베스트셀러로 본 사회 트렌드 분석

AI빅데이터융합경영학과 20182797 김진호

1. 연구 배경

인간의 역사는 문자와 함께 기술되기 시작되었다고 해도 과언이 아니다. 문자가 보편적으로 퍼진 이래 인간과 책은 떼려야 뗄 수 없는 관계를 유지하게 되었다. 책을 통해 인간은 지식을 얻고 새로운 세계를 여행한다. 현대에 이르러 책은 시·공간적으로 정보를 매개하는 매개체로써 중요한 역할을 담당하기도 한다. 하지만 이제는 책이 아니더라도 '정보'는 넘쳐난다. 대중들은 너무나도 쉽게 정보를 구할 수 있다. 검색 포털 사이트에서 클릭만 하면 대중들은 자신들이 원하는 정보를 얼마든지 얻어갈 수 있다. 그렇게 본다면 '베스트셀러'라는 책에 대한 구매는 단순히 정보를 얻기 위한 대중들의 수단으로만 보아서는 안 될 것이다. 베스트셀러란 대중들의 심층적인 코드를 반영한 것이기 때문이다. 한 권의 책이 베스트셀러가 되기 위해서는 수많은 사람이 비용을 지불하고 책을 구매해야만 한다. 또한, 그 분야 혹은 그 책에 대한 대중들의 지속적인 관심이 필요하다. 대중들은 정보의 홍수라 해도 과언이 아닌 다양한 책 중 그들이 원하는 것을 취사 선택할 자유가 있다. 한 시대, 한 사회를 풍미하는 책들, 일명 베스트셀러는 그런 독자의 선택이 집적된 결과이다. 또한, 베스트셀러는 독자의 선택 결과와 더불어 출판사들의 기획이 어우러진 결과물인데, 자본주의 사회에서 출판사가 베스트셀러를 만들기 원하는 즉 많이 팔린 책을 만들기 원하는 것은 당연한 일이다. 그리고 출판사가 책을 많이 팔기 위해서 좋은 글을 출판하는 것도 있겠지만 시대를 읽고, 독자의 수요를 파악하는 것 또한 당연한 일이 될 것이다. 즉 베스트셀러는 근본적으로는 다수의 독자 선택에 의해 만들어지는 것이지만 그 속에는 이미 출판사들이 독자들의 마음을 읽고 기획된 것이기도 하다. 그렇게 본다면 베스트셀러는 단순히 많이 팔린 책이 아니라 대중들의 마음을 이중 삼중으로 파악한 결과물이며, 한 사회의 정신을 반영한 목록인 것이다.

인터넷의 발전으로 서점에서 오프라인으로 책을 구매하기보단, 온라인 서점에서 책을 구매하는 경우가 많아지고 있다. 따라서 인터넷 서점 중 '교보문고'에서 최근 1년간 베스트셀러로 선정된 책의 제목, 저자, 줄거리, 카테고리, 선정 기간을 크롤링하고 워드클라우드, TF-IDF 기법을 사용하여 베스트셀러가 담고 있는 사회의 정신과 트렌드를 파악해보고자 한다.

2. 관련 연구

권보드래(2014)는 1910년부터 집계된 '국민이 가장 많이 읽은 책'을 나열하며 책은 시대의 정신을 담고 있는 거울이라고 주장하고 있다. 이를테면 1900년대부터 1920년대는 애국과 계몽, 1930년부터 1940년은 일본의 식민지로 인한 일본 문학책의 유입, 1950년 이후 광복 후 자주적인 국가 발전을 위한 지식과 문화, 1960년대 이후 자유와 민주주의를 지키기 위한 청년들의 투쟁 등의 키워드로 시대의 정신을 정립하였다. 예시로 1980년대는 전 국민이 '민주화'를 쟁취하기 위해 싸우고 투쟁했던 시대이다. 대학진학률이 이미 30%를 넘었기 때문에 이

시대의 책 소비는 청년들이 주를 이뤘다. 청년들의 대규모 시위와 정치•사회적 관심도의 증가로 『해방 전후사의 인식』 같은 사회•정치 분야의 서적에 대한 수요가 증가하였고, 『자본주의 경제의 구조와 발달』 등과 같은 마르크시스트들의 개론서가 번역되어 많이 읽혔다.

