The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 21, No. 2, pp.195-203, Apr. 30, 2021. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2021.21.2.195
JIIBC 2021-2-27

LDA 기반 사용자 감정분석을 위한 문서 토픽 추출 시스템에 대한 연구

A Study on the Document Topic Extraction System for LDA-based User Sentiment Analysis

안윤빈*, 김학영*, 문용현*, 황승연**, 김정준***

Yoon-Bin An*, Hak-Young Kim*, Yong-Hyun Moon*,
Seung-Yeon Hwang**, Jeong-Joon Kim***

요 약 최근 IT 분야의 주요 기술인 빅데이터는 다양한 산업 분야로 확장되고 있으며 활용 방안에 대한 연구가 활발하게 진행 중이다. 대부분의 인터넷 산업 분야에서 사용자 리뷰는 이용자가 상품 구매를 결정하는 데 많은 도움을 준다. 그러나 방대한 제품 리뷰에서 긍정, 부정적 의미와 도움이 되는 리뷰를 선별하는 과정은 제품 구매 결정에 있어 많은 시간을 요구한다. 따라서 본 논문에서는 빅데이터 분석 기술인 LDA를 이용해 키워드를 분석 및 종합하여 사용자에게 의미 있는 정보를 제공하는 시스템을 설계하고 구현한다. 문서 토픽 추출을 위해 본 연구에서는 국내 도서 산업을 도메인으로 데이터를 크롤링하고, 빅데이터 분석을 실시한다. 이를 통해 사용자 리뷰의 토픽 및 감정단어를 바탕으로 상품에 대한 종합적인 정보를 제공함으로써 구매자에게 도움을 주고 나아가 리뷰 현황 분석을 통해 상품의 전망 또한 파악할 수 있다.

Abstract Recently, big data, a major technology in the IT field, has been expanding into various industrial sectors and research on how to utilize it is actively underway. In most Internet industries, user reviews help users make decisions about purchasing products. However, the process of screening positive, negative and helpful reviews from vast product reviews requires a lot of time in determining product purchases. Therefore, this paper designs and implements a system that analyzes and aggregates keywords using LDA, a big data analysis technology, to provide meaningful information to users. For the extraction of document topics, in this study, the domestic book industry is crawling data into domains, and big data analysis is conducted. This helps buyers by providing comprehensive information on products based on user review topics and appraisal words, and furthermore, the product's outlook can be identified through the review status analysis.

Key Words : Big Data, Data Analysis, LDA, Topic Modeling, Web Crawling, Sentiment Analysis

*준회원, 한국산업기술대학교 컴퓨터공학과
**준회원, 안양대학교 컴퓨터공학과 석사과정
***정회원, 안양대학교 ICT융합학부 소프트웨어전공 교수 (교신저자)
접수일자 2020년 11월 26일, 수정완료 2021년 3월 8일
게재확정일자 2021년 4월 9일

Received: 26 November, 2020 / Revised: 8 March, 2021 / Accepted: 9 April, 2021
***Corresponding Author: jjkim@anyang.ac.kr
Dept. ICT Convergence Engineering at Anyang University, Korea.

Ⅰ. 서 론

최근 IT 경향을 주도하는 사물인터넷(Internet of Things), 인공지능(Artificial Intelligence) 등 최신 IT 기술들을 공통으로 관통하는 키워드는 '빅데이터'이다.[1] 빅데이터는 기업, 공공, 통신, 금융, 의료, 유통 들 점점 다양한 산업 분야로 뻗어 나가고 있으며, IT 기술의 발전이 가속화될수록 빅데이터 시장도 확대되고 있다.[2] 우리나라도 현재 13대 혁신성장동력 분야를 선정하여 지원하고 있으며 그 중 정보통신기술(ICT) 분야 안에 빅데이터가 주요 분야로 자리 잡고 있다.[3] 빅데이터는 많은 양의 정보 그 자체만을 의미하는 것이 아니라 그 정보를 의미 있는 값으로 활용하는 방법이 중요하다.[4] 정보들을 활용하는 방법들에는 여러 가지가 있는데, 본 연구에서는 LDA 토픽모델링을 이용한다. LDA 토픽모델링을 포함한 데이터마이닝은 숫자 같은 구조화(structured database)되거나, 문자, 문서같이 비구조화(unstructured database)되어있는 데이터베이스를 분석하여 의미 있는 규칙이나 패턴을 도출하는 방법이다.[5]

국내 서적의 거래 규모는 약 3조 원이고, 〈대한출판연감〉에 따르면 2010년에 이미 전체의 40%의 도서가 온라인을 통해 거래되었다고 한다. 즉, 3조 원 중 1조 원이 넘는 거래가 인터넷을 통해 거래되었다고 한다.[6] 인터넷에서 도서를 구매할 때 책을 고르는 기준에는 여러 가지가 있고, 서평도 그중 하나일 것이다. 서평이란 문헌비평, 도서비평, 도서평론 등을 통틀어 부르는 말로 이용자 리뷰를 포함한 비평에 관한 포괄적인 개념이다.[7] 본 연구에서는 대표적인 서평인 "이용자 리뷰"를 감정분석 하여 인터넷 도서 구매에 도움을 주고자 한다.

