워드 임베딩 모델을 이용한 문서 이해 및 유사문서 추천

조정민[1], 강승식[2]
[1]국민대학교 소프트웨어학부 학부생
[2]국민대학교 인공지능학부 교수
min27999@kookmin.ac.kr, sskang@kookmin.ac.kr

Document Understanding and Similar Document Recommendation Through Word Embedding Model

Jeongmin Cho[1], Seungshik Kang[2]
[1]Dept. of Computer Science, Kookmin University
[2]Dept. of Artificial Intelligence, Kookmin University

요 약

문서의 내용을 쉽게 이해하기 위해서는 문서의 핵심 단어, 또는 핵심 문장을 빠르게 파악하는 것이 중요하다. 또한 유사한 문서를 참고하여 같이 읽는다면 해당 문서 내용을 파악하는 시간을 단축시켜주거나 해당 문서에 대한 이해도를 증가시킬 수 있다. 이를 위해서 wordcloud, textrank, Doc2Vec, softmax regression, cosine similarity과 같은 기법을 활용한다. 최종적으로 어떠한 문서를 입력받으면 문서의 명사를 기반으로 한 워드클라우드 시각화 및 핵심 문장 추출, 같은 카테고리를 가지는 유사한 문서를 추천해주는 연구를 수행하였다.

1. 서론

정보가 넘쳐나는 현대 사회에서 중요한 문제 중 하나는 많은 문서를 빠르고 쉽게 이해하는 것이다. 문서의 핵심 내용을 신속하게 파악하는 것은 학생, 연구자뿐만 아니라 모든 사람들에게 매우 중요한 능력으로, 개개인이 받아들이는 정보의 질을 결정짓고, 지식을 쌓는 데에 있어 시간을 절약할 수 있게 해준다. 이러한 상황을 바탕으로 본 프로젝트는 문서의 이해도를 높이는 데에 초점을 맞추고 있다. 특히, 문서의 핵심 단어나 문장을 빠르게 인식하고 이해하는 기술은 문서의 전체적인 맥락을 빠르게 파악하는 데 도움을 줄 수 있다. 또한 유사한 문서를 함께 참고하면서 읽으면 더욱 풍부한 정보와 배경 지식을 얻을 수 있으며, 이는 해당 문서에 대한 이해도를 한층 더 깊게 만들어준다. 이러한 과정을 사람이 아닌 컴퓨터가 대신 해준다면, 사람의 피로도가 감소하며 이는 곧 생산성의 증가로 다가갈 수 있다.

이 연구에서는 wordcloud, textrank, Doc2Vec와 같은 기술을 이용한다. 먼저 wordcloud를 이용하여 문서의 핵심 단어를 시각화하여 문서의 주요 단어를 한눈에 파악할 수 있게 한다. 이때 단어의 출현 빈도로 wordcloud를 구현하지 않고 TF-IDF를 이용한다. Textrank는 문서 내에서 핵심 문장을 추출하기 위해 필요하다. Doc2Vec은 문서를 벡터화하여 카테고리 분류 및 유사 문서를 추천하기 위해 이용한다.

2. 데이터 전처리와 핵심 문장 추출

이 연구에 사용한 데이터는 AiHub의 '요약문 및 레포트 생성 데이터'이며, 뉴스기사, 보도자료, 역사_문화재, 보고서, 회의록, 사설, 간행물, 연설문, 문학, 나레이션 총 10가지 카테고리로 구성되어 있다. 총 문서의 수는 73,431개이다. 이 데이터셋에서 각 문서는 json 형식으로 저장되어 있는데, 본 연구에서 필요로 하는 문서 제목, 카테고리, 본문을 나타내는 doc_name, category, passage 값만 추출하였다. 본문 내용에 포함된 한자나 이모티콘은 정규표현식을 이용하여 한글, 알파벳, 숫자, 공백, 줄바꿈, 괄호, 문장부호만 남기고 제거한 후에 python의 pickle 라이브러리를 이용하여 데이터를 저장하였다. 각 문서는 python의 dictionary 형태로 보관되며, 이 문서들은 전부 list로 관리된다.

핵심 단어를 추출하기 위해 TF-IDF 기법을 이용하였다. KoNLPy의 Okt 토크나이저를 이용하여 명사를 추출하여 각 문서별 명사 집합을 만들고, 이를 list로 묶어서 저장하였다. 이 list는 TF-IDF를 계산하는 과정에서 IDF(Inverse Document Frequency), 즉 특정 명사가 출현하는 문서의 빈도수를 계산할 때 쓰인다. 핵심 단어 시각화를 할 문서를 입력받아서, 위 과정과 유사하게 KoNLPy의 Okt 토크나이저를 이용해 문서의 명사를 추출한다. 그리고 추출한 명사를 단어 집합으로 변환하는 것이 아닌 단어와 단어의 빈도수 형태, 즉 사전의 형태로 변환한다. 이 사전을 이용하여 TF(Term Frequency)의 값을 계산할 수 있다. 이 과정을 통해 입력 문서의 모든 명사의 TF-IDF값을 계산하고, 이 계산값을 python의 wordcloud 라이브러리에 전달하여 워드클라우드를 출력한다. 아래 표 1은 샘플 문서에서 추출한 명사와 TF-IDF 값의 일부이고 그림 1은 이를 기반으로 출력한 워드클라우드이다.

