정책자료 2019-11

2019년 소셜 빅데이터 기반 보건복지 이슈 동향 분석

오미애 · 전종준 · 천미경

KIHASA
한국보건사회연구원
Korea Institute for Health and Social Affairs

【책임연구자】
오미애 한국보건사회연구원 연구위원

【주요 저서】
기계학습(Machine Learning)기반 사회보장 빅데이터 분석 및 예측모형 연구
한국보건사회연구원, 2017(공저)
기계학습(Machine Learning)기반 이상 탐지(Anomaly Detection) 기법 연구
-보건사회 분야를 중심으로
한국보건사회연구원, 2018(공저)

【공동연구진】
전종준 서울시립대학교 통계학과 교수
천미경 한국보건사회연구원 연구원

정책자료 2019-11

2019년 소셜 빅데이터 기반 보건복지 이슈 동향 분석

발 행 일 2019년 12월
저 자 오 미 애
발 행 인 조 흥 식
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1994년 7월 1일(제8-142호)
인 쇄 처 경성문화사

ISBN 978-89-6827-626-2 93510

발간사

보건복지 분야는 공급자 중심에서 수요자 중심의 맞춤형 서비스 체계로 변화하고 있다. 이러한 여건 변화를 빠르게 인지하고 보건복지 분야의 이슈를 파악하기 위해서는 소셜 빅데이터 활용이 중요하다. 빅데이터 분석 환경이 갖추어지고 기하급수적으로 늘어나는 비정형 빅데이터를 수집·분석할 수 있게 되면서 소셜 빅데이터 활용 수요는 증가하고 있다.

비정형 빅데이터는 모바일, 소셜네트워크서비스(SNS), 센서 등의 결합을 통한 새로운 형태의 데이터 생성으로 기존의 정형 데이터로는 파악하기 어려운 변화를 감지하고 정책 욕구를 즉시 확인하고 활용할 수 있는 근거를 제공한다. 비정형 빅데이터는 데이터에서 얼마나 많은 부가가치를 창출할 수 있는가의 관점에서 소셜 빅데이터 분석으로부터 새롭게 얻을 수 있는 지식 또는 부가가치의 양과 차이는 크지 않지만, '왜', '어떻게'에 대한 방향성을 제시해 줄 수 있다는 면에서 가치가 있다.

이 연구에서는 보건·복지 분야의 주요 이슈인 저출산 주제와 건강보험 주제가 어떻게 이슈화되고 있는지를 알아보기 위해 저출산, 건강보험 키워드 관련 문서를 수집하고 다양한 분석결과를 살펴보았다. 그리고 키워드 추출 방법론에 대해서도 자세히 설명하였다. 본 연구를 위해 조언을 해 주신 많은 전문가들과 원고 집필에 참여해 주신 전종준 교수님께 감사드린다.

끝으로 본 보고서에 수록된 모든 내용은 우리 연구원의 공식적인 견해가 아니며 연구에 참여한 연구진의 의견임을 밝힌다.

2019년 12월

한국보건사회연구원 원장

조 흥 식

목 차

그림 목차

Abstract

Social big data trend analysis based on health and welfare issues in 2019

Project Head: Oh, Miae

The health and social welfare sector is changing from a provider-oriented to a consumer-oriented customized service system. It is important to use social big data to readily recognize these changes and to identify issues in the health and welfare sector. The use of social big data is on increasing demand as big data analytics has improved rapidly and an ever-growing amount of unstructured big data is available for collection and analysis.

To analyze trends related to low fertility and health insurance, we searched through the word cloud to see which keywords were extracted most monthly, and network analysis was performed to analyze the relationship between keywords. Also, the characteristics of the subjects were examined using a topic model. We explain the method of extracting keyword based on type. Also, as the advanced technology of keyword extraction methodology, word feature extraction using structural coherence, feature extraction reflecting emotional classi-

Co-Researchers: Jeon, Jongjun · Cheon, Mikyung

fication information, and online learning method are also described. This can help in the study of advanced keyword analysis methods for refining articles in health and welfare fields.

Social big data analysis can serve as an important competitive edge in understanding the current situation of issues of national and social interest in the area of health and welfare policy.

요 약

1. 연구의 배경 및 목적

보건복지 분야는 공급자 중심에서 수요자 중심의 맞춤형 서비스 체계로 변화하고 있다. 이러한 여건의 변화를 빠르게 인지하고 보건복지 분야의 이슈를 파악하기 위해서는 소셜 빅데이터 활용이 중요하다. 빅데이터 분석 환경이 갖추어지고 기하급수적으로 늘어나는 비정형 빅데이터를 수집·분석할 수 있게 되면서 소셜 빅데이터 활용 수요는 증가하고 있다.

이 연구에서는 보건·복지 분야의 주요 이슈인 저출산 주제와 건강보험 주제가 어떻게 이슈화되고 있는지를 알아보기 위해 저출산, 건강보험 키워드 관련 문서를 수집하고 다양한 분석결과를 살펴보았다. 그리고 키워드 추출 방법론에 대해서도 자세히 설명하였다.

2. 주요 연구결과

2장에서는 저출산 관련, 3장에서는 건강보험 관련 동향분석을 위해 월별 어떤 키워드가 가장 많이 추출되었는지 워드클라우드를 통해 살펴보았고, 키워드간에 어떤 연관관계가 나타났는지 분석하기 위해 네트워크 분석을 하였다. 그리고 토픽 모형으로 주제들의 특성을 살펴보았다.

저출산, 건강보험 키워드로 수집된 문서들은 저출산과 건강보험과 관련된 직접적인 정책도 있었지만 일자리, 국민연금 키워드처럼 항상 이슈가 되는 키워드들이 포함된 문서들도 함께 수집되었다. 이는 저출산과 건강보험 이슈는 일자리와 국민연금 등 주요 이슈들과 함께 같이 언급되고

있다고 볼 수 있다.

4장에서는 기존에 개발된 주제추출 및 키워드 임베딩 방법에 기초한 키워드 추출 방법론을 살펴보았다. 우선 문서 내의 단어 특징 추출 및 주제 추출 방법론으로 네트워크 기반 단어분석, 단어 특징 추출 방법, 토픽 모형을 설명하였다. 그리고 키워드 추출에 사용되는 최신 방법론으로 네트워크 기반, 단어 특성 추출 기반, 토픽 모형 기반 키워드 추출 방법론을 살펴보았다. 키워드 추출방법론의 고도화 기술로 구조적 성김성을 이용한 단어 특징 추출, 감성 분류정보를 반영한 특징 추출, 온라인 학습 방법도 기술하였다. 이는 보건·복지 분야의 기사 주제 정제를 위한 키워드 분석방법 고도화 연구에 도움을 줄 수 있다.

3. 결론 및 시사점

소셜 빅데이터를 활용하여 살펴본 위 분석결과는 당연한 결과일 수 있다. 그럼에도 불구하고, 소셜 빅데이터 분석은 보건복지 정책 영역에서 국가적·사회적으로 관심이 있는 이슈에 대해 현 상황을 파악하는 데 중요한 경쟁력으로 작용할 수 있으며, 앞으로의 정책 관련 이슈를 도출하고 연구 전략을 세우는 데 근거자료로 활용될 수 있다. 다양한 소셜 빅데이터 분석 기술을 바탕으로 주요 보건복지 정책에 관한 사회적 관심도, 영향력 등을 분석하고 그 변화 과정을 살펴본다면 시의성 높은 보건복지 정책 연구의 기반을 마련할 수 있을 것이다.

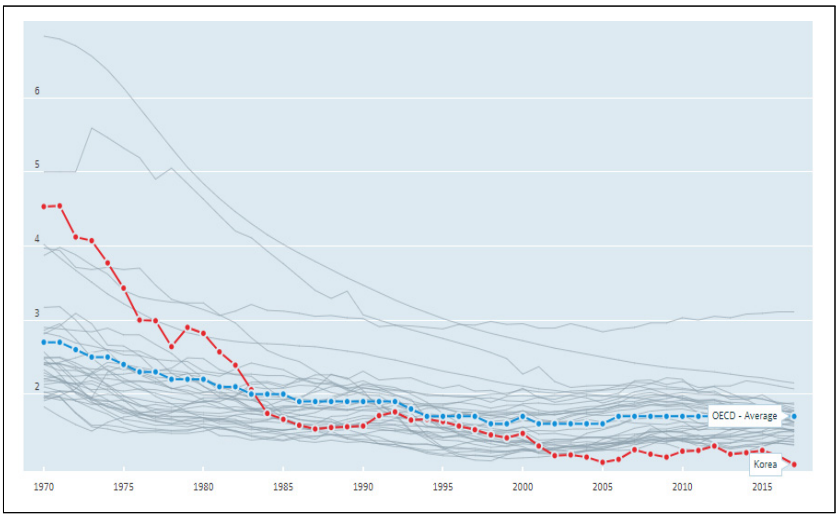
***주요 용어: 소셜 빅데이터, 저출산, 건강보험, 워드클라우드, 네트워크 분석, 토픽 모형**

제 1 장 서 론

1 서 론

우리나라는 1970년에는 합계출산율이 4.53명으로 OECD 평균 2.70명보다 1.83명이나 높은 국가에 포함되었으나 2001년 1.3명을 시작으로, 2017년에는 1.05명으로 계속해서 낮아져서 OECD 국가 중 합계출산율이 가장 낮은 유일한 초저출산국가(합계출산율 1.3명 이하)이다. 이러한 저출산문제를 해결하기 위해 2005년 저출산·고령사회기본법을 제정하였고, 법 제정 이후 대통령을 위원장으로 하는 「저출산·고령사회위원회」를 구성했으며, 현재 3차 저출산·고령사회기본계획(2016~2020년)을 수립하여 시행하였다. 또한 최근에는 저출산 고령사회위원회와 보건복지부에서는 삶의질 제고의 패러다임 전환을 구체화하였으며, 국정과제 및 정책목표와의 연관성을 고려하여 기존과제를 정비하고 범정부적으로 위원회에서 집중 추진할 핵심과제를 발굴하고 중장기 핵심과제 등 4차 계획(2021~2025년)을 위한 준비를 위해 제3차 기본계획을 수정하였다(저출산·고령사회위원회, 보건복지부(2019.2.) 제3차(2016~2020) 저출산·고령사회 기본계획(수정)). 이처럼 합계출산율이 가파르게 낮아졌고, 정부의 다양한 저출산 대책들이 쏟아져 나오고 있는 가운데, 저출산과 관련된 기사들에서 어떤 관련어들이 나오는지 살펴보는 것도 의미가 있을 것이다.

〔그림 1-1〕 OECD 국가의 합계출산율 변화(1970-2017)



자료: OECD Data(data.oecd.org/pop/fertility-rates.htm 2019.10.18. 인출)

건강보험은 대표적인 사회보장제도 중 하나로 지난 5월에 제1차 국민건강보험 종합계획(2019~2023)을 발표하였다. 이는 2017년에 발표한 건강보험 보장성 강화대책의 연장선으로 볼 수 있다. 종합계획은 현재 세대와 미래 세대 모두가 공평하게 건강보험의 혜택을 누릴 수 있도록 제도 안의 비효율적인 부분들을 점검하고 재정 관리를 효율적으로 하기 위한 구체적인 방안을 포함하고 있다[1]. 이 연구에서는 저출산과 관련된 기사들에 대한 분석과 함께 건강보험과 관련된 기사 분석 결과도 제시하고자 한다. 정형데이터로는 파악하기 어려운 단기간적인 이슈에 대한 동향을 파악하여 정책 욕구에 활용 할 수 있는 근거를 제공하고자 한다.

텍스트 분석의 이론적 측면을 생각해보면, 문서에 대한 전통적인 기술연구는 필요한 시점에 그것을 읽을 수 있게 저장하고, 표시, 출력하는 기

1) 노홍인. (2019). 전 국민 건강보험 30주년 성과와 과제 . 보건복지포럼, 통권 272호, 02-03. 재인용

록물 보관의 수단을 개발하는 방향으로 발전해왔다. 문서의 체계적 분류, 문서의 보관, 디지털 미디어를 통한 문서의 저장과 같은 작업들이 이에 해당한다. 이런 기술적 발전은 문서저장의 물리적 공간제약을 크게 줄여줌으로써, 문서의 보관 및 이용의 측면에서 효율성을 크게 높여주었다.

네트워크 기술의 발전은 문서의 생산 방식이 다양하게 만들었고, 그에 맞는 문서의 보관, 출력기술도 함께 발전시켰다. 특히 인터넷을 통한 문서 생산방식의 급격한 변화는 문서의 자동 분석이라는 새로운 기술적 수요를 낳게 되었다. 저장된 많은 텍스트 데이터들을 이용할 수 있게 되면서 주어진 텍스트를 표현하는 효율적인 수리적 방법이 개발을 가속화 시키고 그에 따라 '텍스트 분석'은 데이터 과학의 중요한 분야로 자리 잡게 되었다. 텍스트 분석은 자연어 처리, 텍스트 임베딩, 텍스트 주제추출, 텍스트 생성, 대화형 텍스트 처리 방법 등 수많은 관련 연구들이 문서 분석이라는 하나의 목적으로부터 파생되어 나왔다. 여기에서는 저출산과 건강보험 키워드 분석 외에 보건·복지 뉴스기사 주제 분석을 위해서 기존에 개발된 주제추출 및 키워드 임베딩 방법에 기초한 키워드 추출 방법론을 정리해보고 고도화 방안을 검토해보고자 한다.

본 보고서의 구성은 2장에서 저출산 키워드, 3장에서는 건강보험 키워드 분석 결과를 제시하고자 한다. 4장에서는 키워드 추출 방법론을 기술하고 5장은 결론으로 마무리하고자 한다.

제 2 장 저출산관련 소셜 빅데이터 분석

2 저출산관련 소셜 빅데이터 분석

제1절 데이터 수집 및 분석방법

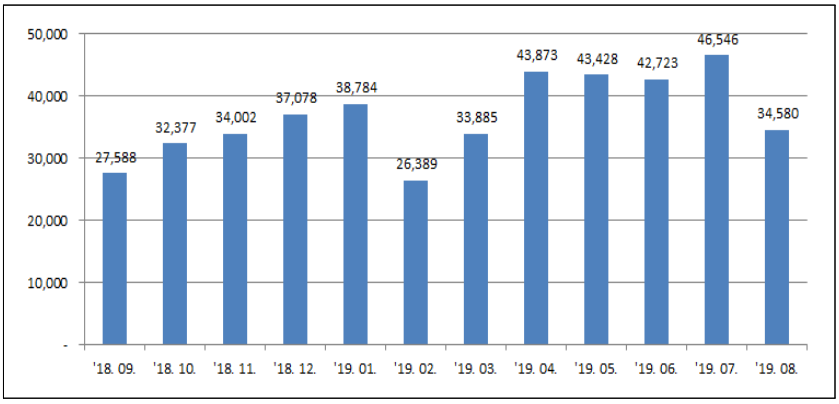
저출산에 대한 동향분석을 하기 위해 우선적으로 '저출산' 키워드 이외에 저출산과 관련있는 키워드 '아동수당, 베이비붐, 양육비용, 양육수당, 여성노동, 인구정책, 일가정양립, 임산부, 출산장려금, 출산율'의 키워드를 포함하였다.

먼저 소셜 빅데이터 수집을 위해서 비플라이소프트(Bflysoft)에서 운영하고 있는 '아이서퍼(eyesurfer)[2]' 플랫폼을 이용하여 2018년 9월 1일부터 2019년 8월 31일(약 1년)까지 관련 키워드가 포함된 온라인 매체(신문, 잡지 온라인 뉴스) 제목 리스트를 추출하였다. 추출된 리스트의 텍스트 데이터를 특수문자 제거, 단어 사전 생성 추가, 형태소 분석 처리, 불용어 제거 등의 데이터 전처리 과정을 거쳐 최종적으로 분석 가능한 정형데이터로 변환하여 최종 수집 되었다. 수집된 총 441,253건의 텍스트 문서를 본 연구의 분석에 포함시켰다.

2) 아이서퍼는 국내 최초 지능형 신문스크랩편집기를 이용하여 원하는 정보를 신문, 잡지, 온라인뉴스(약 1,024개 매체)에서 자동 수집하여 스크랩&모니터링 업무를 편리하게 해주는 서비스 플랫폼임(Bflysoft홈페이지. bflysoft.com; eyesurfer홈페이지. eyesurfer.com)

〔그림 2-1〕 월별 수집된 '저출산' 텍스트 문서

(단위: 건)



이 장에서 활용한 방법은 워드클라우드, 네트워크 분석, 토픽 모형이다[3].

워드클라우드(word cloud)는 대표적인 텍스트 자료 시각화 방법으로, 문서에서 사용된 단어의 빈도수를 분석하여 어떤 단어를 많이 사용했는지를 보여줄 수 있는 방법이다. 이 방법은 태그 클라우드(tag cloud)라고 부르기도 한다. 단어의 빈도수에 비례하여 단어의 글자 크기를 키워서 강조하는 방법이다.

네트워크 분석은 수학의 그래프 이론(graph theory)을 기반으로 한 분석 방법으로 사회학, 경제학, 경영학을 비롯한 사회과학 분야뿐만 아니라 생물학, 물리학, 의학 등의 자연과학 분야에 걸쳐 광범위하게 적용되고 있다. 네트워크는 노드(node)와 간선(edge)으로 구성되어 있다. 네트워크는 두 노드 간 관계의 강도를 고려하여 가중그래프로 표현할 수 있는데, 가중 그래프의 표현은 간선의 강도를 써 주는 방법이 있고, 선의 굵기

3) 분석 방법론에 대한 기술은 오미애·최현수·송태민·이상인·천미경.(2017). 2017년 소셜 빅데이터 기반 보건복지 이슈 동향 분석. 세종: 한국보건사회연구원. 보고서의 일부 내용을 요약·정리하였음.

혹은 색상으로 간선의 강도를 표현할 수도 있다.

토픽 모형(topic model)이란 비정형화된 방대한 문서 집단에서 관측할 수 없는 토픽을 문서 내 단어들의 분포를 통해 찾고자 하는 통계적 모형이다. 토픽 모형은 각 문서 내에 특정 토픽들이 존재하고 그 토픽에 따라 각 문서가 생성되었다고 가정하는 문서들에 대한 생성확률 모형(generative probabilistic model)으로 해석할 수 있다. 즉, 한 문서의 토픽 분포와 토픽별로 특정 단어를 생성할 확률을 모형화하여 그 문서가 만들어질 확률을 계산하는 것이다(Blei, Ng, & Jordan, 2003). 토픽 모형 분석은 방대한 문서 자료에서 맥락과 관련된 단서들을 이용하여 해석 가능성이 높은 주제들을 추출해주는 특성이 있다. 네트워크 분석과 토픽 모형의 방법론에 대한 설명은 4장에서 자세히 기술하였다.

저출산 관련 동향분석을 위해 월별 어떤 키워드가 가장 많이 추출되었는지 워드클라우드를 통해 살펴보았고, 키워드간에 어떤 연관관계가 나타났는지 분석하기 위해 네트워크분석을 하였다. 그리고 토픽 모형으로 주제들의 특성을 살펴보았다. 이러한 분석은 R version 3.6.1을 사용하였다.

제2절 분석 결과

1. 워드클라우드

2018년 9월의 저출산 관련 키워드는 아동수당의 빈도가 제일 높았으며, 일자리, 대통령, 임산부 등의 키워드가 주를 이루었다. 10월에는 임산부 키워드 빈도가 제일 높았고, 아동수당, 일자리, 행복, 인구정책 등의 키워드가 상위 빈도로 나타났다. 11월에는 예산안 키워드 빈도가 제일 높았고, 일자리, 아동수당, 대통령 등의 키워드, 12월에는 일자리 키워드 빈도가 가장 높게 나타났다.

2019년 1월의 저출산 관련 키워드는 2019년도와 마찬가지로, 아동수당, 일자리 키워드가 높은 빈도로 나타났고 대통령 신년사 및 지역별 저출산 정책 관련 사업들이 주요 키워드로 도출되었다. 2월에는 일자리, 임산부 키워드 빈도가 높게 나타났으며 인구정책, 맞춤형, 서비스, 지자체 관련 키워드가 상위 빈도로 도출되었다. 3월에는 경사노위, 특례시, 임산부, 일자리, 자립수당 등 저출산 이외의 정책과 관련된 이슈도 상위 키워드로 나타났다. 4월에는 일자리, 임산부 등의 키워드 외에 낙태죄 키워드가 이슈로 등장했다. 이는 2019년 4월 11일에 헌법재판소의 낙태죄 헌법불일치 결정이 주요 언론에 기사화되었기 때문이다. 5월은 임산부, 대통령, 인구정책 등의 키워드가 높은 빈도로 나타났고, 6월에는 지자체에서 저출산 관련 경진대회와 관련된 이슈도 언급되었다. 7월에 임산부, 고위험, 의료비 키워드가 상위 키워드로 나타났는데, 이는 2019년 7월 15일부터 고위험임산부 의료비 지원 대상 질환이 확대되는 것과 관련이 있다. 8월에는 아동수당 키워드가 압도적으로 높은 빈도로 나타났는데, 이는 2019년 9월부터 아동수당 지급 대상이 만 7세 미만 모든 아동으로 확대된 것과 관련이 있다.

[그림 2-2] 2018년 9월 '저출산' 워드클라우드

[그림 2-3] 2018년 10월 '저출산' 워드클라우드

[그림 2-4] 2018년 11월 '저출산' 워드클라우드

[그림 2-5] 2018년 12월 '저출산' 워드클라우드

[그림 2-6] 2019년 1월 '저출산' 워드클라우드

[그림 2-7] 2019년 2월 '저출산' 워드클라우드

[그림 2-8] 2019년 3월 '저출산' 워드클라우드

[그림 2-9] 2019년 4월 '저출산' 워드클라우드

〔그림 2-10〕 2019년 5월 '저출산' 워드클라우드

〔그림 2-11〕 2019년 6월 '저출산' 워드클라우드

〔그림 2-12〕 2019년 7월 '저출산' 워드클라우드

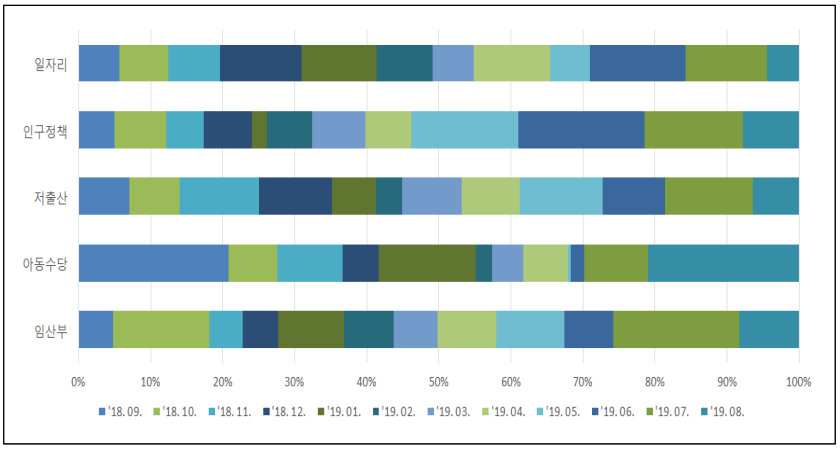
〔그림 2-13〕 2019년 8월 '저출산' 워드클라우드

월별로 가장 많은 빈도를 차지하는 상위 100개 단어의 워드클라우드를 살펴보면 가장 눈에 띄는 키워드는 '임산부, 아동수당, 저출산, 일자리, 인구정책'이다. 이 중 아동수당 키워드가 추출 빈도가 높은 월을 보면 다음과 같다.

앞서 언급한 것처럼, 2018년 3월 27일 공표된 「아동수당법」에 근거하여 소득재산 하위 90%인 가구에게 2018년 9월 21일 처음 아동수당 10만원씩 지급되었고, 이렇게 선별하여 지급하던 방식에서 2019년 1월 1일 보편적으로 만 6세 미만 모든 아동에게 아동수당을 지급하는 방식으로 변화되어 눈에 띄게 '아동수당' 키워드가 추출되었다. 이후 2019년 9월부터 지급 대상 아동의 연령기준이 만 7세까지로 확대되었기 때문에 이전에 중단되었던 아동수당 지급대상자들을 포함하여 각 지자체별로 아동수당이 확대되었다는 점을 홍보하기 위해 2019년 8월 다시 한 번 아동수당이 많이 언급된 것으로 나타났다.

〔그림 2-14〕 월별 '저출산' 주요 키워드의 추출 비중

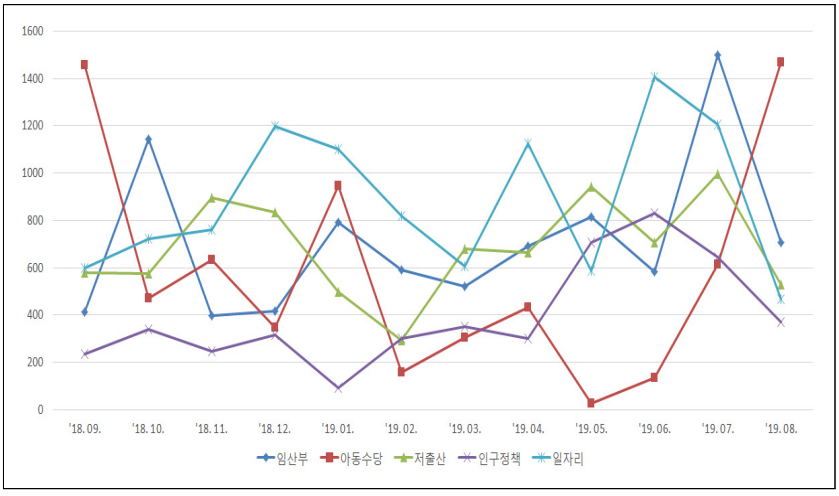
(단위: %)



월별 주요 키워드 추출 건수를 살펴보면 1년간 5가지 키워드 중 일자리 키워드가 10,598건으로 가장 많이 추출되었고, 임산부(8,569건), 저출산(8,194건), 아동수당(6,996건), 인구정책(4,742건)순으로 나타났다. 일자리 키워드는 '여성노동'과 관련된 기사에서 다수 추출되었으며, 저출산 현상의 대책으로 일과 가정의 양립정책이 있으며, 현 정부에서는 저출산 및 고령화와 더불어 일자리 정책도 사회구조적 문제 해결을 위해 노력을 하고 있다.

[그림 2-15] 월별 '저출산' 주요 키워드의 추출 건 수

(단위: 건)



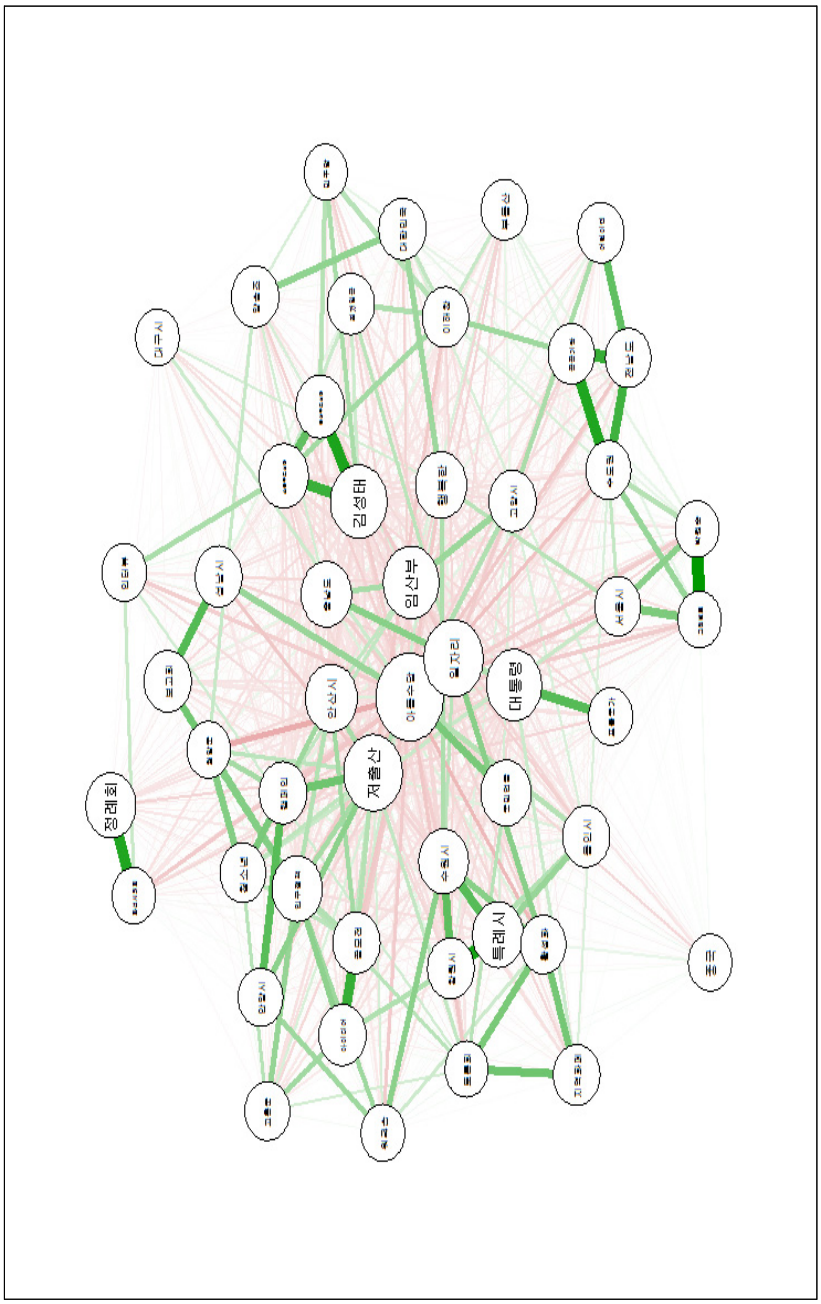
2. 네트워크분석

다음은 네트워크 분석 결과이다. 매 월별 출현 빈도가 높은 상위 50개 단어들 간의 네트워크 분석을 실시하였고, 각 월별 네트워크 그림을 제시하였다.

