비교문화연구 제22집 1호, (2016) pp. 89~135 서울대학교 비교문화연구소

일기자료 연구에서 토픽모델링 기법의 활용가능성 검토*

남춘호**

최근 들어 각종 문헌자료들의 디지털화가 급속히 진행되고 있으며 일상생활사 자료로서의 의의가 새롭게 부각되어온 일기자료 역시 예외는 아니다. 그러나 디지털화된 텍스트자료들은 그 방대한 규모로 인하여 전통적인 텍스트분석방법으로는 소화해내기에 한계가 있다. 본 연구에서는 해당 분야에 대한 별다른 사전적 전문지식이 없이도 방대한 디지털 텍스트자료로부터 소수의 의미 있는 토픽을 추출해주는 알고리즘으로 알려진 토픽모델링 기법의 특징과 이론적 전제에 대해 살펴보고, 이를 농민일기 분석에 예시적으로 적용해보았다. 토픽모델링 기법을 적용하여 아포일기에서 추출된 토픽들은 해석가능성이나 외적 타당도 측면에서 유의미한 것으로 드러났다. 전통적 텍스트분석방법에 의한 연구결과와의 비교에서도 대체로 일맥상통하는 것으로 나타났으며, 더 나아가 기존연구에서는 간과하였던 새로운 토픽을 발견해낼 수도 있음을 보여주었다. 이런 연구결과에 기반하여 향후 일기자료 연구에 토픽모델링 기법이 본격적으로 활용되기 위해서는 검토해야 할 부분이 무엇인지 토픽모델링의 주요 특징으로 알려진 1) 연구분야에 대한 사전적 지식을 요구하지 않는 점, 2) 멀리서 읽기, 3) '어휘자루' 가정과 관계적 의미 전제를 중심으로 논의해 보았다.

〈주요개념〉: 일기, 디지털텍스트, 토픽모델링, 타당도, 사전적 지식, 어휘자루, 멀리서 읽기

* 이 논문은 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2014S1A3A2044461). 심사자들의 세심하고 유익한 논평에 감사를 표한다.
** 전북대학교 사회학과 교수. 개인기록과 압축근대 SSK연구단.

1. 서론

최근 들어 전 세계적으로 아날로그시대에 생산된 텍스트자료의 디지털화가 빠른 속도로 진행되고 있다. 대표적인 것으로는 신문이나 학술논문, 문학저작들을 들 수 있으며, 본 연구에서 집중적으로 다루고자 하는 일기자료 역시 예외가 아니다. 방대한 텍스트 자료의 디지털화는 그러한 텍스트를 연구해온 사회과학자나 문학 및 역사연구자들에게 새로운 도전이 되고 있다. 왜냐하면 이들이 주로 사용해온 종래의 '가까이서 읽기(close reading)' 방법으로는 책상에 배달된 엄청난 규모의 디지털텍스트들을 도저히 소화해낼 수 없기 때문이다. 이에 따라 자연스럽게 디지털 자료의 연구에 컴퓨터를 활용하려는 노력들이 이루어졌는데, 가장 초기에 시도된 방법은 개별 연구자들이 관심을 가진 핵심키워드를 투입하고 이에 관련된 텍스트 내용만을 추출해서 분석하는 것이었다. 그러나 이 방법은 연구자가 이미 연구 관심분야에 대한 핵심키워드들을 사전에 알고 있고, 또한 해당 텍스트 자료에 그 키워드가 어떤 어휘로 표현되어 있는지 알고 있어야만 활용가능하다는 점에서 상당한 한계가 있다.

그런데 신문이나 왕조실록 같은 자료의 연구에 비하여 디지털화된 일기자료 연구에서는 또 한 가지 검토할 측면이 있다. 사회과학자나 역사학자들에게 일기, 특히 일반인들의 일기가 연구의 대상으로 떠오른 것은 일상생활에 대한 관심의 증대에 기인한다. 일기 속에는 근대 혹은 압축근대를 경험해온 일기 저자의 일상생활과 내면세계가 서술대상과 시공간적으로 가장 근접한 지점에서 '날 것' 그대로 기록되어 있기 때문이다(정병욱 2013: 265; 정병욱 · 이타가키 2013: 4-7). 그러나 일기를 쓴다는 행위 자체가 연구의 대상이 되어야 한다는 주장(이케다 2014: 25-26)에서 잘 드러나듯이 일기를 쓰는 동기나 내용, 형식은 일기에 따

라서 매우 다양하다. 거칠게 이분법적으로 표현해 보자면, 개인기록도 있고 집단일지의 성격을 띤 것도 있으며, 사건일기도 있고 내면일기도 있고, 보여주는 일기도 있고 숨기는 일기도 있다. 또한 일기저자의 성별이나 계층, 직업, 출신 지역 등에 따라서도 일기의 내용은 상당한 차이를 보인다(니시카와 2014: 8~45). 따라서 설령 일기 중의 특정주제만 연구를 하고자 할 경우에도 해당 일기의 저자와 일기의 특성에 대한 이해가 선행되어야 한다. 앞에서 예로 든 왕조실록이나 신문은 이미 해당 분야의 연구자들에 의해서 자료의 생산과 기록, 수정, 보존 및 저장 과정이 상세하게 파악되어 있는 반면에 일기자료는 생산과 기록과정 자체가 개별 일기마다 새롭게 연구되어야 하는 과제를 안고 있는 것이다.[1] 그러나 유명 정치인이나 문인의 일기를 제외하면 대부분 일반인들의 일기의 경우, 해당 인물에 대한 연구나 또는 해당 일기의 특성에 대한 연구가 이루어져 있지 않으며, 특히 일기 저자가 사망하였고 저자에 대한 별도의 자료가 없는 경우, 우리가 기댈 수 있는 것은 일기 전체를 통독해서 역으로 그것이 어떤 종류의 일기인지 파악해보는 도리밖에 없다. 다시 말해서 전체 일기 내용 중의 세부주제에만 관심이 있는 경우에도 먼저 일기 전체를 통독하여 일기의 특성을 파악해야 한다는 점에서 보면 관심분야의 핵심키워드 추출이라는 방법은 디지털화된 일기 자료 연구에서 뚜렷한 한계를 지닌다.

1) 디지털화 과정을 거친 종래의 텍스트 자료들은 대규모의 전자화된 자료라는 측면에서 보면 흔히 언급되는 빅데이터와 유사한 점이 있다. 그러나 한신갑(2015: 168-169)이 적확하게 지적하고 있듯이 빅데이터 – 예컨대 카드사용 관련 기록, 인터넷 페이지 뷰 기록들, 이메일이나 SNS의 기록 – 가 디지털화된 환경 속에서 일상생활의 자연스러운 과정에서 생산된 전자흔적들(digital trace)이라는 점을 감안해보면 양자 사이에는 본질적 차이가 있다. 그리고 전통적인 사회조사 자료와 비교해 볼 때 소셜미디어 분석에 사용되는 자료들은 자료의 생산과 수집과정이 연구자에 의해 전혀 통제되지 않은 상태에서 이루어졌기 때문에, 연구와 무관한 잡음이 무수히 포함되어 있으며, 이에 대한 연구가 활성화되기 위해서는 빅데이터의 생성과 수집, 기록 저장 과정에 대한 연구가 선행되어야 한다. 디지털화된 일기자료의 경우 일상생활의 전자적 흔적이 아니라는 점에서는 디지털화된 역사기록과 유사하며, 자료의 생성과정과 성격에 대한 연구가 보완되어야 한다는 점에서는 빅데이터 자료들과 유사한 과제를 안고 있다.

그런데 최근 비정형화된 텍스트 자료에 대한 데이터 마이닝 기법의 일부로 고안되어 발전하고 있는 토픽 마이닝 기법은 텍스트 자료가 담고 있는 내용에 대한 사전지식을 최소한으로 요구하면서 기계적으로 텍스트 전체의 내용을 분석하여 의미있는 주제들(토픽)을 추출해준다는 점에서 디지털 문서들, 특히 신문이나 소설, 학술논문, 역사적 사료 등의 내용분석에 활용되기 시작하고 있다. 그렇지만 아직 일기 자료의 연구에 토픽모델링 기법을 적용해본 사례는 블레빈스(Blevins 2010) 외에는 거의 없다. 19세기 미국사를 전공한 블레빈스 역시 토픽모델링 기법을 통하여 새로운 연구를 수행하기보다는 일기를 기존의 방법으로 연구한 것과 토픽모델링 기법을 통하여 연구한 결과를 비교하면서, 토픽모델링 기법으로 추출된 토픽들의 타당도를 검증해보는 수준에 그치고 있다.

이하에서는 먼저 토픽모델링 기법에 대하여 소개하고 관련한 선행연구들을 검토해 본다. 특히 토픽모델링 알고리즘의 이론적 전제에 초점을 맞추어서, 토픽모델링 기법이 의미상으로 해석가능성이 높은 토픽들을 추출해주는 메커니즘을 살펴보고자 한다. 다음으로는 한국 농촌의 일기 자료『아포일기』를 대상으로 토픽모델링 기법을 예시적으로 적용해보고 이를 통하여 일기자료의 연구에 있어서 토픽모델링 기법 적용의 가능성과 한계를 검토하고 향후의 과제를 제시해보고자 한다.

2. 토픽모델링이란 무엇인가?

토픽모델링은 데이터 마이닝 기법들 중의 하나로서 비구조화된 텍스트 자료들의 뭉치로부터 의미있는 주제(토픽)들을 추출해주는 확률모델 알고리즘이다. 초기에는 확률적 잠재의미분석(pLSA) 기법이 사용되었으나, 블레이와 동료들(Blei, Ng and Jordan 2003)이 LDA(Latent Dirichlet Allocation) 알고리즘을 발표한 이후로는 주로 LDA기법 혹은

LDA의 변용 기법들이 사용되고 있다.

토픽모델링 기법은 다른 텍스트 마이닝 기법들과 마찬가지로 초기에는 주로 산업이나 경영분야에 적용되었다.[2] 이후 방대한 텍스트 자료로부터 맥락과 관련된 단서들을 이용하여 해석가능성이 높은 주제들을 추출해주는 특성 때문에 디지털화된 문헌연구의 기법으로 점차 적용분야가 확대되고 있다. 가장 활발하게 적용된 분야는 문헌정보학에서 논문의 초록을 분석하여 주제들을 추출하고 저자를 식별하거나, 시간의 흐름에 따른 주제 분포의 변화를 통해 해당 학문분야의 연구동향을 파악하는 연구들이다(김하진 · 정효정 · 송민 2014; 박자현 · 송민 2013; Gerrish and Blei 2010; Griffiths and Steyvers 2004). 다음으로는 신문기사를 분석하여 매체에 따른 보도의 정파성을 분석하거나(강범일 · 송민 · 조화순 2013), 시기별 주제 변화를 포착하고 이를 사회적 역사적 사건이나 상황과 연관시켜서 고찰하는 연구들(Bonilla and Grimmer 2013; DiMaggio, Nag and Blei 2013; Nelson 2010; Newman and Block 2006; Yang, Torget and Mihalcea 2011)에도 적용되어 왔다. 토픽모델링 기법은 문학저작들의 주제분석에도 비교적 활발하게 활용되어 왔으나(Jockers 2014; Jockers and Mimno 2013; Rhody 2012), 그 외의 자료에는 적용된 사례가 많지 않은데, FDA 회의록 분석을 통하여 리더쉽의 스타일을 연구한 경우(Broniantowski and Magee 2011)와 중국 청조의 실록을 분석하여 청 왕조가 반란이나 소요를 어떻게 인식하고 규정하였는지 연구한 사례가 주목할 만하다(Miller 2013). 특히 후자는 반란이나 소요에 대한 사전정의된 코딩지침 없이 사료 속에서 폭력과 관련하여 나타난 토픽을 분석함으로서 역으로 청 왕조가 일상적

2) 비구조화된 텍스트 자료를 분석하는 기법인 텍스트 마이닝에는 정보추출, 개념추출, 문서분류, 군집화, 정보검색 등의 분야가 포함되는데 토픽모델링은 개념추출에 해당하며, 차원축소의 기능도 있어서 문서분류 분야로 볼 수도 있다(Miner, Delen, Elder, Fast, Hill and Nisbet 2012: 30-34). 한편 토픽모델링 기법들 중에서 가장 흔히 사용되는 것이 LDA 알고리즘이어서 때로는 양자가 혼용되기도 한다(Weingart 2012).

범죄나 반란 소요를 어떻게 이해하고 분류하고 규정하였는지 귀납적으로 포착하고자 시도하였으며, 또한 고전한문 사료에도 토픽모델링을 적용할 수 있음을 잘 보여주었다.

본 연구에서 다루고자 하는 일기자료에 토픽모델링 기법을 적용한 사례는 블레빈스(Blevins 2010)의 연구가 유일하다. 그렇지만 그는 일기연구를 통해서 새로운 연구주제를 탐색하기보다는 이미 18세기 미국여성사 전문연구자인 울리치(Urlich 1991)에 의해서 일상사 연구의 주목할 만한 연구성과로 발표된『조산원 이야기(*A Midwife's Tale*)』의 주요 사료인 마사 발라드(Martha Ballard)의 일기를 토픽모델링 기법을 통해 분석하고 그 결과를 Urlich의 연구결과와 비교함으로써 그가 명시적으로 표현하고 있지는 않지만 추출된 토픽들에 대한 일종의 타당도 검증을 하는데 치중하고 있다. 두 번째 연구(Baird and Blevins 2013)에서는 마사 발라드의 일기와 부유한 필라델피아 퀘이커교도의 부인이었던 드링커(Drinker)의 일기를 비교하고 있다. 드링커는 일기에서 여성의 관점에서 미국혁명의 극적인 영향을 연대기적으로 기록하고 있는데, 블레빈스와 그의 동료는 18세기 미국여성사 연구에서 중요한 사료로 자리잡은 두 일기자료에 대한 토픽모델링 분석을 통하여 두 일기에서 지속적으로 나타나는 이슈는 무엇인가? 계급과 교육수준의 차이, 메인주 농촌과 필라델피아 도시라는 환경의 차이 등은 두 일기의 토픽분포에서 어떤 영향을 미치는가? 사람들과의 일상적 상호작용의 범위는 어떤 차이를 보이는가? 등을 탐구하고자 시도하고 있다. 제인 오스틴의 소설『오만과 편견(*Sense and Sensibility*)』이 119,394단어에 불과한데 비하여 드링커의 일기는 8,178일에 걸쳐 975,131 단어로 구성되어 있음을 감안하면 이들의 작업은 토픽모델링 기법의 활용이 없이는 불가능했을 것으로 판단된다.

