



국민대학교
전자정보통신대학
컴퓨터공학부

CONFIDENTIALITY/SECURITY WARNING


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

캡스톤 디자인 I 종합설계 프로젝트

프로젝트 명	뉴스 키워드 추천
팀 명	해보자
문서 제목	결과보고서

Version	1.0
Date	2018-MAY-25


이름	이 가영 (조장)
----	-----------

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

문서 정보 / 수정 내역


Filename	결과보고서_산학분반_20153204_이가영.doc
원안작성자	이가영
수정작업자	이가영

수정날짜	대표수정 자	Revision	추가/수정 항 목	내 용
2018-05-24	이가영	1.0	최초 작성	

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

목 차

1	개요	4
1.1	프로젝트 개요	4
1.2	추진 배경 및 필요성	4
1.2.1	키워드 뉴스의 시장 현황	8
1.2.2	키워드 설정 기능이 개발된 시스템의 문제점, 개발할 시스템의 필요성	8
2	개발 내용 및 결과물	6
2.1	목표	6
2.2	연구/개발 내용 및 결과물	7
2.2.1	연구/개발 내용	7
2.2.2	시스템 기능 및 구조 설계도	8
2.2.3	활용/개발된 기술	9
2.2.4	현실적 제한 요소 및 그 해결 방안	9
2.2.5	결과물 목록	10
2.3	기대효과 및 활용방안	10
3	자기평가	11
3.1	차별성	11
3.2	개선 필요성	11
4	참고 문헌	12
5	부록	13
5.1	사용자 매뉴얼	13

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

1 개요

1.1 프로젝트 개요

지금 우리는 정보화 시대에 살고 있으면서, 아주 많은 정보 속에서 살고 있다. 하지만, 과도하게 많은 정보는 우리에게 도움이 된다고 볼 수 없다. 특히, 과도하게 많은 정보를 가지고 있는 가장 대표적인 예로는 신문 기사를 들 수 있다. 이제는 이렇게 많은 정보 속에서 사용자가 원하는 것을 찾아서 보여주는 것 또한 하나의 기술로 자리 잡고 있다. 따라서, 우리는 매일 새롭게 업데이트가 되는 뉴스 기사들 사이에서 사용자들이 원하는 정보들을 구별할 수 있는 소프트웨어를 계획하게 되었다.

Xsports 뉴스와 스포츠 한국 뉴스의 연예 분야 기사를 Python BeautifulSoup 라이브러리를 이용한 크롤링으로 Word Embedding시 필요한 데이터를 수집한다. 수집한 대용량 기사에서 명사만을 추출하고, Doc2vec 모델을 이용하여 명사 사이의 연관 관계를 분석한 벡터 모델을 구축한다.

Python Flask를 이용하여 구축한 웹 페이지에서 사용자가 입력한 키워드를 받아서 그것과 연관이 되어 있는 키워드들을 추천한다. 이후에 사용자는 추천 받은 키워드 중에서 원하는 신문사(Xsports, 일간스포츠, 스포츠월드, 스포츠서울)와 키워드를 선택하면, 해당 키워드가 포함된 뉴스 기사의 목록을 보여주는 소프트웨어를 개발한다.

1.2 추진 배경 및 필요성

1.2.1 키워드 뉴스의 시장 현황

키워드 알림 기술은 구글, 카카오톡, SNEK 등의 웹과 앱 애플리케이션에서 다양하게 개발되어 쓰이고 있다. 시장의 현황을 정확하게 파악하기 위해 몇 가지를 실제로 사용해 보았다. 사용해 본 몇 가지의 애플리케이션 중 간단하게 두 가지만 소개를 하자면, SNEK과 구글의 ‘알리미’가 있다.

