creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조선대학교
CHOSUN UNIVERSITY

[UCI]I804:24011-200000267149

2019년 2월
석사학위 논문

단어의 중요도와 연관성을 고려한 키워드 추천 방법

조선대학교 산업기술융합대학원

소프트웨어융합공학과

김 재 현

조선대학교
CHOSUN UNIVERSITY

2019년 2월 석사학위논문

단어의 중요도와 연관성을 고려한 키워드 추천 방법

김 재 현

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

단어의 중요도와 연관성을 고려한 키워드 추천 방법

Method of Keyword Recommendation Considering Importance and Correlation of Words

2019년 2월 25일

조선대학교 산업기술융합대학원

소프트웨어융합공학과

김 재 현

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

단어의 중요도와 연관성을 고려한 키워드 추천 방법

지도교수 신 주 현

이 논문을 공학석사학위신청 논문으로 제출함

2018년 11월

조선대학교 산업기술융합대학원

소프트웨어융합공학과

김 재 현

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

김재현의 석사학위논문을 인준함

위원장 조선대학교 교수 반 성 범 (인)

위 원 조선대학교 교수 김 판 구 (인)

위 원 조선대학교 교수 신 주 현 (인)

2018년 11월

조선대학교 산업기술융합대학원

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

목 차

ABSTRACT

I. 서론 ····· 1
A. 연구 배경 및 목적 ····· 1
B. 연구 내용 및 구성 ····· 3

II. 관련연구 ····· 4
A. LDA ····· 4
B. Word2Vec 및 Doc2Vec ····· 6
1. Word2Vec ····· 6
2. Doc2Vec ····· 7
C. TF-IDF ····· 9

III. 단어의 중요도와 연관성을 고려한 키워드 추천 방법 ····· 11
A. 시스템 구성도 ····· 11
B. 뉴스 기사 토픽 추출 ····· 13
1. 전처리 과정 ····· 13
2. 토픽 추출 ····· 15
C. 연관키워드 추출 ····· 16
1. 키워드 유사도 ····· 16
2. 중요도 가중치 추출 ····· 19
3. 결합 유사도 ····· 24

IV. 실험 및 결과 ····· 26

V. 결론 ····· 29

참고문헌 ····· 30

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

표 목 차

[표 3-1] 한글 형태소 품사 태그표 예시 ····· 14
[표 3-2] 이산가족 연관 키워드 유사도 ····· 17
[표 3-3] 연관 뉴스 기사 예시 ····· 17
[표 3-4] 제목과 내용 키워드 예시 ····· 22
[표 3-5] 중요도 가중치 결과 ····· 23
[표 3-6] 이산가족 연관 키워드별 유사도 ····· 24
[표 4-1] 실험 환경 ····· 26
[표 4-1] '이산가족' 키워드 연관성 평가 ····· 27

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

그 림 목 차

[그림 1-1] Naver 뉴스 기사의 언론사별(위), 분야별(아래) 목록 ····· 1
[그림 1-2] Naver 급상승 검색어 예시 ····· 2
[그림 2-1] LDA 모델 ····· 4
[그림 2-2] (좌) CBOW 모델 (우) Skop-gram 모델 ····· 6
[그림 2-3] Distribued Memory(DM) 모델 ····· 7
[그림 2-4] Distributed Bag of Words(DBOW) 모델 ····· 8
[그림 3-1] 시스템 구성도 ····· 11
[그림 3-2] 데이터 셋 ····· 13
[그림 3-3] LDA 토픽 모델링기반 토픽 추출 ····· 15
[그림 3-4] Doc2Vec 사전구축 예시 ····· 16
[그림 3-5] TF-IDF 계산 알고리즘 ····· 19
[그림 3-6] TF-IDF를 적용한 연관 키워드 추출 ····· 20
[그림 3-7] 제목과 내용의 키워드 가중치 적용 예 ····· 21
[그림 3-8] 이산가족 연관 키워드 ····· 25
[그림 4-1] 연관성 평가 그래프 ····· 28

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

ABSTRACT

Method of Keyword Recommendation Considering Importance and Correlation of Words

JaeHyun Kim
Advisor : Prof. JuHyun Shin, Ph.D.
Department of Software Convergence Engineering
Graduate School of Industry Technology Convergence, Chosun University

You can share information in real time anytime, anywhere, check for real-time fast-growing search terms through the Internet, and check news articles. As smartphones and the Internet have developed, a variety of service functions are increasing for users. However, if you look at online news articles, they are categorized by press and field, and if you look at Naver associated search terms, you can see only a few search words.

To address these limitations, we collect the title and content of news articles from Naver that correspond to politics by web crawl in Java environments. The collected data is processed in Korean using the KoNLP package, extracted only two or more common nouns, and extracted using LDA Topic Modeling and Doc2Vec to obtain similarity values by keyword.

To compensate for Doc2Vec resulting limits, TF-IDF was used to vector the titles and contents respectively, and the TF-IDF values were weighted to match the key word's equivalent values. In order to take into account the meaning and frequency of the words, the association of keywords was more accurately identified by generating a combined similarity by combining the similarity values of the

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

keywords in Doc2Vec with the weight of importance as presented in this paper.

In order to evaluate the association of keywords, the PMI technique was used to evaluate the similarity between words and words, which is similar to the combined similarity results proposed in this paper, unlike the similarity values by Doc2Vec. This will allow for more relevant keywords to be recommended with keywords of association similarity.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅰ. 서론

A. 연구 배경 및 목적

현재 사용자들은 언제 어디서나 정보를 공유하며 인터넷을 통해 실시간으로 급상승한 뉴스를 확인할 수 있다. 이렇듯 인터넷 환경이 좋아짐으로써 사용자들의 편의성을 고려한 다양한 서비스 기능들이 늘어나고 있다.

인터넷 사이트에 새로 올라온 글들을 사용자가 필요로 하는 정보만 볼 수 있게 되어 있는 RSS(Really Simple Syndication)[1], 사용자가 마음대로 정해 놓은 단어가 다른 사용자에게 유용한 정보로 검색할 수 있는 Tag 기능들이 있다[2]. 하지만 대부분의 사이트에 뉴스 기사들은 날짜와 언론사 그리고 분야별로 카테고리 되어 해당되는 기사들만 볼 수 있을 뿐 연관되어 있는 기사는 찾기 어려운 실정이다.

[그림 1-1]은 네이버에서 제공하는 뉴스 기사의 언론사·분야별 목록이다. 언론사별로 뉴스를 확인할 수 있고 분야별로 내가 원하는 주제에 대한 뉴스를 찾아볼 수 있다.

[그림 1-1] 네이버 뉴스 기사의 언론사별(위), 분야별(아래) 목록

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 1-2]는 네이버에서 제공하는 연관검색어 기능이다. 그림과 같이 '이산가족 상봉'을 검색했을 경우 다음과 같이 몇 가지 연관 검색어들만 볼 수 있다. 따라서 본 논문에서는 네이버에서 제공하고 있는 뉴스 기사를 사용하여 Doc2Vec과 TF-IDF 가중치를 이용해 보다 정확하고 많은 연관 키워드를 추천하는 방법을 제안한다.

[그림 1-2] 네이버 연관검색어 예시

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. 연구 내용 및 구성

본 논문은 국내 대표적인 포털사이트인 네이버에서 제공하는 뉴스 기사를 Java 환경에서 웹 크롤링하여 수집한 후 수집된 문서로부터 Python 기반으로 LDA 모델을 사용하여 뉴스에 대한 토픽을 추출하고 Doc2Vec을 통해 키워드 유사도 값을 구한다. 제목과 내용에 각각 벡터화를 진행해 키워드별 TF-IDF 값을 만든 후 Doc2Vec의 키워드 유사도 값과 비교하여 일치하는 키워드에 중요도 가중치 값을 적용시켜 결합 유사도의 키워드들을 통해 연관 키워드를 추천하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 1장에서는 연구의 배경 및 목적에 대해 서술하고, 2장에서는 LDA와 Doc2Vec, 그리고 TF-IDF에 대하여 기본적인 이론과 관련 논문에 대해 서술하고, 3장에서는 본 논문에서 제안하는 단어의 중요도와 연관성을 고려한 키워드 추천 방법에 대해 서술한다. 4장에서는 제안한 방법의 실험 및 성능평가를 서술한다. 마지막 5장에서는 실험 및 결과에 대해 기술하며 마무리한다.

