회의록 한글 문서가, “pdflink”에는 회의록 PDF 문서가 실재하는 링크가 담겨 있다. 따라서 본 연구는 “회의록 목록 파일”에 있는 회의록 개별 파일, 특히 한글 문서에 접근하여 회의록 문서를 수집했다. 수집된 한글 문서를 토대로 JSON 형식으로 전환한다.

한글과 컴퓨터에서 배포하는 파일 형식은 HWP 3.0, HWP 5.0, HWPX 등이며, 국회 사무처에서 제공하는 한글 문서 회의록은 HWP 5.0 형식이다.[3] HWP 5.0 텍스트 형식은 리틀 엔디언을 기본으로 zilb 압축을 채택하고 있으며,[4] 그 안의 텍스트는 UTF-16LE 형태이다.[5] 따라서 본 연구는 파이썬의 zilb 모듈을 통해

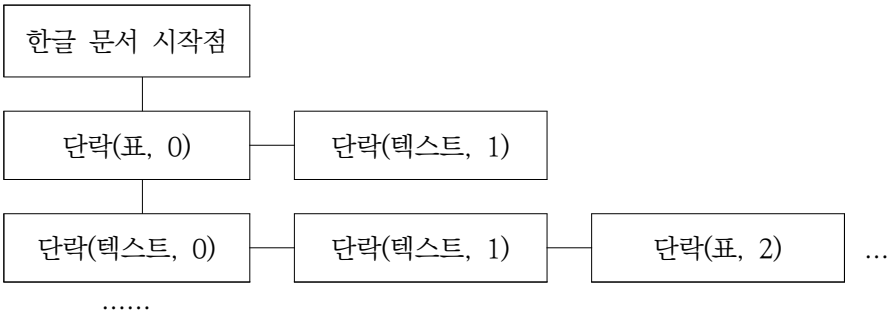
3) HWP 5.0에 대한 기술문서는 한글과 컴퓨터 홈페이지 “HWP 파일포맷” 기술문서를 참조했다(https://www.hancom.com/etc/hwpDownload.do, 방문일: 2022년 1월 4일).

4) 오픈소스 압축 방식으로 범용적으로 사용하고 있는 압축 방식이다(참고: https://zlib.net, 방문일: 2022년 1월 4일).

zilb 압축을 풀었고, 바이트 단위로 한글 문서에 접근했다.

본 연구는 HWP 5.0에 내재한 단락 단위의 속성을 이용하여 문서를 정제했다. 즉 HWP 5.0은 단락(Paragraph) 단위로 속성이 부여되어 있는데, 이를 이용하여 텍스트와 기타 객체(그림, 표, 선 등)가 들어있는 단락을 파악한 후 회의록 텍스트를 추출했다. 구체적으로 보면, 각 단락의 헤더 부분(처음 32bit)에는 각 단락의 타입과 레벨, 길이를 알 수 있는 정보가 담겨 있다. 특히 레벨의 경우, 레벨이 높은 단락은 낮은 단락의 하위 수준임을 의미하고, 이는 곧 한글 문서구조가 일종의 트리 구조임을 의미한다. 따라서 단락의 구조는 단락 헤더에서 제공하는 크기와 헤더의 길이(32bit)씩 가산하는 방식으로 단락 단위로 분리를 수행할 수 있다(한글과컴퓨터, 2018).

[그림 3]에서 제시한 바와 같이, 각 단락의 타입을 파악한 후, 타입이 텍스트에 해당할 경우, 해당 단락의 텍스트를 추출하였다. 예를 들어, “단락(텍스트, 0)”은 텍스트이기 때문에 추추출할 단락에 해당한다. 따라서 본 연구는 “단락(텍스트, 0)” 중 텍스트에 해당하는 부분, 즉 “단락(텍스트, 0)”과 “단락(텍스트, 1)”의 정보를 추출한다.[6] 이러한 과정을 통해 회의록에 있는 의원 발언 모두를 추출했다.



[그림 3] 한글 문서 내부 구조

5) 32bit(4bytes)의 길이를 가진 유니코드(Unicode) 인코딩 방식이다.
6) 본 연구가 사용하는 기법으로 표에 해당하는 부분을 가져올 수도 있다.

2021년도 7월 18일 농림축산식품해양수산위원회 회의록 일부를 예로 들어 회의록 정제 작업을 설명하고자 한다. 예시의 회의록 원본은 [그림 4]와 같다. 국회 회의록은 “회의 시작”, “회의 본문”, “회의 이후” 세 부분으로 구분되며, 이를 기준으로 회의록을 구분하여 처리했다. “회의 시작” 부분에는 회의 정보가 표에 기재되어 있다. 현재까지 처리된 시작 부분의 표 유형은 7가지이며, 이에 따라 정제하였다

제389회국회 (임시회) **농림축산식품해양수산위원회회의록** 제 1 호

국 회 사 무 처

일 시 2021년7월13일(화)

장 소 농림축산식품해양수산위원회회의실

의사일정

1. 2021년도 제2회 추가경정예산안(의안번호 2111282)
   가. 농림축산식품부 소관
   나. 해양수산부 소관

상정된 안건

1. 2021년도 제2회 추가경정예산안(의안번호 2111282) ---------- 1
   가. 농림축산식품부 소관
   나. 해양수산부 소관

(10시12분 개의)

○위원장 이개호 위원님들, 자리를 정돈해 주시기 바랍니다.

성원이 되었으므로 제389회 국회(임시회) 제1차 농림축산식품해양수산위원회를 개의하겠습니다.

회부된 법률안 등 보고사항은 위원님들께 배부해 드린 유인물을 참고해 주시기 바랍니다.

(보고사항은 끝에 실음)

본격적인 회의에 들어가기에 앞서 간단히 한 말씀 드리도록 하겠습니다.

여러 언론 보도에 따르면 최근 외국인 근로자의 숙소와 관련해서 그 설치기준이 강화되고 있는 것으로 파악이 되고 있습니다. 이는 외국인 근로자의 처우를 개선하는 데 그 목적이 있겠습니다마는 우리 농촌 현장에서는 많은 어려움을 토로하고 있는 것도 사실입니다. 기존에 허용되어 오던 농지에 설치된 기숙사가 인정이 되지 않음에 따라서 시설철거 농가 등에 많은 부담이 발생할 것으로 예상되기 때문입니다.

또한 새로운 숙소를 마련하는 데 적지 않은 시간과 비용이 소요됨에도 불구하고 시설을 갖추지 못할 경우 외국인 노동자 배정 대상에서 배제되어 영업을 영위하는 데에도 큰 어려움이 발생할 것으로 우려가 되고 있습니다.

농식품부에서는 이러한 현장의 어려움을 잘 살피셔서 적절한 해결책이 조속히 마련될 수 있도록 각별히 노력해 주시기를 당부드립니다.

다음으로 독도입도지원센터와 관련하여 짧게 한 말씀 더 드리겠습니다.

도쿄올림픽조직위원회에서 독도를 일본의 영토로 표기하는 등 일본의 영유권 도발행위가 계속되고 있습니다. 이에 일본의 도발행위를 규탄하는 결의안이 지난 본회의에서 여야 합의로 채택이 된 바 있습니다.

그럼에도 불구하고 해양수산부의 독도입도지원센터 건립사업은 계속해서 지연이 되고 있습니다. 우리 땅 독도를 찾고 지키려는 국민들의 의지와 관심이 매우 커져 올해 독도를 찾는 방문객이 전년에 비해 두 배 가까이 증가했다는 사실을 고려하여 해양수산부는 입도지원센터 건립사업이 조속히 추진될 수 있도록 특별한 노력을 해 주실 것을 당부드리겠습니다.