이용자 리뷰에는 문서의 내용에 관한 단어, 문서를 읽고 느낌 감정에 대한 단어 등 여러 형태로 분석될 수 있는 단어들이 존재한다.[8] 그 데이터들 속에서 의미 있는 데이터를 찾아내는 방법의 하나가 데이터 마이닝 기법이다.[9] 본 연구에서는 이용자 리뷰 속 감정에 관한 단어만을 골라낸 후, 감정분석하여 리뷰가 부정적 의미가 있는지 긍정적 의미를 지니는지 판단할 수 있게 한다.[10]

감정분석을 통해 직접 이용자 리뷰를 읽지 않아도 리뷰가 긍정적인 평가를 하고 있는지, 부정적인 평가를 하고 있는지 알 수 있게 된다면 빨리 많은 리뷰들을 분석하여 특정 문서에 대한 다수 의견을 알 수 있게 된다. 감정분석 할 리뷰를 모으는 방법으로 본 연구에서는 웹 크롤링을 사용한다. 웹 크롤링이란, 인터넷 웹 검색 엔진(Web Searching Engine)에서 이미지 데이터를 포함하여 사용자가 원하는 데이터를 자동으로 수집하는 기법이다.[11] 웹 크롤러의 사용도 빅데이터가 주목받기 시작하면서 늘어나고 있는데, 대표적인 활용 사례로는 2008년 미국 대통령 선거, 구글의 맞춤형 광고, 아마존닷컴의 추천 상품 표시 등이 있다.[12]

본 논문은 다음과 같이 구성된다. 2장에서는 LDA기법을 포함한 관련 연구들에 대해 설명하고 3장에서는 웹 크롤링부터 토픽 추출 시스템 구조와 감정분석에 대해 설명한다. 그리고 4장에서 결론을 맺는다.

Ⅱ. 관련 기술

1. 빅 데이터 분석

빅 데이터 기술 분야는 다양하며 시스템적으로 복잡한 기능과 구성을 가지고 있어서 분야별로 전문화된 기술과 제품의 통합이 필요하다. 빅 데이터 기술은 크게 데이터 수집 기술, 저장기술, 처리 기술, 분석기술, 표현 및 활용 기술, 관리(Infra, Biz) 기술 분야로 나눌 수 있다. 그중 분석 기술은 검증된 통계적 기법 기반의 고급 분석과 실시간 분석, 사용자와 상호작용하는 탐색적 데이터 분석 기술 등이 요구된다. 그러나 데이터 분석을 위한 많은 연산 시간과 높은 비용, 통계 분석 기법의 프로그램 구현과 검증 문제, 분석 전문가의 부재 등은 빅 데이터 분석의 어려운 점으로 꼽힌다[13].

2. 웹 크롤링

웹 크롤링은 자동화된 방법으로 웹에서 데이터를 수집하는 것을 의미하며, 웹 페이지에 대량의 정보를 수집할 때 유용하게 사용된다. 본 연구에서는 교보문고 웹 사이트에 있는 책의 리뷰와 책 이미지를 가져오는 웹 크롤링 시스템을 소개한다.

3. 감정분석

감정 분석은 자연어 처리, 텍스트 분석, 컴퓨터 언어학 및 생체 인식을 사용하여 정서 상태와 주관적인 정보를 체계적으로 식별, 추출, 정량화 및 연구하는 것을 뜻한다. 감정 분석은 리뷰 및 설문 조사 응답, 온라인 및 소셜 미디어와 같은 고객의 음성 자료, 마케팅에서 고객 서비스, 임상 의학에 이르기까지 다양한 분야에 널리 적용된다.

본 연구에서는 자연어 처리를 이용해 수집된 책 리뷰

를 단어로 정제한 뒤, 감정사전에 있는 단어와 매칭 시켜 긍정, 부정, 중립으로 나누어 감정을 분석하여 시각화하였다.

4. LDA

LDA(Latent Dirichlet Allocation)는 대표적인 토픽 모델링 기법으로 문서들이 하나 이상의 토픽을 포함하고 있다고 가정한다. 단어의 순서는 중요하지 않고 출현 빈도만으로 토픽을 추출하며 미리 알고 있는 토픽별 단어 수 분포를 바탕으로, 주어진 문서에서 발견된 단어 수 분포를 분석하여 해당 문서가 어떤 토픽들을 가지고 있을지를 예측하는 기법이다. 본 연구에서는 사용자 리뷰 분석을 위해 베이즈 추론을 이용한 깁스 샘플링 방법을 이용하여 토픽을 할당한다[14].

5. R

R 프로그래밍 언어는 뉴질랜드 오클랜드 대학의 로스 이하카(Ross Ihaka)와 로버트 젠틀맨(Robert Gentleman)에 의해 1995년에 최초 버전이 소개되었다. 오픈 소스 임에도 고성능의 컴퓨팅 속도와 데이터 처리 능력, 구글이나 아마존 클라우드 서비스와의 API 제공으로 연동성과 호환성이 좋다. R은 통계 분석, 데이터 마이닝, 시각화를 효율적으로 제공하는 언어이다. 특히, 빅데이터 분석을 목적으로 주목을 받고 있으며, 5,000개가 넘는 패키지들이 다양한 기능을 지원하고 있으며 수시로 업데이트되고 있다.

