단어 임베딩(Word Embedding) 기법을 적용한 키워드 중심의 사회적 이슈 도출 연구: 장애인 관련 뉴스 기사를 중심으로*

A Study on the Deduction of Social Issues Applying Word Embedding: With an Empasis on News Articles related to the Disables

최가람 (Garam Choi)**
최성필 (Sung-Pil Choi)***

초 록

본 논문에서는 온라인 뉴스 기사에서 자동으로 추출된 키워드 집합을 활용하여 특정 시점에서의 세부 주제별 토픽을 추출하고 정형화하는 새로운 방법론을 제시한다. 이를 위해서, 우선 다량의 텍스트 집합에 존재하는 개별 단어들의 중요도를 측정할 수 있는 복수의 통계적 가중치 모델들에 대한 비교 실험을 통해 TF-IDF 모델을 선정하였고 이를 활용하여 주요 키워드 집합을 추출하였다. 또한 추출된 키워드들 간의 의미적 연관성을 효과적으로 계산하기 위해서 별도로 수집된 약 1,000,000건 규모의 뉴스 기사를 활용하여 단어 임베딩 벡터 집합을 구성하였다. 추출된 개별 키워드들은 임베딩 벡터 형태로 수치화되고 K-평균 알고리즘을 통해 클러스터링 된다. 최종적으로 도출된 각각의 키워드 군집에 대한 정성적인 심층 분석 결과, 대부분의 군집들이 레이블을 쉽게 부여할 수 있을 정도로 충분한 의미적 집중성을 가진 토픽들로 평가되었다.

ABSTRACT

In this paper, we propose a new methodology for extracting and formalizing subjective topics at a specific time using a set of keywords extracted automatically from online news articles. To do this, we first extracted a set of keywords by applying TF-IDF methods selected by a series of comparative experiments on various statistical weighting schemes that can measure the importance of individual words in a large set of texts. In order to effectively calculate the semantic relation between extracted keywords, a set of word embedding vectors was constructed by using about 1,000,000 news articles collected separately. Individual keywords extracted were quantified in the form of numerical vectors and clustered by K-means algorithm. As a result of qualitative in-depth analysis of each keyword cluster finally obtained, we witnessed that most of the clusters were evaluated as appropriate topics with sufficient semantic concentration for us to easily assign labels to them.

키워드: 키워드 추출, 클러스터링, 토픽 모델링, 단어 임베딩, TF-IDF
keyword extraction, clustering, topic modeling, word embedding, TF-IDF

* 이 논문은 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017M3C4A7068188).
** 경기대학교 일반대학원 문헌정보학과 석사과정(garam1310@kyonggi.ac.kr) (제1저자)
*** 경기대학교 문헌정보학과 조교수(spchoi@kgu.ac.kr) (교신저자)
■ 논문접수일자 : 2018년 3월 8일 ■ 최초심사일자 : 2018년 3월 12일 ■ 게재확정일자 : 2018년 3월 22일
■ 정보관리학회지, 35(1), 231-250, 2018. 〔http://dx.doi.org/10.3743/KOSIM.2018.35.1.231〕

1. 서 론

2015년 한국정보화진흥원에서 수행한 정보격차 실태조사에 의하면 일반 국민의 정보화수준을 100%라 할 때 장애인의 정보화수준은 86.2% 수준에 미치는 것으로 조사되었다. 이에 따라 많은 기관 및 국가 차원의 정보 격차 해소를 위한 노력들이 이루어지고 있다.

〈표 1〉에서와 같이 2004년 57.5% 수준에 머물렀던 장애인 정보화 수준은, 2015년 86.2% 수준까지 미치는 모습을 보여주었다. 이와 같이 장애인의 정보화수준은 꾸준히 증가하고 있는 추세지만, 이에 대한 근본적 해결을 위해서는 정보격차를 일으키는 세부 요인 및 이슈 파악이 지속적으로 이루어져야 한다. 이는 실시간으로 제공되는 장애인 관련 자원을 활용할 수 있다. 그러나 실시간으로 제공되는 대다수의 관련 이슈들은 방대한 양의 비정형 텍스트이다. 이러한 텍스트를 통해 정보 자원에 대한 주제별 이슈 제공이 가능하도록 토픽 모델링(Topic Modeling)을 활용할 수 있다.

토픽 모델링은 특정 문서 집합의 '주제'를 발견하기 위한 통계적 모델이다. 매일 생성되는 방대한 정보는 인간이 직접 처리할 수 있는 수준을 넘어서는데, 토픽 모델을 활용한다면 자동적으로 비정형 텍스트의 집합을 이해하기 쉽도록 조직하고 정리할 수 있다(Blei, 2012).

또한 모델링을 위한 클러스터링 방법으로는 K-평균(K-means) 알고리즘이 존재한다. K-평균은 클러스터(cluster) 개수를 파라미터로 지정하고 데이터를 지정된 k개의 클러스터로 분류하는 알고리즘이다. 각 클러스터와 거리 차이의 분산을 최소화하는 방식을 통해 데이터를 분류한다. 즉, 어느 군에도 속하지 않은 입력 데이터에 적절한 토픽을 부여하는 역할을 수행한다. 예를 들어 n차원의 공간으로 표현된 단어가 m개 존재한다면 사전에 정해진 k 값만큼의 그룹으로 단어를 분류하는 과정이라고 할 수 있다. 앞서 설명한 n차원의 공간으로 표현된 단어는 단어 임베딩 벡터(Word Embedding Vector)로 설명할 수 있다. 단어 임베딩 벡터의 개념은 2장에서 자세히 다룬다.

본 연구는 앞서 언급한 장애인 관련 이슈를 실시간으로 살피고 이를 활용하기 위하여 장애인 관련 뉴스 미디어 자원을 기반으로 주제별 주요 이슈를 추출할 수 있는 방안에 대하여 제시하고자 한다. 토픽 모델링을 위해 K-평균 알고리즘을 사용하였고, 추가적으로 단어 간의 의미적 관계 도출에 용이한 임베딩 벡터를 구성 및 활용하였다. 토픽모델링 분야에서 단어 임베딩 벡터를 사용한 연구는 미미하므로, 기 구성된 임베딩 벡터의 적용 결과를 확인하고자 한다.

〈표 1〉 2004년~2015년 장애인 정보화수준 변화(한국정보화진흥원, 2015)

(단위: %)

장애인 정보화 수준	2004년	2005년	2006년	2007년	2008년	2009년
	57.5	65.2	73.9	76.0	78.8	80.3
	2010년	2011년	2012년	2013년	2014년	2015년
	81.3	82.2	83.4	83.8	85.3	86.2

* 일반국민의 정보화 수준을 100으로 할 때, 장애인 정보화 수준을 의미함

또한 이를 통해 도출된 주제별 키워드들은 의미적으로 다른 주제를 가지고 있을 뿐, 특정 주제에 대해 레이블이 달려있지 않으므로 이를 세부적으로 분석하고 주제명을 정성적으로 레이블링(labeling) 하는 작업을 수행하였다.

2. 관련 연구

2.1 키워드 추출(Keyword Extraction) 및 클러스터링(Clustering)

텍스트 마이닝(Text Mining), 정보 검색(Information Retrieval), 자연어 처리(Natural Language Processing, NLP) 등에서 키워드 추출은 중요한 역할을 한다. 문서에 포함된 관련 정보 중 단어, 구, 문장 등의 형태로 문서 주제를 가장 잘 나타내는 표현을 자동으로 식별해 주는 역할(Beliga, Meštrović, & Martinčić-Ipšić, 2015)을 하기 때문이다. 이러한 이유로 키워드 추출은 다음과 같은 다양한 기법을 중심으로 활발히 연구되고 있다.

먼저, 단어의 가중치를 주는 통계적 기법인 TF(Term Frequency)-IDF(Inverse Document Frequency) 기법이 존재한다. TF-IDF는 키워드 추출을 위해 흔히 사용되는 방식이다. 국내에서는 인터넷 신문 기사 기반의 변형된 TF-IDF 키워드 추출 연구(이성직, 김한준, 2009), 문헌에 존재하는 문장 및 문단에서 키워드의 역할에 따라 가중치를 부여하여 주제어를 추출하는 연구(안희정, 최건희, 김승훈, 2015) 등이 존재한다. 추가적으로 통계적 기법 중의 하나인 t-점수를 활용하여 신문 기사에서 자동으로 키워드를 추출하는 방법 또한 연구되었다(김일환, 이도길, 2011).

확률 기법의 일종으로, 문서의 토픽을 파악할 수 있는 기계 학습 분야의 LDA(Latent Dirichlet Allocation) 기법을 통해 문서에 존재하는 잠재 키워드를 추출하는 연구(조태민, 이지형, 2015), 신문기사를 활용하여 사회적 문제를 해결하기 위한 카테고리 별 키워드 제공 연구(정다미 외, 2013) 등이 존재한다. 추가적으로 딥 러닝 기법을 기반으로 단어 벡터를 활용하여 주요 키워드를 추출하는 연구(김성진, 김건우, 이동호, 2017) 또한 진행되었다.