그리고 최근 코로나19가 크게 확산이 되고 있습니다. 오늘 회의에 참석하여 주신 분들께서는 마스크 착용 등 방역수칙을 보다 더 철저하게 준수해 주실 것을 부탁드리겠습니다.

그러면 회의에 들어가도록 하겠습니다.

---

1. 2021년도 제2회 추가경정예산안(의안번호 2111282)

[그림 4] 농림축산식품해양수산위원회 회의록 일부

“회의 본문” 부분의 특징은 각 발언의 시작에 ◯가 포함되어 있다는 것이다. 따라서 ◯ 문자를 기준으로 발언의 시작 여부를 구별할 수 있다. 예를 들어, “◯위원장 이개호 위원님들, ...”는 새로운 발언 시작을 의미한다. 하나의 회의 주제가 끝난 경우, ‘선’(――――――) 객체가 추가되어, 이를 기준으로 새로운 회의 주제 시작 여부를 구별할 수 있다. 예를 들어, “1. 2021년도 제2회 추가경정예산안(의안번호 2111282)”의 바로 위에 있는 선(――――――――)으로 회의 주제를 구분할 수 있다.

본 연구에서 개발한 국회 데이터베이스는 해당 발언이 국회의원의 발언인지 여부를 구분하여 국회의원의 발언만을 추출한다. 즉 본 연구는 전술한 “국회 국회사무처_의원 정보” API를 통해 수집한 국회의원 정보에서 고유값인 “dept_cd” 등을 이용하여 국회의원마다 새로운 고유값을 만들어 국회의원과 발언을 1:1로 배정했다. 다만 동명이인의 국회의원이 존재할 경우, 회의록 표기법(즉 성명이 한자로 표기되거나 선출방식이 병기)에 맞추어 사전 코드화를 실시했다.

“회의 이후” 부분의 경우 출결 정보를 제외할 때, 들여쓰기 수준에 따라 표기 정보가 다르다는 점을 활용하여 정리하였다. 해당 회의에서 처리한 법안, 제안서 등의 대표발의자와 공동발의자, 발의일, 소관 상임위원회 등을 판단할 수 있도록 정리하였다.

한글 문서 회의록에 대한 정제 작업을 거친 JSON 변환 결과는 [그림 5]와 같다. 회의 시작 부분은 “head”에, 회의 본문 부분은 “body”에, [그림 5]에 제시되어 있지 않으나 회의 이후 부분은 “tail”에 담겨 있다.

[그림 5] JSON으로 변환된 농림축산식품해양수산위원회 회의록 일부

"head" 부분에는 회의 의사일정인 "1. 2021년도 제2회 추가경정예산안(의안번호 2111282)", "가. 농림축산식품부 소관", "나. 해양수산부 소관" 정보 등이 담겨 있다. "body"에는 "Type"에 따라서 개의 시각, 지위, 의원 고유번호, 의원 발언 등 다양한 정보가 순서대로 저장되어 있다. 예를 들어, 회의 개의 시각은 "body"의 첫 번째 요소에 있는 "Content"에, "body"의 두 번째 요소에는 코드화한 의원 고유번호(Name), 발언자의 지위(Position), 발언(Content)에 담겨 있다. 따라서 "Name" 정보를 활용하여 국회의원에 대한 세부 정보와 해당 국회의원의 발언에 쉽게 접근할 수 있다.

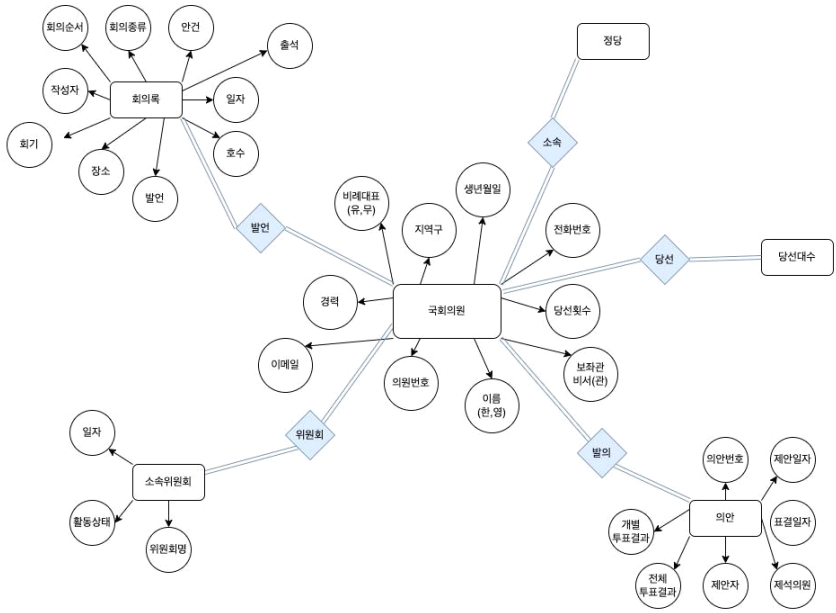
정제된 회의록 한글 문서는 JSON 형태로 임시 저장되어 다음 단계인 데이터베이스 저장 절차로 넘겨진다.

Ⅲ. 국회 데이터베이스 스키마 모델링

국회에서 제공하는 자료는 '국회의원-회의록', '국회의원-입법안' 등 데이터 간 관계가 명확하고, 데이터베이스 구축 시 데이터 간 관계를 사전에 정의할 수 있는 형태로 제공하고 있다. 또한 국회 일정에 따라 순차적으로 수집할 수 있는 형태로 제공하고 있다. 따라서 본 연구는 다양한 데이터베이스 유형 중 설계자가 정형 데이터를 저장하기에 용이하고, 인덱싱(indexing)과 조인(join) 기능을 통해 사용자가 대량 데이터를 쉽게 가져올 수 있는 관계형 데이터베이스(RDB: Relational DataBase)를 사용했다. 또한 관계형 데이터베이스를 쉽게 관리할 수 있도록 도움을 주는 관계형 데이터베이스 관리시스템(RDBMS: Relational DataBase Management System) 중 Array, JSON 등 다양한 자료구조를 지원하는 PostgreSQL을 채택했다.