R은 통계 프로그래밍 언어인 S 언어 기반으로 만들어졌으며 통계 계산과 결과 생성 그래픽을 위한 프로그래밍 언어에 해당한다. R이 다른 개발 언어와의 연계 호환이 가능하고, 웹과 연동하여 실시간 처리가 가능하다는 점에서 개발자에게 매력적이다. R은 비용 절감에 따른 경제적 이익이 수반되는 새로운 애플리케이션을 개발하거나 웹 서비스로 제공하는 데 유용하다[15].

Ⅲ. 시스템 구조 및 상세 프로세스

문서 토픽 추출 시스템은 웹 크롤링, 문서 토픽 추출, 감정 분석, 웹 시각화 4가지 과정으로 구성된다. 웹 크롤링 프로세스는 Python 언어의 selenium을 이용해서 구현하였고, 분석 과정과 웹 시각화는 R 언어를 이용하였다.

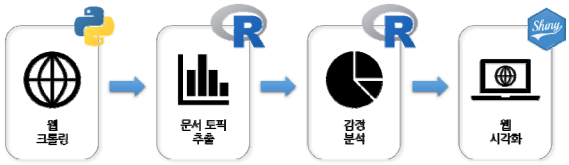


그림 1. 문서 토픽 추출 시스템 흐름도
Fig. 1. Document Topic Extraction System

1. 웹 크롤링

웹 크롤링 과정은 토픽 추출과 감정 분석에 사용할 수 있도록 사용자가 수집할 데이터를 선택하고 해당 데이터를 텍스트 파일로 저장한다.

그림 2는 python의 pyqt5 라이브러리를 이용하여 구현한 gui 프로그램이다. 사용자가 빈칸에 검색하고 싶은 책을 입력하고 리뷰 검색 버튼을 누르면 selenium 라이브러리를 이용한 웹 크롤링을 시작한다.

그림 2. 웹 크롤링 GUI
Fig. 2. Web crawling GUI

그림 3은 리뷰 데이터의 분석 과정에 불필요한 특수문자를 제거하고, 책 이름의 폴더로 저장하는 코드이다.

크롤링이 완료된 리뷰 데이터를 저장하기 위해 selectedBookName 변수에 저장되어 있는 책 이름으로 폴더를 만든다. reviewC 변수에 저장되어 있는 텍스트만 reviewContent에 저장하고, replace 함수를 이용해 특수문자를 제거한다. savedDir에 경로 설정을 위한 selectedBookName 폴더 이름을 경로로 설정하고, 리뷰 텍스트 파일의 이름을 savedName으로 저장한다.

그림 4는 크롤링이 완료된 리뷰 데이터 3개를 예시로 보여준다. 교보문고 사이트에서 '아몬드' 책의 리뷰를 크롤링 하여 저장된 데이터이다.

```
    Path("./" +
selectedBookName).mkdir(parents=True,
exist_ok=True)
    savedDir = "./" + selectedBookName + "/"
    savedName = str(reviewNum) + ".txt"
    reviewContent = reviewC.get_text()
    reviewContent = " ".join(reviewContent.split())
    reviewContent = reviewContent.replace('〉',
'').replace('〈', '').replace('*'
'').replace('-','').replace('#','').('^','').replace('~','')
    path = savedDir + savedName
    with open(path, "w", encoding="cp949",
errors='ignore') as f:
      f.write(reviewContent)
      print(reviewNum, "번째 리뷰가 저장되었습니다.")
    f.close()
```

그림 3. 리뷰 데이터 저장
Fig. 3. Save review data

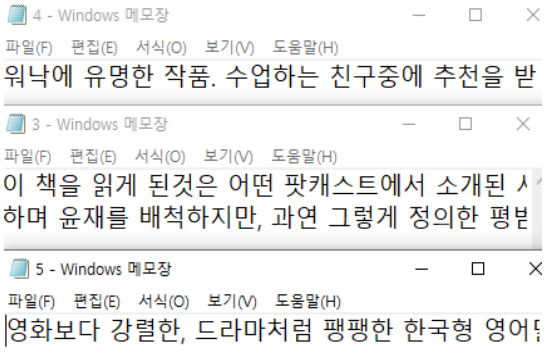


그림 4. 리뷰 데이터 샘플
Fig. 4. Review data sample

2. 문서 토픽 추출

문서 토픽 추출 과정은 웹 크롤링 과정을 통해 저장된 텍스트 파일을 R 언어를 이용하여 추출에 사용할 수 있도록 전처리 과정과 통계를 계산 후 LDA 알고리즘을 이용해 토픽을 추출한다.

그림 5는 크롤링이 완료된 텍스트 파일이 들어있는 폴더를 리스트로 저장하는 코드이다.