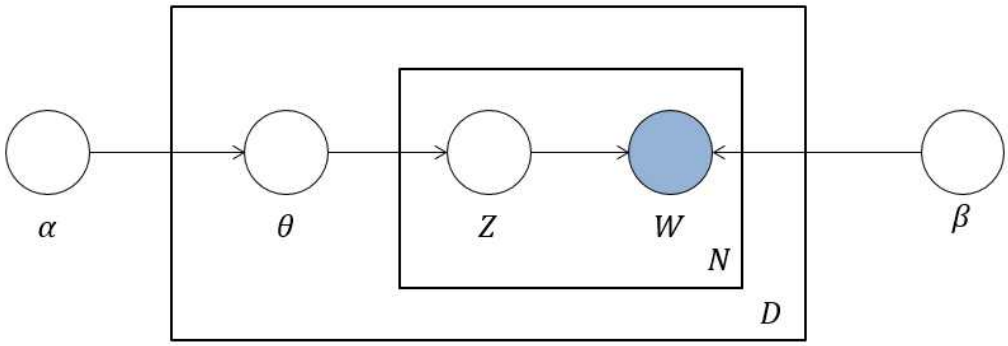
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅱ. 관련연구

A. LDA

잠재 디리클레 할당(LDA)은 Andrew Ng, Michaeall Jordan, David Blei이 2003년 제안한 모델이다. 자연어 처리 단계에서 주어진 문서들에 대해 각각의 문서에 어떠한 주제들이 있는지에 대한 확률모형이다. LDA는 문서별로 토픽 분포, 토픽별로 단어 분포 모두를 추정해 낸다[3].



[그림 2-1] LDA 모델

LDA 모델은 [그림 2-1]과 같이 표현되며 각 기호의 뜻은 다음과 같다.

α : 토픽의 Dirichlet 분포

- 어떤 패턴의 토픽들인지 나타난다.

β : 단어들의 Dirichlet 분포

- 어떤 패턴의 단어들인지 나타난다.

θ : 문서의 토픽어 분포

Z : 주어진 토픽에 대한 단어가 나올 확률

W : 주어진 토픽에 대한 실제 나온 단어

N : 단어의 개수

D : 문서의 개수

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

김은희의 연구에서는 LDA를 기반으로 TV 프로그램 자동 추천을 위해 TV시청 내역 기반으로, 은닉 변수를 시청자들의 주요 관심 토픽으로 설정하고, 이 토픽에 대한 여러 가지 양성을 보기위해 비대칭 디리클레 분포로 실험하여 TV 프로그램 추천 성능을 높인 연구를 하였다[4].

이호경의 연구에서는 LDA 토픽 모델링을 사용하여 주제별로 나뉘어진 클러스터의 상위 단어들을 기반으로 다른 단어들을 분해, 병합해 단어 집합을 확장시켜 SNS(트위터) 데이터를 10 가지 카테고리로 분류하였다[5].

조태민의 연구에서는 LDA를 기반으로 해당 문서와 그 문서와 비슷한 문서들을 구성한 후, 문서들의 키워드를 후보 키워드로 설정하여 후보 키워드를 구성하는 각각의 단어들을 가지고 그들의 중요성을 평가 하였다[6].

심준식의 연구에서는 종합법률정보 사이트의 API를 이용해 판례를 수집하여 LDA 토픽 모델링 기법을 사용해 판례의 단어들에 대해서 확률이 가장 높았던 판례 주제어 태그를 부여한 뒤 유사한 판례를 그룹 별로 분류 하였다[7].

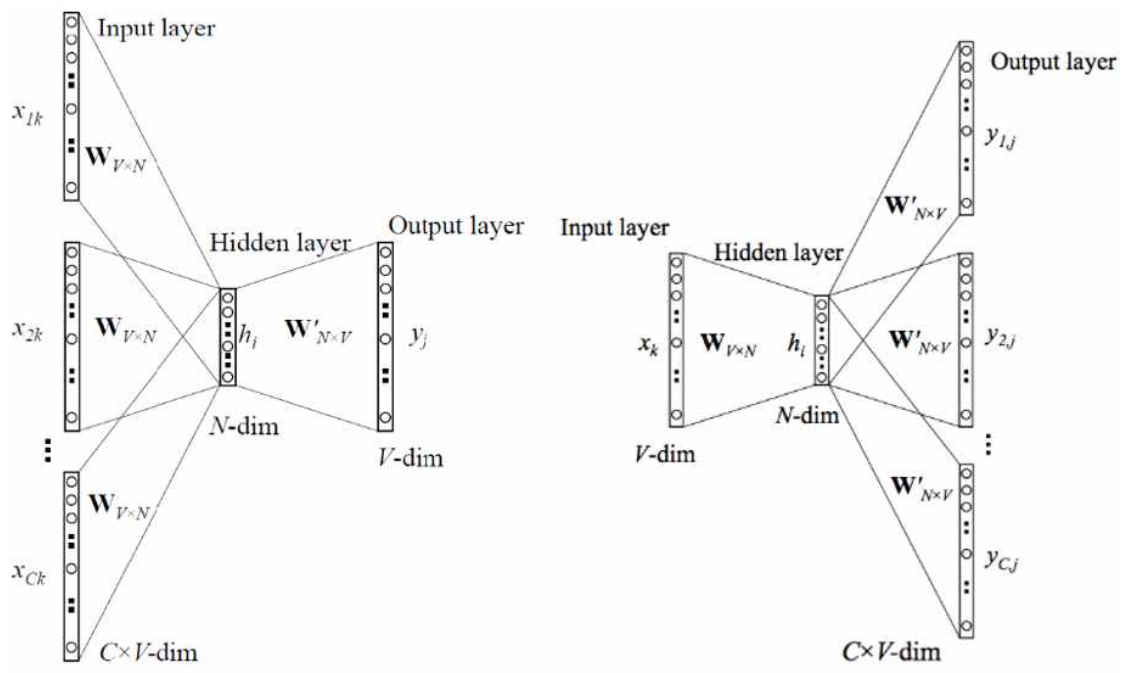
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

B. Word2Vec 및 Doc2Vec

1. Word2Vec

Word2Vec모델은 2003년 Tomas Mikolov이 여러 연구원들이 모여 만든 Continuous Word Embedding 학습 모형으로 쉽게 말해 단어를 벡터로 바꿔 표현해주는 알고리즘이다[8]. Word2Vec 모델은 [그림 2-2]와 같이 CBOW 모델과 Skip-gram 모델이 있다.



[그림 2-2] (좌) CBOW 모델 (우) Skop-gram 모델

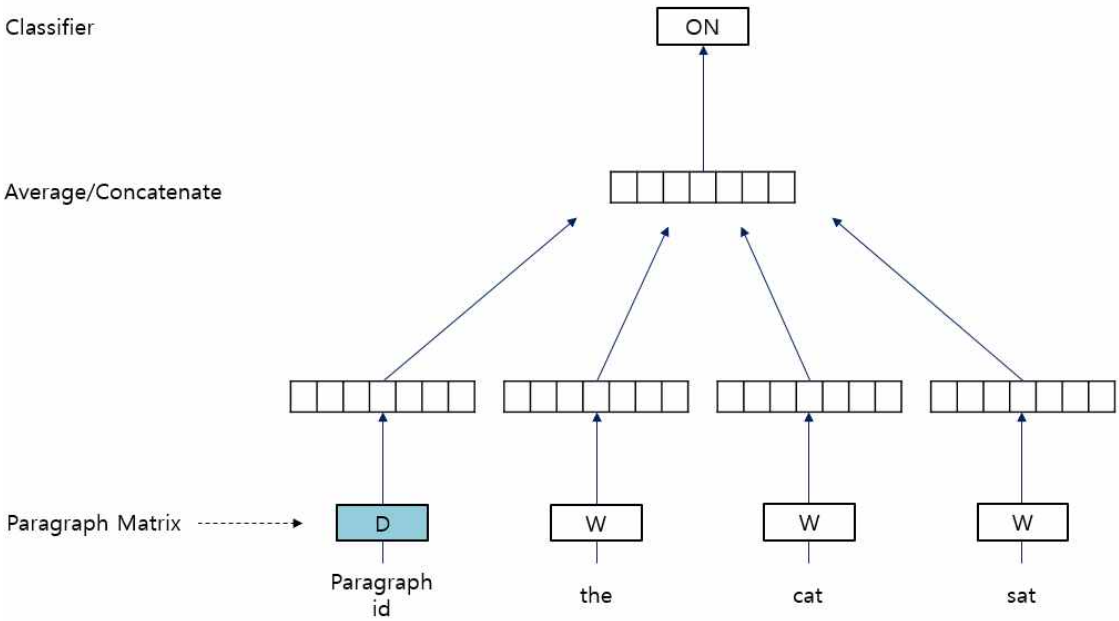
CBOW모델은 텍스트로부터 찾으려고 하는 단어를 예측하는 모델이고 Skop-gram은 CBOW모델과 반대로 단어 하나를 가지고 주변의 단어들을 유추하는 모델이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. Doc2Vec

