독후감 텍스트의 토픽모델링 적용에 관한 탐색적 연구*

A Study on the Application of Topic Modeling for the Book Report Text

이 수 상(Soo-Sang Lee)**

〈목 차〉

초 록

이 연구는 독후감 텍스트의 주제분석에 토픽모델링의 활용방안을 탐색하는 것을 목적으로 하고 있다. 텍스트의 주제분석 방안으로서 토픽모델링 분석방법을 이해하고, R에서 제공하는 "topicmodels" 패키지의 LDA 함수를 사용하여 23건의 사례 독후감 텍스트들을 대상으로 실제의 분석작업을 수행하였다. 토픽모델링 분석결과 16개의 토픽들을 추출하였고, 토픽과 구성 단어들의 관계에서 토픽 네트워크, 사례 독후감과 토픽들의 관계에서 독후감 네트워크를 구성하였다. 이후 토픽 네트워크와 독후감 네트워크를 대상으로 중심성 분석을 수행하였으며, 분석결과는 다음과 같다. 첫째, 16개의 토픽들이 1개의 컴포넌트를 가지는 네트워크로 나타났다. 이것은 16개 토픽들이 상호 연관되어 있다는 것을 의미한다. 둘째, 독후감 네트워크에서는 연결정도 중심성이 높은 독후감들과 낮은 독후감들로 구분이 되었다. 전자의 독후감들은 다른 독후감들과 주제적으로 유사성을 가지며, 후자의 독후감들은 다른 독후감들과 주제적으로 상이성을 가지는 것으로 해석하였다. 토픽모델링의 결과를 네트워크 분석과 결합함으로써 독후감의 주제파악에 유용한 결과들을 얻게 되었다.

키워드: 독후감, 토픽모델링, 주제분석, 네트워크 분석, 중심성

ABSTRACT

The purpose of this study is to explore application of topic modeling for topic analysis of book report. Topic modeling can be understood as one method of topic analysis. This analysis was conducted with texts in 23 book reports using LDA function of the "topicmodels" package provided by R. According to the result of topic modeling, 16 topics were extracted. The topic network was constructed by the relation between the topics and keywords, and the book report network was constructed by the relation between book report cases and topics. Next, Centrality analysis was conducted targeting the topic network and book report network. The result of this study is following these. First, 16 topics are shown as network which has one component. In other words, 16 topics are interrelated. Second, book report was divided into 2 groups, book reports with high centrality and book reports with low centrality. The former group has similarities with others, the latter group has differences with others in aspect of the topics of book reports. The result of topic modeling is useful to identify book reports' topics combining with network analysis.

Keywords: Book report, Topic modeling, Topic analysis, Network analysis, Centrality

* 이 논문은 2015년도 부산대학교 인문사회연구기금의 지원을 받아 연구되었음
** 부산대학교 문헌정보학과 교수(sslee@pusan.ac.kr)
•논문접수: 2016년 11월 18일 •최초심사: 2016년 11월 29일 •게재확정: 2016년 12월 21일
•한국도서관·정보학회지 47(4), 1-18, 2016. [http://dx.doi.org/10.16981/kliss.47.201612.1]

Ⅰ. 서론

독후감(book report)은 '독후감상문'의 준말이며, 뜻을 그대로 풀이하면 책이나 글을 읽은 후 느끼게 된 감상의 내용을 작성한 보고서이다. 독서감상문이라는 용어로도 많이 사용한다. 이 용어들에서 키워드는 '독서', '감상', 그리고 '문'이다. 독서(讀書)는 책을 읽는 것을 의미하고, 감상(感想)은 마음속에서 일어나는 느낌이나 생각을 말하며(표준국어대사전), 문(文)은 말 그대로 글(텍스트)이다. 그러므로 독서감상문은 책을 읽고 마음속으로 느끼고 생각하는 것을 표현한 글이다. 독서 후 독자의 "느낌과 생각"을 표현한 언어 텍스트를 분석하면, 책을 읽은 독자 개개인의 느낌과 생각이 무엇인지, 그리고 독서라는 행위가 가져오는 개인적 변화의 일면을 파악할 수 있는 좋은 근거가 된다.

한편, 감상문은 보고/듣고/읽은 후 느끼게 된 것을 글로 적은 보고서이다. 영화를 보고 난 다음 쓰는 영화감상문, 여행을 하고 난 다음 쓰는 여행감상문, 음악을 듣고 난 다음 쓰는 음악감상문처럼 책을 읽고 난 다음 쓰는 감상문이 독서감상문이다. 이처럼 감상문은 어떤 형태에 구애받지 않고 감상자 본인이 느낀 점을 자유롭게 표현하는 글이다. 개인적인 감상에 따른 글이기에, 원칙적으로 잘 쓰고 못 쓰고 하는 식으로 평가를 내릴 수 없다. 사람들마다 감상한 것이 다르며, 독후감으로 표현된 내용도 다르고, 비교할 수도 없다. 이처럼 동일한 대상이라도 다른 감상이 존재할 수 있다는 점은 분석의 욕구와 가치를 느끼게 하는 근거가 된다. 많은 사람들이 동일한 책을 읽고 난 다음, 제각기 느끼고 생각한 바를 직접 작성한 독후감들이라면, 무엇을 감상하였는지 분석하는 것은 매우 가치가 있다.

책을 읽고 무엇을 어떻게 감상하였는지를 표현한 텍스트를 개인별 또는 집단별로 분석하고 비교한다면, 학교나 도서관의 교사와 사서의 독후활동 지도에 좋은 참고자료를 얻을 수 있을 것이다. 독서 프로그램의 운영자는 책을 읽은 후 독자들이 표현한 감성적인 내용을 파악하고, 각자의 감성적 차이에 대응하면서, 개인별/집단별 맞춤형의 독후지도를 할 수 있다. 이처럼 독후감 분석의 결과는 책 중심보다 독자 중심의 독서활동 프로그램 운영에 유용한 정보가 될 수 있다.

독후감 텍스트 분석에 참조할만한 국내사례로는 구성요소에 대한 분석(박동진 2010)과 언어 네트워크 분석(이수상 2016)이 있다. 전자는 독후감 19편을 대상으로 4가지 유형의 독후감 요소로 구성되는 분석프레임을 사용하여 빈도를 분석하였다. 후자는 독후감 23편을 대상으로 명사를 추출하고 고빈도 명사를 자질로 하는 키워드들을 선정하고 동시출현관계를 바탕으로 하는 키워드 네트워크들을 구성하여 독후감을 구성하는 중심적인 키워드들을 탐색하고 있다.