한편 클러스터링 기법으로는 군집 간의 거리를 이용하여 클러스터링을 하는 K-평균 알고리즘과 계층적 군집 분석(Hierarchical Clustering), 군집 간의 밀도를 기반으로 클러스터링을 하는 밀도 기반 클러스터링(Density-based Spatial Clustering Of Applications With Noise) 방법 등이 존재한다. 클러스터링은 텍스트 분야 이외에도 영상 및 다양한 데이터를 분석하고 처리하는데 사용될 수 있다.

이러한 클러스터링 방식 중 상대적으로 계산량이 적어 데이터 활용에 용이한 K-평균 알고리즘을 활용한 연구는 주성분 분석을 함께 활용하여 군집의 효율을 높일 수 있는 문서군집화 연구(김우생, 김수영, 2014), 인터넷 서비스 사용 시 네트워크 공격에 대한 해결 방안으로 유해트래픽을 탐지하는 연구(신동혁 외, 2016) 등이 존재한다.

키워드 추출과 클러스터링 기법은 앞서 설명한 바와 같이 매우 다양한 방법으로 연구되고 있다. 그러나 본 연구에서는 다양한 방법 중 기계 학습 분야에 비하여 비교적 적은 양의 데이

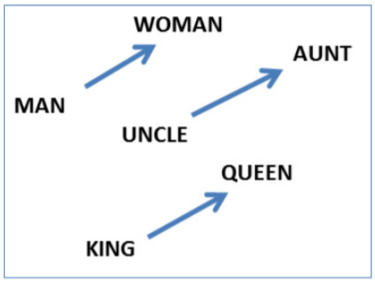
터로도 키워드 추출이 가능한 TF-IDF 가중치 기법과 데이터 계산 시 계산량이 적어 데이터 활용에 용이한 K-평균 알고리즘을 활용하였다. 또한 뉴스 기사를 활용한 다양한 연구들을 바탕으로 본 연구에서는 추가적으로 단어 임베딩 벡터를 구성하였고, 이를 추출된 키워드와 매핑하는 방식을 통해 의미적 유사도를 높일 수 있는 키워드 별 토픽 추출 연구를 진행하였다.

2.2 단어 임베딩 벡터(Word Embedding Vector)

신경망에서 이미지 또는 오디오 데이터를 사용할 경우, 신경망에 입력되는 모든 데이터는 벡터 값으로 변환되어 처리된다. 그러나 자연어의 경우, 데이터 집합에 포함된 모든 단어를 개별적 의미를 지니는 기호(symbol)로 처리해 왔다. 예를 들어 고양이는 'Id537', 강아지는 'Id143' 등으로 변환하여 처리하는 것이다. 이러한 변환 방법은 무작위로 기호를 부여하기 때문에 각 단어 사이에 존재하는 관계 파악이 어렵고, 특정 단어에서 학습한 특징을 이용할 수 없다는 단점이 존재한다. 또한 기호들 사이에 규칙성이 없어 데이터의 군집성이 떨어지므로 충분한 학습을 위해 대량의 데이터가 필요하다는 문제점이 존재한다. 이러한 문제점은 단어를 벡터로 표현하여 해결할 수 있다. 단어를 벡터로 표현하는 과정을 단어 임베딩이라고 부르며, 단어 임베딩을 위해서는 벡터 공간 모델(Vector Space Model)이 주로 사용된다(Tensorflow, 2018).

벡터 공간 모델은 벡터 공간 내 주변에 등장하는 단어는 서로 비슷한 의미를 가진다는 분산 가설(Distributional Hypothesis)(Harris, 1954)에 기반을 두다. 의미상 유사한 단어를 서로 인접시켜 임베딩 시키므로 벡터로 표현된 단어들은 구문론적인 규칙뿐만 아니라 의미적인 부분까지 반영된다. 결론적으로, 단어 임베딩을 통해 생성된 벡터에 따르면 의미론적으로 유사한 단어는 서로 비슷한 벡터를 가지게 된다. 또한 워드 벡터에 포함된 단어들은 모두 수치화되어 있기 때문에 단어와 단어 간의 거리를 활용한 벡터 연산이 가능하다.

〈그림 1〉의 경우 3개의 벡터 오프셋은 성별과 관련된 관계를 나타낸 것임을 확인할 수 있다. 위 벡터공간의 수치를 이용하여 'KING 벡터' - 'MAN 벡터' + 'WOMAN 벡터' = 'QUEEN 벡터'의 오프셋을 계산할 수 있게 된다.



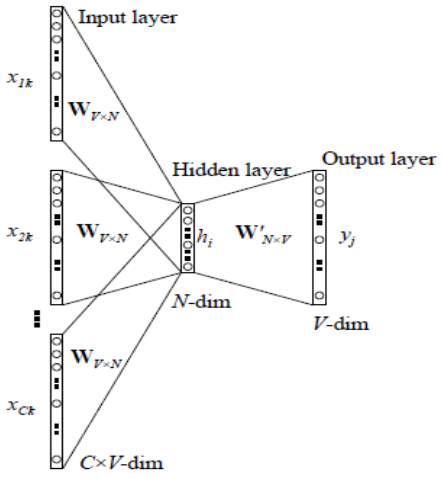
〈그림 1〉 성별 관계 벡터 오프셋(Offset) 예시
(Mikolov, Yih, & Zweig, 2013)

워드 벡터는 기존 임베딩 모델과 비교하였을 때, 학습의 효율이 높은 특징이 있다. 이러한 워드 벡터를 구성하는데 흔히 사용되는 모델은 CBOW(Continuous Bag of Words), Skip-Gram 모델이 있다. CBOW 모델은 문맥을 통해 현재 단어를 예측할 수 있는 방법이며 Skip-Gram 모델은 특정 단어로부터 전체 문맥을 예측할 수 있는 방법이다. 본 논문에서는 연구에서 사용된

CBOW 모델만 자세히 설명한다.

2.2.1 CBOW(Continuous Bag of Words) 모델

CBOW는 문맥을 기반으로 현재 단어를 예측하는 워드 벡터 모델로 입력 계층, 은닉 계층, 출력 계층으로 구성되어 있다. 〈그림 2〉는 CBOW 모델의 구조 예시이다. CBOW 모델의 입력 값은 타겟이 되는 단어의 앞/뒤 단어를 고려하여 원 핫(One Hot) 구조로 인코딩(encoding) 한 단어의 문맥 $\{x_1,...,x_c\}$이 된다. 여기서 C는 윈도우 사이즈(window size)를 의미한다. 입력 값에 곱해지는 행렬은 공통적으로 적용되는 V × N 크기의 가중치 행렬(weight matrix) W이다. 이때 V는 어휘 크기(Vocabulary Size), N은 사용할 벡터의 크기를 의미한다. 입력 계층은 모든 입력 값에 가중치 행렬들을 곱한 뒤 그 벡터들의 평균을 은닉 계층으로 보낸다. 은닉 계층은 입력 벡터와 또 다른 가중치 행렬 W′를 곱한 값을 출력 계층으로 전달한다. 출력 계층은 은닉 계층의 출력 값에 소프트맥스함수(softmax function)를 적용한 결과를 목표 값과 비교하여 오차를 계산한다. 이러한 알고리즘에 기반 한 반복 학습을 통해 데이터 집합에 대한 CBOW 모델 단어 벡터가 생성된다.

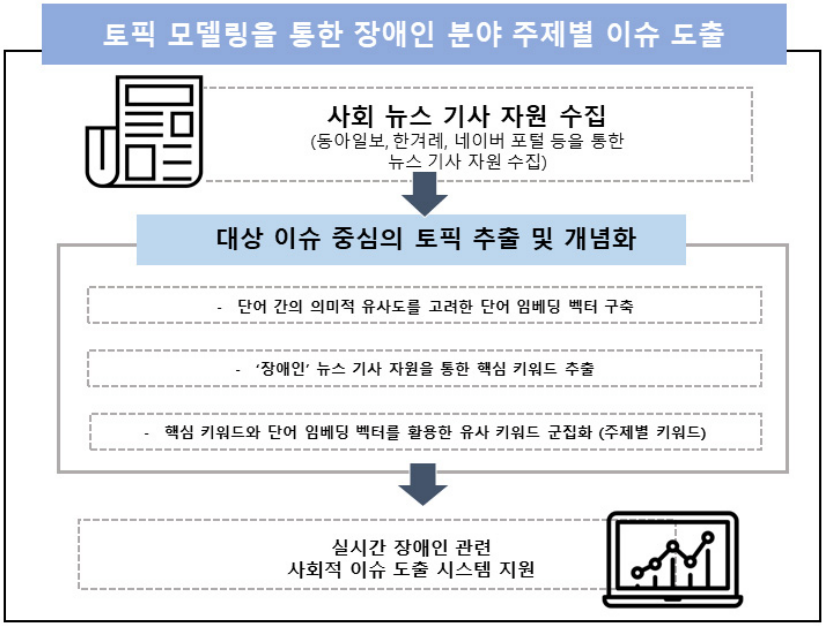


〈그림 2〉 CBOW(Continuous Bag of Words) 모델 구조 예시(Rong, 2014)

3. '장애인' 분야 뉴스 토픽 추출(Topic Extraction)

3장에서는 '장애인' 관련 뉴스 기사 자원을 활용한 토픽 추출의 전반적인 방법론에 대하여 설명한다. 장애인 관련 뉴스에 대한 핵심 키워드를 추출한 뒤 추출 키워드를 클러스터링 하는 방식으로 의미적으로 유사한 키워드 간의 군을 제시한다. 이러한 키워드 군은 서로 다른 주제를 지니므로 이를 통해 주제별 장애인 관련 주요 이슈를 확인할 수 있다.