Doc2Vec모델은 Word2Vec의 확장형으로써 Mikolov가 제안한 모델이다. Word2Vec와 다르게 문장이나 문서에 대하여 vector값으로 나타내는 비지도 학습 방법이다[9]. Doc2Vec 모델도 Word2Vec 모델과 같이 두 가지 알고리즘으로 나뉘는데 [그림 2-3]과 같이 Distribted Memory(DM) 모델과 [그림 2-4]와 같이 Distributed bag of words (DBOW) 모델이 있다.

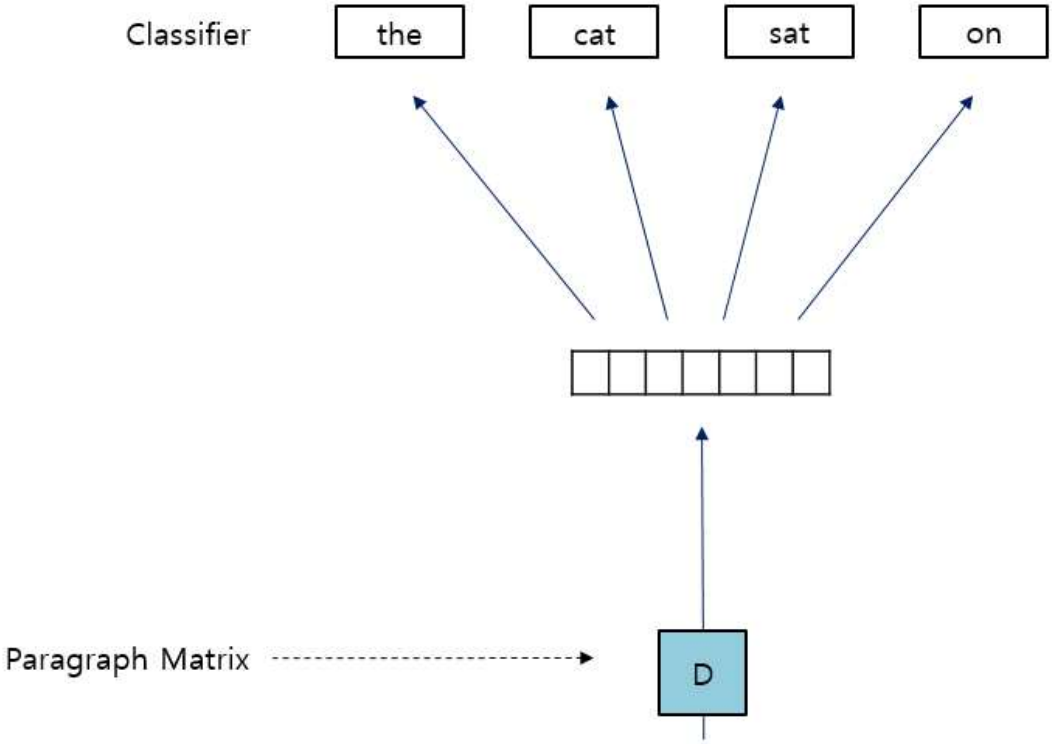


[그림 2-3] Distribued Memory(DM) 모델

Distrebuted Memory(DM) 모델은 단어를 학습시킬 때 각 학습 단계를 vector에 기억하게 하고 최종 학습된 vector를 정의하는 방식이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 2-4] Distributed Bag of Words(DBOW) 모델

Distrebuted Bag of Words(DBOW) 모델은 어떤 문서가 주어졌을 때 그 문서에 포함되어 있는 단어를 예측할 수 있다.

육지희의 연구에서는 LDA와 Doc2Vec을 활용해 단순 통계량이 아닌 의미적 특성을 추출하고 단어의 위치와 자질집합의 크기를 k-NN, Ridge, NB, SVM 분류 알고리즘을 사용하여 성능을 비교 하였다[10].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

C. TF-IDF

TF-IDF는 TF와 IDF를 합친 합성어이며 텍스트 마이닝에서 활용하는 가중치로써, 문서들이 있을 때 해당 단어가 어떠한 문서에서 어느 정도 중요성이 있는지를 나타내는 수치이다[11].

Term Frequency(TF)는 문서에 해당하는 단어가 얼마나 등장하는지 나타낸 값이고, Document Frequency(DF)는 한 문서가 아닌 전체 문서에 대해 해당하는 단어가 나타난 문서의 수이다. Inverse Document Frequency(IDF)는 역문서 빈도로 단어가 문서들 안에서 얼마나 많이 등장하는지 나타낸다.

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \tag{1}$$

식 (1)은 TF-IDF 값을 수식으로 표현한 것이다. d는 문서를 t는 단어를 의미하며 tf(t,d)는 문서 d 안에서 단어 t의 총 빈도수를 의미한다. idf(t,D)는 전체문서(D) 중 단어 t가 포함되어 있는 문서의 수로 나누어 log를 취해 얻는다. idf 수식은 식 (2)와 같이 표현한다.

$$idf(t,D) = \log\frac{|D|}{|d \in D : t \in d|} \tag{2}$$

역문서 빈도에 log를 취하는 이유는 역수를 취해서 값이 무한하게 커질 수 있기 때문에 그것을 방지하기 위해 값을 줄이는 것이다.

유은순의 연구에서는 소설 텍스트 100권의 데이터를 수집한 후 머리말, 비대화문, 대화문, 맺음말 총 4개로 분리하고 TF-IDF 가중치를 계산하여 소설 텍스트에 대한 주제어를 추출하는 방법에 관한 연구를 하였다[12].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

노연우의 연구에서는 시간적 속성을 추가시켜 변형시킨 TF-IDF를 이용해 급상승한 후보 키워드 집합을 추출한 후 Twitter 사용자의 전문성과 영향력을 가중치로 적용하여 토픽의 중요성을 파악하고 이슈 토픽 예측 지수값을 계산하여 미래에 나올 이슈 토픽을 예측 하였다[13].

김정훈의 연구에서는 TF-IDF를 사용하여 농식품과 관련된 SNS의 데이터를 2차원으로 사상하고 이것을 기반으로 SVM 분류자를 구축하는 방법을 제안 하였다[14].

오유리의 연구에서는 용어 매칭과 TF-IDF를 적용하고 Word2Vec을 통해 유의어를 확장시켜 적용한 뒤 특허에 관한 문서의 청구항을 기준으로 유사한 청구항을 선별하는 기법을 비교분석 하였다[15].

이성직의 연구에서는 전체 문서집합의 중요 키워드를 TF-IDF를 변형하여 산출하였다[16]. 하지만 전체 문서집합에 대해서 추출한 키워드는 단일 문서를 요약하거나 탐색하는 등에는 활용할 수 없으므로 문서별로 키워드를 추출하고 이를 활용할 기술이 필요하다.