이 연구는 텍스트 마이닝의 한 방법인 토픽모델링을 통해 특정한 사례의 독후감 텍스트들에 표현되는 주제 또는 관심사가 무엇인지를 파악하려고 한다. 이 연구의 주된 목적은 독후감 분석에 토픽모델링을 활용하는 방안을 모색하려는 것이다. 토픽모델링으로 분석되는 결과들을 활용하여 독후감에 나타나는 주제적 특성을 파악하고자 한다. 특정한 책을 읽고 작성한 독후감들에 표현되는 전체의 주제 요소들이 무엇인지 확인하고, 개별 독후감들에서 표현되는 주제들이 어떻게 유사하고 다른지 파악하는 것도 가능하다. 이러한 결과들은 도서관 사서들이 독후감을 기초로 독서지도에 활용할 수 있는 모티브를 발견하려는 의도로 시도되었다.

Ⅱ. 독후감의 구성요소

독서감상문으로서 독후감은 독서대상의 도서와 독자의 상호작용으로 만들어지는 결과물이며, 편지, 일기, 시, 수필, 비평 등 다양한 형식으로 작성된 산문(에세이)이다. 독후감 쓰기는 학교도서관이나 공공도서관의 주요 독서활동 중 하나이다. 모두들 독후감을 잘 쓰라고 요구하지만, 독후감이라는 것 자체가 개인이 독서라는 체험에서 얻은 느낌이나 감정을 진솔하게 표현하는 것(박동진 2010)이기에, 잘 쓰고 못 쓰고 하는 판단기준을 설정하기가 어렵다.

그런데, 각종 독후감대회, 독후감 공모전 등과 같은 행사에서는 평가지표를 통해 독후감을 평가하여 우수한 독후감을 선정한다고 하지만, 평가기준(심사기준)의 적절성이 문제가 될 수 있다. 그리고 학교현장에서는 독서교육의 일환으로 학생들에게 독후감을 독서교육종합지원시스템에 제출토록 하고, 사서교사나 선생들이 그것에 대해 적절한 평가내용을 제공하고 있다. 독후감이라는 것이 개인이 느끼고, 생각하고, 깨달은 것(감상)을 표현한 것이고, 감상의 주체는 서로 다른 개성을 가진 개인인데, 이것을 심사하고 평가하는 방식으로 접근하는 것은 문제이다. 개인의 감상적 체험을 진솔하게 표현만 하면 좋은 독후감 또는 우수한 독후감이라 해도 충분하다. 독자가 무엇을 감상하였는지 확인한다면, 그리고 다른 독자들과 감상한 내용의 유사점과 차이점을 구분할 수 있다면, 독서지도자는 그것을 근거로 추수적인 독서활동을 진행할 수 있을 것이다.

일반적으로 독후감은 독서대상의 도서에 관한 내용과 필자가 감상한 내용으로 크게 구분할 수 있다. 이러한 관점에서 박동진은 독후감의 구성요소를 <표 1>과 같이 도서와 관련된 정보요소와 필자와 관련된 감상요소로 구분하고 있다(박동진 2010).

〈표 1〉 독후감의 구성요소(박동진의 구분)

구성요소		세부 구성요소
도서와 관련된 정보요소	형식적 정보요소	-서명 -저자/역자 -쪽수 -출판사 -출판년도 -판형
	내용적 정보요소	-목차 -주요내용(줄거리) -주제/저자의 의도 -구성/문체 -장르유형
필자와 관련된 감상요소	도서내용과 관련된 감상요소	-도서내용에 대한 해석 -도서내용에 대한 비판이나 평가 -도서내용에서 비롯한 느낌이나 감동 -도서내용과 관련된 필자의 의견이나 생각(관점) -독서 후의 변화된 생각이나 태도 -독서 후에 새롭게 알게 된 지식 -도서내용과 관련된 필자의 경험이나 배경지식
	독서행위와 관련된 수행요소	-독서를 하게 된 이유나 동기 -독서의 방법이나 과정에 대한 성찰 -도서내용과 관련이 있는 자료나 도서

한편, 김라연과 박은경은 독후감의 구성요소를 이해, 감상, 해석의 영역으로 구분하고 각 영역별 하위요소를 정리하고 있다(김라연, 박은경 2011). 그 내용은 <표 2>와 같다. 이해의 영역은 <표 1>의 도서와 관련된 정보요소, 감상의 영역은 필자와 관련된 감상요소와 유사하다.

〈표 2〉 독후감 구성요소(김라연과 박은경의 구분)

영역	하위요소
이해	줄거리, 구성, 주제, 형식적 정보, 문체, 장르
감상	독서 동기, 경험, 감동, 성찰, 의견과 견해, 비평, 인상, 관련 자료
해석	'이해'와 '감상'의 연결고리

이와 같은 구성요소에 따라 독후감의 내용을 분석하려면 기본적으로 수작업으로 수행해야 하며, 주관적으로 판단하여야 한다. 그러기에 구성요소에 의한 내용분석은 독후감의 양이 방대하거나 보다 객관적인 분석을 요구할 경우, 분석을 어렵게 한다. 대안으로 자동화된 환경에서의 텍스트 주제분석 방법이 고려되며, 독후감을 구성하는 주제적 구성요소와 특성을 자동으로 분석할 수 있을 것이다.

Ⅲ. 텍스트의 주제분석 방법

언어로 된 텍스트의 내용을 분석하는 방법은 다양하다. 전통적인 방법으로 정성분석, 내용분석 등이 있으며, 컴퓨터를 활용하는 분석방법으로 언어 네트워크 분석, 텍스트 마이닝 등이 있다. 최근에는 데이터의 규모에 상관없고, 자동화된 언어처리 기술을 활용하면서, 텍스트에 내재되어 개념이나 의미까지 분석 가능한 방법으로서 텍스트 마이닝이 부각되고 있다.

1. 텍스트 마이닝

텍스트 마이닝(text mining)은 텍스트 형태의 데이터에 마이닝 기법을 적용한 것으로, 대규모 비정형 형태의 대규모 텍스트 내에 숨겨진 가치있는 지식을 탐색하는 것을 말한다. 일종의 지식발굴(knowledge discovery) 행위이다. 텍스트 마이닝이 부각되는 배경에는 이메일, 전자신문, 인터넷 포털, 블로그, SNS 등과 같은 소셜 데이터, 학술논문, 학위논문, 특허, 판례 등과 같은 학술 데이터, 사건/정책 보고서, 업무보고서, 법령 등의 공적 데이터, 개인기록, 일기, 면담/상담 텍스트 등의 사적 데이터 등 다양한 디지털 텍스트들이 폭발적으로 증가하고, 관련된 분석기법들이 등장하였기 때문이다. 숨은 지식(가치가 있고 의미가 있는 정보)을 탐색하고 발굴하기 위해서는 통계분석, 군집분석, 네트워크 분석, 시각화 분석 등의 다양한 분석기법을 사용하게 되는데, 이들을 적용하는 프로그래밍 도구나 분석도구들도 많으며, 어렵지 않게 활용이 가능하다.

