『한국사회학』 제57집 제2호 (2023년), pp. 131~157

연구노트

전산사회과학 연구과정의 블랙박스 열기:
아카데믹 데이터베이스를 활용한 비교사회학 연구를 중심으로*

전 준** · 김 병 준*** · 김 재 홍**** · 김 란 우*****

데이터과학에 기반한 지식사회학 연구는 어떤 과정으로 수행되는가? 이러한 접근의 강점은 무엇이며, 연구 과정에서 염두에 두어야 할 사안들은 어떤 것이 있는가? 데이터과학과 사회과학을 융합하는 시도가 전 세계적으로 증가하고 있는 가운데, 실제로 연구를 수행하는 과정에서 사회과학 연구자들이 실질적으로 마주하게 될 어려움들을 소개한 연구는 드물다. 본 연구 노트는 저자들이 수행하고 있는 사회과학 지식장에 대한 비교사회학 연구를 사례로 하여 전산사회과학의 데이터 수집 및 분석 과정에서의 암묵지를 드러내고, 이와 관련된 방법론적 시사점을 강조한다. 저자들의 지식사회학 연구는 KCI와 SSCI의 사회과학 논문들을 다양한 데이터베이스를 활용해 수집하고 정제하는 것으로 시작되었다. 한국연구재단 데이터베이스와 OpenAlex 데이터베이스를 동일한 차원으로 분석할 수 있도록 전처리하는 과정 또한 필요했다. 한국과 국제 사회과학 학문장의 거시적, 미시적 관계성을 알아내기 위해 우리는 BERTopic과 STM을 모두 시도하며 더 설득력 있는 결과물을 얻어내기 위한 시행착오를 거쳤다. 이 과정에서 사전 훈련된 말뭉치의 선정, 토픽의 개수 지정, 토픽 뭉치의 의미 해석 등 방법론에 대한 연구자들의 선택과 반복이 필수적이었다. 즉, 전산사회과학의 사회학적인 응용 가능성을 높이기 위해서는 코딩 테크닉 자체에 대한 고민뿐 아니라, 데이터, 분석 도구, 조작화 전략, 연구질문, 주장의 신뢰성과 한계에 대한 고찰 등 사회학자들이 오랫동안 천착해 온 방법론적인 질문들이 여전히 가장 중요한 것이다.

주제어: 전산사회과학, 과학의 과학, 암묵지, 자연어 처리, 텍스트 분석

* 연구를 함께 하며 도움을 주고 있는 Maida Aizaz, 최서영에게 감사드린다. 또한 우리의 부족한 원고를 읽고 조언을 준 조원광, 박재혁, 그리고 심사 과정에서 소중한 조언을 아끼지 않은 세 분의 익명의 심사자에게도 감사드린다. 이 논문은 2022년도 과학기술정보통신부의 재원으로 한국과학창의재단 과제 "인문사회-디지털(SW·AI) 융합연구소 지원"(D30300002)으로 수행한 연구이다.

** 충남대학교 사회학과 조교수, 제1저자(jjeon@cnu.ac.kr).
*** KAIST 디지털인문사회과학센터 연구조교수, 공동저자(kuntakim88@gmail.com).
**** KAIST 문화기술대학원 석사과정, 공동저자(luke.4.18@kaist.ac.kr).
***** KAIST 디지털인문사회과학부 조교수, 교신저자(lanukim@kaist.ac.kr).

Ⅰ. 들어가며

데이터과학에 기반한 지식사회학 연구는 어떤 과정으로 수행되는가? 이러한 접근의 강점은 무엇이며, 연구 과정에서 염두에 두어야 할 사안들은 어떤 것이 있는가? 데이터과학과 사회과학을 융합하는 시도가 전 세계적으로 증가하고 있는 가운데(Edelmann, Wolff, Montagne, and Bail, 2020; Lazer, Pentland, Adamic, Aral, Barabási, Brewer, Christakis, Contractor, Fowler, Gutmann, Jebara, King, Macy, Roy, and Van Alstyne, 2009; Mann, 2016; Shah, Cappella, and Neuman, 2015), 폭발적으로 늘어나고 있는 연구 결과물들에 비해, 실제로 연구를 수행하는 과정에서 사회과학 연구자들이 실질적으로 마주하게 될 어려움들을 소개한 연구는 매우 드물다. 본 연구 노트는 저자들이 함께 수행하고 있는 사회과학 학문장에 관한 비교사회학적인 연구를 사례로 하여 데이터과학 기반 사회과학 연구의 과정을 복기하여 드러내 보이고, 이와 관련된 방법론적, 이론적 시사점을 논의하고자 한다.

지식사회학은 과학기술학, 교육사회학, 문화사회학, 현상학과 긴밀한 관계를 맺으며 20세기 중후반에 걸쳐 급격한 성장을 이룬 분야이다(Berger and Luckmann, 1966; Coser, 1968; Mannheim, 1936; Merton, 1937). 특히 과학기술사회학분야에서는 '과학 지식에 대한 사회학(Sociology of Scientific Knowledge, SSK)' 연구가 큰 주목을 받았고, 과학기술의 사회적 구성 과정을 정치적, 문화적, 경제적 맥락에서 설명하는 외재적 과학기술학(STS) 연구가 널리 수행되었다(Bijker, Hughes, and Pinch, 1987; Collins, 1974; Collins and Evans, 2002; Shapin, 1995). 지식사회학의 흐름을 일목요연하게 요약하는 것은 불가능하지만, 가장 주목할 만한 연구 패러다임의 변화는 '지식'을 내면적인 현상으로 보고 그것의 사회적 근원의 탐색하던 연구로부터, 지식 그 자체를 본질적(inherently)으로 사회적인 것으로 보고, 지식 행위의 사회적 맥락과 제도적 배태성에 관해 연구하는 시각으로의 전환을 들 수 있다. 메카시는 이러한 전환을 두고 "마음의 사회성(mind's sociality)" 연구를 극복하는 흐름이라고 했으며(McCarthy, 2005), 금융시장의 위기를 지식사회학의 관점에서 설명한 맥킨지는 역사적, 사회적 상황에 의해 집합적으로 공유되고 동의되어 물질적으로 구현되기도 하는 지식의 속성에 관해 설명하며 지식의 본질적인 사회성에 대해 조명하기도 했다(MacKenzie, 2011).

최근 급부상하고 있는 과학의 과학(Science of science, SoS)은 지식의 과정과 행위의 본질적인 사회성에 대해 주목하고 있다(강정한·권은낭, 2021; Fortunato, Bergstrom, Börner, Evans, Helbing, Milojević, Petersen, Radicchi, Sinatra, Uzzi, Vespignani, Waltman, Wang, and Barabási, 2018; Wang and Barabási, 2021). SoS 학자들은 논문서지정보 데이터베이스를 활용하여 지식 생산과 파급효과의 확산 구조를 사회적 관계망의 형태로 구성해 시각화하거나, 다양한 조작화 전략을 통해 학술장 내부에서 작동하고 있는 사회구조적 변수들을 추출해 낸다추출한다. SoS의 강점은 전수조사에 가까운 빅데이터를 활용해 가장 거시적인 수준으로 학술장을 분석할 수 있다는 점이다. 이러한 연구가 가능해진 것은 Clarivate 사를 비롯한 기존의 출판사들이 일목요연하게 정리된 형태로 수많은 학술자료를 데이터베이스화해 둔 덕분이다. 즉, SoS 학자들의 다양한 연구 질문들은 그들이 활용하는 데이터가 어떤 방식으로 이미 가공되어 있는지와 밀접한 연관을 맺고 있다. 에반스와 포스터가 주장하였듯이, “전산학적 방법론은 상호작용의 미시성을 포착하기에는 아직 무딘(blunt)” 것이다(Evans and Foster, 2019: 12).

데이터의 수집, 가공, 분석 과정의 어려움에 대한 문제는 전산사회과학(Computational Social Science) 전반에 걸쳐서 꾸준히 제기되고 있다. 사회학적으로 높은 가치가 있는 데이터들은 개인 연구자의 접근을 불허하거나, 고도의 분석 및 코딩 능력을 필요로 하는 등 현실적인 어려움들이 존재한다(Edelmann et al., 2020). 데이터뿐 아니라 데이터 분석을 위해 활용되는 다양한 도구들과 관련된 불확실성 또한 존재한다. R과 파이썬을 비롯한 프로그래밍 언어의 패키지들을 활용한 간편한 데이터 분석할 수 있게 되고 있음과 동시에, 이러한 도구들이 구체적으로 어떤 방식과 원리로 작동하는 것인지에 대한 투명한 이해는 점점 어려워지고 있기 때문이다. 전산사회과학적 접근에 기반한 기존의 연구들이 점차 늘어나고 있으나(김란우·송수연, 2020; 주혜진, 2022), 구체적인 연구 과정에서의 어려움은 상대적으로 덜 부각되고 있다. 데이터와 데이터 분석 도구의 폭발적인 발전은 역설적이게도 그것을 기반으로 하는 사회과학 연구의 전 과정을 점점 블랙박스화 하고 있다. 전산사회과학의 사회학적 가능성은 따라서 데이터와 연구자 사이의 거리, 그리고 데이터 분석 방법론의 투명성과 이에 대한 연구자들의 집단적인 이해도의 향상에 의해 좌우될 가능성이 크다. 과학기술사회학자들이 과학기술의 수행 방식을 블랙박스에서 꺼내야 한다고 주장했듯이(Latour, 1987), 전산사회과학 또한 블랙박스에서 꺼내야 하는 것이다.[1)]

본 연구 노트는 전산사회과학의 블랙박스를 열어젖히기 위한 시도로, 연구자들 사이 암묵지의 공유가 중요하다고 주장한다.[2] 이를 위해 현재 저자들이 공동으로 수행하고 있는 한국과 국제 사회과학 학술장 사이의 관계에 관한 연구를 사례로 하여, 우리의 연구질문과 연구 전략이 데이터 수집 및 전처리 방법, 분석 과정에서의 시행착오, 그리고 불완전한 결과물들을 통해 어떻게 구체화하고 있는지를 공개하고자 한다. 구체적으로, 우리는 먼저 논문 서지데이터 수집 및 전처리 방법을 자세히 서술하고, 이후 이 데이터를 통해 어떠한 분석 과정을 거치게 되었는지를 설명한다. 이를 통해 불확실하고, 지저분하며, 몇 번이고 회귀하는 연구 수행과정을 공유함으로써, 본 연구 노트는 전산사회과학의 암묵지(tacit knowledge)의 중요성을 보이고자 한다. 전산사회과학의 사회학적 응용 가능성을 높이기 위한 우리의 시도는, 코딩 테크닉 자체에 대한 고민 뿐 아니라 데이터, 분석 도구, 조작화 전략, 연구질문, 주장의 신뢰성과 한계에 대한 고찰 등 사회학자들이 오랫동안 천착해 온 방법론적인 질문들이 여전히 가장 중요하다는 점을 시사한다.