	특징	장점	단점
SNEK	<ul style="list-style-type: none"> - 주식 관련 정보에 특화되어 있음 - 웹 애플리케이션 	<ul style="list-style-type: none"> - 관심 콘텐츠에 들어가면, 관련된 콘텐츠를 지속해서 받고 싶은 경우 설정하면 효율적인 키워드를 추천해줌 	<ul style="list-style-type: none"> - 공통으로 키워드 알림 애플리케이션은 사용자가 지정한 키워드를 포함한 결과만을 보여주고 있음. 하지만, 한 가지의 단어를 표현하는 방법은 아주 때문에, 사용자가 원하는 정보를 최대한 많이 받기 위해서는 모든 경우의 수를 전부 키워드로 설정해주어야 함
구글 ‘알리미’	<ul style="list-style-type: none"> - 구글에서 지원하는 서비스 - 웹 애플리케이션 	<ul style="list-style-type: none"> - 매우 많은 데이터를 가지고 있어서 다양한 콘텐츠를 볼 수 있음 - 수신 빈도, 출처, 언어, 지역, 검색결과와 개수, 수신 방법 등을 옵션으로 선택 가능 	

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

1.2.2 키워드 설정 기능이 개발된 시스템의 문제점, 개발할 시스템의 필요성

키워드 뉴스의 시장 현황에서 나타나 있는 것과 같이 키워드 알림 애플리케이션은 사용자가 지정한 키워드를 포함한 결과만을 보여주는 단점을 파악했다. 따라서 기존의 키워드 알림 서비스에 적용할 수 있는 키워드 추천 알고리즘을 개발하기로 했다. 사용자가 키워드 단어 하나를 설정하면 그것과 연관된 키워드들을 추천해 주고, 사용자가 선택한 키워드가 포함된 기사의 목록을 보여줄 것이다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

2 개발 내용 및 결과물


2.1 목표

기존 키워드 알림의 기능이 있는 웹, 앱 애플리케이션과 차별화된 키워드 간의 연관 관계 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 계획했다.

사용자가 웹 페이지에서 키워드를 하나 설정하면, 그것과 관련된 연관 단어들을 추천해주고, 사용자가 그 중에서 알림을 받길 원하는 키워드를 선택한다. 이후에 신문사 별로(Xports 뉴스, 일간스포츠, 스포츠월드, 스포츠서울) 선택한 키워드가 포함된 뉴스 기사들의 목록을 보여주는 페이지로 link를 해준다.

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	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

2.2 연구/개발 내용 및 결과물

2.2.1 연구/개발 내용

2.2.1.1 뉴스 기사 크롤링

Xports 신문사의 연예 분야 기사 데이터를 Python3 BeautifulSoup 라이브러리를 이용하여 수집한다. 연예분야로 데이터를 한정시킨 이유는 다양한 카테고리 벡터모델을 구축할 경우 그 결과 정확도가 매우 떨어진다는 점이다. 또한, 대용량 데이터를 이용하는 학습의 시간이 오래 걸리기 때문에 시간 제약상 한 가지의 카테고리만 학습 데이터로 이용한다.

(총 Data 596.6MB)

2.2.1.2 연관 관계 분석

모아놓은 신문 기사 데이터를 이용하여, 키워드 사이의 연관 관계를 분석한다. 연관관계를 분석하는 방법은 단어를 벡터화하고, 벡터 사이의 거리를 0부터 1 사이의 수치로 나타낸 후 1에 가까울수록 연관 관계가 높다고 판단한다. 이때 단어를 벡터화하는 방법으로는 Python3 gensim의 Doc2vec기술을 이용할 것이다.

input으로는 정확한 결과값을 얻고, 훈련시간을 줄이기 위하여 수집한 뉴스 기사 데이터를 명사만 추출하여 사용한다. (총 Data 596.6MB → 406MB) 훈련 반복의 횟수와 alpha 값 등의 가중치를 조정하여 훈련 결과의 정확도가 가장 높은 것을 찾도록 한다.