사람들은 이러한 텍스트 형태의 문헌(또는 문서)을 작성하면서, 주제, 감정, 의견 등과 같이 자신이 의도하는 바를 여러 가지 단어들을 사용하여 표현한다. 텍스트 마이닝은 텍스트에 표현된 단어들을 근거로 이러한 사람들의 의도를 파악하게 된다. 이처럼 텍스트 마이닝은 많은 사람들이 작성한 문헌 텍스트 집합에서 그들의 의도를 파악하는 텍스트 내용분석 방법이다. 대상이 되는 문헌 텍스트 집합의 규모가 방대한 경우 보다 자동화된 기술들을 동원하여 분석작업을 한다. 즉, 텍스트 마이닝은 대규모 문헌 텍스트 집합의 자동화된 내용분석 기법이다. 언어 텍스트 영역의 빅데이터 분석의 대표적이 방법이 바로 텍스트 마이닝이다.

텍스트 마이닝이 기존의 언어 텍스트 분석과 차이가 나는 점은 분석대상의 규모가 방대하고, 다양한 언어처리 기술을 활용한다는 것이다. 그런데, 방대한 규모를 나누는 기준이 명확하지 않기 때문에, 텍스트 집합에 언어처리 기술을 사용하여 키워드들을 추출하고, 자동화된 환경에서 통계나 네트워크 관점에서 분석하는 자체를 의미한다고 보는 것이 가장 타당할 것

이다. 분석과정에 자동화 기술을 좀 더 많이 사용한다는 의미이다. 분석대상 텍스트의 양이 문제가 중요한 것이 아니라, 무엇을 어떻게 분석할 것인가 하는 문제가 중요함을 알 수 있다.

2. 토픽모델링

텍스트 마이닝의 한 방법인 토픽모델링(topic modeling)은 언어 텍스트 집합을 가장 잘 표현하는 토픽(topic) 또는 주제범주를 구분하는 기법이다. 언어 텍스트 집단의 주제분석 방법인 것이다. 토픽모델링에서 단위 텍스트는 문헌 또는 문서(document)라 한다. 그리고 분석 대상이 되는 대량의 문서 집합을 코퍼스(corpus)라 한다. 토픽은 코퍼스의 주제범주(또는 주제 프레임)이며, 각 토픽마다 그 주제를 표현하는 단어들의 리스트로 구성된다.

토픽모델링에서 표현하는 형식으로 풀어쓰면, 코퍼스는 N개의 문서들로 구성되고, 각 문서에는 N_d 만큼의 고유한 단어(키워드)들이 등장한다. 전체 문서에서 등장하는 고유한 단어들의 수 V는 어휘(vocabulary)라고 한다. 토픽모델링은 이렇게 각 문서에서 V개 단어들의 사용 패턴을 근거로 전체 K개의 토픽을 확률적으로 추정하여 제시한다. K개의 토픽은 코퍼스를 대표하는 전체의 토픽들이며, 각 문서의 입장에서는 K개 토픽 중 일부의 토픽들이 나타난다. 그러므로 토픽모델링의 결과는 ① K개의 토픽 추정, ② 각 문서에 해당하는 토픽들(document topics) 추정, 그리고 ③ 각 토픽에 포함되는 단어들(topic words)도 추정이 된다. 코퍼스에 등장하는 단어들로 이러한 잠재된 변인들을 추정하는 것이 토픽모델링이다.

토픽모델링으로 추정한 토픽들은 언어 네트워크 분석의 키워드 선정 과정에서 활용할 수 있다. 먼저 토픽모델링으로 문서의 집단 전체를 구성하는 K개의 토픽들을 도출하여 이것을 주제 프레임으로 정한다. 그리고 각 문서들을 대표하는 토픽들을 확인하면, 토픽을 키워드로 하는 문서와 토픽의 언어 네트워크를 구성할 수 있다. 더불어, 각 토픽을 구성하는 주요 단어들의 분포관계를 이용하여 토픽과 단어의 언어 네트워크도 구성할 수 있다. 토픽모델링은 잠재의미색인(LSI, Latent Semantic Indexing) 기법을 확장한 확률적 잠재의미색인(PLSI, Probabilistic Latent Semantic Indexing), 잠재 디리클레 할당(LDA, Latent Dirichlet Allocation) 기법들을 활용하여 토픽을 추정하는 주제분석 방법이다. 최근의 토픽모델링에서는 LDA 기법을 기초로 LDA를 변형한 다양한 기법들까지 사용한다.

현재의 토픽모델링을 대표하고 기초가 되는 LDA 기법은 주어진 문서집단인 코퍼스에 대하여, 각 문서에 어떤 토픽들이 존재하는지를 확률적으로 판단하는 확률모형이다. LDA 기법의 토대를 이루는 것으로, 사람들이 필요에 의해 어떤 토픽들을 나타내는 문서를 작성하는 과정에서, 문서-토픽-단어 사이에 존재하는 다음과 같은 가정을 전제한다. 토픽은 관련이 있는 단어들로 구성된다. 한 문서가 복수의 토픽을 다룬다면, 각 토픽마다 관련된 단어들을

동원하여 사용한다. 이와 같은 가정을 역으로 하여 문서의 관점에서 보면, 하나의 문서는 여러 가지 토픽들로 구성되어 있고 각 토픽들은 그 토픽에 속하는 여러 단어들을 조합하여 구성된다. 물론 특정한 단어는 하나의 토픽만을 위해 사용하지 않고, 복수의 토픽에 사용될 수도 있다. 그래서 문서는 토픽의 조합(mixture of topics)이고, 토픽은 단어의 조합(mixture of words)이라고 한다. 다시금 정리하면, 문서 내에 등장하는 단어들을 기반으로 토픽을 구성하고, 이것을 토대로 문서의 토픽들을 추정한다.

Blei 등에 의해 소개된 LDA 모델(Blei et al. 2003)은 LDA 기법을 적용한 토픽모델링 방법이다. 실제의 적용과정을 설명하면 다음과 같다. ① N개의 문서들로 구성되는 코퍼스를 구축한다. ② 적절한 방법을 통해 문서에서 주요한 단어(키워드)들을 추출(선정)한다. ③ 추출된 각 단어들을 모아서 전체 V개의 고유한 단어들의 집합인 어휘를 구성한다. ④ 문서(행)와 어휘(열)의 출현빈도(행렬의 셀 값)로 구성되는 문서-단어의 행렬(DTM, document-term matrix)을 생성된다. 이것은 N x V 행렬이 된다. ⑤ 각 문서에 등장하는 단어들의 조합관계를 통해 확률적으로 K개의 토픽을 배당한다. K개 토픽마다 구성되는 단어들의 집합을 확률로 추정하여 조합하게 된다. 어떤 단어가 특정 토픽에 속하는 정도를 확률로 계산하는 것이기에, 그 토픽에 속하는 정도를 알 수 있게 된다. 이러한 과정을 통해 얻어지는 결과는 N개의 각 문서에 대한 토픽들, K개의 토픽에 대한 단어들에 대한 추정확률을 가지는 행렬이 된다.