토픽모델링은 텍스트뭉치(corpus)[3]의 내용을 자동적으로 코딩해서 실질적으로 의미있는 소수의 범주들(토픽)로 추출해주는 절차를 제공해

준다. 토픽모델링 알고리즘의 실행을 위해서는 사전에 정의된 코드나 의미의 범주를 정해줄 필요가 없으며, 단지 토픽의 숫자만 정해주면 자동적으로 텍스트뭉치로부터 지정된 수의 토픽을 추출해준다는 점에서 전통적 텍스트 분석방법들에 비해서 더 귀납적이다. 토픽의 수만 정해주면 토픽을 구성하는 어휘들과 각 어휘들이 토픽에 속할 확률을 산출해주며, 동시에 전체 텍스트뭉치에서 토픽들이 어떻게 분포되는지는 물론 개별 문서들은 어떤 토픽으로 구성되는지 보여준다.

그러면 토픽모델이 어떻게 실질적으로 해석가능성이 높은 토픽들을 추출해주는가? 이는 LDA 알고리즘에 들어 있는 '의미는 관계적이다.'라는 전제에 기인한다. 의미는 어휘자체에 내재해 있는 것이 아니라 각 어휘들이 어떤 어휘들의 군집 속에 있는가에 따라서 정해진다. 동일 주제에 속하는 어휘들은 대화 속에서 동시에 나타날 가능성이 크며, 토픽모델은 구문법(syntax)이나 서사(narative), 혹은 텍스트 내의 위치에 상관없이 한 문서 내에서 어떤 어휘들이 동시발생(co-occurence)하는가를 측정한다(Mohr and Bogdanov 2013: 546-547). LDA 알고리즘에서 텍스트뭉치 속에 있는 각각의 문서는 각 저자가 말하고자 하는 주제들에 따라서 생성된 어휘들의 자루(bag of words)로 간주된다.[4)]

LDA 알고리즘을 식재료와 음식에 비유해서 설명해보자면 다음과 같다. 식재료창고에 다양한 음식메뉴의 100인분 식사재료가 준비되어 있는데 각 메뉴의 식재료들은 같은 바구니에 들어 있으며 바구니의 크기 즉 음식들의 비율은 상이하다고 가정해보자. 그리고 100명의 사람들이 식재료창고에 가서 마음에 드는 메뉴의 바구니를 찾아서 필요한 식재료들을 골라서 자루에 담아 왔다. 여기서 식재료창고에 있는 전체 식

3) corpus는 사전적으로는 말뭉치로 번역되어 사용되지만 여기서는 텍스트의 뭉치로 표현하기로 한다. 왜냐하면 corpus는 텍스트 마이닝에서는 주로 여러 문서들의 집합(collection)을 의미하기 때문이다.

4) 인문사회과학자를 위한 LDA 알고리즘에 대한 기초적 설명으로는 Weingart(2012)와 Rhody(2012), Jockers(2014: 163-165)를 참조하기 바란다.

재료들은 텍스트뭉치에 해당하며, 각자가 가져온 식재료의 자루들은 문서에 해당하고, 각각의 식재료들은 어휘에 해당한다. 그리고 음식메뉴별 식재료 바구니들은 토픽에 해당한다. 배추라는 식재료(어휘)는 김치라는 음식메뉴바구니(토픽)에도 들어있고, 배추된장국이라는 음식바구니에도 들어있다. 배추는 고춧가루, 소금, 새우젓, 멸치젓, 마늘 등의 식재료와 조합을 이룰 때는 김치(토픽)가 되고, 된장, 마늘, 물과 조합을 이룰 때는 배추된장국이 된다. 동일한 배추(어휘)라도 함께 사용되는 식재료(어휘)의 조합과 그 비율에 따라서 음식(토픽)이 달라진다.

100명의 사람들이 각자의 식재료자루(문서)에 식재료들을 담을 때는 식재료창고(텍스트뭉치)의 음식바구니 비율에 따라 음식메뉴(토픽)가 뽑힐 확률이 달라지며, 또한 각 음식바구니에서 식재료를 고를 때도 음식바구니(토픽) 안의 식재료(어휘) 비율에 따라서 개별 식재료가 뽑힐 확률이 달라진다. 어떤 사람은 김치와 보리밥과 무국을 선택할 수도 있고 다른 사람은 김치와 된장국을 선택할 수도 있다. 또한 같은 김치(토픽)를 골랐더라도 배추, 소금, 멸치젓을 선택할 수도 있고, 무, 소금, 마늘, 새우젓을 선택할 수도 있다. 이제 100인의 식재료자루 속에 들어있는 재료들을 관찰하고 나서 원래 식재료창고에 있던 음식메뉴바구니가 무엇이었으며 그 비율은 어떠했는지, 그리고 식재료창고에 있던 각각의 음식바구니에는 어떤 식재료들이 어떤 비율로 들어 있었던 것인지를 추론해내는 것이 LDA 알고리즘이다. 그리고 결과적으로 각자의 식재료 주머니(문서)는 어떤 음식(토픽)들로 구성되어 있고, 각 식재료들(어휘)은 무슨 음식의 재료인지 찾아내는 것이다. LDA의 알고리즘에서는 동일한 음식(메뉴)에 속하는 식재료들(어휘)은 동일한 식재료자루(문서)에 함께 들어 있을 가능성이 높을 것이라고 전제하며, 식재료자루 속의 동시출현빈도를 측정할 뿐 식재료의 위치나 순서는 고려하지 않는다. 그리고 원래의 식재료창고의 음식(토픽)비율이나 각 음식메뉴바구니(토픽)의 식재료구성비가 어떠했을 때 현재와 같은 100개의 식재료자루들

이 결과적으로 관찰될 가능성이 가장 높을 것인지를 추정하는 방식으로 확률모형의 파라미터들을 역으로 추정한다.

이 같은 LDA 알고리즘은 동일한 어휘(배추)라도 함께 사용되는 어휘들의 조합이 다르면 다른 토픽(음식)으로 분류하며, 다른 어휘라도 동일한 의미를 가지고 있어서 유사한 어휘조합에서 사용되면 – 마치 새우젓과 멸치젓이 위의 예에서 동일한 김치(토픽)로 분류되었듯이 – 동일한 토픽(음식)으로 분류해준다. 또한 LDA는 모든 어휘를 하나의 신호로 처리하기 때문에 텍스트뭉치에 사용된 언어가 영어든 한국어든 중국어든 동일한 알고리즘으로 토픽을 산출해준다.

토픽모델링 기법이 텍스트자료의 분석에 적용되기 시작한 것은 디지털화된 대규모 텍스트자료의 출현에 직접적으로 기인한다. 전통적인 텍스트 분석의 절차를 살펴보면 1) 흔히 학자들은 텍스트를 세밀하게 읽으면서 획득한 통찰력에 기반하여 전문가적인 해석을 내어놓는다. 그러나 이 방법은 개별 연구자의 통찰력에 크게 의존하고 있기 때문에 다른 연구자에 의해서는 동일한 결과가 산출되기 쉽지 않다는 재생가능성의 문제점을 내포하고 있다. 2) 그리고 텍스트의 양이 많아지면 보다 흔히 사용하는 방법은 연구질문, 사전적인 이론의 도움, 텍스트일부의 정독 등을 통하여, 한 묶음의 테마를 뽑고 코딩 규칙을 작성하여, 대개는 다수의 연구보조원들로 하여금 텍스트 전체를 코딩하게 한다. 그러나 이 방법은 코더들 간의 신뢰성 확보가 곤란하며, 연구자가 '전체' 텍스트를 읽기 전에 이미 무엇을 발견할 것인지 알 수 있다는 것을 전제로 한다. 그리고 여전히 규모가 아주 큰 텍스트자료에는 적용하기가 곤란하다. 3) 디지털 텍스트인 경우에는 흔히 연구질문이나 사전적인 이론적 기반에 기초하여 키워드를 설정하고, 컴퓨터를 이용하여 텍스트 속에서 키워드를 탐색하여 해당 부분만을 추출한 후 이를 세밀하게 정독하는 가까이서 읽기 방법이 사용되기도 하였다. 이 방법은 대량의 텍스트를 소화할 수 있다는 장점은 있지만, 여전히 텍스트의 내용을 읽어보기 전

에 무엇이 핵심 키워드인지, 해당 키워드가 어떻게 표현되고 있는지 사전적으로 알고 있어야 한다는 문제가 있다. 또한 어휘의 내용은 어휘 자체에 내재해 있기도 하지만 상당 부분 사용되는 맥락에 따라 의미가 달라진다는 점에 비추어 보면 한계를 지니고 있다(DiMaggio et al. 2013: 576-577). 앞의 비유적 설명을 빌리자면 김치라는 주제와 관련된 텍스트를 추출하려고 배추라는 핵심키워드로 탐색하면 엉뚱하게 배추된장국을 추출해줄 수도 있다.

그런 점에서 보면 토픽모델링 기법은 토픽의 추출 방법이 명료해서 타 연구자에 의한 재생가능성이 높으며, 사전 지식에 크게 의존하지 않고 의미의 관계성을 찾아낼 수 있다는 점에서 디지털 텍스트자료의 분석에 매우 유용한 기법이라고 하겠다. 토픽모델링 기법 중에서 비교적 널리 사용되는 표준적 LDA 알고리즘의 실행에 투입되는 정보는 분석 대상인 문서들(documents)의 집합인 텍스트뭉치(코퍼스)와 토픽 수뿐이다. 텍스트의 내용과 관련된 사전적인 지식을 크게 필요로 하지 않는다는 점에서 귀납적인 성격이 강하다고 할 수 있다.

연구자가 사전에 지정해 주어야 할 유일한 정보는 토픽의 개수이다. 그런데 최적의 토픽 수에 대한 통계적 해법은 없다. 토픽 수 결정은 산출된 토픽들의 해석가능성과 타당도 및 연구질문과 관련한 유용성에 따라 좌우된다. 토픽 전체에 대한 통계적 방식의 평가는 곤란하다. 왜냐하면 토픽모델링은 흔히 잡음 자료들을 소수의 토픽에 몰아넣음으로써 나머지 다수의 토픽들을 좀 더 해석 가능한 것으로 만들어준다. 따라서 전체 토픽을 대상으로 평가하여 최적의 모델을 선택하기는 곤란하며, 오히려 얼마나 많은 수의 의미있고 분석적으로 유용한 토픽을 산출해 주는가에 따라서 토픽 수를 선택하는 것이 나으며, 그런 의미에서 모든 토픽에 대한 최적화 시도는 무의미하다(DiMaggio et al. 2013: 582-583). 또한 토픽수의 결정은 일종의 렌즈 선택과 유사하다. 연구 관심에 따라서 원경에 대한 전체적 조망이 필요할 경우에는 소수의 토픽으로

묶어서 망원렌즈처럼 쓸 수도 있고, 보다 세밀한 관찰이 필요할 경우에는 많은 수의 토픽을 추출하게 하여 현미경처럼 사용할 수도 있다. 결국 토픽 수의 결정에서는 추출된 토픽의 해석가능성과 타당도, 연구질문에 비추어본 유용성 및 분석의 용이성 등이 중요한 기준이 되고 이 단계에서는 해당 연구영역에 대한 전문적 식견이 요구된다.

이와 관련하여 산출된 토픽의 타당도를 검증하기 위해서는 1) 토픽모델링이 과연 어휘들의 의미를 제대로 분간하고 포착해내는지 내적인 의미 타당도를 살펴보는 방법이 있다. 그리고 2) 토픽모델링으로 산출된 토픽이 외부의 변수와 예측된 대로 반응하는지 확인하여 일종의 외적 타당도를 검증해 볼 수 있다. 토픽모델링으로 산출된 토픽이 과연 내적인 의미타당도가 있는지 검토하는 가장 간단한 방법은 수작업으로 파악한 특정 사건이나 내용과 토픽모델링으로 추정한 사건이나 내용이 얼마나 유사한지 비교해 보는 것이다. 넬슨(Nelson 2010)은 남북전쟁 당시 지역신문에 대한 토픽모델링 분석을 통하여 도망노예광고 토픽을 추출하고 이 토픽의 타당도를 검증하기 위하여 수작업으로 세어본 도망노예광고의 연도별 추세와 도망노예토픽으로 추정한 도망노예광고의 추세를 비교하여 양자가 거의 일치함을 보여주었다. 마사 발라드의 일기에 토픽모델링을 적용한 블레빈스는 같은 일기자료를 이용하여 조산원이야기를 발표한 울리치의 연구결과와 비교해봄으로써 토픽모델링으로 산출된 토픽들의 타당성을 검토하였다. 블레빈스에 따르면 조산원이었던 발라드가 분만에 참여한 회수를 수작업으로 계산한 울리치의 연구결과와 자신의 연구에서 발견된 분만 관련 토픽의 년도 별 추세가 거의 유사한 분포를 보여주었다. 또한 블레빈스는 문서(일기자료)의 기록 일자 자료와 토픽들의 분포를 비교하여 추운 날씨에 관한 토픽의 분포는 겨울철에 빈번하게 나타났으며, 정원 가꾸기에 관한 토픽의 분포는 주로 봄부터 여름 사이에 빈번하게 나타났다고 보고하였다. 블레빈스 자신은 타당도라는 용어를 사용하지 않았지만 위의 방법은 추출된 토픽들

이 외부변수(날짜나 계절)에 예측 가능한 또는 설명 가능한 방식으로 반응하는가를 살펴본 것이므로 일종의 외적 타당도를 측정해본 것이라고 할 수 있다.