1) bookName에 분석할 책의 이름을 입력한다. R script 파일이 들어있는 경로를 setwd 함수를 이용해 기본 폴더로 설정한다. path에 크롤링이 완료된 리뷰 텍스트 파일들을 불러오기 위해 현재 폴더 경로를 저장한다. 2) bookName의 폴더에 있는 파일들을 불러와 list로 저장한다. content에 책의 내용이 들어있는 "content.txt"을 불러온다. 각 리스트 목록에 폴더에 저장되어 있는 파일명을 이름으로 설정하고, 데이터 필드 구분을 “ ”로 설정한다.

```
1   bookName <- "아몬드"
    setwd(dirname(rstudioapi::getSourceEditor
Context()$path))
    path
    <-file.path(dirname(rstudioapi::getSourceE
ditorContext()$path))

2   kor <- list.files(file.path(path,bookName))
    kor.files <- file.path(path, bookName, kor)
    txt <- lapply(kor.files, readLines)
    content
    <-readLines(file.path(path,bookName,"cont
ent.txt"))
    textCnt<-length(txt)-1
    topic <- setNames(txt, kor.files)
    topic <- sapply(topic, function(x) paste(x,
collapse = " "))
```

그림 5. 리뷰 데이터 삽입
Fig. 5. Insert review data

그림 6은 “content.txt" 파일의 일부로, 리뷰 웹 크롤링 과정에 저장된 책 내용 텍스트 파일이다.

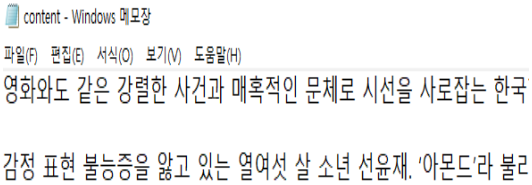


그림 6. 책 내용 데이터 샘플
Fig. 6. Book content data sample

그림 7은 불용어 처리 라이브러리를 사용하여 리뷰를 저장한 리스트에서 단어를 추출하기 위해 전처리를 하는 코드이다.

1) 불용어 목록을 사용하기 위해 텍스트 마이닝 패키지를 불러오고, “SMART" 불용어 사전을 불러와 stop_words에 저장한다. 2) gsub 함수를 이용해 topic에 있는 불용어와 불필요한 숫자를 제거한다. doc.list에 정제한 단어들을 ” “로 구분하고 저장한다.

```
1   library(tm)
    stop_words <- stopwords("SMART")
    stop_words <- c(stopwords("SMART"), "")

2   topic <- gsub("'", "", topic)
    topic <- gsub("[[:punct:]]", " ", topic)
    topic <- gsub("[[:cntrl:]]", " ", topic)
    topic <- gsub("^[[:space:]]+", "", topic)
    topic <- gsub("[[:space:]]+$", "", topic)
    …
    topic <- gsub("[A-Za-z]","",topic)
    doc.list <- strsplit(topic, "[[:space:]]+")
```

그림 7. 리뷰 데이터 정제
Fig. 7. Review data purification

그림 8은 리뷰에서 추출한 단어들을 출현 빈도수와 함께 테이블로 변환하여 저장하는 코드이다.

1) useSejoingDic 함수로 세종 사전을 사용한다. doc.list에 unlist 함수를 사용한다. 2) doc.list를 테이블로 변환하여 단어 출현 빈도수와 함께 term.table에 저장하고 내림차순으로 정렬한다. 불용어 또는 5회 미만으로 언급된 단어들을 제거하고 del에 저장한다. term.table에 있는 단어 중 del에 존재하는 단어를 제거한다. vocab에 term.table에 있는 단어를 저장한다.

```
1   useSejongDic()
    doc.list <- sapply(doc.list, extractNoun, USE.NAMES=F)
    doc.list <- unlist(doc.list)
    doc.list <- Filter(function(x){nchar(x)>1}, doc.list)

2   term.table <- table(doc.list)
    term.table <- sort(term.table, decreasing = TRUE)
    del <- term.table < 5
    term.table <- term.table[!del]
    vocab <- names(term.table)
```

그림 8. 자료구조 변환
Fig. 8. Data structure conversion

그림 9는 데이터 분석을 위한 LDA 모델을 생성하는 코드이다.

vocab의 인덱스를 index에 저장하고, index에서 결측치를 제거한 후, 인덱스 번호를 매겨 각 인덱스의 길이와 결합하는 get.terms 함수를 생성한다. doc.list에 get.terms 함수를 적용하고 documents에 저장한다.

```
get.terms <- function(x) {
  index <- match(x, vocab)
  index <- index[!is.na(index)]
  rbind(as.integer(index-1), as.integer(rep(1, length(index))))}
documents <- lapply(doc.list, get.terms)
```

그림 9. LDA 모델 생성
Fig. 9. Create an LDA Model

그림 10은 LDA 분석에 활용되는 통계를 계산하고, 그에 해당하는 변수를 지정하는 코드이다.