〈그림 3〉은 '장애인' 관련 토픽 모델링의 전체적인 과정이다. 토픽 모델링의 결과는 실시간으로 '장애인' 관련 이슈 도출 시스템의 지원 방안으로 활용이 가능하다. 본 연구에서는 가장 먼저 뉴스 기사 자원을 수집하였다. 그다음 수집된 자원에 기반 하여 핵심 키워드 추출을 진행하였다. 이 단계에서는 가장 먼저 TF-IDF 가중치 알고리즘을 활용하여 뉴스 기사 자원에서 핵심이 되는 키워드를 추출하고, 추가적으로 키워드 간의 의미적 연관성을 나타낼 수 있는 단어 임베딩 벡터를 구축하였다. 단어 임베딩의 원리 및 개념은 2.2에서 자세히 다루었다. 마지막으로 '장애인' 관련 주제별 주요 이슈를 도출한다. 주제별 주요 이슈 도출을 위해 추출

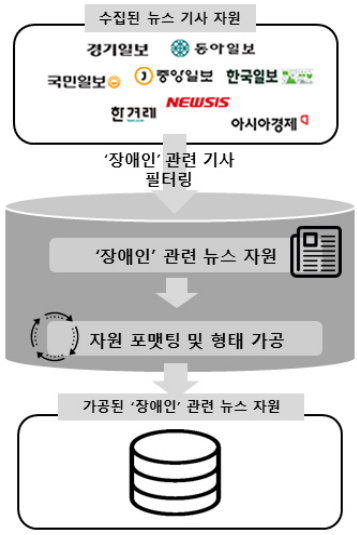


〈그림 3〉 '장애인' 관련 주제별 이슈 도출을 위한 토픽 모델링 개요도

된 키워드와 단어 임베딩 벡터를 기반으로 K-평균 알고리즘을 활용하여 클러스터링 과정을 수행한다. 클러스터링의 결과는 키워드들이 주제 군에 따라 분류된 형태를 의미한다.

3.1 토픽 추출 및 이슈 분석을 위한 뉴스 기사 자원 수집

3.1에서는 토픽 추출 및 이슈 분석을 위해 신문 기사 자원을 수집하였다. 자원의 수집 및 가공 과정은 〈그림 4〉와 같다. 본 연구에서는 동아일보, 한겨레를 대상으로 일차 자원 수집을 수행하였으나, 주제 분야 필터링 후 관련 기사가 양적으로 충분하지 못하다고 판단 되 추가적으로 네이버 포털에 존재하는 뉴스를 활용하였다. 수집된 자원을 기반으로 '장애인' 관련 기사들을 필터링하였으며, 필터링은 주제 분야와의 연관성을 높이기 위해 기사 제목 위주로 '장애인' 키워드 포함 여부를 통해 이루어졌다.



〈그림 4〉 뉴스 기사 자원 수집 및 가공 과정

가장 먼저 토픽 모델링을 위해 수집된 '장애인' 관련 뉴스 자원의 세부 사항은 다음과 같다. 전체 수집된 뉴스 기사 자원에서 필터링을 통해 추출된 장애인 관련 기사는 〈표 2〉와 같다. 네이버 뉴스는 동아일보, 한겨레 등을 포함한 다양한 신문사의 기사를 포함하고 있으므로, 이에 대해 기사 중복 제거 작업을 실시하였다. 장애인 관련 기사는 총 5,088건이나 추가적으로 중복 기사 182건을 제외한 총 4,906건의 기사를 기반으로 전반적인 연구를 진행하였다.

〈표 2〉 키워드 '장애인' 관련 뉴스 기사 자원 수집 결과

데이터 명	'장애인' 데이터 건수
동아일보	1,130
한겨레	197
네이버 뉴스	3,761
총합	5,088
중복 제거	182
총 '장애인' 데이터 건수	4,906

〈표 3〉은 앞서 수집된 기사를 바탕으로 가공된 자원의 실제 예시이다. 최초 수집된 자원은 기사별로 분리된 CSV(Comma-separated Values) 포맷이었으나, 자원 처리 과정에서의 용이함을 위해 아래와 같이 하나의 단일 파일 형태로 변형하는 작업을 수행하였다. 제목과 내용 사이는 분절 자 Tab(\t)을 통해 분리하였으며, 현재 기사와 다음 기사는 개행 문자(\n)로 분리하였다.

3.2 TF-IDF 가중치를 활용한 '장애인' 관련 자원의 핵심 키워드 추출

3.2에서는 수집된 대상 자원의 핵심 키워드 추출 작업을 수행하였다. 가장 먼저 〈그림 5〉와 같이 '장애인' 관련 뉴스 텍스트에 대한 형태소 분석을 진행하였다. 그다음으로, TF-IDF 방법을 활용하여 형태소 분석을 통해 추출된 명사 키워드 별 가중치를 계산하였으며, 가중치 값이 높은 상위의 키워드를 중심으로 추출 작업을 진행하였다.

〈표 3〉 가공된 뉴스기사 자원에 대한 예시

양식	기사제목 \t 기사 내용 \n
실제 데이터 예	"IPTV 등 유료방송에서도 수화방송 나와요" - 케이블티브이(TV)·위성·인터넷티브이(IPTV), 스마트 수화 실험방송 실시- 청각장애인을 위한 스마트 수화방송 상용화 기반 마련[이데일리 김현아 기자] 방송통신위원회(위원장 최성준)는 CJ헬로비전(037560), KT스카이라이프(053210), SK(034730)브로드밴드 등 3개사를 통해 유료방송에서의 스마트 수화방송 실험방송을 11월 16일부터 12월 31일까지 실시한다. 스마트 수화방송서비스는 방송영상과 수화영상을 방송망과 인터넷망으로 각각 제공하고, 수신기에서 두 영상을 동시에 한 화면에 재생하는 방식으로 수화영상의 크기·위치의 조정과 제거가 가능하다. 청각장애인들은 수화화면의 크기가 너무 작아 수화내용을 제대로 이해하기 어렵다는 애로사항과 반면, 수화화면이 너무 커서 방송 내용을 가리는 문제점을 해결하게 됐다. ..(중략).. (사)한국농아인협회와 협력하여 서울농아인협회 양천지부와 수도권에 위치한 수화통역센터 등 17곳에 스마트 수화방송 수신 환경을 설치하고, 청각장애인이 직접 스마트 수화방송을 체험하여 만족도를 조사한 결과를 향후 정책에 반영할 예정이다. ..(중략).. 김현아 (chaos@edaily.co.kr)

〈그림 5〉 '장애인' 뉴스 기사 자원을 활용한 키워드 추출 과정

가장 먼저 수집된 장애인 관련 뉴스 기사에 대하여 KoNLPy 파이썬(Python) 패키지를 사용하여 형태소 분석을 진행한 후 명사만을 추출하는 작업을 수행하였다. KoNLPy에는 여러 형태소 분석 패키지가 존재한다. 해당 자원에 적합한 형태소 분석기 선택을 위해 KoNLPy의 Twitter, Hannanum, Kkma 패키지를 사용하였다.

〈표 4〉는 문장 "IPTV 등 유료방송에서도 수화방송 나와요"에 대한 각 패키지 별 형태소 분석에 기반을 둔 명사 추출 결과이다. 세 패키지 중 명사 추출에 적합하다고 판단된 Twitter를 통해 신문 자원에 대한 형태소 분석을 진행하였다. Hannanum은 특수 문자를 포함하고 있는 문장에 대해 형태소 분석이 잘 이루어지지 않았고, Kkma의 경우 같은 용어를 여러 번 반복하여 형태소 분석이 되는 결과를 도출하였다.

그다음으로는 형태소 분석을 통해 추출된 명사에 대한 가중치 계산을 진행하였다. 가중치는 정보검색(Information Retrieval) 분야에서 문서 및 문헌의 단어 중요도 추출을 위해 사용되는 TF-IDF 모델을 활용하였다.