2.2.1.3 HTML 및 테스트 환경 구축

사용자가 키워드를 입력할 수 있는 텍스트 입력 창과 연관된 키워드들을 보여주고, 결과적으로 키워드가 포함된 기사들의 목록을 보여줄 수 있는 HTML UI를 제작한다. 이때 HTML은 Jinja2의 문법에 맞도록 작성한다. 또한, 사용자가 입력한 키워드에 대한 결과들을 보여주는 방법으로는 python 라이브러리인 Flask를 이용할 것이다. Flask는 인턴 기간에 산학과의 연계과정에서 다루어본 적이 있으며, 벡터 모델 구축을 python 언어를 사용하고 있다는 점과, 한국어 인코딩의 문제를 고려하여 사용하기로 했다.

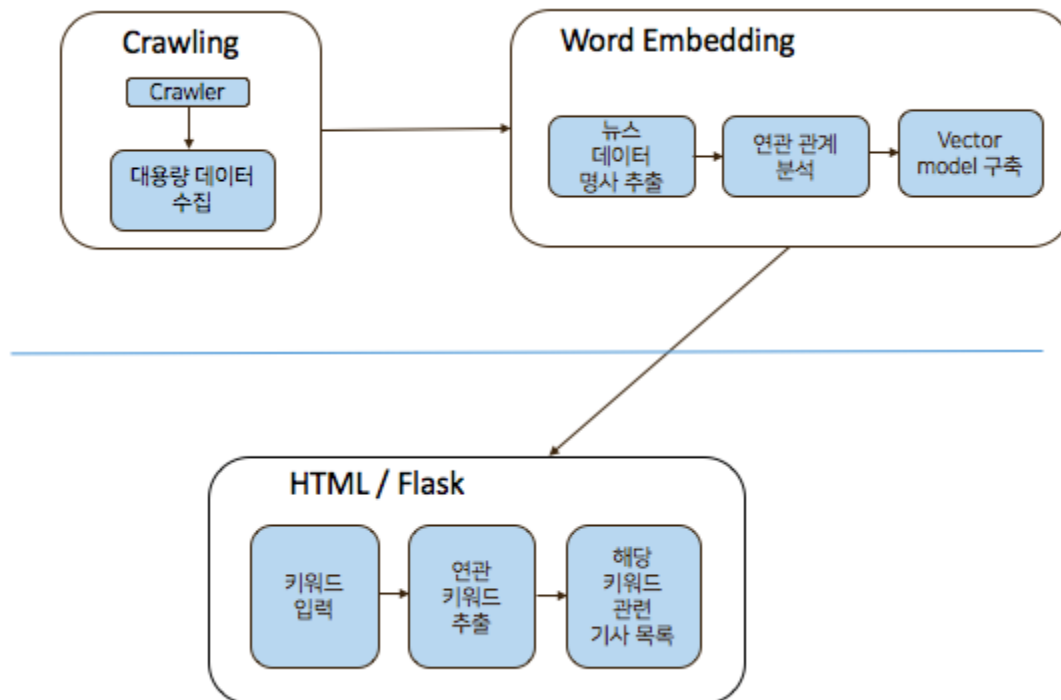
2.2.1.4 키워드가 포함된 뉴스 기사 분류 방식


추천 받은 키워드들 중 사용자가 원하는 키워드를 선택한 후 그것이 포함된 뉴스 기사를 보여주는 방식이다. 기존에 one-hot encoding 방식을 사용할 계획이었으나, 결과의 정확도가 예상보다 매우 낮았다. 따라서 추천 키워드를 신문사 별로(Xports, 일간 스포츠 클릭하면 해당 신문사의 검색창에 키워드가 자동으로 입력이 되도록 Link를 하여 추천 키워드가 포함된 목록을 보여주는 방식으로 변경했다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