---

1) 블랙박스에 대한 비유는 과학기술사회학에서 널리 사용됐다. 블랙박스는 작동 원리가 지나치게 복잡하거나, 중요하지 않은 것으로 치부되는 나머지 일반적으로 입력과 출력값에만 관심을 기울이게 되는 형태의 일련의 복합적인 과정을 뜻한다(Latour, 1987). 즉, 우리는 전산사회과학이 블랙박스화되고 있다고 말함으로써, 전산사회과학의 실제 수행과정은 알 필요가 없는 것으로 생각되는 와중에, 연구자들의 연구 소재와 그 성과물들 자체만이 지나치게 관심을 받고 있다고 지적하는 것이다. 사회과학 전반에서는 단순 상관관계가 인과관계인 것으로 논증되는 과정에서, 구체적인 사회적 기작이 드러나지 않은 경우, 사회현상이 블랙박스화되었다는 표현 또한 사용한다. 에반스와 포스터의 논문은 특히 이러한 의미에서 전산사회과학이 “무딘” 점을 지적하였다. 이러한 용례 또한 라투어적인 블랙박스로 이해할 수도 있는데, 사회과학의 과학적 수사로 인해 엄밀한 기작에 대한 설명 없이 인과관계에 대한 논증이 이루어지곤 한다고 해석할 수 있기 때문이다. 즉, 블랙박스의 개념에 대한 서로 다른 의미들은 양립 가능하다고 본다. 이하 본 연구 노트에서는 라투어적인 맥락에서 “블랙박스”라는 용어를 사용하고 있음을 밝혀둔다.

2) 우리는 암묵지(Tacit knowledge)를 단순히 표현 불가능한 지식(Non-explicable knowledge)의 의미로 사용하지 않는다. 왜냐하면 암묵지 또한 많은 경우 문자와 말로 표현할 수 있기 때문이다. 이 연구 노트에서도 우리는 실제로 우리가 경험한 암묵지를 표현하고 기록하였으므로, 표현 불가능한 것만을 암묵지로 부른다면 자기 모순적인 상황에 놓이게 된다. 따라서 우리는 암묵지를 “표현이 가능하지만 많은 경우 표현되어야 할 것으로 여겨지지 않는 종류의 지식”의 의미로 사용한다. 지식사회학자 해리 콜린스를 이러한 암묵지를 약한 암묵지(Weak Tacit Knowledge), 혹은 관계론적 암묵지(Relational Tacit Knowledge)라고 개념화했다(Collins, 2010). 암묵지는 전산사회과학에만 존재하는 것이 아니다. 과학기술사회학자들은 인간의 모든 행위에 넓은 의미의 암묵지가 개입한다고 보고 있으며, 암묵지의 존재로 인해 과학기술의 재현 가능성의 문제까지도 발생한다고 본다. 따라서 전산사회과학에도 암묵지가 존재한다는 주장 자체는 새로운 것은 아니다. 다만, 전산사회과학은 빅데이터 및 인공지능의 발달로 인해 실제보다도 지나치게 자동화되고 엄밀한 분야인 것으로 오해받고 있다. 본 연구 노트는 전산사회과학 또한 인간이 수행하는 사회과학으로서 본질적으로 내포하고 있는 암묵지적 속성으로부터 예외가 될 수 없음을 드러내 보인다고 볼 수 있다. 또한 우리의 암묵지는 표현 없이 통용되는 인간 사회의 비언어적 소통행위 전반까지를 포괄하는 의미는 아니다. 우리는 과학기술사회학에서 정의하는 좁은 의미의 암묵지만을 다룬다.

Ⅱ. 논문 서지데이터 수집 및 전처리 방법에서의 암묵지

우리의 연구 주제는 한국과 국제 사회과학 학술장의 비교에 초점이 맞춰져 있다. 이 장에서는 구체적으로 연구 질문에 답하기에 앞서, 지식사회학 연구를 위한 국내외 학술지 논문서지 데이터 수집과 전처리(preprocess) 방법을 소개한다. 전산사회과학 연구의 가능성과 효율성을 극대화하기 위해서는 데이터 분석 이전 데이터 구축 단계부터 데이터/디지털 방법론 전공자들과 사회과학 연구자들의 적극적인 소통이 필요하다. 데이터 수집과 전처리를 '용역'을 주는 것이 아닌, 연구 설계를 온전히 이해한 연구진과의 적극적인 협업만이 전산사회과학 연구 과정의 블랙박스를 여는 방법이기 때문이다.

1. 한국학술지인용색인 데이터의 수집 과정

한국학술지인용색인(Korea Citation Index, 이하 KCI)는 2007년 11월 시스템 시범 공개 이후 국내 인문사회과학 학술장 내에서 학술 커뮤니케이션 도구로서 영향력을 미치고 있다. 2023년 3월 현재 사회과학 대분류로 분류된 논문은 총 500,817건에 달하며 2022년 게재된 논문(비정규 논문 포함)만 30,906건이다. 2022년 생산된 인문학 논문이 17,817건, 복합학 대분류 논문이 8,572건임을 고려하면 매년 범 인문사회과학 논문이 50,000건 이상 생산된다고 볼 수 있다.

KCI 논문 서지데이터를 수집하는 방법은 수집하고자 하는 데이터의 특성과 크기 등에 따라 약 세 가지로 요약할 수 있다. 첫째, KCI 홈페이지에서 검색된 정보를 엑셀 파일로 내보내는 '서지정보 내보내기' 기능을 사용할 수 있다. 이 방법은 빠르게 원하는 조건의 데이터를 소량으로 수집하고자 할 때 가장 유용하다. 둘째, KCI에서 제공하는 Application Programming Interface(이하 API)를 사용하는 방법이 있다. API는 간단한 논문 서지정보를 넘어서 논문 상세 정보 및 참고문헌 정보 수집을 대량으로 할 수 있다는 가능하다는 장점이 있으나, 연구자의 일정 수준 이상의 프로그래밍 기술이 요구된다. 셋째, 우리 연구진이 API로 데이터를 수집하던 도중 수집의 한계(일일 요청량 제한)로 인하여 한국연구재단에 데이터 제공을 여러 번 문의한 결과 2022년 7월부터 데이터 직접 요청 서비스가 시작되었다. 따라서 현재 구체적으로 필요한 데이터의 조건과 형태를 인지하고 있다면 웹사이트를 통해 데이터를 직접 신청할 수 있다.[3)]

앞에서 설명한 세 가지 방법(서지정보 내보내기, API 사용, 직접 데이터 요청)으로 거의 모든 KCI 논문 서지 데이터에 접근할 수 있지만, 연구자의 인구사회학적 정보는 얻기 어렵다. 지식사회학에서 연구자의 성별, 세대, 최종학위 학교 등의 정보는 학술계의 불평등이나 분화를 확인할 수 있는 요긴한 변수이다. 우리는 논문 저자의 인구사회학적 정보를 수집하기 위해 KCI의 저자 정보와 한국연구자정보 홈페이지를 연계해 웹 크롤링하는 작업을 진행했다.

연구자의 인구사회학적 정보는 우리가 현재 진행하고 있는 연구 주제인 한국과 국제 학문장의 비교작업에 직접적으로 연관된 데이터는 아니다. 하지만 우리는 연구 자료 수집 단계에서부터 연구자 정보를 데이터베이스 일부로 포함했다. 이는 현재 연구의 미래 확장 가능성을 고려하여 데이터 수집을 한 번에 진행하기 위함이었다. 대용량 서지데이터 수집은 현재의 주제에만 맞춰져 진행하기에는 비효율적인 측면이 많다. 효율적인 연구 진행을 위하여 우리는 최대한 넓게 데이터를 수집하여 데이터베이스를 구축하고, 그 데이터 내부에서 필요한 정보를 추출하고자 하고 있다. 따라서 전산사회과학의 경우, 데이터베이스 구축에 노력과 시간이 많이 들기 때문에 데이터베이스를 확보한 연구팀을 중심으로 논문들이 집중적으로 생산되는 경우를 흔히 볼 수 있다.

2. OpenAlex 데이터의 수집 과정

‘SoS’ 연구에서 가장 일상적으로 사용되는 데이터는 Clarivate 사에서 제공하는 Web of Science라는 이름으로 묶여있는 Science Citation Index-Expanded (SCI(E)), Social Science Citation Index(SSCI), Art and Humanities citation Index (A&HCI)에 포함된 저널들에 실린 논문이다. 한국에서는 KCI 제도가 국가에 의해 관리되고 있는 것에 반해 Web of Science에 색인된 저널들은 Clarivate이라는 사기업에 의해 관리되고 운영되고 있다. 따라서 이 논문들의 정보 역시 사적 재산이기 때문에 무료로 사용할 수 없으며 대규모 웹 스크래핑(Web Scraping) 역시 원칙적으로 금지되어 있다. Web of Science 데이터가 가장 완전한 데이터에 가까우므로 이를 구매

3) 이 세 가지 접근방법에 대한 더 구체적인 내용은 김병준의 깃헙 페이지를 통해 자세한 내용을 확인할 수 있다(https://github.com/ByungjunKim/OpenBlackBox). 이 깃헙 페이지는 이후 논의되는 데이터 수집 방법과 관련된 생략된 내용들 역시 광범위하게 포함되고 있다. 따라서 실질적으로 이와 같은 데이터 수집법을 사용하고자 하는 연구자라면 해당 웹페이지를 적극 참고하는 것이 도움이 될 것이다.