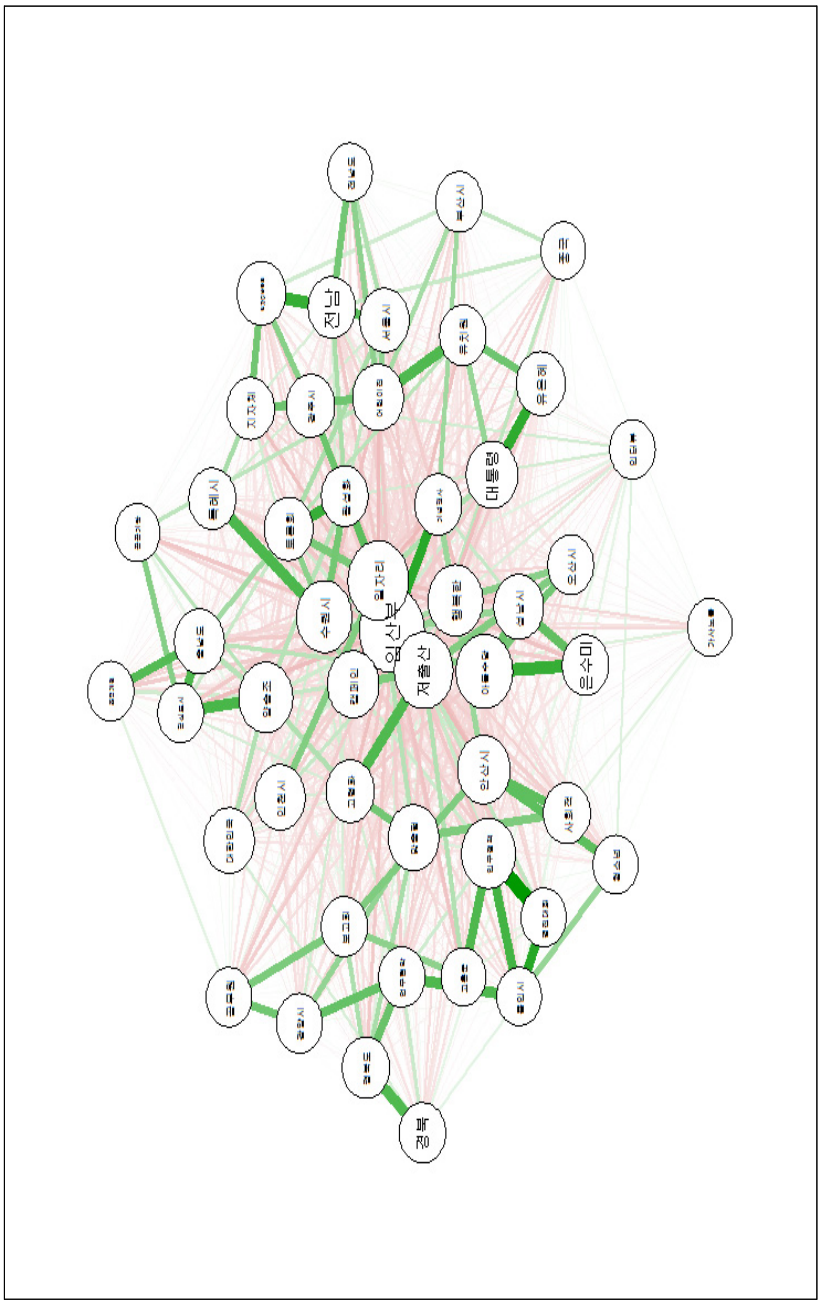
네트워크 상에 나타나는 주요키워드는 지자체에서 홍보하는 정책 및 홍보과 관련된 키워드 및 임산부 관련 키워드가 중심으로 나타났는데, 특징적인 2019년 1월과 2019년 8월의 네트워크 분석결과와 관련된 정책은 다음과 같다.

2019년 1월 보건복지부와 각 지자체에서 고위험 임산부들에게 건강한 출산을 위해 의료비 지원대상자를 확대하는 정책을 펼쳤으며, 정부는 난임부부의 난임 시술관련 건강보험 비급여와 본인부담금 등에 대해 지원을 확대하기로 하여 임산부, 고위험, 난임부부, 의료비 키워드들간의 연관성이 간선(edge)의 굵기로 굵게 나타났다.

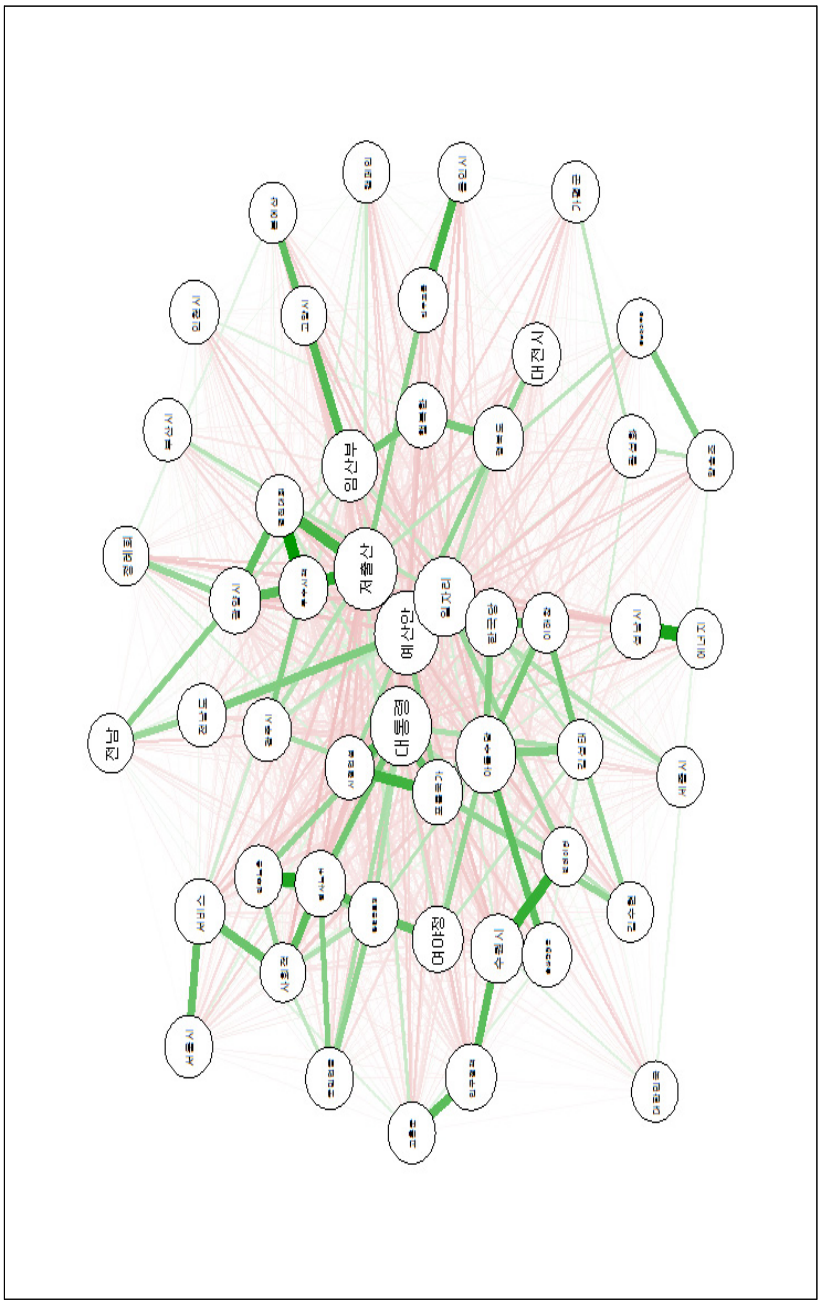
2019년 8월 각 지자체에서는 다양한 인구문제(결혼, 출산, 육아, 교육환경 개선, 일자리 창출 등)를 극복하기 위한 인구정책 아이디어 공모전을 개최하는 홍보가 증가하여 아이디어, 공모전, 인구정책 키워드들간의 연관성이 간선(edge)의 굵기로 굵게 나타났다.