2.2.2 시스템 기능 및 구조 설계도

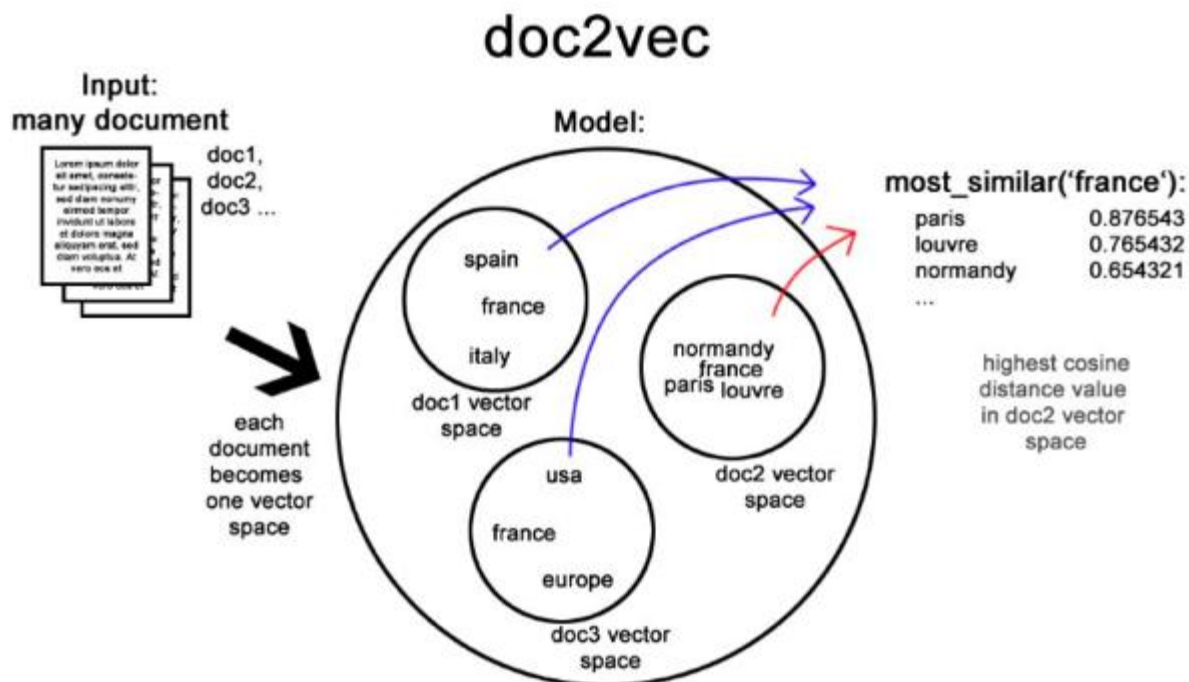
1. Crawling
크롤러를 이용하여 대용량 데이터를 수집한다.
2. Word Embedding
수집한 데이터에서 명사만을 추출한다. 명사만 추출한 데이터를 입력으로 하여 단어 사이의 연관 관계를 분석하고 vector model로 만든다.
3. HTML / Flask
사용자가 키워드 하나를 입력하면 뉴스 데이터를 기반으로 하여 연관된 키워드 10개를 추천해준다. 원하는 신문사의 추천 키워드를 누르면 해당 키워드가 포함되어 있는 기사의 목록을 보여주는 Link로 연결해준다.



 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

2.2.3 활용/개발된 기술

1. 연관 관계 분석 (Doc2vec)



Doc2Vec 은 Word Embedding 기법의 하나로, 단어 간의 연관 관계를 분석해서 연관된 단어들끼리 군집이 되는 벡터를 만드는 방법이다. 즉, 뉴스 기사 본문과 같은 큰 텍스트 블록에 대하여 vector 값으로 변환시키는 것으로 쉽게 word2vec 의 확장된 버전이다.


Doc2vec 의 입력은 Labeled Sentence 객체의 Iterator 이다. 우리는 크롤링한 뉴스 기사 본문 데이터에서 명사를 추출한 것을 input 으로 했다.

이것을 이용하여 연관 관계 분석이 된 벡터 모델을 만들고, 사용자가 키워드를 입력했을 때 그 키워드를 중심으로 연관된 단어들을 사용자에게 보여주는 방식으로 키워드를 추천한다.