Collection @ chosun

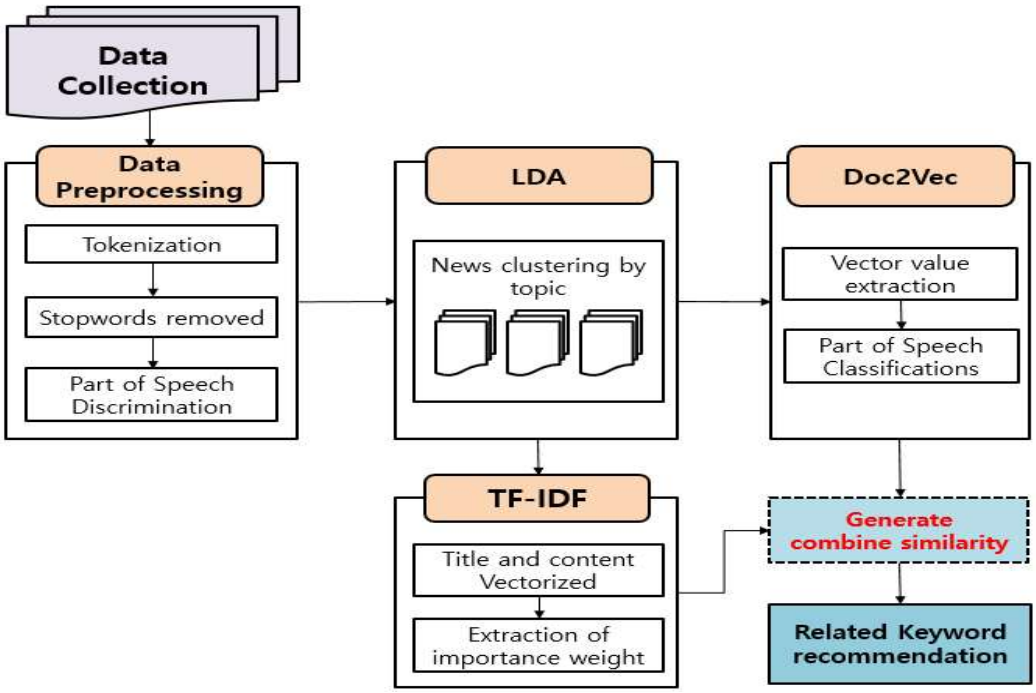
조선대학교
CHOSUN UNIVERSITY

Ⅲ. 단어의 중요도와 연관성을 고려한 키워드 추천 방법

본 논문에서는 연관 키워드 추천을 위해 LDA를 사용해 토픽을 추출하고 Doc2Vec을 통해 키워드 유사도 값을 구한다. 뉴스 기사의 제목과 내용에 각각 벡터화를 진행해 키워드별 중요도 가중치를 만든 후 Doc2Vec의 결과와 비교하여 일치하는 키워드에 중요도 가중치를 적용시켜 결합 유사도의 키워드들을 통해 연관 키워드를 추천하는 방법을 제안한다.

A. 시스템 구성도

[그림 3-1]은 본 논문에서 제안하는 시스템 구성도이다. 시스템 구성도는 크게 LDA 토픽 추출 단계와 Doc2vec 키워드 유사도 값 추출단계, TF-IDF 중요도 가중치 적용 단계, 결합 유사도 생성 단계로 구성된다.



[그림 3-1] 시스템 구성도

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

본 논문에서는 실험을 위해 Java기반 환경에서 네이버 뉴스 기사 중 정치 분야에 해당하는 기사의 제목과 내용에 대해 22,779개 수집하여 실험을 진행하였다.

수집된 데이터는 전처리 단계로 R studio 환경에서 KoNLP 패키지를 통해 한국어 처리를 하고 Corpus를 생성한 후 특수문자, 영어 등 불용어를 제거하였다. 다음 단계로 LDA를 통해 주제별 토픽을 추출하고, Doc2Vec을 사용하여 키워드별 유사도 값을 추출한다.

Doc2Vec의 한계점을 보완하기 위해 뉴스 기사의 '제목과 내용'에 단어별 TF-IDF 값을 추출하였고 내용에서 나온 키워드들은 제목과 일치하는 키워드들만 추출해 빈도수 기반 평균을 구하여 중요도 가중치 값을 구하였다. 이후 Doc2Vec과 비교하여 일치하는 키워드에 중요도 가중치를 적용시키고 단순 빈도수가 아닌 단어의 의미와 빈도수를 고려하기 위해 Doc2Vec의 키워드별 유사도 값과 중요도 가중치를 결합해 키워드별 결합 유사도를 생성해 보다 정확한 연관 키워드를 추천하는 방법을 제안한다.

Collection @ chosun

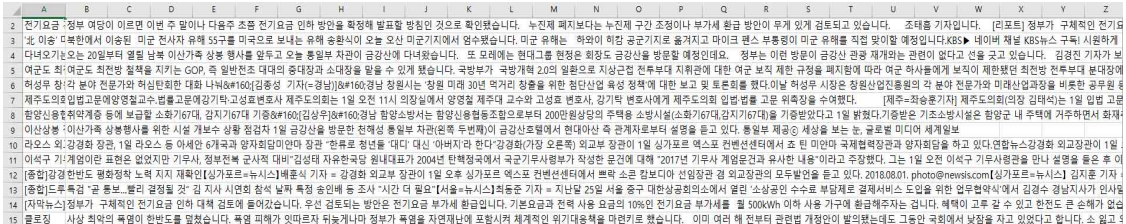
조선대학교
CHOSUN UNIVERSITY

B. 뉴스 기사 토픽 추출

본 절에서는 네이버 뉴스 기사를 수집하고 전처리한 후 LDA를 사용하여 토픽을 추출하는 과정에 대해 서술한다.

1. 전처리 과정

본 논문에서 사용된 데이터 셋은 네이버에서 제공하는 뉴스 기사 중 2018.07.19.부터 2018.08.19.까지 정치 분야에 해당하는 기사의 제목과 내용을 총 22,779개를 Java환경에서 웹 크롤링하여 실험을 진행하였다. 수집된 문서들은 R studio 환경에서 KoNLP 패키지를 사용하여 한국어 처리를 진행하고 품사를 판별해 두 글자 이상의 보통 명사들만 추출한 후 Corpus를 생성해 특수문자, 영어, 완전한 글자가 아닌 단어, 뉴스, 기자 등을 불 용어로 포함시켜 처리하였다. 다음 [그림 3-2]는 본 논문에서 사용한 데이터 셋 이다.



[그림 3-2] 데이터 셋

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-1] 한글 형태소 품사 태그표 예시

Tag	name	description	example
NNG	보통 명사	일반적인 사물의 이름	나무, 하늘, 희망
NNP	고유 명사	특정한 사람이나 사물의 이름	무등산, 한강, 신라
NNB	의존 명사	스스로 뜻을 지니고 있어 다른 말의 도움 없이 쓰이는 명사	따름, 나름, 뿐
NR	수사	사물의 수량이나 순서	하나, 둘, 첫째, 둘째
NP	대명사	인칭 대명사 : 사람의 이름을 대신 가르킴 지시 대명사 : 사람이외의 것을 대신 가르킴	귀하, 어르신, 이것
VV	동사	동작이나 과정	자다, 먹다, 가다
VA	형용사	사물의 모습이나 상태	둥글다, 빨갛다, 같다
IC	감탄사	놀람, 감동등을 나타냄	파이팅, 앗, 호호
MD	관형사	체언앞에 놓여서 그 체언의 내용을 꾸밈	**저** 어린이, **한** 사람, **순** 살코기
MAG	일반 부사	용언 또는 다른 말 앞에 놓여 그 뜻을 분명하게 함	-
MAC	접속 부사	문장 수식의 접속사와 단어나 구 등을 이어주는 접속사	**그**리고, **그**러면, **그**러나
:			