하기 위해 한국의 Clarivate 지사에 데이터 구매의뢰를 하였으나 데이터 구매 비용은 천문학적인 수준이었다. 논문 1건의 세부 정보당 비용은 $0.1이었으며, 2021년까지 출판된 모든 데이터를 구매하기 위해서는 약 10억 원 이상의 연구비가 필요하였다. 이를 개인 연구자의 수준에서 구매할 수는 없었고, 연구자 소속 대학교와 같은 연구자 집단이 구매하는 것이 현실적인 해결책이었으나, 한국에서 이와 같은 커뮤니티를 찾기는 쉽지 않았다.[4)]

위와 같은 문제로 그동안 'SoS' 연구자들은 마이크로소프트가 무료로 제공해온 Microsoft Academic Graph(이후 MAG)를 연구 데이터로 애용해왔다. 하지만 MAG가 2021년 12월부로 더 이상의 업데이트를 하지 않기로 선언하면서 연구자들은 새로운 대체 데이터베이스를 구축하기 시작했고, 그 결과물이 바로 OpenAlex이다(Priem, Piwowar, and Orr, 2022). OpenAlex는 무료 오픈 데이터 세트로 MAG와 Crossref를 주요 기반으로 PubMed, ORCID, 웹사이트 크롤링 등 여러 곳의 데이터를 한데 모은 거대 데이터베이스이며, 2022년 첫 공개 기준 약 2억 건이 넘는 논문 정보를 포함하고 있다. OpenAlex에는 Web of Science와 SCOPUS, arXiv 같은 preprint 서지정보도 포함되어있다. 최근 연구에 따르면 OpenAlex가 MAG만큼 계량서지학 연구에 쓸만한 정도로 데이터 품질이 좋다고 알려져 있다(Scheidsteger and Haunschild, 2023). OpenAlex 데이터 역시 API를 통해 접근하거나, 현재까지 모인 전체 데이터(snapshot)를 한 번에 다운로드 받을 수 있다.

OpenAlex 데이터가 갖는 무료 수집 비용 및 방대한 수집 범위라는 장점에도 불구하고 실제로 분석에 적용하기 위해서는 문제점도 많았다. KCI 데이터는 국가기관에 의해 관리되고 제공되는 것에 반하여, OpenAlex가 기반으로 하는 MAG 데이터와 추후 업데이트되는 데이터는 여러 데이터 소스에 기초하고 있다. 이와 같은 특징으로 인해 우리는 OpenAlex 데이터의 정확도와 관련된 몇 가지 문제점을 발견하였다. 첫째, 본 데이터에서 구분하고 있는 연구 논문, 서평, 편집자 서문 등의 분류가 정확하지 않았다. 둘째, pdf 파일의 초록을 수집한 것으로 추정되는 데이터의 경우 띄어쓰기가 올바르게 이루어지지 않았다. 셋째, 웹상에 보이는 초록이 축약된 경우 수집된 데이터 역시 축약되기 전까지의 부분만이 수집되었다. 넷째, 논문의 제목과 초록 간의 구분이 잘못된 경우가 종종 발견되었다. 여기까지가 현재 우리의

4) 데이터 저작권 문제를 해결하기 위해서는 Web of Science 데이터를 소유하고 있는 국내외 연구기관에 소속된 연구자를 공저자로 포함하여 연구를 진행하는 방법도 있다. 하지만 이 경우 본 연구진이 원하는 연구 방향을 지속적으로 수정해야 할 필요가 있다는 문제점이 있었기 때문에 선택하지 않았다.

분석 과정에서 필요한 데이터의 특성을 확인하는 부분에서 발견된 문제들이며, 우리의 분석이 아직 닿지 않은 다른 변수들(저자 정보, 인용 정보 등)의 경우 얼마든지 비슷한 문제점들이 존재할 것이란 점을 추론할 수 있다.

우리는 다양한 데이터 전처리 방식으로 위의 문제점들을 해결하려 노력하였으나, 완벽한 해결은 불가능하였다. 대신 우리가 선택한 방법은 OpenAlex의 데이터의 정확성을 벤치마크 데이터를 사용하여 확인하는 것이었다. 가장 정확한 벤치마크 데이터는 위에서 기술했듯이 Clarivate사가 보유하고 있는 Web of Science 데이터였다. 역시 위에서 기술한 연구비의 한계로 인해 우리는 서지정보 50,000건에 한정된 데이터만을 Clarivate사로부터 구매할 수 있었으며, 이를 통해 연구자들이 가장 익숙하게 알고 있는 사회학을 사례로 전체 서지 데이터를 수집할 수 있었다. 우리는 이 데이터로 OpenAlex 데이터의 초록과 Web of Science 데이터의 초록 간의 유사성을 비교하였다. 최종적으로 코사인 유사도(cosine similarity)를 사용하여 텍스트 간의 유사성을 측정한 결과, 유사도 1이 완벽한 동일성을 뜻할 때, .9가 넘는 논문이 사회학 논문의 95% 이상을 차지함을 밝혀내었다. 이와 같은 데이터의 질을 확인하는 과정을 수행한 후, 최종적으로 OpenAlex 데이터를 좀 더 신뢰한 상태로 분석을 진행할 수 있었다.

Ⅲ. 텍스트를 데이터로 분석하는 과정에서 존재하는 암묵지

위의 데이터 수집과정을 통해 우리는 한국과 국제 학계에 해당하는 서지정보를 모두 모을 수 있었으며, 이를 통해 텍스트 분석을 진행하고자 하였다. 텍스트를 데이터로 활용하여 사회과학적 분석의 대상으로 삼는 연구 방법 자체는 이미 다양한 분야에서 보편화되었지만(Grimmer, Roberts, and Stewart, 2021; Grimmer and Stewart, 2013), 구체적으로 분석에 적용하는 과정에서의 발생하는 조작적 정의에 관한 판단은 주로 연구자 개인에게 맡겨져 있다. 특히 연구 논문이라는 장르의 특성상 연구자는 시도해본 여러 다른 연구 방법에 따른 결과를 모두 제시하기보다는 가장 최종적으로 사용한 방법론에 대해서만 구체적으로 설명하며, 그 선택만을 정당화하는 경우가 대부분이다. 그 과정에서 연구에 실질적으로 필요한 암묵지들이 삭제되고, 이 지식은 연구자 개인의 노하우로만 남게 되는 한계점이 있다. 본 장에서는 빅데이터 연구, 혹은 전산사회과학 연구 대부분이 탐색적 데이터 분석

(Exploratory Data Analysis, EDA)과 가설 검정이 순차적으로 이뤄지기보다는 끊임없이 상호작용한다는 점을 인지하고, 그 과정을 좀 더 상세하게 기술하여 암묵지로만 남아있는 연구 노하우를 구체화하는 사례를 남겨보고자 하였다.

1. 우리의 사례: 연구 질문 “거시적인 관점에서 KCI와 SSCI 간에 담론의 차이가 있는가?”에 데이터를 통해 답하는 과정

본 연구의 첫 번째 연구 질문은 거시적인 관점에서 보았을 때, 사회과학 분야에서 KCI와 SSCI의 담론 구조에 얼마나 차이가 있는지를 알아보고자 하는 것이었다.[5] 좀 더 자연어 처리 기법에 맞추어 우리의 연구 질문을 조작화 하면 다음과 같다: KCI와 SSCI 라는 사전정보 없이, 오로지 논문 제목과 초록 텍스트를 기준으로 분류했을 때, KCI와 SSCI 논문이 얼마나 다른 군집(집단)으로 분류되는가? 예를 들어, KCI와 SSCI가 100% 다른 집단으로 분류된다면 KCI와 SSCI가 관심을 가지는 담론의 성격은 완전히 다르다고 결론 내릴 수 있을 것이다.

최종적으로 우리의 분석 목표는 같은 방법론을 다양한 사회과학 분야에 적용하는 것이었으나, 이에 앞서 연구진들이 가장 익숙하게 알고 있는 학문 분야인 사회학을 사례로 연구를 시작하였다. 이를 위해 우리는 구축된 데이터베이스에서 2004년부터 2021년까지 KCI 및 SSCI에 색인된 사회학 저널에 출판된 논문들의 정보를 추출하였다. 다양한 논문 관련 정보 중에서 우리는 논문의 담론 구조를 파악하기 위한 수단으로 논문 제목 및 초록을 사용하였다. KCI 색인 저널의 경우 등재지 조건을 유지하기 위하여 영문 논문 제목 및 초록이 필수였기 때문에, 우리는 KCI와 SSCI의 담론 구조를 영문 텍스트를 활용하여 직접적으로 비교할 수 있었다.

위에서 언급하였듯이 국가에서 관리하는 KCI의 경우 초록 데이터가 깨끗하게 보관되어 있으나 OpenAlex를 통해 수집된 SSCI 데이터의 경우 여러 단계의 텍스트 전처리 과정이 필요하였다. 먼저 영어가 아닌 언어를 사용한 논문 정보를 제외하기 위하여 언어 감지 알고리즘(language detection algorithm)을 적용하였다. 또한 연구 논문이 아닌 경우를 제외하려고 초록의 길이가 지나치게 짧거나 긴 경우를 제외하였다. 이와 같은 과정을 거쳐 최종적으로 2004년부터 2021년까지 게재된 총 6,086건의 KCI 및 26,232건의 SSCI 사회학 논문을 추출하였다. 이는 전수조사 데이터이

5) 이 연구 질문은 연구진 중 한 명의 기존 논문에서 사회학에 한정되어 다뤄진 바 있다(김란우·송수연, 2020). 하지만 본 연구 질문은 이전의 연구를 다른 사회과학 분야(경제학, 지리학 등) 및 한국을 포함한 대만으로 확장하고, 그 과정에서 이론적 함의를 찾는다는 점에서 의의가 있다.

기에 실질적으로 두 학계의 담론구조를 파악하기 위한 목적으로서는 가장 신뢰할 수 있는 데이터라고 볼 수 있다.

흔히 자연어 처리 방법론을 사용하게 되는 계기는 데이터의 규모가 너무 커서 소수의 연구자가 편향과 기복 없이 작업을 수행하기가 어렵기 때문이다. 우리의 데이터 크기 및 구조도 비슷한 상황에 해당하였기에 자연어 처리 기법을 사용하는 데는 연구진 사이에 큰 이견이 없었다. 또한 본 작업은 컴퓨터에 미리 KCI/SSCI 분류에 대한 사전지식을 주지 않고 컴퓨터 스스로 카테고리를 찾아내고 논문들을 분류하도록 주문하고자 했다는 점에서 비지도 토픽 모델(Unsupervised Topic Model) 이 적합하다고 판단하였다.[6] 우리는 모든 분석에서 KCI와 SSCI에 실린 논문들을 합쳐서 분석을 수행함으로써 두 서로 다른 학계의 이질성을 파악해보고자 하였다.