요약하면, N개의 문서에 등장하는 단어들(N x V 행렬, 각 문서별 단어의 등장빈도)을 토대로 N개 문서의 토픽들(N x K 행렬, 각 문서별 토픽들의 비율)과 K개 토픽의 단어들(K x V 행렬, 각 토픽별 단어들의 비율)을 확률적 추정으로 구하게 된다(정한조 2015; 박자현, 송민 2013). 한편, 문서 텍스트 집단이 방대하고 시기적 변수에 따라 구분할 수 있으면, 문서 텍스트를 시기별로 나누어 각 시기별 토픽들의 특성을 파악할 수 있다. 이것을 동적 토픽모델링(dynamic topic modeling)이라고 하며, 각 시기별 토픽의 변화양상을 파악할 수 있다. 그리고 각 문서의 메타데이터(저자, 대상 등)를 활용하여 저자별 토픽모델링, 대상별 토픽모델링 등을 수행할 수도 있다.

문서집단(코퍼스)의 토픽모델링 결과로 얻어지는 토픽(주제범주)과 구성단어 리스트, 그리고 문서에 분포하는 토픽 리스트를 이용하면, 토픽과 단어 관계 그리고 문서와 토픽 관계의 행렬에서 이원모드 네트워크를 구성할 수 있다. 또한 이러한 이원모드 행렬들에서 문서들, 토픽들, 단어들 각각에 해당되는 일원모드 행렬을 구하여, 각각의 언어 네트워크들을 구성할 수도 있다. 이처럼 토픽모델링은 언어 네트워크 분석과 혼합하여 분석의 범위를 확장시킬 수 있다.

토픽모델링 실행 도구는 대체로 MALLET라는 토픽모델링 패키지를 사용하거나 R에서 제공하는 패키지들을 사용한다. 자바 기반의 MALLET는 토픽모델링뿐만 아니라 자연어처리,

문서분류, 클러스터링 등과 같은 기계학습을 가능하게 하는 응용 프로그램이다. R에서 토픽 모델링을 가능하게 하는 패키지는 “topicmodels”, “lda”, “RMallet” 등이 있다.

MALLET 또는 R의 패키지들을 사용하려면 사전에 토픽수(K값)를 결정해야 한다. 대체로 두 가지 방법을 사용한다. 첫째, 여러 개의 토픽수를 부여하여 얻어지는 결과를 가지고 주제 범주화가 가장 잘 되었다고 분석자가 판단하는 토픽수를 선택하는 방법이다. 이 경우 분석자가 판단해야 하는 부담이 있다. 둘째, 토픽수를 알고리즘에 의해 결정하는 방법이다. 토픽수 결정 알고리즘은 대체적으로 확률분포의 퍼플렉서티(perplexity) 값을 이용한다. 퍼플렉서티는 당황, 당혹 등과 같이 확률분포 모형에서 결과의 만족도가 낮을 때 큰 값을 나타내며, 만족도가 높을 때 작은 값을 나타낸다. 따라서 LDA에서 최적의 토픽수는 퍼플렉서티가 낮은 값을 가지는 K값을 찾는 문제가 된다. 결국 다양한 K값을 선택하여 퍼플렉서티 값이 작게 나타나는 K값을 선택하는 방식으로 토픽수를 결정한다(이원상, 손소영 2015).

이러한 토픽모델링과 관련된 국내 연구는 토픽모델링 자체에 대한 연구보다 그것을 활용하는 연구들이 많다. 학술논문을 중심으로 정리한 국내의 연구사례들은 다음과 같다. 첫째, 학술논문의 초록 텍스트를 대상으로 특정 주제분야의 연구동향 분석을 다루고 있다(박자현, 송민 2013; Su Yeon Kim et al. 2015; 유소영 2015; 이기헌 등 2015). 둘째, 특정한 텍스트를 대상으로 하는 주제분석에 관한 연구들이다. 트위터 데이터(진설아 등 2013), 신문기사(강범일 등 2013), 신문기사와 연구논문(안주영 등 2016), 일기자료(남춘호 2016), 동인지 문학작품(이재연 2016) 등이 해당된다.

Ⅳ. 독후감의 토픽모델링 분석

독후감의 대상이 되는 책은 <가족의 두 얼굴>이다. 2012년 출판 된 책으로 한세대학교 상담대학원 교수이자 트라우마가족치료 연구소장인 최광현이 저자이다. 이 책은 개인의 상처에는 대부분 가족과 연결되어 있다고 하면서, 가족에 관한 다양한 문제들을 심리치료의 관점에서 소개하고 설명하고 있다. 분석대상 독후감은 전체 23편이며, 일반부 8편, 고등부 9편, 중등부 6편으로 구분된다.

1. 분석과정

독후감 텍스트는 R의 토픽모델링 분석 패키지인 “topicmodels”에서 제공하는 LDA 함수를 사용하였다. 구체적인 분석과정은 다음과 같다.