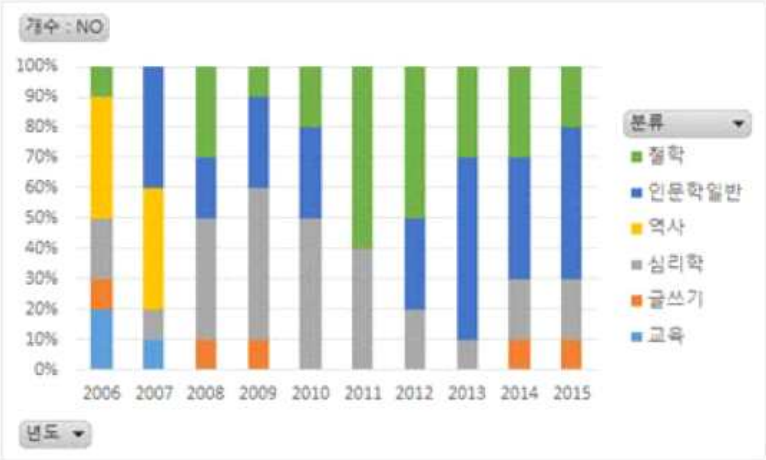

Fig 1. 인문학 베스트셀러 연간 분야별 비중

<Fig 1.>은 2006년부터 2015년까지 연도별로 나타나는 인문학 베스트셀러의 분야별 경향과 특징을 순서대로 정리한 것이다. 2006년도에는 faction과 같은 역사물에 대한 대중들의 관심이 많았다. 사회구조가 지금보다 취업이 힘든 것도, 경쟁이 치열한 것도 아니었기 때문에 본인의 삶과 이질적인 분야에 관심을 뒀음을 알 수 있다. 하지만 그림에서 알 수 있듯 점차 시간이 지나면서 '역사'보다는 '철학', '인문학 일반'과 같은 분야가 대세를 이루었다. 2008년부터 약 10년간 인문학 도서의 주름을 잡은 것은 '힐링' 문화였다고 말할 수 있다. 예전보다 삶이 바빠지고 치열한 경쟁 사회 속에서 실질적인 '나'를 가꾸기보다 대외적인 '나'를 가꾸는데 치중한 사회구조 상, 책을 통해 자기계발과 힐링을 하는 사람들이 늘어났다고 할 수 있다.

베스트셀러란 한 사회의 체제와 구조, 유행 속에서 염원하는 사회적 산물로써 그 시대와 사회를 드러내는 하나의 지표이다. 그런 만큼 하나의 책이 베스트셀러가 되는 특정 한 가지 원인을 꼽을 순 없다. 따라서 베스트셀러의 텍스트 정보를 통해 시대를 특정할 수 있는 단어를 추출하여 해당 단어가 얼마나 가중치가 있는지를 파악해 시대의 흐름과 책의 흐름을 비교해보고자 한다. 이때 사용되는 텍스트 분석 기법이 TF-IDF이다.

TF-IDF를 이해하기 위해선 DTM의 개념을 정리해야 한다. DTM(문서 단어 행렬, Document-Term Matrix)이란 다수의 문서에서 등장하는 각 단어의 빈도를 행렬로 표현한 것을 말한다. 즉, 각 문서에 대한 BoW를 하나의 행렬로 만들 것으로 생각하면 되고, BoW와 다른 표현 방법이 아니라 BoW 표현을 다수의 문서에 대해서 행렬로 표현하고 부르는 용어이다. 다음과 같은 4개의 문장 있다고 가정하자.

Ⅰ. 나는 오늘도 밥을 먹는다.
Ⅱ. 나는 오늘도 출근한다.
Ⅲ. 오늘도 날씨가 좋다.

각 문장을 토큰화해서 DTM으로 만들면 다음과 같은 표가 생성된다.