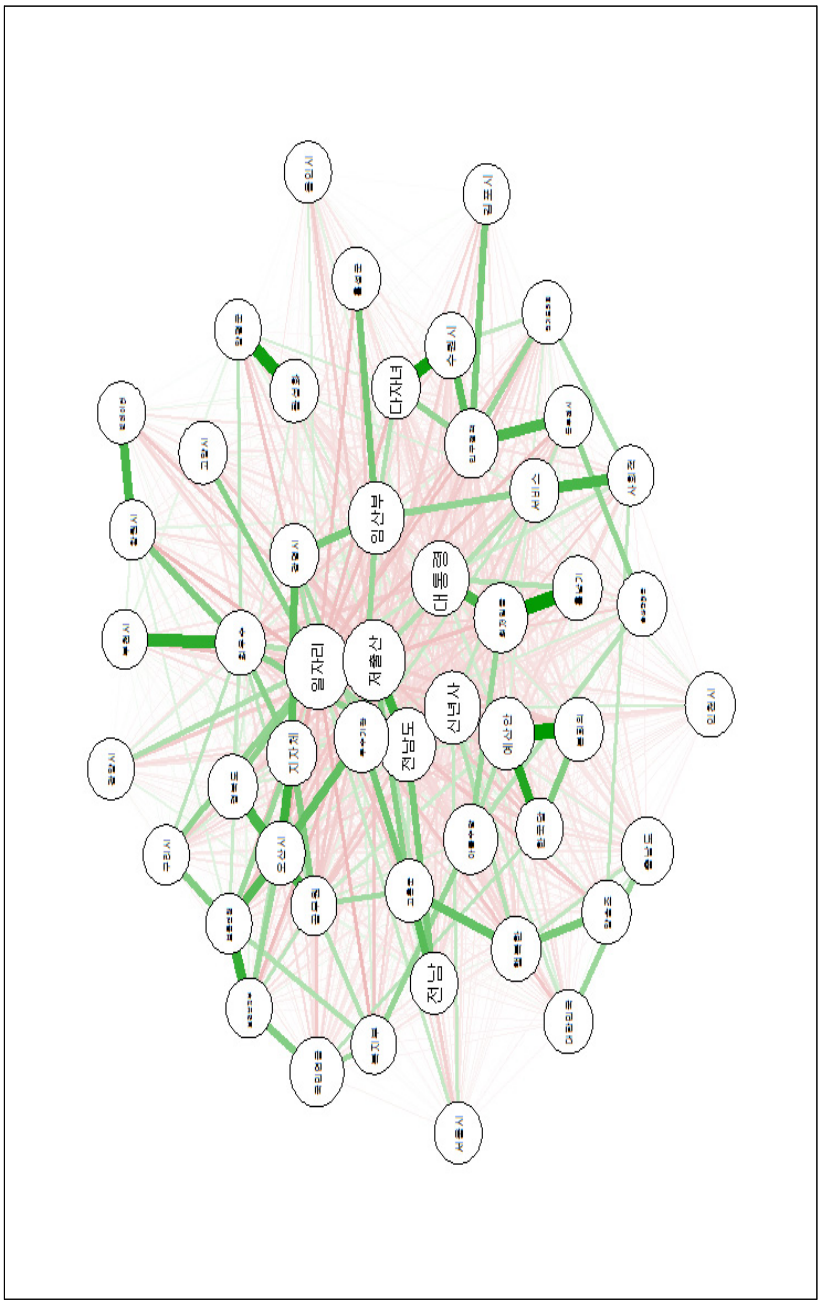
[그림 2-16] 2018년 9월 '저출산' 네트워크 분석



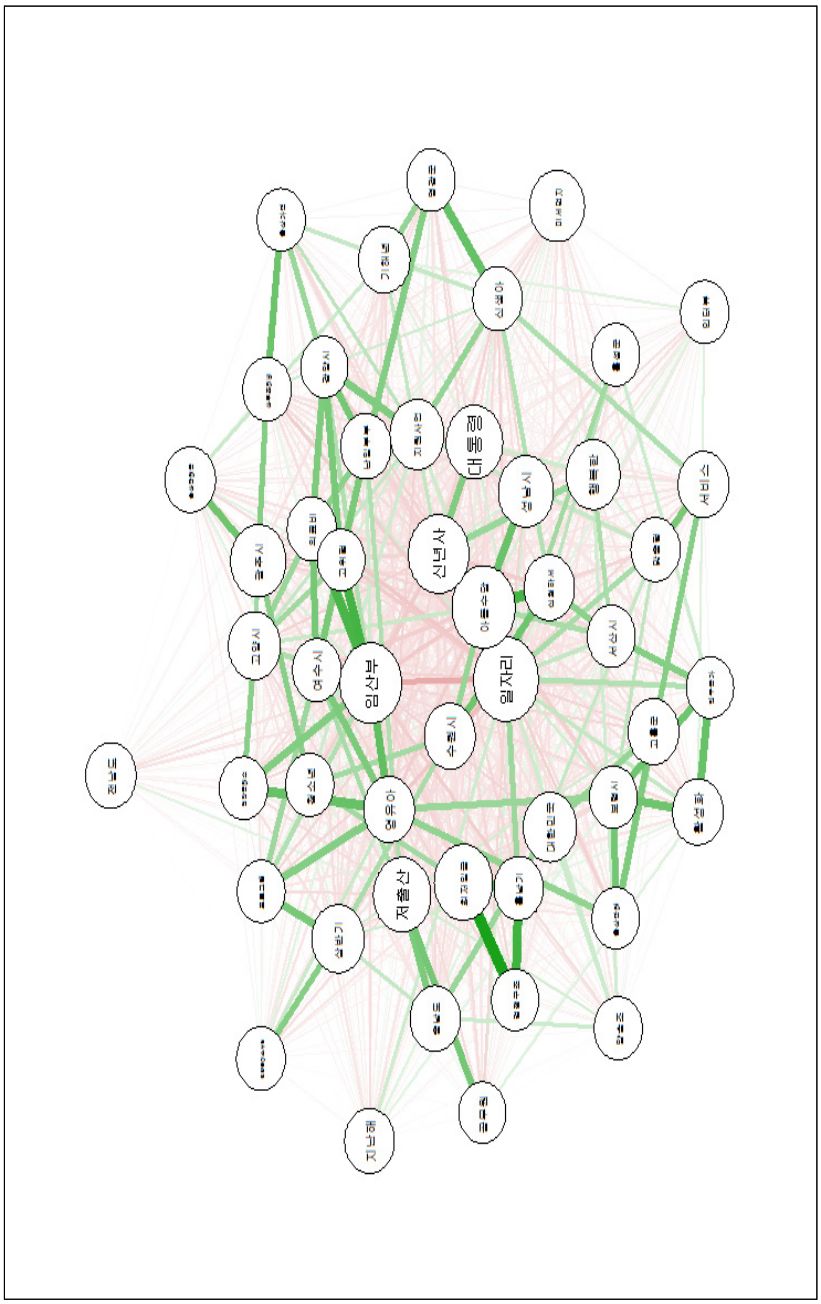
[그림 2-17] 2018년 10월 '저출산' 네트워크 분석



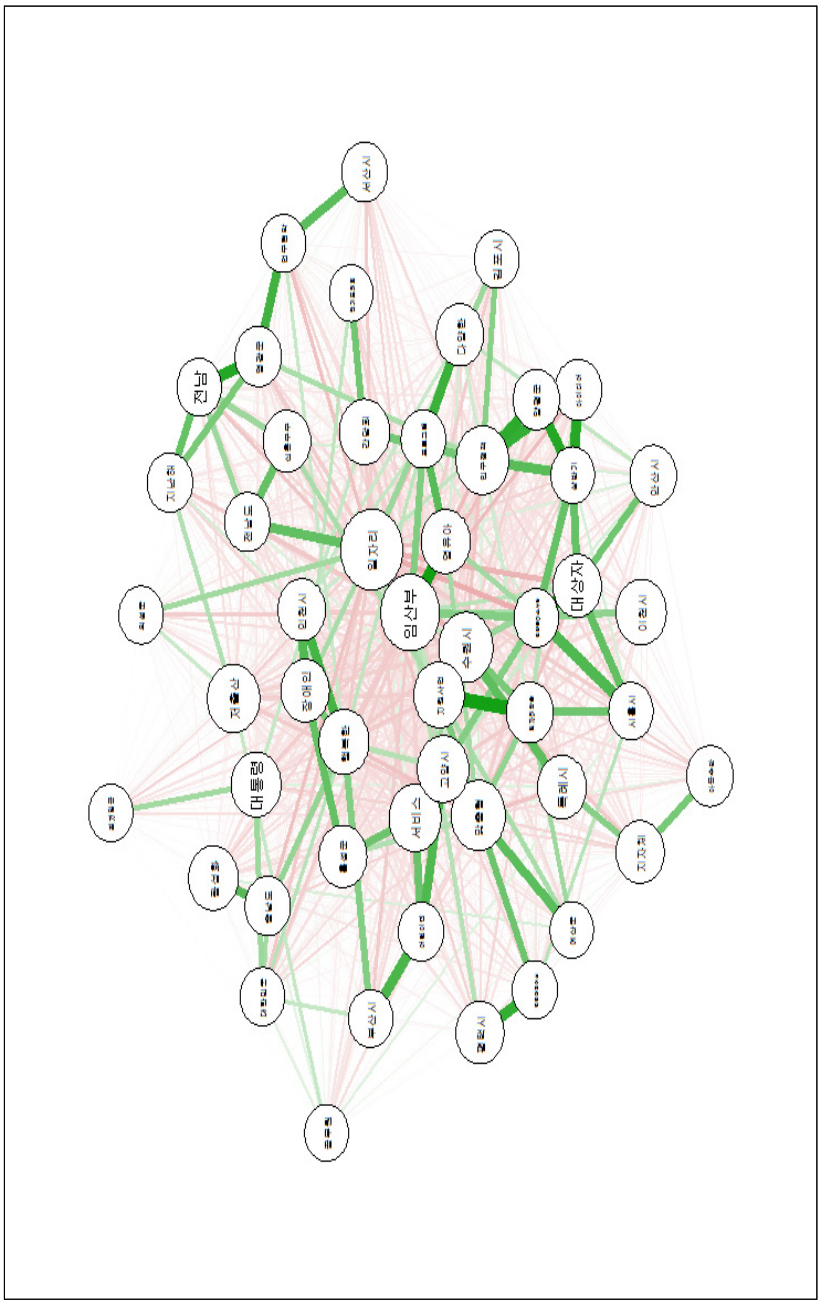
[그림 2-18] 2018년 11월 '저출산' 네트워크 분석



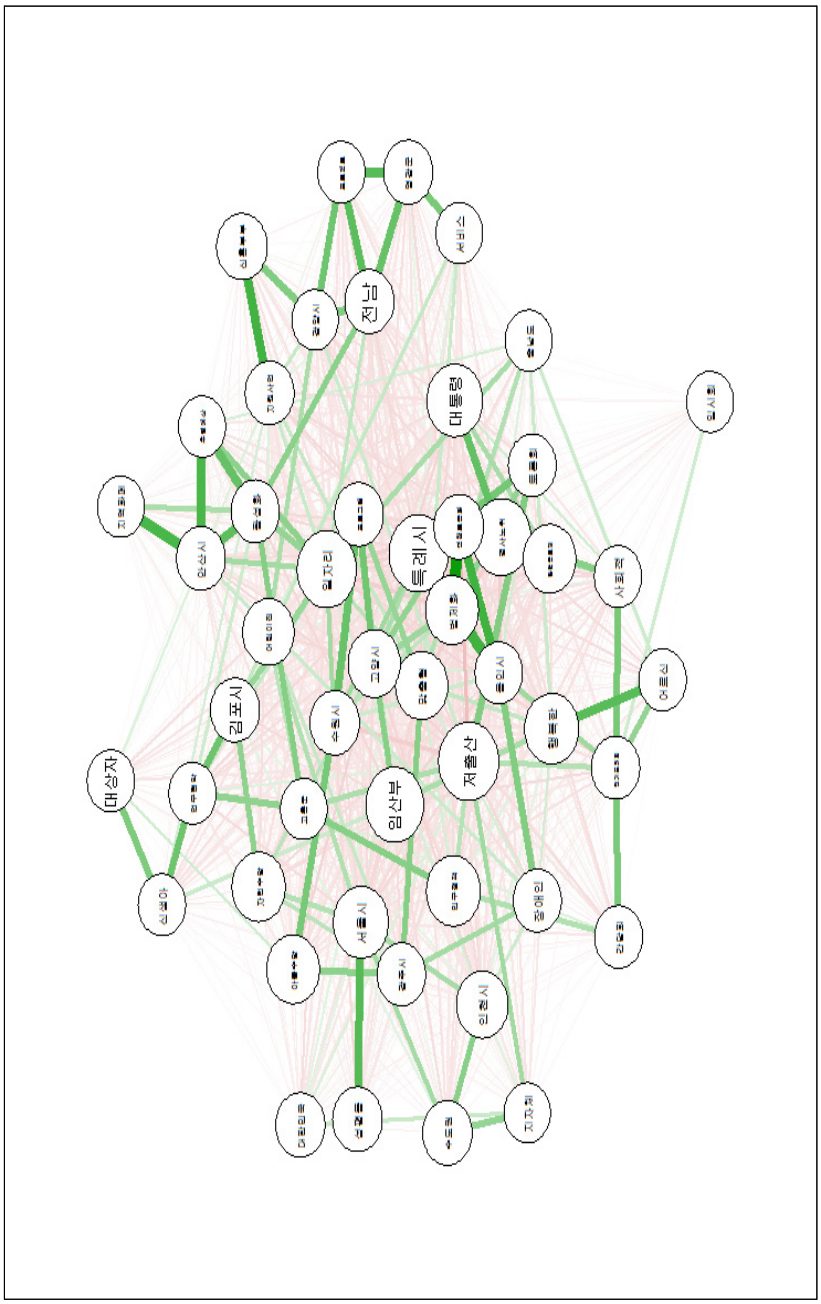
[그림 2-19] 2018년 12월 '저출산' 네트워크 분석



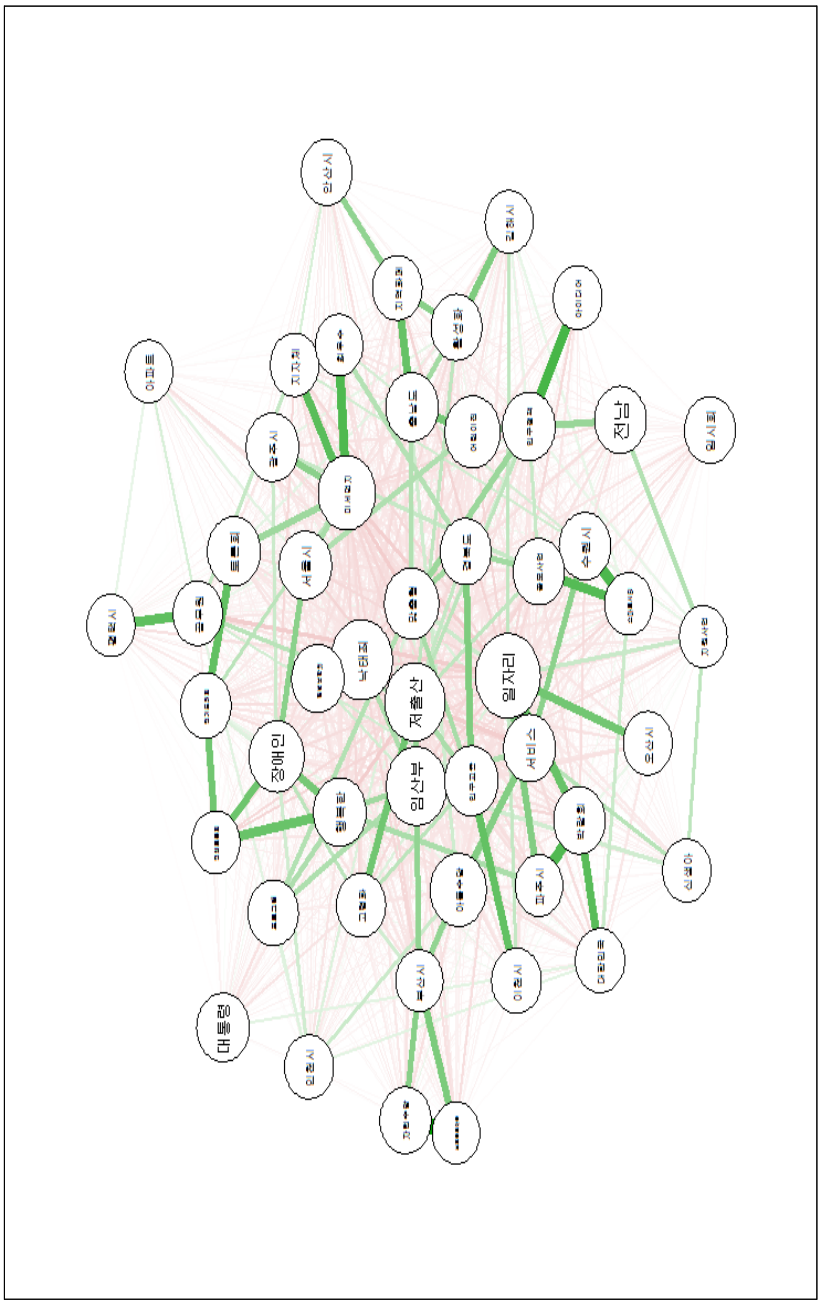
[그림 2-20] 2019년 1월 '저출산' 네트워크 분석



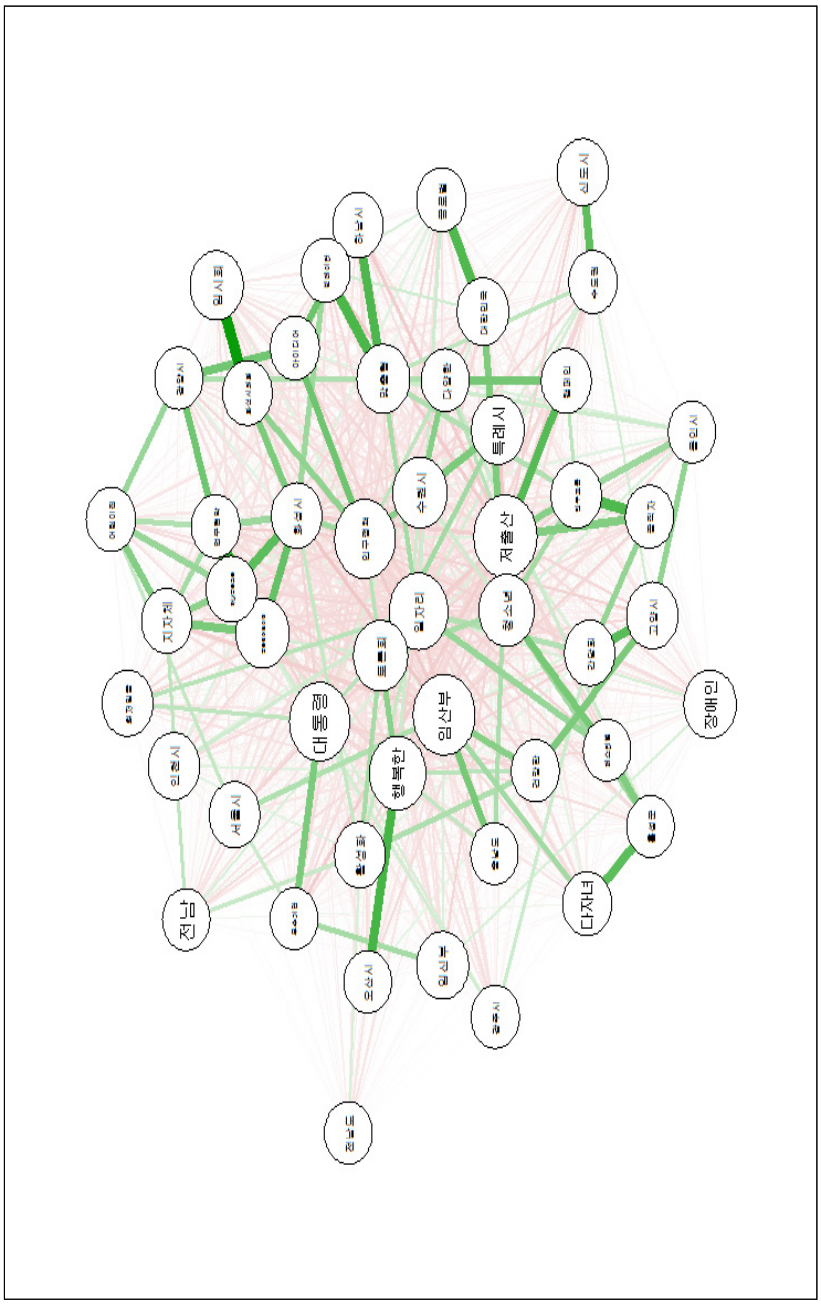
[그림 2-21] 2019년 2월 '저출산' 네트워크 분석



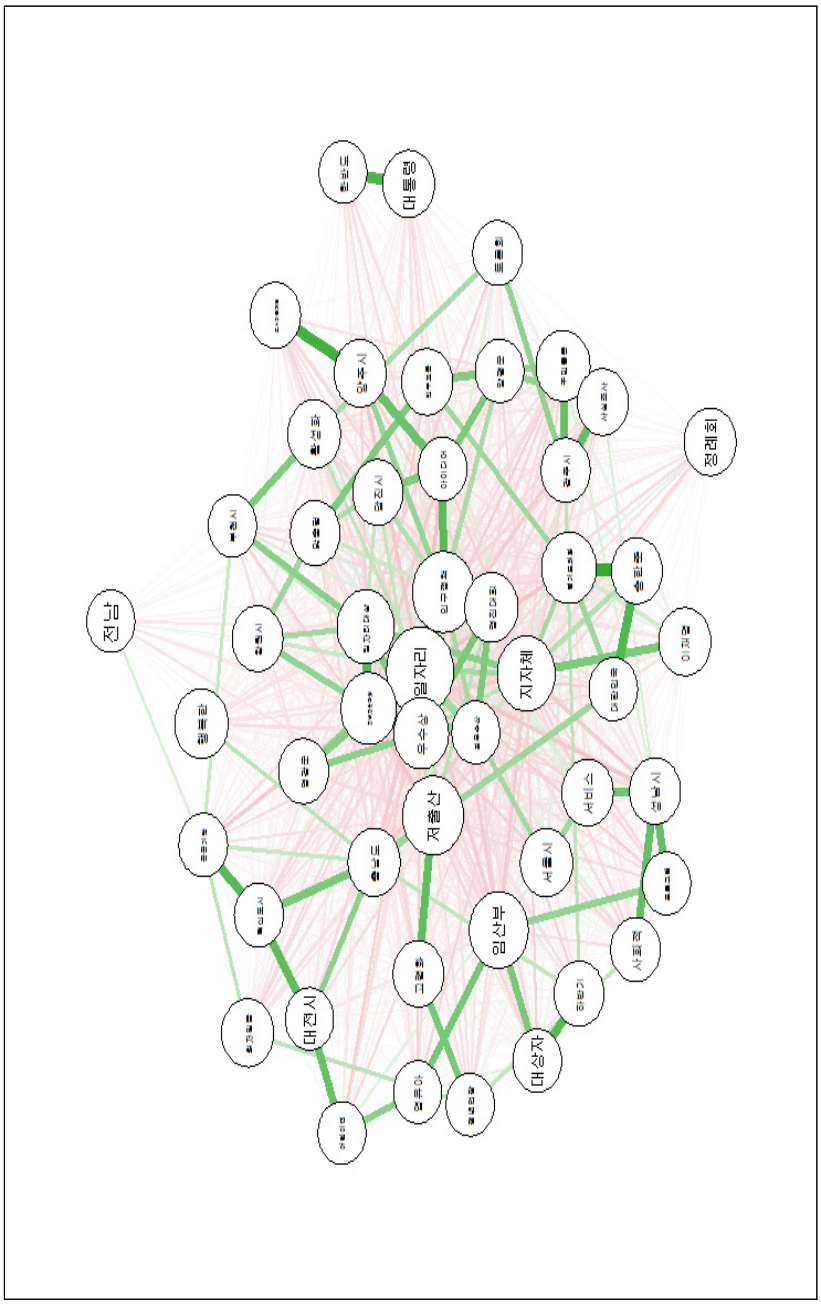
[그림 2-22] 2019년 3월 '저출산' 네트워크 분석



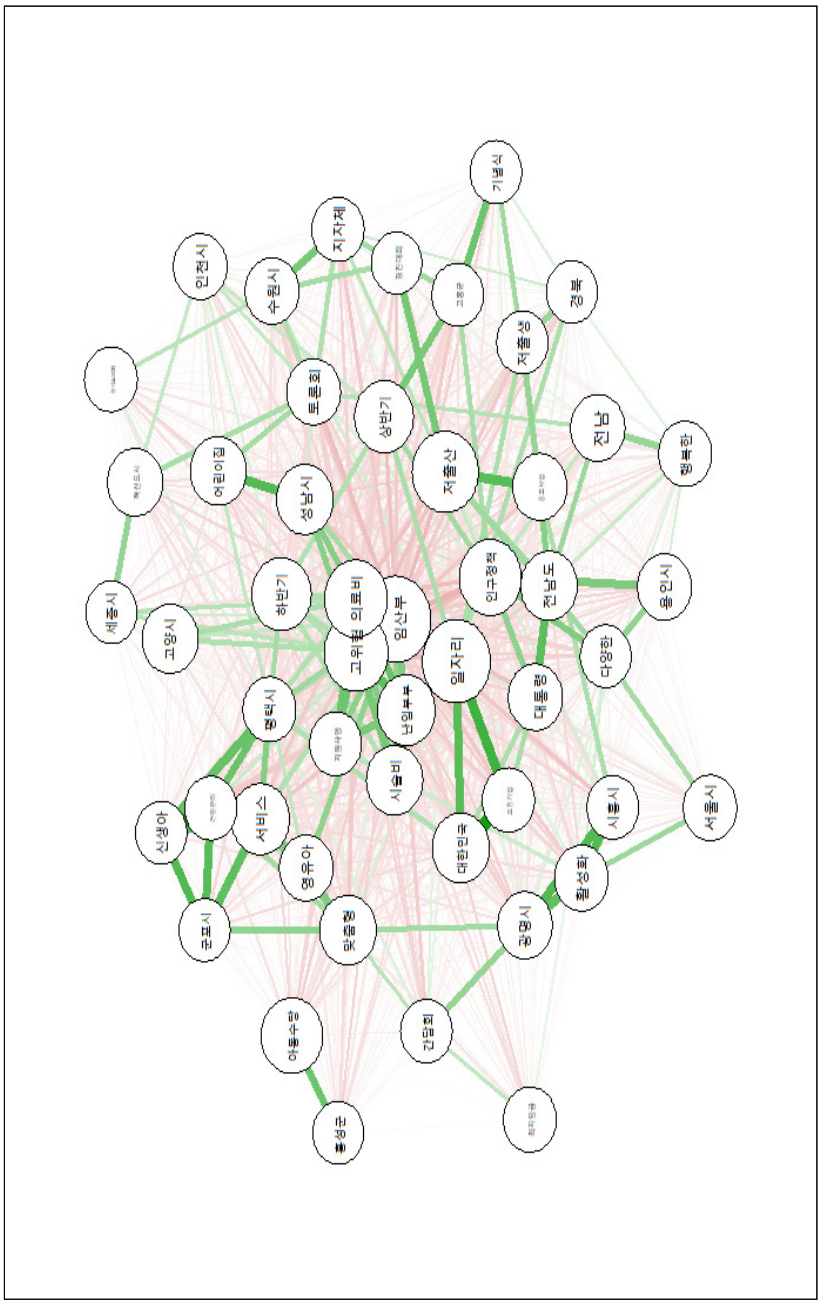
[그림 2-23] 2019년 4월 '저출산' 네트워크 분석



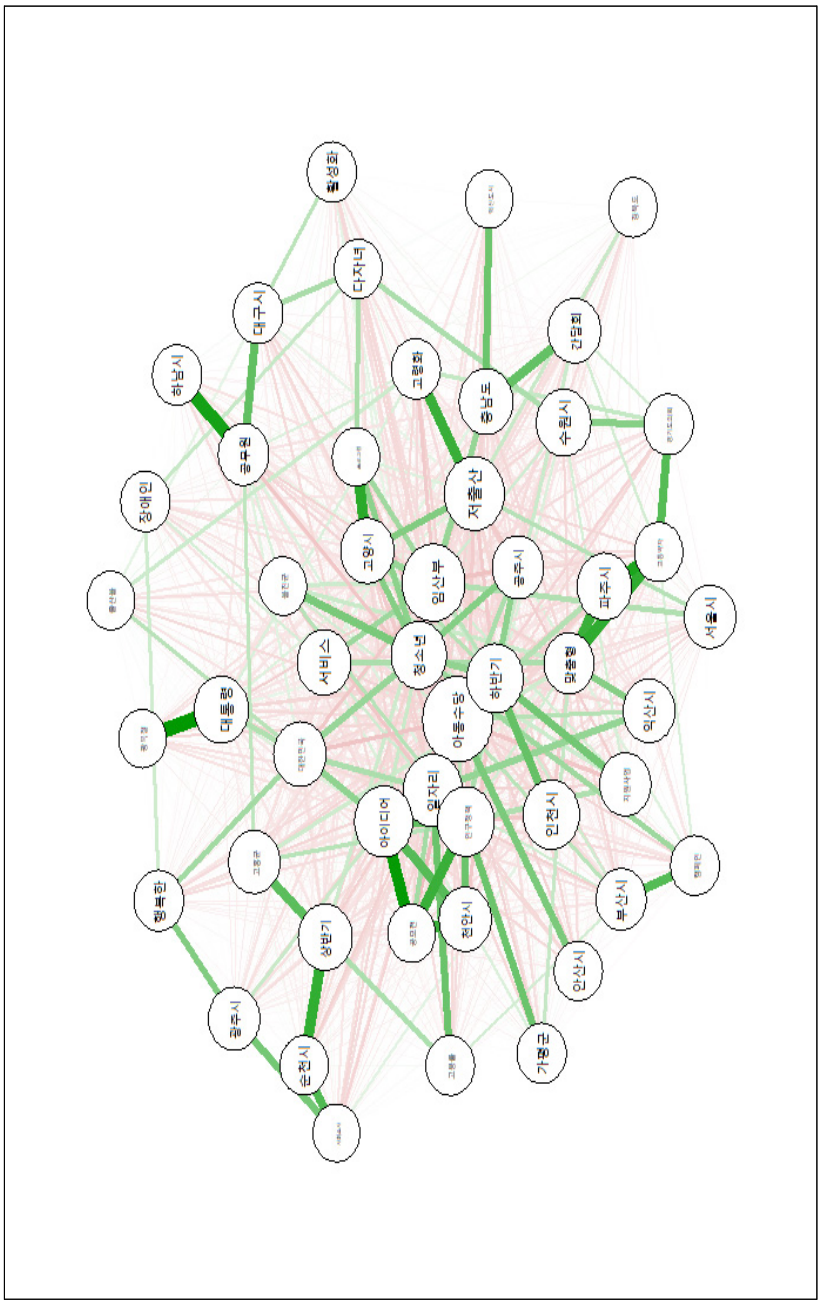
[그림 2-24] 2019년 5월 '저출산' 네트워크 분석



[그림 2-25] 2019년 6월 '저출산' 네트워크 분석



[그림 2-26] 2019년 7월 '저출산' 네트워크 분석

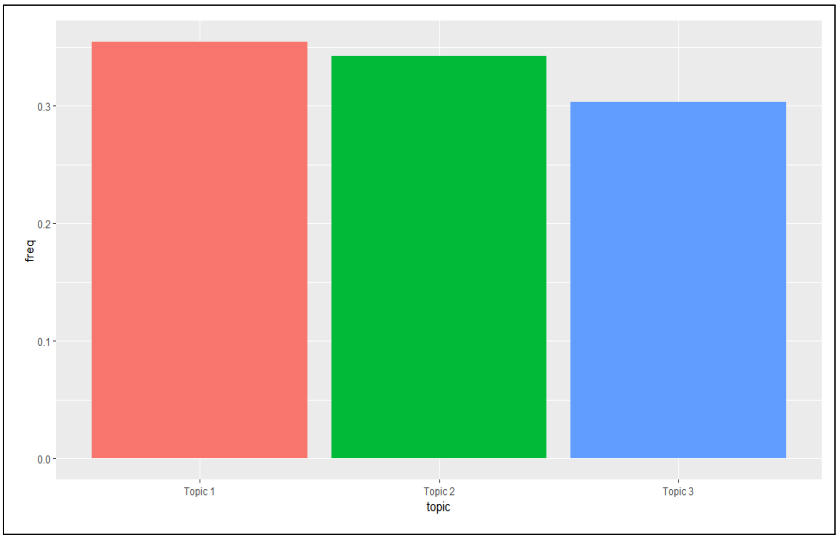


[그림 2-27] 2019년 8월 '저출산' 네트워크 분석

3. 토픽모형

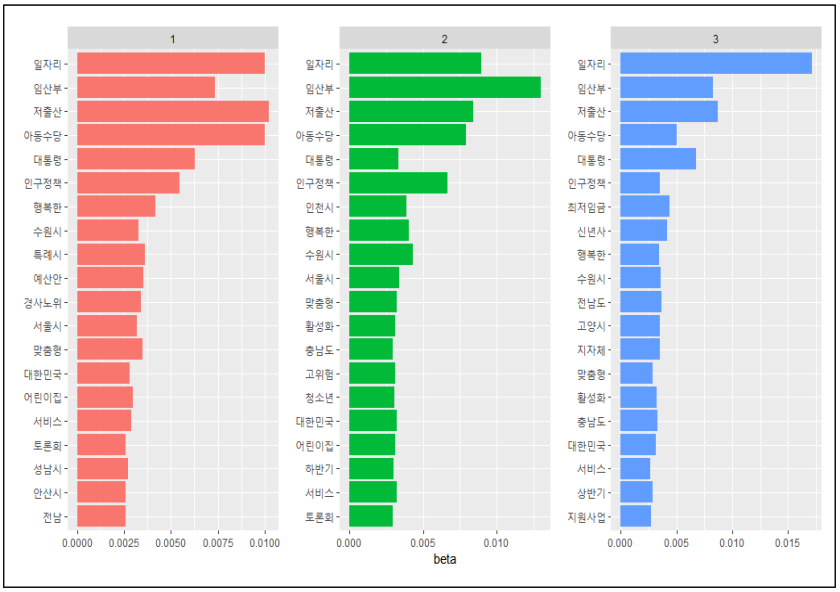
다음은 토픽모형에 대한 분석 결과이다. 토픽은 주제를 나타내며, 토픽 수는 연구자가 결정해야 한다. 토픽 수를 3개부터 6개까지 변경해가며 결과를 살펴보았고 최종적으로 토픽 수 3개로 결정하였다. 토픽별 문서 비중을 보면 3개의 토픽이 비슷한 수준으로 언론에 언급되었음을 알 수 있다.

〔그림 2-28〕 토픽별 '저출산' 문서 비율



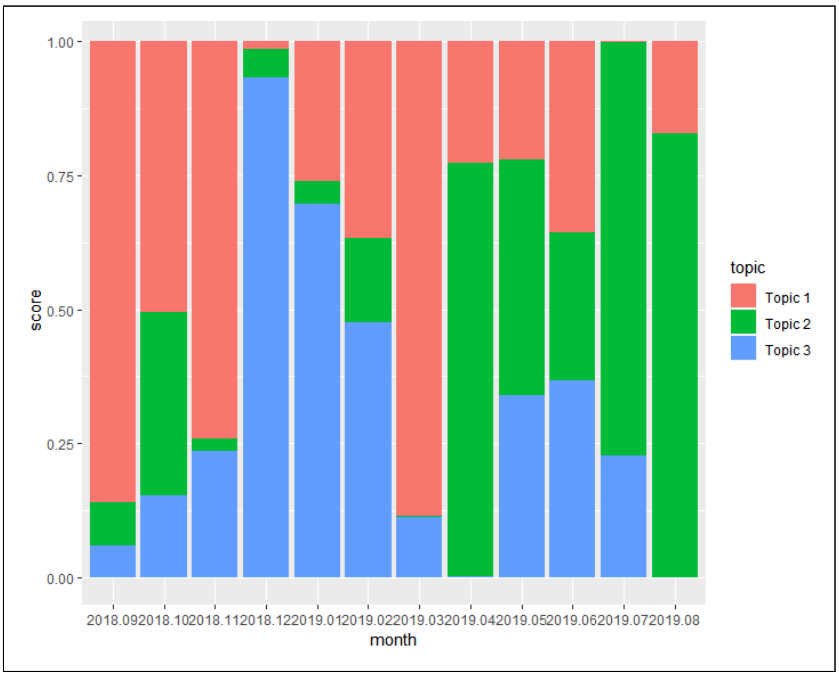
토픽 1은 저출산, 아동수당과 관련된 키워드가 상대적으로 많이 언급되었고, 토픽 2는 임산부, 맞춤형, 인구정책 관련 키워드가 상대적으로 많이 언급되었다. 토픽 3은 일자리, 대통령과 관련된 키워드가 상대적으로 많이 언급된 것으로 보아 신년에 발표한 정책과 관련이 있다.

[그림 2-29] 토픽별 '저출산' 상위 키워드



월별 토픽 변화를 문서비율로 나타내보면 토픽1의 경우 2019년 9월, 2019년 3월의 비중이 높아 워드클라우드의 결과와 일치한다. 토픽2는 2019년 4월, 2018년 7월, 2019년 8월에서 비중이 상대적으로 높았다. 토픽 3은 월별로 비교했을 때 2018년 12월에 가장 큰 비중을 차지하고 있다.

〔그림 2-30〕 월별 '저출산' 토픽변화(문서 비율)



제 3 장

건강보험관련 소셜 빅데이터 분석

3 건강보험관련 소셜 빅데이터 분석

제1절 데이터 수집 및 분석방법

건강보험에 대한 동향 분석을 하기 위해 우선적으로 '건강보험' 키워드 이외에 건강보험과 관련있는 키워드 '건강보장성, 건강보험재정, 급여, 노인장기요양보험, 미충족, 보험료, 비급여, 상대가치, 의료급여, 의료보장, 의료비용, 의료수가, 의료인력' 의 키워드를 포함하였다.

소셜 빅데이터 수집은 저출산 키워드와 마찬가지로 비플라이소프트(Bflysoft)에서 운영하고 있는 '아이서퍼(eyesurfer)[4]' 플랫폼을 이용하여 2018년 9월 1일부터 2019년 8월 31일(약 1년)까지 관련 키워드가 포함된 온라인 매체(신문, 잡지 온라인 뉴스) 제목 리스트를 추출하였다. 추출된 리스트의 텍스트 데이터를 특수문자 제거, 단어 사전 생성 추가, 형태소 분석 처리, 불용어 제거 등의 데이터 전처리 과정을 거쳐 최종적으로 분석 가능한 정형데이터로 변환하여 최종 수집 되었다. 수집된 총 732,177건의 텍스트 문서를 본 연구의 분석에 포함시켰다. 분석은 워드클라우드, 네트워크분석, 토픽모형으로 진행하였다.

4) 아이서퍼는 국내 최초 지능형 신문스크랩편집기를 이용하여 원하는 정보를 신문, 잡지, 온라인뉴스(약 1,024개 매체)에서 자동 수집하여 스크랩&모니터링 업무를 편리하게 해주는 서비스 플랫폼임(Bflysoft홈페이지. bflysoft.com; eyesurfer홈페이지. eyesurfer.com)

제2절 분석 결과

1. 워드클라우드

2018년 9월 워드클라우드 결과를 보면, 국민연금, 메르스, 건강보험, 일자리 등의 키워드 빈도가 높았는데, 메르스의 경우, 중앙방역대책본부의 메르스 대응 중간현황 발표와 관련이 있다. 2018년 10월, 11월에는 국민연금, 일자리, 건보공단, 등 보건복지정책과 관련된 키워드가 상위빈도로 도출되었는데, 트럼프 키워드가 상위에 도출된 이유는 미국 중간선거을 앞두고 실시된 여론조사에서 유권자들이 중요하다고 생각하는 이슈가 건강보험으로 나타났기 때문이다. 2018년 12월에는 국민연금, 일자리, 최저임금, 부양의무자, 기초생활보장 등의 보건복지정책 및 일자리 관련 키워드가 높은 빈도로 나타났는데, 이는 2019년 새해에 달라지는 정책에 대한 이슈와 관련이 있다.

2019년 1월에는 건강보험 키워드와 관련된 문서에서 상승세, 일자리 등의 키워드가 높은 빈도로 나타났는데 이는 국토교통부의 표준주택 가격공시 발표와 함께 건강보험료, 기초연금 및 기초생활보장 급여 등의 영향이 최소화되도록 중·저가는 점진적 개선한다는 기사와 관련이 있다. 2월에는 저소득층, 일자리, 의료비, 만성폐쇄성폐질환 키워드 빈도가 높았는데, 만성폐쇄성폐질환은 건강보험심사평가원에서 만성폐쇄성폐질환 4차 적정성 평가 결과 공개 기사와 관련이 있다. 3월에는 건강보험, 서울시, 저소득층, 보험료 등의 키워드 빈도가 높았다. 서울시 청년수당 신청 공고가 3월이었고 청년수당의 자격 소득요건이 건강보험료 월 부과액 기준이기 때문에 관련 이슈로 수집되었다. 4월에는 강원, 대통령, 산불피해, 특별재난지역 키워드가 상위 순위에 있었는데 이는 4월 4-5일 강원

도 동해안 지역에서 발생한 산불과 관련이 있다. 복지부는 이재민들의 생활안정 지원을 위해 건강보험료를 50% 범위 내에서 3개월분을 경감하고, 병원·약국 이용 시 본인부담금 면제·인하하며, 노인들의 틀니 재제작에도 건강보험을 적용할 계획이라고 밝혔다. 5월에는 일자리, 유방암, 아프리카돼지열병 키워드가 상위 키워드였는데, 유방암 키워드는 건강보험심사평가원의 유방암·위암치료 모두 잘하는 1등급 병원 발표와 관련이 있다. 6월에도 아프리카돼지열병 키워드가 높은 빈도로 나타났는데, 이는 아프리카돼지열병의 국내 유입을 막기 위해 정부가 검역에 노력하고 있다는 기사와 관련이 있고, 주요 이슈로 건강보험 키워드와 함께 수집된 것으로 보인다. 항생제 키워드는 건강보험심사평가원의 약제급여 적정성 평가와 관련이 있다. 2019년 7월에는 의료비, 고위험 임산부, 난임부부, 시술비 키워드가 상위 빈도로 노출되었는데 이는 건강보험 보장성 강화와 관련된 하반기 정책으로 난임시술비 지원을 최대 10회에서 17회까지 확대하고, 고위험임산부의 대상질환을 11종에서 19종으로 늘리는 등 진료비 지원을 확대하는 정책과 관련이 있다. 8월에는 복지사각지대, 지역사회, 관상동맥우회술 등의 키워드가 높은 빈도로 나타났는데, 이 달에 건강보험심사평가원은 허혈성심질환 환자에게 실시한 '관상동맥우회술 5차 적정성 평가' 결과를 공개하였고, 지자체는 '여름철 복지사각지대 위기가구 집중 홍보 및 발굴기간'을 운영하여, 단전, 단수, 건강보험 체납 등의 빅데이터로 위기가구를 발굴하여 지원한다는 기사와 관련이 있다.

〔그림 3-1〕 2018년 9월 '건강보험' 워드클라우드

〔그림 3-2〕 2018년 10월 '건강보험' 워드클라우드

〔그림 3-3〕 2018년 11월 '건강보험' 워드클라우드

〔그림 3-4〕 2018년 12월 '건강보험' 워드클라우드

〔그림 3-5〕 2019년 1월 '건강보험' 워드클라우드

〔그림 3-6〕 2019년 2월 '건강보험' 워드클라우드

〔그림 3-7〕 2019년 3월 '건강보험' 워드클라우드

〔그림 3-8〕 2019년 4월 '건강보험' 워드클라우드

〔그림 3-9〕 2019년 5월 '건강보험' 워드클라우드

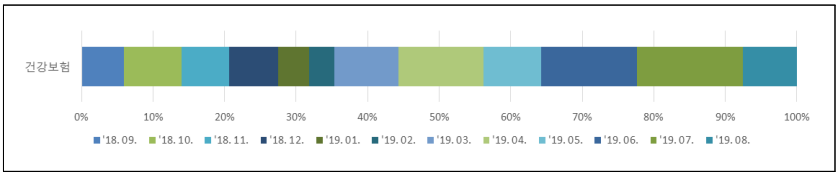
〔그림 3-10〕 2019년 6월 '건강보험' 워드클라우드

〔그림 3-11〕 2019년 7월 '건강보험' 워드클라우드

〔그림 3-12〕 2019년 8월 '건강보험' 워드클라우드

〔그림 3-13〕 월별 '건강보험' 키워드의 추출 비중

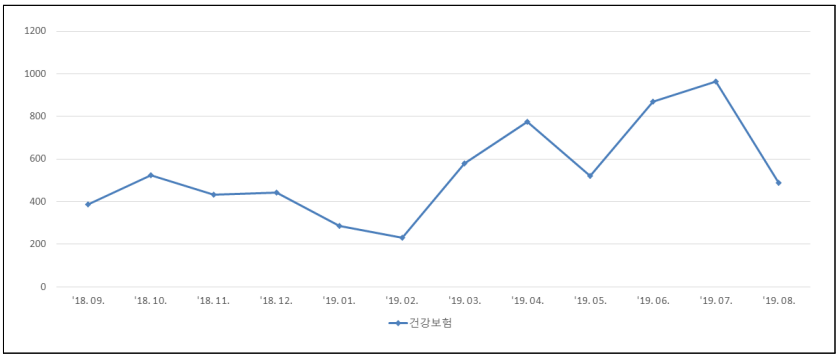
(단위: %)



건강보험 키워드 관련하여 상위 키워드는 일자리, 연금보험 등의 이슈도 많았기 때문에 월별 건강보험 키워드의 추출 비중을 살펴보았다. 2019년 7월, 6월, 4월, 2018년 10월 순으로 건강보험 관련 키워드가 많은 비중을 차지하였다. 월별 추출 건수로 보면 2019년 7월, 6월, 4월 순으로 추출 건수가 많았다.

〔그림 3-14〕 월별 '건강보험' 키워드의 추출 건 수

(단위: 건)



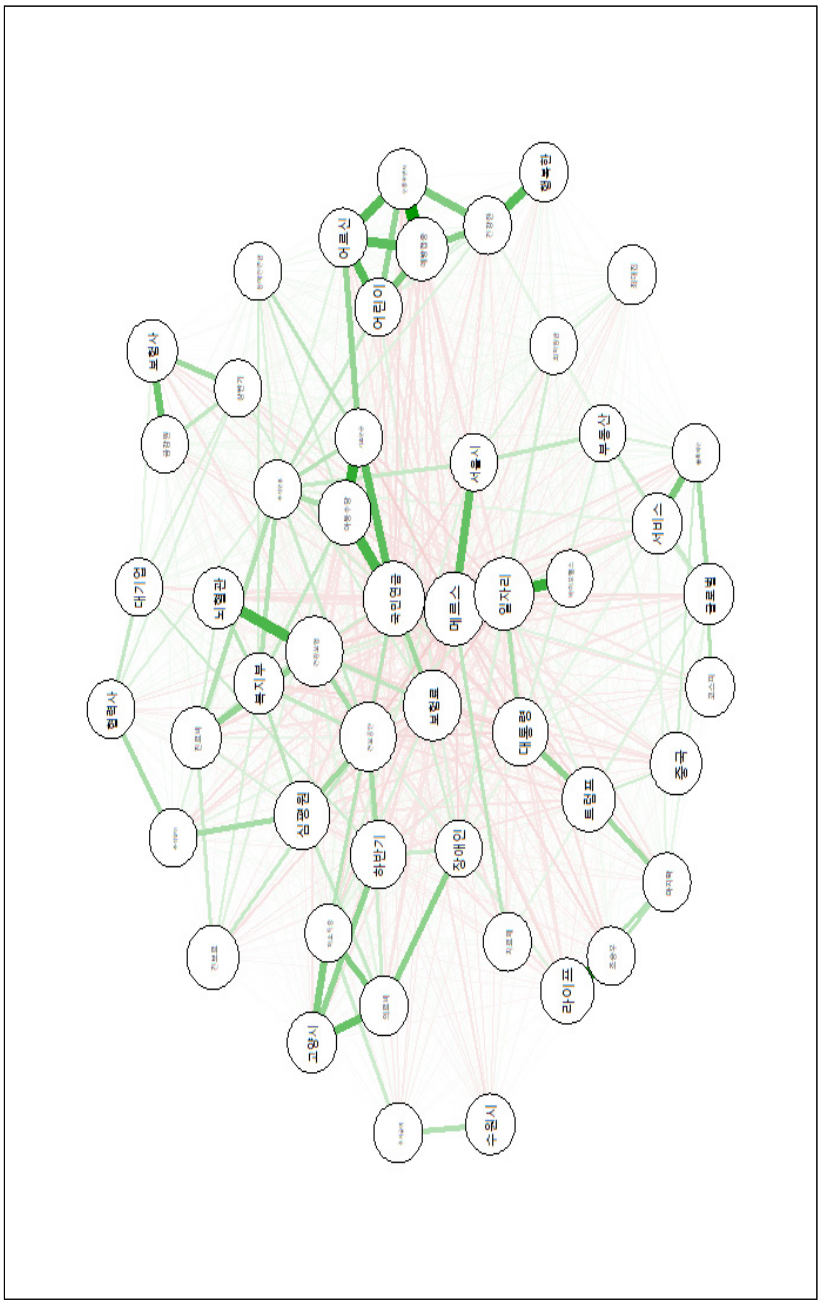
2. 네트워크분석

다음은 네트워크 분석 결과이다. 저출산 분석과 마찬가지로 매 월별 출현 빈도가 높은 상위 50개 단어들 간의 네트워크 분석을 실시하였고, 각 월별 네트워크 그림을 제시하였다.