2. Python Flask / jinja2 를 적용한 HTML
3. Python BeautifulSoup 을 이용한 크롤링

2.2.4 현실적 제한 요소 및 그 해결 방안

- 연관 관계 분석을 할 때 단어간의 연관 관계가 뚜렷하지 않을 수 있다는 문제점이 있다. 또한, 대용량 데이터를 이용하기 때문에 우리가 사용하고 있는 하드웨어의 속도를 고려했을 때 여러 개의 카테고리를 학습할 경우 시간적 제약이 따른다. 따라서 그 관계

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

가 비교적 뚜렷하게 나타나며, 시간적인 한계를 고려했을 때 연예 분야만을 대상으로 프로젝트를 진행한다.


- 부족한 연예분야의 데이터를 더 많이 확보하기 위해 크롤링을 진행할 때, 각 신문사마다 html 태그가 다르기 때문에 크롤링 소프트웨어를 새롭게 수정해주어야 한다. 하지만 모든 신문사에 맞춰서 소프트웨어를 수정하는 것은 시간적으로 부족하다. 또한, 크롤링으로 대용량 데이터를 수집하는 것 자체에 많은 시간이 필요하기 때문에 신문사를 한정하여 데이터를 수집할 것이다. 우리가 선정한 신문사는 'Xports 뉴스와 스포츠 한국'이다.

2.2.5 결과물 목록

결과물 목록	기술 문서 유무
뉴스 데이터 596.6MB	무
뉴스 본문 명사 추출 데이터 406MB	무
Doc2Vec Model	무
Flask web page	무

2.3 기대효과 및 활용방안

1. 과도한 데이터들 사이에서 하나의 키워드 설정으로 사용자가 생각하지 못했던 연관된 키워드들을 추천해주기 때문에 시간과 비용을 절약할 수 있다.
2. 뉴스 기사 이외에도 학교 공지사항 알림이나 중고거래 사이트, 혹은 많은 광고메일들 속에서 원하는 정보를 찾고 싶을 때, 등 이 기술을 다양한 목적으로 활용할 수 있다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

3 자기평가


최종 결과물 : 연관관계 분석을 기반으로 사용자가 입력한 키워드와 관련된 키워드 10개를 추천해주고, 신문사별로 각 키워드가 포함된 뉴스 기사의 목록을 볼 수 있는 Web page이다.

3.1 차별성

기존에 있는 키워드 알림 서비스와는 다르게 연관 관계 분석을 기반으로 하여 뉴스 기사 목록을 보여준다. 또한, 연관 관계 분석의 내용을 시각적으로 보여주기 때문에 사용자 입장에서 연관 관계 분석의 정도를 이해하기 쉽다.


3.2 개선 필요성

수집한 데이터에서 명사를 추출할 때 조사 등이 제대로 분리되지 않았다. 따라서 훈련의 결과에 조사가 함께 등장하거나 명사가 아닌 키워드가 나오는 경우가 있다. 또한, 수집한 뉴스 본문 기사 내용 자체에 일부 단어만 바뀐 내용이 많았다. 따라서 훈련한 데이터가 충분히 정제되지 않았다. 이러한 이유로 정확도가 떨어진 것을 확인할 수 있었다. 기존의 수행계획서 상에는 키워드 '알림' 기능으로 사용자의 이메일과 웹 페이지 상에서 바로 뉴스 기사의 목록을 볼 수 있도록 계획을 세웠다. 하지만, 시간의 제약으로 해당 키워드가 포함된 신문사의 Link로 대체를 했기 때문에 사용자 입장에서 사용성이 떨어진다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