	나는	오늘도	밥을	출근을	날씨가	먹는다	한다	좋다
Ⅰ	1	1	1	0	0	1	0	0
Ⅱ	1	1	0	1	0	0	1	0
Ⅲ	0	1	0	0	1	0	0	1

표1. 문서Ⅰ, Ⅱ, Ⅲ의 DTM

해당 표와 같이 각 행은 각 문서를 의미하며 각 열은 모든 문서에서 나온 단어의 종류이다. 각 값은 해당 문서에서 해당 단어가 나온 빈도수를 의미한다. 이러한 DTM은 문서들을 서로 비교할 수 있도록 수치화할 수 있다는 장점이 있다. 필요에 따라서는 형태소 분석기로 단어 토큰화를 수행하고, 불용어에 해당하는 조사들도 제거하여 더욱 정제된 DTM을 만들 수 있다.

이러한 DTM도 한계점이 존재한다. 각 문서에서 각기 다른 단어들이 많이 존재한다면 열 공간에 들어가는 단어의 수가 많아지며 DTM 행렬이 매우 sparse 해지기 때문에 많은 양의 저장 공간과 높은 계산 복잡도를 요구할 수 있다. 이러한 이유로 전처리를 통해 단어 집합의 크기를 줄이고 task에 필요한 단어들의 행렬로 모델을 구현하는 것이 효율적이다. 추가로 불용어 제거, 어간 추출, 정규화 등을 통해 단어 집합의 크기를 줄일 수 있다. 여러 문서에서 등장하는 단어에 대해 가중치를 계산할 때도 문제가 발생한다. 예를 들어 불용어인 'the'는 어떤 영문서이든 자주 등장하는 단어이기 때문에 모든 문서에서 동일하게 the의 빈도수가 높다고 해서 문서들이 유사한 문서라고 판단해서는 안 된다.

각 문서에는 해당 문서를 잘 설명하는 단어들이 분명 존재한다. 불용어와 같은 단어들의 빈도수가 상대적으로 높으므로 표현력이 높은 단어들이 자연어 처리에 있어서 의미가 없을 수 있다. 이러한 점들을 고려하여 만든 자연어 처리 분석기 중 하나가 바로 TF-IDF이다.

TF-IDF(Term Frequency-Inverse Document Frequency)는 단어의 빈도와 역 문서 빈도를 사용하여 DTM내의 단어마다 중요한 정도를 가중치로 주는 방법으로 사용방법은 DTM을 만든 후 TF-IDF 가중치를 부여한다. 각각의 텍스트가 토큰화되어 TF-IDF 방법을 적용한다면, 여러 문서로 이루어진 텍스트 단어에 대해서 특정 문서 내에 영향력을 얼마나 갖는지 알 수 있다. 여기서 문서를 d, 단어를 t, 문서의 총 개수를 n으로 표현한다. tf(d.t)는 특정 문서 d에서 특정 단어 t의 등장 횟수를 의미한다. DTM이 각 문서에서의 각 단어의 등장 빈도를 나타내는 값이므로 tf(d,t)와 같은 의미가 있다. df(t)는 특정 단어 t가 등장한 문서의 수를 의미한다. 여기서 특정 단어가 각 문서, 또는 문서들에서 몇 번 등장했는지 관심 두지 않으며 오직 특정 단어 t가 등장한 문서의 수에만 관심을 둔다. idf(d,t)는 df(t)의 반비례하는 수로 log(n/(1+df(t)))로 나타낼 수 있다.

TF-IDF는 위의 수식에 의거 모든 문서에서 자주 등장하는 단어는 중요도가 낮다고 판단하고, 특정 문서에서만 자주 등장하는 단어는 중요도가 높다고 판단한다. 즉, 'the', 'a'와 같은 불용어는 모든 문서에서 자주 등장하기 때문에 중요도가 상대적으로 낮을 것이다.

이 연구는 TF-IDF를 통해서 시대별 베스트셀러가 어떤 단어의 중요도가 높은지, 해당 시대에는 어떤 사건이 이슈였으며 사회 구조적으로 어떠한 관념이 지배적이었는지를 파악해 베스트셀러에 담긴 사회의 흐름을 알아보고자 한다.

