


캡스톤 디자인 I 종합설계 프로젝트

프로젝트 명	뉴스 키워드 추천
팀 명	해보자
문서 제목	뉴스 키워드 추천 중간 보고서

Version	1.0
Date	2018-04-12

팀원	이 가 영
	이 승 언
지도교수	윤 성 혜 교수님

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-APR-12


CONFIDENTIALITY/SECURITY WARNING

이 문서에 포함되어 있는 정보는 국민대학교 전자정보통신대학 컴퓨터공학부 및 컴퓨터공학부 개설 교과목 캡스톤 디자인I 수강 학생 중 프로젝트 "뉴스 키워드 추천"를 수행하는 팀 "해보자"의 팀원들의 자산입니다. 국민대학교 컴퓨터공학부 및 팀 "해보자"의 팀원들의 서면 허락없이 사용되거나, 재가공 될 수 없습니다.

문서 정보 / 수정 내역


Filename	중간보고서-뉴스 키워드 추천 중간 보고서.doc
원안작성자	이가영
수정작업자	이가영

수정날짜	대표수정자	Revision	추가/수정 항목	내 용
2018-04-12	이가영	1.0	최초 작성	

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-APR-12

목 차

1	프로젝트 목표	4
2	수행 내용 및 중간결과	5
2.1	계획서 상의 연구내용	5
2.2	수행내용	6
3	수정된 연구내용 및 추진 방향	7
3.1	수정사항	7
4	향후 추진계획	8
4.1	향후 계획의 세부 내용	8

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-APR-12


1 프로젝트 목표

기존에 있는 키워드 알림의 기능이 있는 웹, 앱 애플리케이션과 차별화된 키워드 간의 연관 관계가 분석을 기반으로 하여 뉴스 키워드를 추천해주는 소프트웨어를 개발한다.

기존에 개발된 시스템은 사용자가 지정한 키워드를 포함한 결과만을 보여주고 있다. 하지만, 한 가지의 단어를 표현하는 방법은 매우 많다. 따라서 사용자가 원하는 정보를 최대한 많이 받기 위해서는 사용자가 키워드를 표현할 수 있는 모든 경우의 수를 직접 키워드로 설정 해주어야 한다는 문제점이 공통으로 있었다. 우리는 이런 문제점을 파악하고, 사용자에게 더 좋은 편의성을 부여해주기 위해 뉴스 키워드를 추천해주는 소프트웨어를 계획하게 되었다.

사용자가 웹 페이지에서 키워드를 하나 설정하면, 그것과 관련된 연관 단어들을 추천해주고, 사용자가 그중에서 알림을 받길 원하는 키워드를 선택한다. 이후에 그 키워드가 포함된 뉴스 기사 중 이를 내로 업데이트가 된 기사들의 목록을 보여줄 것이다.

연관 단어들을 추천해주는 알고리즘은 Doc2Vec 기법을 사용할 것이다. 대용량의 데이터를 반복적으로 훈련을 시킨 후, 연관된 단어들끼리 군집화하여 벡터 모델을 구축하는 방법이다. 우리는 이번 프로젝트에서 이 벡터 모델의 정확도를 최대화하여 사용자가 원하는 결과값이 나오도록 하는 것을 가장 큰 목표로 하고 있다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-APR-12

2 수행 내용 및 중간결과


2.1 계획서 상의 연구내용

연관 단어들을 추천해주는 알고리즘은 Python 라이브러리 gensim의 Doc2Vec을 이용할 것이다. Doc2Vec은 Word Embedding 기법의 하나로, 단어 간의 연관 관계를 분석해서 연관된 단어들끼리 군집이 되는 벡터를 만드는 방법이다. 이것을 이용하여 연관 관계 분석이 된 벡터 모델을 만들고, 사용자가 키워드를 입력했을 때 그 키워드를 중심으로 연관된 단어들을 사용자에게 보여주는 방식으로 키워드 추천을 할 계획이다.

이 때, Doc2Vec을 하기 위해서는 훈련을 할 대용량의 데이터가 필요하다. 인터넷 기간에 모아두었던 대용량 데이터는 현재 우리가 필요한 스포츠, 연예 분야 부분에서 훈련을 시키기에 부족하기 때문에 새롭게 크롤링을 진행할 예정이다. 데이터는 많을수록 좋지만, 시간적인 한계로 인하여 신문사를 '스포츠 경향', '일간 스포츠'으로 한정하도록 한다.

키워드를 추천받은 사용자는 그것 중에서 알림 받기를 원하는 키워드를 선택한다. 이후에 우리는 그 키워드가 포함된 뉴스 기사를 선별하여 목록을 보여줄 것이다. 이때 키워드가 포함이 되어 있는지는 one-hot encoding을 이용한 Naïve Bayes 방식을 사용한다. 뉴스 기사를 크롤링한 본문 데이터 전체에서 특정 키워드가 있으면 1로, 없으면 0으로 나타내는 식으로 뉴스 기사에 대한 벡터를 만들 것이다. 따라서 1로 표시가 된 기사는 사용자가 원하는 문서라고 판단이 되므로 데이터 베이스에 따로 저장해서 사용자에게 알림을 주도록 한다.