1) 리뷰 파일의 수(documents)와 리뷰에 등장한 단어의 개수(vocab)를 각각 D, W에 저장한다. 각 리뷰 파일에 등장한 단어의 개수를 doc.length에 저장한다. 문서에 있는 토큰의 총개수를 N에 저장한다. 말뭉치(corpus)가 출현한 빈도수를 term.frequency에 저장한다. 2) LDA를 적용하기 위해 필요한 토픽의 수 K, 반복 횟수 G 와 alpha, eta 값을 지정한다.

```
1   D <- length(documents)
    W <- length(vocab)
    doc.length <- sapply(documents, function(x) sum(x[2, ]))
    N <- sum(doc.length)
    term.frequency <- as.integer(term.table)

2   K <- 3
    G <- 5000
    alpha <- 0.02
    eta <- 0.02
```

그림 10. 통계 계산 및 변수 지정
Fig. 10. Statistical calculations and variable assignment

그림 11은 R에 있는 LDA 라이브러리를 이용하여 깁스 샘플링을 사용함으로써 토픽을 할당할 수 있는 LDA 분석 코드이다.

1) lda.collapsed.gibbs.sampler 함수에 문서(documents), 토픽의 수(K), 단어(vocab), 반복 횟수(num.interations), 문서별 주제가 생성될 확률(alpha), 각 주제별 특정 단어가 생성될 확률(eta), 단어에 대한 초기 주제 지정 목록(initial), gibbs의 수를 나타내는 스칼라 정수(burnin), TRUE 변수에 대한 각 교체 이후에 샘플러가 단어의 로그우도를 계산하게 하는 스칼라 논리(compute.log.likelihood)를 파라미터로 넣어서 fit에 저장한다. 2) alpha에 문서별 주제가 생성될 확률을 계산해 theta에 행렬로 저장한다. 3) eta에 각 주제별 특정 단어가 생성될 확률을 계산해 phi에 행렬로 저장한다.

```
1   set.seed(357)
    fit <- lda.collapsed.gibbs.sampler(documents = documents, K = K, vocab = vocab,num.iterations = G, alpha = alpha,eta = eta, initial = NULL, burnin = 0,compute.log.likelihood = TRUE)

2   theta <- t(apply(fit$document_sums + alpha, 2, function(x) x/sum(x)))

3   phi <- t(apply(t(fit$topics) + eta, 2, function(x) x/sum(x)))
```

그림 11. LDA 분석
Fig. 11. LDA analysis

그림 12는 LDA 결과를 웹에서 시각화하기 위한 리스트를 만드는 코드이다.

각 변수들에 이전에 지정한 phi, theta, doc.length, vocab, term.frequency 값을 파라미터로 지정하고, list를 만들어 result에 저장한다.

```
result <- list(phi = phi,
               theta = theta,
               doc.length = doc.length,
               vocab = vocab,
               term.frequency =
term.frequency,encoding='UTF-8')
```

그림 12. 웹 시각화를 위한 데이터 저장
Fig. 12. Save data for web visualization

3. 감정 분석

감정 분석 과정은 웹 크롤링 과정을 통해 저장된 텍스트 파일을 감정사전(긍정어, 부정어 사전) 과 대조하여 문서의 감정을 분석한다.

그림 13은 감정 분석에 사용할 긍정어, 부정어 사전을 불러오는 코드이다.

```
positive <- readLines("positive.txt")
positive=positive[-1]
negative <- readLines("negative.txt")
negative=negative[-1]
```

그림 13. 감정 분석 사전 파일 불러오기
Fig. 13. Load sentiment analysis dictionary file

그림 14는 감정 분석의 전처리 과정과 긍정 부정 판단 과정을 포함하는 함수의 코드이다.

```
sentimental = function(sentences, positive,
negative){
scores = laply(sentences, function(sentence,
positive, negative)
{
sentence = gsub('[[:punct:]]', '', sentence)
sentence = gsub('[[:cntrl:]]', '', sentence)
sentence = gsub('\\d+', '', sentence)
word.list = str_split(sentence, '\\s+')
words = unlist(word.list)
pos.matches = match(words, positive)
neg.matches = match(words, negative)
pos.matches = !is.na(pos.matches)
neg.matches = !is.na(neg.matches)
score = sum(pos.matches) -
sum(neg.matches)
return(score)
}, positive, negative)
scores.df = data.frame(score=scores,
text=sentences)
return(scores.df)
}
```

그림 14. 감정 분석용 함수 생성
Fig. 14. Create function for sentiment analysis

gsub 함수를 사용하여 리뷰 텍스트 파일의 문장부호, 특수문자, 숫자를 제거한다. str_split 함수로 공백 기준으로 단어를 생성하고 word.list에 저장한다. unlist 함수를 사용하여 word.list를 vector 객체로 구조를 변경한다. match 함수를 사용하여 긍정어, 부정어 사전과 words의 단어를 대조하여 리뷰의 긍정어와 부정어의 개수를 계산한다. score의 값(긍정어의 수 - 부정어의 수)의 값이 양수이면 긍정적 리뷰, 음수이면 부정적 리뷰, 0이면 중립적 리뷰로 판단한다.

그림 15는 생성한 토픽의 단어들을 단어 감성사전과 매칭하는 과정이다.