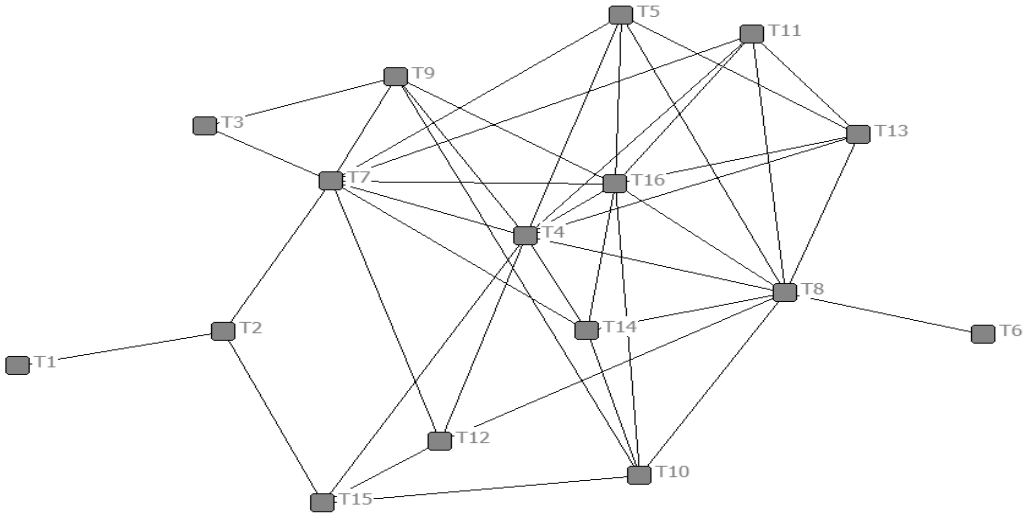
첫째, 형태소 분석 과정으로 KLT(Korean Language Technology)라는 한글형태소 분석기를 통해 명사형 단어들을 키워드로 추출, 불용어로 판단되는 단어들을 삭제하였다. 둘째, 코퍼스 구성 과정으로 독후감 텍스트를 추출한 명사 리스트로 재구성하였다. 그리고 재구성된 독후감 텍스트를 입력파일로 하여 코퍼스로 구성하였다. 셋째, DTM 생성 과정이다. 코퍼스에서 23건의 문서와 456건의 키워드로 구성되는 DTM을 생성하였다. 넷째, 토픽수 결정 과정으로 전체 입력문서에 해당되는 토픽수를 사전에 결정하여야 한다. 이 연구에서는 Ponweiser가 제안한 최적의 토픽수 결정 기법(Ponweiser 2012)을 Meza가 구현한 알고리즘(Meza 2015)을 활용하여 토픽수 K값을 산정하였다. 즉 토픽수를 2개에서 100개까지 순차적으로 LDA 함수(반복횟수 1,000회)를 실행하여 얻어지는 우도(log-likelihood, 가능도) 값의 조화평균(harmonic mean)이 최대가 되는 K값으로 토픽수로 결정한다. Meza 알고리즘은 기계학습 기법의 유형이므로, 적용할 때마다 다른 결과가 나타난다. 따라서 Meza 알고리즘을 전체 3회 실행하였고, 최적의 토픽수는 18, 16, 17이 산출되었다. 이 중에서 토픽수가 가장 적은 16을 토픽수(K값)로 선택하였다. 그리고 이 과정에서 은닉된 잠재변인들의 추론은 Liu 등이 추천한 Gibbs 샘플링 기법을 사용하였다(Liu et al. 2016). 다섯째, LDA 기법을 적용하는 과정으로 “topicmodels” 패키지에서 제공하는 LDA 함수를 사용하였다. 앞에서 생성한 최종 DTM을 입력 데이터로 하여, 토픽수(K값)를 16으로 하고, 반복횟수는 1,000회, 다른 인수들은 기본값으로 하며, Gibbs 샘플링 기법으로 LDA 함수를 적용하였다.

2. 분석결과

가. 토픽과 주요 단어의 분석결과

LDA 함수의 적용결과 16개 토픽을 구성하는 단어들, 23건의 독후감 문서들의 토픽분포의 비율을 산출하였다. 먼저 토픽을 구성하는 주요 단어들(분포비율 상위)과 토픽명은 <부록>과 같다. 전체적으로 보면, 책(<가족의 두 얼굴>), 가족, 사람들과의 관계, 상처 등에 대한 주제들을 나타내고 있다. 이 중에서 두 개 토픽 T8(나와 책), T16(말에 의한 상처)의 비율이 상대적으로 높게 나타났다. 전자는 독후감의 대상인 책에 관한 토픽이고, 후자는 가족 간에 있어 말에 의한 상처를 나타내는 토픽이다.

이들 토픽과 주요단어들의 행렬에서 토픽들의 네트워크를 구성하여 보았다. 각 토픽별 주요단어들의 출현빈도를 구하고, 그것의 동시출현빈도의 평균값(1.32)을 기준으로 유사도를 측정하여, <그림 1>과 같은 이진형(binary) 토픽 네트워크를 구성하였다. 밀도 0.342, 평균연결거리 5.125를 가지는 하나의 컴포넌트로 구성되는 네트워크이다. 전체가 하나의 컴포넌트로 구성됨을 알 수 있다.



〈그림 1〉 토픽 네트워크 사례

16개 토픽들의 중심성(연결정도 중심성, 근접 중심성, 매개 중심성) 분석결과는 <표 3>과 같다. 23건의 독후감에서 추정한 16개 토픽 중에서 연결정도 중심성의 순위가 높은 토픽은 T4, T7, T8, T16 등이며, 낮은 토픽은 T1, T3, T6이다. 근접 중심성은 T4, T7, T8 등이 높고, T1, T6 등이 낮다. 매개 중심성은 T2, T4, T7, T8 등이 높고, T1, T3, T6, T13 등이 낮다.

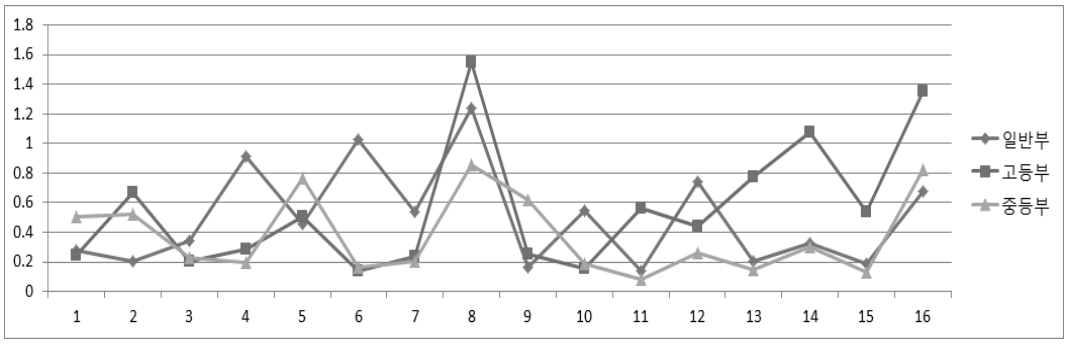
결국 독후감에서 중심적인 역할을 하는 토픽은 T4, T7, T8 등이며, T1, T3, T6 등은 상대적으로 중심성의 순위가 낮다. 중심성이 높은 토픽들은 다른 토픽들과 연관성이 높다는 의미이고, 낮은 토픽들은 연관성이 낮다는 의미이다.

