빅데이터 분석을 통한 트렌드 파악 및 사용자 맞춤 도서 추천[1)]

윤경서[1], 강승식[2]
[1]국민대학교 AI빅데이터융합경영학과 학부생
[2]국민대학교 인공지능학부 교수
rudtj0107@naver.com, sskang@kookmin.ac.kr

A Trend Analysis and Book Recommendation through Bigdata Analysis

Kyungseo Yoon[1], Seungshik Kang[2]
[1]Dept. of AI Bigdata Convergence Management, Kookmin University
[2]Dept. of Artificial Intelligence, Kookmin University

요 약

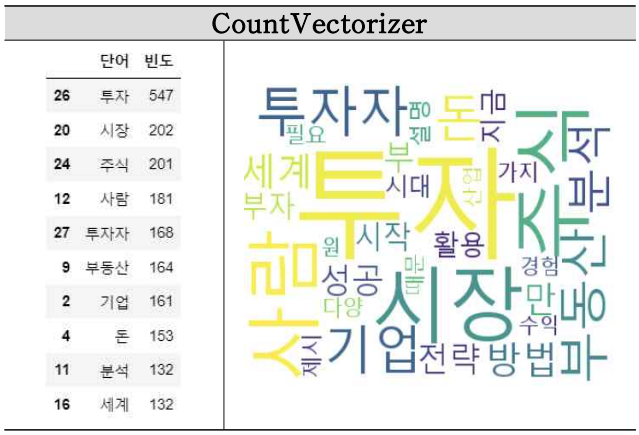
카테고리별 베스트셀러를 통해 트렌드 파악 및 사용자 맞춤형 도서 추천을 위해 카테고리별로 도서 데이터를 수집하고, 대용량 데이터인 위키피디어 데이터를 이용하여 워드임베딩 모델을 구축한다. 도서 데이터에 대한 키워드 분석 및 LDA 주제분석 기법에 의해 카테고리별 핵심 단어 분석을 통해 도서 트렌드를 파악하고, 사용자 맞춤형 도서 정보 제공 및 도서를 추천하는 기능을 구현한다.

1. 서론

기계학습 기법에 의한 추천 시스템, 텍스트 마이닝 연구의 기술 발전과 함께 도서 추천 시스템은 독자들에게 더 많은 도서 정보를 제공하고, 개인 맞춤형 독서 경험을 제공함으로써 도서출판 산업에 기여하고 있다. 기존의 도서 추천 시스템은 주로 독자가 검색한 정보를 바탕으로 관련 도서를 추천하는 방식이었다. 이 방법은 독자의 이전 구매 이력이 없거나, 리뷰 혹은 평점 등의 사용자 활동이 없는 경우에는 적용할 수 없다는 단점이 존재한다. 이와 같이 기존 연구들은 독자의 개별 취향과 니즈를 고려함에 있어 한계가 있었다. 따라서 독자들의 도서 선택과 만족도 향상을 위해 본 연구를 수행하였다. 독자들의 도서 시장의 트렌드로 본인의 니즈와 적합한 단어 혹은 카테고리를 파악하고, 독서 후 사람들이 느끼는 만족감을 극대화할 수 있는 방향으로 추천 방향을 설정하였다. 교보문고의 실시간 베스트셀러 리뷰를 분석한 결과, 도서 내용에 대한 리뷰가 대부분인 것을 알 수 있었다. 본 연구는 현재 트렌드가 되는 키워드들을 파악함으로써 도서 시장의 흐름을 파악하고, 이를 통해 독자들에게 최신 카테고리별 정보를 제공하는 도서 트렌드 분석을 시도하였다.[1,2]

2. 도서 데이터 수집 및 시각화

교보문고 웹사이트에서 각 카테고리별로 도서 데이터를 크롤링하여 수집하였다. 수집한 데이터는 도서 제목, 저자, 카테고리, 도서 내용 정보를 포함한다. 교보문고 웹사이트의 '국내도서' 항목에서 '소설', '시/에세이', '인문', '가정/육아' 등 총 20개의 카테고리를 선정하였다. 각 카테고리별 240권씩 총 4800권에 대한 데이터를 파이썬 selenium 라이브러리를 사용하여 데이터셋을 구축하였다. 각 도서 카테고리별 트렌드를 파악하기 위해 키워드를 추출하여 워드클라우드 시각화를 수행하였다.