3. 제안방법론

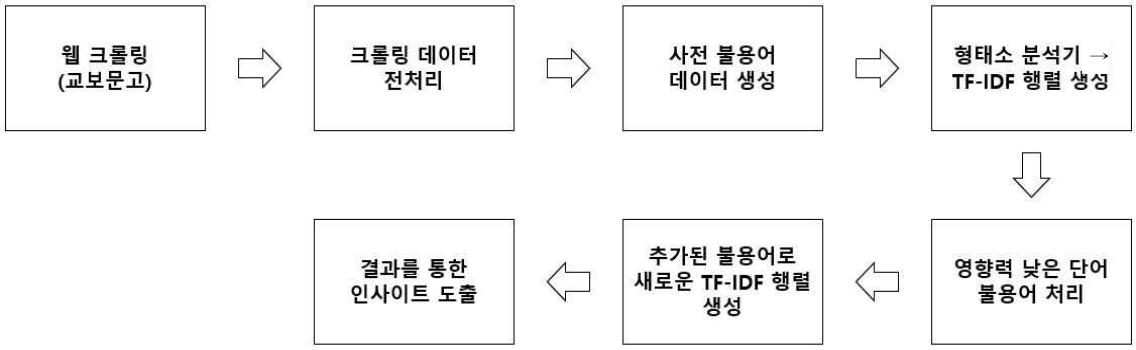

Fig 2. 방법론 모형

본 연구는 교보문고 사이트에서 2021년부터 2012년까지의 베스트셀러 데이터 중 책 제목, 저자, 간략한 줄거리, 카테고리, 베스트셀러 선정 기간으로 총 5가지의 데이터를 크롤링하여 진행하였다. 책 제목만으로는 베스트셀러와 시대의 사회적 흐름 파악에 한계가 존재한다고 생각하여 간략한 줄거리를 중심으로 연구를 진행하였다. 연도별 베스트셀러를 통한 사회 흐름의 분석이 주 연구목적이기 때문에 크롤링해온 데이터를 2012년부터 2021년으로 분리하여 각 데이터에 관해서 연구를 진행하였다.

줄거리의 분석을 위해 사전에 불용어를 설정한 후, CountVectorizer, TfidfVectorizer를 사용하여 각각 등장 단어의 빈도와 중요도를 구하였다. 추가로 CountVectorizer를 통해 구한 값을 WordCloud로 나타내 도식화된 그림을 통해 결과를 쉽게 보고자 하였다. 각각의 Vectorizer 생성 이후 영향력이 낮은 조사, 감탄사 등의 단어를 불용어 사전에 추가하여 새로운 Vectorizer를 만드는 과정을 반복해 결론을 도출할 수 있는 단어들을 나열하고, 이를 통해 인사이트를 도출하는 과정을 진행하였다.

4. 실험내용

본 절에서는 연도별 베스트셀러가 내포하고 있는 사회의 트렌드를 파악해보는 실험을 TF-IDF 비교를 통해 진행한다. 연도별 실험에서의 시사점은 TF-IDF를 통해 나온 상위 20개의 단어가 해당 연도를 얼마나 잘 나타내고 있는지다. 따라서 크롤링, 전처리, TF-IDF 변환 과정을 간단히 살펴본 후, TF-IDF를 통해 나온 결과물을 비교해보며 실험의 의의를 파악하려 한다.

1) '교보문고' 웹 크롤링

인터넷 서점 사이트인 '교보문고'에서 연간 베스트셀러를 크롤링하기 위해 각각의 CSS 경로를 지정해주었다. <표 2>와 같이 로그인, 베스트, 연간 등의 링크를 지정하여 클릭하도록 하고 각 제목을 클릭하여 링크에 들어가 제목, 저자, 간략한 줄거리, 카테고리, 베스트셀러 선정 기간을 지정된 리스트에 추가하는 과정을 반복한다. 페이지를 바꿔가며 총 상위 100개의 책 크롤링이 완료되면 다음 연도로 넘어가 위와 같은 과정을 반복하도록 설정하였다.