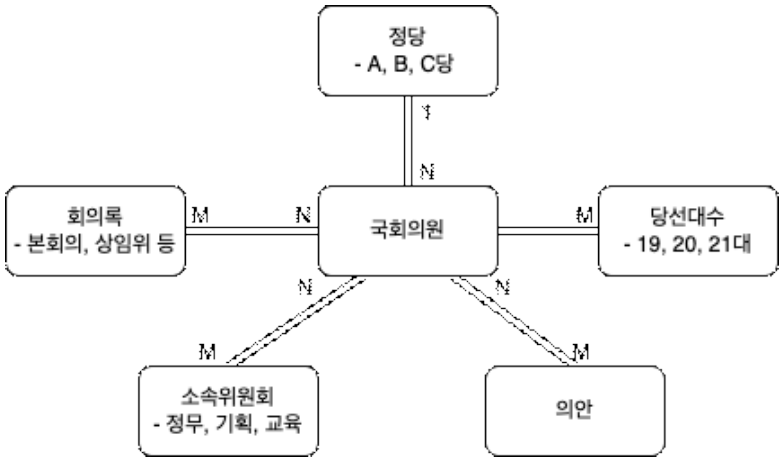
데이터베이스 구축 과정은 크게 기획, 설계, 구현, 운영 및 유지 보수 단계로 구성된다(한국데이터베이스진흥센터 1999). 제Ⅲ장은 이 중 설계와 구현 단계에 초점을 두어 살펴보고자 한다. 먼저 설계 단계는 개념적 모델 설계, 논리적 구조 설계, 물리적 구조 설계 과정으로 구분한다. 개념적 모델 설계는 표현하고자 하는 데이터 정보를 개체-관계(Entity-Relationship, 이하 E-R) 모델을 통해 단순화하고 개념화하는 과정을 일컫는다. 후술하는 바와 같이 국회 데이터베이스는 '국회의원'이란 개체, '정당'이란 개체, '회의록'이란 개체 등이 존재하며, 이들 간의 관계가 존재한다. 논리적 구조 설계는 개념적 모델 설계로 정의했던 개체를 데이터베이스 유형에 맞게 변환(mapping)하는 단계로, 1:N(일대다), N:M(다대다) 등과 같은 관계 설정과 데이터의 원자성(atomic value), 유일성(unique value) 등을 통해 정규화(normalization)를 확보하는 과정이다. 정규화는 테이블 간에 중복된 데이터를 허용하지 않기 위해 진행하며, 테이블의 컬럼(column)이 하나의 값을 갖도록 분해하는 원자성, 테이블의 레코드(record)가 서로 구별할 수 있는 값을 갖도록 하는 유일성을 통해 확보될 수 있다. 끝으로 물리적 구조 설계는 사용하는 데이터베이스 관리시스템의 특성과 성능을 고려해 실제 데이터베이스에 저장될 구성과 구조를 정의하는 것을 말한다.

본 연구에서 구성하는 데이터베이스 스키마 중 일부를 들어 설명하고자 한다. [그림 6]은 데이터 정보를 개체(entity)와 개체 간의 관계(relationship), 개체가 가지는 속성(attribute) 등을 포함해 E-R 다이어그램으로 표현한 개념적 모델 설계이다. 일반적으로 E-R 다이어그램에서 개체는 사각형, 개체의 속성은 타원형, 관계는 마름모형으로 나타낸다. [그림 6]에서 개체는 '국회의원', '정당', '당선대수', '의안', '위원회', '회의록'이며, 개체 각각은 각자의 속성을 가지고 있다. 예를 들어, '국회의원' 개체는 '의원번호', '지역구', '경력' 등과 같은 속성들을 가지고 있다. 개체 간의 관계는 마름모형으로 표시한다. 예를 들어, '국회의원' 개체와 '의안' 개체 사이는 '발의'의 관계로 연결되어 있고, '국회의원' 개체와 '소속위원회' 개체 사이는 '소속'의 관계로 연결되어 있다. 이와 같은 관계는 1차원 또는 N차원의 관계를 형성할 수 있다. 예를 들어, '국회의원'은 '의안', '소속위원회', '회의록', '정당', '당선대수'와 관계를 갖는 5차원 관계를 형성하고 있지만, '당선대수'는 '국회의원'과만 관계를 갖는 1차원 관계를 형성하고 있다.



[그림 6] 국회 데이터베이스 E-R 다이어그램

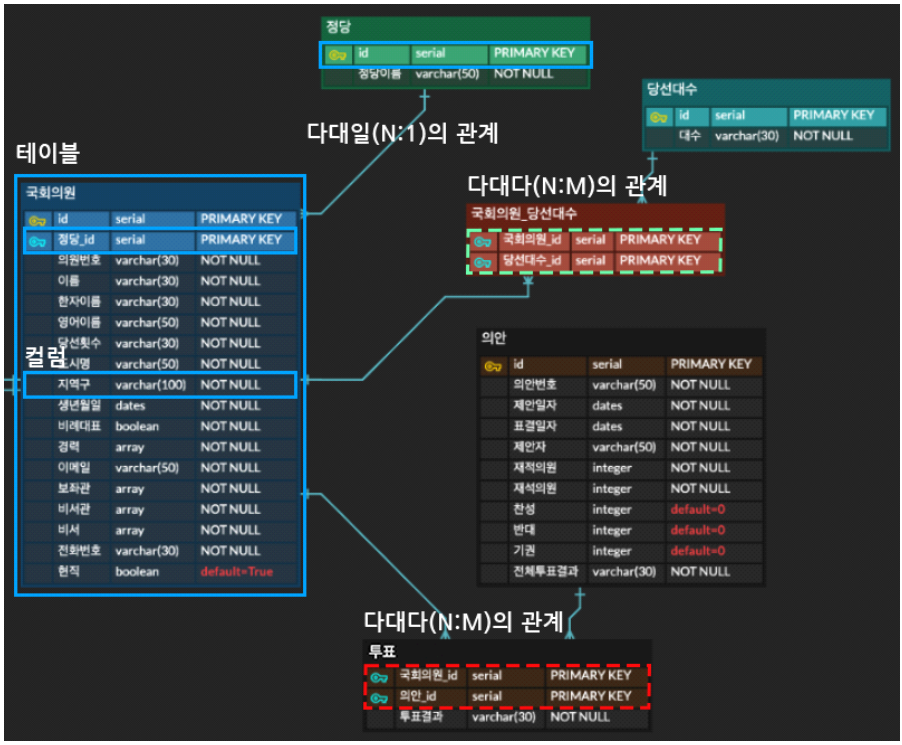
개체 간의 관계는 관계형 데이터베이스 모델 설계에 있어 중요한 요소이다. 따라서 [그림 7]에 제시된 바와 같이, '국회의원' 개체를 통해 관계를 더욱 자세히 살펴보고자 한다. 국회의원은 무소속을 포함한 오직 1개의 정당에 소속될 수 있다. 반면 정당에는 다수 국회의원이 포함될 수 있다. 따라서 '국회의원' 개체와 '정당' 개체 간의 관계('소속')는 일대다(1:N)이다.



[그림 7] 국회 데이터베이스 개체별 관계도

한편 '국회의원' 개체와 '당선대수' 간의 관계('당선')는 다대다(N:M)의 관계를 한다. 다선의원인 국회의원은, 예를 들어 제19대 초선, 제20대 재선, 제21대 3선 등 여러 당선대수와 관계를 가지며, 당선대수는 대수별 최대 국회의원 300명과 연결되어 있기 때문이다. 마찬가지로 [그림 7]의 '발의' 관계, '위원회' 관계, '발언' 관계도 다대다(N:M)의 관계를 맺는다.

논리적 구조 설계 과정은 앞서 설계한 개념적 모델의 '개체', '속성', '관계'를 관계형 모델의 '테이블', '컬럼', '키' 등으로 변환하는 과정이다. 예를 들어, [그림 8]은 [그림 7]에서 정의한 '국회의원' 개체를 논리적 구조 설계로 변환한 결과이다. [그림 8]의 '국회의원' 테이블에는 '속성'이 컬럼으로 정의되어 있고, 개체 간의 '관계'는 외래키(즉 정당_id)로 정의되어 있다. 관계형 데이터 모델에서는 컬럼 값이 단일 값으로 존재해야 하며(즉 원자성), 행의 집합인 레코드는 테이블 내에 중복하는 경우가 없도록 유일성이 보장되어야 한다(즉 id). 이를 위해 테이블 별로 정규화를 거쳐 최소 단위로 테이블을 구분하고, 테이블 간의 관계를 기본키(즉 id)와 외래키(즉 정당_id)를 통해 설정했다.