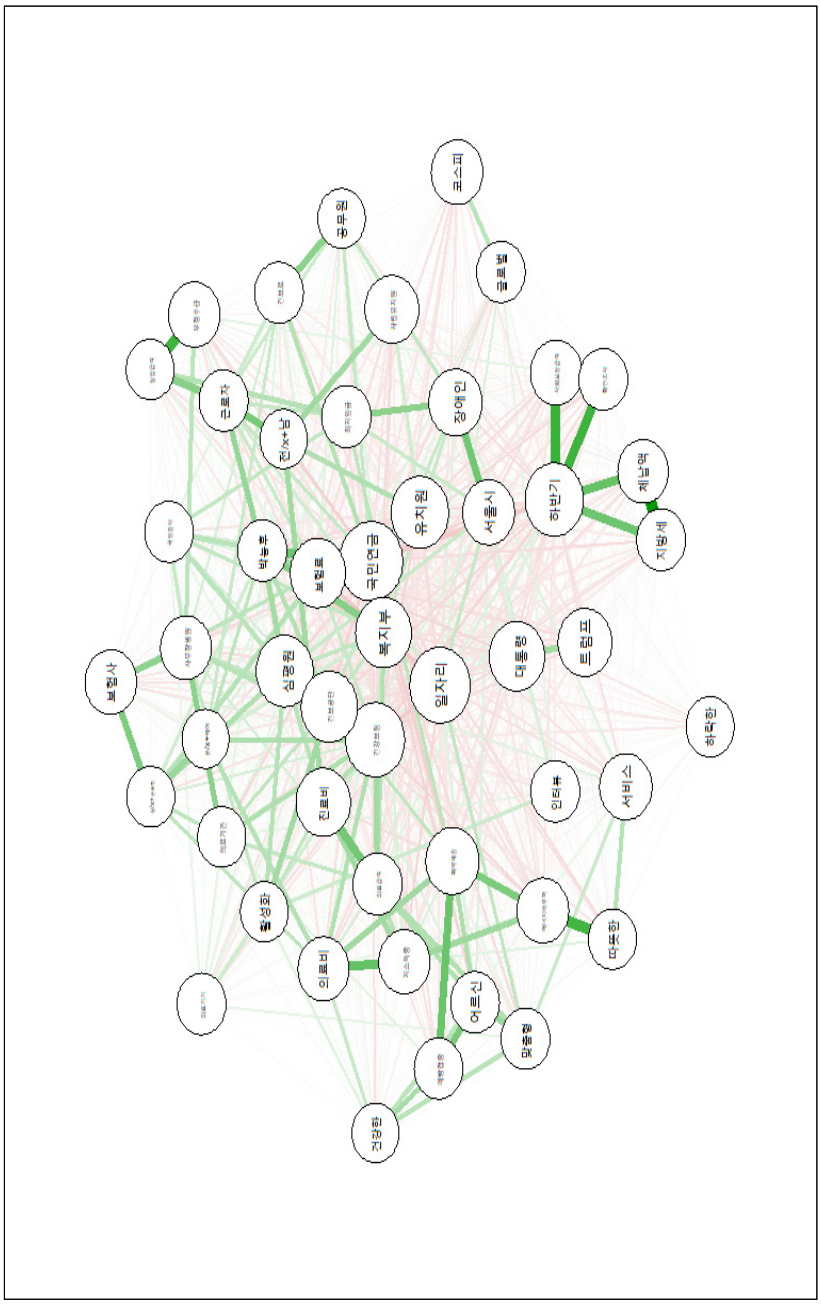
네트워크 상에 나타나는 주요키워드는 건강보험과 직접적으로 연관된 키워드 외에도 간접적으로 연관되어 있는 키워드(예를 들어 트럼프)도 나타났는데, 이는 건강보험만의 단독 기사뿐만 아니라, 주요 이슈와 함께 여러 정책이 발표된 기사 내용도 함께 수집되었기 때문이다.

네트워크 분석 결과 중에서 특징적인 2019년 7월 관련된 정책은 다음과 같다.

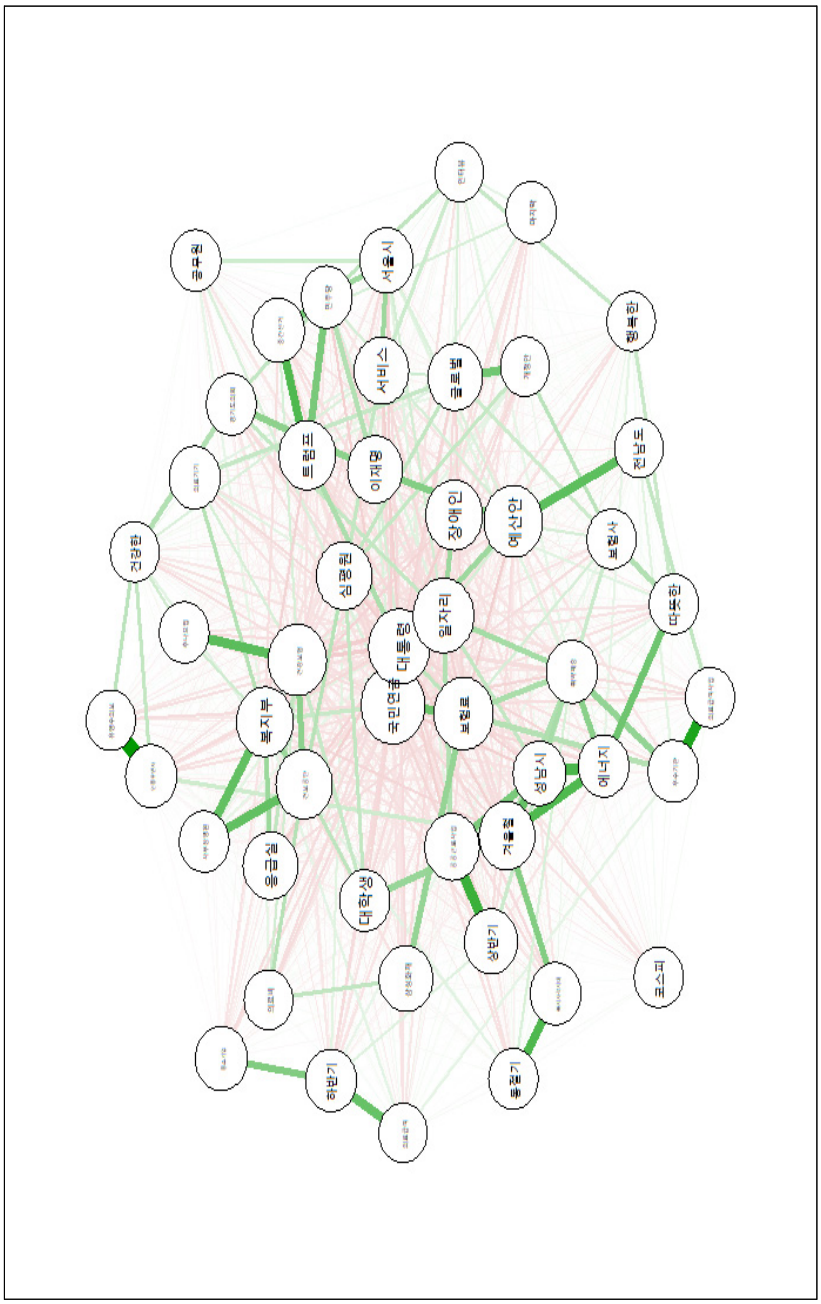
2019년 7월에 정부는 2019년 하반기 경제정책방향을 발표하면서 의료비 부담 경감과 건강관리서비스 활성화를 위한 건강보험 보장성 강화를 추진한다고 밝혔다. 워드클라우드 결과에서도 나타난 것처럼 난임시술비 지원 확대와 고위험 임산부 진료비 지원, 방문진료 시범사업, 중증 정신질환자 보호, 재활지원 강화 등의 방향을 제시하였다. 네트워크 상에서도 임산부, 하반기, 고위험, 난임부부, 시술비, 의료비들 간의 연관성이 높게 나타났다.