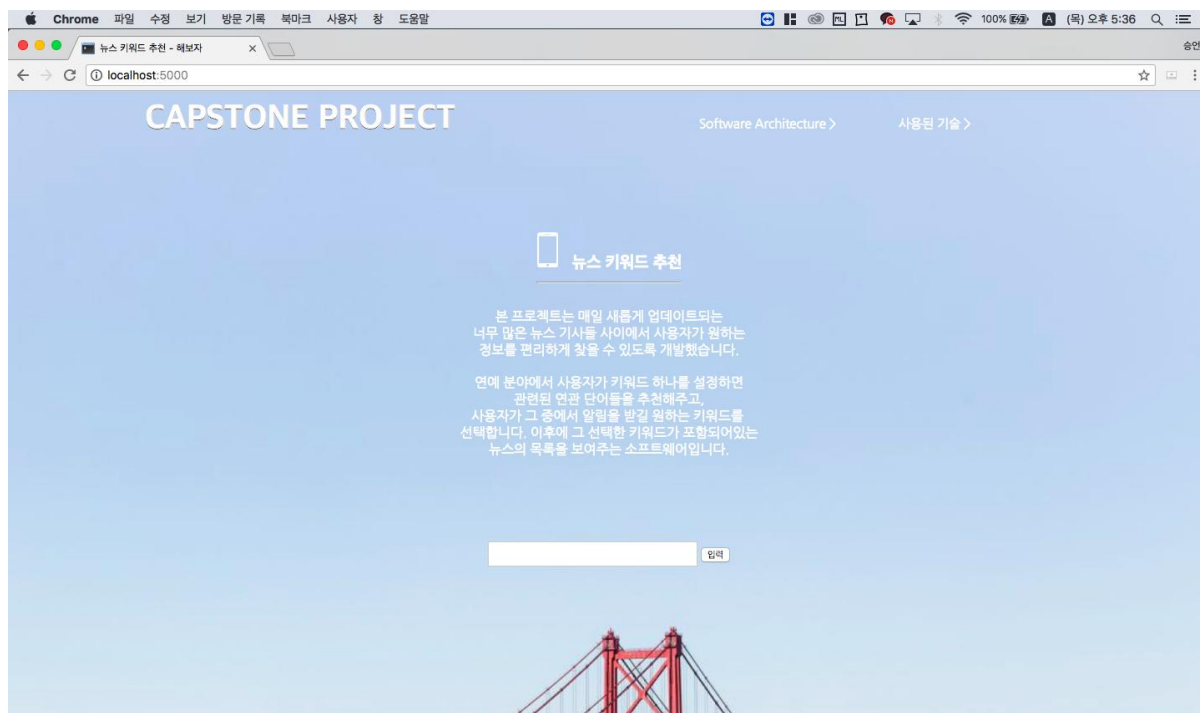
4 참고문헌

- (1) 김도우, 구명완. (2017). Doc2Vec과 Word2Vec을 활용한 Convolutional Neural Network 기반 한국어 신문 기사 분류. 정보과학회논문지, 44(7), 742-747.
- (2) 정영미, 이재윤. (1998). Preliminary on the Analysis of Term Associations in Korean Text. 한국정보관리학회 학술대회 논문집, , 243-246.
- (3) 구글 검색 알고리즘 원리 <http://markov.tistory.com/3>
- (4) Text mining(python) <https://www.lucypark.kr/courses/2015-ba/text-mining.html#3-load-tokens-with-nltktext>
- (5) 연관도 분석을 이용한 데이터 마이닝 <https://www.slideshare.net/ocworld/ss-60871182>