여기서 우리 연구진이 선택한 전략 또한 시행착오를 거친 선택이었다. 두 학계의 모든 말뭉치를 한데 모아 비지도학습을 진행하는 전략을 최종적으로 선택하였으나, 또 다른 방법으로는 KCI와 SSCI의 말뭉치에 대해 각각 토픽 모델을 수행하는 것도 가능한 선택지였다. 그러나 우리는 이러한 방식으로 연구를 수행했을 때 지나치게 자명한 결과가 나오는 점을 두고 고민하지 않을 수 없었다. 왜냐하면 서로 다른 두 말뭉치로부터 기인한 토픽 모델의 결과 각각 다르게 나오는 것은 당연한 것이고, 이 당연한 결과만을 바탕으로 KCI와 SSCI가 서로 다른 내용을 연구하고 있다고 주장하는 것은 오히려 과학적이지 않기 때문이다. 연구진 중 한 명은 이를 두고 '클리어한 결과물을 가지고 거짓말을 하는 것 같은 기분'이라고 고백했고, 나머지 연구진도 이러한 심경에 동의했다. 따라서 우리는 보다 '자명하지 않은 결과가 나올 수 있는 방법'을 의도적으로 선택하기로 했다. 두 학계의 말뭉치를 한데 모아서 토픽 모델을 수행하되, 각각의 논문이 KCI인지 SSCI인지에 대한 식별정보를 유지하면, 최종적으로 도출된 거대한 토픽 모델에서 각각의 주제에 대해 KCI와 SSCI가 어느 정도로 분포하고 있는지 역산할 수 있다.

---

6) 다른 한 편, 이미 논문들이 KCI와 SSCI로 구분되어 있으므로 Labeled Topic Model을 사용한 접근을 통해서도 얼마나 이 두 학계가 구분 가능한지를 측정할 수 있다. 다만, 이에 대한 분석은 연구 노트의 분량의 한계로 인해 다루지 못했다. 또한 본 논문에서는 분량 관계로 다루지는 못했지만, 이에 더해 우리는 같은 개념이 각각의 학계에서 얼마나 서로 다른 개념들과 관계를 맺고 있는지도 분석해 두 학계의 '다름'의 맥락을 다면적으로 보이고자 했다.

2. 사전학습 말뭉치 선택의 문제

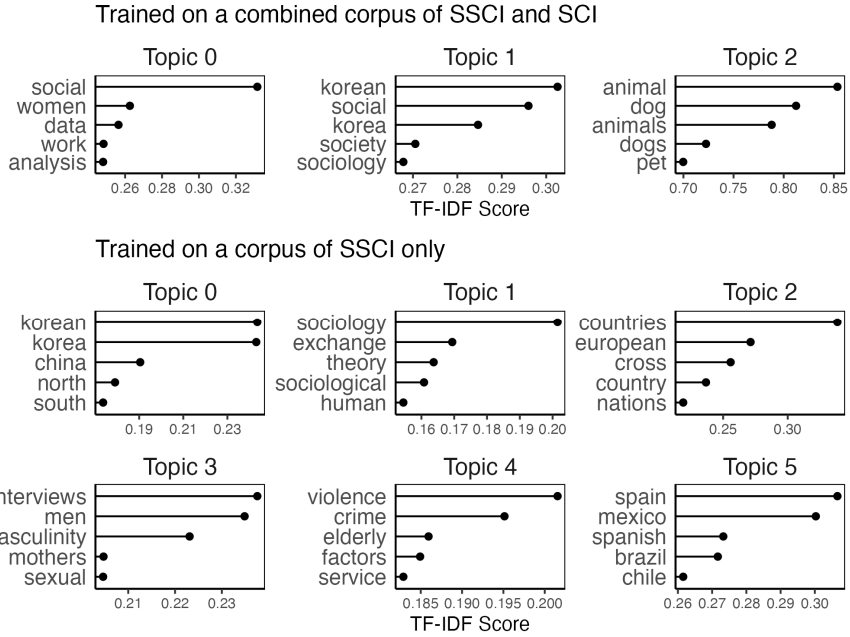
다음 문제는 여러 토픽 모델 중에 어떤 모델을 선택해 연구를 수행할 것인지, 또한 이때 이 모형의 사전학습을 위한 말뭉치로는 무엇을 활용할 것인지에 대한 것이었다. 토픽 모델 중 가장 흔히 사용되는 모델은 구조적 토픽 모델(Structural Topic Model, STM)이었으나, 비교적 최신의 토픽 모델링 기법은 BERTopic으로 BERT (Bidirectional Encoder Representations from Transformers) 모델을 기반으로 한 토픽 모형이었다(Grootendorst, 2022). 해당 토픽 모델링 기법은 딥러닝을 통해 사전훈련된 모델을 불러와 토픽모델링에 사용하는 방식이다. 이 방법론은 단어 단위의 의미뿐만 아니라, 문맥을 파악할 수 있는 우수한 사전학습 모델을 토픽 모델링에 사용함으로써 인간이 텍스트를 읽는 형태와 유사하게 컴퓨터에 분석을 의뢰할 수 있는 장점이 있다. 또한 이 접근은 데이터 전처리에 사용되는 자원을 줄일 수 있다. STM 모델은 어근 처리(stemming), 합성어 생성 등의 과정을 거치며, 텍스트를 있는 그대로 분석에 사용하기 보다는 한 번 더 컴퓨터가 읽기 좋은 형태로 '처리'하는 과정이 필요하지만, BERTopic은 이와 같은 과정이 불필요했다.

우리 연구팀은 두 가지 토픽 모델 중에 좀 더 최신 자연어 처리 방법론인 BERTopic 모형을 시도해보기로 하였다. BERTopic 모형은 미리 대규모 말뭉치를 통해 데이터를 '사전학습'(pre-train) 시키는 과정이 필요하였는데, 현재 우리의 데이터 총량만으로 모델을 학습시키기에는 부족하였다. 그렇다면 어떠한 데이터로 BERT 모형이 학습되어야 할 것인가? 좀 더 인간 친화적인 표현을 사용하면, 어떠한 글을 읽고 배운 주체(여기서는 컴퓨터)가 우리가 하고자 하는 문서분류작업을 가장 효과적으로 해낼 수 있을 것인가? 이상적으로 생각하였을 때, 우리는 사회과학 논문을 영어로 많이 읽어본 주체가 본 분류작업에 적절할 것으로 판단할 수 있다.

다행스럽게도 학계에서 연구 결과뿐만 아니라 과정 역시 투명하게 공유하고자 하는 개방형 과학(open science) 사조가 유행하면서 미리 학습을 마친 BERT 모형을 온라인으로 공유하는 경우가 흔하게 발견되었다. 다양한 말뭉치를 학습한 사전학습된 BERT 언어 모형들이 존재하는 가운데, 우리는 크게 두 가지 유용한 사전학습 모형을 찾을 수 있었다. 첫 번째는 SSCI 및 SCI 논문을 모두 학습한 언어 모형이었으며, 두 번째는 SSCI 논문만을 학습한 언어 모형이었다. SSCI 및 SCI 논문을 모두 합친 모델의 경우, SCI 논문은 바이오, 컴퓨터공학, 화학 등과 같이 사회과학을 벗어나는 범위의 글들이긴 하지만 그 규모가 워낙 방대하므로 언어의 규칙들을

더 자세하게 학습할 수 있다는 점에서 장점이 있는 것으로 생각되었다. SSCI 논문만을 학습한 모델의 경우 주제 측면에서 더 적절하였으나, 사회과학 논문의 수가 상대적으로 적다는 단점이 있었다. 이 정도의 사전지식만으로 두 가지 모델 중 더 적절한 모델을 선택하기에는 우려가 있었으므로, 우리는 두 가지 모형 결과를 모두 살펴본 후 최종 모델을 결정하고자 하였다.

〈Figure 1〉 The Comparison of BERTopic Results Using SCI-SSCI Pre-Trained Model and SSCI Pre-Trained Model



<Figure 1>은 사전학습 모델이 어떤 말뭉치를 통해 훈련되었는지에 따라, 토픽 모델 결과가 어떻게 다르게 나타나는지를 사회학 분야의 사례를 통해 보여주고 있다. 우리는 사전학습 모델에 사용된 말뭉치의 크기가 워낙 크고, 결국 학술장에서 쓰이는 언어라는 측면에서 두 가지 사전학습 모델의 결과가 크게 다를 것이라고 예상하지 않았다. 하지만 실제 결과는 예상과 달랐다. BERTopic 모형의 최적화된 추천에 따르면, SCI-SSCI 말뭉치로 학습된 모델의 경우, 사회학 분야는 총 3개의 토픽으로 분류하는 것이 최적화된 선택이며, SSCI 말뭉치만을 사용할 때 총 43개로 분류할 수 있다. <Figure 1>은 분류된 토픽 중 몇 가지 사례에서 가장 많이 쓰인 키

워드를 보여주고 있다. SCI와 SSCI 말뭉치로 학습된 토픽 모델 결과(<Figure 1>의 위 그림)에 따르면, Topic 0은 광범위한 사회학 주제와 관련된 단어들(social, women, data, work 등)을 보여주고 있으며, Topic 1은 지역적 특성이 나타나는 단어들(korean, korea)을 보여준다. Topic 2의 경우 동물과 관련된 특성이 드러나는 단어들의 집합으로 이루어져 있다(animal, dog, pet). 이에 비하여 SSCI 말뭉치만을 학습하여 나온 결과의 경우, 좀 더 토픽 모델의 결과가 직관적이고 설득력이 있는 점을 찾아볼 수 있다(<Figure 1>의 아래 그림). Topic 0의 경우 동북아시아의 지정학적 특성과 관련된 주제로 이루어져 있으며(korea, china, north), Topic 3의 경우 질적인 방법론을 통해 젠더와 관련된 연구를 보여주고 있음을 알 수 있다(interviews, masculinity). Topic 2와 5는 각각 유럽 및 스페인어권 국가들과 관련된 주제를 다루고 있다. 즉, 정리하자면 SCI와 SSCI 말뭉치로 학습된 BERTopic 모형에 기반한 토픽 모델 결과의 경우, SSCI 말뭉치만을 사용한 모델에 비하여 최적의 토픽 모델 숫자도 우리의 상식보다 적으며, 토픽 모델의 결과도 납득하기 힘든 것으로 나타났다.