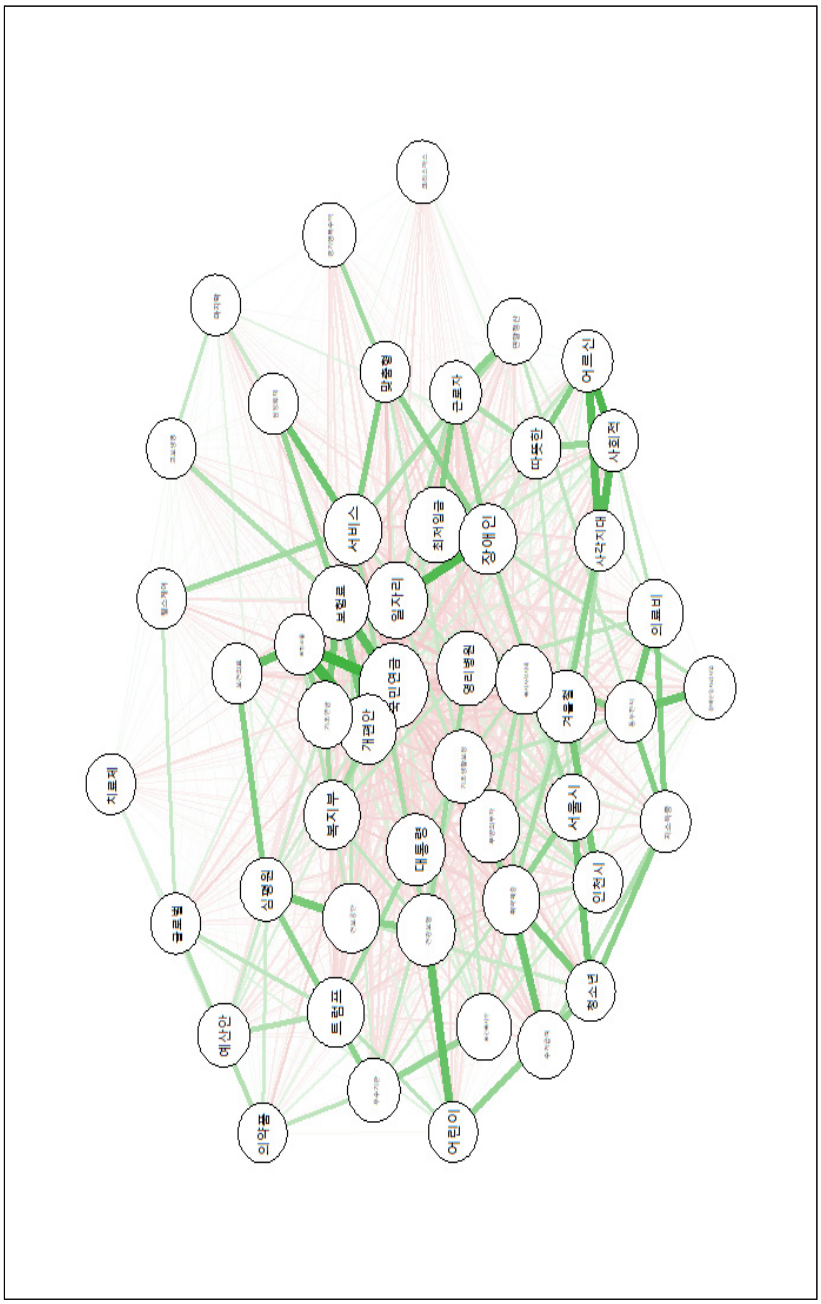
[그림 3-15] 2018년 9월 '건강보험' 네트워크 분석



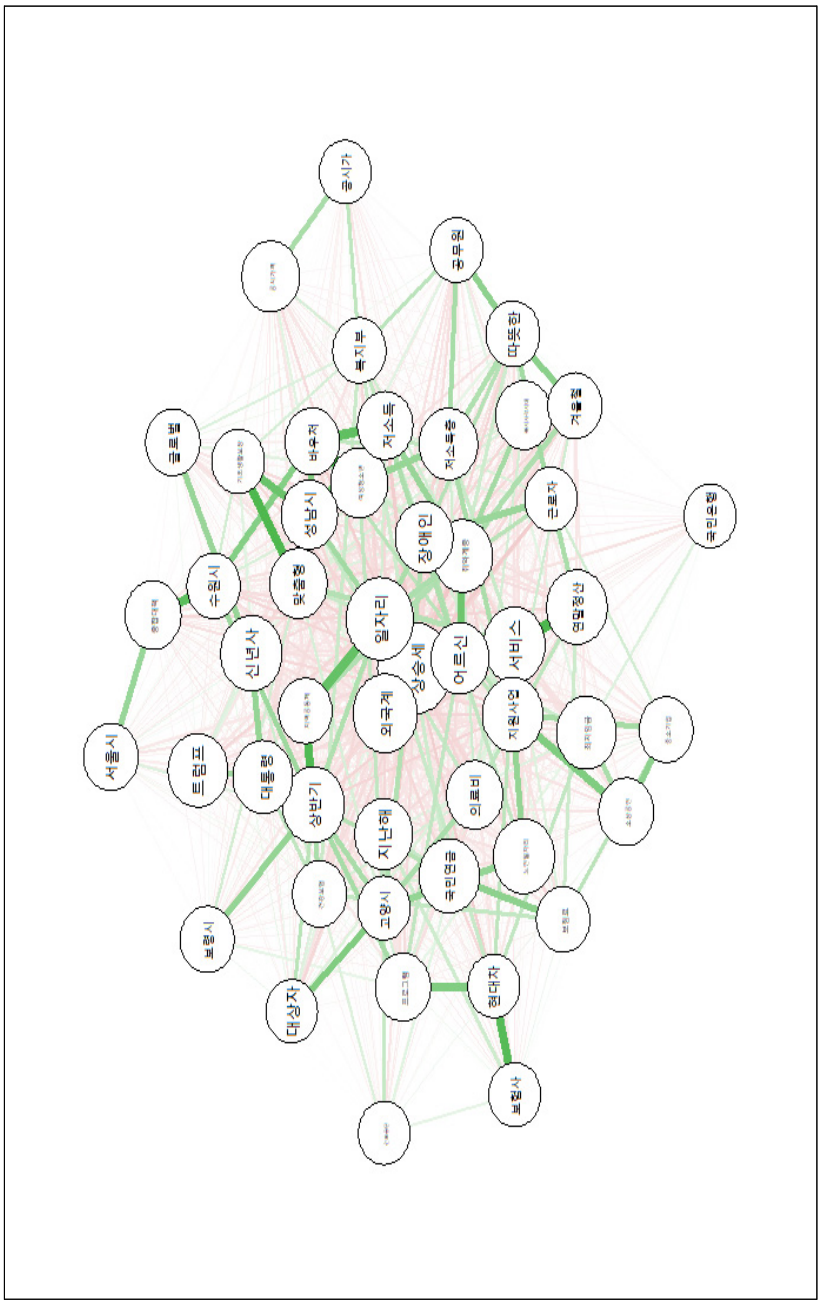
[그림 3-16] 2018년 10월 '건강보험' 네트워크 분석



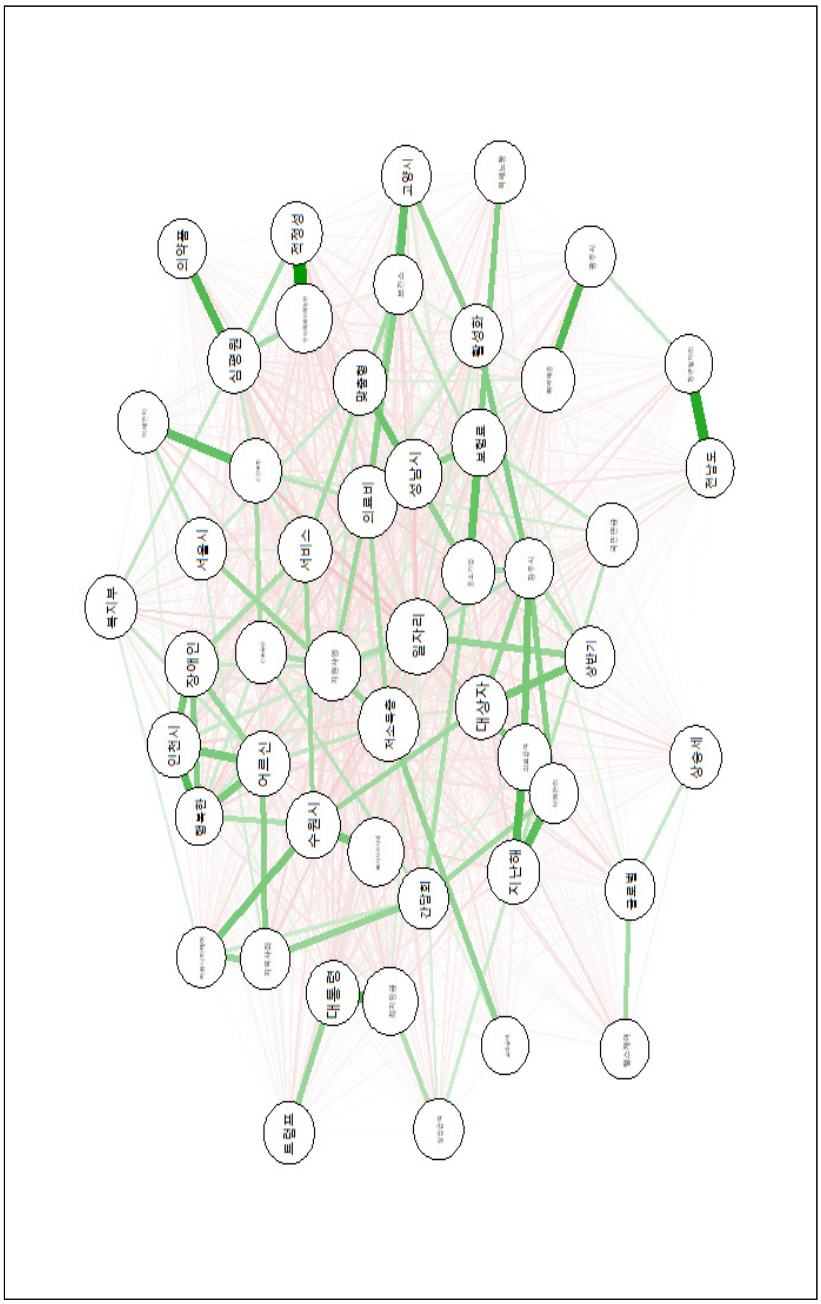
[그림 3-17] 2018년 11월 '건강보험' 네트워크 분석



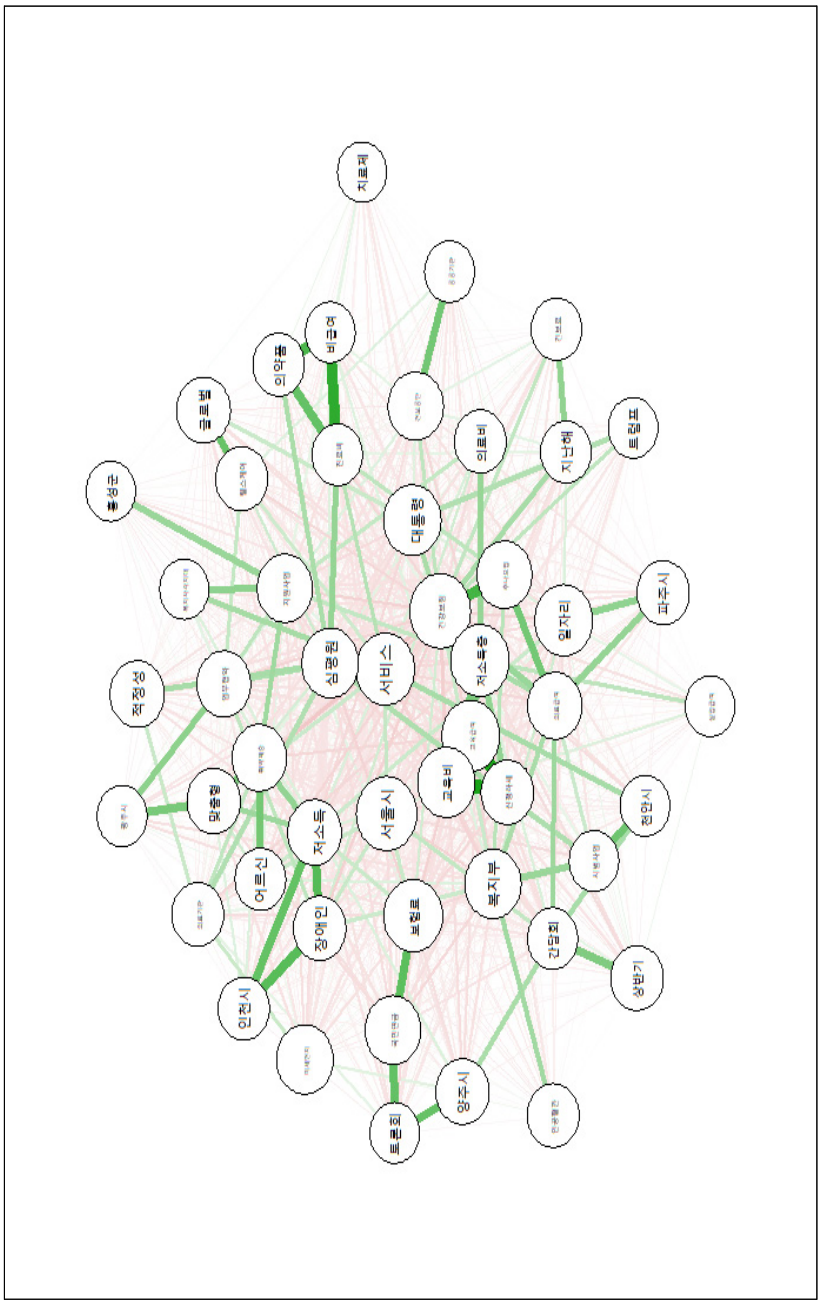
[그림 3-18] 2018년 12월 '건강보험' 네트워크 분석



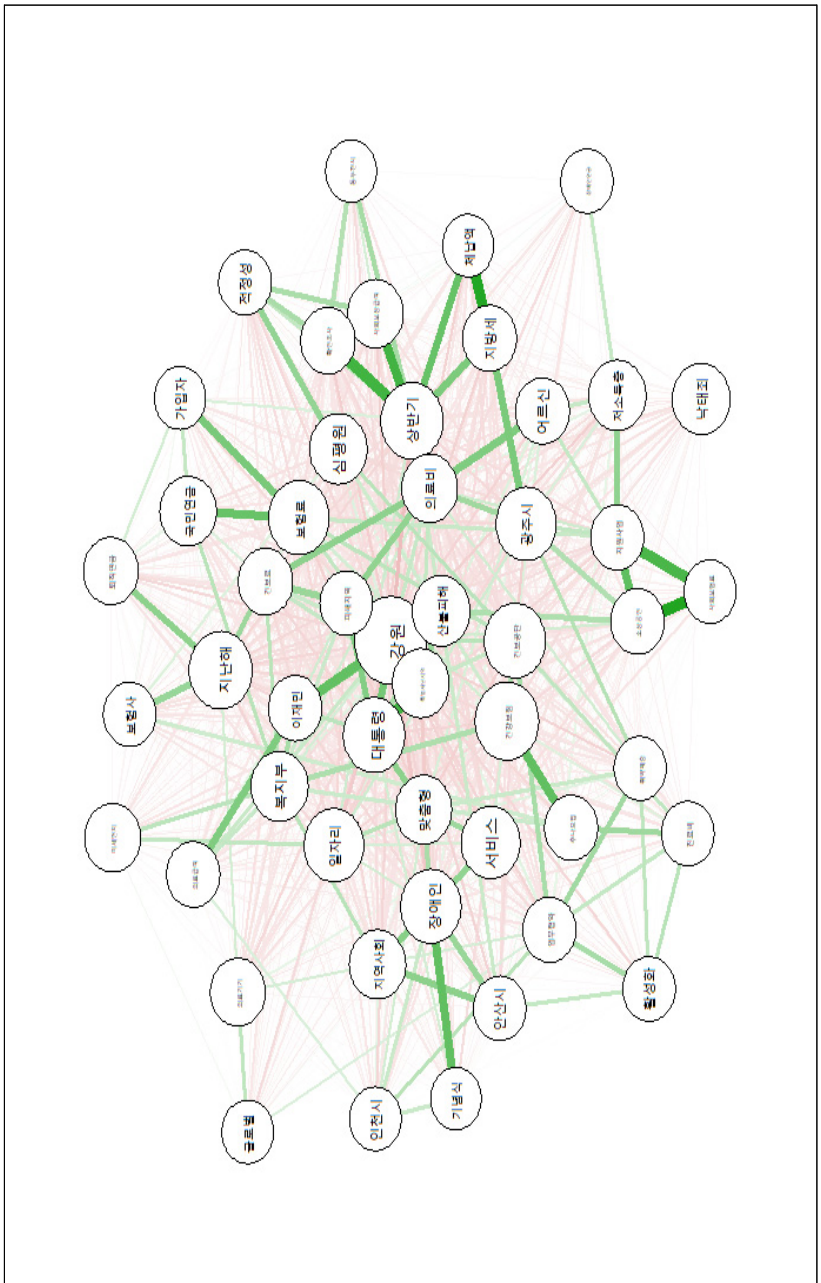
[그림 3-19] 2019년 1월 '건강보험' 네트워크 분석



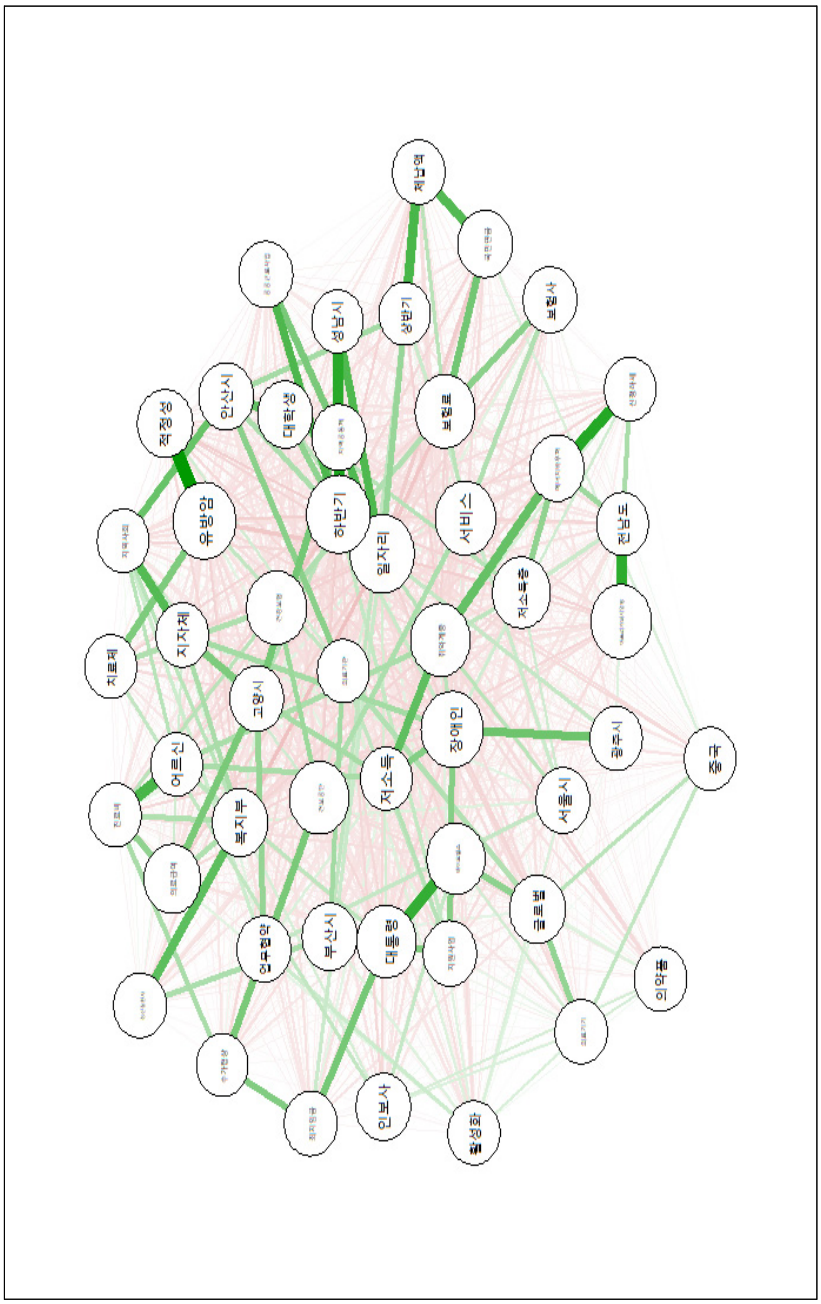
[그림 3-20] 2019년 2월 '건강보험' 네트워크 분석



[그림 3-21] 2019년 3월 '건강보험' 네트워크 분석

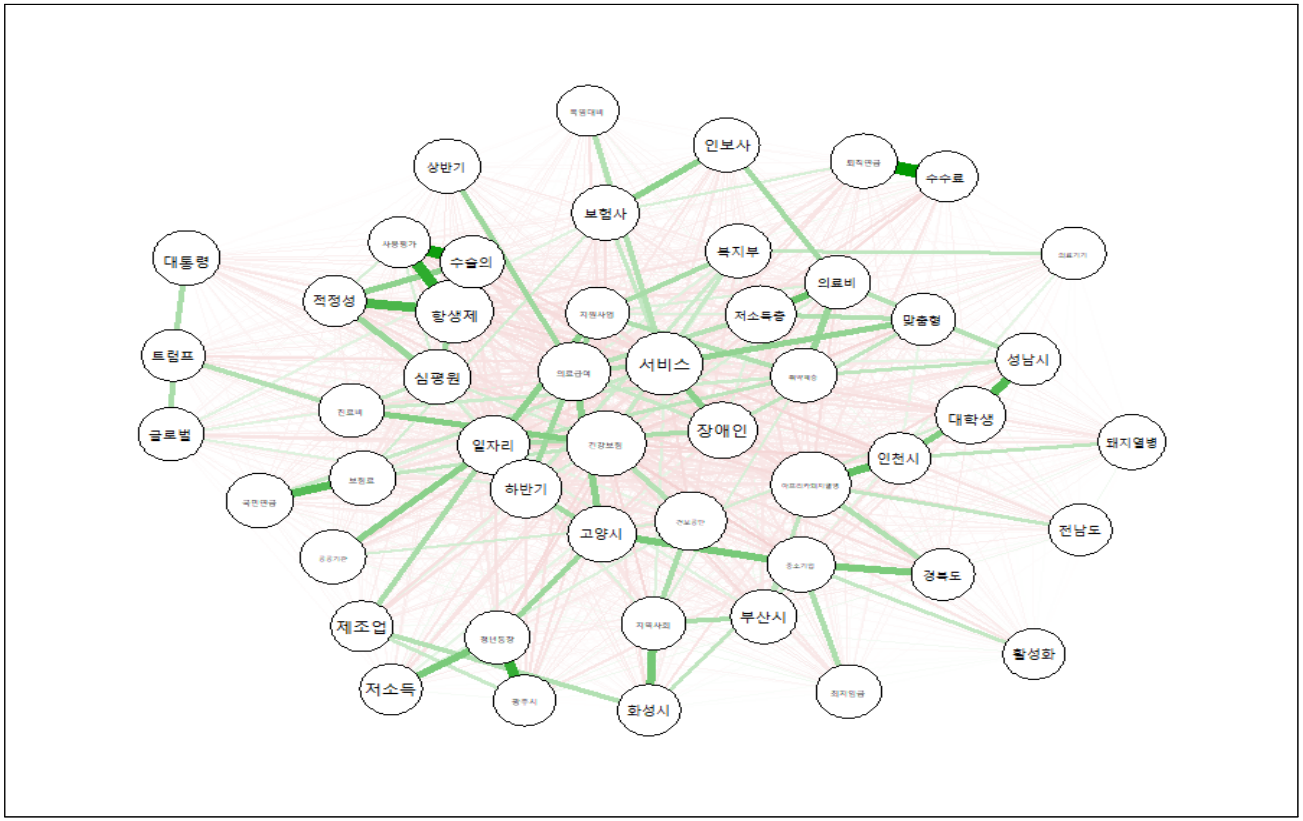


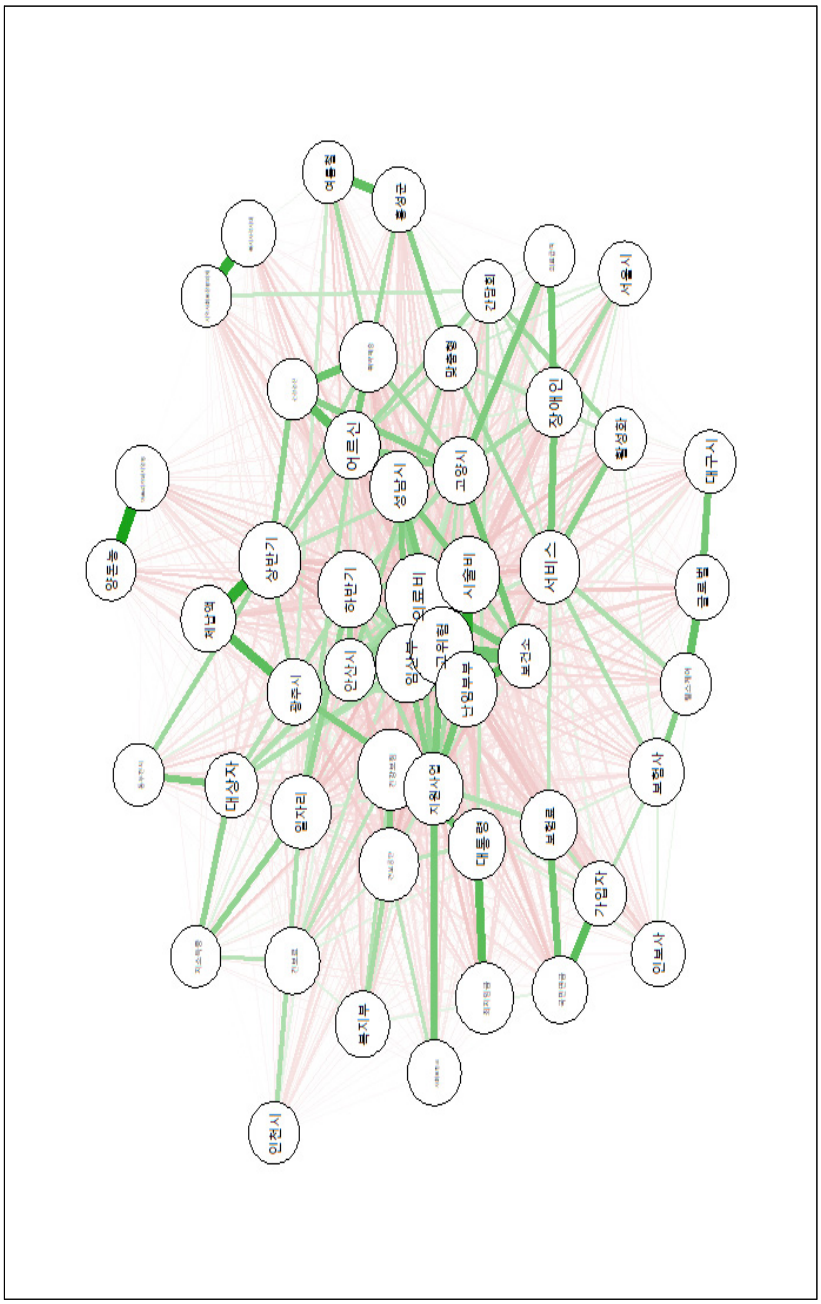
[그림 3-22] 2019년 4월 '건강보험' 네트워크 분석



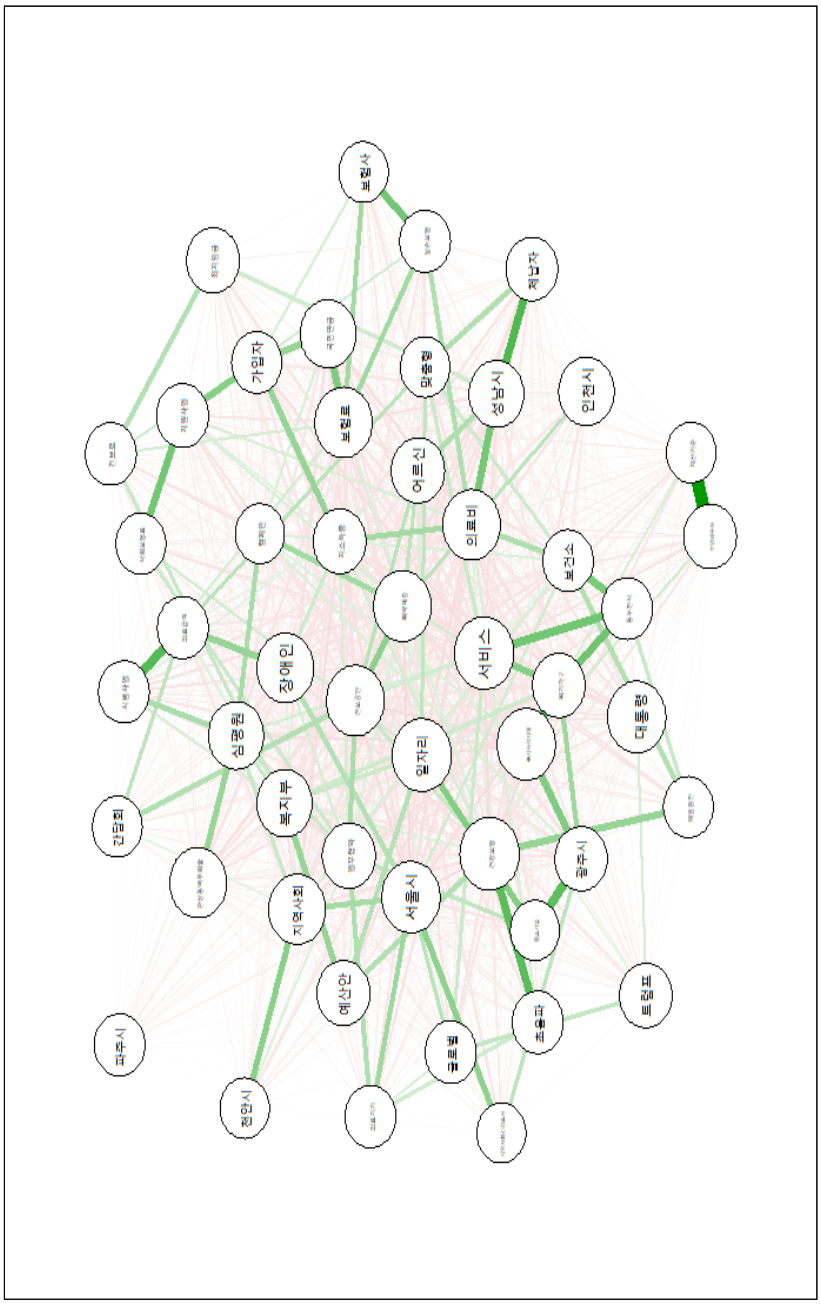
[그림 3-23] 2019년 5월 '건강보험' 네트워크 분석

[그림 3-24] 2019년 6월 '건강보험' 네트워크 분석





[그림 3-25] 2019년 7월 '건강보험' 네트워크 분석

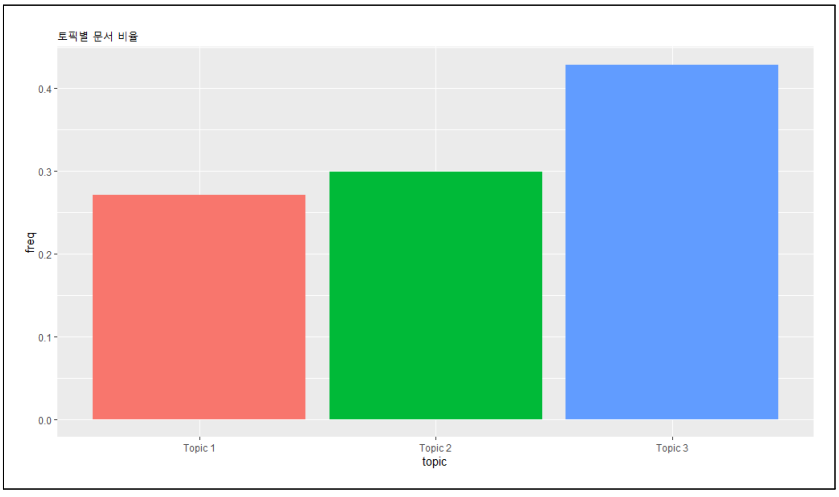


[그림 3-26] 2019년 8월 '건강보험' 네트워크 분석

3. 토픽모형

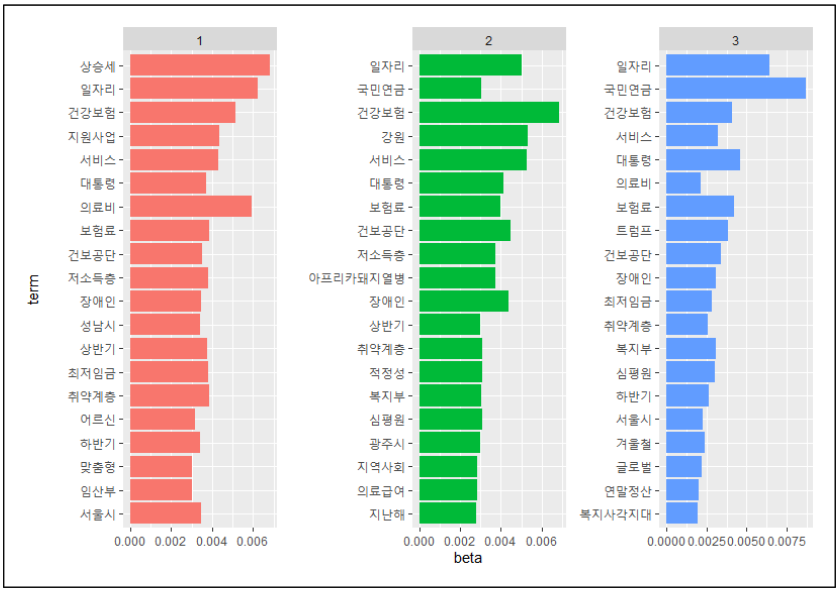
다음은 건강보험 키워드의 토픽모형에 대한 분석 결과이다. 저출산 키워드 분석과 마찬가지로 토픽 수를 3개부터 6개까지 변경해가며 결과를 살펴보았고 최종적으로 토픽 수 3개로 결정하였다. 토픽별 문서 비중을 보면 3번 토픽의 문서 비중이 상대적으로 높음을 알 수 있다.

〔그림 3-27〕 토픽별 '건강보험' 문서 비율

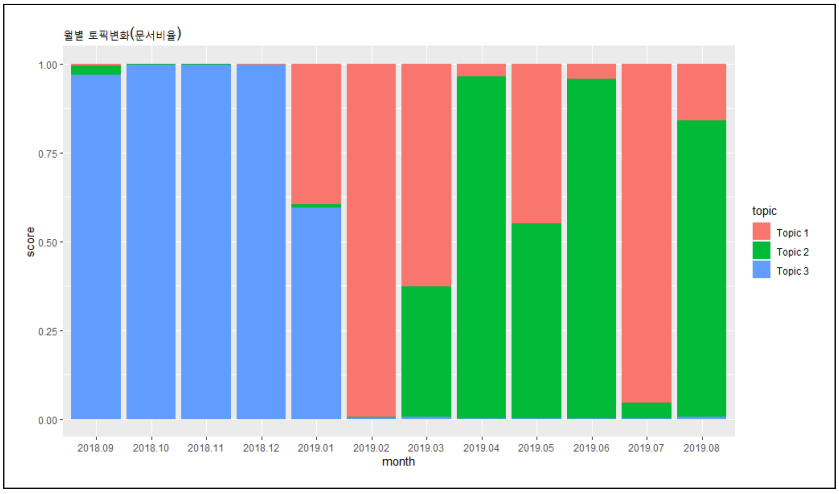


토픽 1은 건강보험 뿐만 아니라, 보건복지 정책과 관련된 키워드가 상대적으로 많이 언급되었고, 토픽 2는 건강보험과 직접적으로 관련있다고 볼 수 있는 키워드 중심으로 상대적으로 많이 언급되었다. 토픽 3은 토픽 1과 비슷한 것으로 보이지만, 월별 토픽 변화를 보면 토픽 3은 2018년도의 이슈임을 알 수 있다. 2019년 2월과 7월에 토픽1의 문서 비중이 높은 것은 정부에서 상·하반기 주요 정책들을 발표하는 것과 관련이 있다.

〔그림 3-28〕 토픽별 '건강보험' 상위 키워드



〔그림 3-29〕 월별 '건강보험' 토픽변화(문서 비율)



제 4 장 보건복지 뉴스기사 주제 분석을 위한 키워드 추출 방법론 연구

4 보건복지 뉴스기사 주제 분석을 위한 키워드 추출 방법론 연구

제1절 토픽과 키워드

텍스트 분석 연구주제들 중 하나로 문서의 주체 추출 방법이 있다. 문서의 주제 추출은 특정 문서 집합의 정보를 이용하여 어떤 문서의 주제를 추정하는 방법으로 다수의 문서들은 몇 개의 주제 분류군으로 군집화하거나, 새롭게 관측된 분서의 주제를 할당하는 것을 목표로 한다. 문서에 대한 주제 추정 방법론은 문서 내 식별 가능한 메타 정보 없이 자동으로 의미 있는 군집을 만들어 주어 기존에 사람이 수행했던 문서들을 분류하는 작업량을 크게 줄여주었다.

문서의 주제를 추정하는 가장 유명한 방법 중 하나로 LDA (latent dirichlet allocation)가 있다. LDA는 기본적으로 문서 생성에 대한 모형이다. LDA 모형은 문서들이 우리가 직접 관찰할 수 없는 주제들이 연속된 확률적 출현으로 구성되며, 실제로 관찰되는 문서 내 단어는 각각의 주어진 주제들에 대해 사용되는 단어들이 확률적으로 출현하여 이루어진다는 가정을 하고 있다. 즉, 우리가 읽고 있는 문서 내 단어의 이면에 숨겨져 있는 주제들이 있고, 실제로 그 주제들로부터 랜덤하게 표현된 단어들의 집합을 우리가 문서라고 부른다는 것이다.

이 LDA는 문장 생성방식을 개별적인 단어가 아닌 여러 개의 주제들의 랜덤한 출현 빈도로 먼저 모형화하였다는 점이 특이하다. 예를 들면, 어떤 문서가 3개의 어절로 구성이 되었다고 가정해보자, 이때 각 어절마다 주제들이 랜덤하게 하나씩 생성된다. 물론 이 주제는 우리가 관찰할 수

없는 잠재변수로 취급한다. 3개의 어절에 각각의 주제가 결정이 되면, 주제에 따른 단어들이 생성되어, 주어진 3개의 어절에 단어가 관찰된다. LDA는 이러한 방식으로 문서의 생성 모형을 가정한다. 이를 통해 LDA는 하나의 문서가 여러 개의 주제를 가질 수 있게 모형이 구성되어 있으며, 모형 내에서 문서의 주제의 비율을 추정함으로써 문서의 주제별 특징을 표현할 수 있게 하였다. 문서의 특징이 주제로서 표현되었기 때문에 우리는 쉽게 문서들을 주제별로 군집화 하여 분석할 수 있다.

LDA 내에서 정의된 주제는 우리가 알고 있는 어떤 의미를 가진 단어가 아니라 관찰된 문서내의 단어집합 내에서 단어 분포로 정의된다. 즉, LDA 모형으로부터 문서의 주제가 추정되었다 하더라도, 이 추정한 주제를 정성적으로 기술하기 위해서는 단어분포의 해석이 필요하다. LDA모형은 문서 내 주제의 분포를 알 수 있다는 점에서 매력적이지만, 문서 주제를 탐색하여 최종적으로 의미 있는 주제를 도출하기 위해서는 사람의 노력이 필요하다는 점에서 문서의 자동 분석에 한계가 있다.

문서 데이터 분석에서 주제 분석 이전에는 문서의 요약 방법으로 문서 내 키워드 추출 방법론이 연구되어 왔다. 키워드 추출은 문서의 주제추출보다 더 오래 전부터 연구되어 왔으며, 문서를 요약하고 특징을 추출하는데 큰 기여를 해 온 연구주제다. 문서를 요약하는 핵심적인 단어를 추정하는 방법이라는 점에서 문서의 주제 분석의 도구로서 이해할 수 있으며, 키워드 추출은 문서 내의 중요 단어를 직접 지정해주기 때문에 결과가 직관적이라는 장점을 가지고 있다. 하지만 문제 전체에서 키워드 후보군이 매우 많고, 문서의 주제가 단어로서 표현된다는 점은 주제를 군집화하는데 큰 어려움이 되었다. 이에 따라 키워드 추출 방법론 연구는 문서 분석의 초기 단계에서 활발히 연구가 되었지만, LDA가 널리 사용되면서 연구결과가 감소하는 추세를 보였다.

문서의 주제 추출이라는 점에서 LDA와 키워드 추출은 유사한 성격을 가진다. 키워드 추출이 문서의 주제를 나타내는 중요한 단어를 선택하는 방법이라면, LDA는 문서의 주제를 표현하기 위해 주요 단어의 비중을 제공하는 방법으로 볼 수 있다. 이는 문서 내의 주제를 표현함에 있어 단어를 선택하는 것과 같은 단어 집합의 불연속적인 특성값을 이용할지 혹은 단어의 분포를 계산하는 방법 즉 단어의 연속적인 특성값을 이용하는 방법을 선택하느냐의 문제로 볼 수 있다. 전자는 키워드 추출 방법으로, 후자는 LDA 방법에 해당한다.

문서의 주제 추출을 위한 수단으로서 LDA와 키워드 추출 방법은 분명한 장단점을 가지고 있다. 먼저 LDA는 문서 내 여러 개의 주제를 가진 모형을 가정한다. 이 모형을 통해 추정한 문서의 주제가 문서 집합내의 단어 분포로 표현된다는 점에서 주제를 이루고 있는 고차원에 있는 단어의 패턴을 잘 모형화 할 수 있다는 장점이 있다. 하지만, LDA는 단어의 분포에 대한 정성적인 해석이 필요하기 때문에, 문서의 주제를 파악하는데 전문적인 지식이 필요하다는 단점이 있다. 반면 키워드 추출 방법은 문서의 주제가 몇 개 단어들의 집합으로 표현되므로, 즉각적으로 주제를 판단할 수 있다는 장점이 있다. 하지만, 다수의 단어들의 조합의 만들어내는 주제들은 유사 문서의 군집화, 주제의 표현에 어려움이 있다는 단점이 있다. 두 방법 모두 공통적으로 결과물에 대한 주제 해석을 위해서는 다른 방식의 노력이 필요하다는 한계를 가지고 있다.

현재 LDA가 텍스트 분석에 많이 이용되고 있다는 점을 고려한다면, 연구키워드 추출 방법은 LDA의 한계점을 보완하는 방편이 될 수 있을 것으로 보인다. 키워드 추출방법은 문서 내에서 가장 중요한 단어가 무엇인지 추정해주기 때문에, LDA와 같은 주제 추정모형의 주제 해석과 같은 문제

에 유용하게 사용될 것으로 보인다. 이런 점은 최근 연구를 통해 주목받으면서 LDA를 주제 추출 방법론을 보완하는 측면에서 키워드 추출방법론이 다시 연구되기 시작하였다.

키워드는 중요한 단어라는 뜻으로 중요도를 어떻게 평가하는가에 따라 다르게 정의될 수 있어, 문서의 주제를 표현하는 일반어, 특수어의 관계로 다르게 해석해 볼 수 있다. 예를 들어 문서 집합 내에서 공통적으로 중요도를 가지는 단어가 키워드가 될 수 있는 반면 문서 집합 내에서 문서간 주제를 구별해주는 특수한 단어가 키워드가 될 수도 있을 것이다. 이렇게 다르게 정의된 중요도에 따라서 문서의 키워드는 다르게 정의될 수 있고, 문서의 주제를 다른 방향에서 해석할 수 있다. 이를 위해서는 네트워크 기반의 단어 중요도를 평가한 기존의 연구결과가 도움이 될 것이라 생각된다.

한편 키워드가 단어 벡터가 정의된 고차원 공간이 아닌 문장 내에서 의미를 이용하여 저차원에서 키워드를 추출하는 방법이 필요하다. 이는 기존의 키워드 추출이후 주제를 해석하는데 필요한 추상적인 개념을 완성하기 위해 필요한 작업이며, 이러한 접근방법은 단어 임베딩과 같은 단어 특성 추출방법의 연구결과를 이용할 수 있을 것이다.

이 장에서는 보건/복지 뉴스기사 주제 분석을 위해서 기존에 개발된 주제추출 및 키워드 임베딩 방법에 기초한 키워드 추출 방법론을 알아보고자 한다. 이를 위해 다음과 같은 순서에 따라 논의를 진행하고자 한다.

1) 문서 내의 단어 특징 추출 및 주제 추출 방법론: 문서의 단어추출과 주제추출을 위해 널리 사용되었던 기존 방법론을 살펴본다. 여기서는 네트워크 기반 단어분석, 단어 특징 추출 방법, 토픽 모형을 다룬다.

2) 키워드 추출방법론 소개: 키워드 추출에 사용되는 최신 방법론을 살펴본다. 네트워크 기반, 단어 특성 추출 기반, 토픽 모형 기반 키워드 추출 방법론을 살펴본다.

3) 키워드 추출방법론의 고도화 기술 조사: 구조적 성김성을 이용한 단어 특징 추출, 감성 분류정보를 반영한 특징 추출, 온라인 학습 방법을 살펴본다.

4) 결론: 보건/복지 분야의 기사 주제 정제를 위한 키워드 분석 방법 고도화에 대한 제언을 담고 있다.

제2절 배경 연구

이 절에서는 후에 서술할 키워드 추출의 방법들을 자세히 소개하기 위해 필요한 기반이 되는 배경 연구들을 소개한다. 서술할 배경 연구들은 단어의 관계를 그래프로 표현한 네트워크 기반 단어분석, 문서 내의 단어를 저차원 공간의 원소로 변환하는 워드 임베딩 방법, 그리고 문서의 주제를 추출하는 LDA 방법이 있다.

1. 네트워크 기반 단어 분석

가. 네트워크 모형

네트워크는 어떤 관계를 가진 구조를 표현하기 위해 사용되는 유용한 도구다. 네트워크로 모형화 된 대상은 미시적으로는 하나의 단어 내의 음소가 될 수도 있고, 더 크게는 문서 내 집합에서 문서가 될 수도 있다. 단어의 중요도를 파악하는 네크워크 모형에서는 문장 내 단어들이 분석 단위가 된다. 네트워크 분석에 사용되는 대표적인 모형으로 그래피컬 모형 (graphical model)이 있는데 확률 변수 사이의 조건부 독립성 (conditional independency)을 추정하고 표현하기 위해 널리 사용된다. 일반적인 네트워크 모형에서 확률 변수는 노드 (node)로, 변수 사이의 조건부 독립 관계는 엣지(edge)를 통해 표현한다. 일반적으로 그래프 G는 노드 집합 V와 엣지 집합 E를 이용하여 $G=(V,E)$와 같이 표현한다.

네트워크 모형은 크게 구조적으로 두 가지 형태로 나눌 수 있으며 확률 변수 사이의 관계를 추정하는 비 방향성 그래프 (undirected graph)와 변수 사이의 관계와 더불어 변수 사이의 순서 (ordering)를 표현할 수 있

는 방향성 그래프 (directed graph)가 있다. 방향성이 있는 모형은 인과성 연구에 중요하게 사용되며, 대표적으로 연구분야로 방향 그래피컬 모형, 베이지안 네트워크 등이 있다. 문서 데이터와 같이 데이터의 차원이 큰 경우에는 방향성 그래프 모형을 적합하기 어렵기 때문에 이 절에서는 키워드 추출에 주로 사용되는 비 방향성 그래프 모형 방법을 소개한다.

먼저, 단어 사이의 조건부 종속성을 추정하기 위한 대표적인 비 방향성 그래프 모형인 가우시안 그래피컬 (Gaussian graphical) 모형을 소개한다. 작성된 모든 단어의 집합을 $W = \{w_1, \cdots, w_p\}$라고 하면 가우시안 그래피컬 모형은 $(w_1, \cdots, w_p)^\top \sim N(\mu, \Omega^{-1})$을 가정한다. 이 때, 반응변수를 w_j, 설명변수를 w_i, $i \in \{1, \cdots, p\} - \{j\}$인 회귀 모형을 적합하고, w_k, $k = \{1, \cdots, p\} - \{i\}$에 대응되는 최소제곱추정량 (least square estimator) β_k^{-j}를 계산한다. 즉, 다음을 만족하는 계수를 추정한다.

$$\widehat{\beta^{-j}}(\lambda) = \mathrm{arg}min_{\beta^{-j}} \sum_{i=1}^{n} (w_{ij} - \beta_0^{-j} - \sum_{k \neq j} \beta_k^{-j} w_{ik})^2$$

단어의 빈도에 대응되는 확률 변수가 다변량 정규분포를 따른다는 가정 하에서 추정된 회귀계수가 $\beta_k^{-j} = 0$일 때 $(\Omega)_{jk} = 0$임과 동치이다. 가우시안 그래피컬 모형은 이러한 이론적 배경을 이용하여 추정된 회귀계수가 $\beta_k^{-j} = 0$인 경우 단어 j와 단어 k 사이의 엣지를 연결하지 않는 방식으로 작동한다. 그러나 실제 데이터에 모형을 적합할 때는 변수 사이에 연관성이 없더라도 오차 때문에 최소 제곱추정량 계수를 정확하게 0으로 추정하기 어렵다는 문제점이 발생한다. 이러한 경우 두 변수 사이에 관계가 없음에도 불구하고 연관성이 존재한다는 결과를 도출하게 되어 문제가 된다. 이런 점을 보완하기 위해 그래피컬 라쏘 (graphical lasso) 모형

이 제안되었다.

그래피컬 라쏘 (graphical lasso) 모형은 가우시안 그래피컬 모형에서 회귀계수를 정확하게 0으로 줄 수 있는 분석방법이다. 그래피컬 라쏘 모형은 가우시안 그래피컬 모형에서 회귀 계수를 추정할 때 라쏘 (lasso) 벌점 함수를 이용하여 w_k에 대응되는 회귀계수 β_k^{-j}를 추정할 때 계수를 0으로 추정하여 연결 관계에 희소성 (sparsity)을 줄 수 있다. 라쏘 회귀 계수는 고정된 $\lambda > 0$과 모든 j에 대해서 아래의 식을 이용하여 추정한다.

$$\widehat{\beta^{-j}}(\lambda) = \arg min_{\beta^{-j}} \sum_{i=1}^{n} (w_{ij} - \beta_0^{-j} - \sum_{k \neq j} \beta_k^{-j} w_{ik})^2 + \lambda \sum_{k \neq j} |\beta_k^{-j}|$$

대표적인 비 방향성 그래프 모형인 가우시안 그래피컬 모형과 그래피컬 라쏘 모형은 변수 사이의 관계를 회귀계수를 이용하여 추정하는 모형이다. 즉, 문서 데이터에 그래피컬 모형을 적합하여 단어들 간의 관계를 추정하고 이 관계로부터 키워드를 추출할 수 있다. 이 때, 어떤 단어가 중요한 단어인지 혹은 중심이 되는 단어인지를 파악하기 위한 측도로 중심성 (centrality)이 사용될 수 있다.

나. 중심성 (Centrality)

중심성은 그래프 $G=(V,E)$에서 엣지의 연결 여부로부터 가장 중요한 노드를 찾는 측도이다. 즉 중심성은 그래프에서 엣지의 실함수로 정의되며 다양한 형태의 중심성 측도가 존재한다. 예로, 연결 중심성 (degree centrality)는 그래프의 노드에 연결되어 있는 엣지의 개수로 정의되며 다른 많은 노드들과 연결되어 있으면 중요한 노드라는 관점에서 이해할

수 있다. 아래는 연결 중심성 외에도 대표적으로 사용되는 몇 가지 중심성에 대해서 소개한다.