 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

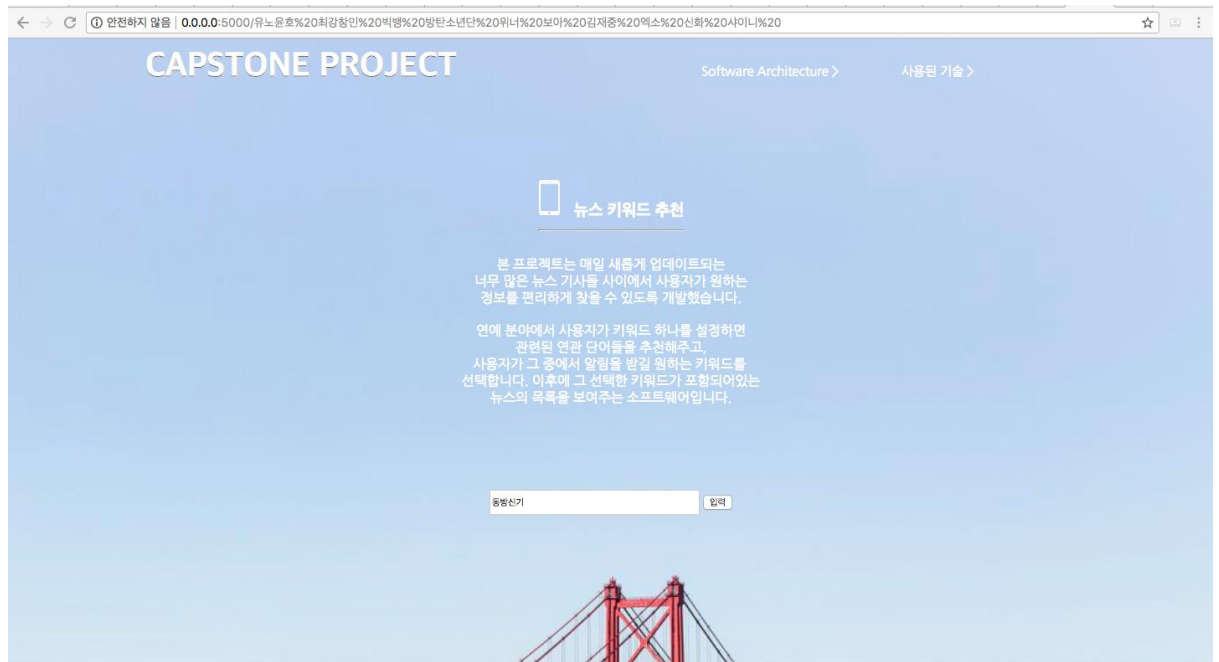
5 부록

5.1 사용자 매뉴얼




해당 페이지에 접속하면 다음과 같은 메인 화면을 볼 수 있다.

 <div> 국민대학교 컴퓨터공학부 캡스톤 디자인 I </div>	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24



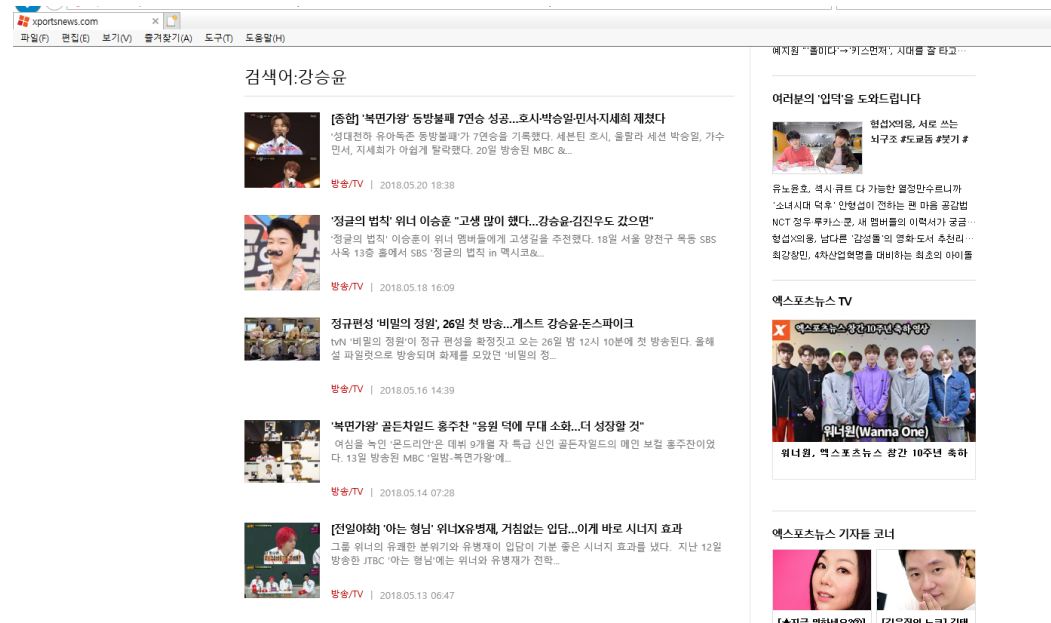
다음과 같이 연예 분야의 특정 키워드를 적고 입력을 누른다.