마지막으로 분류된 토픽들의 외부변수에 대한 예측력 내지 설명력을 직접 검토함으로써 타당도를 검증해볼 수도 있다. 조커스 외(Jockers et al. 2013: 756-760)는 소설 텍스트들의 토픽을 분류한 후 저자의 성별과 토픽분포의 관계를 분석하였다. 이들에 의하면 "여성-패션" 토픽은 여성작가의 소설에서 더 자주 나타났으며, 그밖에도 상당수의 토픽들이 작가의 성별과 관계가 있음을 발견하였다.[5] 이들은 이를 이용하여 문서별 토픽분포비율을 가지고 작가의 성별을 추정하는 모델을 수립한 후, 역으로 이 모델로 하여금 각 문서의 작가의 성별을 추정하게 하여 추정치와 실제 작가의 성별을 비교해 보았다. 그 결과 전체의 81%는 성별을 바르게 추정하였으며, 19%만 잘못 분류한 것으로 나타났다. 이처럼 문서의 토픽분포가 작가의 성별을 예측내지 추정하는 정도를 살펴봄으로써 토픽모델링 기법에 의해 추출된 토픽의 외적타당도를 보여주

5) 앞에서도 말했듯이 토픽모델에서는 표본조사의 결과에서 사용하는 통계적 유의도 검증방법을 통해 전체모델의 통계적 유의성을 검증할 수는 없다. 그러나 개별 토픽과 관련된 분석에서는 치환(permutation) 방법이나 부트스트랩 기법을 활용하여 결과의 신뢰도를 가늠해볼 수 있다. 조커스 외(Jockers et al. 2013)의 분석에서 '여성패션' 토픽의 평균 출현빈도는 여성작가의 소설에서 남성작가의 소설보다 높게 나타났다. 그런데 이 차이가 유의미한 것인지 살펴보기 위하여, 개별소설에 대하여 작가의 성을 인위적으로 무작위로 배정한 후에, 여성작가의 소설과 남성작가의 소설에서 '여성패션' 토픽의 평균출현빈도가 어떻게 나타나는지 반복적인 실험을 통해 검토하였다. 그 결과 작가의 성이 무작위로 배정된 경우에도 평균출현빈도가 다소 차이를 보였으나, 원래의 평균출현빈도 차이는 그보다 훨씬 커서 남녀별 평균출현빈도의 차이가 유의미하다는 결론을 도출하였다. 한편 여성작가 소설에서 '여성패션' 토픽의 평균출현빈도가 높다고 하더라도 이는 한두 개의 특이한 사례에 기인하는 것일 수도 있다. 이 점을 확인해보기 위하여 여성소설과 남성소설에서 각각 일정비율을 무작위로 표집하여 평균을 비교한 후 여전히 '여성패션' 토픽의 평균출현빈도에 차이가 나는지 비교하는 부트스트랩 기법을 1000회 반복해서 실시해 보았으며 그 결과 여전히 남녀작가별 평균출현빈도에 차이가 나타남을 보여주었다. 이 두 가지 방법을 통하여 '여성패션' 토픽은 여성작가의 소설에서 남성작가의 소설보다 유의미하게 자주 나타난다는 점을 입증해보였다. 이러한 방법은 일반적인 표본조사에서 사용하는 통계적 유의미성 검증과 유사한 절차라고 할 수 있으며 이를 통해 결과의 신뢰도를 검증해 볼 수 있다(Jockers et al. 2013: 756-760).

기도 하였다.

토픽모델링 기법은 기존의 텍스트 분석과는 달리 선험적 이론에 기초한 사전적 코딩범주의 투입을 요구하지 않으며, 전통적인 가까이서 읽기의 방법으로는 도저히 소화할 수 없는 방대한 양의 텍스트뭉치에서 유의미한 토픽들을 자동적으로 산출해준다. 이는 토픽모델링 기법이 '어휘들의 의미는 어휘 자체에 내재하기보다는 어휘가 사용되는 맥락 혹은 함께 사용되는 어휘들과의 관계에 기초한다.'는 전제하에서 만들어진 알고리즘이기 때문이다. 의미를 맥락 속에서 찾는다는 특성으로 인하여 토픽모델링 기법은 동음이의어, 다의어는 물론 사적인 일기표기에서 자주 나타나는 약자표기나 방언으로 인한 문제도 해소해준다. 그렇지만 토픽모델링 기법의 활용에서 연구 해당 분야의 전문적 지식이 전혀 필요하지 않다는 의미는 아니다. 토픽모델링 기법에서는 토픽의 수를 사전에 지정해 주어야 하는데 토픽수의 지정은 주로 산출되는 토픽들의 해석가능성, 타당도, 연구문제에 비추어본 유용성 등에 의존하기 때문이다. 다음 장에서는 27년간에 걸쳐 기록된 농민일기를 대상으로 토픽모델링 기법을 예시적으로 적용해보고자 한다.

3. 아포일기 토픽분석

아포일기는 경북 김천시 아포읍에 거주하는 농민 권순덕이 25세가 되던 1969년 1월 1일부터 2000년 12월 31일까지 기록한 총 27년에 걸친 9,468일의 기록이다. 본 장에서는 먼저 아포일기 텍스트 전체에 토픽모델링 기법을 적용하여 아포일기 전반을 원경에서 조망해보고, 다음에는 자녀에 관한 텍스트만 따로 추출하여 권순덕의 자녀관이나 교육관에 대해 분석해보고자 한다. 그리고 토픽모델링 기법에 의한 분석 결과를 전통적인 텍스트 분석 방법을 이용한 연구의 결과들과 비교해봄으

로써 토픽모델링으로 산출된 토픽들의 해석가능성이나 타당도를 중심으로 일기에 대한 활용가능성과 문제점들을 검토해보고자 한다. 토픽모델링을 수행하기 위해서는 먼저 한글자연어처리과정을 거쳐야 한다. 이하에서는 전처리과정, 토픽모델링 실행 및 산출된 토픽의 해석 순으로 논의를 진행하고자 한다.[6)]

1) 전처리 과정

본 연구에서는 우선 토픽모델링에 들어가기에 앞서서 한글자연어처리기 KoNLP를 사용하여 전처리과정을 수행하였다. 전처리과정에서는 아포일기 텍스트를 분석하여 품사분류와 어근분리를 수행하고 명사를 추출하였다. 아포일기에는 동음이의어나 다의어가 자주 등장하고, 경북 김천지역의 방언들이 구어체로 표기되어 있거나 소리나는 대로 표기된 경우가 많다. 그런데 토픽모델링 기법에서는 동음이의어나 다의어, 혹은 다양한 표현의 문제들을 대부분 해소해주었다. 토픽모델링 기법은 어휘 자체가 아니라 어휘가 사용되는 맥락에 따라 토픽을 분류하므로 동음이의어나 다의어는 사용되는 맥락에 따라서 상이한 토픽으로 분류해주며, 동일한 의미의 다른 표현들은 사용맥락이 동일하면 같은 토픽으로 묶어주기 때문이다. 그러나 명사와 조사의 품사분리에서는 사전에 명사로 포함되지 않은 경우, 명사와 조사의 분리상에서 문제를 일으키는 경우도 발생하였다.[7)] 따라서 본 연구에서는 KoNLP에 탑재된 세종

6) 한글자연어처리에는 R기반의 KoNLP_Ver 0.76.5(Jeon 2012) 패키지를 이용하였으며, 토픽모델링 분석에는 R기반의 LDA프로그램인 'topicmodels' 패키지를 사용하였다(Grun and Hornik 2011).

7) 블레빈스(Blevins 2010)가 분석한 발라드의 일기에서도 딸을 의미하는 daughter은 Dauther, Dagts, Daftr, Daug, Dagt, Daugts, Daght, Dats, Daughter, Daughts, dt, Daught, daugt, dgt, datr, daugrtr, dafters, dafter, dtr 등으로 다양하게 표기되었으나 토픽분류에서는 대부분 동일한 토픽으로 묶여 나왔다. 개인일기의 특성상 방언이나 다양한 약자표기가 많은데 이런 다양한 표현에서 발생하는 문제들은 토픽모델링 알고리즘에 의해서 대부분 해소되었다. 품사의 분류과정에서 특히 동사방언들의 어근

사전에 아포일기의 명사방언들을 추가시켜 전처리과정을 수행하였다. 명사만을 사용한 것은 아직 한글자연어 처리 프로그램들의 동사나 부사 형용사의 어근분리 성능이 안정되지 않았기 때문인데, 방언이 많은 아포일기의 경우 그 문제가 더 심각하였다. 또한 기존 연구사례를 보면 명사만 사용했을 경우에도 의미있는 토픽들을 추출해주는 것으로 보고되어 명사만을 분석에 포함시켰다(강범일 외 2013: 319-320; Jockers et al. 2013: 754). 명사사전의 추가 후에도 계속적으로 오분류되어 나오는 어휘들은 불용어휘집(stop word list)에 넣어서 전처리 과정에서 제거시켰다.

아포일기의 텍스트는 모두 981,065개의 낱말(중복포함)로 구성되어 있는데 전처리 과정을 거쳐서 명사를 뽑은 결과 43,799개의 명사가 추출되었다. 이중에서 어휘 총빈도(GF), 문서별 어휘빈도(TF), 해당 어휘의 출현 문서수(DF)를 활용하여 3,619개의 어휘를 추출하여 토픽모델링 분석에 사용하였다.[8)]

이 명사로 오분류되는 경우가 상당수 있었는데. 이런 오분류 어휘들은 발생빈도가 대부분 아주 낮아서 전체분석의 경우 발생빈도를 8이상으로 제한하는 방법으로 분석에서 제외시켰다. 또한 일음절 단어나 6음절 이상의 단어에서 오분류의 문제가 심각하여 본 연구의 대상은 2~5음절 명사로 제한하였다. 다만 6음절 이상의 명사 중 유의미한 언어는 결합명사나 고유명사가 많았는데 김천내과의원 같은 경우 김천, 내과의원이라는 명사를 추가하여 두 단어로 분리시켰다.

8) 대부분의 개인일기와 마찬가지로 아포일기의 어휘사용패턴을 보면 소수의 어휘가 집중적으로 사용되는 지수곡선 커브를 그리고 있다. 이는 발라드의 일기에서도 나타나는 특징이다(Blevins 2010). 따라서 본고에서는 1차로 해당 어휘가 일기전체에 나타난 총 빈도수(GF)가 8회 이상인 어휘를 추출하고 2차로 TFIDF가 0.2보다 큰 어휘를 재추출하였다. TFIDF는 각 문서별 해당 어휘빈도(TF)를 해당 어휘가 나오는 전체 문서수(DF)로 나눈 값이다. 이렇게 하면 빈도가 낮은 어휘와 더불어, 특별한 문서나 주제에 집중되지 않고 모든 문서에 공통적으로 등장하는 어휘(예컨대 나, 자신, 생각)도 제거해주기 때문에 각 문서 특유의 잠재적 의미 토픽을 더 잘 추출할 수 있다(Grun et al. 2011: 7).

2) 날짜별 전체일기 토픽분석

토픽모델링 실행에서 첫 번째로는 매일의 일기를 문서(document)로 지정하여 9,468일 전체 텍스트뭉치에 나오는 3,619개 어휘를 R기반 topicmodels 패키지로 분석하여 일기전반의 내용을 조망해보았다. 토픽수는 추출된 토픽의 해석가능성을 위주로 40개로 정하였다.[9] 각 토픽별 구성어휘는 〈부표 1〉과 같다.

(1) 농사관련 토픽

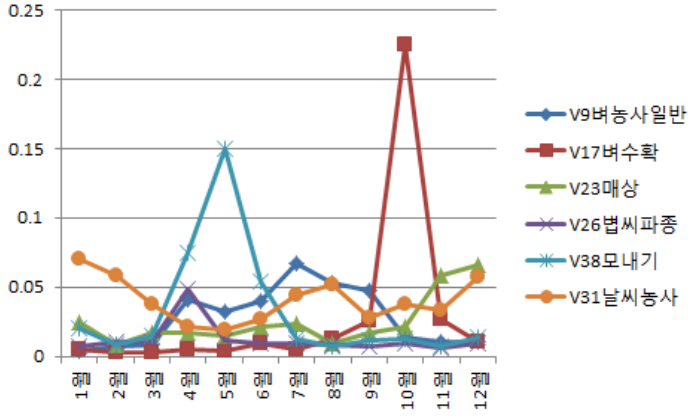
아포일기에서 추출된 40개 토픽들을 전반적으로 살펴보면 농민의 일기답게 농사관련 토픽이 가장 많이 추출되었다. 농사관련 토픽주제들을 보면 벼농사와 관련하여 벼농사일반(9), 벼수확(17), 볍씨, 기타작물 파종(26), 모내기(38)가 추출되었고, 과수 및 채소농사와 관련하여 과일수확 · 자두(8), 과실수확 · 복숭(10), 과수 · 교육(12), 배추무우농사(20), 포도수확(25), 보리농사(29), 노타리 · 파종(33), 과수전지(37) 등의 토픽이 추출되었으며, 그밖에 시비(2), 매상(23), 농약살포(24), 양계(32) 등의 토픽이 추출되었다. 토픽 16은 경운기 · 기계 · 누님 토픽인데 농사용 장비나 기계 및 그 작업과 관련한 인물이 추출된 토픽으로 보인다. 여기에 농업정책에 관한 토픽(1), 농사와 날씨(31)에 관한 토픽을 합치면 40개 토픽 중 절반 가까이가 농사에 관한 것이라고 볼 수 있다.

9) LDA 알고리즘을 실행해주는 R 패키지 'topicmodels'는 내부적으로 VEM(variational expectation - maximization) 추정법과 Gibbs 샘플링을 이용한 추정법을 제공해주는데 각 추정법을 사용하여 모델을 추정할 때 참고할 수 있는 펄플렉시티(perplexity) 값을 산출해준다. 펄플렉시티 값이 낮을수록 좋은 모델이라고 할 수 있다. 본 연구에서는 날짜별 전체일기 분석에서는 VEM_estimation 추정법을, 그리고 뒤에 나올 세 자녀문서 분석에서는 Gibbs 샘플링에 의한 추정법을 선택하였다. 이는 전자는 토픽수 40 근처에서 낮은 펄플렉시티값을 보이고 후자는 토픽수 15 근처에서 낮은 펄플렉시티 값을 보였기 때문이다. 토픽수의 결정은 토픽의 해석가능성과 추정법별 펄플렉시티 값을 동시에 고려하여 결정하였다. Gibbs 샘플링의 경우 반복회수는 1000으로 지정하였다(Grun et al. 2011: 7-8, 12-18).