[그림 8] 국회 데이터베이스 스키마 세부 설명

예를 들어, [그림 8]에서 테이블 간의 관계는 테이블명_id(예를 들어, 정당_id, 국회의원_id, 당선대수_id) 이름인 키를 통해 표현하고 있다. 즉 다대일(N:1) 관계인 '소속' 관계는 국회의원 테이블 내 정당_id이란 외래키로 관계를 설계했다. 한편 다대다(N:M) 관계는 다대일(N:1) 관계처럼 테이블 내에 관계를 표현할 수 없다. 따라서 각각의 테이블 id 값을 컬럼값으로 하는 중간 테이블을 생성해서 관리한다. 예를 들어, '국회의원' 개체와 '당선대수' 간의 관계인 '당선' 관계는 다대다(N:M) 관계이다. 다대다 관계에서는 국회의원 테이블 또는 당선대수 테이블 내에 다른 테이블을 연결하는 외래키를 직접 삽입할 수 없어서 중간 테이블('국회의원_당선대수')을 생성하여 두 개체 간의 관계를 표현한다. 같은 방식으로 '발의' 관계를 표현하기 위해 중간 테이블인 '투표'

테이블을 생성하여 '국회의원' 개체와 '의안' 개체 간의 관계를 표현했다.

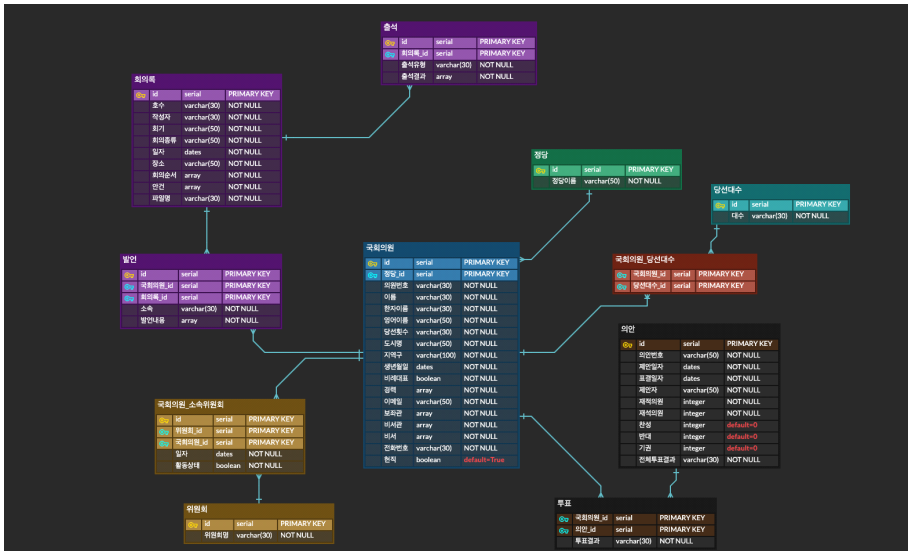
논리적 구조 설계가 데이터베이스에 물리적으로 구현되는 방식을 살펴보자. [그림 9]는 다대일(N:1)인 '소속', 다대다(N:M) 관계인 '당선'이 데이터베이스에 물리적으로 구현되는 방식이다. 국회의원 테이블 내 정당_id 값은 정당 테이블 내 id 값을 참조하는 방식으로 국회의원의 소속 정당을 표현하고 있다. 예를 들어, 국회의원 테이블에서 김00의원은 정당_id 값 1을 갖고 있으며, 이는 정당 테이블 내 id 값과 연결되어 김00의원이 A당 소속임을 표현하고 있다.

[그림 9] 물리적 데이터베이스 내 데이터 저장방식 예시

다대다(N:M) 관계인 '당선'은 중간 테이블인 '국회의원_당선대수'를 통해 관계를 표시하고 있다. 예를 들어, 국회의원 테이블의 id 값이 1인 김00의원은 국회의원_당선대수 테이블의 국회의원_id 1 값과 연결되어 있다. 같은 방식으로 당선대

수 테이블의 id 값 1은 21대 대수인 것을 의미하며, 이는 국회의원_당선대수 테이블의 당선대수_id 1 값과 매핑되어 있다. 따라서 국회의원_당선대수 테이블에서 국회의원_id와 당선대수_id는 한 명의 국회의원이 어떤 대수와 연결되는지를 보여준다. 예를 들어, 김00의원(국회의원_id 값 1)은 21대와 20대(당선대수_id 값 1과 2) 의원, 이00의원(국회의원_id 값 2)은 21대(당선대수_id 값 1) 의원, 박00의원(국회의원_id 값 3)은 21대와 19대(당선대수_id 값 1과 3) 의원인 것을 알 수 있다.

본 연구는 개념적 모델의 '개체', '속성', '관계'를 관계형 모델의 '테이블', '컬럼', '키' 등으로 변환하고 정규화 과정을 거쳐 [그림 10]과 같은 국회 데이터베이스 스키마를 최종적으로 설계했다.



[그림 10] 국회 데이터베이스 전체 스키마

본 연구는 [그림 10]에서 제시한 데이터베이스 스키마를 바탕으로 물리적 데이터베이스를 구축했다. 일반적으로 관계형 데이터베이스는 SQL(Structured Query Language)이라는 언어를 통해 물리적 데이터베이스

를 구축한다. 본 연구에서 구축하는 국회 데이터베이스는 Django 백엔드 프레임워크를 사용하여 서버를 구축하므로, Django 프레임워크에서 제공하는 ORM(Object Relation Mapping) 기능으로 SQL 문법을 자동 생성하여 데이터베이스를 구축했다.

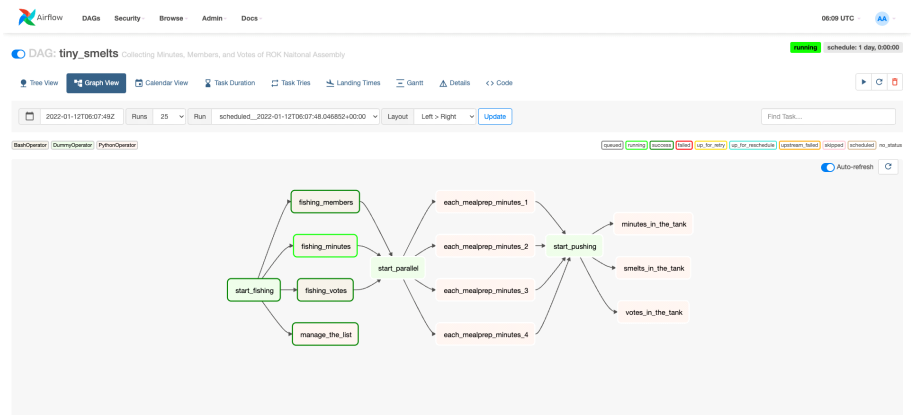
IV. 국회 데이터베이스 자동화 시스템

본 연구에서 개발한 국회 데이터베이스는 수집, 정제, 저장의 단계를 거친다. 제II장과 제III장에 기술한 방법으로 원자료에 대한 수집, 정제, 저장이 이루어지고 체계화와 통합화 단계를 거쳐 최종 사용자가 다양한 방식으로 국회 자료를 소비할 수 있도록 했다.

본 연구의 목적은 국회가 생산한 자료를 지속해서 손쉽게 사용할 수 있도록 하는 시스템을 구축하는 데 있다. 이를 달성하기 위해 반드시 해결해야 하는 요소는 자동화이다. 국회는 끊임없이 데이터를 생산한다. 법률안을 비롯한 의안, 회의록, 표결, 그리고 국회의원 정보 등 다양한 자료가 존재하고 정보의 양 또한 지속적으로 증가하고 있다. 예를 들어, 제19대 국회에 접수된 의안은 총 17,822개이고, 제20대 국회에서는 총 24,141개의 의안이 접수되었다. 제21대 국회는 2022년 1월 13일 현재 14,000개 의안이 접수되었다. 한편 국회의원의 소속 상임위원회 정보 등과 같이 변동 이력이 누락된 경우가 있다. 즉 일정한 간격으로 변동하는 국회의원의 소속 상임위원회 정보는 게재 시점의 소속 상임위원회 정보는 제공하나 변경된 이력을 제공하지는 않는다. 따라서 이러한 형태의 정보를 추적하여 축적하기 위해서는 국회 데이터베이스를 자동화하여 일정한 간격으로 업데이트할 필요가 있다.