〈표 3〉 토픽 네트워크의 중심성

토픽	연결정도	근접	매개	토픽	연결정도	근접	매개
T1	1	0.333	0	T9	5	0.556	3.933
T2	3	0.484	14.533	T10	5	0.577	3.95
T3	2	0.441	0	T11	5	0.577	1.817
T4	10	0.714	14.4	T12	4	0.556	2.644
T5	5	0.577	1.817	T13	5	0.517	0.2
T6	1	0.385	0	T14	5	0.577	1.533
T7	9	0.682	29.317	T15	4	0.536	6.433
T8	9	0.6	19.067	T16	9	0.682	7.356

나. 독후감과 토픽의 분석결과

토픽모델링 결과로 얻어지는 23건의 독후감들과 16개 토픽들의 분포비율 데이터를 3집단(일반부, 고등부, 중등부)의 독후감들로 구분한 토픽분포 비율은 <그림 2>와 같다. 이 그림에서 보면, T8과 T16의 비율이 상대적으로 높음을 알 수 있다. 그리고 일부 토픽을 제외하고는 3집단(일반부, 고등부, 중등부)의 독후감에서 나타나는 토픽들에 다소 차이가 있다. 3집단의 독후감에서 강조하는 토픽들이 서로 조금씩 다르다는 것을 알 수 있다.



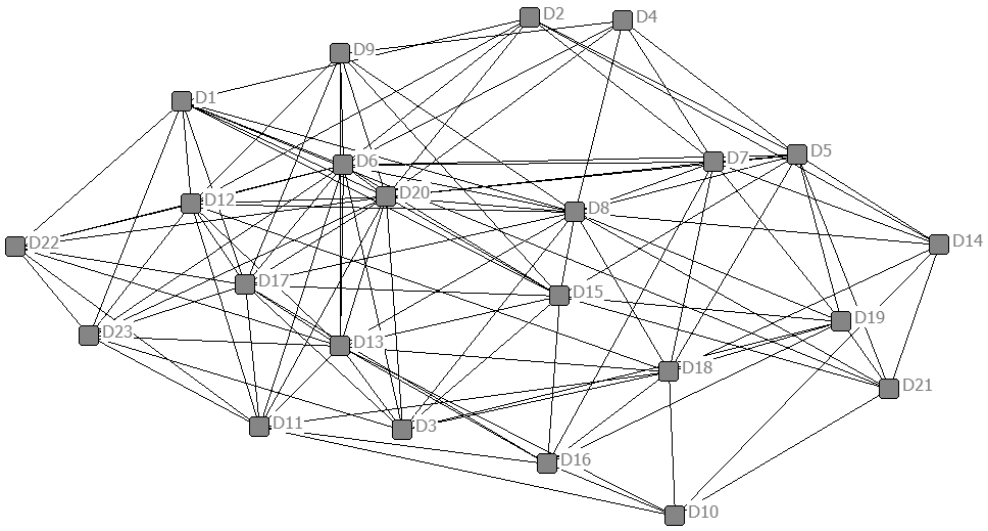
<그림 2> 3집단(일반부, 고등부, 중등부) 독후감들의 토픽비율

그리고 독후감별로 분포되는 토픽들의 분포비율이 높은 5순위 토픽까지 추출한 결과는 <표 4>와 같다. 1순위 토픽만으로 볼 경우, 전체 16개의 토픽 중에서 T2, T12, T14는 2건의 독후감에서, T8, T16은 3건의 독후감에서 등장하고 있다. 이들을 제외한 나머지 11건은 모두 한 번씩 나타났다.

<표 4>의 독후감과 토픽의 분포순위 관계에서 구성한 독후감의 네트워크는 <그림 3>과 같다. 동일한 토픽이 3회 이상 등장하는 조건으로 연결관계를 설정하였다. 이 네트워크는 23건 독후감들이 모두 연결되어 하나의 컴포넌트로 구성된다는 것을 알 수 있다. 동일한 책을 읽고 작성한 독후감들이기에 주제적으로 유사할 수밖에 없는 당연한 결과이다. 만일 동일한 책의 독후감들로 구성되는 네트워크에서 복수의 컴포넌트들로 분리된다면, 그것은 주제적으로 서로 다른 내용을 담고 있다고 판단할 수 있다. 만일 복수의 컴포넌트들로 분리된다면, 이들은 서로 다른 주제의 독후감들이 된다.

〈표 4〉 독후감과 토픽의 관계

독후감 \ 토픽순위		1순위	2순위	3순위	4순위	5순위
1	일반부-1	T12	T8	T5	T14	T1
2	일반부-2	T8	T6	T1	T4	T14
3	일반부-3	T10	T8	T16	T5	T9
4	일반부-4	T7	T6	T16	T5	T4
5	일반부-5	T8	T6	T4	T12	T16
6	일반부-6	T6	T8	T16	T5	T14
7	일반부-7	T3	T16	T8	T4	T6
8	일반부-8	T4	T5	T16	T12	T8
9	고등부-1	T15	T16	T14	T5	T4
10	고등부-2	T11	T8	T16	T13	T2
11	고등부-3	T16	T8	T14	T7	T2
12	고등부-4	T14	T8	T5	T4	T2
13	고등부-5	T16	T2	T15	T5	T8
14	고등부-6	T13	T8	T16	T1	T4
15	고등부-7	T12	T5	T16	T8	T15
16	고등부-8	T2	T8	T15	T16	T3
17	고등부-9	T14	T8	T15	T16	T5
18	중등부-1	T9	T16	T4	T8	T2
19	중등부-2	T16	T8	T3	T12	T9
20	중등부-3	T5	T8	T16	T14	T6
21	중등부-4	T1	T16	T8	T13	T12
22	중등부-5	T2	T5	T14	T7	T8
23	중등부-6	T8	T5	T14	T2	T10



〈그림 3〉 독후감 네트워크 사례

<표 5>는 <그림 3>의 독후감 네트워크에 대한 3가지 중심성 분석의 결과이다. 근접 중심성은 값의 차이가 크지 않은 상태에서 순위를 나타내고 있으며, 상대적으로 매개 중심성은 값의 차이가 크게 나타났다. 네트워크의 중심성에서는 값의 차이보다 순위가 중요하지만, 연결정도 중심성은 값의 차이에도 의미를 부여할 수 있다. 따라서 연결정도 중심성에서 보면 D6(일반부), D8(일반부), D13(고등부), D20(중등부) 등은 높은 값을 가지며, 이들은 그만큼 연결되는 다른 독후감들과 상호 주제적으로 유사하다고 해석할 수 있다. 반면에 연결정도 중심성이 낮은 D2(일반부), D4(일반부), D10(고등부), D14(고등부), D21(중등부)은 상대적으로 다른 독후감들과 주제적인 유사성이 낮다는 해석이 가능하다. 이처럼 중심성 분석의 결과에서는 독후감들이 주제적으로 유사한 것들과 그렇지 않은 것들을 판단할 수 있다.