Algorithm 1. 교보문고 웹사이트 크롤링

for 베스트셀러 기간 (2012~2021)
  for 페이지 수 (1~5)
  for 책 개수 (20개)
  검색할 책 링크 이동
  * 책 제목 저장
  * 책 저자, 설명 저장
  * 카테고리 저장
  * 베스트셀러 선정 기간 저장
  이전 페이지로 back
  다음 기간으로 이동(100개 크롤링 완료 시)

표2. 크롤링 Algorithm 설명

2) 데이터 전처리

크롤링 된 데이터를 전처리하는 과정을 거쳤다. '2021.03.01.~ 2021.03.31.' 의 형태로 되어있는 '기간' 열을 보기 쉽게 '21.03'의 형태로 변경했다. '카테고리' 열에 대해서도 '소설 > 한국소설'의 형태를 각각 대분류, 중분류로 나눠, 이후에 카테고리별 추이를 보기 위해 새로운 열을 생성하였다. 또한, 각 문서에 '★', '※'과 같은 특수 기호들이 사용된 데이터가 많았기 때문에 정규표현식을 사용하여 해당 값들을 이후 TF-IDF 과정에 끼칠 영향을 고려해 제거하였다. 여기서 정규화를 토큰화 이전에 한 이유는 형태소 분석기가 해당 단어를 어떻게 분류하는지 모르기 때문에 필요 없다고 판단되는 문자들을 사전에 제거하였다. 이후 각각의 연도별로 새로운 데이터 프레임을 지정하여 2012년부터 2021년까지의 베스트셀러 정보를 각각의 데이터 프레임에 할당하였다.

3) 연도별 베스트셀러 분석

본격적으로 연도별 베스트셀러를 분석하면서 단어의 빈도수, 중요도, 카테고리별 빈도수를 파악해보았다.

1] 2012년

2012년의 데이터를 워드클라우드, TF-IDF 행렬 변환을 한 결과는 다음과 같다.

Fig 3. 2012년 베스트셀러 WordCloud / TF-IDF (max_features = 20)

<Fig 3.>을 보면 '토익', ' 해커스', '모의고사'와 같은 자격증 시험과 관련된 단어가 많이 나오고 중요도 또한 높음을 알 수 있다. 이는 2012년 토익•토플을 대체하고 대학 입시자료로 활용되는 '국가 영어 능력 시험'이 2012년부터 본격 시행됐기 때문이다. 등급제로 이뤄지는 본 평가는 취업, 유학, 대학입시 등에 사용되는 참고자료로 사용되었고, 본 평가를 수능 영어 과목을 대체한다는 교육과학기술부의 발표로 큰 화두를 불러일으켰다. 2012년에 처음 치르는 시험이었기 때문에 마땅한 기출문제가 없었으므로, 가장 신빙성 있고 자료가 많은 '토익' 시험 교재가 많이 판매되었음을 알 수 있다.

2] 2013년

2013년의 데이터를 워드클라우드, TF-IDF 행렬 변환을 한 결과는 다음과 같다.

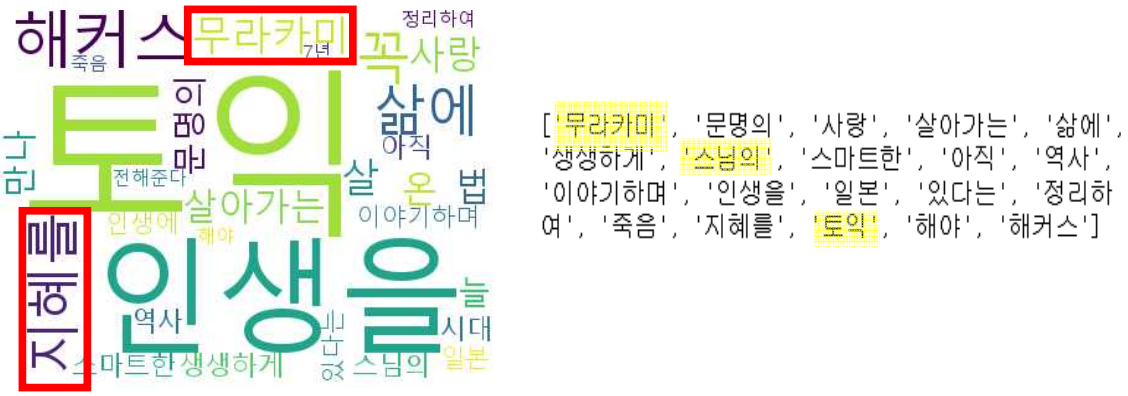

Fig 4. 2013년 베스트셀러 WordCloud / TF-IDF (max_features = 20)