〈표 5〉는 TF-IDF의 가중치를 구하는 기본 공식 및 변형 공식이다. 표의 기본 공식에서 설명한 것과 같이 TF는 본래 각 단어가 하나의 문서를 기준으로 몇 번 출현하였는지 그에 대한 빈도수를 기반으로 측정된다. 그러나 본 연구에서 도출하려는 키워드는 문서 집합 전반에 대한 핵심 키워드이므로 하나의 문서가 아닌 전체 문서 집합을 대상으로 확장하여 TF 가중치를 측정하였다(이성직, 김한준, 2009). 〈표 5〉에서의 변형된 TF 측정 방식은 단일 문서가 아닌 전체 문서 집합으로 확장하여 단어의 빈도수를 측정하는 BTF 가중치 측정 방법과 BTF 값을 최

〈표 4〉 KoNLPy 형태소 분석 패키지 성능 비교

KoNLP 형태소 분석 예시		
문장: "IPTV 등 유료방송에서도 수화방송 나와요"		
Twitter	Hannanum	Kkma
등, 유료, 방송, 수화, 방송	"IPTV 등 유료방송 수화방송 나와요"	등, 유료, 유료방송, 방송, 수화, 수화방송, 와요

〈표 5〉 TF-IDF 가중치 모델 공식(이성직, 김한준, 2009)

TF (Term Frequency)	$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: 단어 ti가 문서 dj에서 출현한 회수 $\sum_k n_{k,j}$: 문서 dj에서 모든 단어가 출현한 회수
IDF (Inverse Document Frequency)	$idf = \log \frac{\|D\|}{\|\{d_j \| t_j \in d_j\}\|}$ $\|D\|$: 문서집합에 포함되어 있는 전체 문서의 수 $\|\{d_j \| t_j \in d_j\}\|$: 단어 t_j가 등장하는 문서의 수
TF-IDF	$TFIDF_{i,j} = tf_{i,j} \times idf_i$
변형된 TF (Modified Term Frequency)	$BTF_i = \sum_{j=1}^{\|D\|} n_{i,j}$
	$NTF_i = \frac{BTF}{Max\{BTF_1, BTF_2, BTF_3 \cdots\}}$
변형된 TF-IDF	$TFIDF_i = NTF_i \times IDF_i$ $NTFIDF_i = (\log(NTF_i) + 1.0) \times IDF_i$

댓값으로 나누어 정규화한 NTF 가중치 측정 방법이 존재한다. IDF의 경우 기존 방식을 그대로 유지한다. 최종적으로 NTF 가중치 측정 방법에 IDF를 단순히 곱한 $TFIDF$ 값과 NTF에 log를 씌운 후 1.0의 값을 더 해 정규화 한 뒤 IDF 값을 곱한 $NTFIDF$ 측정법이 존재한다.

3.3 키워드 간 의미적 연관성 고려를 위한 단어 임베딩 벡터 구축

추가적으로, TF-IDF 가중치 방법을 통해 추출된 핵심 명사 간의 의미적 연관성을 효과적으로 고려하기 위한 단어 임베딩 벡터를 구성하였다. 한국어 단어 임베딩 벡터 구축을 위해 본 논문에서는 네이버 신문 기사 100만 건의 데이터를 사용하였다. 신문기사 데이터는 공백 단위의 어절 분리 작업을 거친 후 벡터로 생성되었다.

네이버 신문기사 데이터는 앞서 설명한 〈표 3〉과 같이 '기사제목\t기사내용\n' 형태로 1차 가공이 수행되었다. 1차적으로 가공된 신문기사 데이터에 추가적으로 특수기호를 포함한 텍스트에 대한 기타 전처리 과정을 수행하였다.

단어 벡터 생성을 위해서는 Google의 Word2 Vec 중 문맥을 통해 단어를 예측하는 CBOW 모델을 사용하였다. 〈표 6〉에서와 같이 네이버 신문 기사 어절 단위 워드 벡터는 dimension 300, window size 5, min-count 2, epoch 70으로 설정되었다. 여기서 dimension은 단어 임베딩 차원, window size는 단어 임베딩 학습 시 한 번에 학습되는 단어의 수, min-count는 문서에서 단어가 min-count 이하 빈도만큼 존재할 경우 단어 임베딩 생성에서 제외되는 수치 값, epoch는 학습의 반복 횟수를 의미한다. token은 생성할 단어 임베딩 파일에 존재하는 전체 단어 수

〈표 6〉 네이버 뉴스 기사 어절 단위 데이터 단어 임베딩 정보

단어 임베딩 파라미터				단어 임베딩 단어 정보	
dimension	window size	min count	epoch	token	vocab
300	5	2	70	237,872,111	2,895,103

이며, vocab는 token에서 공통 단어를 제외한 단어 수이다. 한편 token의 수는 237,872,111개이며, vocab 수는 2,895,103개이다.

3.4 '장애인' 관련 핵심 키워드 기반의 키워드 별 주제(Topic) 추출

3.4에서는 '장애인' 관련 주요 이슈 파악을 위한 최종 단계로 추출된 '장애인' 핵심 키워드를 기반으로 키워드 별 토픽 화 연구를 진행하였다. 앞선 3.3에서 구성된 단어 벡터와 TF-IDF 상위 값 키워드에 대해 매핑 작업을 수행한 후, 매핑 키워드를 기반으로 K-평균 알고리즘을 활용하여 의미적 유사도를 고려한 키워드 클러스터링을 진행하였다. 이에 따라 키워드를 주제별로 분류하고, 이를 통한 주제별 핵심 키워드를 제공한다.

뉴스 벡터를 통해 매핑된 상위 키워드 데이터를 기반으로 sklearn 패키지 중 K-평균 알고리즘을 사용하여 유사 키워드 간의 클러스터링을 진행하였다. 클러스터링은 입력 데이터에 적절한 토픽을 부여하는 역할을 수행하게 된다.

본 연구에서는 클러스터링을 통하여 상위 키워드를 200개의 토픽으로 분류하였다. 〈표 7〉의 클러스터의 수와 최대 학습 수는 본 데이터를 사용하여 여러 수치 값을 적용해본 후 실제 클러스터링 된 키워드들의 결과를 비교하여 가장 최적의 값이 도출될 수 있도록 한 파라미터이다.

〈표 7〉 K-평균 클러스터링에 사용된 파라미터 값

K-평균 알고리즘 파라미터	파라미터 값
데이터 벡터 차원	300
클러스터의 수	200
최대 학습 수	300

4. 토픽 추출(Topic Extraction) 결과

4.1 TF-IDF 가중치 기반의 키워드 추출 결과

'장애인' 뉴스 기사 자원에서 추출된 전체 명사는 850,731개 이며 중복되는 명사를 제거하여 카운트 한 결과 총 25,310개의 명사가 도출되었다. 본 연구에서는 추출 명사 25,310개에 대한 TF-IDF 계산을 진행하였다.

〈표 8〉에서와 같이 변형된 TF-IDF 가중치 계산법을 통해 도출된 결과를 비교해 보았다. 본래 앞서 설명한 두 가지 TF-IDF 가중치 계산 방법 중 상위 10개의 키워드를 기반으로 더 나은 결과를 도출하는 계산 방법에 기반을 둔 연구를 진행하려 했으나, 두 가지 방법이 거의 동일한 키워드 순위를 도출하는 결과를 보였다. 추가적으로, TF-IDF 가중치 계산 결과는 한 음절 문자를 제외한 결과이다.

〈표 8〉 변형된 TF-IDF 가중치 계산 별 키워드 추출 결과 비교

	TFIDF	*NTFIDF*
1	장애인	장애인
2	**기자**	**기자**
3	시설	시설
4	장애	장애
5	**서울**	**서울**
6	지원	지원
7	사회	사회
8	복지	복지
9	**뉴스**	센터
10	센터	**뉴스**

*** 굵은 글씨는 불용어 수준의 명사를 뜻함**

10개의 키워드 중 두 방법 모두 불용어(stop words) 수준의 명사와 형태소 분석 시 잘못 분석된 명사를 상위에 가지고 있었다. 불용어는 보통 영어에서 관사, 전치사, 대명사와 같이 자주 사용되는 단어로, 정보 검색 시 색인어로 사용되지 않는 용어를 의미한다. 불용어 수준의 명사는 수집된 신문기사 자원에 포함된 '기자', '서울', '뉴스' 등의 키워드를 의미한다. 이러한 키워드는 모든 신문기사에 포함된 명사이므로 전반적으로 높은 가중치 값을 얻어 상위에 링크된다. 두 가중치 측정 방법 모두 유사한 결과를 도출하였기 때문에 본 연구에서는 전반적으로 가중치 값을 정규화 한 *NTFIDF* 값을 사용하여 연구를 진행하였다.