매개 중심성 (betweenness centrality)은 그래프에서 최단 경로의 개념을 사용하는 중심성이다. 모든 두 개의 노드에 대해서 두 노드를 이어주는 최단경로가 존재한다고 하면 한 노드를 지나는 최단경로의 개수를 매개 중심성으로 정의한다. 즉, σ_{st}를 노드 s에서 노드 t로 향하는 모든 최단경로의 개수로 정의하고 $\sigma_{st}(v)$를 σ_{st}중 노드 v를 지나는 최단경로의 개수라고 정의하면 매개 중심성은 $C(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$와 같이 표현할 수 있다. 이는 다른 노드들의 관계 사이에 포함되는 노드가 중요하다는 관점에서 이해할 수 있다.

근접 중심성 (closeness centrality)은 매개 중심성과 마찬가지로 최단 경로의 개념을 사용하지만 매개 중심성과는 다르게 노드 v로부터 다른 모든 노드와의 평균 최단 경로의 길이를 계산하는 방법이다. 예를 들어, 단어 간의 네트워크 구조에서 한 단어가 다른 단어와의 거리가 가깝다면 해당 단어는 중요한 단어라고 판단할 수 있다. 즉, $d(y,v)$를 노드 y와 노드 v 사이의 거리라고 하면 근접 중심성은 $C(v) = \frac{1}{\sum_{y \neq v} d(y,v)}$와 같이 표현할 수 있다.

고유벡터 중심성 (eigenvector centrality)는 네트워크에서 노드의 영향력을 측정하는 측도로 사용된다. 노드 y와 노드 v가 연결되어 있으면 $a_{y,v} = 1$의 값을 가지고 연결되어 있지 않다면 $a_{v,t} = 0$인 원소를 모은 인접 행렬 (adjacency matrix)을 $A = (a_{v,t})$라고 하자. 고유벡터 중심성은 $C(v) = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} C(t)$을 통해 계산된다. 키워드 추출 측면에서는 중요한

단어들과 연결된 단어가 중요한 단어라는 관점에서 사용할 수 있다.

마지막으로 PageRank 중심성은 고유벡터 중심성을 확장한 개념으로 $C(v) = \alpha \sum_j a_{ji} \frac{C(j)}{L(j)} + \frac{1-\alpha}{N}, L(j) = \sum_i a_{ji}$와 같이 정의되는 중심성 측도이다.

네트워크 모형을 이용하여 중심단어를 도출하는 방법은 직관적으로 이해하기 쉽고 이론적 배경이 뒷받침되고 있지만 데이터의 차원이 큰 경우 많은 계산이 필요하기 때문에 쉽게 사용하기 어렵다. 이러한 이유 때문에 최근 많은 키워드 도출 방법들은 단어 벡터를 저차원의 숫자 벡터로 변환하는 방법인 워드 임베딩 방법이 선행되는 경우가 많다. 아래에서는 이러한 워드 임베딩 방법에 대해서 소개한다.

2. 워드 임베딩 방법

단어를 모형 적합에 사용하기 위해서 단어를 숫자 벡터로 변환하는 과정이 반드시 필요하다. 전통적인 방법에서는 전체 단어 집합의 개수를 V개라고 하면 하나의 단어를 해당하는 단어 위치에 1의 값을 가지고 있는 지시 벡터 (indicator vector)로 변환하는 방법을 사용했다.

아래에서 지시 벡터를 특정한 규칙을 통해 가중치 벡터로 변환하는 대표적인 방법인 TF-IDF를 소개한다. 또한 최근 연구되고 있는 신경망 (neural network) 모형을 통해 단어 벡터를 V보다 작은 차원을 가지는 숫자 벡터로 변환하는 방법과 마지막으로 LDA를 소개한다.

가. TF-IDF

TF-IDF는 문서를 단어 단위로 분해하여 생성된 문서-단어 행렬(document-term matrix)에서 사용할 수 있는 방법이다. 문서-단어 행렬의 각 행은 문장에서 단어의 등장 여부를 0과 1을 통해 표현하는 지시벡터이거나 단어의 등장 횟수를 지시벡터의 원소에 곱해준 숫자로 표현한다.

TF-IDF는 문서-단어 행렬의 모든 원소를 특정한 규칙을 통해 변환하는 방식으로 TF (term frequency) 방법과 IDF (inverse document frequency) 방법으로 행렬을 변환한 뒤 변환된 값을 곱해서 가중치를 출력하는 방법이다. 아래에서는 TF 방법과 IDF 방법에 대해서 각각 소개한다.

TF 방법은 단어가 문서에 등장하는 빈도가 많으면 가중치를 크게 할당하는 방법으로 특정한 단어가 해당 문서에서 많이 등장할수록 중요한 단어로 고려하는 가중치로 직관적으로 이해할 수 있다. 문서 d에서 등장한 단어 t의 빈도를 $f_{t,v}$라고 하면 다음과 같은 TF 변환 방법이 사용된다.

1) binary: $I(f_{t,d} \geq 1)$
2) raw frequency: $f_{t,d}$
3) log normalization: $1 + \log(f_{t,d})$
4) double normalization K: $K + K\frac{f_{t,d}}{\max_{t' \in d} f_{t',d}}$)

IDF 방법은 단어가 모든 문서에서 등장하는 빈도가 많다면 가중치를 작게 할당하는 방법으로 전체 문서에서 적게 등장하는 단어일수록 중요한 단어로 고려하는 가중치로 직관적으로 이해할 수 있다. 단어 t를 포함

하는 문서의 개수를 $n_t = |\{d \in D : t \in d\}|$로, 문서의 개수를 $N = |D|$라고 하면 다음과 같이 다양한 변환 방법이 있다.

1) unary: 1

2) frequency: $\log \frac{N}{n_t}$

3) frequency smooth: $\log \frac{N}{1+n_t}$

4) frequency max: $\log \frac{\max_{t' \in d} n_{t'}}{1+n_t}$

5) document frequency: $\log \frac{N-n_t}{n_t}$

정리하면 문서-단어 행렬을 $W = (W_{i,j})_{i=1,\cdots,D, j=1,\cdots,V}$라고 하면 TF-IDF 변환을 이용하여 출력된 행렬의 각 원소는 $TFIDF(W_{i,j}) = TF(W_{i,j}) \times IDF(W_{i,j})$와 같이 표현할 수 있다. 이는 각 행, 각 열을 따로 계산한 뒤에 곱을 해주는 방식으로 계산할 수 있기 때문에 병렬 처리가 가능하며 실증적으로 모형에 사용할 때 더 좋은 성능을 갖는다는 장점이 있다.

그러나 문서-단어 행렬을 변환하는 방법인 TF-IDF 방법은 본질적으로 단어의 개수가 증가하면 변환 행렬의 차원 또한 증가하기 때문에 문서-단어 행렬을 모형에 적합할 때와 마찬가지로 차원이 큰 데이터에 대해서 모형 적합이 어렵다는 단점이 있다. 최근 연구되는 많은 방법은 데이터의 차원을 단어의 개수보다 작도록 구성하며 아래에서 소개하는 두 방법 또한 마찬가지로 변환된 벡터의 차원을 작은 값으로 고정하여 단어의 특징 벡터를 추출하는 방법이다.

나. Word2Vec [Mikolov, Tomas, et al. 2013.]

Word2Vec은 텍스트 마이닝 분야에서 신경망을 이용하여 특징 벡터를 추출할 때 사용되는 대표적인 방법으로 크게 Continuous Bag of Words (CBOW)와 Skip-gram 방법이 있다. CBOW 모형은 주변 단어로부터 해당 위치의 단어를 예측하는 모형으로부터 특징 벡터를 추출하는 방법이며 이와 반대로 Skip-gram 방법은 해당 위치의 단어로부터 주변 단어를 예측하는 모형으로부터 특징 벡터를 추출하는 방법이다. 두 방법은 공통적으로 입력 단어가 주어졌을 때 출력 단어가 등장할 확률을 최대화하는 방향으로 학습한다. 이 때, 단어의 개수에 비례하여 모수의 개수가 증가하는 것을 보완하기 위해 Negative sampling 등의 방법을 사용한다. 아래에서는 CBOW와 Skip-gram 모형에 대해 간략히 소개한다.

두 모형을 설명하기 위해 $C=[w_1,\cdots,w_N]$를 단어 $w_i \in W$로 구성된 벡터라고 하고 단어 $w_i \in W$ 주변의 l개의 단어 벡터를 context라고 부르며 $(c_{i-l},\cdots,c_{i-1},c_{i+1},\cdots c_{i+l}),\ c_j \in C,\ j \in [i-l,i+l]$와 같이 정의한다.

CBOW는 로그 가능도 함수

$$L_{cbow}=\sum_{i=1}^{N}\log P(w_i|h_i),\ P(w_i|h_i)=\frac{\exp(\overrightarrow{w_i}^{\top}\overrightarrow{h_i})}{\sum_{w\in W}\exp(\overrightarrow{w}^{\top}\overrightarrow{h_i})}$$

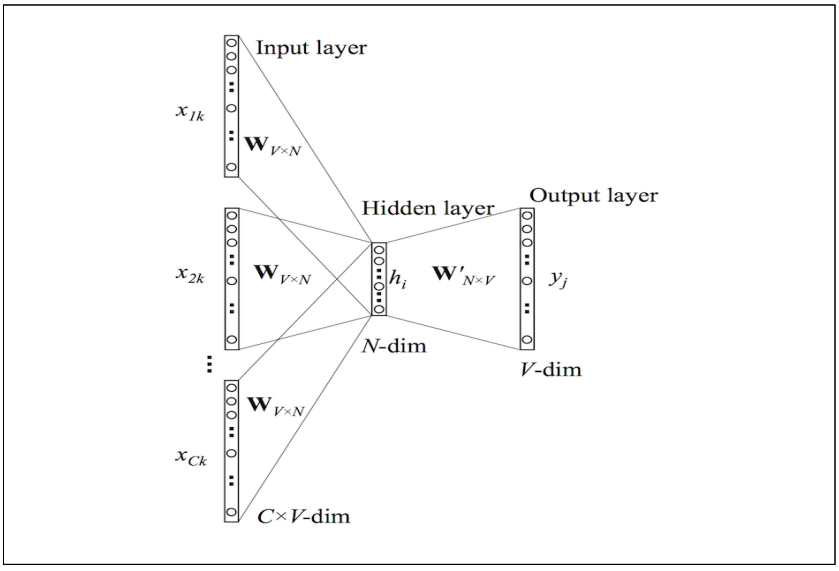
를 최대화하여 특징 벡터를 도출한다. 여기에서 $\overrightarrow{h_i}$는 c_j의 평균 벡터인 $\overrightarrow{h_i}=\frac{1}{2l}\sum_{j=i-l,j\neq l}^{j=i+l}\overrightarrow{c_j}$이며 $\overrightarrow{w}$는 w의 선형변환으로 생성되는 특징벡터이다.

Skip-gram 모형은 CBOW와 유사하게 로그 가능도 함수

$$L_{sgram} = \sum_{i=1-l \le}^{N} \sum_{c \le l, c \neq 0} \log P(w_{i+c}|w_i)$$

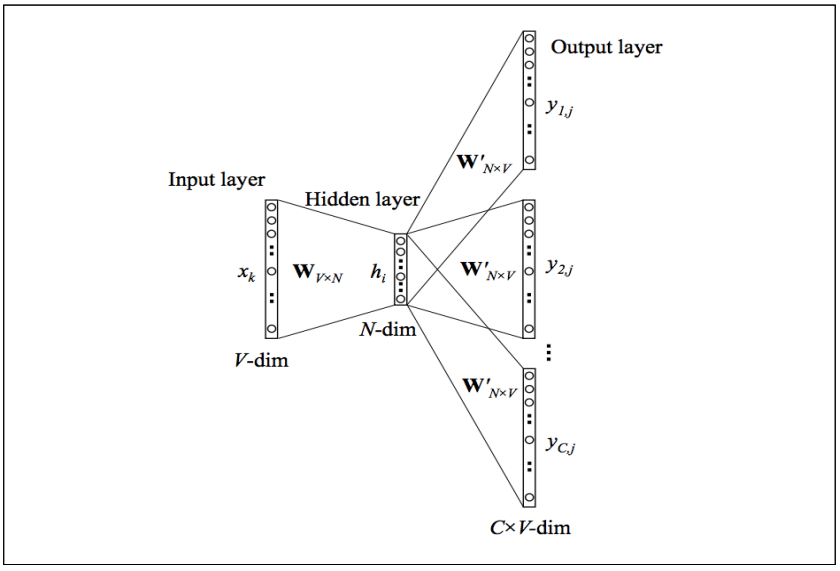
을 최대화하여 특징 벡터를 도출하는 방법이다. 두 방법의 큰 차이점은 입력값과 출력값의 방향성이 서로 반대로 되어 있다는 점이다. 이러한 차이점은 두 모형을 도식화한 아래의 그림으로부터 명확히 파악할 수 있다.

[그림 4-1] CBOW 도식화



자료: Sarwan, N. S. (2017).

〔그림 4-2〕 Skip-gram 도식화



자료: Sarwan, N. S. (2017).

다. GloVe [Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014.]

GloVe는 Word2Vec의 한 단어의 주변 l개 단어만을 모형 구성에 사용한다는 점을 단점으로 보고 주변 단어뿐만 아니라 전체 문서에서 동시 발생 (Co-occurrence) 확률을 보존하며 특징 벡터를 학습하도록 제안된 방법이다. 동시 발생 확률은 동시 발생 빈도 행렬 (co-occurrence matrix)을 X_{ij}라고 하면 $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$로 정의된다.

GloVe에서 동시 발생 확률을 보존하기 위한 모형 구성 과정은 다음과 같다. 단어의 특징 벡터를 w, 문서의 특징 벡터를 $\tilde{w}$라고 하면 두 단어의

특징 벡터 차와 문서 특징 벡터의 내적의 함수 F가 동시 발생 확률과 같아지도록 하는 변환을 고려한다. 즉, 이는 $F((w_i - w_j)^\top \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$와 같이 표현된다. 이러한 목적으로부터 함수 F에 대칭성 (symmetry), 준동형 사상 (homomorphism) 등의 조건들을 고려하면 만족하는 함수는 지수함수이기 때문에 F를 지수함수로 치환하고 식을 변형하면 최종적으로 다음과 같은 목적함수를 구할 수 있다.

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^\top \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

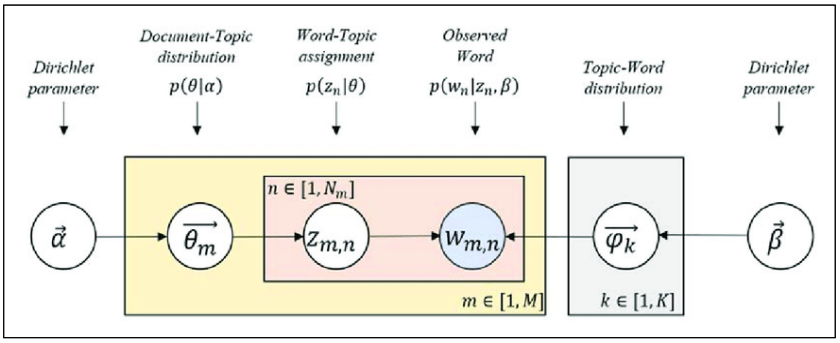
$$f(x) = \begin{cases} (\frac{x}{x_{\max}})^\alpha \text{ if } x < x_{\max} \\ 1 \qquad else \end{cases}$$

목적함수에서 $f(X_{ij})$는 X_{ij}에 대한 가중치로 X_{ij}가 큰 값을 가졌을 때 모형에 주는 영향을 줄이기 위해 곱해준 함수로 해석할 수 있다.

3. 토픽 모델링 (LDA)

LDA는 문서에서 주제를 발견하기 위해 사용된 생성 모형 (generative model) 이다. LDA는 전체 문서를 주제들의 혼합 (mixture)로 표현하며 각 주제는 단어의 분포이다. 이를 쉽게 표현하면 LDA는 전체 문서에서 각 단어를 특정한 토픽으로부터 생성하는 모형이다. 이를 도식화 하면 아래와 같다.

[그림 4-3] LDA 모형



자료: Lee, Junseok, et al. (2018)

위 그림으로부터 단순한 형태의 LDA의 작동원리를 다음과 같이 설명할 수 있다. 먼저 $\theta_d,\ d = 1,\cdots,D$는 문서 d에서 주제들의 분포이며 모수가 α인 디리클레 분포를 따른다. 전체 주제의 개수가 K개라고 하면 α는 K차원의 벡터이다. 또한 $\phi_k,\ k = 1,\cdots,K$는 한 주제에서 단어의 분포를 의미하며 모수가 β인 디리클레 분포를 따른다. 즉, 전체 단어 집합의 개수를 W라고 하면 β는 W차원의 벡터이다. 이 때, 단어는 θ_d로부터 하나의 토픽 $z_{d,n}$을 생성하고 해당 토픽 $\phi_{z_{d,n}}$에서 생성한다. 위 내용을 수식으로 요약하여 정리하면 다음과 같다.

$$\phi_i|\beta \sim Dir(\beta),\ \ i \in \{1, \cdots, K\},\ \ \phi_i \in R^V$$

$$\theta_j|\alpha \sim Dir(\alpha),\ \ j \in \{1, \cdots, D\},\ \ \phi_i \in R^K$$

$$z_{j,t}|\theta_j \sim Multi(1, \theta_j),\ \ z_{j,t} \in \{1, \cdots, K\}$$

$$w_{j,t}|\phi_{z_{j,t}} \sim Multi(1, \phi_{z_{j,t}}),\ \ w_{j,t} \in \{1, \cdots, V\}$$

LDA에서 위의 변수들 ϕ, θ, z는 잠재 변수로 추정해야 한다. 즉, $P(\theta,\phi,z|w,\alpha,\beta) = \frac{P(\theta,\phi,z,w|\alpha,\beta)}{P(w|\alpha,\beta)}$의 식을 해결해야 하지만 이 분포를

직접 계산하는 것은 매우 어렵다. 따라서 LDA 모형을 추정하기 위해 Markov chain Monte Carlo 방법 중 하나인 Gibb's sampler 방법을 사용한다. Gibb's sampler는 고차원 분포의 표본을 저차원 조건부 분포로부터 순차적으로 추출하는 방법이다. Gibb's sampler에 대한 자세한 부분은 뒤의 키워드 추출 방법론에서 소개한다.

제3절 키워드 추출 방법론

1. 그래프를 이용한 키워드 추출

그래프 구조를 이용한 키워드 추출 방법 중 첫 번째는 단일문서에서 중심성을 이용하는 알고리즘이다. 키워드를 뽑을 때 주어진 문서 데이터를 그래프로 표현하는데, 이 그래프의 노드는 문서 데이터에서 단어의 빈도가 되며 엣지는 문서에서 동시에 발생하는 빈도로 표시된다. 그래프 구조에서 단어의 중심성이 높은 단어는 다른 단어와의 연관성이 높기 때문에 전체 문서 데이터에서 키워드가 될 가능성이 높다.

분석에서는 데이터의 간결성을 위해 데이터에 대한 사전 정제가 필요하다. 정제 과정은 분석 방법마다 달라질 수 있지만 공통적으로는 다음의 과정을 거친다. 사전 정제 과정은 1) 약어 사용, 2) 숫자 제거, 3) 불용어 제거, 4) 물음표, 마침표, 느낌표 등 구분자를 제외한 특수문자 삭제, 5) 빈도수가 적은 단어 제거, 6) 어간 추출과 같은 순서로 진행된다.

중심성을 이용하여 키워드를 추출하기 위한 아래 세 가지 방법론을 살펴보자.

키워드를 추출하는 첫 번째 방법은 편심 기반 키워드 식별 (Eccentricity-based keyword identification)이다. 그래프 $G=(V,\ \mathrm{E})$에서 엣지의 거리 함수를 $\omega \in (0,1]$라고 하자. 엣지의 거리 함수는 두 단어 u, v의 동시 발생 빈도를 $c(u,v)$라고 했을 때 동시 발생 빈도의 함수로 다음과 같이 정의된다.

$$\omega(u,v) = \frac{1}{c(u,v)} I(c(u,v) \neq 0)$$

두 단어의 동시 발생 빈도가 커질수록 단어 사이의 거리가 작아짐을 알 수 있다. 즉 $c(u,v) = 0$ 이면 그래프 구조에서 두 단어 u, v간에 거리가

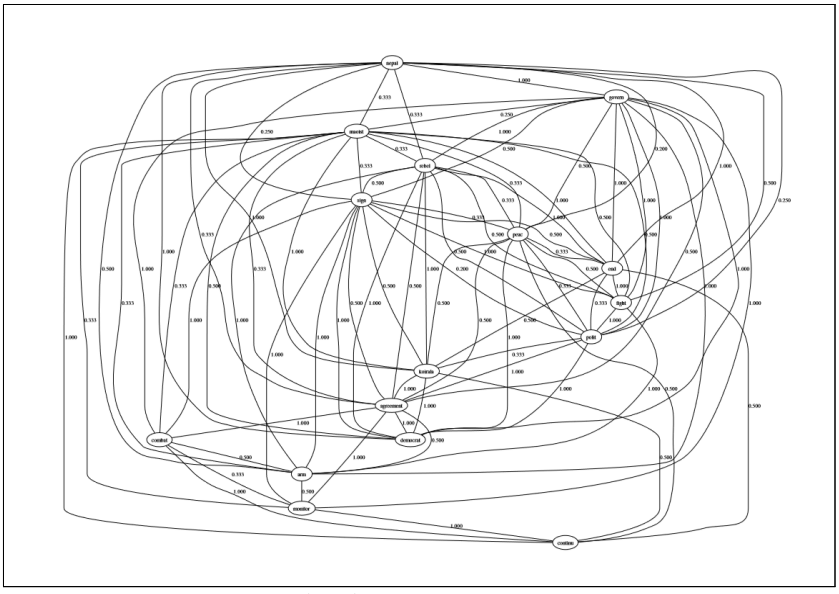
정의되지 않으며 두 단어가 동시에 출현한 문장의 수가 많아질수록 $w(u,v)$ 값이 작아져 단어 간의 유사성이 커진다.

키워드의 중심성을 계산하기 위한 측도로 사용되는 편심 $\epsilon(u)$는 아래와 같이 정의된다.

$$\epsilon(u) = \max\{d(u,v)|v \in G\}$$

여기에서 $d(u,v)$는 단어 u,v간 최단거리에서의 엣지의 거리들의 합이며 직관적으로 다른 단어 사이의 거리가 작으면 작을수록 중심이 되는 단어임을 알 수 있다. 즉, 편심의 값이 작은 단어는 문서 데이터의 다른 단어들까지의 거리가 작은 단어가 된다고 할 수 있다. 이 방법을 사용하는 경우 다음과 같은 그래프가 그려진다.

[그림 4-4] 네트워크를 이용한 키워드 추출 예시



자료: Palshikar, Girish Keshav. (2007)

두 번째 방법은 키워드 추출하는데 근접성(closeness) $C(u)$을 사용하

는 것이다. 2절에서 소개한 것처럼 중심성은

$$C(\mu) = \sum_{v \in V} d(u, v)$$

으로 정의되며, 여기서 $d(u, v)$는 u에서 v로의 최단 경로의 길이이다. 편심과 유사한 원리로 근접성 값이 작은 단어일수록 다른 단어들까지의 모든 거리의 합이 작은 단어이므로 그래프 상에서 중심에 위치하며, 중요한 단어임을 알 수 있다.

세 번째는 인접성(proximity)을 사용하여 키워드를 식별하는 방법이다. 인접성은 $prox(u, v) = \sum_{p \in V(u, v, m)} \prod_{e_i \in p} \hat{P}(e_i | a_j \; \forall j < i)$와 같이 정의된다. 여기에서 $V(u, v, m)$은 노드 u에서 노드 v로 가는 길 중 거리가 m 이하인 경로의 집합이며, e_i는 $V(u, v, m)$의 원소이다. 또한 확률의 추정값 $\hat{P}(v|u)$는 u라는 단어가 있을 때, v도 있을 확률의 추정치로 아래와 같이 표현한다.

$$\widehat{P}(v \mid u) = \frac{\sum_{L : \{u, v\} \subseteq L} \left(\frac{1}{|L| - 1}\right)}{\sum_{L : u \in L} 1}$$

$|L|$은 문장 내 노드의 수이며 가중치는 노드 u로부터 노드 v로의 연결수와 노드 u의 비율에 비례하는 값이므로 가중치가 높을수록 동시에 자주 발생한다는 의미를 갖는다. 하지만 실제로는 두 단어 간의 거리(비유사성)를 측정하는데 $\hat{P}(v|u)$값의 역수를 사용하여 것이 적절하다.

그래프를 이용한 키워드 추출 방법 중 다른 한 가지는 순위 모형인 TextRank이다. 이 모형은 키워드와 단어 추출에 대한 비 지도학습(unsupervised method)을 사용한다. 앞의 중심성 측도와 마찬가지로 TextRank 방법에서도 각 단어들마다 점수를 할당하여 해당 점수를 통해

키워드를 도출하는 방법을 사용한다.

각 단어에 대한 점수는 반복적인 추정법을 통해 계산된다. 그래프 $G=(V,E)$에서 하나의 노드가 $In(V_i) \rightarrow V_i \rightarrow Out(V_i)$로 이루어져 있다고 한다면, V_i의 점수는 아래의 식을 반복적으로 계산하여 구할 수 있다.

$$S(V_i) = (1-d) + d \sum_{V_j \in In(V_i)} \frac{1}{Out(V_j)} S(V_j)$$

여기서 d는 0과 1 사이 값을 가지며 알고리즘의 수행 도중 $S^{k+1}(V_i) - S^k(V_i)$가 임계값 보다 작아지면 알고리즘을 중단하고 그 값을 해당 노드의 점수로 표현한다.

그래프에서 두 노드 V_i, V_j 사이의 엣지에 할당된 가중치 w_{ij}가 있다면 다른 방식으로 단어의 점수를 추정할 수 있다. 가중치가 있을 때의 각 꼭짓점의 점수는 다음의 식을 반복적으로 계산하여 추정하며

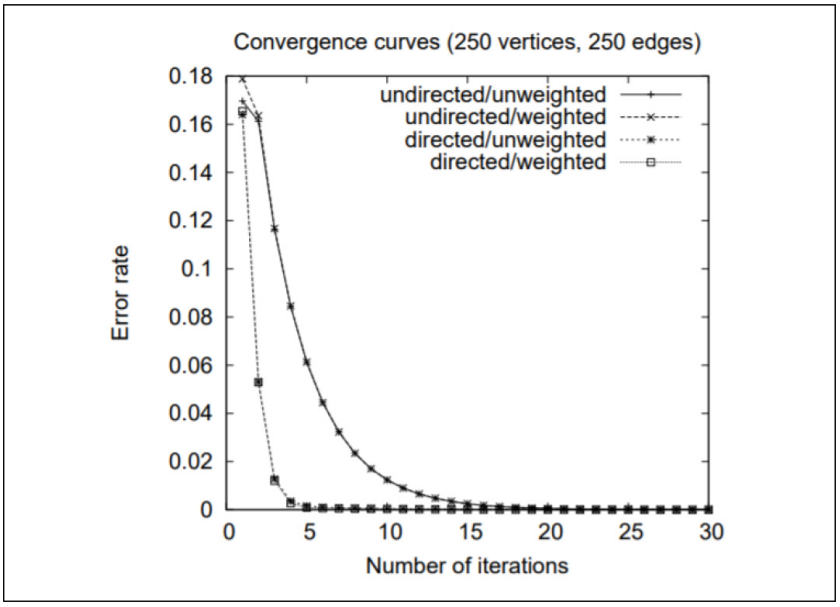
$$WS(V_i) = (1-d) + d \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

앞의 알고리즘과 마찬가지로 $WS^{k+1}(V_i) - WS^k(V_i)$가 임계값보다 작아지면 알고리즘을 중단한다.

점수를 추정하는 반복적인 계산법은 비 방향성 그래프에서도 똑같이 적용할 수 있다. 방향성과 가중치의 유무에 따라서 수치적 방법의 반복수와 오차를 시각화한 결과는 아래 그림과 같다. 그림에서 확인할 수 있듯 비 방향성 그래프의 수렴속도는 방향성 그래프의 수렴보다 느리며 가중치의 유무에 따라서 수렴속도와 수렴의 추세는 서로 비슷함을 알 수 있다.

정리하면 문장 데이터에서 단어와 단어 사이의 관계를 그래피컬 모형을 통해 추정한 뒤 수렴할 때까지 알고리즘을 반복하고, 모든 단어의 점수를 내림차순으로 정렬하면 키워드를 도출할 수 있다.

[그림 4-5] 방향성과 가중치 유무에 따른 반복수 및 오차 결과



자료: Mihalcea, Rada, and Paul Tarau. (2004)

그래프를 이용한 키워드 추출 중 마지막 방법론은 Rapid Automatic Keyword Extraction (RAKE)이다. RAKE는 개별 문서들에 대하여 언어에 무관한 (language-independent) 비 지도학습 키워드 추출 알고리즘이다. 위에서 제시한 TextRank보다 RAKE가 더 속도가 빠르고 재현율 (recall)이 높은 방법론임이 경험적으로 알려져 있다.

RAKE의 문장 분할 알고리즘은 다음과 같다. 먼저 알고리즘의 입력값은 불용어와 사용자 정의 단어 목록, 구분자 집합으로 구성된다. 다음으로 문서를 구분자와 불용어를 기준으로 문장을 단어 단위로 분석하여 후보 키워드의 배열을 생성하는 방식으로 작동한다.