2013년은 일본의 한 세대를 풍미한 '무라카미 하루키' 작가의 도서가 인기 있던 한 해였다. 이와 더불어 '혜민 스님'의 『멈추면, 비로소 보이는 것들』 등과 같이 삶에 관련된, 사랑과 관련된 서적이 인기를 끌었다. 또한, '토익', '해커스' 단어가 중요성이 높은 것으로 보아 영어시험 관련 서적이 꾸준히 높은 판매량을 유지했음을 알 수 있다. 2013년은 책으로 출판될 만한 큰 사건이 발생하지 않은 해였다. 따라서 베스트셀러에도 자기계발 서적, 에세이, 소설 등이 선정됨을 유추할 수 있다.

3] 2014년

2014년의 데이터를 워드클라우드, TF-IDF 행렬 변환을 한 결과는 다음과 같다.

Fig 5. 2014년 베스트셀러 WordCloud / TF-IDF (max_features = 20)

2014년의 결과 중 '김우중'에 집중해보자. '김우중'은 전 대우그룹 회장이다. 현재는 별세하였지만, 우리나라에 끼친 영향은 실로 엄청나다. 故 김우중 전 회장은(앞으로 김 전 회장이라 명칭 하겠다.) 1982년 ㈜대우를 설립하여 창업 30여 년 만에 매출 91조 원, 자산총액 76조 원으로 현대그룹에 이어 재계 2위를 차지하는 국가적 기업의 총수였다. 당시 대한민국 총수출의 14%를 대우그룹이 차지할 정도로 당시 영향력이 엄청난 기업이었는데 문제는 국제통화기금(IMF) 외환위기가 찾아올 때 위기가 찾아왔다. 국가신용등급이 추락하자 해외채권자가 그간 빌려줬던 자금을 상환하라고 압박하였고, 이 과정에서 400%가 넘는 대우그룹의 부채비율이 고스란히 드러났다. 이에 2000년에는 분식회계까지 적발되면서 대우그룹 신화는 신기루가 되었고, 이로 인해 김 전 회장은 2006년 징역을 선고받았다.

중요한 건, 2014년 '대우 특별포럼'에서 김 전 회장은 "방만한 경영 때문에 대우그룹이 해체했다는 건 사실과 다르다"라며 "외환위기로 발생한 자금난을 해소하기 위해서 기업어음, 회사채를 발행하려 했지만, 한국 정부가 이를 가로막는 조치를 취하면서 자금난이 심화했다"라는 주장을 내세웠다. 김 전 회장의 주장과 정부의 주장이 엇갈리고 법원 판결까지 넘어가며 사건이 마무리되었다.

본 사건은 당시 국가적 기업이었던 ㈜대우의 파산에 대한 속사정이 낱낱이 드러나는 중대한 사건이었다. 따라서 과거 김 전 회장이 출판한 에세이가 베스트셀러로 선정되었고 2014년 TF-IDF 결과로 중요도가 높게 측정되었다고 판단된다.

4] 2015~2016년

2015~2016년의 데이터를 워드클라우드로 도식화한 결과는 다음과 같다.

Fig 6. 2015~2016년 베스트셀러 WordCloud (좌 : 2015년 / 우 : 2016년)

2015~2016년은 시대를 내포하는 단어가 적고, 베스트셀러에서도 적게 선정되어 그렇다 할 결과를 도출해내지 못했다. '교보문고' 사이트에 한해서 상위 100위의 베스트셀러만을 크롤링하여 연구를 진행하였기 때문에 모든 한정된 데이터로 연구를 진행하면서 발생하는 당연한 한계점이라 판단한다. 눈여겨볼 점은, 2012년부터 2016년까지 '토익'은 꾸준히 많은 빈도를 보이는 것을 알 수 있다. 대한민국에서 토익을 응시하는 사람이 많다는 걸 유추할 수 있고, 그만큼 높은 토익 점수가 많은 사람이 기본으로 가져가는 역량이라고 판단할 수 있다.

