Korean Journal of English Language and Linguistics, **Vol 24, September 2024, pp. 979-1010**
DOI: 10.15738/kjell.24..202409.979

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS
ISSN: 1598-1398 / e-ISSN 2586-7474
http://journal.kasell.or.kr

AI 기반 영문학 텍스트 분석: 제인 오스틴의 소설 엠마를 중심으로 한 자연어처리 연구

오영교 (전남대학교)

Received: July 29, 2024
Revised: September 2, 2024
Accepted: September 23, 2024

Young-Kyo Oh
Lecturer, Dept. of Education,
College of Education,
Chonnam National University; 77,
Yongbong-ro, Buk-gu, Gwangju,
Korea, 61186
Email: 5young-kyo@hanmail.net

ABSTRACT

Oh, Young-kyo. (2024). AI-powered text analysis of English literature: A Natural Language Processing (NLP) study centered on Jane Austen's novel *Emma*. *Korea Journal of English Language and Linguistics* 24, 979-1010.

This study attempts to conduct data-driven text analysis on the text of Jane Austen's novel *Emma* (1815). Text analysis is a research method that utilizes Natural Language Processing (NLP) techniques to extract meaningful content and information from large-scale unstructured text data, and to discover new meaning and insights at the contextual level by considering the relationship between text and words. The main text analysis techniques include text network analysis, topic modeling, and sentiment analysis. In this study, we tried to analyze the text of the novel *Emma*, which is representative among English literary texts, according to NLP algorithms. To do so, we first analyzed the text of the novel *Emma* through term frequency (TF) analysis and term frequency-inverse document frequency (TF-IDF) analysis to determine the relative importance of words according to word frequency. Then, to examine the relationships between characters in the novel, we conducted co-occurrence and network centrality analysis through text network analysis. Next, we applied topic modeling using Latent Dirichlet Allocation (LDA) to classify *Emma*, a novel consisting of three volumes and 55 chapters, into four topics. Finally, sentiment analysis was conducted to calculate the degree of positivity and negativity for each volume to quantify the sentiment score. This study aims to help non-literature majors understand and appreciate English classical texts by objectively quantifying the literary content and character relationships inherent in 19th-century English fiction texts, and furthermore, to gain implications for English education in terms of effective English text comprehension education.

KEYWORDS

text analysis, *Emma*, network text analysis, topic modeling, sentiment analysis

1. 서론

최근 고전 문헌들의 디지털화가 확대되면서, 이에 따른 디지털 텍스트의 다양한 활용 방안에 대한 논의가 활발해지고 있다. 비정형 데이터인 텍스트 대상 분석 접근방법은 텍스트에 숨겨진 깊은 의미를 찾기 위한 이론적이고 방법론적인 발전을 가져왔으며, 비정형 데이터인 텍스트에 통계 및 수학적 모델을 적용하는 분석 알고리즘을 통해 새로운 의미를 발굴하려는 시도가 나타나고 있다. 특히, 영문학 분야에서는 영문 고전들을 디지털 형태로 전환하는 작업과 이 데이터를 어떻게 분석하고 활용할지에 대한 논의가 진행 중이다. 이 과정에서 언어학, 통계학, 인공지능 딥러닝 등의 다양한 학문적 지식을 융합하여 영문 고전 텍스트에서 필요한 정보와 지식을 추출하고자 하는 연구가 늘고 있다(원영선 2023, Smeet 2021).

테크놀로지에 기반을 둔 문학연구로서 초기 디지털 인문학(Digital Humanities)의 연구주제는 주로 말뭉치(Corpus) 연구나 단어빈도 처리를 통한 저자 판별이나 스타일 패턴 판독인 반면, 2000년대 이후에는 그 영역이 문학연구로 확대되고 있다. 특히, 서사구조나 장르, 문학사 등과 같은 문학연구에 초점을 둔 디지털 문학연구(Digital Literary Studies)가 디지털 인문학 내에서 중요한 영역을 담당하고 있다. 이에 따라 디지털 인문학 및 문학연구의 이론화와 전문화, 방법론 발달, 그리고 전통적 문학비평과의 융합 방법론에 이르는 다양한 노력이 이루어지고 있다(Jockers 2013, Moretti 2013, Ramsay 2011, Schreibman et al. 2008).

전통적 문학비평을 가까이 읽기로 정의하고 전산언어학(Computational Linguistics) 기법을 적용한 멀리서 읽기(Distant Reading)와 통합적 연구를 제안하는 입장이 있지만(Smeets 2021), 이 연구에서는 제인 오스틴(Jane Austen)의 소설 엠마(Emma)를 자연어처리 기법의 관점에서 텍스트 분석(Text Analysis)으로 접근한 선행연구와 맥을 같이한다(원영선 2023). 선행연구의 주요 내용은 소설 엠마의 배경인 하이베리(Highbury) 사회의 다양한 인간관계 양상과 텍스트 속에서 가려져 있던 베이츠 양(Miss Bates)의 특별한 사회적 위치와 역할을 객관적 수치로 규명하는 것이었다.

텍스트 분석은 언어학, 통계학, 기계 학습의 융합을 기반으로 한 자연언어 처리기술을 활용하여 대규모 텍스트 데이터를 분석하는 방법으로 알려져 있다. 이 분석기법은 빅데이터 중 온라인 네트워크상에서 제품과 서비스에 대한 고객 리뷰 또는 사회 이슈에 대한 대중의 댓글과 의견 및 여론 등을 분석하는 데 주로 적용되어 왔다(Chae et al. 2015, Cho et al. 2017, Kim and Kang 2018). 이 기법은 텍스트 내에 존재하는 키워드를 도출하거나 키워드 간의 관계를 규명하는 데 초점이 맞춰져 있다(Kim et al. 2016).

구체적으로 텍스트 분석은 NLP와 형태소 분석기술을 바탕으로 하여 비정형화된 다량의 텍스트에서 유의미한 단어를 추출하고, 이들 단어의 빈도수를 분석하여 문맥 수준의 의미를 찾아내는 방법이다. 또한 유의미한 단어를 추출한 것을 기초로, 다른 다양한 정보와 연계하여 텍스트를 분석함으로써, 새로운 관계와 정보 그리고 지식을 찾는 기법이다. 이 연구는 소설 엠마를 데이터 분석가나 영어교육자 입장에서 이해하기 위해 작품 속 등장인물들의 관계와 역할, 주제, 감정 플롯을 파악하기 위해 텍스트 네트워크 분석(Text Network Analysis), 토픽 모델링(Topic Modeling), 감성분석(Sentiment Analysis)과 같은 NLP 기법을 활용하였다.

먼저, 텍스트 네트워크 분석은 소설 내에서 인물 간의 관계를 노드(Node)와 링크(Link)의 형태로 시각화하여 인간관계의 도식적 구조를 그래프로 표현하는 방법이다. 이를 통해 등장인물 간의

연결성과 그 관계성에서의 중요도를 수치화한다. 다음으로 토픽 모델링은 문서 내 숨겨진 주제들을 찾아내기 위해 텍스트 데이터를 분석하는 기법으로, 특히 LDA 방법을 사용하여 소설의 각 장을 대표하는 키워드를 도출하고, 이를 통해 소설의 주요 테마와 구조를 파악한다. 이 기법은 문학작품의 다양한 주제와 구조적 요소를 파악하는 데 유용하다. 마지막으로 감성분석은 텍스트 내의 감정적 요소를 분석하여, 각 텍스트가 표현하는 감정의 극성(긍정 또는 부정)을 평가한다. 이 방법은 소설 속 발화자나 서술자의 감정적 태도를 점수로 수치화하고, 이를 통해 텍스트가 전달하려는 감정적 색채와 의도를 파악하는 데 유용하다.

이 연구에서는 이러한 다양한 최신 NLP 기법들을 적용하여 소설 엠마에서의 등장인물 간 관계, 역할 중심과 서사구조, 그리고 감정의 변화를 포괄적으로 이해하고 해석하고자 하였다. 이를 통해 기존의 영문학 비평에 인공지능 딥러닝 기법을 보완하여 영문학 텍스트 이해와 분석에 대한 깊이를 넓히고자 하였다. 이 연구는 디지털 인문학 측면에서 NLP 방법론을 활용하여 영어 문학 텍스트를 체계적으로 분석하고, 텍스트 이해에 대한 새로운 통찰을 제공하여 영문학 텍스트와 영어교육 연구의 접점을 찾고자 시도하였다.

이 연구에서는 다양한 NLP 기법들을 영문학 정전 텍스트인 제인 오스틴의 소설 엠마에 적용하여, 기존에 각기 개별적으로 진행되었던 텍스트 분석기법을 하나로 통합하고자 하였다(Elson et al. 2010, Heuser and Le-Khac 2012, Mimno 2012, Rhody 2012). 즉, 영문학 대표적인 소설 엠마 텍스트에 대해 주요 NLP기법인 네트워크 텍스트 분석, 토픽 모델링 및 감성분석을 실시하여, 영문학 작품의 주요 주제와 인물 간의 관계 발전, 작품 분위기를 체계적으로 탐색하고자 하였다. 이처럼 한 작품을 대상으로 한 다각적 텍스트 분석 접근은 텍스트 이해를 단순히 인간의 직관에 의존하는 것에서 벗어나, 객관적으로 수량화함으로써 인공지능 기술을 영어 텍스트 이해 및 학습에 적용하는 연구에 실질적인 기여를 할 수 있을 것으로 기대된다. 이러한 접근은 인공지능 기술의 발전을 통해 얻은 데이터 기반 분석 방법론이 학문적 연구뿐만 아니라 실제 영어교육 현장에서도 유용하게 활용될 수 있음을 시사한다.

2. 이론적 배경

2.1 영어 텍스트로서 소설

텍스트 분석은 대규모 텍스트 데이터에서 일정한 패턴과 특징 등을 추출함으로써 새로운 정보나 유용한 지식을 발굴하는 작업이다. 이 때문에 광산에서 광물을 채굴하는 것에 빗대어 데이터 마이닝(Data Mining)이라고도 한다. 최근에는 해석학적 방법론을 주로 사용하는 인문학에서도 텍스트 분석이 적용되고 있다. 이는 디지털 테크놀로지와 인문학의 융합 분야인 디지털 인문학으로 연결된다. 이 연구에서는 제인 오스틴의 소설 엠마를 멀리서 읽기로서의 NLP 기법을 통한 통계적 분석과 해석을 시도하였다. 멀리서 읽기는 새로운 문학 연구법으로서, 텍스트 자체의 문학성에 대한 비평 대신에 수량적 자료들의 분석을 통한 형식주의 읽기 특징을 나타낸다(Moretti 2012).

소설 문학작품에서 등장인물들의 관계를 바탕으로 네트워크의 분석을 한 연구가 있다(Elson et al.

2010). 즉, 소설의 등장인물들에 대한 인간관계 네트워크를 구축하고 이 연결망의 특성을 분석하였다. 등장인물 네트워크는 대부분 단어 수준의 텍스트를 기반하여 수행된다(Hassan et al. 2012). 이 작업은 작가의 단어 사용의 스타일과 어휘 패턴에 집중되고 있으며, 소설이 현실 세계를 객관적으로 표현하기 위한 양식으로 문학적 이론을 검증하거나 평가하는데 활용될 수 있다.

주요 선행연구로 원영선(2016)의 연구에서는 주인공 엠마, 하이베리 공동체, 계급, 관계망, 베이츠와 페리의 숨겨진 역할을 규명하고자 하였다. 소설 엠마의 작품 속 등장인물의 중심 무대인 하이베리 사회는 지리적 이동이나 공간 확장이 적어 사회관계망으로 분석하기에 적합하다. 연구에서 소셜네트워크 분석의 주요 대상은 하이베리 공동체의 계급적 인물 관계망, 그리고 주인공 이외의 주변 인물들, 특히 베이츠 양과 하이베리의 공동체 사회 간의 특별한 역할 관계 탐색이었다.

디지털 문학연구로 분류될 만한 초기 연구 사례로 셰익스피어 작품에 기초적인 디지털 분석을 실험한 모레띠의 시도(Moretti 2013)가 대표적이며, 최근에는 디지털 문학연구의 방법론적 방향이 단일 작품보다는 소설 장르와 작가 및 작품 간 관계망 등으로 그 대상을 넓히고 있다(Fischer and Skorinkin, 2021). 이외에도 통계학자들이 멀리서 읽기로서의 텍스트 마이닝을 적용한 국외 연구로는 제인 오스틴의 작품들에서 남성 인물과 여성 인물에 의해 표현된 단어들이 만들어내는 감성을 비교 분석하여 작품 속에서 젠더의 역할을 계량화한 연구가 있다(Silge and Robinson 2017). 국내에서도 영문학 멀리서 읽기 선행연구로 셰익스피어의 텍스트에 코퍼스를 도입한 연구(하명정 2013)와 아동 모험 소설에 대한 코퍼스적 문체론 분석을 시도한 연구가 있다(최은샘, 정채관 2021).

영문학 텍스트는 비정형화된 데이터 구조이기 때문에 텍스트 분석을 위해서는 통계학, 데이터 마이닝 그리고 NLP의 학제적 연계가 필요하다. 실제로 NLP 기법에 의해 영문학 텍스트 데이터는 구조화 및 정제, 그리고 텍스트의 계량화를 통하여 분석이 진행된다. 이 연구는 소설 엠마 텍스트의 네트워크 구조와 토픽 및 감성을 수치화하고, 오스틴이 설정한 작품 속 텍스트가 나타내는 양적 특성을 인공지능 기반 통계학적으로 이해하고자 한다. 이 연구는 거시적인 관점에서 여러 문학 텍스트에 대한 분류와 목록화에 초점을 둔 선행연구들과 달리, 미시적인 관점에서 하나의 문학 텍스트 내의 플롯에 대한 구조화에 집중하고자 하였다. 즉, 소설 엠마 대한 미시적 텍스트 분석을 통해 작가 오스틴의 문학 텍스트를 계량화하고 특성을 추출하여 인공지능 통계학적 관점에서 재구성하는 것이다. 이러한 분석을 통해 소설에 대한 질적 연구를 양적연구와 접목시키고, 디지털 인문학의 가능성을 모색하는 것이 이 연구의 의의 중 하나다.

2.2 텍스트 분석기법

소설 속 등장인물들의 관계를 규명하고 이를 통해 작가의 서술 방식을 이해하기 위해서 텍스트 네트워크 분석이 사용된다. 텍스트 네트워크 분석은 현실 사회에서 관찰될 수 있는 인간관계의 도식적 구조를 텍스트 속에서 발견하려는 접근방식으로서, 텍스트 데이터에 존재하는 키워드 간 관계 네트워크를 그래프 혹은 수치로 표현하는 기법이다. 텍스트 네트워크에서 관계의 중심을 노드(Node), 그들 간의 상호관계를 링크(Link)라고 부른다. 즉, 개인이나 집단을 점으로, 그들의 연결 관계를 선으로 개념화하여 점과 선이 형성하는 네트워크의 수학적·시각적 패턴을 도출하고, 이들이 만들어 내는 관계 양상을 네크워크로 구조화하여 개인 및 집단의 관계방식과 그 영향력을

측정하는 분석 방법이다(Borgatti et al. 2009). 이처럼 텍스트 네트워크 분석은 노드와 엣지가 형성하는 연결 관계에 대한 개념을 바탕으로 사회적 관계의 구조를 모형화하여 탐색하는 작업이다(Scott 2012).

실제 사회 속 인간관계와 같은 연결성을 텍스트 구조속에서 발견하기 위해서는 텍스트에 존재하는 데이터들의 특성을 추출하고 그들 중 관계의 중심이 될 수 있는 키워드를 찾아냄으로써 가능하다. 네트워크 분석은 키워드 간 관계를 파악하기 쉬우며 소설을 비롯한 사회 이슈에서 인물들 간 관계를 파악하는 연구에서 자주 활용되고 있다. 선행연구에서는 소설 데이터 속 등장인물들의 거리를 계산하여 사회 연결망을 분석하였다(Park et al. 2013). 네트워크는 텍스트 내에서 키워드들 간 동시 발생(Co-occurrence) 정도를 계산함으로써 구현이 가능하다. 이 연구는 소설 엠마에 등장하는 핵심 등장인물을 선정하여 그들 간 텍스트 내 동시 발생 빈도수를 매트릭스(matrix)로 계산하였다. 네트워크의 정적인 구조를 분석한 선행 연구들은 리어왕과 헴릿 등과 같은 셰익스피어 작품에 등장하는 인물들의 네트워크(Stiller et al. 2003), 그리스-로마 신화 네트워크(최연무 2004), 박경리의 대하소설 토지에 나오는 인물들에 대한 네트워크(김상락 2005) 등이 있다.