NTFIDF 가중치 방식을 통해 추출된 상위 100개의 키워드는 〈표 9〉와 같다. 굵은 글씨는 불용어 수준의 명사를 의미하며, 밑줄 글씨는 형태소 분석 시 잘못 분석된 명사를 의미한다. 상위 키워드는 '장애인', '시설', '장애', '지원', '사회', '복지', '센터', '한국', '사업' 등의 순이다. 상위에 존재하는 키워드들은 사회적으로 이루어지는 복지 사업 및 지원과 관련된 키워드들이다. 뉴스 기사의 경우 복지 사업 및 지원에 대한 내용을 많이 다루고 있기 때문이다.

추출된 키워드에서 특정 장애와 관련된 키워드는 '발달장애(15위)', '시각장애인(18위)', '지적장애(82위)', '청각장애(169위)', '시각장애(315위)', '정신장애인(636위)', '무장애(902위)', '장애우(1168위)', '언어장애(1887위)', '정신장애(2044위)', '장애자(2335위)', '절단장애인(5114위)' 등의 순이었다.

'대회(16위)'와 관련된 키워드는 '스포츠(230위)', '체전(279위)', '패럴림픽(357위)', '장애인올림픽(399위)', '스페셜올림픽(1319위)', '선수권대회(1446위)', '골프대회(2699위)', '전국체육대회(2720위)', '농구대회(6884위)' 등의 순이다. '체전' 키워드의 경우 다양하게 열린 '장애인 체전', '장애인체전'과 관련된 기사를 통해 추출되었으며, '선수권대회'의 경우 최근 2017년에 열린 '제7회 장애인 트라이애슬론 선수권대회' 관련 기사를 통해 추출된 키워드로 추측된다. 또한 지적 장애인 발달 대회인 '스페셜올림픽' 키워드 또한 살펴볼 수 있었다.

또한 '정책(80위)' 키워드와 관련 있는 '대통령(454위)'의 실제 이름은 '문재인(1019위)', '박근혜(2501위)', '이명박(2892위)' 등의 순이었다. '문재인', '박근혜' 대통령과 관련된 기사는 수집 시점에서의 '장애인' 정책 관련 기사가 주를 이루었으며 '박근혜' 전 대통령의 기사에는 관련 시위에 참가한 장애인 연합 및 연대에 대한 기사 또한 다수 포함되어 있었다.

〈표 9〉 *NTFIDF* 가중치 계산을 통해 추출된 키워드 상위 100개

	NTFIDF		*NTFIDF*		*NTFIDF*		*NTFIDF*		*NTFIDF*
1	장애인	11	위해	21	고용	31	지난	41	단체
2	**기자**	12	한국	22	지역	32	**부산**	42	**제공**
3	시설	13	사업	23	교육	33	**배포**	43	재활
4	장애	14	애인	24	전국	34	**무단**	44	휠체어
5	**서울**	15	발달장애	25	서비스	35	활동	45	경기
6	지원	16	대회	26	학교	36	만원	46	대상
7	사회	17	대한	27	**제보**	37	**전재**	47	운영
8	복지	18	시각장애인	28	**금지**	38	기관	48	행사
9	센터	19	**사진**	29	이번	39	선수	49	경찰
10	**뉴스**	20	**뉴시스**	30	통해	40	**클릭**	50	사람
51	서울시	61	학생	71	오후	81	정부	91	주민
52	계획	62	문화	72	예정	82	지적장애	92	최신
53	조사	63	관계자	73	경제	83	경우	93	관련
54	생활	64	차별	74	보조	84	협회	94	문제
55	시간	65	올해	75	인권	85	기업	95	비장
56	**광주**	66	정보	76	직업	86	개최	96	이용
57	지난해	67	공단	77	복지관	87	훈련	97	특수
58	가족	68	**페이스북**	78	직원	88	대표	98	이상
59	**연합뉴스**	69	영상	79	보호	89	협의	99	협약
60	**종합**	70	편의	80	정책	90	**인천**	100	병원

* **굵은 글씨는 불용어 수준의 명사를 뜻함**
* 밑줄 글씨는 형태소 분석 시 잘못 분석된 명사를 뜻함

4.2 '장애인' 관련 상위 추출 키워드와 단어 임베딩 벡터 매핑(Mapping)

앞서 3.3에서 구성된 임베딩 벡터를 본 논문에서는 '뉴스 벡터'라고 명명한다. 먼저 신문 자원에서 추출한 키워드 25,310개에 대한 *NTFIDF* 가중치 값 중 *NTFIDF* 가중치 값이 어느 정도 유효한 범위 내에 있다고 가정되는 10,000개의 키워드만을 뉴스 벡터와 매핑 시키는 작업을 수행하였다. 다음과 같이 임베딩 벡터가 매핑된 키워드를 기반으로 클러스터링을 진행했을 시 클러스터군 내의 대부분의 단어들이 의미적으로 유사한 결과를 보인다는 것을 확인할 수 있었다. 이에 대한 결과는 4.3 및 부록에서 확인 가능하다. 뉴스 벡터 매핑 시 벡터에 존재하지 않는 키워드 382개가 제외되었으며 나머지 9,620개의 키워드에 대한 매핑 데이터가 구성되었다. '*NTFIDF* 추출 키워드에 대한 뉴스 벡터 매핑 데이터'는 〈그림 6〉과 같다.

뉴스 벡터 매핑 시 벡터에 존재하지 않는 키워드들은 〈표 10〉과 같다. 탈락된 키워드의 대다수가 불용어 수준 및 형태소 분석 시 잘못 분석된 키워드이며, 사람 이름과 함께 대부분 쉽게 쓰이지 않는 키워드들이다. 이러한 키워드들은 *NTFIDF* 가중치 값 또한 하위에 존재하는 것을 확인하였다. 그러나 그 외 '동권', '보조

장애인	3.461632	-0.518432	2.594484	2.445098	-0.674228	-0.392791	0.288365	-1.150287
기자	3.252808	2.061334	0.946388	1.789317	-0.776933	-0.532352	-1.684827	1.526453
시설	1.079243	0.613767	-1.937495	1.338347	-0.305160	0.331817	2.609923	2.375892
장애	-0.701520	-2.437190	-0.578029	2.299606	0.115330	-2.137783	-0.778102	-0.787803
서울	4.790877	1.395647	1.210949	0.142963	-1.570034	-0.539855	-0.680606	-0.530639

〈그림 6〉 *NTFIDF* 추출 키워드 뉴스 벡터 매핑 데이터 예시

〈표 10〉 *NTFIDF* 추출 키워드와 뉴스 벡터 매핑 시 탈락된 키워드

	키워드		키워드		키워드
1	증장	11	장애인스포츠	21	꿈드
2	동권	12	유일하	22	노인장
3	인복	13	데리	23	여준
4	노해	14	피스토리우스	24	인승
5	관왕	15	후천	25	김형
6	보조공학	16	폐쇄회	26	예땅
7	노약	17	권법	27	불렀
8	호점	18	라르	28	살피
9	양낙규	19	자법	29	유도기
10	오마이	20	한상민	30	탑산

공학', '장애인스포츠', '피스토리우스' 등과 같이 비교적 중요한 키워드들도 탈락되는 모습을 보였다.

4.3 '장애인' 관련 핵심 키워드 기반의 키워드 별 주제 추출 결과

3.4에서 설명한 클러스터링 방식을 통해 200개의 토픽으로 분류된 키워드들을 본 논문에서 모두 다루기는 어렵다고 판단 되, 그중 키워드들이 주제명을 두드러지게 담고 있다고 판단된 12개의 토픽에 대해 토픽 명을 부여하고 그에 대한 세부 분석을 진행하였다. 결과적으로 9,620개의 키워드를 200개의 토픽으로 분류하는 작업을 수행하였으며 이를 기반으로 12개의 토픽을 정성적으로 분석하였다.

200개의 토픽 중 12개의 키워드 군을 대상으로 토픽 명을 부여한 후 키워드 군의 번호에 따라 오름차순 정렬하였다. 클러스터링을 통한 키워드 주제 분류 전체 결과는 [부록 1]에서 확인 할 수 있다. 전체 결과의 토픽 키워드는 앞서 구한 *NTFIDF* 값 순으로 나열되어 있으므로 등장 순서가 빠를수록 해당 토픽에서 중요한 키워드로 간주될 수 있다.