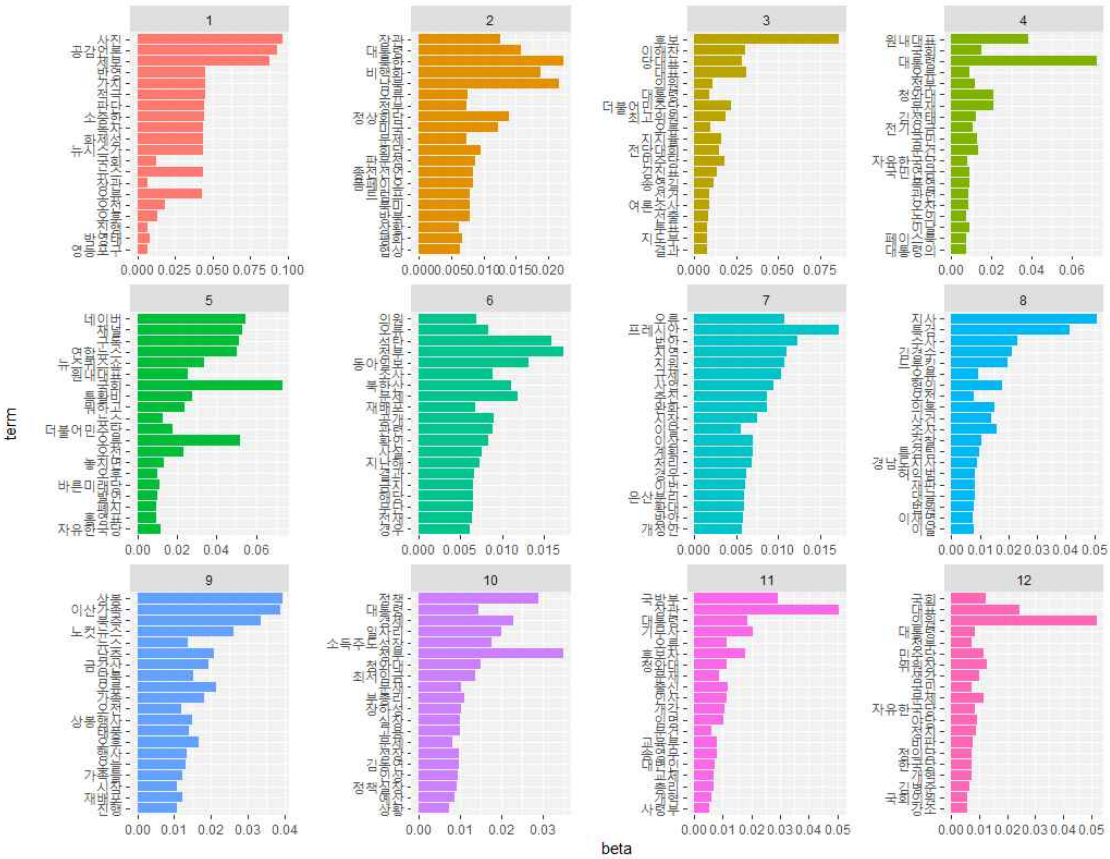
Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. 토픽 추출

전처리 과정을 거친 데이터는 LDA 토픽 모델링을 통해 토픽을 추출한다. LDA의 클러스터 수는 하이퍼 파라미터로서 여러 실험을 통해 최적의 클러스터를 구해야한다.

본 논문에서는 클러스터의 수를 12개로 제한했을 때 가장 성능이 좋았으며 [그림 3-3]은 LDA 토픽 모델링을 통해 토픽을 추출한 결과이다.



[그림 3-3] LDA 토픽 모델링기반 토픽 추출

정치 분야에 해당하는 뉴스 기사 22,779개를 클러스터링 한 결과 12개의 클러스터를 생성했으며 각 클러스터에는 서로 연관된 뉴스 토픽끼리 묶여있는 것을 볼 수 있다. 예를 들면 9번째 토픽은 이산가족, 상봉, 금강산 등, 이산가족에 관련된 단어들로 되어 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

C. 연관키워드 추출

본 절에서는 Doc2Vec을 키워드 유사도 값을 구하고, TF-IDF를 적용하여 연관 키워드를 추출하는 방법에 대해 서술한다.

1. 키워드 유사도

Python API인 Gensim패키지를 이용하여 Doc2Vec으로 토픽의 벡터 값을 추출한 후 연관 키워드 유사도를 추출한다. Doc2Vec에서는 띄어쓰기 단위도 하나의 단어로 학습되기 때문에 같은 단어 일지라도 다양한 뜻을 나타낸다. 22,779개의 데이터 중 20,000개의 train data와 10,000개의 test data로 실험을 진행하였다.

[그림 3-4]는 Doc2vec을 실험하기 위한 사전구축 예시이다.

```
tagged_train_docs=[TaggedDocument(d,[c])for d,c in train_docs]
tagged_test_docs=[TaggedDocument(d,[c])for d,c in test_docs]

from gensim.models import doc2vec
doc_vectorizer=doc2vec.Doc2Vec(vector_size=300,window=10, min_count=5, workers=11,
                               alpha=0.025, min_alpha=0.025, epochs=20)
doc_vectorizer.build_vocab(tagged_train_docs)
doc_vectorizer.train(tagged_train_docs, epochs=doc_vectorizer.epochs,
                     total_examples=doc_vectorizer.corpus_count)

pprint(doc_vectorizer.most_similar('이산가족/Noun'))
```

[그림 3-4] Doc2Vec 사전구축 예시

다음 [표 3-2]는 사전구축된 Doc2Vec을 토대로 '이산가족'이라는 토픽과 관련된 상위 20개의 연관 키워드와 유사도를 출력한 결과이다. [그림 3-4]에서 'most_similar' 함수를 통해 출력할 수 있고, 이 함수는 벡터들 간의 코사인 유사도를 측정하며 값이 높을수록 키워드와 유사하다. 키워드와 단어 간의 유사도를 나타내고 있으며, '작별', '금강산', '개별' 등의 키워드를 확인할 수 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-2] 이산가족 연관 키워드 유사도

No.	연관 키워드	Doc2Vec	No.	연관 키워드	Doc2Vec
1	이산	0.643995	11	눈물	0.545792
2	작별	0.621491	12	만남	0.534576
3	금강산	0.589559	13	점심	0.527799
4	개별	0.584738	14	어머니	0.527620
5	이산가족상봉	0.578467	15	설레다	0.516723
6	이별	0.559387	16	오붓하다	0.506205
7	가족	0.558603	17	식사	0.504146
8	65년	0.556476	18	형제	0.502905
9	68년	0.555606	19	상봉	0.470939
10	오열	0.548075	20	재회	0.466823

[표 3-3] 연관 뉴스 기사 예시

관심 뉴스 기사	이번 **이산가족 상봉** 행사는 오는 20일부터 26일까지 금강산관광지구에서 진행된다. 첫날에는 금강산호텔에서 단체상봉과 북측 주최 환영만찬이 진행된다. 둘째 날 오전에 객실에서 개별상봉을 진행한 다음 객실에서 오붓하게 오찬까지 이어간다. 셋째 날의 경우에도 예전에는 작별상봉을 마친 다음 남북이 오찬을 따로 진행했으나, 이번 행사에서는 작별상봉과 공동오찬을 묶어서 진행하게 된다[17].
연관 뉴스 기사1	2년 10개월 만에 재개되는 **이산가족 상봉**행사를 하루 앞두고 남측 가족들이 오늘 강원도 속초에 모였습니다. 강원도 고성의 동해선 남북출입사무소를 거쳐 북측 통행검사소에서 심사를 받고, 낮 12시 30분쯤 **금강산** 온정각에 도착할 것으로 보입니다. 특히 둘째 날인 21일에는 2시간의 **개별상봉**에 이어 1시간 동안 객실에서 함께 점심을 먹게 될 텐데요[18].
연관 뉴스 기사2	문 대통령은 다음 달로 합의한 김정은 북한 국무위원장과의 평양 남북 정상회담을 통해 분단의 원인이 된 **65년** 간의 전쟁 체제를 종식하는 종전선언을 연내 이루겠다는 뜻을 분명히 밝혔다. 오는 평양 남북 정상회담에서 기존의 4·27판문점 남북 정상회담에서 합의한 비핵화와 종전선언을 구체화시키겠다는 뜻을 공개적으로 밝힌 것이다[19].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

다음 [표 3-3]는 [표 3-2]의 유사도를 비교하기 위해 관심 뉴스 기사와 연관 뉴스 기사를 보여주고 있다. 관심 뉴스 기사와 연관 뉴스 기사1을 보면 공통적으로 들어간 '이산가족 상봉', '금강산', '개별'이라는 연관된 키워드들이 있었으며, 그 외 공통적이진 않지만 '오붓하게', '작별', '점심' 의 키워드를 확인할 수 있었다. 이로써 공통적으로 들어있지는 않은 키워드임에도 연관된 기사임을 알 수 있다.