텍스트 네트워크 중심성은 관계망 속에서 누가 더 중요한지를 수량화 한다(Tsvetovat and Kouznetsov 2011). 중심성은 중요성과 권력과 같이 다양한 해석이 가능한 개념이므로 이는 다시 연결정도 중심성(Degree Centrality), 근접 중심성(Closeness Centrality), 매개 중심성(Betweenness Centrality)과 같은 하위개념으로 나뉜다. 더불어 위세 중심성으로 불리는 에이겐벡터 중심성(Eigenvector Centrality) 개념이 함께 사용되기도 한다(정성훈 2014).

텍스트 네크워크 분석 결과를 이해하기 위해서 먼저 연결정도 중심성은 노드에 연결된 엣지의 총 개수로 계측하여 중심성을 설명하는 개념으로, 노드의 구조적 지위에 대한 가장 간단하면서도 효과적인 지표이다. 즉, 가장 많은 연결 관계를 가진 노드는 네트워크 내에서 가장 많은 노드와 알고 지내는 가장 인기 있는 유명인사(Tsvetovat and Kouznetsov 2011)의 지위를 갖는다. 하지만 연결정도 중심성이 노드에 직접 이어진 엣지 만을 고려하기 때문에 연결 노드가 다른 노드로 확장되는 정도를 알 수 없으므로, 전체 네트워크 내의 상대적 중요성을 드러내는 데 한계가 있다(곽기영 2014).

연결 정도 중심성을 보완한 근접 중심성은 네트워크 내의 간접 연결 관계까지 고려하여 전체 네트워크 속 한 노드와 나머지 모든 노드 사이의 최단 경로를 계량한다. 노드가 얼마나 많은 노드에 연결되는가를 보여주는 연결정도 중심성이 노드의 활동성 지표라면, 근접 중심성은 직접 연결되지 않고도 모든 노드와 가깝게 연결되는 정도를 표현하므로 노드의 독립성 지표라고 할 수 있다(곽기영 2014). 즉, 근접 중심성은 빠르고 긴밀한 연결 관계를 통해 많은 정보를 얻고 널리 퍼뜨릴 수 있는 구조적 지위를 표현한다.

다음으로 매개 중심성 역시 직접 연결되지 않은 노드를 포함한 네트워크 전체를 고려하지만, 서로 연결되어 짝을 이루는 노드 사이의 최단 경로에 위치하는 횟수를 계측한다는 면에서 근접 중심성과 구별되며 보완적 관점을 제공한다. 매개 중심성이 높은 노드는 네트워크 내 노드들 사이에서 소통의 병목 역할을 하거나 네트워크 내 구성원을 이어주는 공동체 다리 역할을 수행할 수 있는 구조적 지위를 갖는다. 이 연구에서는 연결 정도, 근접, 매개, 에이겐벡터 중심성 분석지표를 바탕으로 소설의 등장인물이 가진 지위와 관계를 논의한다.

다음으로 토픽 모델링은 방대한 텍스트를 의미 있고 해석 가능한 언어 단위로 결합하는 텍스트 마이닝 기법으로, 텍스트에 존재하는 맥락을 단서를 통해 발견하고 유사 단어들을 군집화함으로써 구현된다. 이러한 특성으로 인해 문서 간 정보를 분류하거나 특정 사회 이슈나 주제를 구분하고 요약하는 연구에 주로 활용된다. 소설의 토픽과 플롯은 등장인물들이 부여된 성격과 주어진 배경에 따라 전개되는 담화의 연속이라고 했다(Choi and Yoo 2014). 텍스트를 접하는 독자는 스토리를 추측하고 다음에 나올 사건에 대해 관심을 집중하게 된다. 소설 속 등장인물의 활동이 텍스트와 작가, 독자에게 미치는 영향이 중요하다고 볼 수 있으며, 이는 작가가 등장인물을 통해 드러내고자 하는 고유한 텍스트 해석이 가능하다는 것을 의미한다.

마지막으로 감성분석은 텍스트 마이닝의 분석방법으로서, 텍스트 내에 포함되는 발화자의 주관적인 감정 및 태도를 발견하고 이를 객관적 수치와 도식으로 표현할 수 있는 분석기법이다(Pang and Lee 2008). 감성분석은 텍스트에 포함되어 있는 감정 어휘들의 극성값(Polarity)을 측정하여 감성 점수를 부여함으로써 텍스트 발화자의 성향과 태도 등 주관적인 정보를 정확하게 추출할 수 있다. 이를 통해 인지 및 정서적 차원에서만 머물렀던 화자의 감정과 성향을 인공지능 프로그램이 인식할 수 있는 자연어로 기법으로 수치화하려는 시도가 이루어지고 있다(Cho et al. 2016).

3. 연구방법

3.1 자료 수집 및 분석절차

인문학의 텍스트 분석 활용은 언어학, 문학 등 텍스트 데이터가 이미 축적된 분야에서 활발하다. 대규모 인문학 텍스트가 데이터 분석기술의 적용이 가능하도록 디지털화 작업이 진행되었다. 인문학 텍스트를 디지털로 변환하는 대표적인 구텐베르크 프로젝트에서는 셰익스피어나 제인 오스틴 등의 문학작품을 비롯한 59,000여 개의 고전문헌이 전산화되어 전자책으로 제공된다. 이러한 인문학 텍스트 말뭉치(corpus)의 데이터베이스를 텍스트 분석을 통해 연구하는 시도가 이루어지고 있다(Chen and Chang 2019). 이러한 텍스트 분석은 대규모 인문학 프로젝트뿐만 아니라 개별 문학 연구에도 적용되고 있다. 소설 엠마에 대한 텍스트 분석을 위해 데이터 수집→데이터 전처리→데이터 분석과정 순서로 연구를 진행하였다. 분석 도구로 Python 프로그래밍 언어를 사용했고, 데이터 분석과정에서는 소설 엠마의 핵심적인 내용을 도출하기 위해 TF 및 TF-IDF(Term Frequency-Inverse Document Frequency)분석, 텍스트 네트워크 중심성 분석, 토픽 모델링, 감성분석 순서로 텍스트마이닝 분석기법을 적용하였다. 특히, TF-IDF의 경우 단순 빈도보다는 특정 문서에서 더 자주 반복되는 단어에 상대적 가중치를 부여하는 분석기법이다.

3.2 Emma 텍스트 전처리

텍스트 자료는 문장으로 구성된 구조를 가지며, 문장 구조로는 분석의 정확성 및 구체성이 구현되지 않는다. 따라서 본 분석에서는 텍스트 자료의 전처리 작업으로 네 가지의 단계를 진행하였다. 먼저 대문자를 모두 소문자로 변환하는 정규화(Normalization)를 거쳤으며, 유의미한

분석 결과를 얻기 위해 문자에 대한 공통된 기준을 제시하여 일반성을 규정하였다.

표 1. 단어 수정 및 삭제 목록(Mapping Table)

수정(단어의 일부 삭제 및 수정)	
a. 소문자 변경	c. 의미 없는 단어 삭제
TRUE→ true	-_she_, _her_, _you_
b. 잘린 문자 수정	d. 특수문자제거
es → yes	_must_ → must
...	...

표 2. Stop Words

be	it	sir	in	them	that	your	are
have	her	me	my	or	thi	with	ther
do	to	th	she	the	they	at	...
and	sh	him	is	on	don	us	
dear	you	for	of	he	wa	upon	

두 번째 단계로 표 1과 같이 문장 및 문단으로 구성된 데이터 구조를 최소 의미 단위로 끊는 토큰화(Tokenization)를 시행하였다. 형태소 추출 시 파이썬 nltk 라이브러리를 사용하여 명사, 형용사, 동사를 추출하는 작업을 실시하였다(Karl et al. 2015). 세 번째로 표 2와 같이 불용어 제거 단계(stop words)를 통해 분석 결과에 유의미하지 않은 단어와 잡음(Noise)을 제거하였다.

마지막 단계로 어간 추출(Stemming)과 표제어 추출(Lematization)을 시행했다. Stemming시에는 nltk 라이브러리의 PorterStemmer 모듈을 활용하였고 Lemmatization시에는 ntlk의 WordNetLemmatizer 모듈을 사용하였다. 분석 과정에서 중복적으로 계산될 것을 고려하여 공통적으로 등장인물들의 성과 이름을 하나의 이름으로 통일하는 과정을 거쳤다. 예를 들어, mr, miss, mrs는 이름과 함께 사용, 등장인물은 harriet, harriet smith, miss smith를 harriet_smith 형식으로 통일하였다.

3.3 데이터 분석

3.3.1 단어 빈도분석(TF) 및 빈도-역문서 빈도분석(TF-IDF)

텍스트 전처리 과정을 거친 후 추출된 단어들에 대한 우선순위를 매겨 특정 단어의 출현 횟수를 기준으로 중요도를 판단하는 기법이 단어 빈도분석(Term Frequency: TF)이다. 텍스트 마이닝 분석에서 단어 빈도분석은 텍스트 분석 시 데이터의 흐름과 이해를 위해 가장 먼저 기초 분석 자료로 활용되는 기법이다. 즉, 단어 등장 빈도에 기반한 분석은 전체 문서 내에서 특정 단어의 빈도를 나타낸다. 문서 내에 특정 단어가 출현하는 횟수를 나타내며 이 수치값이 클수록 문서에서 자주 사용하는 단어임을 의미한다(장민서, 오수진, 김응모 2018). 높은 빈도를 나타내는 단어는 연구주제와 관련된 함축된 의미를 내재하고 핵심 단어(키워드)로 작용할 가능성이 높다(이승은 2022).

TF-IDF는 여러 문서로 이루어진 문서 군이 있을 때 특정 문서 내에서 특정 단어가 얼마나

중요한지를 나타내는 통계적 수치로 형태소 분석이 많이 활용된다(Oh et al. 2017). TF-IDF값은 전체 문서에서 해당 단어가 출현한 문서 수는 적지만 특정 문서에는 집중적으로 많이 언급되는 경우 값이 증가한다. 따라서 TF-IDF값이 클수록 상대적으로 핵심적인 의미와 내용을 나타낼 가능성이 크다(Chung et al. 2019). 이 연구에서는 텍스트 사전처리 후 Python의 scikit-learn 패키지의 CountVectorizer, TfidfVectorizer 모듈을 활용하여 TF와 TF-IDF를 도출하였다.

3.3.2 텍스트 네트워크 단어 중심성 분석

단어 중심성 분석(Centrality Analysis)은 단어 간의 연결관계 방향과 강도를 분석하는 것으로 어떤 단어가 영향력이 큰지를 알 수 있는 지표이며 대표적으로 연결 중심성(Degree Centrality), 근접 중심성(Closeness Centrality), 매개 중심성(Betweeness Centrality), 위세 중심성(Eigenvector Centrality) 분석방법으로 나뉜다(홍주현, 나은경 2015). 연결정도 중심성은 전체 노드의 수와 실제로 연결되어 있는 노드 수의 비율을 의미하며, 근접 중심성은 노드가 연결될 수 있는 최단거리 합의 역수이며 한 노드가 네트워크의 중앙에 위치해 있는지를 나타낸다(Kang and Lee, 2015). 매개 중심성은 한 노드가 다른 노드와 네트워크를 구축하는데 있어 중개자 역할을 어느 정도로 수행하는지를 나타내며, 위세 중심성은 영향력이 큰 노드와 많은 연결한 비율을 의미한다(Sawng and Lee 2018). 본 연구에서는 연결 중심성, 근접 중심성, 매개 중심성, 위세 중심성을 분석하기 위해 Ucinet 6.775를 이용하였다.

3.3.3 동시출현 단어 네트워크 분석

동시 연결 네트워크 분석은 특정 용어와 동시에 출현한 단어들을 추출하여 연관어 네트워크를 구축하여 분석하는 방법론이다(Kang 2010). 데이터의 규모가 큰 경우 의미를 추출하는 데 시간과 노력을 아끼기 위해서 주로 사용된다. 본 연구는 Ucinet 6.775를 사용하여 네트워크 그래프를 시각화하였다.

3.3.4 토픽 모델링

LDA 기반의 토픽 모델링 기법을 사용하여 소설 엠마와 관련 데이터를 분류할 수 있는 적정 토픽주제를 도출하였다. LDA는 유의미한 주제를 군집화할 수 있는 방법으로 비정형 텍스트 분석에 주로 활용되며(박주섭, 홍순구, 김종원 2017). LDA기반 토픽모델은 주어진 자료나 텍스트 집합에 대하여 각 텍스트에 어떤 주제들이 존재하는지를 보여주는 확률 모형이다. 즉, 단어가 모이면 주제(토픽)가 되고, 주제가 모여서 특정 텍스트를 구성한다. 분석 대상이 되는 모든 텍스트 집합은 일련의 단어 집합으로 임베드되며 단어는 여러 잠재된 주제를 나타내는 실제 관측치이다. 따라서 관찰된 각 단어는 주제를 대표하고, 주제의 집합이 문서가 되기 때문에 각 단어가 어떤 잠재된 주제에 속하는지를 추출해 냄으로써, 주제의 집합인 각각의 텍스트를 분류할 수 있는 것이다. 이를 통해 연구자는 잠재변수를 활용해 각각의 문서에서 드러나지 않은 숨겨진 의미를 찾을 수 있다.

좀 더 구체적으로 살펴보면, 토픽 모델링의 기본 가정은 세 가지이다(Blei et al. 2003). 첫째,

하나의 텍스트는 여러 개의 주제를 포함할 수 있다. 둘째, 주제에는 여러 개의 단어가 포함될 수 있다. 셋째, 하나의 텍스트에 사용된 각 단어는 어떤 주제에 포함된다. LDA는 디리클레 분포를 통해 주제별 단어의 분포와 각 텍스트 별 주제의 분포를 추정하는 확률적 모형이다. 잠재 디리클레 할당이 실제로 하는 작업은 각 텍스트에 나타나는 단어들을 관측하고, 이들 단어가 어떤 주제가 속해 있는지 숨겨진 정보를 추론하는 것이다. LDA의 구조는 그림 1과 같다.

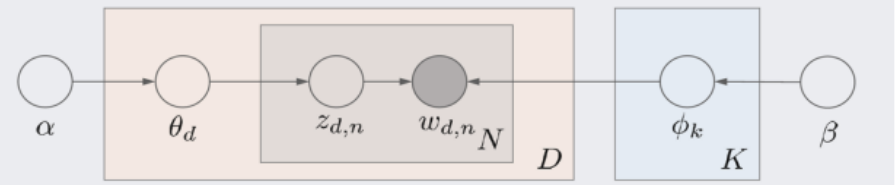

그림 1. 잠재 디리클레 할당 구조(Blei, 2009)

*D*는 분석 대상이 되는 전체 텍스트의 수를 나타내며, *K*는 텍스트 집합이 가지고 있는 전체 주제의 수를 가리킨다. α(초기 파라미터 값)는 디리클레 파라미터로서 텍스트 집합 속에 어떤 단어들이 얼마나 분포되어 있는지를 나타내는 지표이고, β(조정된 파라미터 값)는 주제 하이퍼파라미터로서 텍스트 집합 속에 어떤 주제가 얼마나 분포되어 있는지를 보여주는 지표이다. 디리클레 분포를 따르고 있음을 가정하고 α와 β는 각각 ϕ_k(토픽에 속한 단어)와 Θ_d(문서 내 토픽의 디리클레 분포)에 영향을 준다. α와 β의 값이 클수록 주제의 분포를 비슷하게 만들며, 값이 작아질수록 특정한 주제가 두드러지게 나타나게 된다고 알려져 있다.

위의 구조는 α와 β에서 시작해서 텍스트의 단어를 생성해 나가는 과정을 나타낸 것이다. 그런데 우리가 실제로 관찰해서 측정할 수 있는 값은 $\omega_{d,n}$(관측 가능 값)뿐이다. 따라서 LDA는 관측치인 $\omega_{d,n}$를 잠재변수로 하여 위의 과정을 역으로 추론하는 과정을 수행한다. 이 과정에서 주제의 단어 분포와 텍스트의 주제 분포의 결합 확률이 높으면 높을수록 좋은 모델이 될 수 있다. 즉, 실제로 관찰이 가능한 텍스트 내의 단어들을 이용하여 사후(posterior)확률을 최대로 만드는 $Z_{d,n}$(특정 단어가 속한 토픽 번호), ϕ_k, Θ_d를 계산하는 것이 잠재 디리클레 할당이 하는 추론이다.