9번 '범죄' 토픽의 경우 현재 발생한 피해 및 범죄 관련 키워드들이 나열되어 있다. '경찰', '폭행', '학대', '위반', '범죄', '형사', '성폭력', '성폭행', '횡령' 등의 키워드 순이며, '사기', '노예', '폭력', '욕설', '성범죄', '감금' 등의 키워드들도 포함되어 있다. 이 토픽은 현재 이슈가 되고 있

〈표 11〉 특정 주제 별로 분류된 키워드(요약)

주제명	키워드
범죄(9번)	경찰, 폭행, 학대, 위반, 범죄, 형사, 성폭력, 성폭행, 횡령, 사기, 노예, 폭력, 욕설, 성범죄, 감금
차별(16번)	차별, 지체, 도움, 어려움, 편견, 고통, 형편, 걱정, 지장, 장벽, 구분
인권(33번)	인권, 약자, 권리, 연대, 국민, 철폐, 권익, 노동, 자유, 책임
정치(37번)	국민의당, 한나라당, 새누리당, 김현미, 김성태, 김영주, 박경미, 나경원, 홍준표
사회적 약자(78번)	장애인, 발달장애, 고용, 거주, 자립, 자립생활, 계층, 빈곤, 요양, 의족, 노후, 다문화, 육아, 사회생활, 은퇴, 정보격차
이동 및 교통수단(125번)	휠체어, 택시, 주차, 차량, 버스
교육(153번)	교육, 활동, 직업, 훈련, 체육, 상담, 근무, 수업, 공부, 심리검사, 안전교육, 자기계발, 응급처치, 심폐소생술
스포츠(155번)	선수, 체전, 선수단, 패럴림픽, 출전, 금메달, 올림픽, 스키, 마라톤, 육상, 스페셜올림픽, 아시안, 선수권, 대회, 보치아
질병 및 장애(158번)	장애, 중증, 청각, 치과, 발달, 질환, 자폐, 뇌성마비, 입원, 치매, 마비
장소(161번)	병원, 구역, 건물, 체육관, 공원, 청사, 경기장
대통령(165번)	문재인, 박근혜, 이명박, 노무현, 김대중
복지(191번)	시각장애인, 복지관, 청각장애, 봉사, 나눔, 후원, 기부, 시각장애, 봉사활동, 공헌, 안내견, 흰지팡이, 벽화, 배식, 꽃동네, 적십자, 어린이재단, 생필품, 영정사진

는 관련 범죄 인식 및 예방에 도움이 될 수 있을 것이다.

16번 '차별' 토픽의 경우 사회적인 차별 및 인식과 관련된 키워드들이 나열되어 있다. '차별', '지체', '도움', '어려움', '편견', '고통', '형편', '걱정', '지장', '장벽', '구분' 등의 키워드 순이다. 이를 통해 사회적으로 어떠한 차별을 느끼는지, 또한 관련된 문제점이 무엇인지에 대하여 파악하는데 도움이 될 수 있을 것이다.

33번 '인권' 토픽의 경우 인간에 대한 각종 권리와 관련된 키워드들이 나열되어 있다. '인권', '약자', '권리', '연대', '국민', '철폐', '권익', '노동', '자유', '책임' 등의 키워드 순이다. 이러한 키워드를 활용한다면 현재 문제 및 요구되는 권리가 무엇인지 파악하는데 도움이 될 것이다.

37번 '정치' 토픽의 경우 정치적으로 관련이 있는 정당 및 국회의원 키워드들이 나열되어 있다. 정치 정당은 '국민의당', '한나라당', '새누리당' 등의 순이며, 관련 정치인으로는 '김현미', '김성태', '김영주', '박경미', '나경원', '홍준표' 등의 순이다. '정치' 토픽을 통해 추출되는 키워드는 현재 어떠한 정당 및 정치인이 장애인과 관련해 관심 및 중요도를 지니는지에 대한 확인이 가능하다.

78번 '사회적 약자' 토픽의 경우 장애인을 비롯하여 사회적 약자이거나 그와 관련된 사회적 키워드들이 나열되어 있다. '장애인', '발달장애', '고용', '거주', '자립', '자립생활', '계층' 등의 키워드 순이며, '빈곤', '요양', '의족', '노후', '다문화', '육아', '사회생활', '은퇴', '정보격차' 등의 키워드들도 포함되어 있다. 이 토픽의 키워드는 현재 사회에서 사회적 약자로 인식되는 계층과 그에 대한 사회적 인식 및 문제 파악에 도움이 될 수 있을 것이다.

125번 '이동 및 교통수단' 토픽의 경우 이용

되는 이동 및 교통수단과 관련된 키워드들이 나열되어 있다. '휠체어', '택시', '주차', '차량', '버스' 등의 키워드 순이다. 이를 통해 장애인이 어떤 수단을 활용하고, 활용 시 어떠한 어려움을 겪는지에 대한 문제점 파악 및 그에 대한 해결점 제공에 도움이 될 수 있을 것이다.

153번 '교육' 토픽의 경우 교육과 관련된 키워드들이 나열되어 있다. '교육', '활동', '직업', '훈련', '체육', '상담', '근무', '수업', '공부' 등의 키워드 순이며, '심리검사', '안전교육', '자기계발', '응급처치', '심폐소생술' 등의 키워드들도 포함되어 있다. 이를 통해 현재 장애인들이 어떠한 교육에 관심이 있는지, 필요로 하는 교육의 무엇인지 파악하는데 도움이 될 수 있을 것이다.

155번 '스포츠' 토픽의 경우 스포츠 및 각종 대회에 대한 키워드들이 나열되어 있다. '선수', '체전', '선수단', '패럴림픽', '출전', '금메달', '올림픽', '스키', '마라톤', '육상' 등의 키워드 순이며, '스페셜올림픽', '아시안', '선수권', '대회', '보치아' 등의 키워드들도 포함되어 있다. 이러한 키워드를 활용한다면 현재 어떠한 스포츠 및 대회가 이루어지고 있는지 또한 이에 대한 선수 및 정보 파악에 용이할 것이다.

158번 '질병 및 장애' 토픽의 경우 관련 장애와 질병에 대한 키워드들이 나열되어 있다. '장애', '중증', '청각', '치과', '발달', '질환', '자폐', '뇌성마비', '입원', '치매', '마비' 등의 키워드 순이다. 이 토픽의 키워드는 현재 장애인과 관련된 중요 장애 및 질병을 확인할 수 있을 것이다.

161번 '장소' 토픽의 경우 특정 장소에 대한 키워드들이 나열되어 있다. 장소의 경우 '병원', '구역', '건물', '체육관', '공원', '청사', '경기장' 등의 순이다. 이를 활용한다면 장애인과 관련하여 발생하는 사건 및 문제 장소 파악에 용이할 것이다.

165번 '대통령' 토픽의 경우 관련 기사에서 많이 등장하는 대통령의 이름에 대한 키워드들이 나열되어 있다. '문재인', '박근혜', '이명박', '노무현', '김대중' 순이다. 이를 활용하여 관련 기사들을 살펴본다면 현재 및 과거에 어떠한 대통령이 장애인 관련 문제에 어떠한 관심을 가지는지를 살펴볼 수 있을 것이다.

191번 '복지' 토픽의 경우 현재 복지가 필요한 대상 및 복지 관련 기관들 그리고 그에 대한 키워드들이 나열되어 있다. '시각장애인', '복지관', '청각장애', '봉사', '나눔', '후원', '기부', '시각장애인', '봉사활동', '공헌' 등의 순이며, '안내견', '흰지팡이', '벽화', '배식', '꽃동네', '적십자', '어린이재단', '생필품', '영정사진' 등의 키워드들도 포함되어 있다. 이러한 키워드를 활용한다면 필요로 하는 복지 대상 및 복지 기관 그리고 현재 이루어지고 있는 봉사 및 나눔 관련 키워드 파악에 도움이 될 것이다.

본 연구에서는 단어 임베딩을 활용하여 추출된 핵심 키워드를 기반으로 의미적으로 유사한 키워드들을 분류하는 작업을 수행하였다. 또한 이에 대한 키워드들을 세부적으로 파악하고 분석하여 키워드 별 토픽 명을 부여하였다. 특정 자원을 바탕으로 유사 키워드에 대한 토픽을 추출한 결과인 특정 자원의 토픽 별 이슈 키워드들은 키워드 군 내의 대다수 이슈가 주제 분야와 연관되어 있으며, 이슈 간의 의미적 관계 또한 유사함을 확인 할 수 있었다.

5. 결론 및 제언

본 논문에서는 뉴스기사 자원을 활용하여 이에 대한 주요 키워드 추출과 추출 키워드 간의 의미적 유사도를 고려한 클러스터링 방식을 통해 주제 별 키워드 군을 제시하였다. 이러한 제시 키워드 군을 통해 '장애인' 관련 주요 이슈를 주제별로 확인 할 수 있었다.

가장 먼저 장애인 관련 분야의 사회 미디어 자원을 수집하였고, 수집된 자원에 기반을 두어 TF-IDF 가중치 기법을 활용하여 전체 문서 군에 대한 핵심 키워드를 추출하였다. 추가적으로 특정 키워드 간의 의미적 유사도를 수치화 한 단어 임베딩 벡터를 구성하였으며, 이를 추출된 핵심 키워드에 매핑하는 방식을 통해 키워드 간의 의미적 관계를 도출하였다. 최종적으로 K-평균 알고리즘을 활용하여 키워드 간의 의미적 관계에 따른 주제별 키워드 군을 제시하였다. 이러한 결과는 '장애인' 관련 실시간 주요 이슈 파악을 위한 데이터로 사용 될 수 있으며, 학문적인 관점에서는 계량 정보학, 문헌 분석, 내용 분류 분야에서의 핵심 기술로 활용이 가능하다.