하지만 관심 뉴스 기사와 연관 뉴스 기사2를 비교하였을 경우 공통적으로 '남북'이라는 키워드는 있지만 관심 뉴스 기사의 경우 이산가족에 관한 뉴스 기사이고 연관 뉴스 기사2의 경우는 비핵화에 관한 뉴스 기사이다. 이것을 토대로 Doc2Vec에서는 연관이 없는 기사임에도 단어가 아닌 문장, 문서간의 유사도를 측정하여 공통적인 단어가 있다는 이유로 높은 유사도가 나타난 것을 볼 수 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

2. 중요도 가중치 추출

본 절에서는 관련성이 낮은 키워드가 추천되는 문제점을 보완하기 위해 전체기사에 대한 TF-IDF 가중치를 적용한 실험과 본 논문에서 제안하는 제목과 내용에 대한 TF-IDF를 적용한 실험을 진행한다.

먼저, TF-IDF 가중치를 적용한 실험은 다음 [그림 3-5]와 같이 CountVectorizer를 사용하여 DTM을 구하고 TfidfVectorizer를 사용해 TF-IDF를 계산한다.

```
tf = CountVectorizer()

tf.fit_transform(documents)
tf.fit_transform(documents)[0:1].toarray()

tfidf = TfidfVectorizer(max_features = 100, max_df=0.95, min_df=0)

A_tfidf_sp = tfidf.fit_transform(documents)

tfidf_dict = tfidf.get_feature_names()
```

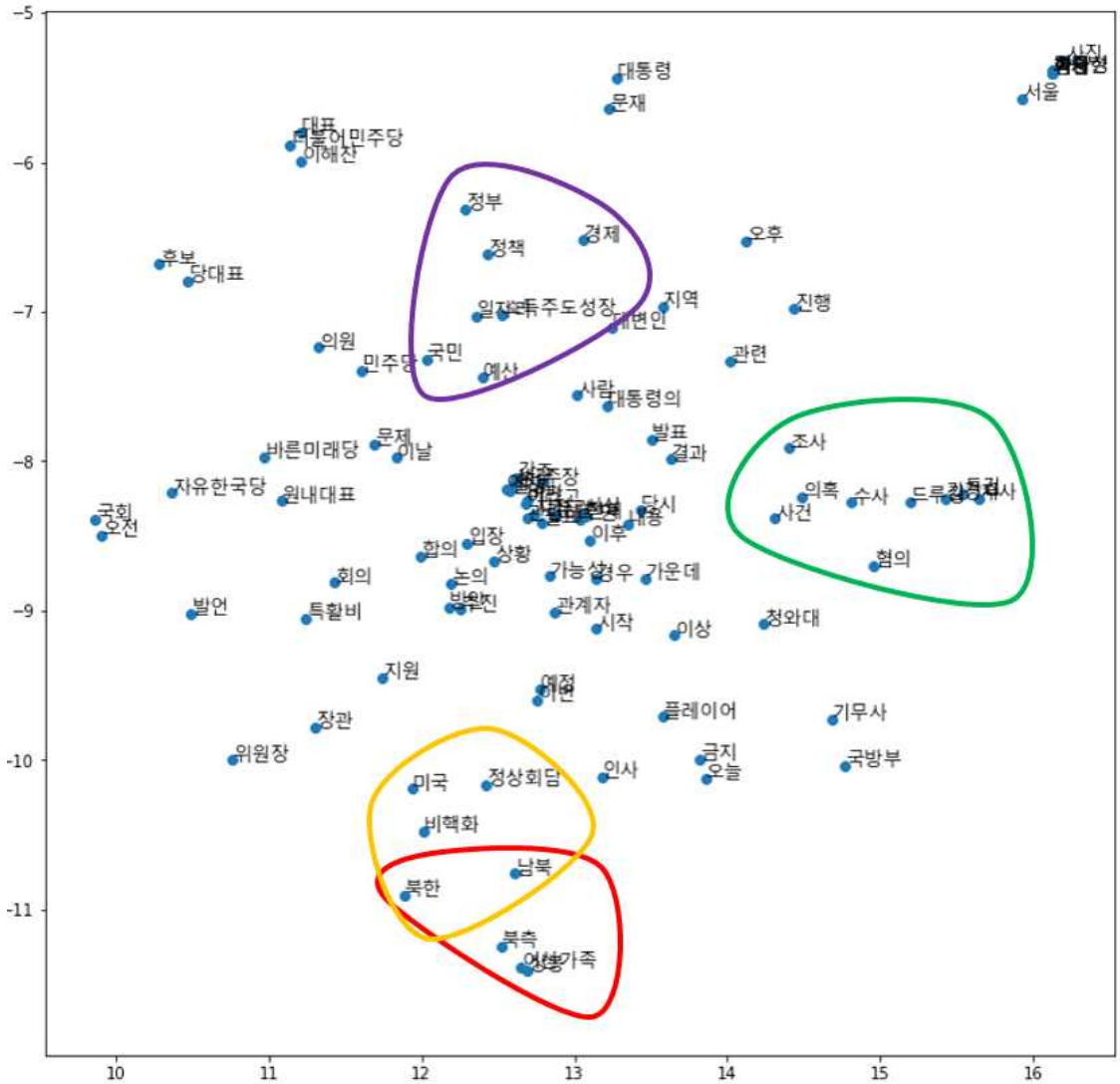
[그림 3-5] TF-IDF 계산 알고리즘

TF-IDF 계산 알고리즘을 통해 중요 단어를 산출하였고, 상위 100개의 키워드를 가지고 고차원상에 있는 데이터의 유사성들을 저차원으로 임베딩 시키는 방법인 t-SNE로 시각화 하였다[20].

[그림 3-6]은 100차원에 있던 22,779개의 모든 기사를 2차원에 표현하였고 그 중 상위 100개의 키워드만 출력한 결과이다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 3-6] TF-IDF를 적용한 연관 키워드 추출

추출된 키워드들을 확인한 결과 크게 4개의 그룹으로 ‘드루킹’, ‘이산가족’, ‘비핵화’, ‘일자리’에 관련된 키워드들이 연관성 있는 것을 알 수 있다. ‘비핵화’와 ‘이산가족’ 그룹에는 ‘남북’과 ‘북한’이라는 공통된 키워드를 확인할 수 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

본 논문에서는 TF-IDF 결과 4개의 그룹 중 '이산가족'에 관한 뉴스 기사를 분류하여 실험을 진행한다. 이산가족에 관한 뉴스 기사 총 1210개의 문서를 분류해 '제목과 내용'에 각각 벡터화를 진행하고 이를 결합해 키워드별 중요도 가중치 값을 추출한다. 식 (3)은 중요도 가중치를 적용하는 식을 나타낸다.

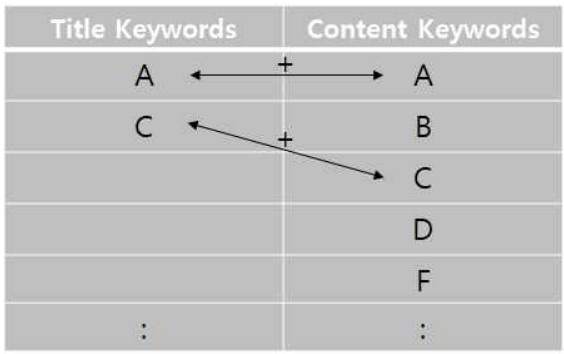
$$V(t,d,D) = tf(t_1,s_1) \times idf(t_1,S_1) + tf(t_2,c_2) \times idf(t_2,C_2) \tag{3}$$

· $V(t,d,D)$: 제목과 내용의 가중치

· $tf(t_1,s_1) \times idf(t_1,S_1)$: 특정제목(s_1)내의 단어(t_1)의 총 빈도수 × 전체제목(S_1) 중 단어(t_1)이 들어간 빈도수

· $tf(t_2,c_2) \times idf(t_2,C_2)$: 특정내용(c_2)내의 단어(t_2)의 총 빈도수 × 전체내용(C_2) 중 단어(t_2)이 들어간 빈도수

이산가족 기사 1210개에 대해 제목과 내용의 TF-IDF를 각각 계산한 결과 제목의 TF-IDF값이 더 높게 나타났다. 뉴스 기사의 제목은 내용의 핵심을 나타내고 문서를 대표하는 키워드들이 분포 되어있기 때문이다.