이처럼 LDA 분석을 통해 단어에 대한 분포를 확인하고, 이를 통해 *K*개의 잠재 토픽을 추출할 수 있다(그림 1). 디리클레 분포는 *K*차원의 실수 벡터 중 벡터 요소가 양수임을 만족하며 모든 요소의 합이 1인 확률값으로 연속확률분포를 나타낸다. 이 연구에서는 Python의 gensim과 sklearn 라이브러리를 사용하여 토픽 모델링을 실시하였다. 적정토픽 수 선정 시 토픽 주제 간 중복 범위와 혼란도(Perplexity)와 일관성(Coherence) 지표를 고려하여 가장 적합한 토픽 수를 선정하였다.

3.3.5 감성 단어 추출

감성분석 기법은 특정 주제에 대해 사람들이 느낀 감성에 대한 텍스트 데이터를 텍스트 마이닝으로 분류하여 분석하는 기법을 의미한다. 또한 텍스트에 내재된 감정, 반응, 평가, 태도 등의 주관적인 정보나 의견을 빅데이터 분석기법을 통해 분류하거나 수치화하여 의미 있는 정보를 찾아내는 기법이다(Medhat et al. 2014). 기존의 감성분석은 전문가와 심층 면접 인터뷰 등을

통해 의견을 구하는 분석이 위주였다. 하지만, 최근 텍스트 마이닝 기법의 발전에 따라 기존 기법의 공간적·시간적인 한계가 사라지면서 대규모의 텍스트 데이터에 대한 분석이 가능해졌다. 이러한 이유 때문에 감성분석은 마케팅, 금융, 정치 등 다양한 분야에서 활용되고 있다. 이 연구에서는 수집된 데이터 전처리 후 NLTK(Natural Language Toolkit) 라이브러리를 통해 형용사 형태소를 추출한 후 긍정과 부정 키워드에 대한 세부감성과 관련된 긍정 및 부정 단어 상위 30위의 감성 어휘 단어를 권(Volume) 별로 추출하였다. 이 연구에서는 소설 엠마의 텍스트 분석을 위해 파이썬(Python) 영문 자연 언어 처리 패키지인 NLTK의 Sentiment Intensity Analyzer와 vader_lexicon 감성 사전을 이용하여 감성분석을 실시하였으며 텍스트의 극성 값을 판단하여 긍정과 부정 카테고리로 분류했으며 감정점수로 계량화하였다.

4. 연구 결과

4.1 소설 엠마 텍스트 빈도 분석

제인 오스틴의 소설 엠마에서 자주 사용된 단어들의 목록을 통해 소설의 주요 테마와 인물, 그리고 이야기의 전개와 개요를 파악할 수 있다. 각 단어의 빈도는 소설에서 해당 단어가 얼마나 중요하게 사용되었는지를 나타내며, 이는 소설의 텍스트 분석을 위한 기초 내용적 통찰을 제공한다. 표 3은 소설 엠마에 나타난 단어와 그 빈도이다. 단어 빈도순으로 살펴보면, 'emma', 'say' 'think', 'harriet', 'jane', 'know', 'good', 'thing', 'make' 순으로 많이 사용되었는데 그 중 빈도가 높은 단어는 주인공 'emma'였다.

구체적으로 소설의 주인공인 'emma'는 1,030회 언급되어 가장 빈번한 단어로 등장한다. 이는 엠마가 소설의 중심 인물로서 매우 중요한 역할을 하며, 소설의 다양한 사건과 관계의 핵심에 있음을 강조한다. 다음으로 자주 등장하는 단어는 'say'(853회)와 'think'(650회)이며, 이는 소설이 대화와 내적 사고를 통해 진행됨을 함축한다. 엠마와 다른 인물들 간의 대화와 엠마의 사고 과정이 소설에서 상당히 중요한 부분을 차지하고 있음을 암시한다.

표 3. Word Count Term Frequency (TF)

순위	단어(term)	빈도	순위	단어(term)	빈도	순위	단어(term)	빈도
1	emma	1,030	41	young	194	81	pleasur	118
2	say	853	42	bodi	190	82	meet	116
3	think	650	43	mrs	185	83	parti	115
4	harriet	563	44	tell	184	84	idea	115
5	jane	530	45	find	184	85	visit	114
6	know	530	46	talk	182	86	oblig	114
7	good	463	47	sure	182	87	sort	112
8	thing	456	48	mr_weston	176	88	campbell	111
9	make	416	49	mean	174	89	someth	111
10	mr	410	50	way	169	90	felt	111
11	come	392	51	hope	165	91	oh	111
12	knightley	390	52	hartfield	160	92	ladi	111

13	go	375	53	first	160	93	return	110
14	littl	350	54	walk	157	94	well	110
15	such	348	55	mrs_elton	154	95	use	110
16	give	348	56	love	152	96	comfort	110
17	see	344	57	like	151	97	few	110
18	great	323	58	believ	150	98	subject	110
19	much	313	59	woman	147	99	sit	109
20	own	304	60	begin	146	100	call	109
21	look	303	61	letter	145	101	miss_bates	109
22	time	301	62	word	141	102	saw	108
23	mrs_weston	293	63	woodhous	140	103	understand	107
24	other	276	64	mani	137	104	bring	106
25	take	257	65	poor	137	105	hour	106
26	friend	256	66	mind	135	106	morn	104
27	feel	256	67	appear	134	107	no	104
28	noth	248	68	leav	133	108	smile	102
29	day	239	69	room	131	109	same	102
30	hear	236	70	marri	131	110	pass	101
31	man	235	71	moment	131	111	allow	100
32	mr_elton	234	72	home	129	112	deal	98
33	more	225	73	last	129	113	natur	98
34	speak	216	74	highbury	125	114	robert_martin	98
35	wish	216	75	manner	125	115	bad	96
36	seem	211	76	suppos	121	116	attent	96
37	happi	206	77	doubt	121	117	present	96
38	frank	205	78	get	121	118	pleas	96
39	father	203	79	kind	120	119	hous	96
40	want	201	80	place	119	120	even	95

'harriet_smith'는 563회 언급되어 엠마의 친구이자 주요한 서브 인물 중 하나임을 나타낸다. 엠마와 해리엇의 관계는 소설의 주요한 줄거리 중 하나로, 엠마가 해리엇의 사랑과 결혼을 조종하려 시도하는 것이 문제를 일으키며 이야기를 전개시킨다. 또한, 'jane', 'knightley', 'mrs_weston', 'mr_elton', 'woodhouse', 'mr_weston', 'mrs_elton', 'campbell', 'miss_bates', 'robert_martin'과 같은 다른 중요 인물들의 이름도 자주 언급된다. 이는 각 인물이 엠마의 삶과 하이베리 사회 내에서 중요한 역할을 수행함을 보여주며, 이들 각각의 인물이 소설의 갈등과 테마 발전에 중요한 기여를 하는 것을 나타낸다. 소설에서 자주 등장하는 단어 'marriage', 'love', 'friend', 'woman', 'man' 등은 소설의 주된 테마인 결혼, 사랑, 우정, 젠더 역할에 관한 사회적 규범과 기대를 반영하고 있다. 이러한 테마는 엠마와 다른 인물들의 인간관계와 개인적 성장을 통해 탐구되며, 오스틴이 19세기 초 영국 사회의 결혼과 성 역할에 대해 다루고 있음을 보여준다. 이 분석을 통해 소설 엠마의 텍스트에서 중심적으로 다루어진 주제와 인물 관계, 사회적 상호작용을 객관적으로 이해할 수 있다. 이처럼 단어 빈도분석은 소설 전체의 주요 테마와 인물의 특성, 그리고 작가의 의도를 개관하는데 도움을 준다.

다음으로 TF와 IDF를 결합한 표 4의 TF-IDF 결과값은 특정 문서 내에서 자주 등장하지만 전체 문서 집합에서는 드물게 나타나는 단어를 높게 평가한다. 따라서 제인 오스틴의 소설 엠마에서 각 단어의 상대적 중요도와 특징을 정확히 반영한다. 높은 TF-IDF 점수를 가진 단어는 해당

텍스트에 특별히 중요하거나 의미가 있다는 것을 나타낸다. 'emma'는 가장 높은 TF-IDF 점수(9.02)를 가지며, 이는 엠마 캐릭터가 소설에서 중심적인 역할을 하며 자주 언급됨을 의미한다. 다음으로 'say'와 'jane', 'harriet'이 높은 점수를 보여주는데 표 3의 단순 TF와 달리 'jane'이 'harriet'보다 상대적으로 높은 점수를 기록했다. 이는 소설에서 대화가 활발하게 이루어지며, 특정 문맥에서 제인 캐릭터가 해리엇보다 주요 인물로서 자주 등장하는 것을 시사한다. 또한 'marriage'의 경우 TF에서 70위에서 66위로 상승하였다. 이는 특정 에피소드에서의 빈도가 상대적으로 높다는 것을 의미한다.

반면, 상대적 순위가 낮아진 단어인 'great', 'hope', 'pleasure', 'friend' 등으로, 이들 단어는 소설 전반에 걸쳐 일반적인 주제로 등장하지만 특정 맥락에서는 반복적으로 사용되지는 않는다는 것을 암시한다. 예를 들어, 'friend'의 경우 26위에서 33순위로 내려가서 소설 전반에 걸쳐 다양한 관계와 갈등 속에서 빈도는 높지만, 특정 장면이나 대화에서 상대적 빈도는 낮다는 의미이다.

표 4. Word Count Term Frequency-Inverse Document Frequency (TF-IDF)

Rank	Word	tf-idf	Rank	Word	tf-idf
1	emma	9.02	51	love	1.76
2	say	7.42	52	hartfield	1.75
3	jane	6.45	53	mr_woodhouse	1.75
4	harriet	6.11	54	mean	1.73
5	think	5.65	55	find	1.70
6	know	4.71	56	walk	1.70
7	thing	4.17	57	first	1.66
8	good	4.05	58	woman	1.66
9	make	3.92	59	talk	1.66
10	knightley	3.75	60	hope	1.66
11	mr	3.62	61	miss_bates	1.63
12	come	3.54	62	highbury	1.61
13	mrs_weston	3.52	63	way	1.60
14	go	3.35	64	like	1.56
15	littl	3.14	65	poor	1.53
16	such	3.11	66	marri	1.52
17	give	3.11	67	room	1.51
18	mr_elton	3.04	68	visit	1.49
19	see	2.97	69	john	1.48
20	great	2.93	70	word	1.44
21	own	2.89	71	believ	1.43
22	much	2.80	72	home	1.40
23	time	2.75	73	mind	1.40
24	look	2.70	74	moment	1.40
25	frank	2.65	75	mani	1.38
26	other	2.49	76	doubt	1.36
27	feel	2.48	77	felt	1.36
28	mrs_elton	2.45	78	appear	1.36
29	mr_weston	2.40	79	ladi	1.35
30	take	2.33	80	begin	1.35
31	man	2.32	81	suppos	1.33
32	day	2.29	82	parti	1.33
33	friend	2.29	83	perry	1.32

34	noth	2.21	84	manner	1.32
35	hear	2.20	85	idea	1.31
36	letter	2.18	86	miss	1.31
37	wish	2.11	87	comfort	1.29
38	mrs	2.09	88	last	1.29
39	happi	2.08	89	pleasur	1.29
40	father	2.07	90	leav	1.27
41	speak	2.01	91	call	1.26
42	more	2.00	92	randalls	1.26
43	young	1.98	93	return	1.26
44	seem	1.97	94	child	1.26
45	want	1.91	95	someth	1.25
46	robert_martin	1.90	96	hour	1.25
47	bodi	1.88	97	place	1.25
48	campbell	1.86	98	oh	1.25
49	tell	1.81	99	isabella	1.24
50	sure	1.76	100	oblig	1.24

4.2 소설 엠마 텍스트 네트워크 분석 결과

하이베리 네트워크 내 인물들의 연결 관계, 그리고 그들의 역할과 지위를 검토하기 위한 방법으로 3가지 네트워크 (연결정도, 근접, 매개)중심성 값을 산출하였다. 분석 결과는 표와 그림으로 표현된다. 먼저 그림 2는 시각화 그래프로 나타나는데, 그래프는 표에 비해 정확한 수치 파악이 어렵지만 한눈에 순위와 정도 차이를 파악할 수 있는 장점이 있다. 반면 각 노드의 중심성을 수치화하여 순위로 정리한 표는 정확한 순위 파악과 비교에 유용하다. 텍스트 네트워크 분석의 가장 큰 장점인 시각화 그래프는 전체 네트워크를 한눈에 파악할 수 있고 중심성이 높은 인물의 위치와 관계 패턴을 직관적으로 조망할 수 있다.

그림 2의 텍스트 네트워크 그래프는 제인 오스틴의 소설 엠마에서 등장하는 주요 인물들 사이의 사회적 연결 관계를 시각화하였다. 그래프는 각 인물(노드)을 점으로, 그들 간의 상호작용이나 관계(엣지)를 선으로 나타내며, 텍스트에서 나타나는 인물 간의 직접적 또는 간접적인 연결을 통해 이들의 사회적 구조를 분석한다. 그래프 중심에 위치한 엠마는 다수의 엣지로 연결되어 있다, 이는 엠마가 소설 내에서 중심적인 인물로서 다양한 인물과 깊은 관계를 맺고 있음을 시사한다. 엠마와 같이 중심에 위치한 인물들은 텍스트 내에서 주요 사건의 발달에 중추적인 역할을 하며, 사회적 상호작용에서 중심적인 위치를 차지한다. 주변 노드인 나이틀리, 해리엇, 프랭크, 제인과 같은 인물들도 많은 연결선을 가지고 있어, 이들 역시 하이베리 사회에서 중요한 역할을 담당하고 있음을 나타낸다. 이러한 인물들은 엠마와 함께 사건의 전개에 큰 영향을 미치며, 그들의 네트워크 위치는 소설의 플롯과 주제 이해에 있어 중요한 역할을 한다.

중심성 분석 결과 엠마 노드에 연결된 선이 두꺼운 나이틀리, 해리엇, 제인, 워스턴 부인 같은 인물들은 네트워크에서 주인공 엠마와 중요한 연결 관계를 가지며 중요한 상호작용을 하는 위치임을 의미한다. 이들은 다양한 인물과의 연결을 통해 정보의 중심지 역할을 하며, 소설의 중심적인 이야기를 이끌어 간다. 엠마와 나이틀리는 그래프상에서 다른 모든 노드와의 거리가 상대적으로 가깝다. 이는 이들이 사회 내에서 손쉽게 정보를 주고받고, 사회적 상호작용에 있어 중요한 역할을 한다는 것을 나타낸다. 엠마는 많은 인물과의 관계를 통해 하이베리 사회 내에서

소통의 중재자 역할을 할 가능성이 높다. 이는 그녀가 사회 내에서 다리 역할을 하며 다른 인물들 간의 소통을 원활하게 하는 중요한 위치에 있다는 것을 시사한다. 그림 2 네트워크 그래프는 엠마 소설의 등장인물들 사이의 복잡한 사회적 연결망을 효과적으로 시각화하고, 각 인물의 사회적 중요성과 역할을 분석하는 데 도움을 준다. 이러한 분석을 통해 소설의 깊이 있는 이해와 인물들 간의 관계 구조에 대한 통찰을 얻을 수 있다.