〈표 5〉 독후감 네트워크의 중심성

독후감	연결정도	근접	매개	독후감	연결정도	근접	매개
D1	10	0.647	6.026	D13	14	0.733	11.503
D2	7	0.595	2.365	D14	7	0.564	3.293
D3	9	0.629	3.305	D15	12	0.688	7.66
D4	6	0.564	0.45	D16	8	0.611	3.861
D5	11	0.667	6.894	D17	13	0.71	5.847
D6	16	0.786	15.08	D18	11	0.667	9.219
D7	10	0.647	7.144	D19	8	0.595	2.433
D8	16	0.786	18.953	D20	16	0.786	15.08
D9	8	0.611	1.16	D21	7	0.595	3.882
D10	6	0.564	3.224	D22	8	0.579	0.292
D11	10	0.647	5.417	D23	9	0.611	1.005
D12	12	0.688	5.907				

Ⅴ. 결론

지금까지 23건의 독후감들을 대상으로 토픽모델링을 통해 독후감의 주제분석을 시도하였고, 그 결과를 통해 독후감에 나타나는 주제적 특성을 탐색해 보았다. 토픽모델링은 R에서 제공하는 “topicmodels” 패키지의 LDA 함수를 적용하였으며, 연구결과는 다음과 같다.

첫째, 16개의 토픽(주제범주)을 도출하였다. 각 토픽들이 구성하는 단어 리스트는 토픽의 분포비율의 순위에 따라 20개까지 선택하였다. 둘째, 토픽과 구성단어의 리스트를 통해 토픽 네트워크를 구성하고, 기본적인 특성과 중심적인 토픽들을 확인하였다. 셋째, 독후감별 토픽

의 분포비율을 이용하여 3집단의 독후감에서 강조하는 토픽들의 패턴에 차이가 있음을 알 수 있다. 넷째, 독후감 네트워크를 구성하여, 기본적인 특성과 중심적인 독후감들을 파악하였다.

LDA 함수를 적용하는 토픽모델링에서 가장 중요한 의사결정은 전체 문서들에 나타나는 토픽수를 결정하는 방법이다. 이 연구에서는 Meza 알고리즘으로 이 문제를 해결하였다. 또한 얻어진 토픽들에 대해 적절한 토픽명을 부여하는 작업도 중요하다. 도출된 토픽을 구성하는 단어들을 보고 분석자가 판단해야 하는 문제이며, 판단과 결정이 쉬운 작업은 아니다. 이 연구에서 사용한 각 독후감들은 평균 200자 원고지 기준으로 15매에 해당되는 규모가 큰 텍스트이기에, 토픽을 구성하는 단어(키워드)들로 토픽들의 의미를 파악하고, 적절한 토픽명을 부여하는 것이 어려웠다. 그래서 토픽과 구성단어들의 관계에서 토픽 네트워크를 구성하여 시각화와 중심성 분석을 통해 중심적인 토픽들을 확인하고, 유사한 토픽들을 범주화하고, 추출된 토픽들의 특성을 탐색하는 작업을 추가하였다.

이 연구는 토픽모델링이라는 비지도 학습 또는 자율학습(unsupervised learning)의 방법을 활용하여 독후감의 주제를 구분하고, 그 결과를 활용하여 독후감의 주제적 특성을 파악하는 탐색적 연구였다. 그러기에 가장 중요한 한계는 박동진 등이 제안한 독후감 구성요소에 따른 내용분석의 수준까지 주제분석을 수행하지 못한 점이다. 대신에 적절한 크기의 토픽(주제범주)을 구분하고, 독후감-토픽-단어의 관계에서 독후감과 토픽의 네트워크들을 구성하여 독후감과 토픽이 나타내는 주제적 특성들을 파악하는 수준은 가능하였다.

향후 다양한 확장연구를 통해 좀 더 유용한 독후감의 주제분석 방법을 탐색하여야 할 것이다. 이를 위한 후속연구들은 다음과 같이 제안할 수 있다. 첫째, 독후감 텍스트의 전체 내용에서 독자의 주관적인 입장이나 감상적 표현이 있는 문장들만 추출하여, 이들을 대상으로 토픽모델링 분석작업이 필요하다. 그리고 장르별로 다양한 도서를 대상으로 하며, 다양한 유형의 집단들에 대한 감상내용을 비교분석하는 작업도 가능하다. 둘째, LDA 기법이 가지는 비지도 학습의 단점을 극복하기 위해, LDA 기법을 변형한 다양한 기법들을 적용하여, 토픽의 적절성을 높이는 작업이 필요하다. 셋째, 독후감의 감상을 나타내는 단어들을 추출하여 그것으로 독후감 감상사전을 구축한 다음, 감상사전을 기반으로 하는 독후감의 감성분석 작업도 필요하다. 이러한 연구들의 결과들이 축적되면, 도서관의 독서지도 프로그램 또는 독서치료 영역에서 독후감의 주제분석 방법과 결과들의 활용성이 더 높아질 것으로 판단한다.

참고문헌

Blei, D.M, Ng, A.Y, Jordan, M.I. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research*, 3: 993-1022.

Liu Lin, et al. 2016. “An overview of topic modeling and its current applications in bioinformatics.” *SpringerPlus*, 5(1). <DOI: 10.1186/s40064-016-3252-8>.

Meza, David. 2015. “Topic Modeling in R.” <http://davidmeza1.github.io/2015/07/20/topic-modeling-in-R.html> (인용: 2016. 11.15).

Ponweiser, Martin. 2012. *Latent Dirichlet allocation in R.* Diploma thesis, Institute for Statistics and Mathematics, WU (Wirtschaftsuniversitat Wien), Austria.

Su Yeon Kim, Sung Jeon Song, Min Song. 2015. Investigation of topic trends in computer and information science by text mining techniques.『정보관리학회지』, 32(1): 135-152.

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

김라연, 박은경. 2011. 독서 감상문의 구성 요소 분석 연구.『교육과학연구』, 17: 17-30.

남춘호. 2016. 일기자료 연구에서 토픽모델링 기법의 활용가능성 검토.『비교문화연구』, 22(1): 89-135.

박동진. 2010. 독서 감상문 쓰기의 실태 - 구성 요소 분석을 중심으로.『새국어교육』, 84: 109-125.

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

안주영, 안규빈, 송민. 2016. 텍스트 마이닝을 이용한 매체별 에볼라 주제 분석.『한국문헌정보학회지』, 50(2): 289-307.

유소영. 2015. 자아 중심 네트워크 분석과 동적 인용 네트워크를 활용한 토픽모델링 기반 연구동향 분석에 관한 연구.『정보관리학회지』, 32(1): 153-169.

이기헌, 정효정, 송민. 2015. 문헌정보학 분야 핵심 학술지들의 가중 주제-방법 네트워크 분석.『한국문헌정보학회지』, 49(3): 457-488.

이원상, 손소영. 2015. 공간빅데이터 연구 동향 파악을 위한 토픽모형 분석. 『대한산업공학회지』, 41(1): 64-73.