문장이 RAKE의 분할 알고리즘을 통해 단어 단위로 분석된 뒤 단어의

동시 발생 빈도 행렬을 생성한다. 이 때, 특정 단어의 배열이 다른 문장에서도 똑같이 등장한다면 원본 문서에서 사이의 불용어 혹은 구분자 등을 반영하여 새로운 단어 배열을 생성한다. 이 과정을 거쳐 생성된 단어 배열로부터 마지막으로 동시 발생 빈도 행렬을 생성하여 각 단어마다 점수를 할당한다. 각 점수 할당 방법은 다음과 같다.

- 단어 빈도(word frequency) : $freq(w)$
- 단어 연결 수(word degree) : $\deg(w)$
- 단어 빈도와 연결 수의 비 : $\deg(w)/freq(w)$

RAKE는 불용어가 포함되어 있는 단어일지라도 문장 내에서 여러 번 반복된 단어이면 다시 키워드로 추출한다는 장점이 있다. 예를 들면, '$axis\ of\ evil$'의 단어에는 'of'의 불용어가 포함되어 있지만 문서 내에서 여러 번 등장했기 때문에 키워드로 도출될 수 있다.

단어들에 점수가 매겨진 후, 점수가 상위권인 단어들 T개를 문서의 키워드로 선정한다. T는 단어 수의 1/3으로 계산한다. 전체적인 과정은 다음의 예로 설명할 수 있다.

1) 다음 예제문에 입력값으로 사용자 정의 단어가 할당된다.

Compatibility of systems of linear constraints over the set of natural numbers

Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types.

Manually assigned keywords:
linear constraints, set of natural numbers, linear Diophantine equations, strict inequations, nonstrict inequations, upper bounds, minimal generating sets

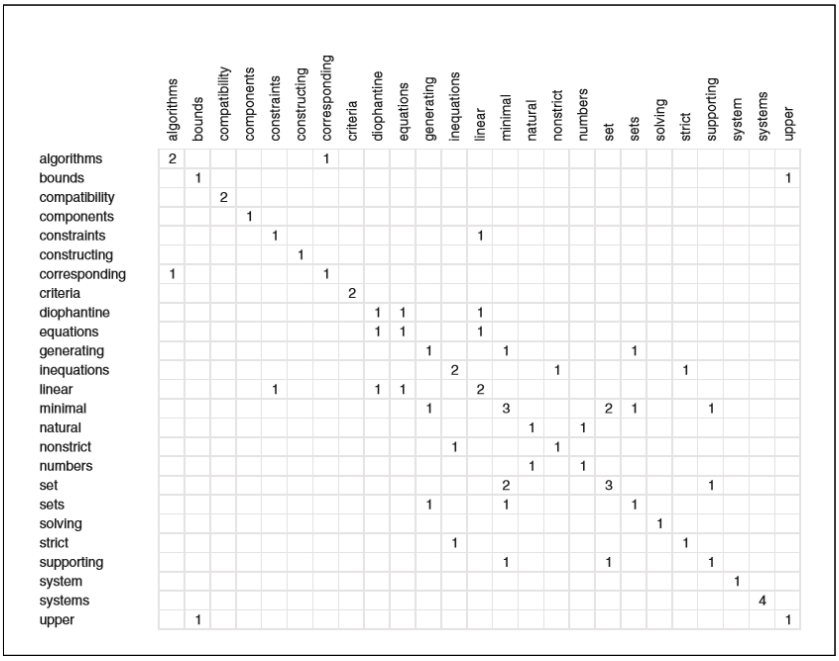
자료: Rose, Stuart, et al. (2010)

2) 구분자, 불용어들에 의해 구분된 단어 배열은 다음과 같다.

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

자료: Rose, Stuart, et al. (2010)

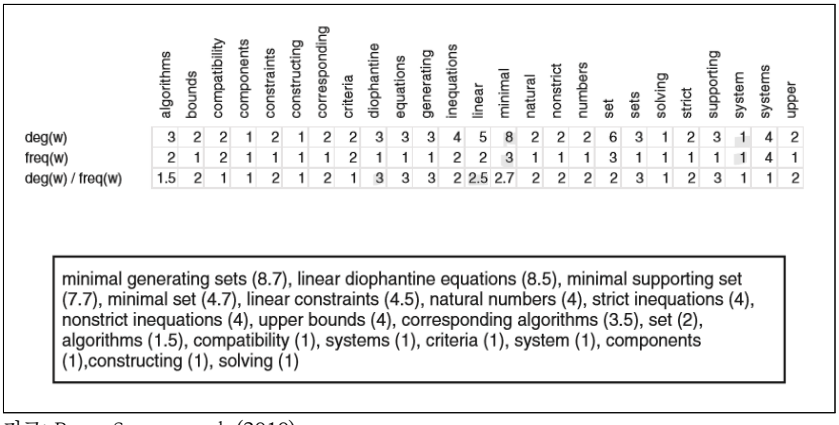
3) 도출된 단어에 대한 동시 발생 행렬을 작성한다.



	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
algorithms	2						1																		
bounds		1																							1
compatibility			2																						
components				1																					
constraints					1								1												
constructing						1																			
corresponding	1						1																		
criteria								2																	
diophantine									1	1			1												
equations									1	1			1												
generating											1			1					1						
inequations												2				1					1				
linear					1				1	1			2												
minimal											1			3				2	1			1			
natural															1		1								
nonstrict												1				1									
numbers															1		1								
set														2				3				1			
sets											1			1					1						
solving																				1					
strict												1									1				
supporting														1				1				1			
system																							1		
systems																								4	
upper		1																							1

자료: Rose, Stuart, et al. (2010)

4) 단어들에 대한 점수 계산하고 T개의 단어를 도출한다.



	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

minimal generating sets (8.7), linear diophantine equations (8.5), minimal supporting set (7.7), minimal set (4.7), linear constraints (4.5), natural numbers (4), strict inequations (4), nonstrict inequations (4), upper bounds (4), corresponding algorithms (3.5), set (2), algorithms (1.5), compatibility (1), systems (1), criteria (1), system (1), components (1),constructing (1), solving (1)

자료: Rose, Stuart, et al. (2010)

2. 워드 임베딩을 이용한 키워드 추출

워드 임베딩은 단어를 숫자 벡터로 변형하는 방법이다. 워드 임베딩을 이용한 키워드 추출에서 첫 번째로 살펴볼 방법은 TF-IDF의 변형을 이용한 키워드 추출 기법이다. TF-IDF (Term Frequency Inverse Document Frequency) 가중치 모델은 문서 내부에 존재하는 단어의 중요도를 평가하기 위해 많은 검색엔진이 채택하고 있는 방법이다. 2절에서 소개한 방식 외에도 TF-IDF를 문서 집합 전체 범위에 적용하기 위해 기존 모델의 원리를 유지하면서 추출 범위를 고려한 6가지 변형식이 제안되었다.

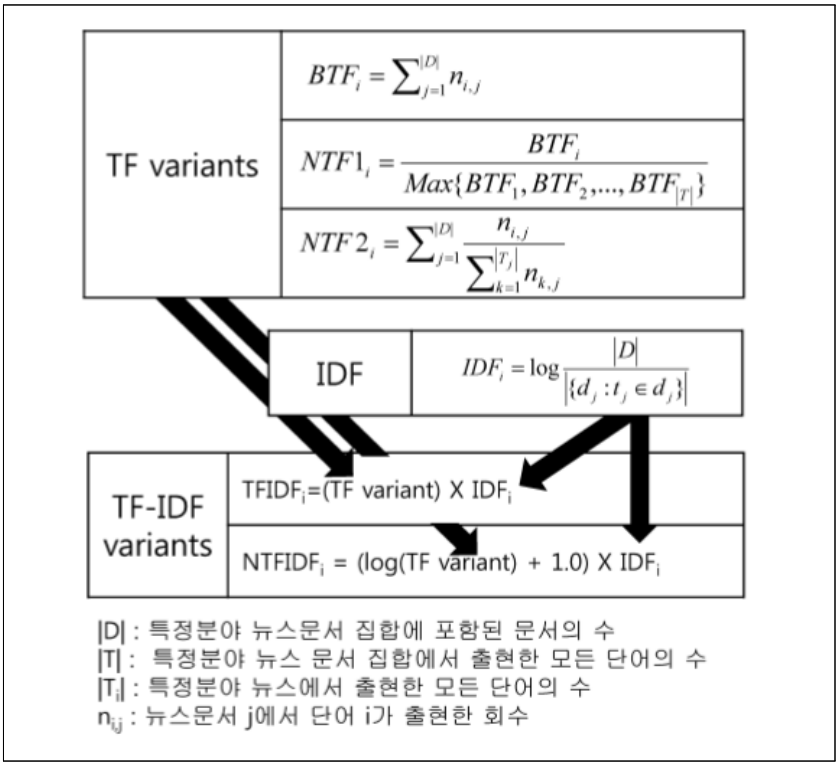
TF-IDF를 활용한 키워드 추출 문제는 문서 데이터에서 각 분야의 주요 키워드를 찾는 것으로 정의한다. 키워드 추출 기법은 2단계로 이루어지는데, 처음으로 전체 문서 집합에 존재하는 단어를 정의된 가중치로 정렬하여 그 값이 주어진 임계값보다 큰 후보 키워드들을 골라낸다. 후보 키워드 집합에 불용어 수준의 단어가 포함될 수 있기 때문에 두 번째 단계로서 각 분야에서 얻어진 후보 단어들의 순위를 교차비교 함으로써 각 분야의 대표단어로서의 키워드 집합을 얻게 된다.

키워드 추출을 위한 TF 변형은 다음 그림과 같이 NTF1, NTF2방법을 제안했다. 전체 데이터에서 특정 단어가 출현한 횟수를 BTF(Basic Term Frequency)라고 하면 NTF1(Normalized Term Frequency 1)는 BTF값을 문서집합 내에서 출현한 모든 단어들의 BTF값 중 최대값으로 나누어 정규화한 변형이다. 그리고 NTF2는 단어의 빈도를 특정한 분야에서 발생한 빈도수로 나누어 더해주는 방식으로 정의한다.

TF-IDF 방법처럼 TF 가중치와 IDF 가중치를 곱해주는 방법 외에도 TF값에도 로그를 취하여 IDF값과 곱하는 방법인 NTIFIDF (Normalized TFIDF)을 사용할 수 있다. 특히 NTF1, NTF2와 같이 사용하는 문서 데

이터의 길이가 다를 경우 발생하는 가중치의 편차를 최소화하기 위해 NTFIDF를 사용한다.

〔그림 4-6〕 키워드 추출을 위한 TF-IDF

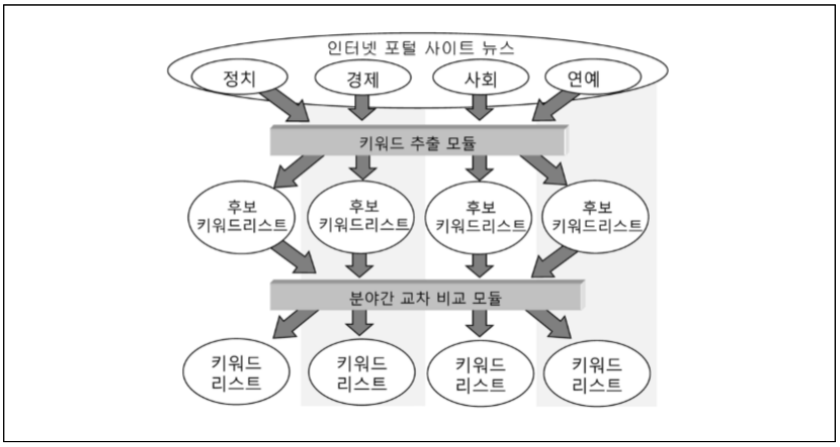


자료: 이성직, 김한준. (2009)

실제 뉴스데이터에 TF-IDF를 활용한 키워드 도출 과정은 아래 그림과 같다. 먼저 정치, 경제, 사회, 연예 분야에서 후보 키워드들을 생성하고 분야 간 교차비교 모듈이 최종 키워드 집합을 생성하게 된다. 분야 간 교차비교는 키워드 순위에 대한 표준편차 값을 계산함으로써 이루어질 수 있다. 각 분야의 뉴스문서집합으로부터 얻어진 후보 키워드집합에 대하

여, 각 분야별 순위값을 계산한 후, 특정 단어들의 순위값들의 표준편차가 주어진 임계값 이하인 경우 키워드 선정에서 제외된다.

〔그림 4-7〕 키워드 추출 시스템 및 과정



자료: 이성직, 김한준. (2009)

워드 임베딩을 이용한 키워드 추출 방법 중 두 번째는 지도학습 알고리즘을 사용한 자동 키워드 추출 방법이다. 여기에는 n-grams, NP-chunk, POS Tag Patterns 등이 있다. 이 방법들은 단어의 빈도, 문서의 빈도, 용어에 할당된 형태소 태그와 같은 특징들이 사용된다. 모형에 언어적 지식을 추가함으로써 경험적으로 전문가들이 할당했던 키워드보다 더 나은 결과를 얻을 수 있다. 아래는 지도학습 알고리즘을 사용한 세 가지 방법에 대해서 소개한다.

□ n-grams

n-grams은 n개의 연속적인 단어 나열을 말하는 것이다. 문서 데이터를 n개의 단어 뭉치 단위로 끊어서 이를 하나의 단어로 간주한다. N-gram의 방법은 모든 유니그램 (unigram), 바이그램 (bigram), 트라이그램 (trigram) 을 포함하는 일반적인 방법이다.

□ NP chunks

수동으로 할당된 키워드를 검사할 때, 대부분은 형용사가 있는 명사나 명사 구문으로 판명된다. 특정 언어적 속성을 가진 키워드들을 더 잘 포착하기 위해 수동으로 할당된 키워드를 하나의 명사구로 할당하여 사용하는 방법이다.

□ POS tag Patterns

POS tagging이란 문서를 형태소 단위로 나누고 각 형태소에 품사 정보를 부착하는 작업을 의미한다. 이 접근 방식은 여전히 특정 구문적인 속성을 갖는 키워드의 개념을 포착하지만, 훈련 데이터의 경험적 증거를 기반으로 한다.

워드 임베딩을 이용한 키워드 추출 방법 중 마지막은 개별 문서에 대해서 단어의 동시 발생 정보를 사용하는 알고리즘이다. 이는 군집화와 동시 발생 분포의 편향을 이용해서 개별 문서 내에서 키워드를 추출하는 방법이다. 편향은 χ^2 통계량을 이용해서 측정하고, 수정된 χ^2 통계량과 군집화는 더 나은 성능을 얻기 위해 사용된다.

단어 동시 발생 빈도를 이용하는 방법의 알고리즘은 전처리, 빈출 단어 선택, 빈출 단어 군집화, 기대 확률 계산, χ^2 통계량 계산, 키워드 산출 순

서로 이루어져 있다. 전처리 단계에서는 다른 알고리즘과 유사하게 불용어를 삭제하고 상위 30% 빈출 단어를 선택한다. 빈출 단어 군집화 단계에서는 동시 발생 행렬을 생성한 뒤 행렬에서 용어들을 군집화 한다. 다음 단계에서는 기대 확률 p_c를 계산한다.

$$p_c = \frac{n_c}{N}$$

여기서 n_c는 $c \in C$인 동시 발생 용어의 수이고, C는 얻어진 군집의 수이며 N은 상위 30% 빈출 단어의 개수이다. χ^2 통계량은 다음과 같은 식으로 계산될 수 있다.

$$\chi'^2(w) = \sum_{c \in C} \left\{ \frac{(freq(w,c) - n_w p_c)^2}{n_w p_c} \right\} - \max_{c \in C} \left\{ \frac{(freq(w,c) - n_w p_c)^2}{n_w p_c} \right\}$$

여기서 $freq(w,c)$는 용어 w에 대하여 $c \in C$인 동시 발생 단어의 빈도이고 n_w는 용어 w를 포함하는 문장의 모든 단어수이다. 마지막으로 χ^2통계량 값이 가장 큰 단어를 선택한다.

3. 토픽 모형을 이용한 키워드 추출

대부분의 키워드 추출연구는 주어진 문서 안에서 후보 키워드를 선택한 후, 후보 키워드가 등장한 위치와 횟수를 바탕으로 후보 키워드의 중요성을 평가하여 키워드를 추출한다. 하지만 기존의 방법들은 주어진 문서에서 나타나는 구문만을 후보 키워드로 선택하기 때문에 문서에서 나타나지 않는 구문은 선택하지 못하는 한계가 있다. 또한, 문서에서 나타나지 않는 후보 키워드를 평가하기에도 적합하지 않다.

잠재키워드는 문서에서는 나타나지 않지만 중요한 개념이나 내용을 함축하고 있어 문서 요약 및 정보 검색에 중요한 역할을 할 수 있다. 잠재키

워드를 찾기 위해, 문서에서 등장하지 않는 구문을 이웃 문서를 통해 후보 키워드로 선택하고 이를 평가하는 방법이 있다. 후보 키워드의 선택 및 평가에는 LDA를 기반으로 생성된 주제 모델을 활용한다.

LDA 모델을 이용한 잠재 키워드를 추출하기 위해서는 다음의 방법을 사용한다.

1) 불용어 제거 등의 전처리 후 문서 집합에서 자주 등장하는 공통어를 제거한다.

2) LDA를 기반으로 주제들의 확률 분포 θ와 각 주제에 속하는 단어들의 확률 분포 z를 추론한다.

3) 후보키워드를 추출하기 위해 이웃 문서를 선별한다. 이웃 문서를 선별하는 방법으로 두 가지가 있다.

3-1) 주어진 문서와 단어 벡터가 유사한 문서들, 즉 서로 비슷한 문서를 이웃 문서로 선별하는 방법이다.

3-2) 주어진 문서와 주제 벡터가 유사한 문서(주제 모델 θ를 이용)를 이웃 문서로 선별한다.

주어진 문서 d_g에 대한 문서 집합 D에 포함되는 이웃 후보 문서의 벡터 간 유사도는 다음의 코사인 유사도 식을 통해 구할 수 있다.

$$sim(\overrightarrow{d_g}, \overrightarrow{d_n}) = \frac{\overrightarrow{d_g} \cdot \overrightarrow{d_n}}{\|\overrightarrow{d_g}\| \times \|\overrightarrow{d_n}\|}, \quad d_n \in D$$

4) 이웃 문서의 키워드를 후보 키워드로 선택한다.

5) 후보 키워드 w를 구성하는 단어 w_i별로 중요도를 평가하여 가장 중요도가 높은 c개의 후보 키워드를 주어진 문서의 잠재 키워드로 선정한다. 주어진 문서 d_g에서 w_i가 등장할 확률을 LDA에서 추론된 주제분포 θ와 단어분포 z를 이용하여 계산한다. w_i는 여러 가지 주제 t에서 중복하여 등장할 수 있다. 각 주제는 서로 독립이며, LDA는 d_g와 w_i가 독립임을 가정한다.

$$P(w_i|d_g) = \sum_t P(w_i|t, d_g)P(t|d_g) = \sum_t P(w_i|t)P(t|d_g)$$

후보 키워드 w의 중요도는 위 식의 기하 평균으로 산출된다. 즉, 후보 키워드 w를 구성하는 단어 w_i가 주어진 문서에서 등장할 확률이 높으면 중요한 단어임을 의미한다.

$$P(w|d_g) = \sqrt[k]{\prod_i^k P(w_i|d_g)}$$

유사도가 높은 이웃 문서에서 선택된 후보 키워드가 더 중요할 수 있으므로, 위에서 구한 유사도를 가중치로 활용하는 것도 가능하다.

$$P(w|d_g) = \sqrt[k]{\prod_i^k P(w_i|d_g)}\ sim(\overrightarrow{d_g}, \overrightarrow{d_n})$$

이렇게 주어진 문서와 유사한 문서의 키워드를 후보 키워드로 선택하고, 후보 키워드를 구성하는 개별 단어들의 등장 확률을 이용하여 후보 키워드의 중요도를 평가하는 방법으로 잠재 키워드를 추출할 수 있다.

기존의 단어 선택에 있어 3가지 지표가 있다. 함수 $f(t_k, c_i)$는 t_k항의

단어 선택 점수를 나타내며 특정 범주 c_i에 지엽적으로 지정된다. 전역적인 t_k의 값을 평가하기 위해서 $f_{\max}(t_k) = \max_{i=1}^{|c|} f(t_k, c_i)$ 값을 이용한다.

첫 번째 지표로는 정보 이득(IG)으로, 이것은 해당 문서에 해당 단어가 존재하는지의 여부를 파악하여 엔트로피의 감소를 측정하는 지표이다.

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \overline{c_i}\}} \sum_{t \in \{t_k, \overline{t_k}\}} P(t,c) \log \frac{P(t,c)}{P(t)P(c)}$$

두 번째 지표로는 카이제곱 통계량(CHI)이 있다. 이는 통계적으로 두 확률변수의 독립성을 측정하는 통계량이다.

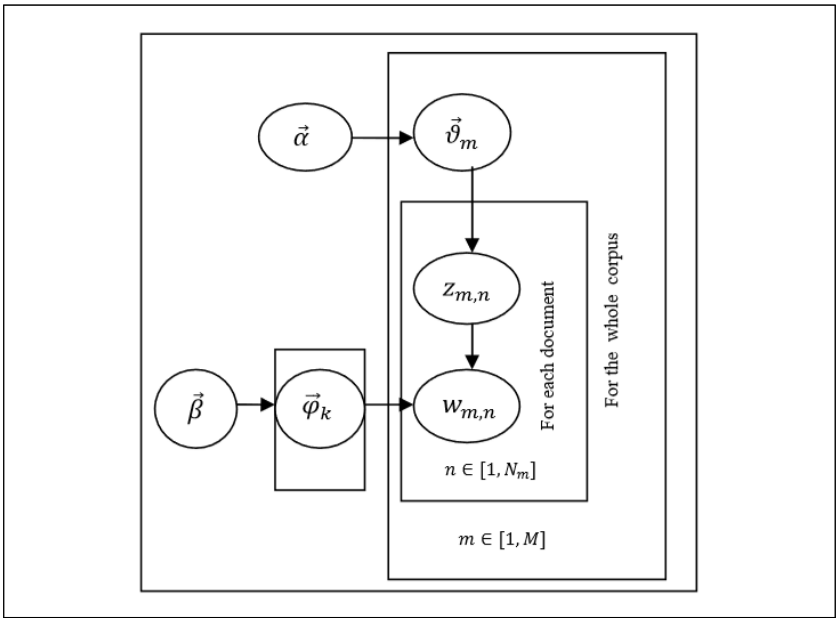
$$x^2(t_k, c_i) = N \times \frac{[P(t_k, c_i)P(\overline{t_k}, \overline{c_i}) - P(\overline{t_k}, c_i)P(t_k, \overline{c_i})]^2}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})}$$

세 번째 지표는 문서 빈도(DF)이다. 이는 자주 나오지 않는 단어가 신뢰도가 떨어지고, 범주 예측에 있어 효과적이지 않다는 가정에 기초한다. 이것은 각 용어가 등장하는 문서의 수를 계산하여 개수가 가장 많은 용어를 선택하는 것이다. 이는 간단함에도 불구하고, 키워드 수가 너무 적지 않다면 IG, CHI와 비슷한 성능을 보여준다.

$$DF(t_k, c_i) = P(t_k, c_i)$$

LDA에서 $w_m = \{w_{m,n}\}_{n=1}^{N_m}$는 디리클레 분포에서의 토픽 v_m뽑음으로써 발생한다. N_m은 문서 m의 길이 이며, v_m은 문서 m의 토픽 분포이다. 그리고 $z_{m,n}$이 문서 m에서 n번째 토픽 인덱스라고 할 때, $Muti(v_m)$분포에서의 특정 토픽 $z_{m,n}$을 표본 추출함으로써 [m,n]의 위치에 할당한다. 마지막으로, 특정단어 $w_{m,n}$은 φ_k가 토픽 k의 단어 분포일 때, 다항분포 $Muti(\varphi_{z_{m,n}})$에서 표본 추출함으로써 [m,n]의 위치에 생성된다.

[그림 4-8] LDA 생성 절차



자료: Tasci, Serafettin, and Tunga Gungor. (2009)

위 그림에서의 생성 절차에 따라 전체 데이터 수집의 가능성을 다음과 같이 쓸 수 있다.

$$P(W|\alpha,\beta) = \prod_{m=1}^{M} \iint P(v_m|\alpha) P(\Phi|\beta) \prod_{n=1}^{N_m} P(w_{m,n}|v_m,\Phi) d\Phi dv_m$$

여기서 Φ는 중심 용어 행렬이고, α, β는 디리클레 분포의 모수이다. 위의 식은 계산하기 힘들기에, 변동 방법(Variational method)와 깁스 샘플링의 방법(Gibbs sampling)을 사용하여 근사한다.

먼저 변동 근사의 로그 가능도 함수는 아래와 같다.

$$\log p(W|\alpha,\beta) = \log \int_{v} \sum_{z} p(w|z,\beta) p(z|v) p(v;\alpha) q(v,z;\gamma,\psi) dv$$

여기서 $q(v,z;\gamma,\psi)$는 $q(v,z;\gamma,\psi)=q(v;\gamma)\prod_{n}q(z_n;\varphi_n)$로 모수 φ_n, γ가 사용된 범주화된 변동 분포이다. 그리고 $q(v;r)$은 $Dir(\gamma)$ 이고, $q(z_n;\varphi_n)$는 $Muti(\varphi_n)$이다.

깁스 샘플링 근사는 아래와 같다.

$$p(z_i=k|z_{i'},w)=\frac{n_{k,i}^{(t)}+\beta_t}{[\sum_{v=1}^{V}n_k^{(v)}+\beta_v]-1}\frac{n_{m,i}^{(k)}+\alpha_k}{[\sum_{j=1}^{K}n_m^{(j)}+\alpha_j]-1}$$

여기서 $n_{k,i}^{(t)}$는 단어 t를 제외한 t가 토픽 k에 할당된 수이고, $\sum_{v=1}^{V}n_k^{(v)}$는 해당 단어를 제외했을 때 그 단어가 토픽 k 에 할당된 단어의 전체 수이며, $n_{m,i}^{(k)}$는 문서 m에서 토픽 k에 해당 단어를 제외한 해당 단어의 수이며, $\sum_{j=1}^{K}n_m^{(j)}$는 해당 단어를 제외 했을 때의 문서 m에서 전체 단어의 수이다.

마지막으로 φ_k는 다음과 같이 계산한다.

$$\psi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{v=1}^{V}n_k^{(v)}+\beta_v}$$

위의 단계를 완료한 후, 용어 토픽 행렬을 사용하여 단어들의 엔트로피를 계산할 수 있다.

LDA와 같이 주제 기반 문서 요약은 일반 사용자가 문서 수집을 빠르게 이해하고 통찰력을 얻을 수 있도록 도와준다. 그러나 LDA를 기반한 결과는 사람이 이해한 것과 다를 수 있기 때문에 LDA 결과를 향상시키기 위해서 주제와 키워드의 순위를 부여하는 방법들이 개발되었다.

문서 집합을 $D=\{d_1,d_2,...,d_M\}$이라 하면, m 번째 문서 d_m는

$W_m = \{w_{m,1}, w_{m,2}, ..., w_{m,N_m}\}$의 단어 시퀀스로 구성되어 있다. 그리고, $V = \{v_1, v_2, ..., v_v\}$는 단어로, v는 단어의 개수이다. z는 관찰된 각 단어와 관련된 잠재된 주제를 나타내는 잠재 변수다.

$\varphi_k = (\varphi_{k,1}, \varphi_{k,2}, ..., \varphi_{k,v})^T \in R^v$이고 여기서 $\varphi_{k,i} = P(w = v_i | z = k)$이다. 그러므로 토픽단어의 분포의 모수는 $\Phi = (\varphi_1, \varphi_2, ..., \varphi_K)^T \in R^{K \times V}$으로 표현할 수 있고, K는 토픽의 개수이다.

정보 검색(IR)에서처럼 단어에 가중치를 재부여하는 기법을 통해 LDA 기반의 두 가지 TF-IDF 점수를 사용하였다.

$$KR_1 = \frac{\hat{\varphi}_{k,i}}{\sum_{k=1}^{K} \hat{\varphi}_{k,i}}, \quad KR_2 = \hat{\varphi}_{k,i} \log \frac{\hat{\varphi}_{k,i}}{\left(\prod_{k=1}^{K} \hat{\varphi}_{k,i} \right)^{1/K}}$$

여기서 가중치 $\hat{\varphi}_{k,i}$는 단어의 빈도에 따라 LDA에 의해 생성된다.

토픽 모델링에서 토픽 선정에 중요하게 고려되는 기준이 분석자마다 다르기 때문에 토픽의 순위를 재조정하는 것이 필요하다. 토픽 순위를 재조정하는 방법은 다음의 4가지 방법이 있다.

첫 번째는 토픽 순위 재조정 방법은 말뭉치 내용의 상당 부분을 다루는 주제가 적은 내용을 다루는 주제보다 더 중요하다고 가정한다. 따라서 모든 문서에 나타나는 주제는 내용의 범위와 분산이 상당히 크지만 일반적이기에 흥미롭다고 간주한다. 즉, 이러한 주제에는 낮은 순위를 매긴다. 결과적으로 주요 토픽 순위 지표는 내용 범위와 토픽 분산의 조합으로 나타낼 수 있다.

두 번째는 라플라스 점수(Laplacian Score)이다. 한 토픽의 라플라스 점수는 다른 범주의 문서를 구분하는데 주요 초점을 둔다. 즉, 차별성이 높은 주제에 높은 순위를 배정한다.