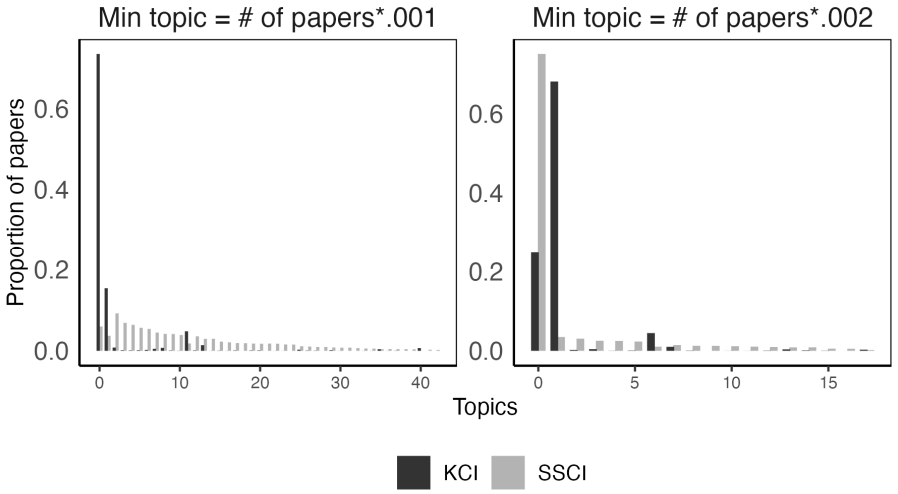
굳이 이 결과를 해석해보자면, BERTopic 모형의 결과는 데이터를 다양하게 많이 넣어 학습된 모델이라 하더라도(SCI + SSCI), 그 결과가 각 상황과 용법에 맞게 다각화되어 나타나지는 않는다는 점을 알 수 있다. 그 대신 말뭉치에서 평균적으로 사용되는 용례를 중심으로 학습이 마무리되는 것으로 보인다. 즉, 위의 결과는 BERTopic 모형을 사용할 때 사전학습 데이터가 크기 때문에 생겨나는 언어 모델의 정확도(accuracy)를 절대적으로 신뢰하기는 어려우며, 사전 학습된 말뭉치가 우리가 분석하고자 하는 말뭉치와 얼마나 유사한 언어 형태를 가졌는지를 고려해야 할 필요가 있음을 보여준다. 다시 말해, 사회과학적 글을 분석하기 위해서는 사회과학을 학습한 언어모델을 활용해야 한다는 것이다. 하지만 이 요소들을 정확히 어떤 기준으로 고려하여 균형을 맞추고, 이를 통해 최종 사전 모형을 선택해야 할지는 명확하지 않으며, 암묵지의 영역에만 존재한다고 볼 수 있다. 우리 연구진 역시 아직도 왜 이런 결과가 나왔는지 100% 이해하기 어렵다. 그리고 이처럼 방법론에 대한 완전한 이해가 부족하다는 점은 결국 분석의 한계점으로 남는다.[7)]

7) 또 다른 선택지로 기존의 언어 모델에 우리가 분석하고자 하는 말뭉치를 더하여 사전 학습된 언어 모형을 미세 조정(fine-tune) 하는 방법론도 존재하였다. 하지만 SCI+SSCI 모형처럼 더 많은 데이터로 학습된 데이터의 부적절성을 확인한 이상 우리는 사전학습 모델을 미세 조정하는 선택지 역시 한계점이 있다고 생각하고 시도하지 않았다.

3. BERTopic 모형 미세 조정을 통한 내적 타당성 검증

우리는 위의 과정을 통해 SSCI 말뭉치만을 학습한 BERTopic 모형으로 분석을 진행하고자 하였다. 우리의 연구 질문(“거시적인 관점에서 KCI와 SSCI 간에 담론의 차이가 있는가?”)에 답하는 과정에서 우리는 사회학 외의 학문에 같은 분석을 적용하여 결과에 차이가 있는지를 살펴보고자 하였다. 즉, 학문 분야의 *크기*(출판된 논문의 수)와 연구 주제가 다른 상황에서도 같은 방법론을 동일하게 적용함으로써 분석의 설득력을 높이는 것이 목표였다.

〈Figure 2〉 Comparison of BERTopic Results by Different Minimum Topic Size



같은 방법론을 동일하게 적용하기 위하여 결정해야 할 BERTopic 모형의 하이퍼파라미터(Hyperparameter)들이 존재하였다. 이 중 하나는 하나의 토픽을 구성하기 위해 몇 가지 문서(이 경우에는 논문 초록)가 필요한지를 결정하는 것이었다. 최소한으로 필요한 문서 개수를 높인다면, 최종 토픽 숫자는 줄어들 것이며, 반대로 최소 문서 개수를 낮춘다면, 최종 토픽 숫자는 늘어날 것이다. <Figure 2>는 이 결과를 잘 보여주고 있다. 최소 토픽 개수를 전체 문서의 0.001(0.1%)로 설정하였을 때, 사회학에서 추출된 토픽의 개수는 43개였으며, .002(0.2%)로 설정하였을 때는 18개였다.

<Figure 2>는 우리가 하이퍼파라미터를 조정하였을 때 총 토픽의 개수는 변하지

만, 우리가 집중하고자 하는 분석 결과의 핵심은 변하지 않는다는 점을 잘 보여준다. 두 개의 막대그래프 모두에서 검은색 막대기는 KCI의 분포를, 회색 막대기는 SSCI의 분포를 보여주고 있다. 최소 토픽 개수를 전체 문서의 0.1%로 설정한 <Figure 2>의 왼쪽 그래프의 경우, 첫 토픽 두 개에서는 KCI 분포가 훨씬 높지만, 그 외 대부분 토픽에서는 SSCI 분포가 우세하였다. 0.2%를 적용한 오른쪽 그래프의 경우, 토픽 1번을 제외한 모든 토픽에서 SSCI의 분포가 훨씬 우세하였다. 이 분석 결과는 사회학 분야의 KCI 주제는 소수의 담론으로 요약되는 것에 반하여, SSCI 주제는 더 다양한 분야에 걸쳐 분포하고 있음을 보여주고 있다.

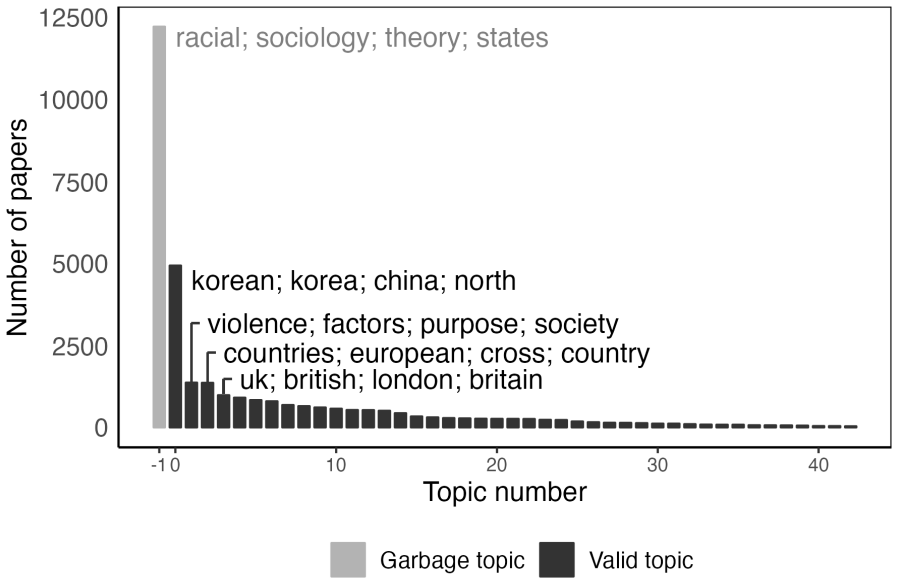
최소 토픽 개수를 전체 문서의 0.1%로 구성해야 하는지, 0.2%로 구성해야 하는지, 혹은 토픽이란 것을 구성하는 절대적인 문서의 개수(예: 100개)가 존재하는지에 대한 이론적인 근거나 경험적인 근거를 찾기는 어렵다. 이와 같은 하이퍼파라미터가 필요한 이유는 컴퓨터가 분석을 수행하는 과정을 쉽게 하기 위한 수단이며, 이를 결정하는 것은 전적으로 연구자의 몫이다. 특히 기계학습에서 하이퍼파라미터 튜닝은 모델의 성능을 좌우하는 데 큰 영향을 미치는 과정임에도 연구자의 경험 법칙(“expert experience, unwritten rules of thumb, or sometimes brute-force search”)에 의존하고 있다(Snoek, Larochelle, and Adams, 2012). 하지만 본 사회과학 연구가 어떤 추상화된 개념을 조작화 하여 분석하는 과정이라고 할 때, 본 변수의 내적 타당성과 외적 타당성이 확보되어야 할 것이다. 이처럼 분석에 결정적인 의미나 역할을 갖고 있지 않은 하이퍼파라미터에 분석 결과가 너무 민감하게 반응한다면, 그 조작화 과정은 내적 타당성을 확보하고 있다고 말하기 어려울 것이다.