이수상. 2016. 독후감 텍스트의 언어 네트워크 분석에 관한 기초연구,『한국도서관 · 정보학회지』,

47(3): 95-114.

이재연. 2016. 키워드와 네트워크 : 토픽 모델링으로 본『개벽』의 주제 지도 분석.『상허학보』, 46: 277-334.

정한조. 2015. 온톨로지와 토픽모델링 기반 다차원 연계 지식맵 서비스 연구.『지능정보연구』, 21(4): 79-92.

진설아 등. 2013. 트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구.『정보관리학회지』, 30(1): 285-302.

협성문화재단 (2013). 책 읽는 당신이 아름답습니다: 제2회 협성독서왕 선발대회 독후감수상작품집. 해성.

국한문 참고문헌의 영문 표기
(English translation / Romanization of reference originally written in Korean)

Beomil Kang, Min Song, Whasun Jho. 2013. "A Study on Opinion Mining of Newspaper Texts based on Topic Modeling." *Journal of the Korean Society for Library and Information Science,* 47(4): 315-334.

Dong Jin Park. 2010. "Condition of Writing Book Reports." *The Academy for Korean Language Education,* 84: 109-125.

Hanjo Jeong. 2015. "A Study on Ontology and Topic Modeling-based Multi-dimensional Knowledge Map Services." *Journal of Intelligent Information Systems,* 21(4): 79-92.

Ja-Hyun Park, Min Song. 2013. "A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling." *Journal of the Korean Society for Information Management,* 30(1): 7-32.

Juyoung An, Kyubin Ahn, Min Song. 2016. "Text Mining Driven Content Analysis of Ebola on News Media and Scientific Publications." *Journal of the Korean Society for Library and Information Science,* 50(2): 289-307.

Keehoen Lee, Hyojung Jung, Min Song. 2015. "Weighted Subject - Method Network Analysis of Library and Information Science Studies." *Journal of the Korean Society for Library and Information Science,* 49(3): 457-488.

Lee, Jae-yon. 2016. "Keywords and Networks - Exploring the Thematic Maps of Kaebyŏk through Topic Modelling." *Sanghur Hakbo - The Journal of Korean Modern Literature,* 46: 277-334.

Nahm, Choon-Ho. 2016. "An Illustrative Application of Topic Modeling Method to a Farmer's Diary." *Cross-Cultural Studies,* 22(1): 89-135.

Ra Yeon Kim, Eun Gyung Park. 2011. "A Study on the analysis of essential ingredient for Book Reports." *EDUCATION RESEARCH STUDIES,* 17: 17-30.

Seol A Jin, et al. 2013. "Topic-Network based Topic Shift Detection on Twitter." Journal of the Korean Society for Information Management, 30(1): 285-302.

Soo-Sang Lee. 2016. "A Preliminary Study on the Semantic Network Analysis of Book Report Text." *Journal of Korean Library and Information Science Society,* 47(3): 95-114.

So-Young Yu. 2015. "Combining Ego-centric Network Analysis and Dynamic Citation Network Analysis to Topic Modeling for Characterizing Research Trends." *Journal of the Korean Society for Information Management,* 31(4): 153-169.

Su Yeon Kim, Sung Jeon Song, Min Song. 2015. "Investigation of topic trends in computer and information science by text mining techniques." *Journal of the Korean Society for Information Management,* 32(1): 135-152.

Won Sang Lee, So Young Sohn. 2015. "Topic Model Analysis of Research Trend on Spatial Big Data." *Journal of the Korean Institute of Industrial Engineers,* 41(1): 64-73.

〈부록〉 16개 토픽과 주요단어 20개

토픽	비율	토픽명	주요단어(20개 추출)
T1	4.478	부모님	인정,부모,반복,어머니,이야기,자녀,막내,스트레스,기대,주연,환경,상담,글,삼각관계,에피소드,둘째,성격,이혼,책임감,후
T2	6.680	부모님 자신	아빠,자신,부모,감정,사실,저,공감,시간,친구,의미,현실,독,부부,충격,결혼,이름,전이,주변,구속,삼각관계
T3	3.381	친구	친구,눈물,때,인생,밤,아이,안데르센,언제,원망,행복,내면,고민,끝,먼로,모,미움,사례,여행,영향,차
T4	6.087	가족	가족,것,모습,마음,행복,자신,이유,행동,노력,변화,비밀,소통,얼마,하나,가정,고통,삶,경우,귀향,속
T5	7.536	사람들	사람,시절,이야기,사랑,자기,누구,얼굴,가족,가지,자신,나,건강,애,적,남,등,부족,용서,함,무엇
T6	5.787	외로움/불행	것,한,외로움,그것,불행,저,해결,집,과거,책,인간,일,날,누구,부담,불문,상황,죄책감,그동안,그때
T7	4.261	아들	너,아버지,아들,사랑,때,미안,아빠,인생,관계,마음,자신,마음속,여동생,함,길,문화,보호,삶,진심,친밀
T8	15.812	나와 책	나,책,내,엄마,적,생각,일,부모님,문제,가족,자신,사람,집,잘못,내용,우리,중요,모습,생활,배려
T9	4.529	아들의 문제	문제,대,아들,세,행복,사례,다짐,부,노력,시간,언제,그것,여자,친근,경우,대물림,말,아버지,이상,프로그램
T10	3.889	아버지/어머니	아버지,어머니,손,언제,자식,트라우마,시골,우리,도시,동네,약,해결,강정,결혼,나중,눈,사람,형제,곳,구성
T11	3.430	희생양/착취	희생양,착취,사회,돌담,문제아,서울대,하나,마음,매커니즘,집단,충격,가족,계급,공부,괴물,깨달음,대학,문제,문화,비
T12	6.270	문제점	나,시작,문제,선택,필요,관계,아내,치료,솔직,인정,삶,트라우마,진실,당신,독서,불안,세계,전,변화,얼마
T13	4.892	가족의 두얼굴	동생,얼굴,존재,사실,마음,가족,심리,언니,가슴,리,심리학,당연,로션,상황,순간,안정,양보,칭찬,부모님,분위기
T14	7.424	가족의 관계	가족,가정,대화,관계,이해,우리,화,원,반응,구성,학원,관심,선생님,숙제,학교,자식,편안,무관심,생기,아들
T15	3.727	감정	기억,부분,사소,감정,남자,사람,무의식,결혼,사례,인간관계,누군가,어른,미래,믿음,소통,시작,얼마,원북원,개입,겨냥
T16	12.415	일에 의한 상처	가족,말,상처,우리,생각,사랑,나,사실,책,어머니,치유,시절,세상,소중,표현,마음,여자,외동,폭력,사회