그러나 본 연구의 방법론을 실용화하기 위해서는 다음과 같은 한계점 해결이 요구된다. 가장 먼저 형태소 분석 시 발생하는 복합 명사 추출 문제이다. 본 연구에서는 형태소 분석의 문제로 인하여 복합 명사를 활용할 수 없었다. 특정 분야에 대한 핵심 키워드는 단일 명사뿐만 아니라 복합 명사 형태로도 다수 존재할 것이다. 이러한 문제점은 형태소 분석 시 전처리 및 후처리 방식을 통해 해결이 가능할 것이라 생각된다. 두 번째로는 단어의 클러스터링을 통한 토픽의 레이블링 문제이다. 대다수의 클러스터링은 문서 및 단어를 분류해줄 뿐 해당 단어에 대한 주제명을 제시해 주지 못한다. 본 연구 또한 토픽 별 클러스터를 구성한 후 데이터 분석을 통하여 토픽 명칭을 정성적으로 레이블링하는 작업을 수행하였다. 그러나 실시간으로 특정 주제 분야를 레이블링 하는 것은 매우 어려운 일이다. 이는 문서 및 주제별 데이터 반자동 분류가 가능한 딥 러닝(Deep Learning) 모델을 활용하여 레이블 학습을 통한 키워드 분류를 진행할 수 있을 것이다. 세 번째로는 단어 벡터 구성과 관련한 문제점이다. 본 연구에서 사용된 단어 벡터는 띄어쓰기 단위를 기반으로 구성된 단어 벡터이다. 만약 형태소 분석의 문제점이 해결된다면 벡터 구성을 위해 수집된 자원을 형태소 분석한 후 활용하는 방식을 통해, 지금보다 유의미한 키워드 군을 도출 할 수 있을 것이다.

참 고 문 헌

김성진, 김건우, 이동호 (2017). 딥러닝 기반의 뉴스 분석을 활용한 주제별 최신 연관단어 추출 기법. 한국정보처리학회 2017년 춘계학술발표대회 논문집, 873-876.

김우생, 김수영 (2014). 주성분 분석과 k 평균 알고리즘을 이용한 문서군집 방법. 한국정보통신학회논

문지, 18(3), 625-630. http://dx.doi.org/10.6109/jkiice.2014.18.3.625

김일환, 이도길 (2011). 대규모 신문 기사의 자동 키워드 추출과 분석. 한국어학, 53, 145-194.

신동혁, 안광규, 최성춘, 최형기 (2016). K-평균 클러스터링을 이용한 네트워크 유해트래픽 탐지. 한국통신학회논문지, 41(2), 277-284. http://dx.doi.org/10.7840/kics.2016.41.2.277

안희정, 최건희, 김승훈 (2015). 키워드 가중치 방식에 근거한 도서 본문 주제어 추출. 한국컴퓨터정보학회 학술발표논문집, 23(1), 19-22.

이성직, 김한준 (2009). TF-IDF 의 변형을 이용한 전자뉴스에서의 키워드 추출 기법. 한국전자거래학회지, 14(4), 59-73.

정다미, 김재석, 김기남, 허종욱, 온병원, 강미정 (2013). 사회문제 해결형 기술수요 발굴을 위한 키워드 추출 시스템 제안. 지능정보연구, 19(3), 1-23. http://dx.doi.org/10.13088/jiis.2013.19.3.001

조태민, 이지형 (2015). LDA 모델을 이용한 잠재 키워드 추출. 한국지능시스템학회 논문지, 25(2), 180-185. http://dx.doi.org/10.5391/JKIIS.2015.25.2.180

한국정보화진흥원 (2015). 2015 정보격차 실태조사. 서울: 미래창조과학부.

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. Journal of Information and Organizational Sciences, 39(1), 1-20.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84. http://dx.doi.org/10.1145/2133806.2133826

Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162. http://dx.doi.org/10.1080/00437956.1954.11659520

KoNLPy. Retrieved from http://konlpy.org/en/v0.4.4/

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. International Conference on Learning Representation 2013.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. NAACL-HLT, 746-751.

Rong, X. (2014). Word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.

Sklearn.cluster.KMeans. Retrieved from http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans

Tensorflow Motivation: Why learn word embeddings? (2018). Retrieved from https://www.tensorflow.org/tutorials/word2vec

Word2vec. Retrieved from https://code.google.com/archive/p/word2vec/

• **국문 참고문헌에 대한 영문 표기**
(English translation of references written in Korean)

Ahn, Hee-Jeong, Choi, Gun-Hee, & Kim, Seung-Hoon (2015). Thematic word extraction from book based on keyword weighting method. The Journal of Korean Society of Computer and Information, 23(1), 19-22.

Cho, Taemin, & Lee, Jee-Hyong (2015). Latent keyphrase extraction using LDA model. Journal of Korean Institute of Intelligent Systems, 25(2), 180-185. http://dx.doi.org/10.5391/JKIIS.2015.25.2.180

Jeong, Dami, Kim, Jaeseok, Kim, Gi-nam, Heo, Jong-Uk, On, Byung-Won, & Kang, Mijung (2013). A proposal of a keyword extraction system for detecting social issues. Journal of Intelligence and Information Systems, 19(3), 1-23. http://dx.doi.org/10.13088/jiis.2013.19.3.001

Kim, Illhwan, & Lee, Do-Gill (2011). Automatic keyword extraction and analysis from the large scale newspaper corpus based on t-score. Korean Linguistics, 53, 145-194.

Kim, Sung-Jin, Kim, Gun-Woo, & Lee, Dong-Ho (2017). A topic related word extraction method using deep learning based news analysis. The Journal of Korea information Processing Society, 873-876.

Kim, Woosaeng, & Kim, Sooyoung (2014). Document clustering technique by k-means algorithm and PCA. Journal of the Korea Institute of Information and Communication Engineering, 18(3), 625-630. http://dx.doi.org/10.6109/jkiice.2014.18.3.625

Lee, Sung-Jick, & Kim, Han-Joon (2009). Keyword extraction from news corpus using modified TF-IDF. The Journal of Society for e-Business Studies, 14(4): 59-73.

National Information Society Agency (2015). 2015 The report on the digital divide. Seoul: Ministry of Science, ICT and Future Planning.

Shin, Dong Hyuk, An, Kwang Kue, Choi, Sung Chune, & Choi, Hyoung-Kee (2016). Malicious traffic detection using k-means. The Journal of Korean Institute of Communications and Information Sciences, 41(2), 277-284. http://dx.doi.org/10.7840/kics.2016.41.2.277

[부록 1] 특정 주제 별로 분류된 '장애인' 관련 추출 키워드 (전체)