4. BERTopic 모형 결과 해석을 통한 외적 타당성 검증

하이퍼파라미터 조정 등 다양한 조작화 방법을 시도하였을 때, 결과가 일정하게 유지되는지가 분석의 내적 타당성을 확보하려는 방법이라면, 분석 결과가 직관적으로 이해되는지를 확인하는 과정은 외적 타당성을 확보하기 위한 과정일 것이다. <Figure 2>는 전체 토픽의 분포만을 비교하였으며, 토픽의 내용을 구체적으로 소개하지는 않았기 때문에 현 모델의 외적 타당성을 확인하기에는 무리가 있다. 따라서 우리는 <Figure 3>를 통해 토픽의 분포 및 내용을 부분적으로 소개함으로써 BERTopic 모형의 또 다른 숨겨진 문제점을 보여주고 있다. 이 분석 모형은 최소 문서 개수를 전체 문서의 0.1%에 해당하도록 하이퍼파라미터를 설정했으며,

<Figure 2>의 왼쪽 그래프와 같은 분석 결과를 보여준다. BERTopic 모형은 하나의 문서에 하나의 토픽을 부여하는 알고리즘에 기반하고 있으므로 하나의 토픽으로 구분하기 어려운 문서들에 대해서는 -1 값을 부여하고, 이를 노이즈 토픽(noise topic)으로 분류한다.[8] 우리 분석의 문제점은 이와 같은 노이즈 토픽이 전체 문서의 약 38%에 달한다는 점이었다. 또한 노이즈 토픽에 포함된 키워드들을 살펴보았을 때, 지극히 사회학적인 용어들(racial, theory, states)이 사용되고 있었다. 따라서 이 38%의 노이즈 토픽이 분석에 포함되지 않는다는 점은 소중하게 모은 유효한 데이터의 상당 부분을 포기해야 한다는 점을 뜻하며, 이는 본 접근의 외적 타당성이 충분히 확보되지 못하였음을 반증하였다.

〈Figure 3〉 Example of BERTopic Model Results



우리 연구진은 -1 토픽의 존재는 문서의 내용에 문제가 있다기보다 많은 토픽이 하나의 문서에 혼재된 과정에서 생겨난 결과라 판단하고, -1 토픽을 최대한 다른 토픽들로 재분배하는 과정을 진행하자고 결정하였다. 노이즈 토픽으로 분류된 문서를 다른 토픽에 재분배하는 과정에서는 문서와 토픽의 최소 거리 혹은 유사도를 사용

8) <Figure 2>는 노이즈 토픽을 제외한 유효한 토픽 분류 결과만을 보여주고 있다.

한다. BERTopic에서는 토픽 재분배를 위한 유사도를 측정하기 위한 여러 척도(확률 분포, c-TF-IDF, Embeddings, Chain Strategies)를 제공한다. 이 척도 중에 무엇을 선택해야 하는지는 또다시 명확하지 않았다. 한 가지 확실한 사실은 자연어 처리를 위해 만들어진 함수들에서 기본 설정값(default)으로 되어있는 방법들이 가장 타당성이 있거나 보편적인 방법론이 아니라는 점이다. 노이즈 토픽을 재분배하는 과정에서도 c-TF-IDF 방법이 최초 설정값으로 되어있었으나, 이는 함수의 컴퓨팅 부담을 최소화하는 방법이었을 뿐, 문서가 특정 토픽으로 분류될 확률값을 이용하는 방식이 연구자의 직관에 가장 가까운 방법이었다.

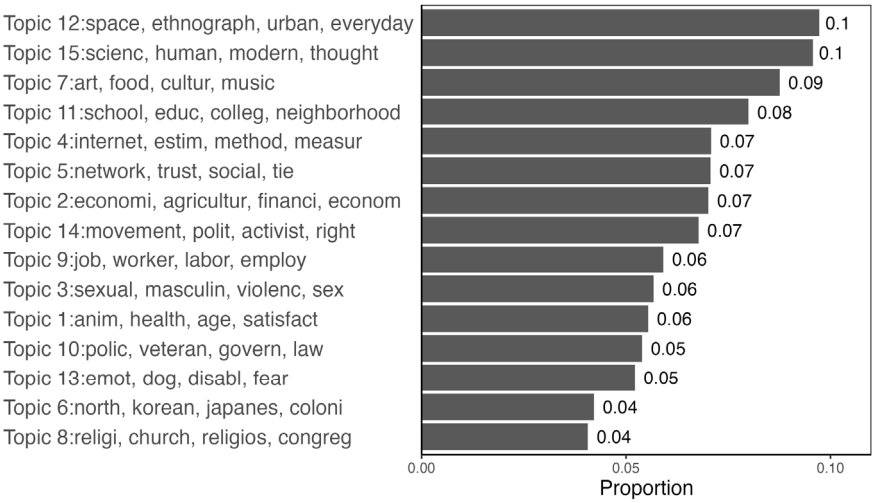
5. 구조적 토픽 모형으로의 선회

<Figure 3>의 결과는 노이즈 토픽의 존재 자체도 문제였으나, 토픽의 구분도 직관적으로 이해하기 힘든 부분이 존재하였다. 구체적으로, 위의 사례에서 토픽의 주된 키워드로 뽑힌 단어들에 대부분 국가 혹은 특정 지역의 이름이 포함돼 있었다.[9] 사회학 분야의 논문들이 다양한 지역을 연구하는 것은 당연하고 긍정적인 현상이나, 이들이 토픽 구분의 주된 기준으로 작용한다는 사실은 직관적으로 이해가 어려운 부분이었다. 좀 더 현실적인 이해를 위해 각 토픽에 해당하는 대표 논문들의 초록을 추출하여 살펴보았으나, 국가 이름을 포함하는 경우도 존재하였지만, 그렇지 않은 경우도 상당히 자주 발견되었다.

우리는 이와 같은 문제를 아예 토픽 모델을 바꿈으로써 해결할 수 있는지를 검정해보고자 하였다. BERTopic 모형이 아닌 구조적 토픽 모형을 쓰면 결과가 어떨까? <Figure 4>는 동일한 데이터로 수행한 구조적 토픽 모형의 결과를 보여주고 있다. 구조적 토픽 모형의 토픽 분류는 사회학 전공자인 연구자들에게 직관적으로 이해가 되었으며, 토픽의 분포 역시 마찬가지였다. 가장 분포가 많은 토픽도 10%를 넘기지 않았으며, 가장 분포가 낮은 토픽도 4% 정도를 차지하여 비교적 동등한 토픽으로 볼 수 있었다.

9) 구체적인 사례는 다음과 같다: 0(korea), 2(european), 3(british), 5(dutch), 6(mexico), 7(czech), 8(israel), 9(australia), 10(germany).

〈Figure 4〉 Example of Structural Topic Model Results



이처럼 구조적 토픽 모형의 결과가 BERTopic과 다르게 도출된 것은 예상 밖의 결과였다. 우리 연구진은 자연어 처리 방법론이 선형적으로 진보하고 있다고 은연중에 가정하고 있었다. 즉, 기존의 구조적 토픽 모형보다 BERTopic 모형이 더 기술적으로 발전한 형태이기 때문에, 토픽 모형의 결과 역시 더 타당하게 출력될 것이라고 가정하였다. 하지만 위의 결과는 가장 최신의 자연어 처리 방법론이 언제나 연구자들의 데이터와 연구 질문에 가장 적합한 연구방법론이 아닐 수도 있다는 점을 보여준다.

그렇다면 왜 이러한 결과가 도출된 것일까? BERTopic은 태생적으로 문서를 한 가지 카테고리로 '분류'하기 위한 목적으로 만들어진 측면이 크다. BERTopic을 만든 제작자의 논문(Grootendorst, 2022)에 따르면 제작자는 토픽 모델을 클러스터링 작업의 일환으로 해석하고 있다. 즉, 알고리즘 내부에서 하나의 문서에는 하나의 토픽만이 존재할 것이라고 가정된 채로 토픽 모델이 진행되고 있다는 특성이 있다.[10] 아마도 이와 같은 특징 때문에 BERTopic 모형은 STM과 유사한 모형인 Latent Dirichlet Allocation 모델과 비교하면 여러 분야에 걸쳐진 짧은 문서를 특정 토픽으

10) 물론 추후에 제작자가 해당 접근의 한계점을 보완하기 위하여 해당 토픽이 아닌 다른 토픽에 문서가 분류될 확률값을 제시하는 옵션을 BERTopic 함수에 추가한 바 있다. 하지만 이는 이미 알고리즘 내부에서 하나의 문서에 하나의 토픽만이 있다고 가정한 알고리즘에서 도출된 확률값이기 때문에, 하나의 문서에 다양한 토픽이 존재할 수 있다는 가능성이 배제되어 있다는 점을 보완할 수는 없다.

로 분류해내는 작업에서 우수한 성능을 보였다(de Groot, Aliannejadi, and Haas, 2022). 이러한 통찰과 유사하게도, BERTopic 모형이 뉴스 기사나 영화 리뷰와 같이 상대적으로 짧은 텍스트에 적용되었을 때 외적 타당성을 확보하는 사례가 발견된 바 있다(Kim, Chun, Jun, Kim, and Lee, 2023). 우리의 분석 대상인 논문은 다양한 관점을 종합하여 기존 연구의 빈틈을 찾아내고 이 부분을 발전시키는 형식으로 구성된다는 점(Foster, Rzhetsky, and Evans, 2015)에서 태생적으로 여러 가지 토픽에 걸쳐있을 수밖에 없다. 또한 이러한 목적을 수행하기 위한 수단으로서 쓰여진 초록 이라는 장르의 글은 BERTopic에서 분석하기에는 각각의 글의 길이가 길다. 우리는 연구 대상의 이러한 특성으로 인하여 BERTopic 모형과 STM 모형의 결과에서 큰 차이가 나타났다고 추론하고 있다.