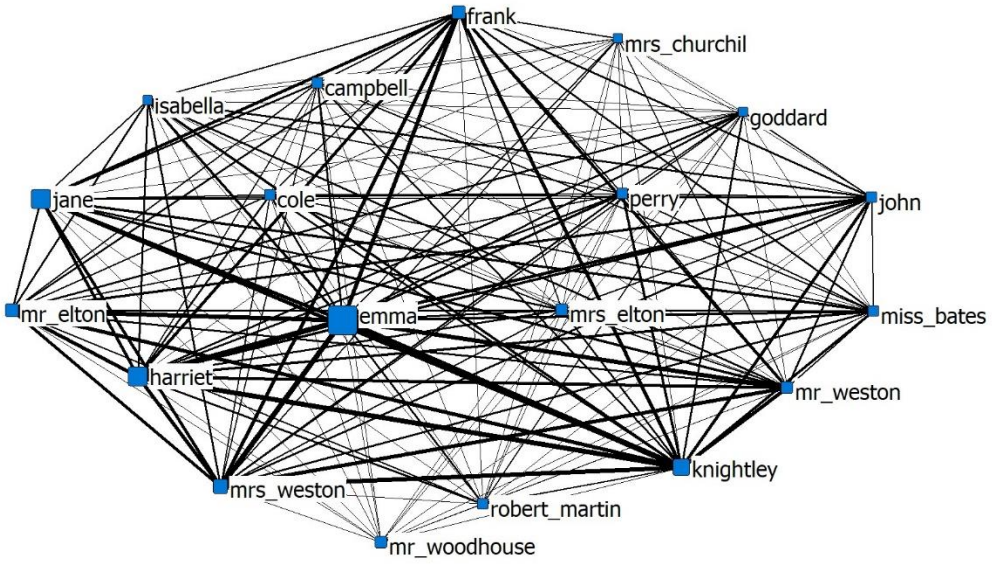

그림 2. 소설 엠마 인물 텍스트 네트워크

다음 표 5는 소설 엠마에 등장하는 주요 인물들의 네트워크 중심성을 정량적으로 분석한 결과표로 각 지표는 소설 내에서 인물들의 사회적 영향력과 중요성을 다른 관점에서 나타낸다. 먼저, 연결정도 중심성에서 주인공 엠마 양이 가장 높은 중심성을 보였고, 나이틀리와 해리엇이 그 뒤를 따른다. 근접 중심성은 우드하우스 씨와 프랭크가 높게 나타났다. 반면 매개 중심성은 우드하우스 씨와 프랭크가 낮게 나타났다. 마지막으로, 위세 중심성은 연결정도 중심성과 같이 엠마가 가장 높았고 다음으로 나이틀리와 해리엇이 뒤를 이었다.

구체적으로 살펴보면, 먼저 연결정도 중심성은 인물이 얼마나 많은 다른 인물과 직접적으로 연결되어 있는지를 나타낸다. 이 지표에서 엠마는 463의 가장 높은 값을 가지며, 이는 엠마가 소설 내에서 가장 중심적이고 활동적인 인물임을 보여준다. 나이틀리는 425의 두 번째 높은 값을 가지며 엠마에 이어 많은 인물들과 연결되어 있다. 이는 두 인물이 소설의 사회적 상호작용에서 핵심적인 역할을 수행함을 시사한다. 다음으로 근접 중심성은 네트워크 내의 모든 다른 인물에게 얼마나 가까이 위치해 있는지를 측정한다. 이 표에서 대부분의 주요 인물들의 근접 중심성 값은 18로 일정하며, 이는 대부분의 인물들이 서로 비슷한 수준의 접근성을 가지고 있음을 나타낸다. 예외적으로, 우드하우스 씨와 처칠 부인, 프랭크는 각각 20과 19의 값을 가지며, 이는 이들이 네트워크 내에서 다른 노드들과 상대적으로 가깝게 위치하여 정보나 자원을 빠르게 전파하거나 수집할 수 있는 위치를 의미한다. 에이겐벡터 중심성은 인물이 네트워크 내에서 중요한 다른

인물들과 얼마나 연결되어 있는지를 고려한다. 엠마는 이 지표에서 0.376의 가장 높은 값을 가지며, 중요한 인물들과 강력하게 연결된 것으로 나타난다. 나이틀리와 해리엇 또한 각각 0.351과 0.324의 높은 값을 가지며, 중요한 사회적 연결고리를 유지하고 있음을 보여준다. 반면, 우드하우스 씨(0.025)와 처칠 부인(0.096)의 에이겐벡터 중심성 값은 매우 낮게 나타났다. 이는 우드하우스 씨와 처칠 부인은 다른 인물들과 효율적으로 연결되어 있지만 영향력은 낮음을 의미한다. 마지막으로 매개 중심성은 인물이 다른 두 인물 간의 상호작용에 얼마나 자주 중재하는 역할을 하는지를 나타낸다. 대부분의 인물들이 0.125의 동일한 매개 중심성 값을 가지고 있으며, 이는 많은 인물들이 소설 내에서 중재자 역할을 공평하게 수행하고 있음을 나타낸다. 그러나 프랭크, 처칠 부인, 우드하우스 씨는 0의 값을 가지며, 이는 이들이 다른 인물들 간의 상호작용에서 중재 역할이나 소통의 병목 역할을 거의 수행하지 않음을 나타낸다.

표 5. 소설 엠마 네트워크 중심성 값

#	Character	Degree	Closeness	Eigenvector	Between
1	emma	463	18	0.376	0.125
2	harriet	389	18	0.324	0.125
3	jane	307	18	0.26	0.125
4	knightley	425	18	0.351	0.125
5	mrs_weston	352	18	0.3	0.125
6	mr_weston	301	18	0.261	0.125
7	mrs_elton	183	18	0.158	0.125
8	mr_elton	313	18	0.266	0.125
9	frank	311	19	0.268	0
10	miss_bates	228	18	0.194	0.125
11	robert_martin	132	18	0.114	0.125
12	john	247	18	0.209	0.125
13	isabella	203	18	0.177	0.125
14	perry	206	18	0.177	0.125
15	cole	189	18	0.159	0.125
16	goddard	173	18	0.15	0.125
17	mrs_churchil	108	19	0.096	0
18	campbell	175	18	0.152	0.125
19	mr_woodhouse	31	20	0.025	0

추가로 표 6의 동시출현 빈도표는 제인 오스틴의 소설 엠마에서 각 등장인물이 함께 언급된 횟수를 수치화한 표이다. 이 표는 소설 내에서 인물들 간의 관계의 밀접함과 상호작용의 정도를 나타내는 지표이다. 분석을 통해 각 인물 간의 연결 강도와 소설 내 주요 사회적 관계를 파악할 수 있다. 단어의 동시출현 빈도를 살펴보면 엠마와 나이틀리의 관계가 가장 높게 나타났고 다음으로 엠마와 해리엇, 그리고 나이틀리와 해리엇, 엠마와 웨스턴 부인 순서로 나타났다.

구체적으로 주인공 엠마와 다른 인물들 간의 연결을 살펴보면, 엠마와 나이틀리는 빈도수(49)가 가장 높아, 엠마와 나이틀리가 서로에게 중요한 존재임을 강조한다. 두 인물은 소설 전반에 걸쳐 서로의 행동과 결정에 큰 영향을 미친다. 엠마와 해리엇은 두 번째로 높은 연결 빈도수(44)를 보여, 이들이 소설에서 매우 긴밀한 관계를 유지하고 있음을 나타낸다. 이는 엠마가 해리엇의 사회적 상승을 돕기 위해 노력하는 중요한 플롯의 일부를 반영한다. 엠마와 프랭크도

빈도수(35)로 둘 사이의 빗나간 상상의 로맨스를 반영한다. 그리고 엠마와 엘튼씨(36) 사이의 높은 동시출현 빈도는 이 둘 사이의 엇갈린 짝사랑 관계를 드러낸다. 다른 주요 인물들의 연결 패턴을 살펴보면, 해리엇과 나이틀리는 비교적 높은 빈도(40)로 함께 언급되며, 이는 해리엇의 나이틀리의 호의에 대한 착각으로 인한 연정을 시사한다. 비교적 낮은 빈도(3)는 우드하우스 씨는 주인공의 아버지임에도 소설 내에서 배경 인물로 남는 것을 암시한다. 또한 전반적으로 낮은 동시출현 빈도는 처칠 부인이 소설 내에서 비교적 고립된 위치에 있음을 시사한다.

표 6. 동시출현 단어 빈도표

	emma	harriet	jane	knightley	mrs_weston	mr_weston	mrs_elton	mr_elton	frank	miss_bates	robert_martin	john	isabella	perry	cole	goddard	mrs_churchil	campbell	mr_woodhouse
emma	0	44	32	49	39	34	18	36	35	23	15	26	21	21	18	19	12	18	3
harriet	44	0	27	40	31	25	15	31	26	20	15	21	18	19	16	19	6	14	2
jane	32	27	0	30	24	21	15	18	25	20	8	13	12	13	14	8	7	18	2
knightley	49	40	30	0	36	30	17	32	31	21	12	25	20	20	17	16	9	17	3
mrs_weston	39	31	24	36	0	31	14	24	29	17	8	17	17	16	13	12	10	13	1
mr_weston	34	25	21	30	31	0	12	19	28	12	6	15	16	13	8	9	11	9	2
mrs_elton	18	15	15	17	14	12	0	13	13	10	4	10	7	8	10	4	4	7	2
mr_elton	36	31	18	32	24	19	13	0	18	14	12	20	15	13	13	16	6	10	3
frank	35	26	25	31	29	28	13	18	0	16	5	15	12	13	11	9	12	13	0
miss_bates	23	20	20	21	17	12	10	14	16	0	6	10	8	10	14	7	6	13	1
robert_martin	15	15	8	12	8	6	4	12	5	6	0	8	4	6	5	10	1	5	2
john	26	21	13	25	17	15	10	20	15	10	8	0	16	12	10	11	7	8	3
isabella	21	18	12	20	17	16	7	15	12	8	4	16	0	10	8	7	6	4	2
perry	21	19	13	20	16	13	8	13	13	10	6	12	10	0	10	11	2	7	2
cole	18	16	14	17	13	8	10	13	11	14	5	10	8	10	0	7	5	9	1
goddard	19	19	8	16	12	9	4	16	9	7	10	11	7	11	7	0	1	6	1
mrs_churchil	12	6	7	9	10	11	4	6	12	6	1	7	6	2	5	1	0	3	0
campbell	18	14	18	17	13	9	7	10	13	13	5	8	4	7	9	6	3	0	1
mr_woodhouse	3	2	2	3	1	2	2	3	0	1	2	3	2	2	1	1	0	1	0

그림 3과 4는 표 5 네트워크 중심성을 막대그래프 나타낸 것으로, 표에서는 한 눈에 들어오지 않는 각 중심성의 패턴을 관찰할 수 있다. 연결 중심성과 위세 중심성은 주인공을 중심으로 차이가 두드러지지만 근접 중심성의 경우 우드하우스 씨와 프랭크, 처칠 부인만 높게 나타났다. 반면 매개 중심성은 반대로 우드하우스 씨와 프랭크, 처칠 부인만 낮고 나머지는 동일하였다. 특이하게 우드하우스 씨와 처칠 부인은 연결 중심성과 위세 중심성이 모두 매우 낮게 나타났지만 프랭크는 둘 다 상당히 높게 나타났다.

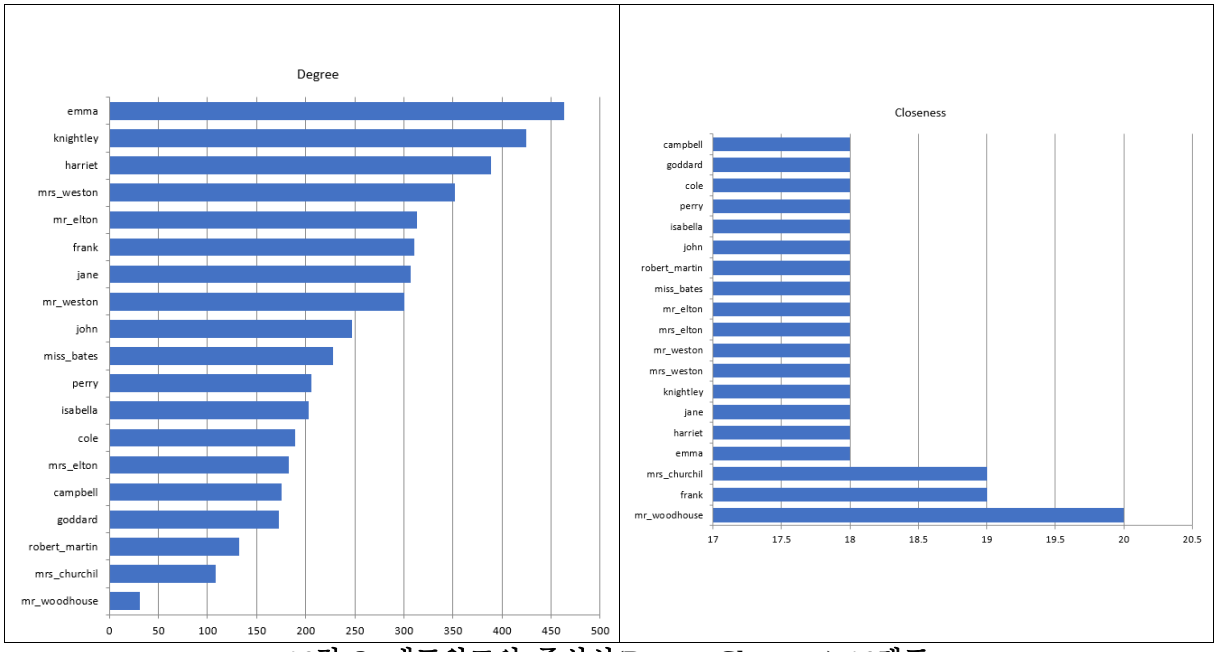

그림 3. 네트워크의 중심성(Degree, Closeness) 그래프

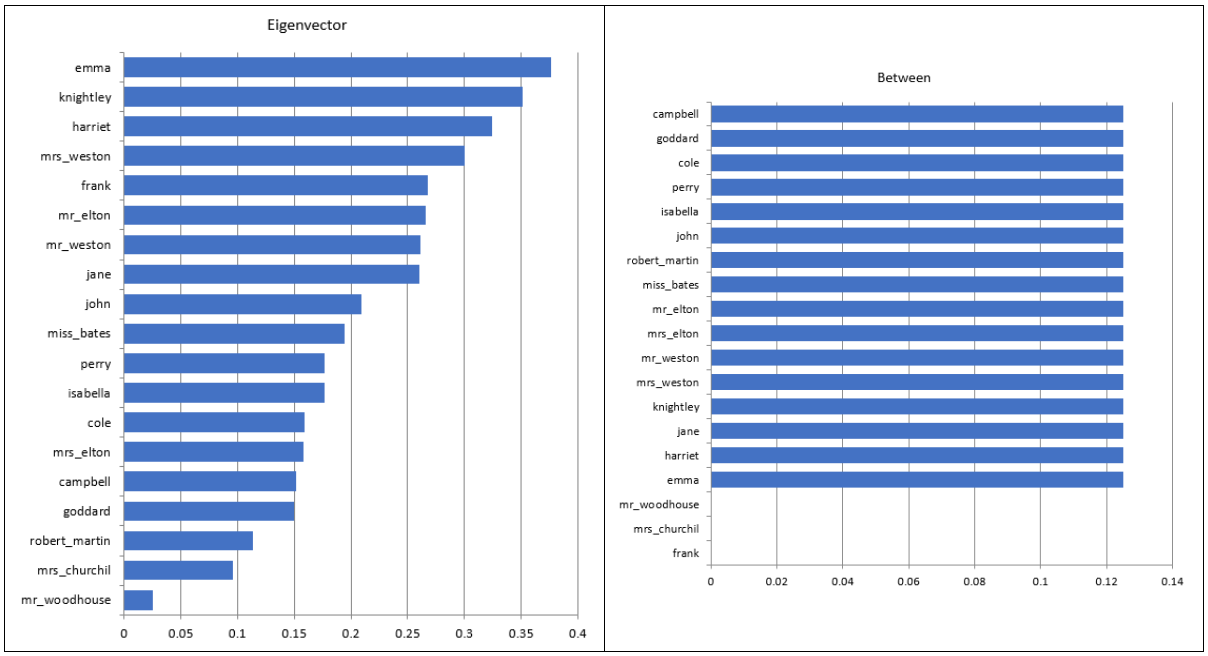

그림 4. 네트워크의 중심성(Eigenvector, Between) 그래프

4.3 소설 엠마 토픽 모델링 분석 결과

오스틴은 소설 엠마를 세 개의 권(Volume)으로 구성하였고, 이는 다시 총 55장(Chapter)인데 1권은 18장, 2권은 18장, 3권은 19장으로 구분되어 있다. LDA 방법에 따라 55개의 장(문서)에 대해 토픽의 수만큼의 토픽 할당 확률을 계산하고, 각 장에 따라 가장 높은 확률을 가지는 토픽을 추출하였다.