5] 2017년

2017년의 데이터를 워드클라우드, TF-IDF 행렬 변환을 한 결과는 다음과 같다.

Fig 7. 2017년 베스트셀러 WordCloud / TF-IDF (max_features = 20)

2017년. 그야말로 혼돈의 해였다. 대한민국 역사상 최초로 대통령 탄핵이라는 사건이 있었다. 비선 실세 최순실과 박근혜 전 대통령의 구속으로 서민들의 대통령이라 불리던 故 노무현 전 대통령(노 전 대통령이라 명칭 하겠다.)의 과거 대통령 시절 행보가 주목받았다. 노 전 대통령의 전 보건복지부 장관이었던 유시민 작가의 에세이들이 인기 있었던 이유이다.

2017년 하면 떠오르는 또 하나의 키워드가 있다. 바로 '페미니스트' 이다. 박근혜 전 대통령이 탄핵 되고 문재인 대통령이 취임하면서 '페미니스트 대통령이 되겠다.'라고 선언했다. 당시 유리천장, 한남충과 같은 단어들이 논란의 시발점이었고, 청년들의 남녀갈등이 최고조에 달했던 시기였기에 당시 페미니스트 책으로 알려졌던, 사회적 약자로 비치는 여성의 모습을 중심으로 쓴 '82년생 김지영'의 키워드가 중요하게 나왔음을 유추할 수 있다.

6] 2018~2019년

2018~2019년의 데이터를 워드클라우드로 도식화한 결과는 다음과 같다.

Fig 8. 2018~2019년 베스트셀러 WordCloud (좌 : 2018년 / 우 : 2019년)

2018~2019년도 베스트셀러의 TF-IDF, WordCloud로 의미 있는 결과물을 출력하지 못하였다. 다만 여기서 특이점은 2017년까지 '토익'이 빈도수에서 항상 우세에 있었지만, 2018년부터는 조금 열세인 모습을 보이는 것이다. 2017년까지는 취업 시장에서 높은 토익 점수가 큰 이점을 주었지만, 시대가 흐르면서 토익은 기본으로 하고 OPIc과 같은 회화능력이 중요시되고 있으므로 상대적으로 토익이 이전보다 베스트셀러에서 중요도가 떨어짐을 유추할 수 있다.

또한, 2019년 미•중 무역갈등, 금리 인하, 대공황의 위험 등으로 세계 경제가 불확실성 속

에 있던 시기이기 때문에 '경제' 키워드가 베스트셀러에서 중요도가 높게 도출됨을 알 수 있다.

7] 2020년

2020년의 데이터를 워드클라우드, TF-IDF 행렬 변환을 한 결과는 다음과 같다.

['1위', '교수가', '그들의', '김호중', '나의', '변화를', '부의', '부자', '성공', '습관', '여러', '역사', '연구', '인생을', '지식을', '코로나', '토익', '투자', '흐름을', '흔한']

Fig 9. 2020년 베스트셀러 WordCloud / TF-IDF (max_features = 20)

2020년을 관통한 핵심 이슈는 '코로나'이다. 코로나로 인해 세계적으로 엄청난 변화가 일어났고, 가장 큰 변화는 다름 아닌 금융시장의 변화였다. 사회적 거리 두기로 인해 제한된 외부 활동으로 활발한 소비경제가 발생하지 못했다. 소비자들 또한 외출을 꺼렸고, 집에 있는 시간이 늘어났다. 국가별 금융시장의 큰 변동이 발생하면서 세계 주식시장에서도 큰 변화가 일어났다. 이에 집에 있던 소위 주식 개미(개인투자자)들이 주식 투자를 하기 시작했다. 주식 관련 뉴스, 영상이 인기를 끌었고, 투자 가이드, 주식 관련 서적 또한 베스트셀러에 등극했다.