Ⅳ. 나가며

전산사회과학 연구에는 암묵지가 존재하는가? 이상의 서술을 통해 우리는 암묵지의 존재와 그 형태를 드러내 보였다. 전산사회과학의 암묵지는 데이터의 수집, 데이터베이스의 선택 및 활용, 데이터의 전처리 및 라벨링 과정에서부터 시작된다. 우리는 KCI와 SSCI의 학술자료들을 모으고, 우리의 연구 목적에 맞게 전처리하는 과정에서부터 시행착오를 거치며 나름의 기준을 갖고 데이터의 완결성(integrity)에 대한 검증을 해 나갔다. 이렇게 일차적으로 구축된 말뭉치에 대해 어떤 토픽 모델을 활용할 것인지를 결정하는 과정 또한 자명하지 않았다. 최신 모델인 BERTopic을 SSCI 말뭉치로 훈련한 컴퓨터를 통해 실행함으로써 일련의 결과물들을 얻어냈지만, 단일 사회과학 분야 내의 토픽의 다양성을 충분히 드러내 보이지 못했기에, 외적 타당도가 높지 않다고 판단하였다. 이에 대한 대안으로 활용한 STM 기법은 의외로 훨씬 풍성한 결과값을 출력해 주었다. 가장 최신의 자연어 처리 모델이 자명하게 더 나은 결과값을 출력해 줄 것이라는 예상이 빗나갔던 것이다. 이처럼 본 연구팀의 비교 지식사회학 연구는 불확실하고 재귀적인 과정을 통해 차차 수행되고 있다. 본 연구 노트에서 기술된 내용(i.e. KCI와 SSCI의 거시적 차이점 도출)은 우리가 묻고자 하는 세 가지 연구 질문 중 하나에 해당하는 내용에 국한되며, 지면 관계상 모두 담지 못한 수많은 좌충우돌의 과정이 있고, 이 순간에도 진행 중이다.[11)]

11) 연구팀이 수행하고 있는 KCI-SSCI 사회과학 학문장의 관계에 대한 종합적인 연구의 모든 요소

한편 이러한 방법론적 좌충우돌의 과정이 단순히 연구진의 방법론적 숙련도가 부족하기 때문인 것은 아닌지 성찰하는 것도 필요하다. 만약 본 연구 노트가 기술한 것과 같은 방법론적 시도, 실패, 재선택의 과정이 우리 연구진만이 경험하는 특수한 사례인 것이라면, 이는 연구 노트의 공간에 쓰일 필요가 없는 것이 된다. 그러나 우리는 이러한 "꼭 엄밀하지만은 않아보이는" 연구 수행의 뒷무대(backstage)가 모든 전산사회과학, 더 나아가 모든 사회과학과 모든 과학 연구에서 어느 정도는 존재한다고 본다. 논리적으로 가장 엄밀한 선택의 과정들이 수행된 이후에도, 논리적으로 동등한 수준의 선택권들이 완전히 소거될 수는 없다. 결국 동등한 수준의 엄밀성을 담보하는 선택지 중 연구진들의 소통과 직관을 통해 특정 경로를 선택하게 된다. 따라서 과학 연구는 어느 정도 과소결정(underdetermine) 된다. 우리는 본 연구 노트를 통해 가장 엄밀하고 정교하게 수행된 연구에서도 암묵지는 존재할 수밖에 없다는 점을 드러내 보이려 하는 것이다. 과학적 통념은 이러한 암묵지적인 뒷무대를 드러내지 않는 전략을 선호하지만, 본 연구 노트는 이조차 함께 공유할 때 전산사회과학은 더욱 진전을 보일 것이라고 본다.

본 연구 노트에 기록된 전산사회과학의 수행과정을 통해 무엇을 고찰할 수 있을까? 첫째, 전산사회과학의 과학성을 지탱하는 것은 그 수행과정의 주도면밀함과 완전무결함이 아닌 커뮤니티의 협동이라는 점이다. 과학기술사회학에서는 자연과학자들이 가설을 증명하거나 부정하는 실험을 수행할 때, 선형적인 과정을 통해 결론에 도달하기보다는, 실험 수행 결과를 바탕으로 실험 설계를 재조정하는 재귀적인 과정, 즉, 실험자의 회귀(experimenter's regress)를 경험한다고 주장한 바 있다(Collins, 1974). 전산사회과학의 수행과정도 이와 유사한 과정을 거친다고 볼 수 있다. 우리는 BERTopic과 STM의 수행과정에서 표준화되어 있지 않은 다양한 선택을 해야 했고, 사회학자로서의 직관에 어느 정도 의존하며 해당 모델이 충분히 만족할만한 답변을 내놓는 것인지 판단해야 했다. 이 과정은 연구자들 사이의 합의를 거쳐 안정화(stabilize)된 것이다(Callon, 1984). 이 안정화와 회귀의 과정에 개입하는 것은 비단 단일 연구팀 커뮤니티 내부의 결속력만이 아니다. 유사한 분야에서

들을 본 연구 노트에서 다 기술할 수는 없었다. 우리는 본 연구 노트에 소개된 바와 같이 거시적 차원의 토픽모델을 수행한 뒤, 더 나아가 같은 개념이 KCI와 SSCI에서 어떻게 다른 의미로 사용되는지, 그리고 KCI과 SSCI 사회과학 담론이 어느 행위자에 의해 주도되는지 분석했다. 더 나아가, KCI와 SSCI에 모두 참여하는 학자들을 학문장의 브로커(Broker)라고 명명하고, 두 학문장 사이의 관계를 매개하는데에 있어서 이들의 역할 또한 밝히고 있다. 현재까지 수행된 연구의 내용과 향후 연구의 방향성은 전준·김병준·김란우(2022)에서 찾아볼 수 있다.

연구하고 있는 동료 연구자들이 공개하는 오픈소스 형태의 모델들, 데이터베이스 서비스 조직의 새로운 정책 또한 연구 과정에서의 의외의 돌파구가 되기도 했다.

둘째, 전산사회과학이 빠르게 발전하고 있는 학문 분야임에는 틀림이 없지만, 그럼에도 불구하고 전산사회과학이 사회과학의 오래되고 근본적인 문제들을 자동으로 해결해주지는 않는다는 점이다. 가령, 자연어 처리 방법론은 문자 그대로 하루가 다르게 빠른 속도로 발전하고 있다. 이 글을 쓰고 있는 현재(2023년 5월) GPT-4가 출시되었고, 많은 사용자가 이는 세상을 바꿀 기술이라 평하고 있다. 그 뿐만 아니라 본 연구진이 SSCI 문헌 정보를 읽어오기 위해 활용한 OpenAlex 서비스도 유료 프리미엄 서비스를 추가하며 사용자들에게 새로운 기능을 제공하기 시작했다[12]. 이 글이 출판될 때에는 어떤 변화가 있을지 또한 예측할 수 없다. 하지만 이처럼 빠른 방법론의 발전과 별개로, 의외로 분석 방법을 직접 적용하는 과정은 여전히 평범하고, 지루한 과정으로 구성되어 있다. 블랙박스 안에 화려하게 자동화된 오토마타는 존재하지 않는 것이다. 하버마스는 현대사회의 정치조직이 점점 과학적 근거를 의사결정의 정당성으로 의존하고 있음을 지적한 바 있는데, 이러한 현상은 학문장에서도 마찬가지로 나타난다(Habermas, 1970; Moore, Kleinman, Hess, and Frickel, 2011). 최근 전산사회과학이 큰 주목을 받는 배경에는 인간이 아닌 알고리즘과 데이터로부터 기인한 도구적 합리성이 연구의 과정적 정당성을 담보해줄 것이라는 기대가 깔려있다. 이와 관련해 크리스 베일(Christopher Bail)을 비롯한 전산사회과학의 개척자들은 사회학 이론과 방법론에 대한 근본적인 고민이 오히려 전산사회과학의 미래를 결정하게 될 것이라는 견해를 내놓은 바 있다(Edelmann et al., 2020). 실제로 우리의 연구 노트가 기술하듯 늘 최신의 방법론이 우리에게 최선의 방법론이 되지는 않았다.[13] 최선의 방법론은 결국 연구 방법의 내적 타당성과 외적 타당성을 끊임없이 탐구하고 이해하려는 노력, 그리고 연구자가 마주하고 있는 사례가 결국 사회학 이론에 비추어 무슨 의미를 제시하고 있는지 고민하는 과정에서 만들어진다고 볼 수 있다.

셋째, 전산사회과학을 수행하는 과정에서 연구팀 내부의 긴밀한 협업은 매우 중

12) 유료 서비스를 활용하면 월간 주기로 업데이트되던 스냅샷이 매시간 API의 형태로 리뉴얼되며, 처리할 수 있는 API의 개수 제한도 사라진다. OpenAlex는 2023년 3월 22일에 이 유료 결제 기능을 공개했다(https://openalex.org/pricing).

13) 최신의 알고리즘이 꼭 더 나은 연구 결과를 보장하는 것은 아니라는 관점은 여러 학문적 맥락의 선행 연구에서 언급된 바 있다(Donoho and Jin, 2008; Hand, 2006; Zhao, Parmigiani, Huttenhower, and Waldron, 2014).

요하면서도 어렵다. 돌발적인 난관에 대응하는 과정에서 연구팀이 함께 동의할 수 있는 돌파구를 선택해야 하는데, 이 과정에서 불협화음이 자주 발생한다면 연구를 완결짓기까지 지나치게 비효율적인 경로를 선택하게 될 가능성이 높을 것이다. 그러나 전산사회과학의 핵심적인 도구들이 점점 더 전문화되고 블랙박스화됨에 따라, 연구자 개개인의 프로그래밍 언어 문해력과 데이터 수집 및 분석 역량이 균일하기 어렵다. 설상가상으로 위에서도 언급한 바와 같이 방법론의 발전이 매우 빠르게 진행되고 있어서, 팀원 중 가장 최신 방법론을 받아들여 학습하는 인원과, 분석된 자료를 바탕으로 해석 및 글쓰기 작업을 주로 하는 인원 사이의 격차가 벌어지는 것은 자연스러운 일이다. 이때 발생할 수 있는 다양한 난제들(i.e. 데이터 버전의 관리, stop words 목록 최신화, 분석 모델의 최신화 혹은 구형 모델의 유지, 논문 공헌도 분배 등)을 해결하며 연구를 종결하는 일은 움직이는 과녁을 명중시키는 것만큼이나 까다로운 과업임을 전산사회과학 연구팀은 염두에 두어야 할 것이다.

전산사회과학의 학문적 연대는 어떻게 가능할 것인가? 학문적 연대는 전산사회과학을 넘어 한국사회학계의 당면 과제이기도 하며, 더 나아가 모든 학문 공동체의 고민이기도 하다(한준, 2022). 사실 우리의 연구를 비롯하여 대부분의 전산사회과학 연구는 이미 다양한 연대와 공유를 통해 수행되고 있다. 전산학자들이 구축해온 다양한 오픈소스들이 없었다면 지금과 같은 전산사회과학의 확산은 불가능했을 것이다. 전산학자들은 깃헙(GitHub)을 비롯한 플랫폼을 통해 진보된 프로그래밍 언어 패키지들과 자신들의 코드들을 이미 공유하고 있고, 이는 유관 분야 학자 사이의 '보이지 않는 협업'을 이끌어내고 있다. 사회과학자들은 어떤가? 위에서 보였듯, 전산사회과학의 과정에서 발생하는 어려움들은 단순히 코딩 자체의 기술적인 어려움으로 환원할 수 없는 성질을 갖고 있다. 전산사회과학의 사회학적인 응용 가능성을 높이기 위해서는 코딩 테크닉 자체에 대한 고민뿐 아니라, 데이터, 분석 도구, 조작화 전략, 연구 질문, 주장의 신뢰성과 한계에 대한 고찰 등 사회학자들이 오랫동안 천착해 온 방법론적인 질문들이 여전히 가장 중요한 것이다. 따라서 전산사회과학을 활용하는 사회학자들은 자신들끼리도 이러한 고민을 함께 나누고 토론해야 하겠지만, 더 나아가 전산사회과학을 활용하지 않는 다양한 사회학자들과 협업하며 토론하는 자세 또한 필요하다. 연구 과정에서 발생하는 좌충우돌을 길들이는 것이 사회과학의 본질이라면, 우리는 그 좌충우돌 또한 드러내고 공유하는 것이 사회과학을 더욱 진보시키는 길이라고 믿어야 하지 않을까? 우리가 암묵지를 강조하는 이유다.