먼저 토픽 개수를 결정하기 위해, 그림 5의 토픽 분류의 혼란도(Perplexity)와 일관성(Coherence) 지수를 살펴보면, 혼란도는 확률 모델이 실제 데이터를 얼마나 잘 예측하는지를 나타낸다. 혼란도 점수가 낮을수록 더 나은 예측 모델을 의미한다. 그래프에서 주제의 수가 3개에서 9개로 증가함에 따라 혼란도가 감소하는 것을 볼 수 있다. 이는 주제의 수를 늘릴수록 데이터의 구조를 더 잘 포착하여 모델이 숨어있는 데이터를 더 효과적으로 예측할 수 있음을 시사한다. 반면, 일관성은 토픽 내에서 높은 점수를 얻은 단어들 사이의 의미적 유사성을 측정한다. 일관성 수치가 높을수록 주제가 더 해석하기 쉬워진다. 이 연구에서 그래프의 일관성 점수는 주제가 4개일 때 최고점을 찍고, 이후에는 대체로 감소하는 경향을 보이나, 7개 주제에서 약간 상승한 후 다시 떨어졌다. 이는 4개의 주제에서 모델이 가장 의미 있는 해석 가능한 주제 집합을 제공한다는 것을 나타냈다. 즉, 혼란도가 계속 개선되는(즉, 감소하는) 동안에도 4개의 주제에서 일관성 점수가 가장 높기 때문이다. 4개를 초과하는 주제를 추가하면 혼란도는 계속 개선되지만 일관성 점수는 감소하여, 주제가 과적합될 수 있음을 시사했다. 따라서 4개의 주제를 선택하는 것이 혼란도를 최소화하면서 일관성을 최대화하여 주제를 예측적이면서도 해석 가능하게 만드는 균형을 이룬다.

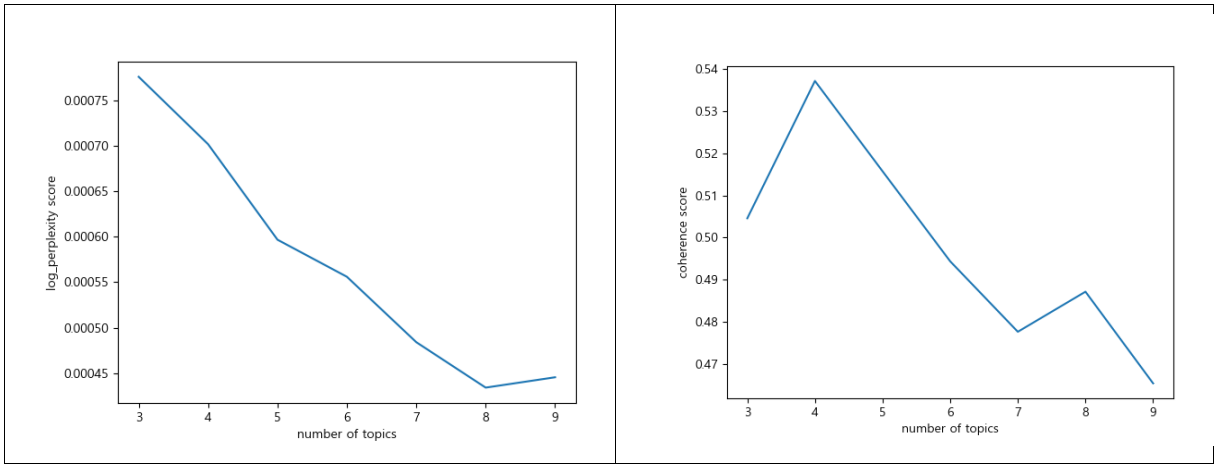

그림 5. 혼란도(Perplexity)와 일관성(Coherence) 지표

표 7에 토픽 별 주요 단어 결과표를 제시하였다.

표 7. LDA 분석 결과 각 토픽 주요 단어

Topic	Words with Weights
Topic 1: 사회적 연결과 개인적 상호작용	mrs_elton (0.009), robert_martin (0.005), goddard (0.004), miss (0.004), john (0.004), girl (0.004), pay (0.003), read (0.003), fine (0.003), draw (0.003), brother (0.003), send (0.003), sens (0.003), isabella (0.003), son (0.003)
Topic 2: 외부 세계와의 상호작용	campbell (0.009), perry (0.007), miss_bates (0.005), cole (0.004), mention (0.004), air (0.004), carriag (0.003), how (0.003), voic (0.003), isabella (0.003), concern (0.003), john (0.003), toward (0.003), suspicion (0.003), london (0.003)
Topic 3: 사회적 이벤트와 개인적 경험	danc (0.009), mrs_elton (0.008), miss_bates (0.007), miss (0.005), read (0.004), ball (0.004), mother (0.004), campbell (0.003), news (0.003), rememb (0.003), suffer (0.003), door (0.003), send (0.003), stop (0.003), cole (0.003)
Topic 4: 계절의 변화와 감정의 교류	john (0.007), robert_martin (0.006), isabella (0.006), miss (0.005), snow (0.004), mrs_elton (0.004), carriag (0.004), sister (0.004), attend (0.003), play (0.003), mother (0.003), quick (0.003), campbell (0.003), danger (0.003), confess (0.003)

구체적으로 그림 6의 토픽 모델링 분석 결과 그래프에서 토픽 1에 대한 주요 내용을 다각적으로 보여준다. 이 그래프는 토픽 1의 핵심 단어 분포와 토픽 간의 거리를 시각화한 것으로, 토픽의 내용과 각 토픽 간의 관계를 파악하는 데 유용하다. 토픽 간 거리 지도(Intertopic Distance Map)는 토픽 간의 관계를 다차원척도법을 통해 시각화하였다. 크기가 큰 빨간 원(토픽 1)은 이 토픽이 다루는 문서들의 분포가 넓다는 것을 나타내며, 다른 주제들과의 거리가 멀다는 것은 토픽 1이 상대적으로 독립적인 특성을 가진 토픽임을 시사한다. 이는 토픽 1이 다루는 내용이 다른 주제들과는 구별되는 독특한 내용을 포함하고 있음을 의미한다.

토픽 1의 핵심 단어 분포를 나타내는 막대 그래프는 토픽 내에서 중요도가 높은 단어들을 상위 30개까지 보여준다. 이 그래프에서 엘튼 부인, 로버트 마틴, 고다드 부인 등의 단어가 높은 빈도로 등장함을 알 수 있다. 이러한 단어들은 소설 엠마 내에서 특정 인물들과 관련된 내용이 토픽 1에서 중요한 비중을 차지함을 나타낸다.

구체적으로 토픽 1은 가장 범위가 크고 '사회적 연결과 개인적 상호작용' 토픽 제목으로 소설 엠마에서 엘튼부인과 로버트 마틴, 그리고 고다드 부인의 사회적 관계 또는 사건들을 중심으로 구성되어 있다. 토픽 1은 주로 엘튼 부인, 로버트 마틴, 고다드 부인과 같은 인물들을 중심으로 소설 내에서의 사회적 상호작용과 계급 간의 교류를 다룬다. 엘튼 부인은 사교계에서의 지위를 확립하려는 그녀의 노력을 통해 소설에서 중요한 역할을 하며, 이는 하이베리 사회 내에서의 긴장과 계급 간의 상호작용을 묘사하는 데 중요한 요소로 작용한다. 로버트 마틴의 경우, 계급적 위치에도 불구하고 개인적인 성취와 사랑을 추구하는 인물로, 그의 이야기는 계급 간의 경계를 넘나드는 사랑과 관계의 가능성을 탐구한다. 고다드 부인은 해리엇의 삶에 긍정적인 영향을 미치는 기숙학교 교장 선생님으로 해리엇의 보호자와 교육자 역할을 한다.

이 토픽에서 나타나는 'read', 'draw', 'pay' 같은 단어들은 인물들이 일상생활에서 어떻게 자신의 취미와 경제적 상황을 통해 사회적 상호작용을 하고 있는지를 보여준다. 특히, 문화 활동에 참여하는 모습은 인물들의 사회적 지위와 개인적 취향이 어떻게 연결되어 있는지를 드러내며, 이는 소설의 사회적 맥락과 계급 구조를 이해하는 데 기여한다. 가족 관계를 나타내는 'brother', 'son', 'isabella' 등의 단어는 가족이 개인의 사회적 행위와 선택에 얼마나 큰 영향을 미치는지를 강조한다. 가족 구성원 간의 관계는 소설에서 인물들의 행동을 이해하고 예측하는 키가 되며,

사회적 행사나 개인적 결정에 중요한 배경이 된다. 이러한 분석을 통해 소설 엠마의 사회적 상호작용, 계급 문제, 개인적 취미 및 가족 관계가 어떻게 복합적으로 얽혀 있으며, 이가 소설의 주요 테마와 갈등을 형성하는 방식을 깊이 있게 이해할 수 있다.

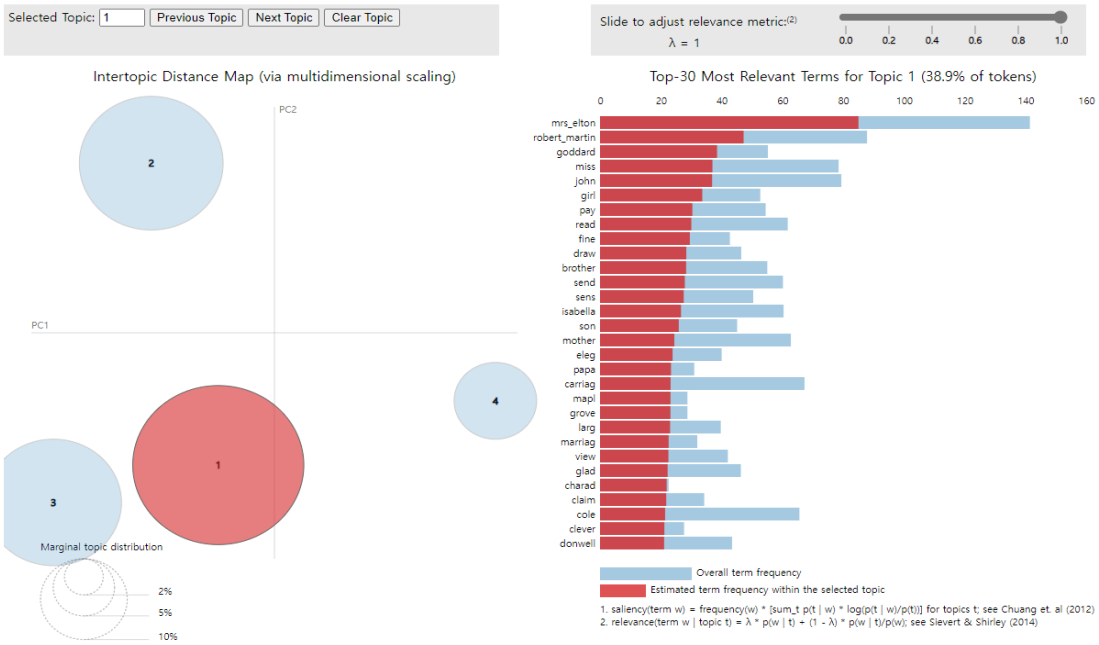

그림 6. 토픽 1 토픽 간 거리 지도(Intertopic Distance Map)

토픽 2는 '외부 세계와의 상호작용'이라는 제목으로, 제인 오스틴의 소설 엠마에서 하이베리 지역의 사회적 상호작용과 외부 세계와의 연결을 다루는 주제를 반영한다. 이 토픽은 캠벨, 페리, 베이츠 양, 콜 등의 인물들과 그들의 일상적인 활동 및 대화를 중심으로 구성되어 있으며, 이는 하이베리의 사회적 맥락과 이 지역을 둘러싼 외부 세계와의 상호작용을 나타낸다. 캠벨과 페리는 소설에서 중요한 지역 인물로서, 지리적 이동성에 기반한 그들의 일상과 대화는 하이베리의 사회적 상호작용을 형성하는 데 중요한 역할을 한다. 특히, 양부모 캠벨은 제인 페어팩스의 성장과 관련된 다양한 사회적 활동을 통해 주요 인물 간의 관계 구축에 기여한다. 페리는 지역의 의사로서 그의 의견과 활동이 다른 인물들의 건강과 관련된 대화에서 중심적인 역할을 하며, 사회적 교류의 중심축을 제공한다.

베이트 양과 콜은 소설에서 하이베리의 소소한 일상과 행사를 통해 커뮤니티의 일원으로서의 역할을 강조한다. 베이츠 양의 지속적인 대화는 종종 그녀의 개인적인 경험과 관찰을 통해 하이베리 사회의 세세한 부분을 드러내며(원영선 2016, 2023), 콜은 사회적 모임과 행사에서 중요한 위치를 차지한다. 이 토픽에서 'carriage', 'London'과 같은 단어는 하이베리와 외부 세계, 특히 런던과의 물리적 연결을 상징한다. 이러한 연결은 하이베리 주민들의 이동과 소통을 가능하게 하며, 외부 세계의 사건과 정보가 하이베리 내의 사회적 상호작용과 어떻게 통합되는지를 보여준다. 이는 하이베리가 고립된 마을이 아니라, 더 넓은 사회적, 경제적 맥락 속에서 존재하고 있음을 강조한다. 토픽 2를 통해 제인 오스틴은 하이베리 커뮤니티의 복잡한

사회적 구조와 외부 세계와의 상호작용을 통해 인물들의 관계와 개인적 성장이 어떻게 영향을 받는지 깊이 있게 탐구한다. 이는 소설 엠마의 중요한 테마 중 하나로, 인물들의 개인적 삶과 사회적 역할이 어떻게 상호작용하는지를 세밀하게 조명한다.

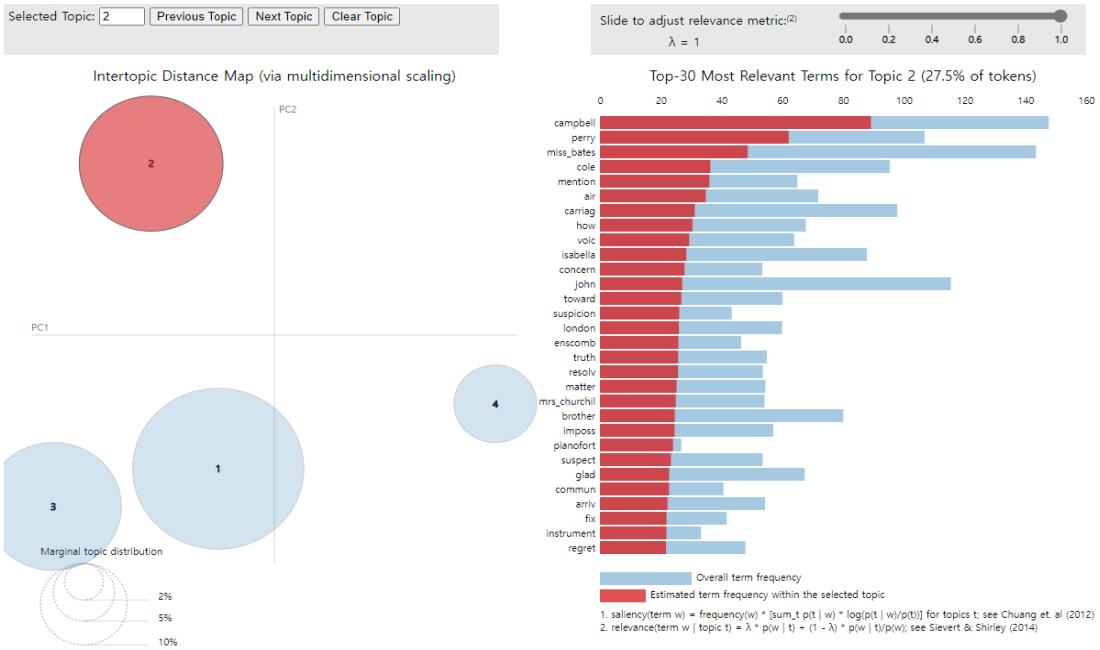

그림 7. 토픽 2 토픽 간 거리 지도(Intertopic Distance Map)

토픽 3 '사회적 이벤트와 개인적 경험'은 토픽 1과 가까운 거리에 위치하면서 소설 엠마에서 중요한 사회적 이벤트와 그에 따른 개인적 감정과 기억의 교류를 다룬다. 이 토픽에 포함된 주요 단어들은 사회적 모임, 특히 댄스와 볼을 중심으로 한 행사와 관련된 내용을 반영하며, 인물들 간의 상호작용 및 개인적인 감정의 표현을 통해 깊이 있는 인간관계의 직물을 짜낸다. 사회적 행사의 중심성으로 'danc'와 'ball'이 토픽의 중요 단어로 등장함으로써, 하이베리에서 열리는 댄스와 볼이 주민들의 사회생활에서 차지하는 중심적인 역할을 강조한다. 이러한 행사들은 소설에서 주요 인물들이 서로를 더 잘 알아가고, 사회적 유대를 강화하는 기회를 제공한다.