<표 1> 키워드 추출 및 TF-IDF 예시

단어	TF-IDF
소방	0.258
임용	0.180
공무원	0.114
신규	0.085
최초	0.084
구급	0.080
정문	0.080
국가	0.076
충남소방	0.070
구조구	0.070
소방청장	0.070
방사	0.068
소방청	0.066

(그림 1) 워드클라우드 생성 예시

문서에서 핵심 문장을 추출하기 위해서는 python의 textrankr 라이브러리를 이용하였다. 이 라이브러리는 textrank 알고리즘을 한국어에서 이용할 수 있는 라이브러리이며, 토크나이저를 추가적으로 필요로 하여 KoNLPy의 Okt를 이용하였다. 샘플 문서의 본문에서 핵심 문장 3개를 추출한 결과는 다음과 같다.

1. '소방청(청장 정문호)은 4월 1일 시행되는 소방공무원 신분 국가직화에 따라 85명의 소방사가 국가직으로 최초 임용되어 충남소방에서 근무를 시작한다고 밝혔다'
2. '이들은 신분이 국가직으로 전환됨에 따라 전국 최초로 신규 발령받는 소방공무원이 되는 것이다'
3. '이날 임용되는 신규 소방공무원은 남자 82명, 여자 3명이다'

3. 카테고리 분류와 유사문서 추천

카테고리 분류나 유사 문서를 추천하기 위해서는 문서를 벡터로 변환하는 과정이 필요하다. 이를 위해 python의 Doc2Vec 라이브러리를 활용했다. Doc2Vec을 이용해 문서 벡터를 구하기 위해서는 Doc2Vec 학습이 필요하다. Doc2Vec에 필요한 데이터는 문서, 문서의 태그이다. 그래서 '요약문 및 레포트 생성 데이터'의 본문과 카테고리를 이용하였다. 조사, 어미, 구두점 등을 제거하고 토큰화하여 학습 데이터를 제작하였고, 벡터크기 200, window 5로 설정하여 Doc2Vec을 학습시켰다. Doc2Vec을 이용하여 각 문서 별로 벡터를 구하고, 이를 이용하여 분류 모델인 softmax regression, SVM, randomforest, kNN를 학습시켰다. 위의 모델을 각각 학습을 진행한 후 정확도를 평가하였으며, 그 결과는 표 2와 같다.

<표 2> 모델별 정확도

모델	정확도
Softmax	0.8245901639344262
SVM	0.8030601092896175
RandomForest	0.7172677595628415
kNN	0.4386885245901639

입력문서와 유사한 문서를 추천하기 위해서는 문서간 유사도 계산이 필요하다. 유사문서 추천은 '요약문 및 레포트 생성 데이터' 내에서 수행한다. 유사도 계산은 cosine similarity를 이용하였다. 입력문서를 토큰화하고, Doc2Vec 임베딩으로 문서 벡터를 얻는다. Softmax regression을 통해 카테고리를 분류하고 문서집합에서 동일 카테고리의 문서를 대상으로 입력문서 벡터와 벡터 유사도를 계산하여, 유사도가 높은 문서 순으로 유사문서를 검색하였다.

4. 결론

이 연구에서는 문서의 핵심 단어를 시각화 한 후, 본문의 핵심 문장을 추출하고, 문서 임베딩 기법을 이용하여 같은 카테고리의 유사 문서를 추천해주는 시스템을 구현하였다. 학습데이터 집합으로 AI Hub에 공개되어 있는 '요약문 및 레포트 생성 데이터'를 이용하여 Doc2Vec, softmax regression 모델을 학습하였다. 입력문서와 본문 내용이 가장 유사한 문서들을 추천함으로써 사용자들이 원하는 문서를 쉽게 찾는데 활용할 수 있을 것으로 기대된다.

Acknowledgment

본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (No. 2022-0-00964)

참고문헌

[1] 진찬용, 신성윤, 남수태, "파이썬을 이용한 비정형 데이터 워드 클라우드 시각화", 한국정보통신학회 종합학술대회 논문집, pp.239-240, 2023.
[2] 홍진표, 차정원, "TextRank 알고리즘을 이용한 한국어 중요 문장 추출", 한국정보과학회 2009 한국컴퓨터종합학술대회 논문집, pp.311-314, 2009.
[3] 고동우, 양정진, "KoNLPy와 Word2Vec을 활용한 한국어 자연어 처리 및 분석", 2018 한국컴퓨터종합학술대회 논문집, pp.2140-2142, 2018.
[4] Jamie J. Seol, textrankr, GitHub repository, https://github.com/theeluwin/textrankr
[5] 박은정, 조성준, "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지", 제26회 한글 및 한국어 정보처리 학술대회 학술발표논문집, pp.133-136, 2014.