테스트 서버 환경으로는 python Flask 라이브러리를 이용할 것이다. 벡터 모델 구축을 python 언어를 사용하기 때문에 테스트 환경을 Python 라이브러리인 Flask로 사용한다면 벡터 모델을 적용하여 키워드를 추출하고, 원하는 결과값을 출력하기 편리할 것으로 예상된다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-APR-12

2.2 수행내용

가장 먼저 Word Embedding과 Doc2Vec의 방법론에 대한 이해가 필요하다고 판단되어 관련된 논문을 읽고 공부를 시작했다. 이후에, Doc2Vec 알고리즘을 작성해보고, 훈련에 필요한 최소한의 데이터로 테스트를 진행해보았다. 그 결과, 벡터 모델을 구축하고 연관된 키워드들이 추출되는 것을 확인했다.

또한, 이 과정을 사용자가 웹 페이지에서 사용할 수 있도록 Python Flask과 Html을 이용하여 서버를 구축하여 테스트를 진행했다. 현재는 사용자가 키워드를 입력하면, 벡터 모델에 의해 가장 많이 연관된 키워드 10개를 보여주는 수준까지 구현했다.

다음은 우리가 구현한 Test 용 웹 화면이다.

Doc2Vec Demo

Recommend Keyword


<사용자 키워드 입력 화면>

Doc2Vec Demo

Recommend Keyword

빅뱅, 권지용, 탐, 태양, 민효린, 지코, 고마츠, 대성, 팔로우, 산다라박


<연관 키워드 추천 화면>

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-APR-12

3 수정된 연구내용 및 추진 방향

3.1 수정사항

- 인턴쉽 기간에 모아두었던 대용량 데이터를 사용할 계획이었으나, 우리가 사용할 연예, 스포츠 분야의 데이터 부족으로 데이터 수집을 따로 진행할 것이다.
- 기존에는 중앙일보, 이투데이, 국민일보에서 데이터를 수집할 계획이었으나, 연예, 스포츠 분야로 특화되어 있는 신문사인 스포츠 경향, 일간 스포츠에서 데이터를 수집할 것이다.
- 서버 환경을 AWS를 이용하여 구축할 예정이었으나, Python Flask로 수정하여 테스트를 진행할 것이다. 인턴쉽 기간에 flask를 이용하여 서버 구축을 한 경험이 있어 더욱 수월하게 진행이 될 것으로 판단이 되며, 시간적인 한계로 인하여 AWS 서버 구축을 하는 것에서 flask로 변경하였다.
- 기존 tensorflow로 개발 예정이었던 벡터 모델 구축 방법을 Python 라이브러리인 gensim으로 변경했다. 새로운 언어를 사용하는 것보다 기존에 알고 있던 언어를 사용하는 것이 더 수월할 것으로 판단되어 Python의 라이브러리 중 하나인 gensim을 사용하기로 했다.
- 사용자가 추천 키워드를 선택한 후에, 이미 업데이트가 되어있는 뉴스 중에서 실시간으로 크롤링하여 뉴스 목록을 보여줄 계획이었다. 하지만, 이 방법은 실시간 크롤링이 진행될 경우 시간이 너무 오래 걸린다는 단점이 있다. 그래서 우리는 새롭게 업데이트 되는 기사를 크롤링하여 DB에 저장한 후, 사용자가 키워드를 선택하면 DB에서 쿼리문을 통해 기사 목록을 가져올 계획이다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-APR-12

4 향후 추진계획

4.1 향후 계획의 세부 내용

인턴십 기간 중 모아놓은 데이터의 연예, 스포츠 분야가 훈련을 시키기 충분하지 않아서 더 많은 데이터를 수집할 것이다. '스포츠 경향' 신문사의 데이터를 모을 수 있는 크롤링 소프트웨어를 개발하고, 연예, 스포츠 분야의 대용량 뉴스 데이터를 수집할 계획이다.

현재 테스트용으로 개발한 Doc2Vec 알고리즘을 활용하여 실제 사용할 수 있는 벡터 모델을 구축할 것이다. 크롤링 소프트웨어를 활용하여 모은 대용량 데이터로 충분히 훈련을 시키고, 최적화된 가중치를 찾아 사용자가 원하는 결과값이 나올 수 있도록 정확도를 높일 것이다.

또한, 사용자가 추천 키워드 중 원하는 키워드를 선택할 수 있도록 버튼 UI를 활성화할 것이다. 사용자가 버튼을 눌러서 선택하면, 해당하는 키워드를 저장하고, 그 키워드를 포함하고 있는지의 여부를 판단할 수 있는 알고리즘을 작성할 것이다. 이것은 뉴스 기사 목록을 보여줄 때 사용된다.