다음으로는 자녀 및 가족 관련 토픽들이 많이 추출되었는데 훈이군대(14), 훈이 · 공부(19), 공부 · 대학(22) 등이 직접적으로 자녀에 관한 토픽들이고, 부인사랑고충(11), 집사람 · 감기몸살(27)은 부인에 관한 토픽이며, 부모님 · 친족(6), 어머님 · 생신 · 처가(28), 형제(39)는 부모와 형제자매에 관련된 토픽이다. 그리고 친지애경사(13)에 관한 토픽 역시 이와 관련된다고 볼 수 있다. 다음에 가족의 질병과 병원 치료에 관한 토픽(35)이 있으며, 여행(5), 농한기 놀이(30), 각종공사(18), 조합 · 부역(4) 등이 발견된다. 그리고 농업정책과는 별개로 정치(7) 토픽이 별도로 추출되었으며, 인간, 생활, 세상 등의 어휘로 이루어진 인생관(15) 토픽이 뒤를 잇는다. 그리고 아포일기에는 일기 하단에 지출항목을 적기 때문에 지출항목 관련 토픽이 2개 추출되었다. 그런데 토픽 21은 자녀선물에 관한 내용과 물류단지 유치반대 활동들에 관한 어휘들이 뒤섞여 있으며, 토픽 3은 태풍, 예보 등 기상에 관한 어휘들과 현주 시집에 관한 어휘들이 뒤섞여 있다. 토픽 34 역시 과수, 전정 등의 작업과 훈이 회사에 관한 어휘가 섞여 있다.[10)]

토픽모델링에서 토픽수를 결정할 때 가장 중요한 기준은 해석가능성과 타당도 그리고 연구질문에 비추어본 유용성이다. 아포일기에서 가장 빈번하게 나타나는 농사관련 토픽들의 타당도를 간접적으로 검증해 보기 위해서 월별로 농사관련 토픽의 출현빈도를 비교해보았다. 먼저 벼농사관련 토픽들의 월별출현빈도를 보면, 5월의 모내기와 10월의 벼수확 토픽이 가장 눈에 띈다. 벼농사 중에서 집중도가 가장 높은 것이

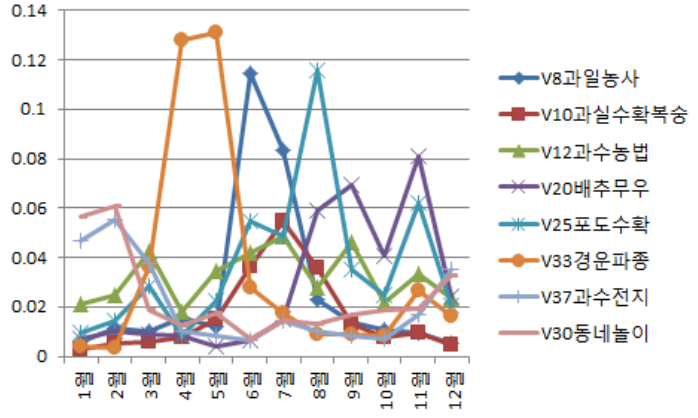
10) 이 세 토픽은 오분류되었을 가능성이 있다. 토픽모델링 기법은 분류가 어려운 어휘들은 일부 토픽에 몰아넣는 방식으로 다른 토픽들의 해석가능성을 높이기 때문에 추출되는 토픽항목들 모두가 해석가능성이 높지는 않다(DiMaggio et al. 2013: 583). 다만 미처 파악하지 못한 새로운 잠재적 의미 구조를 보여주는 것일 가능성을 완전히 배제할 수는 없다. 또한 연구자가 물류단지 유치 반대운동에 관심이 있을 경우에는 토픽의 수를 늘여서 현재 자녀·선물 관련어휘들과 혼재되어 구성된 토픽 21이 어떻게 변화하는지 관찰해 볼 수도 있고, 토픽 21이 주로 등장하는 텍스트만 추출하여 가까이 읽기를 통해 세밀하게 분석해볼 수도 있을 것이다.



〈그림 1〉 월별 벼농사관련 토픽분포

모내기작업과 수확작업임을 감안하면 이러한 결과는 두 토픽의 타당도가 매우 높다는 것을 보여준다. 그리고 상대적으로 집중도는 약하지만 4월에는 볍씨 · 기타작물 파종 토픽이 높은 비율을 보이며 벼농사 일반 토픽은 4월부터 9월에 걸쳐 장기간 상대적으로 높은 비율을 보이고 있다. 마지막으로 매상 토픽은 11~12월에 높은 비율로 나타나는데 이는 이 시기가 추수가 끝나고 본격적인 추곡매상이 이루어지는 시기이기 때문으로 짐작된다. 그리고 날씨에 관한 토픽은 12월부터 2월까지의 겨울철과 7~8월의 여름철에 빈도가 높은데 대체로 추위나 더위 · 장마에 관한 어휘가 주를 이룬다는 점을 감안하면 이러한 결과 역시 타당한 것이라고 여겨진다.

〈그림 2〉의 과수 및 채소농사 관련 토픽을 보면, 경운(노타리) 및 파종 토픽은 4~5월에, 과일농사 · 자두(8)와 과실수확 · 복숭(10)은 6~7월에 포도수확(25)은 8월에 높은 비율을 보인다. 흥미로운 것은 배추무우와 관련된 토픽(20)인데 수확기인 11월이 가장 빈도가 높고 다음이 파종기인 8~9월로 가을 김장채소의 수확기 및 파종기와 정확하게 일치한다. 그리고 12월과 1, 2월에는 과수전지작업 토픽(37)이 상대적으로



〈그림 2〉 월별 과수 및 채소 관련 토픽분포

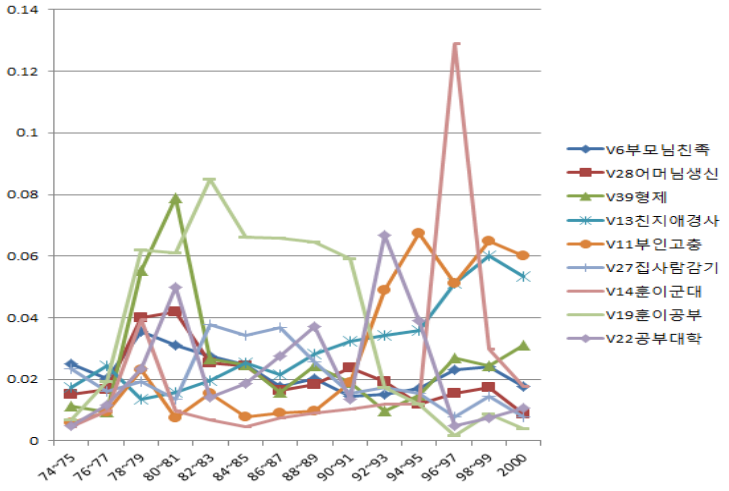
높은 비율을 보인다. 또한 농한기인 이 시기에는 농한기 놀이 토픽(30)도 높은 비율로 나타난다.

한편 농사 관련 토픽들의 '년도별' 변화 추세를 살펴보면 과수나 채소농사의 경우 시기의 변화에 따른 작목의 변화를 뚜렷하게 파악할 수 있으며, 벼농사의 경우 기계화에 따른 모내기노동의 감소가 확연히 드러난다. 이를 통해 장기간에 걸친 농사 관련 토픽의 변화추세 속에서 농업 기계화나 영농방식의 변화, 농산물시장의 변화와 연관된 작목의 변화나 농업노동과정의 변화를 검토해 볼 수도 있다. 다만 여기서는 지면의 제약상 상세한 논의 및 도표의 제시는 생략하기로 한다.

토픽모델링의 결과로 추출된 토픽들의 타당도를 측정하는 방법 중의 하나는 텍스트 외부의 변수와 토픽과의 관계를 비교해서 외적타당도를 살펴보는 것이다. 여기서는 농사일은 계절성을 강하게 띤다는 특성을 이용하여 문서(매일의 일기)의 기록일자라는 외부정보와 농업관련 토픽들의 관계를 주로 비교해보았으며, 그 결과는 농업관련 토픽들의 (외적)타당도가 매우 높다는 것을 보여주고 있다.

(2) 가족관련 토픽

아포일기에서 농사 관련 토픽 다음으로 빈번하게 나타나는 것은 부모, 형제, 부인, 자녀들에 관련된 토픽들이다. 년도별로 가족관련 토픽의 분포를 보면 원가족에 관련된 토픽들[부모님친족(6), 어머님 · 생신 · 처가(28), 형제(39)]은 대체로 자녀들이 초등학교에 입학하기 전인 1981년경까지 많이 나타나지만 이후로는 점차 감소한다. 대신 그 자리에 자녀들의 공부에 관한 토픽[훈이 · 공부(19), 공부 · 대학(22)]과 부인에 관한 토픽이 들어선다. 부인에 관련된 토픽으로는 훈이 · 공부 토픽에 부인이 중요하게 등장하는 것을 별도로 하면 집사람 · 감기 토픽(27)은 주로 1981~87년 사이에 나타나고 부인 · 고충 토픽(11)은 1992년 이후 나타난다. 집사람 · 감기 토픽은 부인이 휴식도 없이 큰집, 형님집에 가서 탈곡 타작 상자 나르기를 하고 자녀 돌보기 등으로 피로하여 몸살감기로 고생하는 내용들이 구체적으로 묘사되어 있다. 반면에 후반부에 주로 나타나는 부인 고충 토픽은 부인의 농사일이나 가사 및 자녀에 관련된 구체적 노동의 내용보다는 남편으로서 느끼는 부인에 대한 사랑과 건강 염려, 과중한 피로, 고충, 고통에 대한 미안함, 여자로서 잠자리, 가족을 위한 노력에 대해 느끼는 따뜻한 심정과 행복을 토로하고 있다. 전자에서는 자기착취적 노동에 동참시켜야만 했던 부인의 과중한 노동일과 고통에 대한 직접적인 묘사가 주를 이루는 반면 생활이 다소나마 안정되기 시작한 후반부에서는 부인의 고통과 피로에 연민과 건강에 대한 걱정, 가족을 위한 노력에 대한 감사와 거기서 느끼는 행복감 등 주로 감정적 · 정서적 묘사가 주를 이루고 있다. 그리고 이 무렵부터는 결혼식, 예식장, 식사, 대접 등으로 이루어진 애경사 토픽(13)이 증가하기 시작한다. 이정덕 외(2014: 2-3)에 의하면 아포일기 저자 권순덕의 특징은 가족, 자녀, 형제들을 범위로 하는 가족주의 이념이라고 지적하는데, 년도별 토픽 분포의 결과에서도 그러한 특징은 뚜렷하게 드러난다. 아포일기의 인간관계에 대한 언급은 초기에는 원가족 부모 형제자매에 관



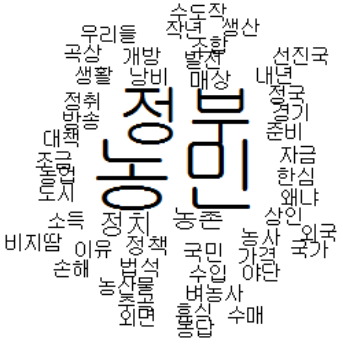
〈그림 3〉 년도별 가족관련 토픽분포

한 토픽이, 자녀가 성장하면서는 자녀와 부인에 관한 토픽이 주를 이루며 가족 이외의 사람들에 대한 토픽은 발견되지 않는다. 다만 생활이 안정된 후 증가하는 애경사 토픽에는 원가족이나 생식가족 이외에 친구 동기생이 등장하기도 하며 농한기놀이 토픽에서도 친구라는 어휘가 등장하지만 기본적으로 아포일기의 토픽 분포에서도 가족주의적 성향이 강하게 드러난다.

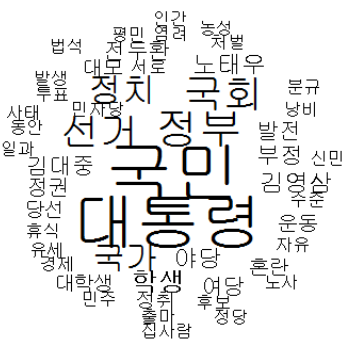
(3) 정치 정책 관련 토픽

농사 관련 토픽이나 가족 관련 토픽 이외에 흥미로운 것은 농업정책 토픽(1)과 정치 토픽(7)이다. 농업정책 토픽은 농민, 정부, 정치, 농촌, 정책, 국민, 법석, 매상, 수입, 가격, 농사, 한심, 상인, 낭비, 준비, 농산물, 발전, 벼농사, 정취(정치) 등의 어휘로 구성되어 있으며, 정치 토픽은 국민, 대통령, 정부, 국회, 선거, 정치, 국가, 학생, 야당, 김영삼, 노태우, 부정, 김대중, 발전, 전두환, 정권, 여당, 정취, 혼란, 운동 등의 어휘로 구성되어 있다. 두 토픽 모두에 국민, 정부, 정치라는 어휘가 커다란 비

중으로 포함되어 있으나 농업정책 토픽에서는 농민, 농촌, 정책, 매상, 수입, 가격, 농사, 한심, 낭비 등의 어휘와 조합을 이루고 있어서 주로 구체적인 농민과 농촌의 현실이라는 맥락 속에서 언급되는 반면, 정치 토픽에서는 대통령, 국회, 선거, 국가, 학생, 야당, 김영삼, 노태우, 부정, 김대중, 발전, 전두환, 혼란, 운동 등의 어휘와 조합을 이루어서 언급되고 있어서 상이한 잠재적 의미 토픽을 이루고 있다. 농촌과 농업에 관하여 정부, 정치를 논의할 경우에는 농민, 농촌, 농사의 현실경험에 기반하여 외국, 선진국, 수입, 개방에 대한 대책을 제대로 수립하지 못하여 추곡, 매상, 가격 측면에서 손해가 크고 한심하다면서 도시에 비해 농가는 비지땀을 흘려도 타격이 크며 영농의 어려움은 정부의 외면과 방치, 정책 실패의 탓이 크다고 불만을 제기한다. 한편 정치 토픽에서는 정당(야당, 여당), 정치인(김영삼, 김대중, 전두환, 노태우), 국민, 대통령, 국회, 선거, 국가, 발전 등의 어휘가 중심을 이루고 있으며, 학생, 운동, 데모는 자유, 혼란, 낭비, 경제, 염려, 농성, 분규, 노사, 법석, 사태 등의 어휘들과 조합을 이루어 정치는 정당과 정치인, 대표자(대통령)가 중심이고 국민은 선거를 통해서만 참여하는 것이라는 인식을 보여주고 있으며, 학생, 운동, 데모, 노사, 분규에 대해서는 자유가 혼란, 낭비, 법석, 사태,



〈그림 4〉 농업정책 토픽(1)

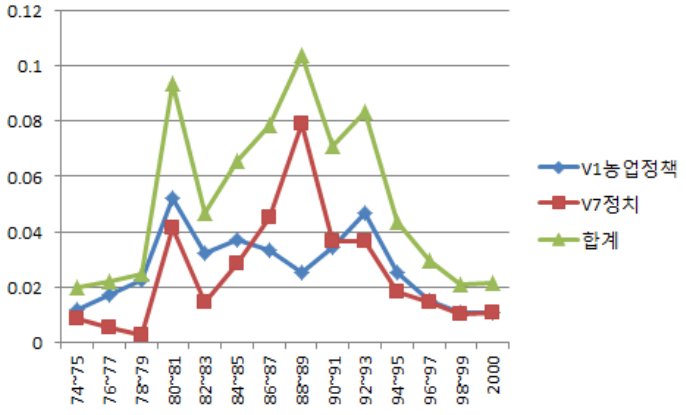


〈그림 5〉 정치 토픽(7)

경제, 염려로 이어지고 있어서 부정적인 시각을 드러내고 있다.

편의상 토픽 1을 농업정책으로, 토픽 7을 정치로 명명하였으나 두 토픽에는 공통적으로 국민, 정부, 정치라는 어휘가 빈도 순으로 6위 내에 포함되고 있어서 국민과 정부 정치의 관계에 대한 견해를 보여주고 있다. 전자에서는 농민과 농촌의 어려운 현실은 농민 개개인의 탓이라기보다 정부의 농업정책 실패라는 구조적 요인에 기인한다는 관점을 보여주고 있다. 그런데 후자에서는 국가의 대통령 국회 정부를 국민의 선거로 구성하는 것이 정치이며 여기서는 정당(여당, 야당)과 정치인이 주요 행위자이고 국민의 정치참여는 선거를 통해서만 이루어져야 하며 학생이나 노동자의 시위와 같은 정치참여방식에 대해서는 부정적 시각을 드러내고 있다. 정치가 무엇인가, 국민 혹은 농민과 정치의 관계는 무엇이며 국민의 정치참여는 어떠해야 하는가에 관해서 아포일기가 보여주는 이러한 시각들은 87년 체제의 성립배경 및 그 성취와 한계를 이해하는 데 중요한 기반을 제공하고 있는 것으로 여겨진다.