(그림 1) 경제/경영 분야의 워드클라우드.

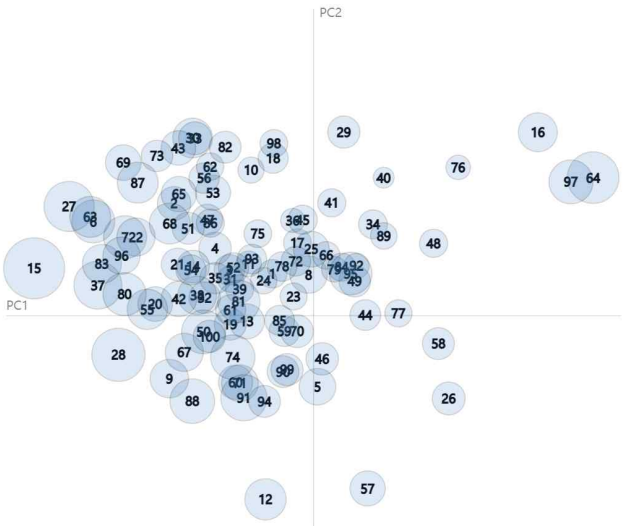
도서 내용 벡터를 만들기 위해 도서 내용에 해당하는 부분을 KLT2023 형태소 분석기[2)]를 사용하여 명사를 추출하고, 불용어 처리 및 중복 제거를 한다. 위키피디아 데이터로 만든 FastText 모델을 사용하여 각 도서의 내용에서 추출한 명사들의 단어 벡터를 구한다. (그림 1)은 경영/경제 카테고리의 단어 빈도와 워드클라우드 예시이며 투자, 주식, 기업 등 주식 관련 내용이 주요 트렌드인 것을 알 수 있다.

1) 본 연구는 2023년도 산업통상자원부 ATC+사업의 지원을 받았음

2) https://cafe.naver.com/nlpkang/41

3. LDA 분석

도서 데이터의 카테고리 특징들을 잘 내포하고 있는지 확인하기 위해 LDA 분석을 수행하였다. LDA 분석은 도서 데이터에서 두 글자 이상인 단어들 중에서 최소 10개의 도서에서 등장하고, 전체 도서 수 중 70% 이상의 도서에서는 등장하지 않는 단어는 제외하였다. 단어별 빈도수를 구하여 말뭉치를 생성한다. 말뭉치 중 90%는 학습용, 10%는 평가용으로 분할하여 LDA 모델 학습을 진행했다. LDA 모델의 topic 수는 100개로 설정했으며, loss는 log_perplexity로 혼란도를 측정하여 사용하였다. 가장 loss가 낮았던 -18.43의 혼란도를 갖는 LDA 모델을 사용하였다.[3]