참고문헌

강정한·권은낭. 2021. “코로나 위기가 학술 논문 생산에 미친 영향: KCI 등록 학술지 논문 분석.”『한국사회학』55(1): 179-199.

Kang, Jeong-han and Eun Rang Kwon. 2021. “The Effect of COVID-19 Pandemic on Research Productivity in South Korea: A Comparative Analysis of Korean Journal Articles across Aademic Fields.” *Korean Journal of Sociology* 55(1):179-199.

김란우·송수연. 2020. “한국 학계의 고유성은 존재하는가? 한국 사회학과 국제 사회학의 지식 담론 구조 비교를 중심으로.”『한국사회학』54(4): 1-40.

Kim, Lanu and Sue-Yeon Song. 2020. “Is Korean Academia Unique?: Comparison of Knowledge Discourses between Korean and International Sociology.” *Korean Journal of Sociology* 54(4): 1-40.

전준·김병준·김란우. 2022. “사회과학 학문장의 주변화 과정.”『한국사회학회 사회학대회 논문집』253.

Jeon, June, Byungjun Kim, and Lanu Kim. 2022. “Peripherializing Social Science Academia.” *Proceedings of Korean Sociological Association* 253.

주혜진. 2022. “대전은 어떻게 ‘노잼도시’가 되었나: 텍스트 마이닝과 의미 연결망으로 본 ‘장소성’ 소비.”『한국사회학』56(4): 51-102.

Chu, Hyejin. 2022. “How Did Daejeon City Become the ‘Snoozefest’ (No-Jam City)? : Applying Text Mining and Semantic Network Analysis for a Study on Trendy Consumption of Sense of Place.” *Korean Journal of Sociology* 56(4): 51-102.

한준. 2022. “2000년 이후 한국 사회학의 사회학: 변동과 과제.”『한국사회학』56(1): 1-28.

Han, Joon. 2022. “Sociology of Korean Sociology since 2000: Challenges and Agendas.” *Korean Journal of Sociology* 56(1): 1-28.

Berger, Peter L. and Thomas Luckmann. 1966. *The Social Construction of Knowledge: A Treatise in the Sociology of Knowledge*. Harmondsworth: Penguin Books.

Bijker, Wiebe E., Thomas P. Hughes, and Trevor Pinch. 1987. The Social Construction of Technological Systems : New Directions in the Sociology and History of

Technology. Cambridge, MA: MIT Press.

Callon, Michel. 1984. Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review*, 32, 196–233.

Collins, Harry M. 1974. "The TEA Set: Tacit Knowledge and Scientific Networks." *Social Studies of Science* 4(2): 165–185.

Collins, Harry M. 2010. *Tacit and Explicit Knowledge*. Chicago: University of Chicago Press.

Collins, Harry M. and Robert Evans. 2002. "The Third Wave of Science Studies: Studies of Expertise and Experience." *Social Studies of Science* 32(2): 235-296.

Coser, Lewis A. 1968. "Sociology of Knowledge." *International Encyclopedia of the Social Sciences*. https://essaydocs.org/sociology-of-knowledge.html

de Groot, Muriël, Mohammad Aliannejadi, and Marcel R. Haas. 2022. "Experiments on Generalizability of BERTopic on Multi-Domain Short Text." *arXiv Preprint*. arXiv: 2212.08459

Donoho, David and Jiashun Jin. 2008. "Higher Criticism Thresholding: Optimal Feature Selection When Useful Features Are Rare and Weak." *Proceedings of the National Academy of Sciences* 105(39): 14790-14795.

Edelmann, Achim, Tom Wolff, Danielle Montagne, and Christopher A. Bail. 2020. "Computational Social Science and Sociology." *Annual Review of Sociology* 46 (1): 61–81.

Evans, James and Jacob G. Foster. 2019. "Computation and the Sociological Imagination." *Contexts* 18(4): 10–15.

Fortunato, Santo, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. "Science of science." *Science*. 359(6379). https://www.science.org/doi/abs/10.1126/science.aao0185

Foster, Jacob G., Andrey Rzhetsky, and James A. Evans. 2015. "Tradition and Innovation in Scientists' Research Strategies." *American Sociological Review* 80(5): 875-908.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3): 267-297.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24: 395-419.

Grootendorst, Maarten. 2022. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure." *arXiv Preprint*. arXiv: 2203.05794.

Habermas, Jürgen. 1970. *Toward a Rational Society: Student Protest, Science, and Politics*. Boston: Beacon Press.

Hand, David J. 2006. "Classifier Technology and the Illusion of Progress." *Statistical Science* 21(1): 1-14.

Kim, Byungjun, Yuwon Chun, Bong Gwan Jun, Minhye Kim, and Wonjae Lee. 2023. "The General Public has the Key to Popularity: A Quantitative Data Analysis of Professional News Articles and Audience Reviews on Rating Platforms about Netflix's Original Series Squid Game." Working Paper.

Latour, Bruno. 1987. *Science in Action : How To Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. "Social Science. Computational Social Science." *Science* 323(5915): 721–723.

MacKenzie, Donald. 2011. "The Credit Crisis as a Problem in the Sociology of Knowledge." *The American Journal of Sociology* 116(6): 1778–1841.

Mann, Adam. 2016. "Computational Social Science." *Proceedings of the National Academy of Sciences* 113(3): 468–470.

Mannheim, Karl. 1936. *Essays on the Sociology of Knowledge*. Routledge.

McCarthy, E. Doyle. 2005. *Knowledge as Culture: The New Sociology of Knowledge*. Routledge.

Merton, Robert K. 1937. "The Sociology of Knowledge." *Isis* 27(3): 493-503.

Moore, Kelly, Daniel L. Kleinman, David Hess, and Scott Frickel. 2011. "Science and Neoliberal Globalization: A Political Sociological Approach." *Theory and Society* 40: 505-532.

Priem, Jason, Heather Piwowar, and Richard Orr. 2022. "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts." arXiv Preprint. arXiv:2205.01833

Scheidsteger, Thomas and Robin Haunschild. 2023. "Which of the Metadata with

Relevance for Bibliometrics are the Same and Which Are Different When Switching from Microsoft Academic Graph to OpenAlex?" *Profesional de La Información* 32(2).

Shah, Dhavan V., Joseph N. Cappella, and W. Russell Neuman. 2015. "Big Data, Digital Media, and Computational Social Science: Possibilities and Perils." *The Annals of the American Academy of Political and Social Science* 659(1): 6–13.

Shapin, Steven. 1995. "Here and Everywhere: Sociology of Scientific Knowledge." *Annual Review of Sociology* 21(1): 289–321.

Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams. 2012. "Practical Bayesian Optimization of Machine Learning Algorithms." *Advances in Neural Information Processing Systems* 25.

Wang, Dashun and Albert-László Barabási. 2021. *The Science of Science*. Cambridge, UK: Cambridge University Press.

Zhao, Sihai D., Giovanni Parmigiani, Curtis Huttenhower, and Levi Waldron. 2014. "Más-o-menos: A Simple Sign Averaging Method for Discrimination in Genomic Data Analysis." *Bioinformatics* 30(21): 3062-3069.

전준은 University of Wisconsin-Madison에서 사회학 박사 학위를 받았고, 현재 충남대학교 사회학과 조교수로 재직중이다. 관심 연구분야는 과학기술사회학, 환경사회학, 사회 이론 분야이다.

김병준은 성균관대학교에서 데이터 사이언스 박사 학위를 받았고, 현재 KAIST 디지털인문사회과학센터 연구조교수로 재직중이다. 관심 연구분야는 디지털인문학, 전산사회과학, 자연어처리 분야이다.

김재홍은 KAIST 문화기술대학원에서 석사과정으로 재학 중이다. 관심 연구분야는 자연어처리를 활용한 전산사회과학이다. 특히 텍스트에 담긴 감정을 분석하는 데 관심이 있다.

김란우는 University of Washington에서 사회학 박사 학위를 받았으며 현재 KAIST 디지털인문사회과학부 조교수로 재직중이다. 주로 빅데이터, 전산사회과학 방법론을 지식사회학적 연구 질문에 적용하는 연구를 이어가고 있다.

[2023.03.29 접수; 2023.05.24 수정; 2023.06.12 게재확정]

Opening the Blackbox of Computational Social Science Research Process: A Case of Comparative Study of Social Science Academia

June Jeon
Chungnam National University
Byungjun Kim
KAIST
Jaehong Kim
KAIST
Lanu Kim
KAIST

How does the data science-driven sociology of knowledge work? What are the strengths and limitations of the computational approach in the sociology of knowledge, and what kinds of methodological challenges exist? Despite the rapid growth of computational social science and related infrastructure, we need more systematic reports on the practical hurdles of computational social science research. This research note reveals tacit knowledge of computational social science by utilizing the case of our research project on a comparative study of Korean and international social science academia. In doing so, it reveals practical challenges of computational social science that are often untold while underlining the methodological significance of such a tacit process. Our comparative project started by collecting and preprocessing academic archival data via KCI and SSCI. For the macroscopic comparison of the thematic difference between KCI and SSCI, we tried both BERTopic and Structural Topic Model. During this process, we had to make non-trivial decisions on pre-trained model and hyperparameters, and interpret meanings of clustered topics. To enhance the strengths of computational social science for the sociological enterprise, we argue that classic methodological conundrums such as evaluating quality and type of data, choice of analytical toolkits, strategies for operationalization, aligning research question and methodology, and harnessing internal and external validity of the method should be prioritized over the computational technique itself.

Keywords: computational social science, science of science, tacit knowledge, natural language processing, text as data