[그림 3-7] 제목과 내용의 키워드 가중치 적용 예

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-7]에서와 같이 제목의 키워드가 A, C가 있고 내용의 키워드가 A~F라고 가정 했을 때 일치하는 A, C 키워드에 대해서만 중요도 가중치를 적용하였다.

[표 3-4] 제목과 내용 키워드 예시

No.	제목 키워드	TF-IDF	No.	내용 키워드	TF-IDF
1	선물	0.899776	1	아버지	0.26336
2	아들	0.843453	2	할머니	0.218488
3	기다림	0.839531	3	가족	0.216677
4	정상회담	0.781644	4	눈물	0.208386
5	이산가족상봉	0.765738	5	개별상봉	0.201079
6	눈물	0.751263	6	단체상봉	0.17964
7	작별	0.686951	7	이산가족상봉	0.177541
8	남북	0.677432	8	북한	0.172477
9	작별상봉	0.673981	9	어머니	0.157506
10	가족	0.654491	10	금강산	0.143367

[표 3-4]는 제목과 내용 상위 10개의 키워드 예시이다. 내용에서 나온 키워드들은 제목과 동시에 출현하는 키워드에 대해 값을 구하였다. 이렇게 구한 TF-IDF 값은 Doc2Vec의 키워드별 유사도 값과 비교하여 일치하는 키워드에 중요도 가중치 값을 적용하였다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[표 3-5] 중요도 가중치 결과

No.	연관 키워드	중요도 가중치	No.	연관 키워드	중요도 가중치
1	가족	0.216677	11	이별	0.193808
2	만남	0.213922	12	이산	0.177541
3	눈물	0.208386	13	이산가족상봉	0.177541
4	상봉	0.201575	14	작별	0.176505
5	개별	0.201079	15	식사	0.174134
6	오열	0.198652	16	어머니	0.157506
7	설레다	0.198652	17	금강산	0.143367
8	오붓하다	0.198652	18	형제	0.094733
9	점심	0.195899	19	68년	0.028064
10	재회	0.195899	20	65년	0.011805

[표 3-5]는 제목과 내용을 결합시킨 결과의 키워드별 중요도 가중치 값을 Doc2Vec의 유사 키워드를 기준으로 Doc2Vec에서 나온 키워드들과 일치하는 값들만 추출하여 적용한 결과이다.

결과를 보면 Doc2Vec에서 이산가족과 연관성이 높았던 키워드인 '65년'과 '68'년은 일치하는 키워드는 있지만 연관성이 낮은 것을 볼 수 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

3. 결합 유사도

단순 빈도수 기반이 아닌 의미와 빈도수를 결합하기 위해 앞에서 추출된 중요도 가중치 값과 Doc2Vec의 키워드별 유사도 값을 적용시켜 결합 유사도를 만든 결과는 [표 3-6]과 같다.

[표 3-6] 이산가족 연관 키워드별 유사도

연관 키워드	Doc2Vec	중요도 가중치	결합 유사도
이산	0.643995	0.177541	0.821536
작별	0.621491	0.176505	0.797996
개별	0.584738	0.201079	0.785817
이산가족상봉	0.578467	0.177541	0.756008
눈물	0.545792	0.208386	0.754178
이별	0.559387	0.193808	0.753195
만남	0.534576	0.213922	0.748498
금강산	0.589559	0.143367	0.732926
점심	0.527799	0.195899	0.723698
:	:	:	:
68년	0.555606	0.028064	0.58367
65년	0.556476	0.011805	0.568281

의미기반의 Doc2Vec의 키워드별 유사도 값과 본 논문에서 제안하는 중요도 가중치 값을 합하여 결합 유사도를 구한 결과 [표 3-2]에서 제시했던 이산가족 연관 키워드 중 '65년'과 '68년'의 키워드들은 유사도가 현저히 낮아져 연관성이 떨어짐을 증명하였다. 연관성이 높지만 연관성이 낮게 나왔던 키워드들이 결합 유사도를 통해 높은 유사도 값을 보여 보다 정확한 키워드를 추천할 수 있게 된다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[그림 3-8] 이산가족 연관 키워드

[그림 3-8]은 [표 3-6]을 한눈에 비교하기 위해 그래프로 표현한 것이다. 본 논문에서 제안한 중요도 가중치를 적용한 결과 '이산가족'에 연관된 키워드들이 뚜렷하게 높은 유사도를 보이는 것을 알 수 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

Ⅳ. 실험 및 결과

본 장에서는 앞서 제안했던 중요도 가중치를 적용한 결합 유사도를 통해 추출한 키워드들에 대해 성능 평가를 수행한다.

[표 4-1]은 본 논문의 실험을 진행한 실험 환경이다.

[표 4-1] 실험 환경

Division		Contents
HW	CPU	Intel(R) Core(TM) i5-6500 CPU @ 3.20GHz
	GPU	Intel(R) HD Graphics 530
	RAM	8.00GB
	HDD	237GB
SW	OS	Windows10
	Program	Java, Python, R studio

연관성 평가를 위해 본 논문에서는 두 단어 A와 B 사이의 연관성을 확률론을 적용시켜 수치화하는 Pointwise Mutal Information(PMI) 기법을 이용한다. 다음 식 (4)는 각각의 발생 확률 $P(A)$와 $P(B)$를 갖는 두 단어 A, B에 대한 $PMI(A,B)$의 수식이다.

$$PMI(A,B) = \log\frac{P(A,B)}{P(A)P(B)} \tag{4}$$

이 PMI 기법은 두 단어가 같이 등장하는 정도를 독립적으로 나올 확률로 나눈 것으로 A단어와 B단어가 어느 정도의 연관성을 나타내는지 확률적으로 계산하는 과정이다 [21].

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

PMI 기법을 사용해 '이산가족'과 연관 키워드들의 관계를 계산한 결과 다음 [표 4-2]와 같다.

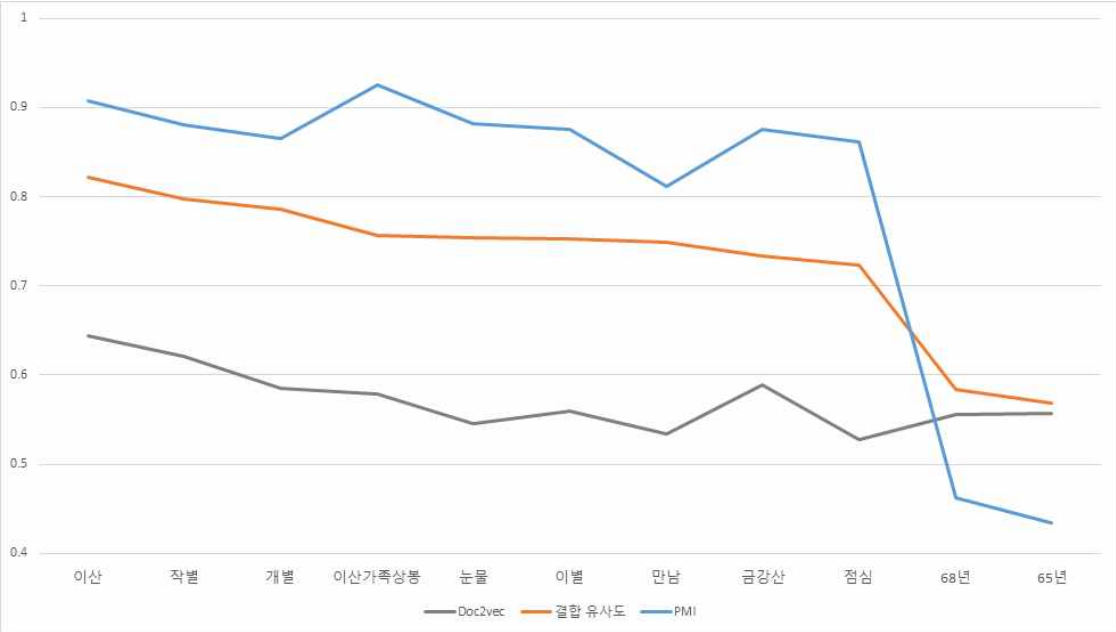
[표 4-2] '이산가족' 키워드 연관성 평가

연관 키워드	Doc2Vec	결합 유사도	PMI
이산	0.643995	0.821536	0.908
작별	0.621491	0.797996	0.881
개별	0.584738	0.785817	0.865
이산가족상봉	0.578467	0.756008	0.926
눈물	0.545792	0.754178	0.882
이별	0.559387	0.753195	0.875
만남	0.534576	0.748498	0.811
금강산	0.589559	0.732926	0.876
점심	0.527799	0.723698	0.862
:	:	:	:
68년	0.555606	0.58367	0.462
65년	0.556476	0.568281	0.434