본 연구의 데이터 처리 자동화는 수집, 정제, 저장을 위해 개별적으로 작성된 컴퓨터 프로그램을 통합해 개별 작업이 논리적 순서에 따라 단절 없이 작동하는 것을 의미한다. 그리고 완성된 작업 흐름이 미리 설정한 일정에 따라 반복적이고 지속적으로 실행한다. 이러한 작업을 가능하게 하는 플랫폼으로 아파치 에어플로우(Apache Airflow)가 존재하고, 이를 활용해 본 연구는 국회 데이터베이스 자동화

를 완성하였다. 에어플로우는 복잡한 작업의 흐름을 제어하고 관리하기 위해 만들어진 오픈소스 프로그램이다. 에어플로우는 유향(有向) 비순환 그래프(Directed Acyclic Graph)를 사용해 작업 간의 관계를 설정하고 실행하게 한다(Apache Airflow Documentation, n.d.). [그림 11]은 에어플로우 프로그램으로 만든 데이터 처리 작업(특히 회의록)을 그래프로 형태로 표현한 것이다.



[그림 11] 에이플로우로 구현한 국회 데이터베이스 자동화

[그림 11]에서 제시하듯, 국회 데이터베이스 자동화에는 3개의 구분된 단락(“start_fishing”, “start_parallel”, “start_pushing”)이 존재하며, 각 단락은 수집, 정제, 저장 작업을 수행한다. 각 단락에서 네모 상자로 표현한 소작업은 논리적 연관성에 따라 화살표로 연결되고, 실행 성공 여부에 따라 연결된 다음 작업이 실행되게 된다. 예를 들어, “start_fishing” 단계에서 국회 회의록 수집 관련 모든 작업이 완료한 이후, 다음 단계로 새로운 회의록에서 필요한 정보를 추출하고 데이터베이스 저장에 적합한 형태로 데이터를 변형하는 정제 단계(“start_parallel”)를 실행한다.

정제 부분은 데이터베이스 구축에서 컴퓨팅 리소스를 가장 많이 필요로 하기 때문에 생산성 향상을 위해 병렬처리를 적용하여 진행한다. 모든 회의록의 정제 작업이 완료되면 필요한 정보가 데이터베이스 설계에 맞게 저장된다

(“start_pushing”). 데이터를 저장하는 단계에서 의도하지 않은 자료의 손실과 오염에 대비하기 위해 원자료와 정제된 파일을 독립적으로 저장한다. 이와 같은 작업 흐름은 원자료, 정제된 파일, 데이터베이스 입력을 마침으로써 마무리된다. 같은 방식으로 국회의원 정보, 의안 정보, 표결 자료 등이 확인되고 수집·정제·저장된다.

수집부터 저장까지의 개별 업무를 하나의 작업 흐름으로 묶어 논리적으로 수행할 수 있게 하는 만드는 것이 첫 번째 작업이면, 정해진 시간에 반복해서 작업 흐름을 동작시키는 것이 두 번째 작업이다. 에어플로우는 작업에 실행 시간을 설정할 수 있다. 본 연구의 경우 첫 시작을 의미하는 가작업(Dummy Task)을 지정하고 매일 동일한 시간에 실행되도록 설정하였다. [그림 11]에 나타나듯 모든 작업은 가장 왼쪽에 위치한 가작업에 연결되어 있다. 즉, 흐름이 시작되면 논리적 연결에 따라 개별 작업이 순차적으로 실행되고 마지막 작업을 마친 후 다음 실행 시간을 기다리는 형태이다.

에어플로우를 이용한 작업 흐름 관리는 몇 가지 장점을 가지고 있다. 첫째, 잠재적 오류에 유연하게 대처할 수 있다. 에어플로우는 개별 작업의 성공 여부와 내용을 기록으로 남긴다. 그리고 논리적으로 연결되어 있는 작업에서 오류가 발생할 시 다음 작업을 실행하지 않는다. 즉, 저장 과정의 오류를 제외한 잠재적 데이터 손실 및 오염의 위험을 방지할 수 있게 한다. 둘째, 작업 흐름을 유연하게 확장할 수 있다. 에어플로우는 파이썬으로 작성되어 있는 프로그램으로 본 연구에서 사용하는 다른 개별 프로그램에 쉽게 덧붙일 수 있다. 그리고 독립된 프로그램으로 만들지 않고 파이썬 함수 또는 클래스를 활용해 작업을 정의하고 실행할 수 있다. 예를 들어, 수집해야 할 회의록 양이 급격히 증가할 경우 병렬처리를 늘리거나, 의안 정보를 추가적으로 수집해야 할 경우 박스로 표현된 작업을 늘려줌으로써 문제를 해결할 수 있다. 이 과정에서 이미 만들어진 작업 흐름을 크게 변경하지 않아도 된다. 셋째, 작업 흐름을 관리할 수 있는 관리자 인터페이스를 제공한다. [그림 11]은 관리자 인터페이스를 통해 작업 흐름을 그래프 형태로 시각화한 것이다. 이와 더불어 색깔로 작업 상태를 표현하고 실행 결과를 포함하고 있는 페이지를 메뉴를 통해 간단히 열람할 수도 있다. 이러한 모니터링 용이성은 데이터베이스 관리를 보다 효율적으로 할 수 있음을 의미한다. 따라서 에어플

로우에서 제공하는 관리자 인터페이스는 안정적인 프로그램에서조차 발생할 수 있는 오류를 신속하게 대처할 수 있게 도움을 준다.

V. 국회 데이터베이스 활용 예시

본 연구의 목적은 국회에서 제공하는 자료를 사용자가 편리하게 검색하고 사용할 수 있는 데이터베이스를 구축하고, 이를 공유하는 데 있다. 이를 통해 국회에 대한 견제와 감시, 그리고 연구 활성화에 기여하는 데 궁극적 목적이 있다. 이와 같은 목적을 위해 본 연구는 국회 자료의 수집·정제·저장을 자동화하여 연구의 지속성을 확보했다.

본 연구는 데이터베이스와 인터넷 웹페이지를 연동하는 웹 사이트를 개발하여 국회 데이터베이스 결과물을 제공하고 있다. 웹 사이트의 기본 구성(Front-End)은 HTML 언어로 작성되었으며, 웹 사이트 내 운용(Back-End)은 장고(Django)를 사용하였다. 제V장은 본 연구가 개발한 국회 데이터베이스 활용 예시를 소개함으로써 사용자의 활용 편의성을 극대화하는 데 있다.

본 연구에서 개발한 국회 데이터베이스는 국회의원, 회의록, 의안명 각각으로 제한하여 검색할 수 있으며, 세 가지 모두를 통괄하여 검색 가능하다. 예를 들어, 국회의원으로 검색 대상을 제한한 후 검색창에 키워드를 입력하면, 국회의원 이름만을 대상으로 해당 키워드와 일치하는 정보를 찾아 반환한다. 예를 들어, 국회의원으로 조건을 제한한 후 검색창에 “강기윤”을 입력하면, 국회의원 “강기윤”에 대한 소속정당, 지역구, 선수 등과 더불어 “강기윤”이 상임위원회, 소위원회 등에서 발언한 내용이 해당 위원회명, 일자 등과 더불어 반환된다. [그림 12]에 해당 예시가 제시되어 있다. 특히 본 연구에서 개발한 국회 데이터베이스는 각종 위원회 회의록을 의원 단위로 분해하였기 때문에 각 위원회에서 행한 의원 발언을 추출할 수 있는 장점이 있다.