2020년에서 한 가지 재밌는 결과가 나왔는데 '김호중'에 집중해보자. '김호중'은 2020년 대한민국을 뒤흔든 트로트 유행의 시발점이라 불리는 '미스터트롯' TV 프로그램에 출연한 출연진이다. 김호중 씨는 팬들을 위한 『트바로티, 김호중』이라는 에세이을 발간하여 팬들과 소통하고자 했다, 팬들에게 많은 사랑을 받았던 김호중 씨의 『트바로티, 김호중』는 2020년 베스트셀러에 등극했고, 따라서 '김호중'이라는 키워드가 2020년 중요도가 높은 키워드로 도출되었음을 알 수 있다.

8] 2021년

2021년의 데이터를 워드클라우드, TF-IDF 행렬 변환을 한 결과는 다음과 같다.

['1위', '경제', '돈을', '돈의', '메타버스', '모닝', '미국', '미라클', '법을', '부의', '살아가는', '새롭게', '시대에', '아침', '인기', '인생을', '주식을', '토익', '투자', '하다']

Fig 10. 2021년 베스트셀러 WordCloud / TF-IDF (max_features = 20)

<Fig 10>을 보면 투자, 돈, 메타버스, 주식에 관련된 단어가 많이 등장하고 중요도 또한 높음을 알 수 있다. 2021년까지 지속된 코로나로 인해 주식 개미들의 투자 열풍은 여전히 뜨거웠고, 주식, 비트코인으로 돈을 많이 벌어간 사람이 늘어남에 따라 '부자 되는 법', '미라클', 과 같은 돈에 관련된 키워드가 중요하게 도출되었음을 알 수 있다.

2021년에는 기존에 등장하지 않았던 '메타버스' 단어의 중요도가 높게 나왔다. 메타버스란 가공, 추상을 의미하는 메타(Meta)와 현실 세계를 의미하는 유니버스(Universe)의 합성어로 가상세계 이용자가 만들어내는 UGC(User Generated Content)가 상품으로서, 가상통화를 매개로 유통되는 특징이 있다. 2021년 5G 상용화에 따른 정보통신기술 발달과 코로나 19 팬더믹에 따른 비대면 추세 가속화로 주목받았다. 실제로 메타버스 관련 업종 채용이 증가하였으며, 메타버스는 프로그래밍으로 생성되는 공간이기 때문에 높은 전문성을 요구한다. 따라서 메타버스 관련 전문 서적이 2021년에 인기를 끌었음을 알 수 있다.

5. 결론

본 연구는 시대별 베스트셀러가 시대의 트렌드와 흐름을 내포하는지를 TF-IDF의 출력 결괏값을 통해 알아보았다. 불용어 처리를 거치면 어느 정도 해당 시대의 트렌드를 파악할 수 있는 단어들이 출력되었음을 실험을 통해 알 수 있었다. 웹 크롤링을 통한 베스트셀러 파악은 시대를 파악할 수 있다는 가능성을 확인했다는 것에 의의가 있다. 하지만, 인터넷 • 오프라인 서점마다 각기 다른 판매량, 서점 업체마다 다른 베스트셀러 선정 기준, 한정된 데이터, 방대한 불용어 처리 등의 이슈로 해당 결과로 전체적인 사회 흐름을 파악하기엔 한계가 존재한다. 따라서 해당 모델에서 제안하는 단어들을 사용하기 위해서는 추가 데이터 수집, 주제 분석, 감성 분석 등을 통한 모델의 안정화가 필요하다.

제안하는 모델은 형태소 분석기의 경우 카카오의 형태소 분석기인 'khaiii'로 tokenizing을 진행하고 LDA를 통한 주제 분석, sentimental을 통한 감성 분석을 사용하여 향후 연구에는 세부적으로 베스트셀러를 분석해 사회 흐름을 중요토픽과 감정 어휘로 파악하는 연구를 진행할 예정이다.

6. 참고자료

[1] 박치완, 박성준(2016). "인문학 분야 베스트셀러 트렌드 분석 (2006년~2015년)."「글로컬 창의 문화연구」제7집, pp. 132~159.

[2] 박소현, 송애린, 박영호, 임선영(2018). "토픽 모델을 사용한 베스트셀러 서적 단문 의미 분석 연구" 숙명여자대학교 IT 공학과

[3] 권보드래(2014). "베스트셀러, 20세기 한국 사회의 축도(縮圖)". 근대서지 2014 제10호 pp. 150~163