1) 토픽별로 상위 20개의 단어를 뽑고, 단어를 리스트로 합친다. 빈도수 테이블에서 단어를 검사하고 단어와 빈도수로 새로운 테이블을 생성한다. 2) 단어 감성사전인 "sorted.txt"파일을 불러온다. 3) 단어를 매칭하고, NA를 제거 후, 위치(숫자)만 추출한다. 단어와 빈도수로 새로운 테이블을 생성하고 sent.frequency에 sent.table을 숫자형으로 바꾸고 저장한다.

```
1   topicwords <-
top.topic.words(fit$topics, 20, by.score =
TRUE)
    topicwords<-c(topicwords[,1],topicword
s[,3],topicwords[,3])
    search<- names(term.table) %in%
topicwords
2   newterm.table <- term.table[search]

    new_sentiment <-
readLines("sorted.txt")
    new_sentiment=new_sentiment[-1]
3   newterm.data<-data.frame(newterm.tab
le)
    newterm.vec<-rep(newterm.data$doc.lis
t)

    sent.matches =
    match(newterm.vec, new_sentiment)
    sent.matches = !is.na(sent.matches)
    sent.table<-newterm.table[sent.matches
]
    sent.frequency <- as.integer(sent.table)
```

그림 15. 빈도수 확인
Fig. 15. Check frequency

4. 웹 시각화

웹 시각화 과정은 분석된 데이터를 RShiny 라이브러리를 이용하여 시각화된 데이터를 서버에 저장하고, 웹 애플리케이션에서 데이터를 불러오며 출력한다.

그림 16은 문서 토픽 추출의 결과를 시각화하기 위해 json 파일을 생성하는 코드이다.

그래프의 한글 깨짐 방지를 위해 json의 인코딩 타입을 utf-8로 설정한다. createJSON 함수에 phi, theta, doc.length, vocab, term.frequency 값을 파라미터로

지정하고 json에 저장한다. serVis 함수를 이용하여 현재 R script 파일이 있는 경로에 vis 폴더를 생성하고 json 파일을 저장한다.

```
json<-options(encoding='utf-8')
json <- createJSON(phi = result$phi,
                   theta = result$theta,
                   doc.length = result$doc.length,
                   vocab = result$vocab,
                   term.frequency = result$term.frequency,encoding='UTF-8')
serVis(json, out.dir = 'vis', open.browser = FALSE)
```

그림 16. 웹 시각화용 json 파일 생성
Fig. 16. Create json file for web visualization

그림 17은 문서 토픽 추출의 결과를 웹에서 시각화하는 코드이다.

1) vis 폴더에 생성된 lda.json 파일을 rjson 변수에 불러온다. 2) 웹 시각화를 위해 서버에서 myChart에 renderVis를 이용하여 시각화 데이터를 저장한다. 3) 서버에서 저장된 myChart 데이터를 visOutput으로 데이터를 출력한다.

```
1  rjson<-readLines(file.path(path,"vis","lda.json"),encoding='utf-8')

2  server <- function(input, output,session) {
   ...
   output$myChart <- renderVis({
       with(result, rjson)
     })
   ...
   }
3  ui <- dashboardPage(...
   tabItem(tabName="topic",h1("토픽분석"),hr(),visOutput('mychart'))
   ...)
```

그림 17. '아몬드' 문서 토픽 추출 웹 시각화
Fig. 17. '아몬드' topic web visualization

그림 18은 수집한 '아몬드' 책의 리뷰에 LDA 알고리즘을 적용해 토픽을 추출한 결과이다.

토픽 추출 결과에서 책의 등장인물인 윤재, 엄마, 할머니라는 단어와 책의 내용과 밀접한 관련이 있는 감정, 공감, 성장 등의 단어가 출현한 것을 확인할 수 있다.

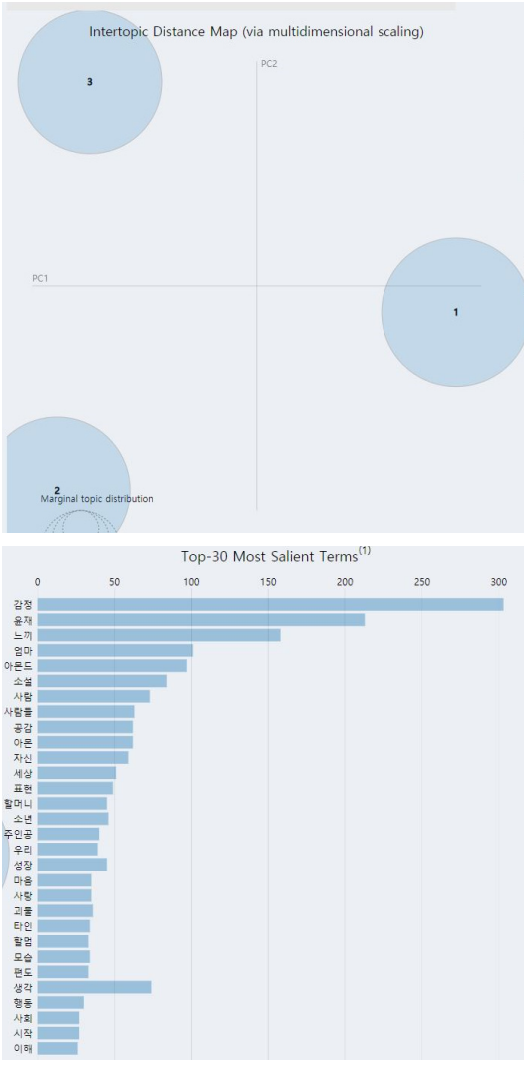


그림 18. '아몬드' 문서 토픽 추출 웹 시각화 결과
Fig. 18. '아몬드' topic extraction web visualization results

그림 19는 감정 분석의 결과를 웹에서 시각화하는 코드이다.