[그림 12] 국회 데이터베이스 예제: 국회의원

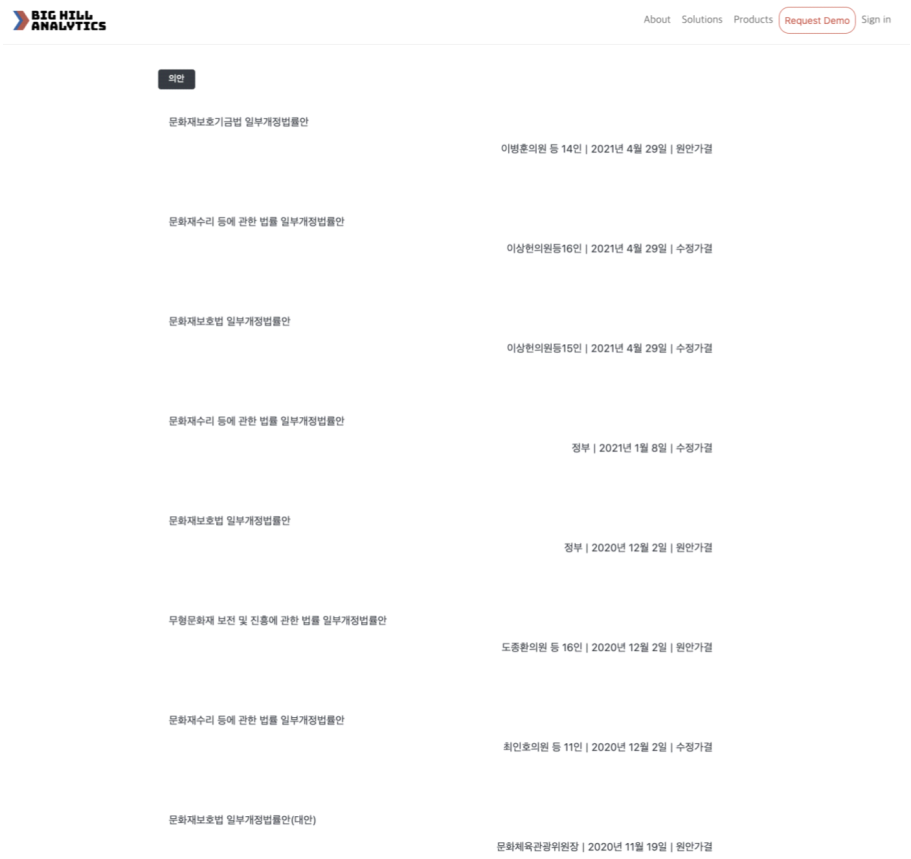
이와 같은 방식은 회의록 및 의안명에서도 적용된다. 예를 들어, 검색 대상을 회의록으로 제한한 후 국회의원 이름 또는 주요 현안 등으로 검색 키워드를 검색창에 입력하면, 해당 조건에 부합하는 결과물이 반환된다. 예를 들어, 검색 키워드를 “코로나19”로 입력하면, 해당 조건에 부합하는 회의록 내 언급이 모두 반환된다. 예시는 [그림 13]에 제시되어 있다.

[그림 13] 국회 데이터베이스 예제: 회의록

다만 해당 키워드가 회의록 또는 의안명에 존재하지 않을 경우, 검색 결과는 반환되지 않는다. 예를 들어, 검색 대상을 회의록으로 한정한 후 검색창에 “강기윤”을 입력할 경우, “강기윤”이 들어가 있는 회의록 내 발언은 반환되어 출력된다. 그러나 검색 대상을 의안명으로 한정한 후 검색창에 “강기윤”을 입력할 경우, 검색에 부합하는 결과물이 없기 때문에 검색 결과는 반환되지 않는다.

의안명으로 제한한 검색도 가능하다. 만약 검색 대상을 의안명으로 한정한 후 검색창에 “문화재”로 입력할 경우, 의안명에 “문화재”가 포함된 모든 의안을 반환

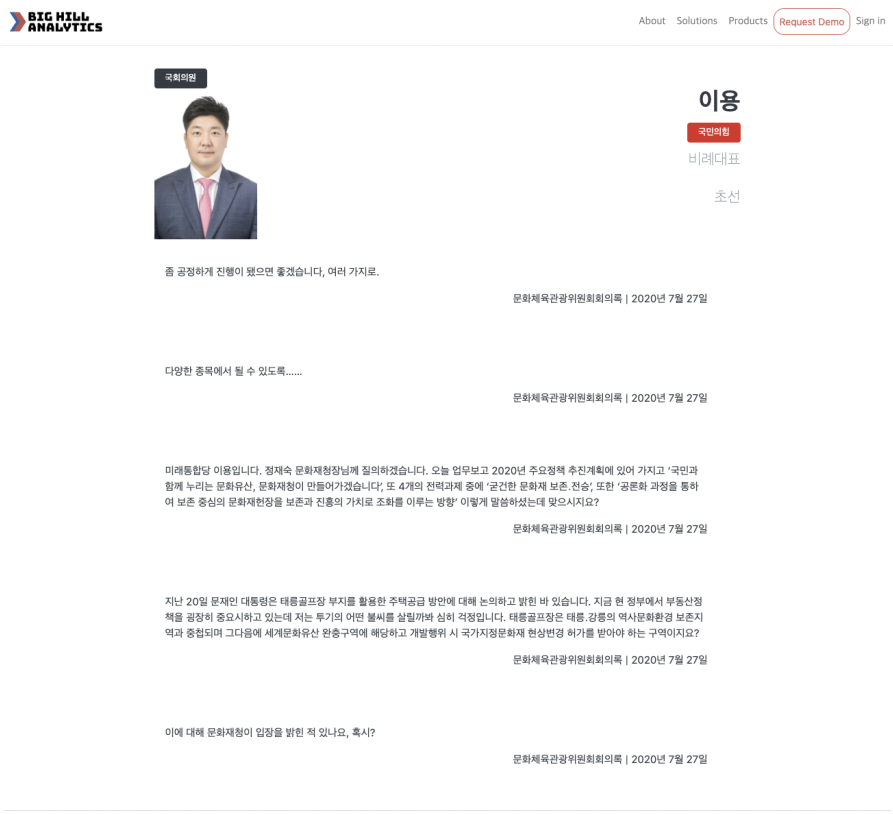
하며, 반환 결과에는 의안명을 포함하여 공동발의자 수, 발의일자, 처리 결과 등이 포함된다. [그림 14]에 해당 예시가 제시되어 있다.



[그림 14] 국회 데이터베이스 예제: 의안명

물론 국회의원, 회의록, 의안명 모두를 대상으로 한 검색도 가능하다. 즉 검색 대상을 “전체”로 할 경우, 국회의원 또는 회의록 또는 의안명에 검색 키워드가 포함된 모든 값을 반환한다. 다만, 검색 대상을 설정하지 않을 경우, 데이터베이스 내 모든 정보를 탐색하기 때문에 다소의 시간이 소요되므로 권장하지

않는다. [그림 15]에 해당 예시가 제시되어 있다.