세 번째 방법은 두 토픽에 대해 상호 정보를 기반으로 하는 토픽 재순

위 방법(Pairwise Mutual Information)이다. 이 지표는 두 항목이 공유하는 정보를 계산한다. 이 지표를 사용하여 두 항목 간의 쌍으로 구성된 상호 정보 양을 측정하여 항목의 순위를 지정할 수 있다.

마지막으로 토픽의 다양성을 극대화하고 중복을 최소화하기 위한 방법(Topic Similarity)이다. 위의 방법들이 토픽-문서의 관계를 사용한 반면, 이 방법은 토픽-단어 의 분포를 이용하여 토픽의 유사성을 계산한다.

위와 같은 여러 방법들을 사용하여 토픽의 순위를 재조정하여 결과를 향상 시킬 수 있다.

제4절 키워드 추출 방법의 고도화 방안 탐색

위에서 제시한 키워드 추출 방법들을 고도화하기 위한 방안으로 다음과 같은 방법이 있다.

1) 워드 임베딩에서 변수 선택 방법과 결합하여 필요한 변수를 선택하고 다양한 제약조건을 이용해서 해석력을 높일 수 있다.

2) 키워드 추출 모형에 감성 분류 모형을 결합하여 단어 혹은 문서들의 감성정보를 이용해서 더 정밀한 모형을 생성할 수 있다.

3) 새로운 데이터가 유입되더라도 학습을 진행할 수 있도록 온라인 학습방법을 적용할 수 있다.

이 절에서는 위 3가지 방법들에 대한 다양한 예를 살펴보도록 한다.

1. 워드 임베딩 정보의 정제

이미지나 텍스트 데이터의 차원축소를 위해서 주성분 분석 (Principal component analysis) 를 사용하는 것은 매우 높은 계산량을 필요로 한다. 그래서 차원축소방안에 대한 대안으로 Random projection을 제안되었다.

$$X_{RP} = XR$$

여기서 X는 $n \times d$ 차원의 데이터이고 R은 $d \times k$차원의 행렬이다. X_{RP}는 $n \times k$차원이다. d는 k보다 크다.

Random projection의 주된 아이디어는 '우리가 가진 데이터를 임의

의 저차원 행렬을 이용하여 사영시켜도 데이터들 사이의 Euclidean 거리가 보존된다.'이다. 임의의 저차원 행렬 성분은 다음과 같이 매우 단순한 분포를 가지는 방식으로 만들 수 있다.

$$r_{ij} = \sqrt{3}\begin{cases} +1 & w.p\ \frac{1}{6} \\ 0 & w.p\ \frac{2}{3} \\ -1 & w.p\ \frac{1}{6} \end{cases}$$

여기서 r_{ij}는 저차원 행렬R의 i행, j열의 원소이다.

위와 같이 워드 임베딩 방법들은 원리에 대한 이해가 직관적이지 않고 결과에 대한 해석이 불가능하다. 따라서 워드 임베딩 알고리즘에 대한 원리를 이해, 해석하고 중요한 변수들을 선택할 수 있도록 하는 모형들이 많이 제안되었다.

처음에 제안된 방법처럼 임의의 행렬을 이용해서 임베딩하기도 하지만 워드 임베딩 방법 중에서 가장 각광받고 있는 Skip-gram with negative sampling(SGNS) 의 원리를 이용해서 저차원으로 임베딩하는 방법도 있다. 어떤 두 단어(w,c)가 함께 관측되는 사건을 D 라고 하면 사건 D의 확률은 다음과 같이 정의한다.

$$P(D\,|w,c) = \sigma(\vec{w},\vec{c}) = \frac{1}{1+\exp^{-\vec{w}\,\bullet\,\vec{c}}}$$

여기서 $\vec{w}$, $\vec{c}$는 각각 단어 w, c에 대응되는 임베딩 벡터들이다.
SGNS 는 주된 아이디어는 함께 관측된 (w,c)의 $P(D\,|\,w,c)$를 높이도록 임베딩하는 것이다. SGNS 의 목적함수를 특정 두 단어(w,c)에 대한 관점에서 보면 다음과 같다.

$$\log\sigma(\vec{w},\vec{c}) + k\ \bullet\ E_{c_N \sim P_D}[\log\sigma(-\vec{w}\ \bullet\ \vec{c_N})]$$

여기서 k는 negative sample의 수, c_N은 negative sample이다. 그리고 negative sample을 추출하는 분포 $P_D(c)$는 단어가 등장하는 비율에 비례하는 분포로 설정한다.

위의 목적함수를 최대화하는 두 벡터$(\overrightarrow{w},\vec{c})$의 내적은 다음과 같이 두 단어의 점 상호 정보 (Pointwise Mutual Information:PMI) 에 대한 식으로 나타낼 수 있다.

$$\overrightarrow{w} \bullet \vec{c} = PMI(w,c) - \log k$$

따라서 SGNS의 최적의 해들은 PMI 과 관련되어있다고 생각할 수 있다. 그렇다면 두 단어들의 PMI 값들로 채워진 행렬M 을 분해해서 단어들의 행렬(W,C)들을 구할 수 있다. 즉, 행렬 M 의 i,j번째 원소 M_{ij}를 다음과 같이 정한다.

$$M_{ij} = W_i \bullet C_j = \overrightarrow{w_i} \bullet \vec{c} = PMI(w_i,c_j) - \log k$$

하지만 PMI 를 사용하면 다음의 두 가지 문제점이 있다.

1) 같이 등장하지 않은 단어쌍에 대해서는 $PMI(w,c) = -\infty$ 을 가진다.
2) 행렬 M 이 희소행렬이 아니게 된다.

이 두 문제점은 행렬을 분해하는데 매우 큰 계산량을 요구하게 된다. 따라서 문제점을 해결하고자 같이 등장하지 않은 단어 쌍에 대해서 $PMI(w,c) = 0$으로 설정하여 희소행렬을 생성한다. 이렇게 했을 경우에도 문제가 생기는데 그것은 적게라도 함께 등장하는 단어들의 PMI 가 같이 등장하지 않은 단어들의 PMI 보다 작아지는 문제가 발생한다.

새로운 대안으로 $Positive\,PMI(PPMI)$를 정의해서 행렬 M_{PPMI}를

생성한다.

$$PPMI(w,c) = \max(PMI(w,c), 0)$$

위와 같이 $PPMI$를 사용해도 문제가 되지 않는 이유는 워드 임베딩에서의 주된 관심사는 함께 많이 등장하는 단어들이다. 즉, (보건, 음악)보다 (보건, 복지)에 더 관심이 많다는 것이다. 여기에 $\log k$ 임계값을 주어 Shifted-$PPMI$($SPPMI$)을 사용해서 적절하게 행렬M을 조정할 수 있다. 최종적인 모형은 $SPPMI$를 이용해서 행렬M_{SPPMI}를 생성한다.

$$SPPMI = \max(PMI(w,c) - \log k, 0)$$

위의 방법으로 생성한 희소행렬을 단어행렬(W, C)들로 분해하기 위해서 대표적인 행렬 분해방법인 특이값 분해 (Singular Vector Decomposit ion:SVD) 를 사용한다. $M_{SPPMI} = U_d \Sigma_d V_d^T$와 같이 분해된다고 가정했을 경우(여기서, d는 임베딩할 차원), 단어행렬 W, C 는 다음과 같이 구해진다.

$$W = U_d \cdot \sqrt{\Sigma_d}, C = V_d \cdot \sqrt{\Sigma_d}$$

위의 방법은 신경망 모형을 이용하지 않고 단순한 행렬분해방법(SVD)를 이용해서 단어들을 임베딩한다.

이 방법은 신경망 모형에 비해 두 가지 장점이 존재한다.

1) 신경망 모형에서 필요한 모수 조율(parameter tuning)과정을 생략할 수 있다.
2) 동시에 발생하는 횟수가 잘 집계된 (count-aggregated) 데이터에서 매우 쉽게 적용할 수 있다.

위의 방법처럼 희소행렬을 분해하는 동시에 주요한 변수를 선택하면서

해석력도 같이 가질 수 있는 방법도 연구되었다. 임베딩 벡터들의 길이(l_2-norm)를 1로 고정하고 비음수의 성질을 통해 해석력을 높인다. 목적함수를 다음과 같이 변형한다.

$$loss = \sum_{i}^{m} \left(\| M_{i,:} - W_{i,:} \times C \|^2 + \lambda \| W_{i,:} \|_1 \right)$$

$$s.t \qquad C_{i,:} C_{i,:}^T \leq 1, 1 \leq i \leq d$$
$$W_{i,j} \geq 0, \, 1 \leq i \leq m, \, 1 \leq j \leq d$$

여기서 행렬M은 동시출현행렬로부터 유도된 입력데이터를 사용하고 위에서 사용한 M_{SPPMI}를 사용해도 무관하다.

위의 목적함수에 따르면 행렬M을 $W \times C$로 분해하면서 W의 각 행에 통계학에서의 대표적인 변수선택 방법인 l_1 정규화 방법(Lasso)을 적용했다. 그리고 비음수 제약조건에 의해서 변수들의 해석력을 부여한다. 위의 방법이 특이값 분해 방법에 비해 축소차원(d)가 커질수록 희소성 정도가 커지는 것이 실험을 통해 확인되었다.

Word2Vec 방법은 자연어처리 분야에서 매우 뛰어난 성능을 가진다. 하지만 임베딩 된 변수들을 해석하는 측면에서는 분명히 한계가 존재한다. 그래서 위와 같이 워드 임베딩 방법인 Continuous Bag of Words(CBOW)와 중요한 변수들을 선택하기 위해서 l_1 정규화 방법(Lasso)을 결합한 모형이 제안되었다. 즉, 목적함수를 다음과 같이 정의한다.

$$L_{scbow} = L_{cbow} - \lambda \sum_{w \in W} \| \overrightarrow{w} \|_1$$

하지만 위의 방법에 SGD 방법을 이용해서 최적의 해를 찾는 것은 적절하지 못하다. 그래서 RDA라는 알고리즘을 이용해서 해를 찾는다.

subgradient들의 합 $\bar{g}_{\overrightarrow{w_i}} = \sum_{i=1}^{t} g_{\overrightarrow{w_i}}^{t}$으로 표현했고 여기서 $g_{\overrightarrow{w_i}}^{t}$는 t시점의 $\overrightarrow{w_i}$의 subgradient이다. RDA 주된 아이디어는 $\bar{g}_{\overrightarrow{w_{i,j}}}$ 이 임계값 이하가 되면 $\overrightarrow{w}_{i,j} = 0$이 되므로 가중치를 없애는 것이다. RDA의 전반적인 과정은 다음과 같다.

〔그림 4-9〕 RDA 알고리즘 절차

Algorithm 1 RDA algorithm for Sparse CBOW

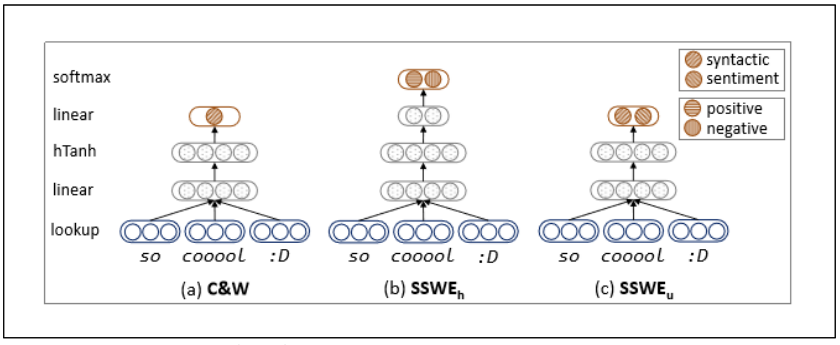
procedure SPARSECBOW($\mathcal{C}$)
  Initialize: $\vec{w}, \forall w \in W,\ \vec{c}, \forall c \in C,\ \bar{g}_{\vec{w}}^{0} = \vec{0}, \forall w \in W$
  for $i = 1, 2, 3, \ldots$ **do**
    $t \leftarrow$ update time of word w_i
    $\vec{h}_i = \frac{1}{2l} \sum_{\substack{j=i-l \\ j \neq i}}^{i+l} \vec{c}_j$
    $g_{\vec{w}_i}^{t} = \left[\mathbb{1}_h(w_i) - \sigma(\vec{w}_i^{t} \cdot \vec{h}_i)\right] \vec{h}_i$
    $\bar{g}_{\vec{w}_i}^{t} = \frac{t-1}{t} \bar{g}_{\vec{w}_i}^{t-1} + \frac{1}{t} g_{\vec{w}_i}^{t}$
    Update $\vec{w}_i$ element-wise according to
    $\vec{w}_{ij}^{t+1} = \begin{cases} 0 & \text{if } |\bar{g}_{\vec{w}_{ij}}^{t}| \leq \frac{\lambda}{\#(w_i)}, \\ \eta t \left(\bar{g}_{\vec{w}_{ij}}^{t} - \frac{\lambda}{\#(w_i)} \mathrm{sgn}(\vec{w}_{ij}^{t})\right) & \text{otherwise,} \end{cases}$
    where, $j = 1, 2, \ldots, d$
    for $k = -l, \ldots, -1, 1, \ldots, l$ **do**
      update $\vec{c}_{i+k}$ according to
      $\vec{c}_{i+k} := \vec{c}_{i+k} + \frac{\alpha}{2l} \left[\mathbb{1}_{h_t}(w_i) - \sigma(\vec{w}_i^{t} \cdot \vec{h}_i)\right] \vec{w}_i^{t}$
    end for
  end for
end procedure

자료: Sun, Fei, et al. (2016)

2. 감성 분류 정보의 이용

기존의 워드 임베딩 모형(C&W, Word2Vec과 GloVe 등)은 함께 출현하는 단어들을 이용해서 문맥적인 의미만을 고려했다. 하지만 이러한 모형들에 단어나 문서의 감성정보를 이용하고자하는 모형들이 제안되었다. 그 중에서 기존의 C&W모형과 GloVe모형에 감성정보를 이용한 모형을 살펴보았다.

[그림 4-10] (a) C&W 모형, (b) $SSWE_h$ 모형 그리고 (c) $SSWE_u$



자료: Tang, Duyu, et al. (2014)

단어들의 문장에서의 구조적인 역할에 기반하여 단어를 임베딩하는 방법으로 C&W모형이 제안되었다. C&W모형은 학습하기 위해 문장에서 학습할 단어를 일부러 임의의 단어로 교체한다. 예를 들어 'cat chills on a mat'라는 문장 t에서 'cat chills w^r a mat'으로 변경한 문장 t^r을 생성한다. t와 t^r를 모두 이용해서 목적함수를 정의한다. 목적함수는 다음과 같다.

$$loss_{cw}(t,t^r) = \max(0, 1 - f^{cw}(t) + f^{cw}(t^r))$$

여기서, $f^{cw}(\bullet)$는 문장에 대한 score 함수이다.

C&W모형의 구조는 4개의 층($lookup \rightarrow linear \rightarrow hTahn \rightarrow linear$)으로 이루어져있다. 문장 t의 score $f^{cw}(t)$는 다음과 4개의 층을 거쳐 구해진다. 먼저, $lookup$층에서 단어들을 임베딩하여 문장 t에 대한 $lookup-table$ L_t를 생성한다. 이후의 과정은 다음과 같다.

$$hTanh(x) = \begin{cases} -1 & \text{if } x \leq -1 \\ x & \text{if } -1 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

$$a = hTanh(h_1 L_t + b_1)$$

$$f^{cw}(t) = h_2(a) + b_2$$

여기서, h_1, h_2, a_1, a_2는 각 신경망 모형에서의 가중치이다.

C&W모형의 신경망 구조와 목적함수를 수정하여 감성정보를 학습하기 위한 신경망 모형($SSWE_h$)이 있다. $SSWE_h$는 C&W모형의 최상단층에 $softmax$층을 추가하여 감성라벨 점수를 계산한다. $SSWE_h$ 목적함수는 다음과 같다.

$$loss_h(t) = -\sum_{k=0,1} f_k^g(t) \log(f_k^h(t))$$

여기서 $f_k^g(t)$는 문장t에 대한 실제 감성라벨이다. $SSWE_h$의 신경망 구조는 그림(8)-(b)를 통해 확인할 수 있다.

$SSWE_h$모형의 $softmax$층을 제외하고 감성라벨 점수의 제약조건을 좀 더 완화한 임베딩 모형($SSWE_r$)이 제안되었고 $SSWE_r$의 목적함수는 다음과 같다.

$$loss_r(t) = \max(0, 1-\delta_s(t)f_0^r(t)+\delta_s(t)f_1^r(t))$$

$$\delta_s(t) = \begin{cases} 1, & \text{if } f^g(t) = [1,0] \\ -1, & \text{if } f^g(t) = [0,1] \end{cases}$$

최종적으로 C&W모형과 $SSWE_r$모형을 결합하여 단어의 구조적인 역할과 감성정보를 모두 이용한 임베딩 모형($SSWE_u$)으로 발전시켰다. $SSWE_u$모형은 문장의 구조적인 점수와 감성점수를 2가지 점수(f_0^u, f_1^u)를 출력한다. $SSWE_u$모형의 목적함수는 다음과 같다.

$$loss_u(t,t^r) = \alpha \bullet loss_{cw}(t,t^r)+(1-\alpha) \bullet loss_{us}(t,t^r)$$

여기서 α는 $loss_{cw}$ 목적함수와 $loss_{us}$ 목적함수에 대한 가중치이고 $loss_{us}(t,t^r)$는 다음과 같다.

$$loss_{us}(t) = \max(0, 1-\delta_s(t)f_1^u(t)+\delta_s(t)f_1^u(t^r))$$

위와 비슷한 방법으로 단어를 포함한 문장의 주변 정보와 문서의 감성정보를 동시에 이용하여 학습하는 방법도 존재한다. 이때, 문서(d)의 감성 정보를 이용하기 위해 문서의 대표벡터($\widetilde{D}_i$)를 문서에 포함된 단어들의 임베딩 벡터들의 합으로 정의한다. i번째 문서 대표벡터는 다음과 같다.

$$\widetilde{D}_i = \sum_{j=1}^{T_i} o_{i,j}$$

여기서 T_i는 D_i 문서에 포함된 단어들의 수이고 $o_{i,j}$는 문서에 포함된 단어들의 조정된 one-hot 벡터들이다. One-hot 벡터를 조정하는 방법은 많이 출현하는 단어들을 0값을 가지도록 한다.

이제 단어의 문맥 정보 학습을 위해서 단어들의 one-hot 행렬에 문서 대표벡터($\widetilde{D}_i$)를 더한 조정된 one-hot 행렬을 이용하여 임베딩한다. 다른 목적함수로 문서 대표벡터에 대한 감성점수 함수를 정의해서 문서의

실제 감성 라벨과의 교차 엔트로피 목적함수도 정의할 수 있다. 이 두 가지 목적함수를 더하여 새로운 목적함수를 정의하면 단어의 문맥적인 의미와 감성 정보를 동시에 이용하여 임베딩할 수 있다.

Word2Vec의 단점을 보완해서 만든 GloVe모형에 단어의 감성 혹은 문서의 감성 정보를 결합한 모형이 제안되었다. 우선 이 모형에서는 문서수준감성과 단어수준감성이 존재한다. 먼저 문서수준감성에서는 $m_{i,k}$는 단어 w_i가 k감성 라벨의 문서에 등장하는 횟수이다. w_i의 감성분포 S_i는 다음과 같다.

$$S_i = \frac{(m_{i,1},...,m_{i,c})}{\sum_{k=1}^{c} m_{i,k}}$$

또는 다음과 같이 두 가지 감성 라벨만 존재하는 경우에는 비를 이용해서 감성정보를 표현할 수 있다.

$$B_i = \frac{m_{i,pos}}{\alpha m_{i,neg}}$$

여기서 α는 조율 모수이다.

단어수준감성에서는 문서의 감성을 이용할 수 없다. 따라서 다음과 같이 단순하게 단어의 감성을 표현한다.

$$B_i = \begin{cases} 1/\alpha & positive \\ 1 & neutral \\ \alpha & negative \end{cases}$$

위처럼 정의된 단어의 감성 정보를 기존의 GloVe 모형의 목적함수와 결합한다. 그 예시로 DLJT2 모형이 있다. DLJT2 모형은 기존의 GloVe 목적함수에서 $\tilde{w}_i^T w_i$에 위에서 제시한 방법으로 추정할 수 있는 감성정보

s_i를 곱하여 감성정보를 추가한다. 목적함수는 다음과 같다.

$$J = \sum_{i,k=1}^{V} f(X_{ik})(\tilde{w}_k^T w_i s_i + b_i + \tilde{b}_k - \log(X_{ik}B_i))^2$$

3. 온라인 학습

일반적으로 신경망 모형 기반의 키워드 추출에 사용되는 Word2Vec 기반의 방법들은 최적화 방법 자체가 확률적 근사에 기반하기 때문에 온라인 학습이 가능하다. 예를 들자면 온라인학습과 동시에 해석력을 높이기 위한 방법으로 앞서 소개한 l_1 정규화를 결합한 방법 외에도 이제 소개할 Projected Gradient Descent이 제안되었다.

기본적인 Skip-Gram 방법의 임베딩 벡터의 갱신은 다음과 같다.

$$w_{i,k}^{n+1} = w_{i,k}^{n} + \gamma \nabla f(w_{i,k})$$

여기서 w_i^k는 i번째 단어의 k 번째 반복에서의 임베딩 벡터이다. γ는 학습률이고 $\nabla f(w_i)$는 w_i에 대한 목적함수의 그래디언트이다.

하지만 Projected Gradient Descent은 해석력을 위해 임베딩 벡터들을 사영시킨다. 가장 단순한 사영으로 Naive projected gradient 방법이 제안되었다. 이 방법은 임베딩 벡터의 원소가 만약 음수라면 0으로 사영시킨다. 다음과 같다.

$$w_{i,k}^{n+1} = \max(0, w_{i,k}^{n} + \gamma \nabla f(w_{i,k}))$$

따라서 모든 임베딩 벡터들의 성분이 0 이상의 값을 가지게 된다. 벡터들의 성분이 비음수라는 성질은 유지되지만 매우 단순한 규칙으로 인해 학습 결과가 매우 불안정하다.

위의 Naive projected gradient 방법의 단점을 보완한 Improved

projected gradient 방법이 제안되었다. 다음의 방법으로 적절한 학습률(γ)을 반복적으로 찾는다.

$$R(\gamma) = \frac{\left|\left\{k | w_{i,k}^n > 0, w_{i,k}^n + \gamma \nabla f(w_{i,k})\right\}\right|}{d}$$

여기서 $R(\gamma)$는 갱신에 의해 0으로 바뀌어야하는 성분의 비율을 의미한다.

γ가 감소하게 되면 $R(\gamma)$ 또한 감소하게 된다. 따라서 갱신 규칙에 의해 0으로 바뀌는 임베딩 벡터의 성분이 줄어들게 된다. 학습률 γ을 다음과 같이 갱신한다.

$$\gamma^{m+1} = \gamma^m \bullet \beta \text{ with } 0 < \beta < 1$$
$$\text{until } R(\gamma^{m+1}) < \delta \text{ or } \gamma^{m+1} \leq \gamma_L$$

여기서 δ은 임계값이고 γ_L은 γ의 하한값이다.

학습률(γ)을 위와 같은 규칙으로 갱신함으로써 임베딩 벡터가 지나치게 변하는 것을 막음과 동시에 성분들이 비음수로 유지할 수 있다. 위와 같이 Improved projected gradient을 이용해서 온라인학습을 진행하면 동시에 해석력을 높일 수 있다.

제 5 장 결론

5 결론

저출산은 모두의 관심사로, 정부의 정책 발표 외에도 언론사에서 특집 기사로 많이 다루고 있는 주제이다. 저출산·고령사회 기본계획은 3차까지 정책 로드맵이 제시되어 있으며 2020년에는 제4차 제4차 저출산·고령사회 기본계획(2021~2025년)'을 수립하여 앞으로 20년, 40년 이상의 미래사회를 종합적으로 대비하는 다양한 핵심과제들을 적극적으로 발굴해 나갈 예정이다.

건강보험은 건강보험 보장성 강화 대책으로, 2022년까지 건강 보험 보장률을 70%까지 끌어올리는 것을 목표로 하고 있다. 이와 관련하여 선택진료비, 상급병실료(2, 3인실), 간병비 등 3대 비급여를 해소하고 초음파, MRI 등 국민 수요가 높은 비급여 항목에 건강보험을 적용하였으며, 재난적 의료비 제도화와 본인부담 상한제 개선 등 취약계층의 의료비 부담 완화를 위한 각종 대책을 시행하였다[5].

저출산, 건강보험 키워드로 수집된 문서들은 저출산과 건강보험과 관련된 직접적인 정책도 있었지만 일자리, 국민연금 키워드처럼 항상 이슈가 되는 키워드들이 포함된 문서들도 함께 수집되었다. 이는 저출산과 건강보험 이슈는 일자리와 국민연금 등 주요 이슈들과 함께 같이 언급되고 있다고 볼 수 있다.

4장에서는 기존에 개발된 주제추출 및 키워드 임베딩 방법에 기초한 키워드 추출 방법론을 살펴보았다. 우선 문서 내의 단어 특징 추출 및 주제 추출 방법론으로 네트워크 기반 단어분석, 단어 특징 추출 방법, 토픽

5) 노홍인. (2019). 전 국민 건강보험 30주년 성과와 과제 . 보건복지포럼, 통권 272호, 02-03. 재인용

모형을 설명하였다. 그리고 키워드 추출에 사용되는 최신 방법론으로 네트워크 기반, 단어 특성 추출 기반, 토픽 모형 기반 키워드 추출 방법론을 살펴보았다. 키워드 추출방법론의 고도화 기술로 구조적 성김성을 이용한 단어 특징 추출, 감성 분류정보를 반영한 특징 추출, 온라인 학습 방법도 기술하였다. 이는 보건·복지 분야의 기사 주제 정제를 위한 키워드 분석방법 고도화 연구에 도움을 줄 수 있다.

소셜 빅데이터를 활용하여 살펴본 위 분석결과는 당연한 결과일 수 있다. 그럼에도 불구하고, 소셜 빅데이터 분석은 보건복지 정책 영역에서 국가적·사회적으로 관심이 있는 이슈에 대해 현 상황을 파악하는 데 중요한 경쟁력으로 작용할 수 있으며, 앞으로의 정책 관련 이슈를 도출하고 연구 전략을 세우는 데 근거자료로 활용될 수 있다. 다양한 소셜 빅데이터 분석 기술을 바탕으로 주요 보건복지 정책에 관한 사회적 관심도, 영향력 등을 분석하고 그 변화 과정을 살펴본다면 시의성 높은 보건복지 정책 연구의 기반을 마련할 수 있을 것이다.

참고문헌

노홍인. (2019). 전 국민 건강보험 30주년 성과와 과제 . 보건복지포럼, 통권 272호, 02-03.

오미애, 최현수, 송태민, 이상인, 천미경. (2017) 2017년 소셜 빅데이터 기반 보건복지 이슈 동향 분석. 세종: 한국보건사회연구원.

이성직, 김한준. "TF-IDF 의 변형을 이용한 전자뉴스에서의 키워드 추출 기법." 한국전자거래학회지 14.4 (2009): 59-73.

조태민, 이지형. "LDA 모델을 이용한 잠재 키워드 추출." 한국지능시스템학회 논문지 25.2 (2015): 180-185.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, Vol.3, 993-1022.

Bingham, Ella, and Heikki Mannila. "Random projection in dimensionality reduction: applications to image and text data." Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001.

Chen, Minmin. "Efficient vector representation for documents through corruption." arXiv preprint arXiv:1707.02377 (2017).

Fu, Peng, et al. "Learning Sentiment-Specific Word Embedding via Global Sentiment Representation." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

Hulth, Anette. "Improved automatic keyword extraction given more linguistic knowledge." Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, 2003.

Lee, Junseok, et al. "Ensemble modeling for sustainable technology transfer." Sustainability 10.7 (2018): 2278.

Levy, Omer, and Yoav Goldberg. "Neural word embedding as implicit matrix factorization." Advances in neural information processing systems. 2014.

Li, Yang, et al. "Learning word representations for sentiment analysis." Cognitive Computation 9.6 (2017): 843-851.

Luo, Hongyin, et al. "Online learning of interpretable word embeddings." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

Matsuo, Yutaka, and Mitsuru Ishizuka. "Keyword extraction from a single document using word co-occurrence statistical information." International Journal on Artificial Intelligence Tools 13.01 (2004): 157-169.

Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Mu, Jiaqi, Suma Bhat, and Pramod Viswanath. "All-but-the-top: Simple and effective postprocessing for word representations." arXiv preprint arXiv:1702.01417 (2017).

Murphy, Brian, Partha Talukdar, and Tom Mitchell. "Learning effective and interpretable semantic models using non-negative sparse embedding." Proceedings of COLING 2012. 2012.

Palshikar, Girish Keshav. "Keyword extraction from a single document using centrality measures." International conference on pattern recognition and machine intelligence. Springer, Berlin, Heidelberg, 2007.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

Rose, Stuart, et al. "Automatic keyword extraction from individual documents." Text mining: applications and theory 1 (2010): 1-20.

Sarwan, N. S. "An intuitive understanding of word embeddings: From count vectors to word2vec." Access in 30 (2017): 07-17.

Sun, Fei, et al. "Sparse word embeddings using l1 regularized online learning." Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. AAAI Press, 2016.

Song, Yangqiu, et al. "Topic and keyword re-ranking for LDA-based topic modeling." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

Tang, Duyu, et al. "Learning sentiment-specific word embedding for twitter sentiment classification." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.

Tasci, Serafettin, and Tunga Gungor. "LDA-based keyword selection in text categorization." 2009 24th International Symposium on Computer and Information Sciences. IEEE, 2009.

OECD Data(data.oecd.org/pop/fertility-rates.htm (2019.10.18. 인출)