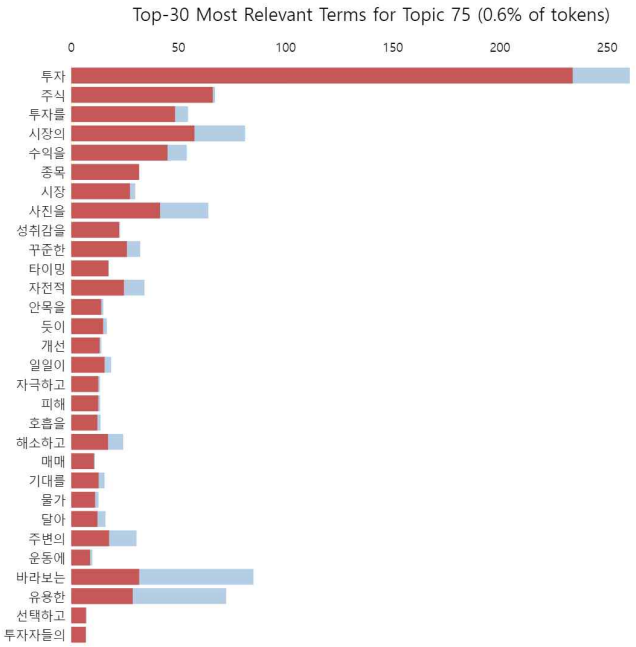


(그림 2) pyLDAvis 결과: distance map

모델 구축 후, pyLDAvis를 통해 LDA 모델을 시각화하는 단계를 거친다. 이때, lambda 값은 키워드 score의 weight와 관련된 것으로 0.4가 적당하다고 판단되어 0.4로 설정하였다. (그림 2)는 pyLDAvis를 실행한 결과로 distance map를 보여주며, (그림 3)은 주요어 30개를 추출한 결과이다.

4. 워드임베딩 기법을 이용한 도서 추천

도서 추천 서비스는 사용자가 입력한 문장 또는 단어들의 조합에 의해 적합한 도서를 추천해 준다. 문장 입력의 경우, 명사를 추출하여 문장 벡터를 구하고, 해당 문장 벡터와 도서 내용 벡터들 간의 코사인 유사도를 구하여 코사인 유사도가 높은 상위 n개의 책을 추천한다. 명사를 벡터화하는 과정에서 위키피디아 데이터로 만든 워드임베딩 모델의 get_vector 함수를 활용하였다. '마음을 편안하게 해주는 위로의 글'이라는 문장을 주었을 때, Word2Vec 모델은 '행복을 담아줄게', '있는 그대로', '보이지 않는 곳에서 애쓰고 있는 너에게' 등의 도서를 추천해 주었고, FastText 모델의 경우, '당신이 좋아지면, 밤이 깊어지면', '나로서 충분히 괜찮은 사람', '당신과 아침에 싸우면 밤에는 입맞출 겁니다' 등의 도서를 추천해 주었다.



(그림 3) pyLDAvis 경제/경영 분야에 대한 결과값.

단어들의 조합에 의한 도서 추천은 사용자가 원하는 도서 내용을 positive word와 negative word로 표현한다. 입력된 단어와 가장 유사한 단어들을 추출하고, 해당 단어들의 단어 벡터 합을 구한다. 단어 벡터들의 합과 도서 내용 벡터 간의 코사인 유사도를 구하여 코사인 유사도가 가장 높은 1권의 도서를 추천한다. 입력으로 positive word는 '모험', '전쟁', '성공'이며, negative word는 '공주', '연애'를 주었을 때, Word2Vec 모델은 '이미 시작된 전쟁'을, FastText 모델은 '로마인 이야기 2: 한니발 전쟁'을 추천해 주었다. 해당 서비스도 각각의 워드임베딩 모델 모두 입력된 단어 조합과 연관이 있는 도서들을 추천해 준 것을 확인할 수 있다.

5. 결론

도서 데이터를 수집하여 카테고리별 트렌드를 파악하고 사용자에게 맞춤형 도서를 추천하기 위하여 도서 데이터를 분석하여 카테고리별 특징을 파악하고, 사용자 취향과 일치하는 도서 추천 연구를 수행하였다. 워드임베딩 기법을 이용하여 도서 추천의 효율성을 높이고자 하였다. 도서 데이터 분석을 통해 트렌드 및 키워드를 파악하고, 독자들의 요구사항에 적합한 도서 추천 방법을 연구하였다.

참고문헌

[1] 정덕영, 이준석, 박상성, "워드클라우드를 이용한 기술 트렌드 분석", 한국지능시스템학회 춘계 학술대회 학술발표논문집, pp.17-18, 2016.

[2] 이은영, 주경희, 이두희, "워드 클라우드 기법을 이용한 최근 소비자학 연구 트렌드 분석", 상품학연구, 37권 1호, pp.1-7, 2019.

[3] 이대영, 이현숙, "LDA 토픽 모델링의 적정 토픽 수 결정 방법 탐색: 혼잡도와 조화평균법 활용을 중심으로", 교육평가연구, pp.1-30, 2021.