두 토픽의 년도별 분포를 살펴보면 또 한 가지 흥미로운 사실을 포착할 수 있다. 먼저 매일의 일기에서 두 토픽이 나타난 비율을 합계해서 년도별 평균추세의 변화를 살펴보면 80~81년, 88~89년, 92~93년의 세 꼭지점이 발견된다. 이 시기는 각각 박정희정권의 군사독재가 끝난 직후, 87년 민주화운동 직후, 92년 문민정부 이행기에 해당한다. 다만 농업정책 토픽(1)과 정치 토픽(7)을 분리해서 살펴보면 80~81년과 92~93년의 작은 꼭지점은 유사하지만 88~89년의 경우에는 정치일반 토픽(7)이 급상승한 반면 오히려 농업정책 토픽(1)은 감소하고 있다. 전체적으로 보면 80년의 봄, 87년 민주화 이후, 92년 문민정부 이행기에는 억압되어있던 정치적 담론공간이 확장되어 활발한 정치적 의사표현이 이루어졌던 것으로 보이며, 80년과 92년에는 본인이 몸담고 있는 농촌의 현실과 관련하여 농민과 농촌, 정부와 정치의 관계를 논의하는 서술들이 많았던 반면에 87년 민주화 이행기에는 농민, 농촌, 정부에 관한 구



〈그림 6〉 년도별 정치관련 토픽분포

체적 논의들이 사회전반의 정치체제 변화에 대한 논의들 속에 흡인되어 들어갔던 것으로 판단된다.

농업정책과 정치 토픽의 어휘구성과 년도별 분포추세의 변화에 바탕을 둔 위와 같은 해석을 좀 더 정밀하게 확인해보기 위해서는 해당되는 일기의 기록을 찾아가서 전통적인 가까이서 읽기의 방법으로 텍스트 분석을 수행해볼 필요가 있다. 이를 위해서는 각각의 날짜별 일기 속에서 농업정책 토픽이나 정치 토픽의 구성비율이 일정 비율보다 큰, 예컨대 50%를 넘는 날짜를 추출하여 해당 날짜의 일기를 추출할 수 있다. 물론 정부나 정치관련 일기를 찾기 위해서는 정부나 정치라는 핵심어를 가지고 해당 텍스트를 추출할 수도 있으나, 핵심키워드 검색 방법에는 한계가 있다. 왜냐하면 앞에서도 보았듯이 농업정책 토픽의 핵심어휘들은 농민, 정부, 정치, 농촌, 정책, 국민, 법석, 매상, 수입의 순으로 구성되어 있어서, 어떤 날의 일기에는 농업정책을 서술하고 있더라도 정부나 정치라는 어휘는 들어있지 않을 수도 있기 때문이다. 마찬가지로 정치 토픽을 다루는 일기에서도 정부와 정치라는 어휘는 없이 국민, 국가, 학생, 야당, 부정, 발전, 혼란, 운동, 처벌, 염려 등으로만 구성되어 있을

수도 있기 때문이다.[11]

여기서는 농업정책 토픽과 정치 토픽이 서술된 해당 날짜의 일기를 찾아서 분석하는 대신 아포일기 전권을 세밀하게 통독하여 전통적인 텍스트분석방법으로 분석한 손현주(2015: 106-112)의 연구결과와 비교하여 토픽모델링 기법에 기반한 위의 발견과 해석들이 과연 타당성을 가지는지 검토해보기로 하자. 아포일기에 나타난 저자의 근대적 경험을 문화적 측면, 정치적 측면, 기술적 측면으로 나누어서 고찰한 손현주에 의하면 권순덕은 80년대 이후 사적인 주제에서 공적인 주제로 관심의 폭을 확장하며, 농촌의 산적한 문제의 원인을 개인의 책임에서 사회구조적 문제로 생각하는 인식의 전환을 보여준다. 70년대까지는 투표행위만 하였으나 80년대 중반 이후 92년까지는 국회의원선거나 대통령선거 유세집회에 참가하기도 하면서 정치의식을 표출한다. 그렇지만 권순덕의 정치참여는 투표, 선거, 정치집회 참여로 국한되며 데모와 시위 같은 정치참여 행위에 대해서는 정치를 혼란하게 하고 국가 전체의 안전을 파괴하는 것으로 간주한다. 권순덕에게 민주주의는 직업정치인들이 중심이 되는 대의제로 간주된다. 정치는 국가가 정책을 실행하는 방향과 방식이라고 보는 협소한 의미의 정치 개념을 가지고 있으며 정부정책의 실패 특히 농업정책의 실패에 대해서는 강한 불만을 가지고 비판하지만, 개인이 다양한 정치활동에 적극적으로 참여하여 정부의 권력을 제한하고 개인의 자유를 최대한 보장하려는 움직임에 대해서는 비판적이다.

이러한 손현주의 연구에서는 정치일반과 농업정책(혹은 농민 · 농

11) 남북전쟁 당시의 신문을 연구한 넬슨(Nelson 2010)은 신문에 난 도주노예현상금 광고의 숫자를 파악하기 위해서 신문기사의 토픽모델링으로 추출한 도주노예 토픽의 비율이 30%가 넘는 기사를 도주노예광고로 간주하고 년도별 분포를 그려보았다. 그리고 이를 수작업으로 도주노예광고를 세어본 결과와 비교한 결과 양자가 거의 일치된 추세를 보인다는 점을 밝힘으로써 토픽모델링 기법이 신문에서 특정한 분야의 기사를 검색하는 유용한 방법임을 입증하였다. 넬슨은 “도주노예”라는 키워드로만 검색할 경우에는 도주노예현상금 광고가 아니라 도주노예에 관한 일반기사가 검색될 가능성이 크기 때문에 토픽모델링의 방법이 더 효과적이라고 주장한다.

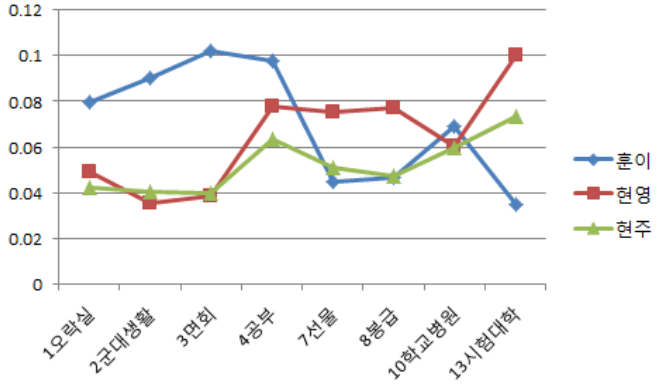
촌·정부·정책) 토픽이 분화되어 나타난 본고의 연구 결과와는 달리 양자가 함께 다루어지고 있으나, 후자의 토픽과 관련하여 강하게 정부 정책의 실패를 비판한다는 점, 정치일반에서는 선거와 투표로 정치참여의 범위를 제한하며 학생이나 노동자들의 다양한 적극적 정치참여에 대해 부정적이라는 점, 80년대 중반 이후 92년까지 적극적인 정치의식의 개진이나 정치참여가 나타난다는 점 등에서는 본고의 토픽모델링에 기반한 해석과 일치하고 있어서 일기에 토픽모델링을 적용한 본 연구의 해석이 상당한 타당성을 지니고 있음을 보여준다.

3) 자녀관련 일기의 문단별 토픽분석

앞에서 살펴본 것처럼 날짜별로 분석한 아포일기에는 자녀들에 관한 토픽이 높은 빈도로 나타난다. 이 절에서는 문서의 단위를 날짜단위가 아니라 문단단위로 좁혀서 자녀들이 언급된 문단을 추출하여 토픽모델링을 실시해봄으로써 보다 세부적인 자녀관 내지 자녀 양육관을 살펴보고자 한다. 토픽모델링 알고리즘은 동일문서에서 나타나는 단어들의 공통출현빈도에 기초하여 의미있는 토픽을 추출하기 때문에 문서의 단위가 작아질수록 세부적이고 구체적인 토픽이 추출되고 문서의 단위가 커질수록 포괄적인 토픽이 추출된다.[12] 토픽의 수는 해석가능성을 고려하여 15개로 정하였다.

자녀의 이름이 언급된 문단만 뽑아서 분석하였지만, 자녀와는 다소 거리가 있는 토픽들도 나타났다. 이중에서 권순덕의 자녀관이나 교육관

12) 토픽모델링에서 토픽의 수를 얼마로 할 것인가 다음으로 연구자가 결정해야할 요소가 문서의 단위를 어떻게 정할 것인가이다. 이 역시 토픽의 해석가능성, 연구질문에 비추어 본 유용성에 비추어 결정할 수밖에 없다. 신문기사등에서는 대체로 기사꼭지를 그대로 한 문서로 사용하므로 큰 문제가 없지만, 원자료에 뚜렷한 단위가 없는 소설 등에서는 문서단위를 무엇으로 할 것인지가 중요한 결정사항 중 하나이다. 토픽 수가 많으면 훨씬 더 세밀한 토픽들이 추출되는 것처럼 문서의 단위가 작아지면 그만큼 세부적인 토픽들이 추출될 가능성이 크다(Jockers 2014: 137-142).



〈그림 7〉 자녀별 주요 토픽분포

을 잘 살펴볼 수 있는 토픽 9개를 골라서 해당 토픽별 구성어휘를 관련이 높은 순으로 제시해본 것이 〈부표 2〉이다. 〈그림 7〉에서 3자녀가 언급된 문단별로 토픽의 분포를 보면 군대생활이나 군대면회에 관련된 토픽은 아들인 훈이문서에서 가장 빈번하게 나타났으며, 오락실, 성질, 약속, 꾸중, 잘못, 거짓말 등으로 구성된 오락실(1) 토픽도 훈이문서에서 빈번하게 나타났다. 이는 오락실에 들락거리는 아이에 대한 꾸중과 걱정 훈육을 담은 토픽이라고 하겠다.[13] 반면에 선물, 엄마, 생일, 졸업 등으로 이루어진 선물(7) 토픽은 공부를 잘해서 상을 타거나 장학금을 받았던 현영문서에서 가장 높은 비율로 나타났으며 그 다음이 현주문서이다. 회사에 취직하여 봉급을 타고 적금을 하는 내용으로 구성된 봉급 토픽(8)도 현영문서에서 가장 많이 나타났으며 학교, 선생님, 병원, 상처, 치료 등으로 구성된 학교 · 병원 토픽은 세 자녀에게서 비슷한 빈도로 나타났다. 공부에 관한 토픽은 4와 13이 있는데. 13번 토픽은 시험, 전문대학, 공부, 학교, 불안, 걱정, 등록금, 전기, 노력, 접수, 지원, 합격,

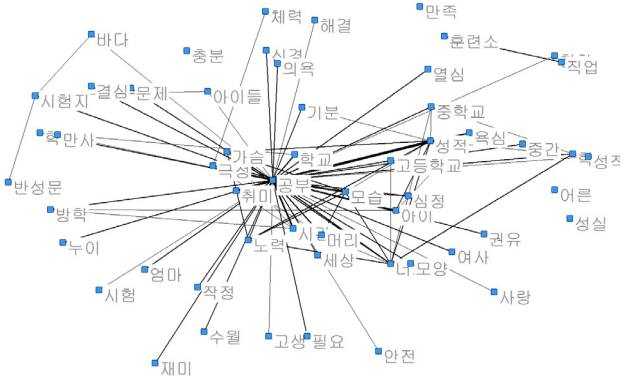
13) 자녀별 학교시기별로 추세를 살펴본 결과에 의하면 훈이문서 중에서도 훈이 중·고시기문서에 오락실 토픽이 가장 빈번하게 나타났으나 지면 제약상 도표의 제시는 생략한다.

소원, 장학생 등으로 구성되어서 구체적으로 대학교 지원을 둘러싼 주제임을 알 수 있다. 반면에 4번 토픽은 공부, 걱정, 학교, 노력, 머리, 시간, 성적, 학원, 욕심, 결심, 엄마, 중(고등)학교, 방학, 문제, 시험, 뒷바라지, 체력 등으로 이루어져서 중·고등학교 시기까지의 공부에 관한 내용임을 알 수 있다. 전자는 현영문서에 가장 높은 빈도로 나오는 반면 후자는 차이는 적지만 상대적으로 훈이문서에 높은 빈도로 나타난다. 상대적으로 공부를 잘해서 장학금을 받기도 했던 둘째딸 현영과 관련하여 전기대학교 입학을 기대하는 등 관심을 많이 보였기 때문에 시험·대학 토픽은 현영문서에 가장 빈번하게 나타나고 다음이 현주문서였던 반면에 어려서부터 공부에 열의를 보이지 않았던 아들 훈이문서에는 구체적으로 대학진학과 관련한 토픽은 상대적으로 자주 보이지 않는다. 그렇지만 중·고시기의 공부에 대한 관심과 걱정은 오히려 아들 훈이문서에 더 자주 나타났다.

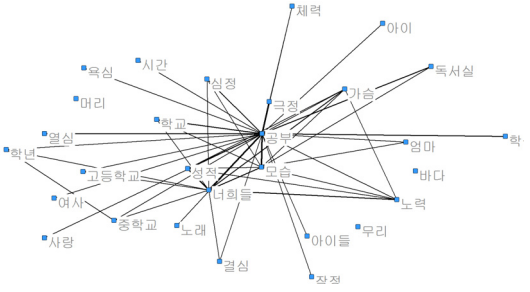
토픽모델링 기법은 개별문서별 단어별로 토픽을 할당해준다. 따라서 이번에는 공부 토픽(4)에 할당된 각문서의 단어를 추출하여 자녀별로 4번 토픽을 구성하는 어휘들의 네트워크를 그려보았다.[14] 네트워크 그래프를 비교해보면 훈이문서의 경우 62개 노트에 연결선(ties)은 176개인 반면, 현주는 29개 노드 94개 연결선, 현영은 20개 노드 76개 연결선을 보였는데 이는 훈이의 경우가 등장하는 어휘의 총수가 가장 많고 어휘간의 연결정도도 긴밀하다는 것을 보여준다.