1) 웹 시각화를 위해 서버에서 renderPlot을 이용하여 감정 분석 결과를 파이 차트로 표현한 시각화 데이터를 sentiment_result에 저장한다. 2) renderPlot을 이용하여 감정 분석 결과를 트리맵 차트로 표현한 시각화 데이터를 treemap에 저장한다. 3) 서버에서 저장된 sentiment_result, treemap 데이터를 plotOutput으로 데이터를 출력한다.

```
1   server <- function(input, output,session) {
    ...
    output$sentiment_result <-
renderPlot({pie(sentiment_result,
main="감정분석
결과",col=c("skyblue2","lightcoral","palegreen2
"),
label=paste(names(sentiment_percent),'',
sentiment_percent,"%"),                    border
= FALSE, radius=1)})
2
    output$treemap <- renderPlot({
    dset<-data.frame(keywords=names(sent.ta
ble), sentiment=sent.frequency)

    treemap(dset, index=c("keywords"),
vSize=c("sentiment"), vColor=c("sentiment"),
type="value",title="감정", title.legend="빈도수",
fontsize.title=15, fontsize.legend=11
,fontsize.labels = 11, fontface.labels =
c("bold"), fontfamily.labels =
"NanumBarunGothic", fontfamily.title =
"NanumBarunGothic", fontfamily.legend =
"NanumBarunGothic", palette = "GnBu",
border.col = "white")
    })
    ...
3   }

    ui <- dashboardPage(...
    tabItem(tabName="sentiment",h1("감정분석
"),hr(),
    splitLayout(
    ...
    plotOutput(outputId =
"sentiment_result"),plotOutput(outputId =
"treemap"))
    ))
```

그림 19. '아몬드' 감정 분석 웹 시각화
Fig. 19. '아몬드' sentiment analysis web visualization

그림 20은 감정 분석의 결과를 웹에서 시각화한 결과이다.

각 리뷰를 긍정어, 부정어 사전과 대조하여 분석한 결과인 파이차트에서 전체 리뷰에서 긍정어가 많이 사용된 리뷰의 비율은 50%, 부정어가 많이 사용된 리뷰의 비율은 34.78%, 긍정어와 부정어 비율이 비슷한 중립적인 리뷰의 비율이 15.22%가 나왔다. 또한, LDA 토픽 추출 결과를 감정사전과 대조한 결과인 트리맵에서 감정, 생각, 이해, 공포, 사랑 등의 단어가 나온 것을 확인할 수 있다.

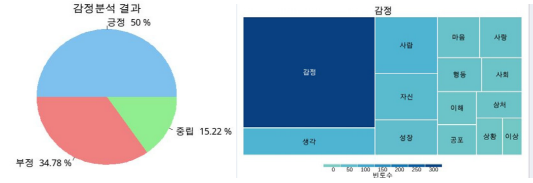


그림 20. '아몬드' 감정 분석 웹 시각화 결과
Fig. 20. 아몬드' sentiment analysis web visualization results

Ⅳ. 결 론

본 논문에서는 LDA 알고리즘과 감정분석을 이용하여 교보문고에 있는 책 리뷰의 키워드들을 분석해보았다. 분석한 토픽과 감정단어를 바탕으로 책의 내용 유추하여 책을 읽으려는 독자들에게 도움이 될 수 있을 것이라 예상된다. 또한, 본 연구에서 쓰인 토픽 추출 시스템을 이용하여 리뷰 데이터 수집에 대한 편리성이 증가할 것이며, 기업은 책 리뷰 분석을 통해 책에 대한 독자들의 반응을 검토 가능하여 책의 전망을 파악할 수 있을 것이라 예상된다.

References

[1] YoungHo Lee, "A Study on Analytical Machine Learning Method Applying Discretization and Hierarchical Clustering Algorithm", The Journal of KIIT, Vol. 19, No. 1, pp. 55-61, 2021.
DOI : 10.14801/jkiit.2021.19.1.55

[2] So-Jin Lee, Chae-Eun Jin, Min-Ji Jeon, Jo-Eun Lee, Su-Jeong and Kim, Sang-Hyun Lee, "De-identification Policy Comparison and Activation Plan for Big Data Industry", The Journal of the Convergence Culture Technology, Vol. 2, No. 4, pp.71-76, 2016.
DOI:http://dx.doi.org/10.17703/JCCT.2016.2.4.71

[3] Chan-Ho Lee, Min-Seung Kim, Jeong-Hee Lee and Tae-Eung Sung, "A Study on the Comparison of Technology Development Trends using Topic Modelling-based LDA Method and Web Search Traffic Analysis", The Korean Institute of Information Scientists and Engineers, pp.1767-1769, 2019.