엘튼 부인과 베이츠 양은 이 토픽에서 두드러지게 나타나는 인물들로, 각각 하이베리 사회 내에서 상이한 사회적 역할을 수행한다. 엘튼 부인은 자신의 사회적 지위를 과시하려는 새로운 이주민으로서의 행동을 통해, 때로는 긴장과 갈등의 원인이 되기도 한다. 반면, 베이츠 양은 보다 친근하고 사교적인 인물로서 커뮤니티 내에서 중재자 및 연결자의 역할을 한다(원영선 2016, 2023). 개인적 감정과 기억의 역할로 'read', 'rememb', 'suffer' 같은 단어들은 사회적 행사가 개인의 내면에 미치는 영향을 나타낸다. 이들 단어는 개인이 사회적 상황을 통해 경험한 감정을 기억하고 때로는 그로 인해 고통받는 과정을 드러내며, 이는 인물들의 개성과 성장에 깊이를 더한다. 가족과의 연결로 'mother', 'send', 'door' 등의 단어는 가족과의 관계 및 가정과의 연결을 상징한다. 특히 'mother'는 엠마의 어머니와의 관계 및 그에 따른 기억과 감정이 어떻게 엠마의 행동과 결정에 영향을 미치는지를 보여준다. 이 토픽을 통해 제인 오스틴은 소설 엠마에서 사회적 행사가 개인의

감정과 기억에 어떻게 영향을 미치는지, 그리고 이러한 경험이 인물들의 인간관계와 자아 발전에 어떻게 기여하는지를 섬세하게 탐구한다. 이는 소설의 풍부한 감정적 및 사회적 관계 구조를 이해하는 데 중요한 열쇠를 제공한다.

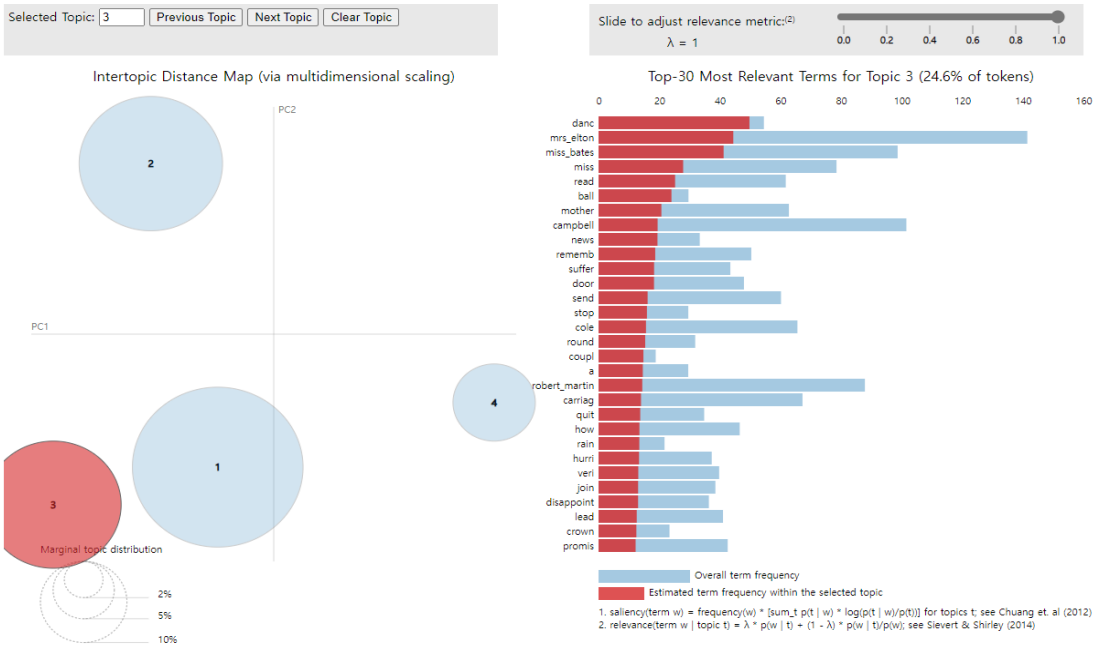

그림 8. 토픽 3 토픽 간 거리 지도(Intertopic Distance Map)

토픽 4 '계절의 변화와 감정의 교류'는 가장 범위가 작고 따로 떨어져 있다. 이 토픽에서는 소설 엠마에서 겨울철에 발생하는 자연적 및 인간적 사건들을 중심으로 하여 이러한 사건들이 인물 간의 감정적 교류와 관계 변화에 어떻게 영향을 미치는지 나타낸다. 이 토픽에서 'snow', 'danger', 'carriag'와 같은 단어들은 겨울철의 도전을 나타낸다. 소설에서 겨울은 물리적인 어려움을 초래할 뿐만 아니라, 인물들 간의 감정적 긴장을 높이는 주요 요소로 작용한다. 특히, 랜들즈에서 열리는 크리스마스 파티 중에 발생한 눈보라는 주요 인물들이 서로의 복지를 위해 협력하면서 관계를 강화하는 계기가 된다. 인간관계의 심화로 존, 이사벨라, 로버트 마틴 등의 인물들은 소설의 중심적인 인물로서, 겨울철의 어려움을 통해 서로에 대한 이해와 감정의 깊이를 더하게 된다.

또한, 감정의 고백과 변화로 'confess'와 같은 단어는 개인의 내면에 숨겨진 감정과 비밀이 겨울철의 특별한 사건을 통해 드러나게 됨을 시사한다. 이러한 고백은 인물들 간의 오해를 해소하고, 갈등을 조정하는 중요한 역할을 수행하며, 소설의 결말에 큰 영향을 미친다. 이 토픽은 제인 오스틴이 소설 엠마에서 계절의 변화와 자연환경이 인물들의 심리적 상태와 사회적 관계에 미치는 영향을 어떻게 포착하고 있는지 보여준다. 겨울철의 시련을 통해 각 인물의 감정적 깊이와 사회적 유대가 어떻게 강화되는지를 세밀하게 탐구함으로써, 계절적 요소가 소설의 감정적 풍경과 인간관계의 변화에 어떻게 중요한 역할을 하는지를 드러낸다.

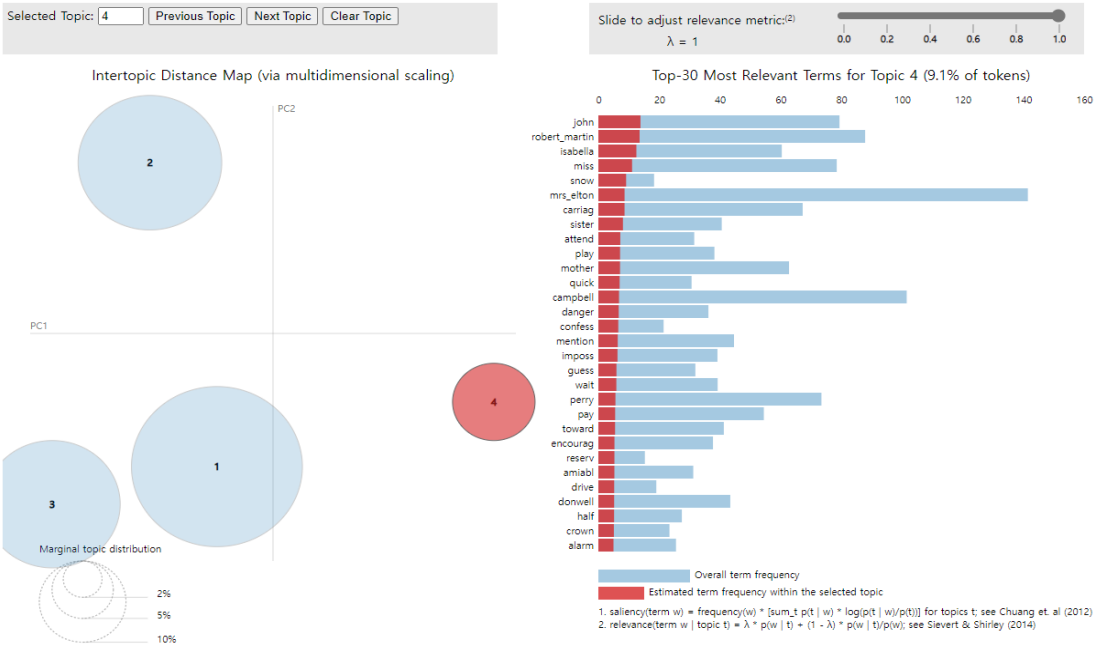

그림 9. 토픽 4 토픽 간 거리 지도(Intertopic Distance Map)

4.4 소설 엠마 감성분석(Sentiment Analysis) 결과

표 8에 소설 엠마의 각 권 별로 사용한 단어에 대한 감성점수(Sentiment Score)를 제시하였다. 제인 오스틴의 소설 엠마는 세 권으로 구성되어 있으며, 각 권에 나타나는 감정 변화는 감성분석을 통해 구체적으로 표현될 수 있다.

첫 번째 권에서는 엠마 우드하우스가 주변 인물들의 연애와 결혼을 적극적으로 조종하려는 모습이 그려진다. 긍정적인 단어인 'good(184회)', 'great(117회)', 'young(73회)', 'sure(46회)', 'best(19회)', 'right(18회)' 등이 부정적인 단어보다 많이 언급되며, 이는 엠마의 낙관적이고 자신감 넘치는 태도를 반영한다. 예를 들어, 'good'이 184회 언급되어 엠마의 활동이 긍정적이고 성공적이라고 믿고 있음을 나타낸다. 그러나 부정적인 단어들인 'poor(52회)', 'last(40회)', 'bad(37회)', 'little(33회)', 'short(25회)', 'afraid(14회)'도 나타나며, 이는 엠마의 과도한 자신감이 문제를 일으킬 가능성을 시사한다. 'last'가 40회 언급된 것은 엠마가 자신의 계획이 끝까지 성공할 것이라고 믿는 동시에, 실패의 가능성도 내포하고 있음을 시사한다.

두 번째 권에서는 엠마의 계획들이 예상치 못한 결과를 낳으며, 그녀의 오만과 착각이 갈등을 유발하는 주된 요인으로 작용한다. 긍정적인 단어들인 'good(128회)', 'great(110회)', 'young(77회)', 'sure(68회)'도 많이 언급되지만, 이는 엠마가 여전히 자신의 행동이 옳다고 믿고 있는 상태를 반영한다. 그러나 부정적인 단어들인 'bad(34회)', 'few(33회)', 'poor(31회)', 'old(31회)', 'afraid(18회)'도 빈번하게 사용되어, 엠마의 개입이 초래한 부정적인 영향과 인물 간의 복잡한 감정적 결과를 나타낸다. 예를 들어, 'bad'가 34회 언급되어 엠마의 행동이 주변 인물들에게 부정적인 영향을 미쳤음을 보여준다. 이러한 감성분석 결과는 엠마가 자신의 행동을 되돌아보게 만들며, 갈등과 실망이 소설의 중반부를 지배한다는 점을 나타낸다.

표 8. 권(volume) 별 감정분석 결과

	Volume 1		Volume 2		Volume 3	
순위	긍정(횟수)	부정(횟수)	긍정(횟수)	부정(횟수)	긍정(횟수)	부정(횟수)
1	good(184)	poor(52)	good(128)	bad(34)	good(149)	poor(50)
2	great(117)	last(40)	great(110)	few(33)	great(95)	bad(24)
3	young(73)	bad(37)	young(77)	poor(31)	sure(46)	afraid(19)
4	sure(46)	little(33)	sure(68)	old(31)	young(43)	wrong(19)
5	best(19)	short(25)	wish(63)	least(26)	superior(24)	ill(16)
6	right(18)	least(22)	same(44)	short(18)	high(18)	secret(15)
7	real(17)	afraid(14)	more(37)	afraid(18)	strong(17)	cold(13)
8	fair(17)	ill(13)	whole(30)	nobodi(17)	long(16)	low(13)
9	pleasant(15)	miss(13)	mean(29)	cold(12)	perfect(15)	evil(11)
10	swee(12)	high(12)	present(26)	evil(12)	smile(15)	sudden(11)
11	happy(12)	evil(11)	first(24)	sad(11)	best(15)	odd(11)
12	fine(11)	cold(11)	right(23)	long(11)	full(15)	hot(10)
13	perfect(11)	soft(11)	best(23)	less(10)	fine(13)	mad(9)
14	clear(9)	difficulty(11)	true(20)	ill(10)	right(12)	inferior(8)
15	warm(8)	nobody(10)	warm(20)	sorry(9)	safe(11)	want(7)
16	admire(8)	wrong(8)	full(19)	wrong(8)	eager(10)	late(7)
17	respect(8)	odd(8)	fine(19)	surpriz(8)	warm(10)	spite(6)
18	pleasure(8)	sigh(8)	superior(19)	asham(7)	glad(9)	avoid(6)
19	safe(8)	consider(8)	happy(13)	regret(7)	quiet(9)	suspect(6)
20	amiable(8)	sad(7)	smile(17)	spite(7)	clear(9)	cry(6)
21	rich(8)	cry(7)	new(17)	off(7)	deep(9)	less(6)
22	quiet(8)	inferior(7)	strong(17)	hurri(7)	pleasant(9)	dull(5)
23	well(7)	uneasy(6)	real(16)	unwil(6)	real(8)	unequal(5)
24	accomplished(7)	unhappy(6)	civil(15)	suspect(6)	felic(7)	unpleasant(5)
25	ready(7)	sorry(6)	equal(15)	silent(6)	satisfact(7)	serious(5)
26	eager(7)	dreadful(5)	pleasant(12)	sudden(6)	thank(7)	separate(5)
27	delightful(7)	sick(5)	proper(11)	late(6)	fortun(7)	pity(4)
28	willing(7)	nervous(5)	safe(9)	low(6)	quick(7)	mere(4)
29	fortunate(7)	regret(5)	certain(9)	inferior(5)	proper(6)	look(4)
30	friendly(7)	weak(4)	well(10)	lose(5)	early(6)	dread(4)

세 번째 권에서는 엠마가 자신의 행동을 반성하고, 진정한 감정을 인정하며 인물들과의 관계를 재정립하는 과정이 그려진다. 긍정적인 단어인 'good(149회)', 'great(95회)', 'superior(24회)' 등이 부정적인 단어보다 많이 언급되며, 이는 갈등의 해결과 인물들의 감정적 회복을 나타낸다. 'perfect(15회)와 같은 단어는 엠마가 자신의 실수를 인정하고 성숙해진 결과로서 긍정적인 변화를 경험하고 있음을 시사한다. 반면, 'poor(50회)', 'bad(24회)' 등 부정적 단어들도 여전히 나타나지만, 그 빈도는 긍정적 단어에 비해 적다. 이는 엠마가 실수에서 배우고 성장하는 과정을 통해 자신과 주변 인물들과의 관계를 더욱 깊이 있고 의미 있는 방향으로 발전시키는 것을 나타낸다.

이와 같이 소설 엠마의 각 권 별 감성분석을 통해 소설의 감정적 흐름과 주요 테마가 어떻게 전개되는지를 개괄적으로 파악할 수 있는 단서를 확인할 수 있다. 이를 통해 제인 오스틴이 그려낸 인간관계의 복잡성과 개인의 심리적 변화 그리고 각 인물의 내면적 성장과 사회적 상호작용이 어떻게 상호 연결되어 있는지를 간접적으로 추론할 수 있다.