딸들의 공부 토픽에서는 공부가 가장 중심적 어휘이며 너희들, 모습, 노력, 성적, 가슴, 심정, 걱정, 학교 등이 상대적으로 연결선이 많다. 그런데 훈이 그래프에서는 공부가 중심을 이루는 점은 동일하지만, 딸들의 경우와 달리 중간 정도의 연결선을 가지는 노드들에 추가로 취미,

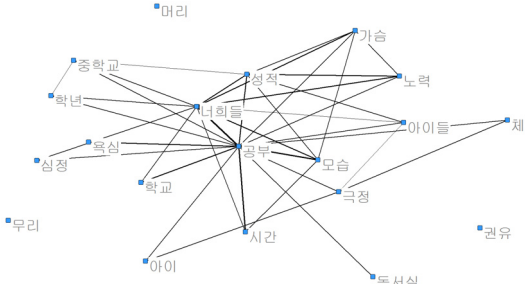
14) 공통출현빈도 4회 이상인 어휘만 골라서 노드로 삼고 여기서 연결선이 없이 고립된 노드들은 제거하였다. 노드(어휘) 간의 거리는 공통출현빈도를 사용하되 연결선은 출현빈도가 8회 이상인 경우만 그려보았다.



〈그림 8〉 공부일반 토픽 핵심어휘 네트워크(훈이)



〈그림 9〉 현주 공부 토픽 핵심어휘 네트워크



〈그림 10〉 현영 공부 토픽 핵심어휘 네트워크

시간, 머리, 아이, 기분, 시험지, 성적표, 고등학교, 중학교, 방학, 학년 등의 어휘가 보이며, 공부와 직접 단선으로 연결된 어휘도 훨씬 많다. 이는 아들의 공부와 관련해서는 훨씬 구체적이고 다양한 어휘들을 다양한 조합으로 동원하여 서술하고 있음을 보여준다.

마지막으로 세 자녀의 교통이동관련 토픽(15)의 분포를 살펴보기 위하여 세 자녀의 교통이동관련 어휘들의 네트워크상의 연결정도중앙성(degree centrality)을 분석해본 결과는 〈부표 3〉과 같다. 어휘의 의미는 결국 함께 사용되는 어휘들과의 맥락에서 나온다는 점을 감안하면 해당 토픽의 내용은 연결정도중앙성이 큰 어휘들에 의해서 좌우될 것이므로 각 어휘의 연결정도중앙성을 살펴보면 해당 토픽의 의미를 파악할 수 있다. 우선 자녀별로 교통이동 토픽 전체의 어휘 구성을 보면 흥미로운 점은 아들문서의 교통이동 토픽 해당 어휘가 만나는 사람이나 목적, 장소, 교통편 등의 측면에서 가장 다양하고 많으며 딸들은 이동의 목적이나 상대, 교통편 모두에서 다양성이 적다는 것이다.

현영문서의 경우에는 연결정도중앙성이 큰 어휘들을 살펴보면 택시나 버스를 이용하여 고사시간에 교실에 도착하는 내용이 전부임을 바로 알 수 있고 하단에 여자, 친구, 남자 등 만나는 상대에 관한 어휘가 나온다. 현주의 경우에는 연결정도중앙성이 100이 넘는 어휘는 전혀 없으며 친구, 시간, 김천, 남자, 마을 등이 상대적으로 중앙성이 높아서 친구, 남자를 만나기 위해 마을에서 김천으로 이동하는 내용이 중심을 이룸을 짐작할 수 있으며, 중앙성이 낮아지면서 총각, 여자, 데이트 등의 대상 및 목적과 구미, 대구 등의 이동지역이 나오고 마지막에 버스, 택시, 구입, 차표 등의 교통편 및 이용관련 어휘가 등장한다.

반면 아들 훈이문서에서는 연결정도중앙성이 100이 넘는 어휘가 훨씬 많다. 상대적으로 연결정도중앙성이 가장 높은 단어들은 친구, 시간, 여자, 김천, 눈물, 택시, 도착, 이용 등이어서 아들의 교통이동 토픽도 친구, 여자를 만나러 김천에 택시로 이동하는 것이 주된 의미 맥락임을

알 수 있다. 그 후로 등장하는 교통편을 보면 버스, 승용차, 열차, 자가용 등으로 다양하고, 구미, 약전, 골목, 대구, 마을, 논산, 대전 등 지역도 더 넓게 퍼져 있고, 접수, 사정, 약전, 약속, 퇴소식, 금오산, 남자, 아파트, 자랑, 한약 등 이동의 목적도 다양하다. 근대화의 가장 큰 축의 하나가 교통 통신의 발달이라고 할 때 교통 및 이동에 관한 토픽의 자녀별 차이는 아포일기 저자의 자녀관을 극적으로 보여주는 것이라고 판단된다.

아포일기에 나타난 자녀관을 전통적인 텍스트 분석방법으로 고찰한 진양명숙(2015: 54~56)에 따르면 권순덕의 자녀관에서 가장 뚜렷한 특징은 '딸보다는 아들이다'로 요약되는데, 이는 공부 토픽(4)이나 교통이동 토픽(15)의 네트워크 분석에서도 잘 드러난다. 상대적으로 어려서부터 공부에 열의를 보이지 않는 아들에 대해서 포기하기보다는 오히려 공부를 잘하는 딸들에 비해서 더 많은 걱정을 하고 더욱 구체적이고 다양한 방법으로 채찍질하고 설득하고 격려하고 지원하며 그 과정에서 훨씬 많은 어휘들을 다양한 조합으로 구사하면서 서술하는 것으로 보인다. 또한 이동과 교통이라는 측면에서도 아들의 경우에는 다양한 목적으로 다양한 교통수단을 이용하여 다양한 지역으로 이동하는 것을 딸들에 비할 수 없을 정도로 상세하게 기록하고 있다는 점에서 토픽모델링의 결과는 진양명숙의 연구결과와 여실히 부합하는 것이라고 하겠다.

5. 결론

최근 들어 텍스트자료의 디지털화가 빠르게 진행되면서 엄청난 양의 텍스트 자료에 들어있는 내용을 분석하는 기법으로 토픽모델링의 활용이 증가하고 있다. 이는 해당 텍스트자료의 내용에 관련된 전문적인 사전적 지식에 기반하지 않고도 방대한 디지털 텍스트 자료로부터 의미

있는 잠재적 주제들을 추출해주는 텍스트 모델링 기법의 기능 때문이다.

본 연구에서는 일상생활사에 대한 학문적 관심의 증대에 기인하여 주요한 연구자료로 떠오르고 있는 일반인들의 개인일기에 토픽모델링 기법을 적용할 수 있는지 검토해보기 위하여 경북 김천의 농민일기를 대상으로 토픽모델링 분석을 수행해보았다.

토픽모델링 기법에 의하여 아포일기 전체를 원거리에서 조망해본 결과에 의하면 아포일기에는 벼농사, 과수 및 채소농사, 시비나 농약살포, 농기계 작업, 농업정책, 날씨 등 농사관련 주제들이 절반 이상을 차지하고 있으며, 다음으로는 부모, 형제자매, 부인, 자녀 등 가족에 관한 주제들이 자주 등장한다. 상대적으로 적은 수의 주제들로는 여행과 놀이, 정치, 각종공사, 부역 · 조합 등이 발견되었다.

토픽모델링으로 추출된 토픽들의 해석가능성은 40개 토픽 중 2~3개를 제외하고는 아주 높았다. 일기전체를 날짜단위로 분석한 경우 문서(각 날짜의 일기)의 외부정보인 날짜정보를 이용하여 월별, 년도별로 주제와의 연관성을 검토하여 외적타당도를 검증해보았는데, 강한 계절성을 띠는 농사일 관련 토픽들은 월별분포가 예상대로 나타났으며, 년도별 변화에서도 기계화에 따른 농사작업의 변화나 시장변화에 따른 작목의 변화를 잘 반영하는 것으로 나타났다. 가족관련 토픽들의 경우에도 결혼과 분가, 자녀 학교 입학 · 진학 · 졸업과 취업 및 군입대 등 생애주기에 따른 변화를 그대로 반영하고 있었다. 흥미로운 점은 정치 관련 토픽들인데 여기서도 80년의 봄, 87년 이후 민주화시기, 92년 문민정부 이행기의 정치적 담론공간의 확장이라는 한국사회의 중요한 시기적 변화와 아포일기에 나타나는 정치관련 토픽들의 빈출시기가 정확하게 일치하였다.

마지막으로 아포일기를 전통적인 텍스트분석 방법으로 연구한 타 연구자들의 연구결과와 본 연구의 결과를 비교해 본 바에 따르면 두 가지 연구방법 모두에서 강한 가족주의적 성향이 발견되었다. 아포일기

저자의 인간관계에 대한 관심은 철저하게 원가족중심성을 보이다가 자녀들이 성장하면서 생식가족중심성으로 이어지고 있으며 가족의 범위를 넘어선 인간관계나 사회적 관계에 대한 관심은 단편적, 산발적으로 나타날 뿐이다.

세 자녀가 포함된 문단만 골라서 문단 단위로 토픽모델링을 실시한 결과와 전통적 텍스트 분석의 결과를 비교해보면 "딸보다 아들"이라는 아포일기 저자의 자녀관이 공통적으로 드러났다. 상대적으로 공부를 잘해서 기대를 모았던 딸의 경우에는 대학 시험을 둘러싼 토픽은 빈번하게 나타나지만 대입이라는 특별한 사건 이전에 이루어진 공부일반에 해당하는 토픽에서는 아들의 경우 훨씬 더 다양한 어휘들을 밀도 높게 구사하면서 다방면으로 관심과 기대, 설득, 걱정하는 모습을 보여주고 있다. 특히 주목을 끄는 것은 교통 및 이동 토픽에서 아들과 딸 사이에 커다란 차이를 보이고 있다는 점이다. 아들의 경우 이동의 목적과 상대, 이동지역, 이동수단 등에서 딸들에 비해서 훨씬 다양하게 서술되고 있다. 근대화의 기본축 하나가 교통의 발달과 이에 따른 공간적 · 사회적 이동 범위의 확장이라고 볼 때 교통 및 이동 주제에서 나타난 아들문서와 딸문서 사이의 차이는 중요한 의의를 지닌다. 그런데 전통적 텍스트 분석방법을 이용한 기존연구에서는 "딸보다 아들"이라는 주제에 주목하였음에도 불구하고 교통 및 이동 측면에서의 차이를 발견하지 못하였다. 이는 토픽모델링 기법이 사전적 지식이나 선입견 없이 텍스트의 내용에 접근하게 해주기 때문에 얻어진 장점의 한 예로 판단된다.

아포일기 저자의 정치의식과 정치행위 측면에 대한 분석결과를 보면 협소한 정치관, 선거 이외의 정치참여방식에 대한 부정적 견해, 어려운 농촌 현실과 관련한 국가나 정부 정책에 대한 책임추궁과 비판 등의 측면에서 기존 텍스트분석 방법에 의한 연구와 토픽모델링에 의한 연구결과가 상당 부분 일치한다. 그런데 이러한 정치의식과 정치행위에 대한 토픽이 아포일기에서 빈번하게 출현한 시점이 한국사회의 정치적 변

화의 주요 길목에서였다는 점을 감안해보면 87년 체제로 일컬어지는 한국사회의 정치적 변화의 성과와 한계에 대한 연구에 주요 시사점을 제공한다고 하겠다.

이처럼 토픽모델링의 기법을 활용하여 일기자료 속에서 비교적 해석가능성이 높고, 타당도가 높은 주제를 추출해내고 경우에 따라서는 사전적 지식이 선입견으로 작용하여 간과할 수도 있었던 주제를 발굴해낼 수 있었던 기반은 토픽모델링 기법이 가지고 있는 몇 가지 특징들에 기인한다. 토픽모델링 기법의 특징을 간추리자면 1) 텍스트의 분석 이전에 사전적 지식을 요구하지 않으며 2) 문서를 어휘의 자루(bag or words)로 가정하여 어휘들의 관계 속에서 잠재적 의미구조를 포착하고 3) 전통적인 가까이서 읽기(close reading) 대신 멀리서 읽기(distant reading)를 가능케 해준다는 점이다. 그러나 사전적 지식을 요구하지 않는다는 것은 빅데이터에 과도한 의의를 부과하는 일각의 주장처럼 데이터 자체가 모든 것을 말하기 때문에 더 이상 이론은 필요 없다는 의미는 아니다. 토픽모델링의 기법에서도 토픽 수의 결정이나 문서 단위의 결정에서 가장 중요한 선택기준은 해석 가능성이기 때문이다. 토픽모델링의 결과로 산출된 토픽들의 해석가능성은 해당 토픽을 구성하는 어휘들의 조합으로부터 판단하는데, 이때 누구에게나 쉽게 다가오는 겉으로 드러난 해석가능성보다는 해당 분야의 전문성이 없이는 포착할 수 없는 토픽들의 해석가능성이 더 중요한 경우가 많다. 많은 경우 누구나 쉽게 해석가능성을 포착할 수 있는 주제들의 경우에는 토픽모델링의 기법을 적용하더라도 대부분 기존 연구의 결과나 이미 알려져 있는 내용을 재확인하는 수준에 그칠 가능성이 큰 반면에, 새롭고 흥미로운 주제들은 겉보기에 해석 가능성이 없어 보이는 토픽들에서 나타나기 때문이다. 겉보기에는 쉽사리 해석하기 어려운 단어들이 조합을 이루고 있을 때 과연 그것이 단순한 오분류 때문인지 아니면 어떤 새롭고 심층적인 의미구조를 반영하는 것인지를 판단하기 위해서는 해당 분야에 대한 전문

적 식견과 통찰력이 요구된다. 그런 측면에서 보자면 토픽모델링의 경우에는 해당 분야에 대한 전문적 지식과 주관적 해석능력이 요구되는 지점이 텍스트 분석의 입구에서 출구로 위치이동한 것이라고 볼 수 있다.