연관성 평가 결과 '이산'이라는 키워드를 포함해 점심까지의 키워드들을 보면 백분율로 환산할 경우 85%가 넘는 유사도를 보였다. 하지만 Doc2Vec에서 연관성이 높았던 '65년'과 '68년'의 결과는 본 논문에서 제안한 중요도 가중치를 적용한 결합 유사도와 비슷하게 유사도가 현저히 떨어지는 것을 확인했으며 결합 유사도의 키워드들을 통해 보다 정확한 연관 키워드 추천이 가능해 진다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY



[그림 4-1] 연관성 평가 그래프

다음 [그림 4-1]은 [표4-2]의 결과를 나타낸 연관성 평가 그래프이다. PMI 기법으로 성능평가 한 결과 본 논문에서 제안하였던 중요도 가중치를 적용한 결합 유사도와 유사한 키워드들의 연관성을 볼 수 있다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

V. 결론

본 논문에서는 단어의 중요도와 연관성을 고려한 키워드 추천을 위해 중요도 가중치 값을 적용한 결합 유사도를 생성하여 연관 키워드를 추천하는 방법을 제안 하였다. 네이버 뉴스 기사 중 정치 분야에 대하여 기사의 제목과 내용을 총 22,779개를 수집하여 실험 데이터로 사용하였다.

엑셀파일로 수집한 뉴스 기사는 Java환경에서 웹 크롤링하여 R studio 환경에서 KoNLP 패키지를 사용하여 한국어 처리를 진행하고 전처리 단계를 진행하였다. 전처리 단계를 거친 데이터는 LDA 토픽 모델링을 이용해 토픽을 추출하였다.

추출된 토픽은 Doc2Vec을 통해 키워드 유사도 값을 구한다. Doc2Vec에서 '이산가족'에 대한 유사 키워드 중 '65년'과 '68년'의 연관성이 없음을 증명하기 위하여 TF-IDF를 적용해 전체기사에 대한 단어를 시각화하였다. 그리고 이산가족에 관한 기사 1210개를 따로 분류하여 제목과 내용에 단어별 TF-IDF 값을 추출하였고 제목과 내용에 동시에 출현하는 키워드를 추출해 중요도 가중치 값을 구하였다. 이후 Doc2Vec에서 추출된 연관 키워드를 기준으로 일치하는 키워드에 중요도 가중치를 적용시켜 연관성이 없는 키워드를 증명해 보였다.

단순 빈도수가 아닌 단어의 의미와 빈도수를 고려하기 위해 중요도 가중치와 의미기반의 Doc2Vec의 키워드별 유사도 값을 결합해 결합 유사도를 생성하여 연관성이 높지만 연관성이 낮게 나왔던 키워드들을 추출해 보다 정확한 연관 키워드 추천이 가능해진다.

키워드의 연관성을 평가하기 위해 PMI기법을 이용해 단어와 단어 간의 유사성을 평가한 결과 Doc2Vec의 키워드별 유사도 값과 달리 본 논문에서 제안한 중요도 가중치를 적용한 결합 유사도의 키워드 결과와 같이 '65년'과 '68년'의 연관성은 현저히 떨어지고, 관련 있는 키워드의 연관성은 높아지는 것을 확인하였다.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

참고문헌

[1] D. Ayers and A. Watt, Beginning Rss And Atom Programming, John Wiley & Sons Inc., 2005

[2] 이강표, 김두남, 김형주. “웹 2.0 환경에서의 태깅기술 동향”, 정보과학회지, Vol.25 No.10, pp. 36-42, 2007.

[3] D. Blei, A. Y. Ng, and M. Jordan. “Latent Dirichlet allocation”, Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[4] 김은희, 표신지, 김문철. “LDA 기반 은닉 토픽 추론을 이용한 TV 프로그램 자동 추천”, 방송공학회논문지, Vol.17, No.2, pp.270-283, 2012.

[5] 이호경, 양선, 고영중. “비격식 문서 분류 성능 개선을 위한 LDA 단어 분포 기반의 자질 확장”, 한국정보과학회 정보과학회논문지, Vol.43, No.9, pp.1008-1014, 2016.

[6] 조태민, 이지형. “LDA 모델을 이용한 잠재 키워드 추출”, 한국지능시스템학회 논문지, Vol.25, No.2, pp.180-185, 2015.

[7] 심준식, 깅형중. “LDA 토픽 모델링을 활용한 판례 검색 및 분류 방법”, 전자공학회 논문지, Vol.54, No.9, pp.67-75, 2017.

[8] T. Mikolov, k. chen, G. Corrado and J. Dean. “Efficient Estimation of Word Representations in Vector Space”, arXiv preprint, arXiv:1301.3781, 2013.

[9] Lau, Baldwin, An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, arXiv.org, 2016.

[10] S. Robertson, “Understanding Inverse Document Frequency: On theroretical arguments for IDF”, Journal of Documentation, Vol.60, No.5, pp.503-520, 2004.

[11] 육지희, 송민. “토픽모델링과 딥 러닝을 활용한 생의학 문헌 자동 분류 기법 연구”, 정보관리학회지, Vol.35, No.2, pp.63-88, 2018.

[12] 유은순, 최건희, 김승훈. “TF-IDF와 소설 텍스트의 구조를 이용한 주제어 추출 연구”, 한국컴퓨터정보학회논문지, Vol.20, No.2, pp.121-129, 2015.

[13] 노연우, 임종태, 복경수, 유재수. “소셜 네트워크 환경에서 변형된 TF-IDF를 이용한 핫 토픽 예측 기법”, 정보과학회 컴퓨팅의 실제 논문지, Vol.23, No.4, pp.217-225, 2017.

Collection @ chosun

조선대학교
CHOSUN UNIVERSITY

[14] 김정훈, 르하그바도르츠 바툴가, 나스리다노프 아지즈. “농식품 SNS 데이터 감성 분석을 위한 TF-IDF 가중치 기반 분류자 생성”, 한국정보과학회 학술발표논문집, pp.283-285, 2018.

[15] 오유리, 민재옥, 김용일, 김대중, 박용균, 이봉건. “워드임베딩 기반 TF-IDF를 이용한 특허 문서의 유사 청구항 도출 기법 비교 분석”, 한국정보과학회 학술발표논문집, pp.1001-1003, 2017.

[16] 이성직, 김한준. “TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법”, 한국전자거래학회지, Vol.14, No.4, pp.59-73, 2009.

[17] “형제·자녀 상봉 닷새 앞둔 이산가족들 ‘살아있어 고맙다’”, NEWSIS, 2018년 8월 15일 수정, 2018년 9월1일 접속, http://www.newsis.com/view/?id=NISX20180814_00003910928&cID=103018&pID=10300

[18] “‘이산 상봉’ 하루 앞으로 …남측 가족들 속초 집결”, KBS NEWS, 2018년 8월 19일 수정, 2018년 9월 1일 접속, http://news.kbs.co.kr/news/view.do?ncd=40182418&ref=A

[19] “북미 비핵화 교착 국면 돌파 의지 …文대통령‘한반도 운전자’재천명”, NEWSIS, 2018년 8월 15일, 2018년 9월 1일 접속, http://www.newsis.com/view/?id=NISX20180815_0000391295&cID=10301&pID=10300

[20] Andreas Muller, Sarah Guido(2017), Introduction to Machine Learning with Python, O‘REILLY,

[21] Turney and M. Littman."Measuring praise and criticism: Inference of semantic orientation from association", Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, pp. 417-424, July 2002.

Collection @ chosun