5. 논의 및 결론

제인 오스틴의 엠마(1815)는 영문학 연구에서 가장 광범위하게 분석된 텍스트 중 하나로, 이 소설의 플롯은 주인공 엠마 우드하우스(Emma Woodhouse)의 일상과 그녀가 맺는 다양한 인간 관계를 중심으로 전개되며, 그녀의 인격적 성장과 인물 관계를 깊이 탐구한다. 이러한 서사 구조의 특성은 엠마가 자신의 잘못을 깨닫고 인격적으로 성숙해가는 과정을 통해 잘 드러난다. 이 연구는 영문학 대표적인 소설 엠마 텍스트에 대해 주요 NLP기법인 네트워크 텍스트 분석, 토픽 모델링 및 감성분석을 실시하여, 텍스트 분석기법이 영어 문학작품의 주요 테마와 인물 간의 관계 발전, 그리고 작품 분위기를 얼마나 효과적으로 반영할 수 있는지를 탐색하였다. 다시 말해, 소설 엠마 텍스트의 주제와 분위기, 그리고 서사 구조 및 내용에서 나타나는 패턴을 인공지능 알고리즘과 통계학적 관점에서 파악해 보고자 하였다. 이러한 NLP 방법론을 통해 영문학 작품을 텍스트 데이터의 형태로 분석하고 작품의 특성과 구조를 계량적으로 평가하고자 하였다. 이를 통해 문학작품의 질적 연구를 객관적인 통계 분석에 기반한 양적연구로 보완하며, 디지털 인문학적 관점에서 영문학 텍스트의 내용분석과 이를 영어교육으로 적용방안에 대해 탐색하고자 하였다.

이처럼 문학작품의 비평과 같은 질적 연구를 객관적인 통계량 계산에 기반한 양적연구로 보완하면서, 작품의 이해와 감상과 작가의 의도를 디지털 인문학적으로 해석하고 영어교육에 적용할 수 있는 시사점을 찾는 것이 이 연구의 주된 목적이라 할 수 있다. 이 연구의 주요 목적은 소설 엠마의 전체 텍스트를 대상으로 주제와 그 특징을 파악하는 것이었다. 이는 방대한 텍스트를 정성적인 방법으로 연구하기 이전에, 정량적인 방법을 통해 텍스트에 담긴 내용을 먼저 파악하려는 시도이다. 본 연구의 의의는 이러한 정량적 접근을 통해 소설 엠마의 텍스트 내용을 분석하고자 한 점에 있다. 최근 디지털화된 텍스트를 분석하려는 연구가 활발해지면서, 영문 고전문헌에 텍스트 마이닝 방법을 적용하여 유의미한 정보를 추출하고 분석하는 작업이 증가하고 있다. 본 연구는 이러한 흐름 속에서 소설 엠마의 영문 텍스트를 대상으로 텍스트 분석 방법을 적용할 수 있는 가능성을 검토하였다.

텍스트 분석 중 NLP 기법은 대규모 텍스트 데이터에서 새로운 정보와 지식을 발굴하기 위해 인공지능 알고리즘을 활용하는 접근방식이다. 이 기법은 빅데이터 환경에서 텍스트 데이터를 수집하고, 이를 구조화하여 의미 있는 패턴과 통찰을 도출하는 데 사용된다. 구체적으로, 텍스트 데이터를 수집하여 다양한 딥러닝 기법을 통해 학습시킨 모델을 새로운 분석에 적용하는 방식으로 작동한다. NLP 과정은 일반적으로 데이터 수집, 전처리, 특징 추출, 모델 학습, 모델 적용, 그리고 결과 해석의 단계를 포함한다.

데이터 수집 단계에서는 웹 크롤링, 문서 저장소, 소셜 미디어 등 다양한 출처에서 텍스트 데이터를 수집한다. 전처리 단계에서는 수집된 텍스트 데이터를 분석에 적합한 형태로 변환하며, 토큰화, 불용어 제거, 형태소 분석, 정규화 등의 과정을 거친다. 이러한 전처리 과정은 NLP에서 가장 중요하고 많은 시간과 노력이 드는 지점으로 분석의 질을 결정짓는 과정이다. 다음으로 특징 추출 단계에서는 전처리된 텍스트에서 의미 있는 특징을 추출하여 벡터화한다. 주요 방법으로는 TF-IDF, 워드 임베딩 등이 사용된다. 이어서 모델 학습 단계에서는 딥러닝 알고리즘을 통해 모델을 학습시키며, 신경망, 순환 신경망, 변환기 등의 기법이 활용된다.

모델 적용 단계에서는 학습된 모델을 새로운 텍스트 데이터에 적용하여 텍스트 분류, 감성 분석, 토픽 모델링, 텍스트 요약 등의 작업을 수행한다. 마지막으로 결과 해석 단계에서는 모델의 분석 결과를 해석하고, 이를 바탕으로 새로운 통찰과 지식을 도출한다. 이러한 NLP 기법은 특히 인공지능과 빅데이터 분석의 발전에 따라 그 중요성이 부각되고 있으며, 다양한 학문분야에서 혁신적인 분석 방법으로 자리 잡고 있다. 영문학 텍스트의 경우, NLP 기법을 활용하면 텍스트의 구조적 및 감성적 특성을 분석하여 영어 텍스트 이해 교육에 유용한 자료를 제공할 수 있다. 실제로, 소설의 주제, 인물 간의 관계, 감정의 흐름 등을 객관적인 수치로 분석함으로써 학생들이 텍스트를 더 깊이 이해하고 감상할 수 있도록 돕는다. 또한, 텍스트 분석 결과를 기반으로 학습자 맞춤형 교육 콘텐츠를 개발하여 효과적인 영어 학습을 지원할 수 있다.

이 연구의 제한점으로는 네트워크 분석, 토픽 모델링, 감성분석 등 다양한 분석방법을 사용하였지만 각각의 텍스트 분석기법의 한계로 인해 모든 문맥적 의미를 완벽히 추출하지 못할 가능성이 있다. 즉, 오스틴 비평가와 문학 전공자에게는 이번 연구 결과가 초보적이고 정교하지 않은 분석일 수 있다. 그러나 순수하게 영어 교육자의 관점에서 문학에 대한 배경지식 없이 영문학 고전 텍스트를 분석하는 이 연구의 제한된 목표를 고려할 때 앞으로 작가의 더 많은 작품을 대상으로 한 연구, 그리고 작품의 문화적·문학사적 문맥을 고려한 엄밀한 연구가 이루어진다면 텍스트 분석이 비평적 접근과는 다른 맥락에서 새로운 질문과 지식 확장의 가능성을 보여줄 수 있을 것이다.

다만, 이러한 계량적 분석 결과의 의미를 해석하는 작업은 여전히 정성적인 분석과 비평적 해석의 영역에 속한다. 디지털 문학 연구는 많은 작가와 작품, 그리고 외부 변인과 조건들을 처리할 수 있는 능력을 주요 강점으로 삼고, 텍스트 분석의 적용이나 도구적 활용을 위해 전문적 통계지식과 코딩능력이 필요하지만 텍스트의 본질적인 의미를 비평적으로 깊이 있게 해석하는 일은 문학 연구에 의존한다. 따라서 문학 연구의 텍스트 분석도 데이터와 문학작품, 그리고 작품의 언어 및 서사에 대한 지식과 시각에서 출발해야 한다는 주장도 타당하다. 즉, 계량적 지식과 비평적 해석의 이분법적 구분을 피하고, 인공지능 전산 적용에 과도하게 의존하거나 문학연구의 의미를 축소하지 않으면서도 디지털 문학연구의 올바른 자리매김을 위해 디지털과 문학연구의 상호 관계규정에 대한 폭넓은 논의가 필요하다.

분석 과정을 살펴보면, 먼저 텍스트 네트워크 분석 전 단계로 동시출현 빈도분석은 제인 오스틴의 소설 엠마에서 각 등장인물이 함께 언급된 횟수를 수치화한 것이었다. 이는 소설 내 인물들 간의 관계 밀접도와 상호작용 정도를 나타내는 지표로, 각 인물 간의 연결 강도와 주요 사회적 관계를 파악하는 데 유용하다. 분석 결과, 엠마와 나이틀리의 관계가 가장 높게 나타났으며, 그 다음으로 엠마와 해리엇, 나이틀리와 해리엇, 엠마와 웨스턴 부인의 순서로 나타났다. 구체적으로 엠마와 나이틀리는 빈도수가 49로 가장 높아, 이들이 서로에게 중요한 존재임을 강조했으며, 엠마와 해리엇은 빈도수 44로 긴밀한 관계를 나타냈다. 엠마와 프랭크는 빈도수 35로 이들의 상상 속 로맨스를 반영하고, 엠마와 엘튼씨의 빈도수 36은 엇갈린 짝사랑 관계를 드러냈다. 해리엇과 나이틀리는 빈도수 40으로 함께 언급되며, 나이틀리가 자신을 사랑한다고 믿는 해리엇의 착각을 시사했다. 우드하우스 씨는 빈도수 3으로 배경 인물임을 암시하고, 처칠 부인은 낮은 동시출현 빈도로 소설 내에서 고립된 위치에 있음을 나타냈다.

텍스트 네트워크상의 연결 정도 중심성 분석에서 엠마는 463으로 가장 높은 값을 기록하며,

이는 그녀가 소설 내에서 가장 중심적이고 활동적인 인물임을 보여주었다. 나이틀리는 425로 두 번째로 높아, 두 인물이 소설의 사회적 상호작용에서 핵심적인 역할을 수행했음을 시사했다. 근접 중심성에서 주요 인물들의 값은 대부분 18로 일정했으나, 우드하우스 씨와 처칠 부인, 프랭크는 각각 20과 19의 값을 가져 네트워크 내에서 상대적으로 가까운 위치에 있음을 나타냈다. 에이겐벡터 중심성에서 엠마는 0.376으로 가장 높아 중요한 인물들과 강하게 연결되어 있음을 나타냈으며, 나이틀리와 해리엇도 높은 값을 기록했다. 반면, 우드하우스 씨와 처칠 부인은 낮은 값으로 다른 인물들과의 연결은 효율적이나 영향력은 낮았다. 매개 중심성에서는 대부분의 인물들이 0.125로 동일한 값을 가져 많은 인물들이 중재자 역할을 공평하게 수행했음을 보여주었지만, 프랭크, 처칠 부인, 우드하우스 씨는 0으로 중재 역할을 거의 수행하지 않았음을 나타냈다.

이 연구에서 활용한 텍스트 네트워크 분석의 핵심은 네트워크의 노드와 엣지, 문학 작품에서는 주로 등장인물과 인물 간 상호관계의 식별에 있다. 정보과학자에 따르면, 문학연구에서 인물 식별은 기술적 어려움 때문에 여러 우회로가 사용된다(Labatut and Bost 2019). 인물사전 등의 자료를 활용할 수 없는 경우 수동 처리를 선호하며, 부분 자동화 처리 시 대명사와 상호참조를 정보가치 없는 데이터로 제거하거나, 너무 많은 인물을 식별해야 할 때 출현 빈도가 낮은 인물을 제거하는 단순한 방법을 택한다는 것이다. 자연어 처리과정에서 대명사와 상호참조 식별 성공률이 높지 않다는 점을 고려할 때, 인물 간 상호관계 식별의 수학적 엄밀성을 얻는 것은 더 어려운 일이다. 실제로 등장하지 않고 언급만 되는 많은 하이베리 주민을 계측 대상에 포함하는 관점에서 보면, 출현 빈도가 낮은 인물을 제거하는 것은 중요한 데이터의 손실로 이어질 수 있다. 분석 대상과 범위, 주제에 따라 적절한 방법을 선택하는 것이 중요하겠으나, 이에 대한 합의된 기준이 부족하며, 많은 연구에서 관련 정보제공과 설명에 소극적이다. 디지털 문학의 계량적 엄밀성 향상이 인공지능 기반 전산 분야의 기술적 발전에 크게 의존한다. 이는 디지털 문학의 융합적 성격을 말해주는 강점이자 제약이 될 수 있다.

이 연구를 통해 소설 엠마의 각 토픽이 소설의 주요 테마와 어떻게 연결되며, 인물들의 상호작용과 사건들이 어떻게 텍스트로 표현되고 있는지에 대한 새로운 이해와 시각을 가질 수 있었다. 이는 소설의 다층적인 이해와 감상을 가능하게 하며, 제인 오스틴의 문학 작품에 대한 학문적 탐구에도 새로운 비정형 양적 데이터 분석 결과와 해석을 제공하였다. 소설 엠마 텍스트에 대한 토픽 모델링 결과를 살펴보면, 첫째, 토픽 1인 사회적 연결과 개인적 상호작용에서는 'Mrs. Elton', 'Robert Martin', 'Isabella', 'John'과 같은 인물들이 주요 키워드로 등장하며, 이들은 소설 내에서 다양한 사회적 상호작용과 개인적 관계를 형성한다. 'Mrs. Elton'과 'Robert Martin'은 각각 소설의 사회적 계급과 결혼 문제를 중심으로 활동하는 인물로, 이들의 동적인 관계가 주요 사건들을 주도한다. 예를 들어, Mrs. Elton의 사회적 야심과 Robert Martin의 결혼 제안은 하이베리 커뮤니티 내에서 다양한 반응을 불러일으킨다.

둘째, 토픽 2에서 'Campbell', 'Perry', 'London', 'Carriage'와 같은 단어들은 토픽 2에서 외부 세계와의 상호작용을 나타낸다. 특히 'London'과 'Carriage'는 인물들의 이동과 여행을 의미하며, 이는 하이베리와 더 넓은 사회적 맥락 간의 연결고리를 상징한다. 소설에서 런던은 문화적, 사회적 기회의 중심지로서 하이베리의 인물들에게 다양한 경험과 도전을 제공한다.

셋째, 토픽 3에서는 'Dance', 'Ball', 'Mother', 'News' 등의 단어를 통해 소설 내의 사회적 이벤트와

개인적 경험을 연결한다. 이 토픽은 특히 사회적 행사에서 발생하는 인간관계의 변화와 개인적 감정의 교류를 다루며, 이러한 사건들은 인물들의 성장과 변화에 중요한 영향을 미친다. 예를 들어, 댄스와 볼은 인물들이 서로를 더 깊이 이해하고 감정적 유대를 강화하는 장으로 기능한다.

넷째, 'John', 'Robert Martin', 'Isabella', 'Snow' 등은 토픽 4에서 계절의 변화와 그에 따른 인물의 감정 교류를 나타낸다. 특히 'Snow'는 겨울철의 자연적 어려움과 인물들 간의 내면적 갈등을 상징하며, 이는 인물들의 개인적 고백과 감정적 변화로 이어진다. 이러한 계절적 변화는 소설 내에서 인물들의 관계를 재정립하고, 갈등을 해결하는 데 중추적인 역할을 한다.

지금까지 데이터 기반의 디지털 인문학적 관점에서 소설 엠마의 스토리 패턴을 미시적으로 분석하였다. 그러나 LDA 방법을 사용하여 문학작품을 토픽으로 군집화하는 과정에서 모든 토픽에 대해 유의미한 문학적 해석을 도출하는 데에는 한계가 있었다. 특히, 토픽 별 주요 단어의 의미를 추론하는 과정에서 등장인물이 아닌 단어들의 명확한 해석이 어려웠다. 또한, 질적 인문학 연구에서는 중요한 쟁점이 될 수 있는 서술자가 의도적으로 감추는 지점이나 숨겨진 문맥을 통계적으로 환산하는 것에는 한계가 있었다.

지금까지 비정형화된 영문 고전문헌을 효과적으로 처리하고 분석하는 방법론에 대한 연구가 부족한 상황에서, 영문학 텍스트의 특성을 딥러닝 기반을 둔 계량적 접근으로 규명하려는 연구가 많지 않았다. 이러한 상황에서 소설 엠마를 텍스트 마이닝의 일종인 토픽 모델링으로 분석한 점에서 이 연구는 의의가 있다. 나아가, 이러한 토픽 모델링 분석이 디지털화된 대량의 영문 고전문헌을 분석하는 데 유용한 방법이 될 수 있다. 토픽 모델링은 연구자에게 해당 텍스트의 전체적인 특성과 주요 내용을 파악하는 데 도움을 주며, 토픽 모델링과 해당 영문 고전문헌에 대한 전문지식이 결합하면 텍스트에 대한 새롭고 다양한 관점을 유도할 수 있다. 또한, 기존 연구에서 예상하지 못한 잠재적 의미구조를 발견할 수 있다. 향후 대량의 텍스트가 보다 정확하게 디지털화되고, 다양한 영문 고전문헌에 대한 형태소 분석과 NLP가 연구자의 요구사항에 맞게 적절히 수정·보완된다면, 토픽 모델링을 이용한 영문 고전문헌 분석이 보다 활성화될 것으로 기대된다.