토픽모델링이 가능하게 해준 방대한 규모의 텍스트 문서에 대한 '멀리서 읽기(distant reading)' 또한 '가까이서 읽기(close reading)'에 대한 대체물은 아니다. 일기, 특히 일상생활사 연구의 중요한 자료로서 의의가 강조되는 일반인들의 일기의 경우 일기 저자나 해당 일기의 특성에 대한 별도의 정보가 없는 경우가 많으므로 일기 전체를 멀리서 읽기로 조망하여 해당 일기의 특성을 파악할 필요가 있다. 그러나 앞에서도 보았듯이 어떤 토픽이 오분류로 인한 것인지 아니면 잠재적 의미구조를 반영하는 것인지를 파악하기 위해서는 일기의 해당 부분을 추출하여 가까이서 세밀하게 읽어보는 과정이 무엇보다 중요하다. 그리고 대부분의 경우 멀리서 읽기로 일기의 전반적 특성을 이해한 후에는 개별 연구자의 관심 영역에 해당하는 텍스트를 추출하여 가까이서 읽기를 수행할 필요가 있다. 그런데 관심 영역의 텍스트를 추출하기 위해서 흔히 사용해온 핵심키워드 탐색방법은 그 키워드가 해당 일기에 어떤 어휘로 어떻게 기록되어 있는지를 사전에 모두 파악하고 있지 못할 경우 적용하기 곤란하다. 또한 동일한 어휘라도 맥락에 따라서 상이한 의미를 지니는 점을 고려하면 키워드 탐색법은 한계가 분명하다. 그런 점에서 보면 토픽모델링 기법은 방대한 텍스트뭉치로부터 특정 분야의 텍스트를 추출하는 작업에서 키워드 탐색방법의 대안이 될 수 있다.[15]

15) 이와 관련하여 최근에는 LDA의 토픽 분류 알고리즘을 약간씩 수정하여 토픽 추출 시에 특정 핵심어와의 관련성을 고려하여 토픽 추출이 이루어지게 하는 알고리즘도 개발되고 있는데 이는 기존의 자율기계학습(unsupervised Machine Learning)의 특징을 가졌던 LDA 알고리즘을 연구자의 필요에 맞도록 (준)지도기계학습((semi) supervised Machine Learning)으로 변용하려는 노력들의 일환이다. 지도학습은 사전에 정보를 주고 기계학습하도록 하는 방법으로서, 앞의 예에서처럼 핵심키워드를 주는 경우도 있고, 긍정적 부정적 반응의 사례를 주어서 인기도나 감정설명에 이용하기도

다음으로는 문서를 ‘어휘의 자루’로 간주하고 관계적 의미를 추적하는 특징에 관해 살펴보자. LDA 알고리즘이 가진 위의 전제는 토픽모델링 기법이 의미상 해석가능성이 높은 토픽들을 자동적으로 추출해주는 핵심적 기반이다. LDA 알고리즘의 이런 전제로 인하여 다의어, 동음이의어, 일기 저자 특유의 다양한 방언이나 약자사용 등의 문제가 해소될 수 있다. 그리고 영어만이 아니라 여러 나라의 언어에 광범위하게 적용가능하게 된 것 역시 위의 특징에 기인한다. 특히 주목되는 점은 토픽모델링 기법을 사용하면 영어의 경우 광학문자판독기(OCR) 처리과정에서 남게 된 불완전 판독된 찌꺼기들로 인한 문제가 해소되었다는 보고이다.[16] 향후 한글 광학문자처리기의 불완전 판독 찌꺼기로 인한 문제도 토픽모델링을 거쳐 해소할 수 있게 된다면 아직 디지털화되지 않은 수많은 텍스트자료의 분석에 획기적 전기가 될 것으로 기대한다.

다만 ‘어휘자루’ 전제와 관련하여 문제가 되는 것은 특정한 목적의 연구에서는 어휘의 순서나 서사구조가 중요한 의미를 지닐 수도 있다는 점이다(Mohr et al. 2013: 559). 이미 생애사 자료의 서사구조에 기반한 네트워크 분석들이 유용한 결과들을 산출하고 있다는 점에 비추어 보면, LDA 알고리즘에 전제된 ‘어휘자루’ 가정을 완화하여 토픽추출과정에서 어휘의 순서나 서사구조가 고려될 수 있도록 보완될 필요가 있다.

마지막으로 본 연구에서는 일기 텍스트의 분석에서 명사만을 추출하여 사용하였다. 그런데 내면세계 생활세계를 서술하는 개인일기자료

한다. 또한 특정한 시점 주위로 토픽이 군집되도록 하여 역사적 사건에 보다 적합하게 토픽이 추출되도록 하기도 한다(Blei 2012: 82-84; Tangherlini and Leonard 2013: 726-731; Templeton, Brown, Battacharyya and Boyd-Graber 2011: 2-5; Wang and McCallum 2006).

16) 활자화된 텍스트 문서의 경우 광학문자판독기를 거쳐서 디지털텍스트로 전환하는데 판독과정의 불완전성으로 정확하게 판독하지 못한 찌꺼기들이 남게 된다. 그런데 양과 동료들(Yang et al. 2011: 99-100)의 연구에 의하면 불완전 판독된 찌거기들은 그 빈도가 낮고 여러 문서에 무작위적으로 발생하므로 토픽모델링 수행과정에서 자동적으로 해소되어, 광학처리 찌꺼기들을 많은 비용을 들여서 제거한 경우와 별 차이를 보이지 않았다고 한다.

의 특성상 감정이나 정서 표현이 많을 수 있으며 감정이나 정서의 표현은 명사만 분석해서는 포착하기가 쉽지 않다. 향후 일기형태의 자료에 대한 토픽모델링 기법의 적용이 본격화되기 위해서는 명사 이외의 품사들을 활용하는 방안을 포함하여 일기텍스트의 전처리과정에서 당면하는 문제점들에 대한 심도있는 연구가 요구된다.[17]

토픽모델링 기법은 디지털화된 대규모 텍스트 자료의 내용을 분석하는 데 유용한 도구로 활용될 수 있다. 그러나 이제까지 살펴본 것처럼 해당 영역의 전문적 지식이 없이 누구라도 자동적으로 의미있는 연구결과를 도출할 수 있는 것은 아니다. 일기자료를 연구할 경우에도 연구질문에 적합한 일기자료를 선택해야 하며 해당 일기저자와 일기의 특성에 대한 이해가 선행되어야 하며 토픽모델링 기법의 적용으로 추출된 토픽들에 대한 해석과 통찰능력이 필수적으로 요구된다. 다만 토픽모델링 기법은 특정 세부영역에 관심이 있는 연구자에게도 일기 전체를 조망하여 해당 일기의 특성을 파악하는 데 도움을 주기도 하고, 관심영역을 추출하는 데 있어서도 다양한 방법으로 활용될 수 있다. 마찬가지로 멀리서 읽기가 가까이서 읽기의 대체물은 아니다. 그렇지만 토픽모델링 기법은 일기연구자에게 다양한 렌즈를 제공하여 연구과정에서 수시로 일기 전체 수준의 분석과 세부영역 수준의 분석을 오가면서 가까이 읽기와 멀리서 읽기를 자신의 연구과정의 진행에 따라서 적절하게 사용하면서 일기 텍스트에 대해서 여러 위치에서 접근할 수 있도록 해준다. 또한 토픽모델링 기법이 가진 귀납적 방법이라는 특성은 때로는 연구자들이 사전에 예상하지 못했던 잠재적 의미구조를 포착할 수 있도록 이끌어 주기도 한다.

향후 텍스트 자료의 디지털화 과정에서 발생하는 광학문자 판독 상

17) 산업 경영분야에서는 다양한 텍스트 마이닝 기법들을 활용하여 주로 SNS자료를 바탕으로 명사 이외의 품사들까지 사용하여 감정분석을 수행하려는 시도들이 이루어지고 있어서(김윤석 · 서영훈 2013), 일기자료의 토픽모델링 적용에서도 적극적으로 참고할 필요가 있다.

의 난제들이 해소되고, 한글자연어처리과정의 안정성이 높아지고, 명사 외의 품사들의 활용 가능성이 높아지고, 인문사회과학 연구자들이 일기 분석에서 당면하는 여러 가지 요구사항에 맞도록 기존의 표준적 LDA 알고리즘이 수정 보완된다면, 토픽모델링을 이용한 디지털 일기자료의 분석은 보다 활성화될 것으로 기대한다.

논문접수일: 2015년 11월 29일, 논문심사일: 2015년 12월 21일, 게재확정일: 2016년 1월 3일

참고문헌

강범일 · 송민 · 조화순
2013 "토픽 모델링을 이용한 신문 자료의 오피니언 마이닝에 대한 연구,"『한국 문헌정보학회지』47(4): 315-334.

김윤석 · 서영훈
2013 "기계 학습을 이용한 한글 텍스트 감정 분류,"『한국엔터테인먼트산업학회 2013 추계학술대회 논문집』, 한국엔터테인먼트산업학회, pp. 206-210.

김하진 · 정효정 · 송민
2014 "토픽모델링을 통한 저자명 식별 성능 비교,"『제21회 한국정보관리학회 학술대회 논문집』, 한국정보관리학회, pp. 149-152.

니시카와 유코
2014 『일기를 쓴다는 것』, 임경택 · 이정덕 역, 전북: 신아출판사.

박자현 · 송민
2013 "토픽모델링을 활용한 국내 문헌정보학 연구동향 분석,"『정보관리학회지』30(1): 7-32.

손현주
2015 “근대적 경험과 양가성,”『압축근대를 경험하는 동아시아』 전북대학교 개인기록과 압축근대 연구단 심포지움 논문집, pp. 93-121.

이정덕 · 소순열 · 남춘호 · 문만용 · 안승택 · 송기동 · 진양명숙 · 이성호 편
2014 『아포일기 1』, 전북: 전북대학교 출판문화원.

이케다 유타
2014 “역사와 개인기록,” 이정덕 · 안승택 편,『동아시아 일기 연구와 근대의 재구성』, 서울: 논형, pp. 25-33.

정병욱
2013 “식민지 농촌 청년과 재일조선인 사회,” 정병욱 · 이타가키 유타 편,『일기를 통해 본 전통과 근대, 식민지와 국가』, 서울: 소명출판, pp. 263-312.

정병욱 · 이타가키 유타 편
2013 『일기를 통해 본 전통과 근대, 식민지와 국가』, 서울: 소명출판.

진양명숙
2015 “남성 농민일기에 나타난 가부장적 젠더 인식,”『압축근대를 경험하는 동아시아』 전북대학교 개인기록과 압축근대 연구단 심포지움 논문집, pp. 47-66.

한신갑
2015 “빅데이터와 사회과학하기,”『한국사회학』 49(2): 161-192.

Baird, Bridget and Cameron Blevins
2013 “Digital Diaries, Digital Tools: A Comparative Approach to Eighteenth-Century Women’s History,” *Women’s History in the Digital World*, Paper 3, http://repository.brynmawr.edu/greenfield_conference/papers/

Blei, David M.
2012 “Probabilistic Topic Models,” *Communications Of The ACM* 55(4): 77-84.

Blei, David M., Andrew Ng and Michael Jordan
2003 “Latent Dirichlet Allocation,” *Journal of Machine Learning*

Research 3: 993-1022.

Blevins, Cameron

2010 "Topic Modeling Martha Ballard's Diary," http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/

Bonilla, T. and Justin Grimmer

2013 "Elevated Threat and Decreased Expectations: How Democracy Handles Terrorist Threats," *Poetics* 41(6): 650-669.

Broniantowski, D.A. and C.L. Magee

2011 "Towards a Computational Analysis of Status and Leadership Styles on FDA Panels," in J. Salerno, S. J. Yang, D. Nau and S. Chai, eds., *Social Computing, Behavioral-Cultural Modeling and Prediction*, Berlin: Springer Berlin Heidelberg, pp. 212-218.

DiMaggio, P., N. Nag and D. M. Blei

2013 "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture," *Poetics* 41(6): 570-606.

Gerrish, S. M. and D. M. Blei

2010 "A Language-based Approach to Measuring Scholarly Impact," *Proceedings of the 27th International Conference on Machine Learning* (ICML-10), pp. 375-382.

Griffiths, T. and M. Steyvers

2004 "Finding Scientific Topics," *Proceedings of the National Academy of Sciences*, 101(suppl 1), pp. 5228-5235.

Grun, Bettina and Kort Hornik

2011 "Topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software* 40(13): 1-30.

Jockers, M. L.

2014 *Text Analysis with R for Students of Literature*, London: Springer.

Jockers, Matthew L. and David Mimno

2013 "Significant Themes in 19th-century Literature," *Poetics* 41(6):

750-769.

Miller, Ian Matthew
2013 “Rebellion, Crime and Violence in Qing China, 1722-1911: A Topic Modeling Approach,” *Poetics* 41(6): 626-649.

Miner, Gary, D. Delen, J. Elder, A. Fast, T. Hill and R. A. Nisbet
2012 *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Boston: Elsevier Inc.

Mohr, J. W. and P. Bogdanov
2013 “Introduction – Topic Models: What They Are and Why They Matter,” *Poetics* 41(6): 545-569.

Nelson, Robert K.
2010 “Mining the Dispatch,” *Mining the Dispatch*, http://dsl.richmond.edu/dispatch/

Newman, D. and S. Block.
2006 “Probabilistic Topic Decomposition of An Eighteenth-century American Newspaper,” *Journal Of The American Society For Information Science and Technology* 57(6): 753-767.

Rhody, Lisa M.
2012 “Topic Modeling and Figurative Language,” *Journal of Digital Humanities* 2(1), http://journalofdigitalhumanities.org/2-1/

Tangherlini, T.R. and P. Leonard
2013 “Trawling in the Sea of the Great Unread: Sub-corpus Topic Modeling and Humanities Research,” *Poetics* 41: 725-749.

Templeton, Clay, Travis Brown, Sayan Battacharyya and Jordan Boyd-Graber
2011 “Mining the Dispatch under Super-vision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus,” *Chicago Colloquium on Digital Humanities and Computer Science.*

Urlich, L.
1991 *A Midwife’s Tale: The Life of Martha Ballard, Based on Her*

Diary, 1785-1812, New York: Vintage Books.

Wang, Xuerui and Andrew McCallum

2006 "Topics over Time: A Non-Markov Continuous-time Model of Topical Trends," *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM.

Weingart, Scott

2012 "Topic Modeling for Humanists: A Guided Tour," http://www.scottbot.net/HIAL/?p=19113/

Yang, Tze-I, Andrew J. Torget and Rada Mihalcea

2011 "Topic Modeling on Historical Newspapers," *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics.

〈참고자료〉

Jeon, Heewon

2012 "KoNLP: Korean NLP Package," R package version 0.76.5.