토픽 번호	주제명	키워드 개수	키워드
9	범죄	130	경찰 폭행 학대 위반 행위 범죄 형사 성폭력 성폭행 횡령 불구속 불법 사기 허위 노예 부정 부당 폭력 욕설 상습 집단 알선 가해자 협박 성범죄 감금 살인 방해 비리 상해 자살 수법 성추행 절도 체벌 위법 조작 성희롱 미수 미성년 갈취 성매매 가짜 금품 심부름 유인 포착 유용 강요 아동학대 폭언 가혹 위조 치사 강도 접촉 음주운전 거액 가담 공무 공갈 모욕 도박 가해 성관계 고의 명예훼손 은폐 이권 추행 행각 강간 인척 재범 강제추행 거짓 대필 방화 부당이득 위법행위 부패 수수 유포 무마 범죄자 소란 실화 회계부정 학교폭력 장물 착복 배임 구타 보이스피싱 가정폭력 침입 방임 무면허 성행위 전과자 사범 묵인 조폭 제압 솜방망이 현행범 술자리 리베이트 대마초 견책 무단이탈 체불 마약 성범죄자 데이트 불법행위 비방 투약 공금 일탈 간음 준강간 사칭 유인물 회유 연루 강압 방조 초범 음주
16	차별	28	차별 지체 도움 어려움 편견 고통 형편 걱정 지장 장벽 구분 구별 두려움 빈틈 무리 책임감 삭발 격의 믿음 욕심 단점 생활고 상관 선입견 가난 시행착오 부족함 죄책감
33	인권	89	인권 약자 권리 연대 국민 철폐 권익 노동 자유 책임 인간 차별금지법 공익 생명 참정권 자치 평등 헌법 독립 생존권 정치 명예 민주 거듭 평화 모성 종교 기본권 통일 존중 평등권 정의 부의 남북 한일 소수자 다자 책무 부와 윤리 보편 남북한 화해 유린 의의 국제사회 공적 촛불 질서 탈핵 진영 자기결정권 고취 불평등 민족 선거권 인종 권리장전 선언문 외교 지성 진리 사생활 존엄성 성소수자 인류 기독교 대국민 민주주의 억압 안보 민생 지방자치 유아교육 존엄 인격권 포용 재산권 이념 권력 후생 약법 분권 국가책임 견제 인민 인도주의 강제노동 수호
37	정치	44	국민의당 비례대표 김현미 한나라당 김성태 새누리당 김영주 정의당 박경미 나경원 홍준표 김상훈 안철수 조원진 최경환 유승민 이용득 김재원 중진 김연 한정애 권미혁 유시민 진선미 김경진 김성수 김부겸 양승조 김영진 김승희 황주홍 도당 전라북도의회 이용주 정종섭 도종환 김종석 민병두 윤영석 박영선 김해영 이태규 천정배 노회찬
78	사회적 약자	101	장애인 발달장애 고용 거주 자립 자립생활 계층 가정 직종 취약 근로 구직 소외 지방 정신장애인 청년 종사 직장 평생교육 생계 지원이 구인 사회보장 빈곤 요양 입소 자활 세대 정신건강 의족 노후 다문화 양육 처우 육아 정규직 가사 일과 사회생활 은퇴 간호 농촌 보육 정보격차 재기 단절 경제활동 비정규직 고령화 빈민 향유 평생교육사 입양 노숙인 서민 체류 임대주택 무기계약직 저출산 빈곤층 농어촌 원거리 학령기 통합교육 반값등록금 한부모 심리상담 이주 경조사 이민자 쪽방 디딤돌 보훈 신혼부부 간병 숙식 시간제 문화생활 모니터링단 생업 퇴소 실업 탈북 농업인 실업자 이직 재외국민 맞벌이 숙련 과밀 조손 새터민 응급의료 살림 노동시장 실직 어촌 장년 산후조리 실직자 미혼모 아르바이트 문해 메디케이드
125	이동 및 교통 수단	81	휠체어 택시 주차 차량 버스 운행 탑승 도로 공항 운전 저상버스 자전거 승차 운전자 승객 정차 열차 고속버스 시내버스 시외버스 대중교통 터미널 승합차 배차 노선 운전사 철도 여객선 좌석 셔틀버스 증차 정류장 통학버스 여객 여행지 통학 왕복 차로 비행기 안전벨트 고속도로 시외 전동차 시티투어 기차 빗길 장거리 가드레일 편도 국내선 승용차 유모차 구급차 보행자 관광버스 스쿨버스 렌터카 교차로 트럭 전철 동승 건널목 수하물 단거리 운반 스타렉스 경찰차 화물차 전세기 폐차 환승 운항 도로교통법 경유 오토바이 안전띠 순환버스 국제선 유람선 광역버스 휴식시간
153	교육	57	교육 활동 직업 훈련 체육 상담 근무 수업 공부 학습 독서 자격증 진로 적응 연수 토론 연습 정규 동아리 실습 학기 재학 수련 예비 진학 수강 학업 담임 이수 수학여행 치유 성교육 모의 장학 유학 심리검사 안전교육 일대일 전문교육 여름방학 학교생활 코칭 방학 자기계발 체험학습 합숙 한국어 입문 저학년 방과 방과후 고학년 응급처치 동기부여 겨울방학 계발 심폐소생술

토픽 번호	주제명	키워드 개수	키워드
155	스포츠	147	선수 체전 선수단 패럴림픽 출전 금메달 올림픽 스키 마라톤 육상 국가대표 수영 성화 은메달 전국체전 축구 탁구 야구 대표팀 골프 스페셜올림픽 프로 아시안 선수권대회 농구 시즌 양궁 동계 보치아 봉송 겨룬 배구 볼링 월드컵 테니스 폐회식 경진 알파인스키 기능올림픽 배드민턴 결승 홍석만 펜싱 스케이트 골퍼 동계올림픽 예선 전국체육대회 하키 아이스하키 컬링 빙상 승마 썰매 태권도 게이트볼 선수촌 달리기 열띤 스쿼트 폐막식 카약 평창동계올림픽 사회인 구장 목발 선수권 투수 입단 스노보드 휠체어컬링 바둑 경마 혼성 관중 메달리스트 체조 라운드 줄넘기 던지기 팔씨름 야구단 달팽이 복싱 배영 강습 이강석 복식 득점 특설무대 운동경기 결승전 국가대표팀 응원단 프로야구단 당구 이용대 타이틀 졸업앨범 스포츠토토 철인 마스코트 풋살 아마추어 바이애슬론 투호 최정 리그 유니폼 다트 트라이애슬론 구단 필승 줄다리기 윷놀이 스포츠인 쇼트트랙 군악대 대항 횡단 친선 대성 홈런 김연아 챔피언십 접영 하프 챔피언 응원전 조로 축구팀 스프린트 토너먼트 배구팀 접전 리허설 시타 택견 스윙 관중석 스타크래프트 조가 단체사진 준결승 대국 티업 권투
158	질병 및 장애	151	장애 중증 청각 치과 발달 질환 인과 자폐 뇌성마비 입원 치매 마비 재발 위험 정신질환 설사 심리 언어장애 질병 신장 만성 정신장애 식이 유발 심신 손상 화상 경증 당뇨 비만 정신과 격리 생존 우울증 자가 외상 후유증 장해 다운증후군 유산 소아 호흡기 전이 합병증 루게릭병 선천성 희귀 안과 통원 우울 소견 난치병 노령 임종 노인성 청소년기 지능지수 고혈압 욕창 태아 뇌졸중 부작용 조현병 과로 투석 언어치료 관절염 파열 조울증 심리치료 기형 급성 낙상 증후군 말기 자폐증 출혈 유전 매개 약시 대인 줄기세포 응급실 요법 정신병 알코올 몸무게 각성 지병 기전 류머티스 심장병 골격계 산후 중독 신체검사 우울장애 간질 바이러스 충동 확률 출생 강박 심장질환 사산 탈수 외래 뇌출혈 탈진 내과 녹내장 충치 외과 카페인 감기 중환자실 해독 빈혈 한의원 백선 보충 당뇨병 병인 병세 병기 결핵 정형외과 광범위 피부병 자각 결핍 결손 저체온증 수혈 인플루엔자 췌장암 흡연 간염 백내장 공격성 미숙아 고열 사춘기 장내 파킨슨병 현기증 보철 난청 과잉행동 말라리아 인격장애
161	장소	148	병원 구역 건물 체육관 공원 청사 경기장 화장실 운동장 지하철 아파트 회의실 작업장 주차장 사무실 엘리베이터 투표소 교회 식당 장소 승강기 출입문 숙소 단지 청소 지하도 횡단보도 천막 옥상 로비 공연장 목욕탕 출입 강당 출입구 입구 창고 휴게소 빌라 요양원 상가 거주지 모텔 승강장 수영장 통로 버스정류장 주거지 현관 터널 이동식 자택 지하철역 현수막 발코니 동사무소 성당 노숙 차도 생활관 부두 업소 객실 병실 동산 잔디 선로 객석 찜질방 울타리 주택가 주차공간 배부 원룸 천변 복도 집무실 비상구 보도블록 경로당 관리소 벽면 야산 장례 회의장 현금인출기 등산로 객차 텐트 매점 방범 수선 피시 분향소 가로등 베란다 화단 집기 에스컬레이터 카운터 외벽 플래카드 딸린 철로 술집 대피소 작업실 대기실 교도소 팻말 주점 관저 현관문 육교 신호등 불로 옥상정원 비어 방이 사격장 펜스 우수관 배회 묘지 병동 전단 공터 공용 휴지통 야구장 휴게 길거리 공공장소 강의실 난간 재판정 애완견 개폐기 정거장 비닐하우스 이동로 용변 관공서 길가 우편물 펜션 가옥 포대
165	대통령	5	문재인 박근혜 이명박 노무현 김대중
191	복지	80	시각장애인 복지관 청각장애 봉사 나눔 후원 기부 시각장애 봉사활동 공헌 전달 자원봉사 모금 돕기 한마음 기증 어르신 재능기부 사단법인 장애우 재능 수익금 독거 성금 봉사자 천사 안내견 기탁 야학 네팔 헌화 흰지팡이 자선 결연 홀몸 불우 벽화 축구단 배식 후원자 꽃동네 위문 적십자 자원봉사단 일손 맞이 어린이재단 생필품 지구촌 적십자사 보육원 재단법인 영정사진 도배 증서 짜장면 연탄 홀트아동복지회 온정 동화책 장기기증 퇴치 독립유공자 어버이날 보금자리 쾌척 가정방문 협찬 말벗 월드비전 키다리 모교 아침밥 청년회 우리동네 농구단 심기 손수 고아원 온누리상품권