BIG HILL ANALYTICS
About Solutions Products Request Demo Sign in

이에 대해 문화재청이 입장을 밝힌 적 있나요, 혹시?

문화체육관광위원회회의록 | 2020년 7월 27일

회의록

그럼에도 불구하고 이런 내용이 있어요. 고용노동부가 근로자 휴양콘도를 사실은 지원해 달라든가 이런 요구는 없었음에도 불구하고 '휴양콘도를 이용할 수 있도록 한다'라는 이런 내용이 들어가 있어서 오히려 택배노동자들이 요구하고 있는 분류 작업에 대체 인력을 투입해야 된다라든가 당일 배송 강요 금지라든가 장시간 노동 근절을 위한 산업안전.근로 감독 실시라든가 이런 요구는 사실은 사용자들에게 정부가 요구할 수 있는 것임에도 불구하고 포함되어 있지 않다라는 지적이 있고요. 저도 사실은 택배노동자들의 건강권을 위해서는 우리 정부가 할 수 있는 게 공동선언문에 담는 내용이 굉장히 중요하다고 생각을 하기 때문에 단순한 휴양콘도 이용 같은 것이 아니라 제도적으로 건강권을 보호할 수 있는 그런 내용들이 기조에 깔려 있어야 된다라고 생각이 듭니다. 그렇기 때문에 이건 너무 미봉책이 아니냐라는 지적이 있다라고 보고요. 지난 5월에 실시한 근로감독 결과에 대한 엄정한 후속조치를 내놔야 하지 않을까라고 생각을 합니다. 산업안전보건법상 산업재해 사망 사고는 중대재해로 규정을 하고 있지요. 택배노동자들은 모두 비슷한 노동환경에서 하루 13시간에서 16시간씩 일하고 있습니다. 이걸 주로 치면 52시간 노동시간제라는 제도가 정말 실현 가능할까 하는 그런 물음을 거꾸로 묻게 되는데요. 택배 없는 날을 하루 정했다라고 해서 택배노동자들의 휴식권이 보장받았다고 할 수 없다, 그렇게 생각을 합니다. 그렇기 때문에 공동선언의 추상적이고 모호한 내용보다는 오히려 보다 구체적인 대책을 마련해야 된다고 보고요. 과로사는 현재까지 대부분 산재로 적용을 받아서 업무상 과로로 인한 재해로 인정을 받고 있습니다. 고용노동부는 택배노동자 과로사를 산업안전보건법상의 중대재해로 규정하고 그에 따른 조치를 취하는 것이 마땅하다 이렇게 생각하는데 어떻게 생각하십니까?

환경노동위원회회의록 | 윤미향 | 위원

잘 알겠습니다. 다음 질의하겠습니다. (영상자료를 보며) 제가 여성 관련 범죄를 말씀드렸는데 실제로 최근 5년간 여성들을 표적으로 하는 여성 대상 5대 범죄, 강간.강제추행, 몰카, 통신매체 이용 음란, 성적 목적의 다중이용장소 침입, 데이트폭력, 화면 보시겠습니다마는 엄청나게 많이 발생하고 있고 줄어들지 않고 있습니다. 무려 17만 8926건입니다, 최근 5년간. 하루 평균 100여 건이 발생되고 있습니다. 청장님께서 방금 말씀을 하셨습니다마는 이 문제에 대해서는 정말 특별한 대책이 그리고 또 범부처 간의 종합적인 대책이 필요하다고 생각합니다. 어떻게 생각하세요?

행정안전위원회회의록 | 이해식 | 위원

그런데 이런 여객운송과 관련해서 노선버스 운전자들의 경우에는 조금 실질적인 보호대책, 문제의 심각성에 비해서 보호대책으로는 미흡한 것이 아닌가 하는 생각이 들어서 말씀 좀 드리려고 합니다. 지금 아시는 것처럼 노선버스가 전국 530개 업체에서 10만여 명의 종사자가 4만 6000여 대의 버스를 운행하는 것으로 나와 있어요. 그런데 작년 코로나19로 인해서 작년 2월부터 12월 마지막 주까지 노선버스 업종별 수익금 감소 현황을 살펴보니까 전 업종에서 2조 2530여억 원의 매출액 감소가 나타나는 것으로 지금 보이는데요. 여기 PPT 자료를 보면 그렇게 나오고 있습니다. 결국은 이걸 보면 버스 이용 승객하고 매출액이 급감해서 심각한 경영난을 겪고 있는 것으로 나타나는데 지금 이 상황에서 노사가 경영위기 극복을 위해서 고용 휴업이라든가 휴직을 하면서 고용유지지원금 제도를 활용을 하고 있는데 지금 상황이 올해도 빨리 종결은 돼야 되겠지만 또 이렇게 급격하게 좋아진다는 보장이 없지 않습니까? 그런데도 지금 고용유지지원금 지원 수준이 67%에서 90%로 올랐다가 다시 또 내렸지요?

환경노동위원회회의록 | 안호영 | 위원

의사일정 제5항 기초연구진흥 및 기술개발지원에 관한 법률 일부개정법률안부터 의사일정 제16항 위치정보의 보호 및 이용 등에 관

[그림 15] 국회 데이터베이스 예제: 전체

[그림 15]에 제시된 바와 같이, 검색 대상을 한정하지 않은 채, 즉 “전체”에 검색 키워드를 입력하면, 국회의원 검색 결과, 회의록 검색 결과, 의안명 검색 결과가 차례대로 반환된다.

Ⅵ. 국회 데이터베이스의 함의 및 추구 과제

본 연구의 목적은 국회에서 제공하는 자료를 수집·정제하여 사용자의 편의성이 극대화된 형태로 데이터베이스를 구축하여 학계·언론·시민단체 등과 공유하는 데 있다. 국회 자료의 방대함과 필요에도 불구하고, 그동안 학계의 연구는 단발적으로 그리고 제한된 공간에서 이루어져 왔다. 예를 들어, 국회를 분석할 때 여러 대수를 분석하여 그 추이를 파악하기보다는 하나의 대수를 분석하는 정적인 차원의 연구에 머무르고 있었다. 또한 상임위원회를 분석하는 데 있어 다양한 유형의 상임위원회를 분석하기보다는 하나의 상임위원회를 분석함으로써 비교 분석의 장점을 살리지 못했다. 주지하다시피 이와 같은 한계의 근본 원인은 국회 자료의 수집 및 정제에 상당한 비용이 들기 때문이다. 본회의 표결 자료를 수천 건의 안건마다 수집한다는 것은 상당한 비용과 시간이 소요되기 때문이다(전진영 2006a; Jun and Hix 2010). 인사청문회에서 의회가 행정부를 견제하는 방식과 다른 상임위원회에서의 견제하는 방식을 비교하는 것은 국회의원 발언의 방대함으로 인하여 불가능 영역으로 간주될 수 있다(최준영·조진만·가상준·손병권 2008; 최준영·조진만 2013).

본 연구는 자동화된 국회 데이터베이스를 구축하였다. 본 연구의 국회 데이터베이스에는 국회의원 표결 자료를 비롯하여 상임위원회 및 소위원회 의원 발언이 포함되어 있다. 또한 데이터베이스에는 법률안, 예산안 등 발의 의안에 대한 대표발의자, 공동발의자 등을 포함한 다양한 자료가 포함되어 있다. 더불어 변경된 이력, 예를 들어 소속 상임위원회 변화, 소속 정당 변화 등을 추적·축적하여 이를 제공하고 있다.