[4] Sejong Oh, Sunghun and Ahn, Jungmin Byun, "A Big Data Study on Viewers' Response and Success Factors in the D2C Era Focused on tvN's Web-real Variety 'SinSeoYuGi' and Naver TV Cast Programming", International Journal of Advanced Culture Technology, Vol.4, No.2, pp.7-18, 2016.
DOI:http://dx.doi.org/10.17703/IJACT.2016.4.2.7

[5] Soon-Uk Yoon and Min-Chul Kim, "Topic Modeling on Fine Dust Issues Using LDA Analysis", Journal of Energy Engineering, Vol.29, No.2, pp.23-29, 2020.
DOI:https://doi.org/10.5855/ENERGY.2020.29.2.023

[6] Seung-hui Baek, Soo-yeon Son, Joo-young Lee, Ji-seon Lee, "A Study on the Influence of Purchasing of books on Internet Bookstore Review", Conf. of Korean Society for Information Society 2015 22th, pp.109-114., 2015.

[7] Kwon, Hyejin, "Content analysis of readers' book reviews on the picture book 『Crow Boy』 from internet bookstores in Korea, USA, and Japan ", The

Korean Society for Early Childhood Education and Care, Vol. 13 No. 2,, pp.29-50, 2018.
DOI:https://doi.org/10.16978/ecec.2018.13.02.002

[8] Jinsu Kim, "Emotion Prediction of Document using Paragraph Analysis", Journal of Digital Convergence, Vol. 12, No.12, pp.249-255, 2014.
DOI:https://doi.org/10.14400/JDC.2014.12.12.249

[9] Pan-Seop Shin, "Emotional analysis system for social media using sentiment dictionary with newly-created words", Journal of The Korea Society of Computer and Information Vol. 25 No. 4, April 2020, pp. 133-140.
DOI: https://doi.org/10.9708/jksci.2020.25.04.133

[10] Raegeun Park, Hyeok-Jin Yun, Ui-Cheol Shin, Young-Jin Ahn, Seungdo Jeong, "Application of Advertisement Filtering Model and Method for its Performance Improvement," Journal of the Korea Academia-Industrial cooperation Society(JKAIS), Vol. 21, No. 11, pp. 1-8, 2020.
DOI: http://dx.doi.org/10.5762/KAIS.2020.21.11.1

[11] Jeong-Bin Hwang, Jin-Woo Kim,, Seok-Ho Chi and Joon-Oh Seo, "Automated Training Database Development through Image Web Crawling for Construction Site Monitoring", Journal of the Korean Society of Civil Engineers Vol. 39, No. 6, pp.887-892, 2019.
DOI:https://doi.org/10.12652/Ksce.2019.39.6.0887

[12] Kyung-Su Kang, Se-Min Park, "Keyword Analysis of KCI Journals on Business Administration using Web Crawling and Machine Learning", Korean Journal of Business Administration, Vol. 32, No. 4, pp.597-615, 2019.
DOI: https://doi.org/10.18032/kaaba.2019.32.4.597

[13] Jae-Saeng Kim, "Big Data Analysis Technologies and Practical Examples", The Korea Contents Association Review, Vol. 12, No. 1, pp.14-20, 2014.

[14] Blei, D., A. Ng, and M. Jordan, "Latent Dirichlet Allocation.", Journal of Machine Learning Research, Vol. 3,, pp. 993-102, 2003.

[15] Yoon-Su Kang, Min-Su Kim, Chang-Hwan Hong, Seong-Baeg Kim and Sang-Cheol Kwon, "Visualizing Educational Material Using a Big Data Analytical Tool R Language", Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, Vol.8, No.3, pp. 915-924, 2018.
DOI: http://dx.doi.org/10.21742/AJMAHS.2018.03.45

저 자 소 개

안 윤 빈(준회원)

•Yoon-Bin An is a student of Computer Engineering at Korea Polytechnic University in 2020. Her research interests include big data, the Internet of Things (IoT), and artificial intelligence (AI).

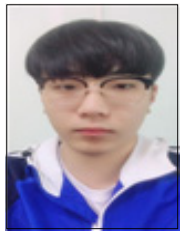
김 학 영(준회원)

•Hak-Young Kim is a student of Computer Engineering at Korea Polytechnic University in 2020.

문 용 현(준회원)

•Yong-Hyun Moon is a student of Computer Engineering at Korea Polytechnic University in 2020. His research interests include big data.

황 승 연(준회원)



•Seung-Yeon Hwang is received his BS in Department of Computer Science at Korea Polytechnic University in 2019. He is currently studying MS in Department of Computer Science at AnYang University. His research interests include Database System, Big Data, Data Analysis, Machine Learning, etc

김 정 준(정회원)

•JeongJoon Kim received his BS and MS in Computer Science at Konkuk University in 2003 and 2005, respectively. In 2010, he received his PhD in at Konkuk University. He is currently a professor at the department of ICT Convergence Engineering at Anyang University, His research interests include Database Systems, BigData, Semantic Web, Geographic Information Systems (GIS) and Ubiquitous Sensor Network (USN), etc.