감성분석 결과에 따르면, 제인 오스틴의 소설 엠마는 각 권마다 다른 감정적 흐름을 보여주었다. 첫 번째 권에서는 엠마가 주변 인물들의 연애와 결혼을 적극적으로 주선하려는 모습이 주를 이루며, 'good(184회)'과 'great(117회)' 등의 긍정적 단어가 많이 사용되어 엠마의 긍정적이고 자신감 넘치는 태도를 반영하였다. 두 번째 권에서는 엠마의 계획이 실패하며 갈등과 실망이 두드러지며, 'bad(34회)', 'few(33회)' 등의 부정적 단어가 빈번하게 나타나 엠마의 오만이 초래한 부정적 영향을 보여주었다. 세 번째 권에서는 엠마가 자신의 행동을 반성하고 성숙해지며 관계를 재정립하는 과정이 그려지며, 'good(149회)'과 'great(95회)' 등의 긍정적 단어가 많아져 갈등의 해결과 감정적 회복을 나타낸다. 이러한 분석을 통해 소설의 전반적인 감정적 흐름과 주요 테마를 이해할 수 있으며, 제인 오스틴이 묘사한 인간관계의 복잡성과 심리적 변화를 객관적으로 추론할 수 있다.

이 연구에서는 감성분석에 감성사전을 활용하였고, 사전에 등록되지 않거나 양면성을 가지는 단어에 대한 감성점수를 따로 설정하였다. 이로 인해 작품의 흐름 속에서 긍정적인 문맥을 가진 문장을 일부 부정적으로 판단했을 가능성도 있었다. 이러한 문제점은 긍정과 부정의 이분법 대신에 감성의 정도를 보다 세밀하게 수치화하거나, 분노, 행복 등과 같은 하위 감성 범주를

추가하여 계산하는 방법을 통해 어느 정도 해결될 수 있을 것이다. 따라서, 이러한 개선 방법을 후속 연구를 위한 제언으로 남기고자 한다.

영문학 작품을 통해 대학 강의실에서 얻고자 하는 목표는 다양하지만, 무엇보다도 외국어로서의 영어 사용 가능성과 이해 능력의 함양이 최우선일 것이다. 또한, 문학적 체험, 문화 예술적 교양 함양, 기타 필요한 지식이나 정보의 입수도 중요하다. 이에 본 논문은 현대 사회에서 절실히 요구되는 국제어로서의 영어교육에 대한 실용적인 의사소통 능력을 함양하고, 이해 능력을 향상시킬 수 있는 방안으로서 문학 작품 텍스트를 활용한 영어교육 방법을 제안하여 그 효과와 구체적인 의의를 논하고자 하였다.

특히, 이 연구는 NLP 기법과 인공지능 알고리즘의 도움을 받아 영문학 작품의 감상과 이해과정을 보여줌으로써 영어교육 방법에 대한 시사점을 얻으려 하였다. NLP 기법을 통해 영문학 작품의 특성을 객관적으로 분석함으로써, 텍스트를 더 깊이 이해할 수 있도록 돕고, 학습자 맞춤형 교육 콘텐츠를 개발하여 효과적인 영어 학습을 지원할 수 있을 것이다. 이를 통해 학생들이 문학작품을 통한 표현 능력과 이해 능력을 함양하고, 영어교육에 있어 실질적인 의사소통 능력을 향상시킬 수 있는 방안을 마련할 수 있다.

또한, 직무 영어, 영어 발표와 토론과 같은 진로 선택과목과 새로운 영어 융합 선택교과목으로 개설되고 있는 세계문화와 영어와 미디어 영어 과목에서 자연어처리 기법과 같은 인공지능 테크놀로지를 활용한 번역, 독해, 말하기 및 쓰기 등의 다양한 디지털 영어 교수학습 활동의 실제 적용 방안이 논의되어야 한다. 따라서, 한국의 영어교육 현실을 고려할 때 영어 문학작품 텍스트 분석을 통한 보다 개선된 수업 방식이 학습자들의 요구에 부응하는 통합적인 영어교육 방법이 될 수 있을 것이다. 특히, 영어교사의 독해수업과 영어 학습지도 시 학생들에게 영어 텍스트 독해지문을 이해시킬 때 (유사)단어의 빈도와 단어 간 상호 연결 망 관계를 그려보고 중심 단어를 찾아내고, 다양한 기준을 통한 단어들의 분류와 유목화를 통해 개별 주제와 토픽을 유추하는 인공지능 알고리즘 학습 방식을 소개하고 실제로 간단한 코딩을 통해 실습 및 적용해 볼 수도 있을 것이다.

이 연구의 의의는 영어 독해와 이해교육 방법에 인공지능 기술을 접목한 새로운 시각을 제공하고자 하는데 있다. 첫째, NLP 기법을 활용한 텍스트 분석은 영어 학습자들에게 보다 명확하고 객관적으로 문학 텍스트를 이해하는 기회를 제공할 수 있다. 텍스트 네트워크 분석을 통해 인물 간의 관계와 상호작용을 수치와 시각적으로 제시함으로써, 학습자들은 문학 작품의 복잡한 내러티브를 보다 쉽게 이해할 수 있게 된다. 이는 전통적인 독해 교육에서 학습자들이 느낄 수 있는 추상적인 영어 텍스트 이해의 어려움을 효과적으로 줄여줄 수 있다. 둘째, 토픽 모델링을 활용한 주요 주제 분석은 학습자들이 문학작품의 핵심 주제와 하위 주제를 근거를 바탕으로 파악하도록 돕는다. 이는 학습자들이 토픽 별 키워드 단어를 중심으로 텍스트의 주요 내용을 효과적으로 정리하고, 이를 바탕으로 논리적 사고를 발전시키는 데 기여할 수 있다. 실제로 소설 작품에서 사회적 연결과 개인적 상호작용, 결혼과 계층 문제 등의 주제를 토픽 모델링으로 분석함으로써, 학습자들은 작품의 주요 테마를 체계적으로 이해하고, 이를 바탕으로 심화 영어학습을 수행할 수 있을 것이다. 셋째, 감성분석을 통해 텍스트 어휘의 감정을 수치로 파악함으로써, 학습자들은 문학 작품의 감정적 변화를 인지하고, 이를 통해 보다 깊이 있는 영어 독해가 가능해진다. 소설의 스토리 전개 과정별 감성분석 결과를 통해 학습자들은 등장인물의

인물 변화와 이야기의 전개 과정을 더 잘 이해할 수 있다. 이러한 접근은 학습자들이 텍스트의 세부적인 부분을 더 잘 이해하도록 도와주며, 이는 영어 독해 능력의 향상으로 이어진다.

결론적으로, 영어 교육에 대한 이러한 디지털 인문학적 접근은 영어 교수자 및 학습자들에게 인공지능 기술을 적용한 텍스트 분석 도구와 방법론을 제시함으로써, 디지털 영어교육을 위한 인공지능 테크놀로지 소양과 에듀테크 활용 역량 함양에 대한 관심을 높일 수 있을 것이다. 또한 NLP 기법을 활용한 문학 텍스트 분석은 전통적인 영어 독해학습 방식에 비해 학생들의 흥미와 동기부여를 야기할 수 있는 디지털 영어학습 환경을 조성한다. 이는 에듀테크 영어 교수학습 환경에서 학습자들의 학습 몰입과 참여도를 높이고, 영어 독해와 이해 능력을 효과적으로 향상시키는 데 도움을 줄 수 있을 것이다.

참고문헌

곽기영(Kwak, K.-Y.). 2014.『소셜네트워크분석』(*Social network analysis*). 서울: 청람(Seoul: Cheongram).

김상락(Kim, S. R.). 2005. 문학 작품에서의 복잡계 연결망 분석: 소설 토지를 중심으로(Complex network analysis in literature: Togi). ≪새물리≫(*New Physics: Sae Mulli*) 50-4, 267-271.

김선영(Kim, S. Y.). 2022. 새마을운동 관련 사회적 이슈 탐색 및 의미에 관한 연구: 뉴스 빅데이터의 LDA 기반 토픽분석을 중심으로(Exploration of the social issues related to Saemaul Undong, and the Meaning: Focusing on LDA topic analysis of news big data). ≪사회적경제와 정책연구≫(*Social Economy & Policy Studies*) 12-2, 151-178.

김용규(Kim, Y. K.). (역). (2012).『멀리서 읽기: 세계문학과 수량적 형식주의』(Franco Moretti의 *Distance Reading*). 서울: 현암사.

박주섭·홍순구·김종원(Park, J. S., S. G. Hong and J. W. Kim). 2017. 토픽 모델링을 활용한 과학기술동향 및 예측에 관한 연구(A Study on Science Technology Trend and Prediction Using Topic Modeling). ≪한국산업정보학회논문지≫(*Journal of the Korea Industrial Information Systems Research*) 22-4, 19-28.

원영선(Won, Y. S.). 2016. 하이베리 탐구: 제인 오스틴의 [엠마](Exploring Highbury in Jane Austen's Emma). ≪19세기 영어권 문학≫(*Nineteenth Century Literature in English*) 20-1, 95-121.

원영선(Won, Y. S.). 2023. 디지털 문학연구의 탐색과 적용: 엠마의 소셜네트워크 분석(Exploring digital literary study, an application: Social network analysis of Emm). ≪19세기 영어권 문학≫(*Nineteenth Century Literature in English*) 27-2, 39-70.

이승은(Lee, S. E.). 2022. 텍스트 마이닝을 활용한 근대 조리서의 분석 연구*(Study of Korean Modern Cookbooks Using Text Mining Analysis)*. 박사학위논문(Doctoral dissertation), 이화여자대학교, 서울, 한국.

장민서·오수진·김응모(Jang, M., S. Oh and U. M. Kim). 2018. TF-IDF를 활용한 k-means 기반의 효율적인 대용량 기사 처리 및 요약 알고리즘(*Article analytic and summarizing algorithm by facilitating TF-IDF based on k-means)*. Paper presented at the Korea Information Processing Society Conference, 271-274.

정성훈(Jeong, S.-H.). 2014. 현대 한국어 부사에 대한 계량언어학적 연구-확률 통계 모형과 네트워크를 이용한 분석(*The quantitative linguistical study on modern Korean adverbs: Using probability-statistical model and network model*). 박사학위논문(Doctoral dissertation), 서울대학교 대학원, 서울, 한국.

최연무(Choi, Y. M.). 2004. 복잡계 네트워크로서의 그리스 신화(Greek myth as a complex network). ≪새물리≫(*New Physics: Sae Mulli*) 49-3, 298-302.

최은샘·정채관(Choi, E. and C. K. Jung). 2021. 영미 아동 모험 소설에 관한 코퍼스 분석 연구: 보물섬을 중심으로(A corpus analysis of British-American children's adventure novels: Treasure). ≪한국콘텐츠학회논문지≫(*The Journal of the Korea Contents Association*) 21-1, 333-342.

하명정(Ha, M. J.). 2013. 코퍼스에 기반한 문학 텍스트 분석(Corpus-based literary analysis). ≪한국콘텐츠학회논문지≫(*The Journal of the Korea Contents Association*) 13-9, 440-447.

홍주현·나은경(Hong, J. H. and E. K. Na). 2015. 세월호 사건 보도의 피해자 비난 경향 연구: 보수 종편 채널 뉴스의 피해자 범주화 및 단어 네트워크 프레임 분석: 보수 종편 채널 뉴스의 피해자 범주화 및 단어 네트워크 프레임 분석(Victim Blaming of Sewal-ferry Disaster on News in Conservative Total TV Programming: Categorization of Victims and Word Network Analysis). ≪한국언론학보≫(*Korean Journal of Journalism & Communication Studies*) 59-6, 69-106.

Blei, D. M., A. Y. Ng and M. I. Jordan. 2003. Latent dirichlet allocation, *Journal of machine Learning research 3*, 993-1022.

Borgatti, S. P., A. Mehra, D. J. Brass and G. Labianca. 2009. Network analysis in the social sciences. *Science 323*(5916), 892-895.

Chae, S. H., J. I. Lim and J. Kang. 2015. A comparative analysis of social commerce and open market using user reviews in Korean mobile commerce. *Journal of Intelligence and Information Systems* 21(4), 53-77.

Chen, C. and C. Chang. 2019. A Chinese ancient book digital humanities research platform to support digital humanities research. *The Electronic Library* 37(2), 314-336.

Cho, H., J. Kang and D. Y. Jeong. 2016. An exploratory study on mobile app review through comparative analysis between South Korea and US. *Journal of Information Technology Services* 15(2), 169-184.

Cho, H. J., S. G. Kim and J. Y. Kang. 2017. An empirical analysis of doppelgänger brand image effects: Focused on the Internet community. *The Journal of Information Systems* 26(1), 21-51.

Choi, S. R. and J. W. Yoo. 2014. Present of the analysis method of the validation between the story proceeding and the character-by the generative trajectory of meaning with Greimass and Enneagram. *Journal of Digital Design* 14(2), 139-147.

Chung, P., H. Ahn and K. Y. Kwahk. 2019. Identification of core features and values of smart phone design using text mining and social network analysis. *Korean Business Association* 32(1), 27-47.

Elson, D. K., K. McKeown and N. J. Dames. 2010. Extracting social networks from literary fiction. In *Workshop Proceedings of TextGraphs-7: Graph-based Methods for Natural Language Processing*, 6-14.

Fischer, F. and D. Skorinkin. 2021. Social network analysis in Russian literary studies. In Daria Gritsenko et al. eds., *The Palgrave Handbook of Digital Russia Studies*, 517-536. Basingstoke: Palgrave Macmillan.

Hassan, A., A. Abu-Jbara and D. Radev. 2012. Extracting signed social networks from text. In *Workshop Proceedings of TextGraphs-7: Graph-based Methods for Natural Language Processing*, 6-14.

Heuser, R. and L. Le-Khac. 2012. *A Quantitative Literary History of 2,958 Nineteenth-century British Novels: The Semantic Cohort Method*. Stanford Literary Lab.

Jockers, M. L. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Kang, B. M. 2010. Constructing networks of related concepts based on co-occurring nouns. *Korean Semantics* 32, 1-28.

Kang, D. J. and K. N. Lee. 2015. A study on co-author networks of journal of Korea trade research association using social network analysis. *Korea trade review* 40(5), 1-23.

Karl, A., J. Wisnowski and W. H. Rushing. 2015. A practical guide to text mining with topic extraction. *Wiley Interdisciplinary Reviews: Computational Statistics* 7(5), 326-340.

Kim, S., H. J. Cho and J. Y. Kang. 2016. The status of using text mining in academic research and analysis methods. *Journal of Information Technology and Architecture* 13(2), 317-329.

Kim, S. G. and J. Kang. 2018. Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews. *Information Processing & Management* 54(6), 938-957.

Labatut, V. and X. Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)* 52(5), 1-40.

Medhat, W., A. Hassan and H. Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5(4), 1093-1113

Mimno, D. 2012. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage (JOCCH)* 5(1), 1-19.

Moretti, F. 2013. *Distant Reading*. NY: Verso Books.

Oh, Y. L., J. O. Min, Y. G. Kim., D. J. Kim., Y. K. Park and B. G. Lee. 2017. A comparative analysis for the extraction of similar patent claims based on word embedding. In *Paper presented at the Korean Institute of Information Scientists and Engineers*, 20-22.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2), 1-135.

Park, G. M., S. H. Kim and H. G. Cho. 2013. Analysis of social network according to the distance of characters statements. *The Journal of the Korea Contents Association* 13(4), 427-439.

Ramsay, S. 2011. *Reading Machines: Toward and Algorithmic Criticism*. University of Illinois Press.

Rhody, L. M. 2012. Topic modeling and figurative language. *Journal of Digital Humanities* 2(1), 17-38.

Sawng, Y. W. and S. J. Lee. 2018. Analysis on the research trend of medical automation industries utilizing the keyword network analysis. *Korean Association of Business Education* 33(2), 225-242.

Schreibman, S., R. Siemens and J. Unsworth. 2008. *A Companion to Digital Humanities*. Oxford: Blackwell Publishing.

Scott, J. 2012. *What is Social Network Analysis?* NY: Bloomsbury Academic.

Silge, J. and D. Robinson. 2017. *Text Mining with R: A Tidy Approach*. Sebastopol: O'Reilly.

Smeets, R. 2021. *Character Constellations: Representations of Social Groups in Present-day Dutch Literary Fiction*. Leuven University Press.

Stiller, J., D. Nettle and R. I. Dunbar. 2003. The small world of Shakespeare's plays. *Human Nature, 14*, 397-408.

Tsvetovat, M. and A. Kouzentsov. 2011. *Social Network Analysis for Startups*. Sebastopol, CA: O'Reilly.

예시 언어(Examples in): English
적용가능 언어(Applicable Languages): English
적용가능 수준(Applicable Level): Tertiary