〈부표 1〉 토픽별 구성어휘(가중치 높은 순)

1:농업정책	농민,정부,정치,농촌,정책,국민,법석,매상,수입,가격,농사,한심,상인,낭비,준비,농산물,발전,벼농사,정취,방송,이유,조금,소득,개방,농업,도시,대책,자금,왜냐,생활,추곡
2:시비	비료,거름,요소,kg,이삭,단지,복합,웃거름,계획,복판,마지기,작년,밑거름,경화,옷들,복합비료,세집,거름량,봉담,포식,전용,유안,내년,작업,작과나무,앞밭,이차,붕사,훈이,욕심,포도
3:현주시집/기상	현주,기상,태풍,예보,다행,대문,시집,지방,회사,휴식,중앙,정미기,고충,서울,발표,지역,장마,전국,이용,안경,부부,대가,포도,만약,생활,부모,며느리,출근,구입,아이,숙모
4:조합/부역	부역,서류,조합,등기,시청,농로,참석,형님,점심,조합장,아포,보수,외성,회사,보상,폭크랜,등본,우리집,주택,전정,계약,주민,점수,선출,불만,만원,갯돌,운영,동원,책임자,친구
5:여행	구경,버스,여행,관광,택시,열차,도착,점심,제주,부부,집사람,직행,현영,이용,경비,여자,구미,손님,대구,침대,건물,공장,출발,만원,경주,식사,인사,신혼,이곳,마중,형수
6:부모님/친족	어머님,아버지,제사,산소,추석,형제,아버님,참석,준비,고향,명절,순덕,성묘,가정,연개,순복,식구,벌초,분주,동생,조카,큰집,부모님,지붕,구정,집안,할머니,형님,친구,한가,동시
7:정치	국민,대통령,정부,국회,선거,정치,국가,학생,야당,김영삼,노태우,부정,김대중,발전,전두환,정권,여당,정취,혼란,운동,대모,대학생,당선,후보,수준,처벌,출마,민자당,휴식,염려,서로
8:과일수확자두	과일,수학,자두,집사람,작과일,숙모,수입,세집,작업,대석,앞밭,만발,상자,작년,장마,이삭,다고,대전,상인,만원,공판장,마지막,후모사,재미,얼마,소낙비,서울,계획,형님,청과,허무
9:벼농사일반	농사,직파,벼논,건답,살포,마지기,참깨,수학,풍년,도열병,내년,벼농사,파종,문고병,부인,농사가,면적,리을,실패,도살,일품,수입,태풍,물꼬,승년을,피살,벼이삭,이삭,논들,올미,욕심
10:과실수확복숭	과실,수학,계짝,복숭,무지,가뭄,수입,살포,참깨,박패,장마,상자,중복숭,구름,홍자두,더위,작물,하늘,자전,서모사,밭작물,해갈,포도,과수,소나기,벼논,고추,자두,시장,발생,불빛
11:부인사랑고충	부인,남편,건강,사랑,부부,여자,피로,고충,휴식,육체,생활,고통,욕심,잠자리,가정,얼굴,가족,노력,하늘,여유,행동,남자,심정,따뜻,세상,인간,행복,새해,인생,몸부림,종일동안
12:과수교육	사과,살포,교육,벼집,나무,약량,포도,약살포,착색,cc,작과나무,전정,포도밭,부인,희석,작과밭,살충,앞밭,약값,훈이,충분,도장지,농약,강사,멸구,바시내이,포도나무,영농,수입,수도작,마지기
13:친지애경사	음식,친구,참석,부주,결혼식,잔치,손님,동서,예식장,대접,결혼,동내,예식,회갑,준비,모친,형님,서울,축의금,점심,장례,동기생,식당,이발,구미,제공,동네,집안,여자,별세,부모
14:훈이군대	훈이,전화,생활,면회,부인,군대,편지,강아지,부모,주일,휴가,일요일,식구,궁금,준비,고충,훈련,계대,부대,대구,외박,건강,하구,친구,인간,배치,통화,날짜,누나,눈물,훈련소
15:인생관	인간,생활,세상,세월,동내,시멘트,인생,어른,과정,일과,성질,작업,답답,가정,상처,친구,고속도로,종석이,집사람,순달,심정,생명,지붕,여자,공동,한심,어이,성격,부엌,모래,옛날
16:경운기기계누님	경운기,누님,상토,집사람,준비,트랙터,벼집,작업,오토바이,발리,자형,실어,바리,기사,형님,대리점,계획,보일라,붕산,수월,기계,큰집,파이프,조합,세집,고장,전화,순달,동안,주인,고기
17:벼수확	탈곡,콤바이,수학,푸대,작년,농사,벼알,마지기,수확량,형님,볏단,생산,봉담,복판,마당,집사람,말루,가을,단지,일반,건조,벼농사,수월,수확,기계,벼탈곡,허무,옷들,통일벼,콤바인,경화
18:공사	작업,공사,수초,농촌,도량,양수,시작,관정,공구,양수기,도구,대신,우물,바닥,현장,양살구,펌프,포도나무,필요,수월,재방,설치,추석,마무리,배수로,고속도로,이웃,나중,낭비,처음,도시
19:훈이공부	훈이,집사람,공부,도박,선생님,아이들,학교,아이,부모,얼굴,오락실,오락,약속,휴식,일과,꾸중,갈치,거짓말,건강,동안,취미,고생,학년,현주,방학,고통,동내,서로,가정,심정,친구
20:배추무우농사	배추,집사람,무우,수학,채소,경운기,채소밭,농사,김장,구미,시장,민정,무지,휴식,수입,벼집,나물,동생,일과,대파,이식,파종,세집,발리,배추밭,축사,가뭄,큰집,벌레,먹이,여유

21:선물자녀물류단지	선물,현영,훈이,물류단지,유황,묘목,친구,유치,운동,반대,교통,석회,위원장,등록금,선생님,적금,낚시,초대,낭비,아이들,형님,현주,주민,기술,공납금,편지,생활,가연,찬성,정신,선생
22:공부 대학	현주,현영,공부,시험,대학,학교,전문,부모,너희들,졸업,대학교,노력,합격,회사,성적,고등학교,아이,학생,아이들,열심,성질,중학교,훈이,학원,전기,취직,휴식,면접,작업,참석,컴퓨터
23:매상	매상,고추,준비,나락,모종,가격,조합,수입,통일벼,공판,일반,등급,가마,가마니,농사,작년,정부,풍구,검사,검사원,금액,신품종,가연,손해,상량,출하,형님,큰집,동장,중앙,자금
24:농약살포	살포,김매기,잡초,집사람,살충제,제초제,포도밭,제거,땅콩,참깨,사용,이화명충,작업,벌레,무성,일과,농약,지섬,콩씨,콩밭,벼논,cc,입제,재초,봉답,고추,해결,수월,제초,재거,채소밭
25:포도수확	포도,사과,수학,부인,포도나무,수확,싸래단,참깨,농사,포도송이,포도밭,작업,상자,바시내이,수확량,선별,송이,작년,출하,저장,공장,박스,시장,껍질,비가,손질,상인,캄벨,수입,내년,가격
26:벼씨파종고구마	벼씨,파종,마리,고구마,침종,날짜,병아리,개알,수입,모재리,사료,대신,작년,일기장,하계,온상,분만,음력,실패,수정,작업,시작,예정,아포,배치,급릉군,지면,예방,해당,토기,권순덕
27:집사람감기몸살	감기,집사람,상자,몸살,큰집,휴식,준비,일과,감기약,육모,동안,간절,피로,신세,하구,내역,현주,몸부림,감사,지면,아래,현영,기온,내용상,훈이,구둘,탈곡,여유,잠자리,접종,소득
28:어머님생신처가	어머님,고기,부모,생신,눈물,처어머님,인사,집사람,부모님,부친,친정,친구,참석,장모,노인,동생,아버지,어른,형제,얼굴,말씀,돼지고기,토기,우리들,음식,사랑,처남,서운,염소,고생,식사
29:보리농사	보리,가을,동장,수입,두엄,타작,보리농사,씨갑씨,농사,기계,봉답,상업,물건,우마차,가연,보리논,작년,가물,논보리,일과,염려,참외,일부,수학,보리밭,풍년,쟁기,이랑,동내,모양,파종
30:동네 놀이	친구,돌부리,화투놀이,화투,가리,낭비,가을,동네,수입,사랑,담배,윷놀이,동신,술집,휴식,술빵치기,상점,오징어,회원,트랙터,총회,어른,참석,동내,경로잔치,여자,형님,동사,손해,안저,계모임
31:날씨	날씨,휴식,겨울,가뭄,인간,하늘,농민,기온,따뜻,전정,동안,마당,시작,여름,방울,장마,가랑비,추위,변덕,봄날,지속,작업,급새,수입,풍년,밭작물,심정,종일동안,작과나무,눈빨,도움
32:양계업	개알,하계,어머님,수입,예정,병아리,사료,인간,친구,어로,말씀,향군,동생,이변,부화,구루마,객지,공상,성공,과수,없시,부친,계획,노력,보람,실패,과수원,생활,모친,결심,고속도로
33:노타리파종	노타리,파종,마늘,트랙터,두룸,작업,마지기,모판,육모상자,보뚝,모재리,수월,경운기,수입,참깨,기계,비누,집사람,김매기,직파,준비,논갈이,내년,여유,본답,부인,작년,해달라고,모내기,정신,훈이
34:전정/회사	전정,수리,훈이,작과나무,계획,경유,지주,포도,회사,봉급,휴식,소장,트랙터,리터,친구,골목,현영,수입,사회,달희,만원,교환,자격증,작업,방위,곤치,앞발,산업체,드람,남방,금액
35:지출항목1	집사람,송아지,수입,현주,물건,국수,생일,과자,현영,일과,아이들,상점,훈이,신발,준비,라면,큰집,쇠고기,장사,비누,양발,시장,합승,모타,반찬,오뎅,돼지고기,마리,성질,손해,치약
36:병원	병원,혈압,치료,수술,병문안,결과,검사,의사,사진,의료,회복,택시,대구,진찰,들거름,집사람,구미,점심,주사,보건소,퇴원,질료,선생님,약방,말씀,부담,복용,경비,이변,검진,필요
37:과수전지	나무,전지,과수,일과,결혼,순복,작과나무,집사람,교회,무의미,하계,앞발,따뜻,친구,일부,작년,두엄,방목,인간,재미,휴식,가연,전쟁,여름,동생,동내,큰집,시계,부모,과수나무,작업
38:모내기	모내기,모판,봉답,모재리,이앙기,객토,집사람,수월,작업,일과,비누,마지기,상자,큰집,복판,준비,장만,사업,성질,이앙,기계,고통,시작,육체,육모,내년,농촌,휴식,고비,욕심,여유
39:형제	순달,동생,형님,구미,비누,아파트,봉지,형수,순서,삼촌,씨우는,이사,순복,형제,노임,큰집,지수,대구,현장,용학,부부,사촌,준비,회사,고기,전화,멸칭,장사,작업,포도,숙모
40:지출항목2	현주,과자,금액,월말,수입,일과,술값,합승,고물상,재방,합승비,분개,감기약,친구,담배,무의미,허가,총지출,총지출금액,총수,예비군,처제,휴식,빵값,열차,수입금액,국수,순달,구담,복구,입금액

〈부표 2〉 자녀문단 토픽별 핵심어휘

1:오락실/아이/꾸중	오락실,성질,아이,약속,오락,꾸중,아이들,소리,잘못,저녁,시간,아들,거짓말,어제,친구,인간,엄마,바람,작정,모습,김천,텔래비전,아침,돼지,미안,공부,머리,문제,그것,귀가,용서
2:생활/고충/군대	생활,하루,고충,시간,군대,인간,교육,날씨,극정,휴가,훈련,제대,고통,건강,날짜,성숙,주일,지루,고생,가슴,개월,바람,원망,정신,군복무,사실,더위,무리,입대,훈련소,운동
3:전화/면회/외박	전화,면회,시간,극정,일요일,외박,주일,부인,내일,식구,궁금,부대,저녁,대구,연락,배치,소리,통화,준비,부탁,군대,자제,하루,편지,목소리,상사,원망,전화벨,엄마,누나,답답
4:공부/걱정/학교	공부,극정,모습,너희들,학교,노력,아이,머리,시간,성적,취미,가슴,학원,욕심,결심,엄마,중학교,기분,작정,고등학교,열심,아이들,심정,시험지,문제,필요,방학,의욕,시험,안전,모양
7:엄마/선물/생일/졸업	엄마,선물,생일,졸업,편지,참석,아들,금일,서운,부인,준비,침대,추석,쇠고기,바다,조금,마음속,반지,행동,물건,누나,생신,너희들,혼수품,가락지,김천,미안,고기,사랑,고등학교,중학교
8:회사/봉급/적금/취직	회사,봉급,컴퓨터,극정,적금,취직,근무,면접,방위,화투,출근,자격증,취업,산업체,문제,공장,직장,학원,돼지,사회,입사,만원,신경,부탁,실정,실습,업체,통장,낭비,구미,예금
10:학교/병원/상처	선생님,학교,병원,아이,집사람,극정,얼굴,상처,아침,가슴,금일,학년,시간,가정,수술,숙제,아이들,안전,치료,학생,사진,안경,신경,방학,퇴원,문제,유치원,소리,대신,방문,염려
13:시험/대학	시험,대학,전문,공부,합격,대학교,학교,불안,극정,등록금,전기,마음속,머리,내일,노력,점수,가슴,고등학교,사실,마지막,상처,장학생,기분,학생,소원,심정,아포,올해,고생,지원,기술
15:친구/교통	친구,시간,버스,여자,택시,자리,남자,김천,이용,구미,승용차,차표,도착,구입,약속,출발,기분,학생,대구,부부,마을,학교,모습,잘못,오전,눈물,사정,열차,한약,총각,오후

〈부표 3〉 교통이동토픽 핵심어휘 연결정도중앙성

훈 어휘	연결중앙성	훈 어휘	연결중앙성	현주 어휘	연결중앙성	현영 어휘	연결중앙성
친구	520	모습	83	친구	64	택시	195
시간	472	대전	81	시간	42	버스	167
여자	305	아들	79	김천	42	고사	128
김천	280	오전	79	남자	36	시간	96
눈물	269	열차	68	마을	27	교실	85
택시	241	퇴소식	68	이용	21	이용	82
도착	241	자가용	66	구미	19	교통	66
이용	236	잘못	62	총각	17	도착	60
접수	198	춘생	62	대구	16	출발	47
버스	197	부부	61	여자	14	자리	46
사정	145	학생	61	버스	13	기분	40
구미	141	차표	53	데이트	13	여자	39
승용차	140	금오산	49	택시	10	친구	17
구입	136	기분	46	자리	5	남자	12
골목	127	처음	46	구입	5		
약전	127	기대	45	차표	4		
출발	123	운영	28				
대구	121	남자	27				
약속	118	개찰	25				
자리	113	자랑	23				
마을	104	아파트	17				
논산	95	한약	12				
안전	86						

〈Key concepts〉: diary, digitized text, topic modeling, validity, prior domain expertise, bag of words, distant reading

An Illustrative Application of Topic Modeling Method to a Farmer's Diary

Nahm, Choon-Ho*

Rapid digitization of text documents, including personal diaries, raised a new puzzle: how can researchers analyze 'large quantities' of textual data efficiently and effectively? The author presents topic modeling as a promising solution to these challenges. The most distinctive feature of topic models is that they provide an automated procedure for coding the content of a corpus of texts into a set of substantively meaningful categories called 'topics'. The author discussed the theoretical presumptions of the topic modeling technique. The author illustrated the strength of topic modeling methods as a means of analyzing large text corpora by applying them to a farmer's diary (Appo diary). Topics extracted by topic modeling method are significant in terms of interpretability and external validity. Most of the results of topic modeling coincide with the results of traditional content analysis. In addition, topic modeling extracted a new topic, which the traditional content analysis had

* Professor, Chonbuk National University

overlooked. Based on this findings, the author discussed the demands and limitations of the methods focusing on three major characteristics of topic modeling methods: Bag of words assumption, no need of a priori coding list (prior domain expertise), and distant reading.