본 연구에서 개발한 국회 데이터베이스의 두 번째 장점은 확장성에 있다. 본 연구에서 소개한 자료 수집 및 정제 방식, 데이터베이스 구축 방식, 그리고 자동화 방식은 선거관리위원회, 통계청 등과 같이 다량의 데이터를 생성하는 기관에 적용할 수 있는 논리로써 데이터베이스의 확장성을 기할 수 있기 때문이다. 더불어 광역의회, 기초의회에 대한 확장을 기할 수 있기 때문에 하나의 데이터베이스에 사회과학 전반에 대해 통합하여 운용할 수 있다는 장점이 있다.

본 연구에서 개발한 국회 데이터베이스의 세 번째 장점, 그리고 가장 중요한

장점은 공유에 있다. 본 연구에서 개발한 국회 데이터베이스는 Big Hill Analytics 공간에 독립된 형태로 구성된 홈페이지를 통해 검색 및 추출을 할 수 있다.[7] 앞서 설명한 바와 같이, 국회의원·위원회·의안명에 따라 전체 또는 상세 검색할 수 있는 사이트를 제공함으로써 학계와 언론·시민단체의 활용성을 높이는 장점이 있다.

본 연구의 시도는 국회 차원에만 머무를 경우 효용성이 높지 않으리라 본다. 즉 본 연구에서 개발한 데이터베이스는 선거관리위원회, 통계청 등에서 제공하는 데이터와 연동할 때 효용성이 높아지리라 본다. 예를 들어, 국회의원의 의정활동이 당선에 미치는 영향을 분석하기 위해서는 선거관리위원회에서 제공하는 지역구 단위 투표결과 자료가 필요하기 때문이다. 국회의원의 표결이 지역구 속성에 따라 변화하는가를 분석하기 위해서는 통계청에서 제공하는 과세자료 등 다양한 지역구 자료가 필요하기 때문이다. 따라서 본 연구의 추후 과제는 이들 기관과의 협업을 통해 데이터베이스의 확장성을 기하는 데 있다.

한편, 본 연구에서 개발한 데이터베이스를 광역 및 기초의회에 대한 확장할 필요가 있다. 주지하다시피 국회와 비교할 때 광역의회 및 기초의회는 감시와 연구의 손길이 덜한 곳이다. 또한 지방의회가 효과적으로 지방정부를 견제하는지, 나아가 지방의회 존재에 대한 부정적 시각이 점증하고 있다. 이와 같은 맥락에서 광역 및 기초의회에 대한 경험적 검증과 감시가 더욱 필요하기 때문이다.

7) 본 연구에서 개발한 국회 데이터베이스는 http://www.bighillanalytics.com에서 제공한다(방문일: 2022년 1월 20일).

참고문헌

가상준·유성진·김준석. 2008.「18대 국회 초선의원과 17대 국회 초선의원 비교연구: 사회 경제적 배경, 이념성향 그리고 역할을 중심으로」 한국의회발전연구회 연구보고서.

강우창·구본상·이재묵·정진웅. 2020. “제21대 국회의원 이념성향과 정책 태도.” 『의정연구』 제26권 3호: 37-83.

문우진. 2010. “국회의원 개인배경과 입법: 입법 메커니즘과 16대와 17대 국회의 입법생산성.”『의정연구』제29권: 35-67.

손병권. 2004. “의원의 의정활동: 의원의 상임위원회 활동 참여에 대한 평가와 전망.”『한국정당학회보』제3권 제2호: 199-224.

이갑윤·이현우. “국회의원 표결과 정당 영향력: 17대 국회를 대상으로.” 『한국정치연구』 제20권 2호: 1-27.

이현우. 2005a. “국회의원의 표결 요인분석: 정당, 이념, 그리고 지역구.” 『한국과 국제정치』 제21권 3호: 187-218.

이현우. 2005b. “한국과 미국의 국회의원 표결요인 비교: 자유무역협정 비준의 경우.” 『국제정치논총』 제45집 3호: 105-125.

임성호·엄기홍·최정원·조진만·가상준·이현우·전진영·전혜원·박경미·최준영·이현출·윤종빈·조성대·손병권. 2010.『한국 국회와 정치과정』 서울: 도서출판 오름.

유승익·김현우. 2003. “국회의원의 사회경제적 배경이 선거구 득표에 미치는 영향.” 2003년 한국정치학회 하계 학술대회. 6월.

윤종빈. 2004. “17대 초선의원들의 사회·경제적 배경.”『의정연구』 제10권 제2호: 59-84.

전용주. 2012. “제19대 국회의원 특성: 사회경제적 배경을 중심으로.”『의정연구』 제18권 제2호: 29-64.

전진영. 2006a. “국회의원의 갈등적 투표행태 분석: 제16대 국회 전자표결을 중심으로.” 『한국정치학회보』 제40집 1호: 47-70.

전진영. 2006b. “조세정책 관련 법안에 대한 국회의원의 투표행태 분석.”『의정연구』 제12권 1호: 131-157.

정회옥·장혜영. 2013. “의원의 사회경제적 배경과 국회 생산성과의 관계: 18대 국회를 대상으로.”『한국정치연구』 제22권 제2호: 49-74.

조진만. 2019. "국회의원-지역구민 이념적 일치도가 선거정치에 미치는 영향 분석.“『미래정치연구』 제9권 1호: 5-30.

최준영·조진만. 2013.『견제와 균형: 인사청문회의 현재와 미래를 말하다』. 서울 써네스트.

최준영·조진만·가상준·손병권. 2008. “국무총리 인사청문회에 나타난 행정부-국회 관계 분석: 회의록에 대한 내용분석을 중심으로.”『한국정치학회보』 제42집 2호: 151-169.

한국데이터베이스진흥센터. 1999.『데이터베이스 구축 방법론』. 서울: 한국데이터베이스 진흥센터, KDPC-PR-1998-05.

한글과컴퓨터. 2018.『흔글 문서 파일 구조 5.0, revision1.3-20181108』. 경기 성남: (주) 한글과컴퓨터.

Apache Airflow Documentation, n.d. “Apache Airflow Documentation.” https://airflow.apache.org/docs/apache-airflow/stable/index.html(검색일: 2022. 1. 13).

Jun, Hae-Won and Simon Hix. 2010. “Electoral Systems, Political Career Paths and Legislative Behavior: Evidence from South Korea's Mixed-Member System.” *Japanese Journal of Political Science* 11(2): 153-171.

Data Management of Korean National Assembly in the Era of Big Data: Establishment of Automated Database for Social Sciences

Kihong Eom*·Daesik Kim**·Daewon Seo***·Wonsik Kang****·Taehun Yuk*****

ABSTRACT

The National Assembly of South Korea continuously generates various data, which facilitates social science studies. However, the application of this information is asymmetric to its utility and necessity. Researchers have to invest a significant amount of time and capital when handling even a small portion of National Assembly data. In addition, studies of the National Assembly possess time and spatial curbs due to the difficulty of manageability. This study aims to overcome these constraints by introducing automated management of the National Assembly database. Furthermore, the automation logic is also suitable for other public institutions, including National Election Commission and National Statistical Office. The study expects expansion of the database beyond the National Assembly to embrace general social sciences. Civic groups, the mass media, and social science researchers benefit from the well-established database of basic social science and eventually contribute to the transparency and integrity of democracy in South Korea.

Keywords: Database of National Assembly, Meeting Minutes, Schema, Automation, Korean Document Parser

투고일: 2022.01.27. 심사일: 2022.02.10. 게재확정일: 2022.03.07.

* Professor, Kyungpook National University
** Lecturer, Kyungpook National University
*** MA Graduate, Kyungpook National University
**** MA Student, Kyungpook National University
***** MA Student, Kyungpook National University