 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

다음과 같이 가장 연관 관계가 높은 키워드 20개에 대한 wordcloud와 추천 키워드 10개와 함께 엑스포츠뉴스, 일간스포츠, 스포츠월드, 스포츠서울 4가지의 신문사 별로 각 키워드가 포함된 뉴스 기사 목록으로 link가 되어있다.

이 키워드 중 원하는 신문사의 원하는 키워드를 클릭한다.

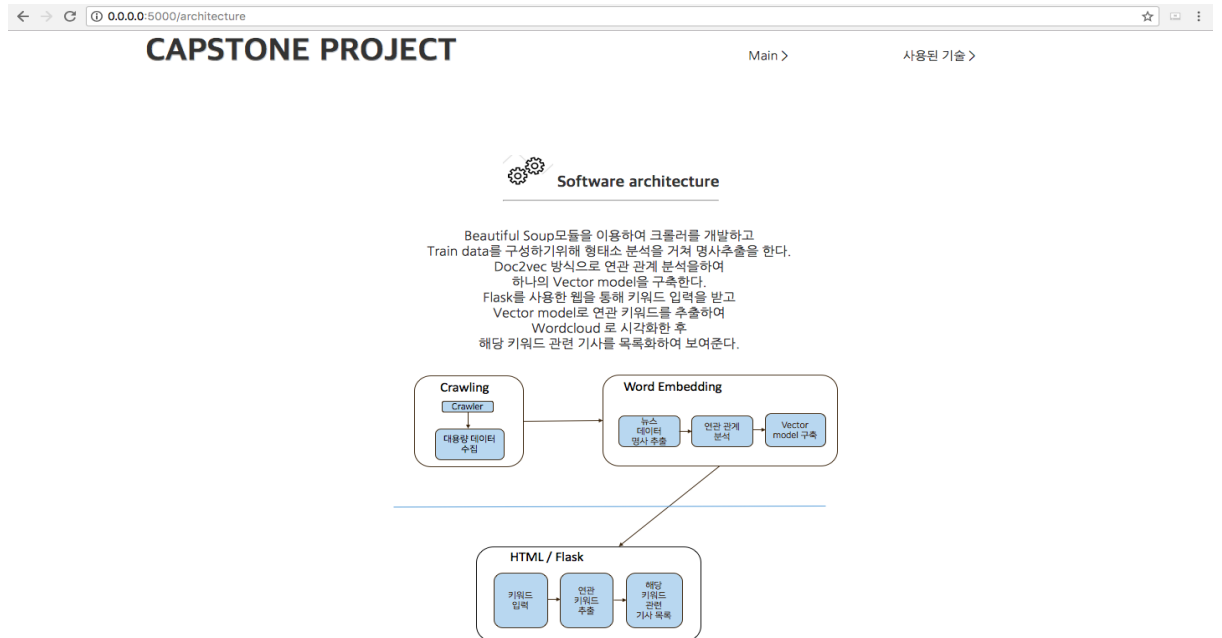


다음과 같은 결과를 확인할 수 있다.

메인 페이지에서는 software architecture와 사용된 기술인 doc2vec에 대한 설명도 볼 수 있는 메뉴가 따로 있다. 상단의 오른쪽에 보면 두 가지의 메뉴가 있다.

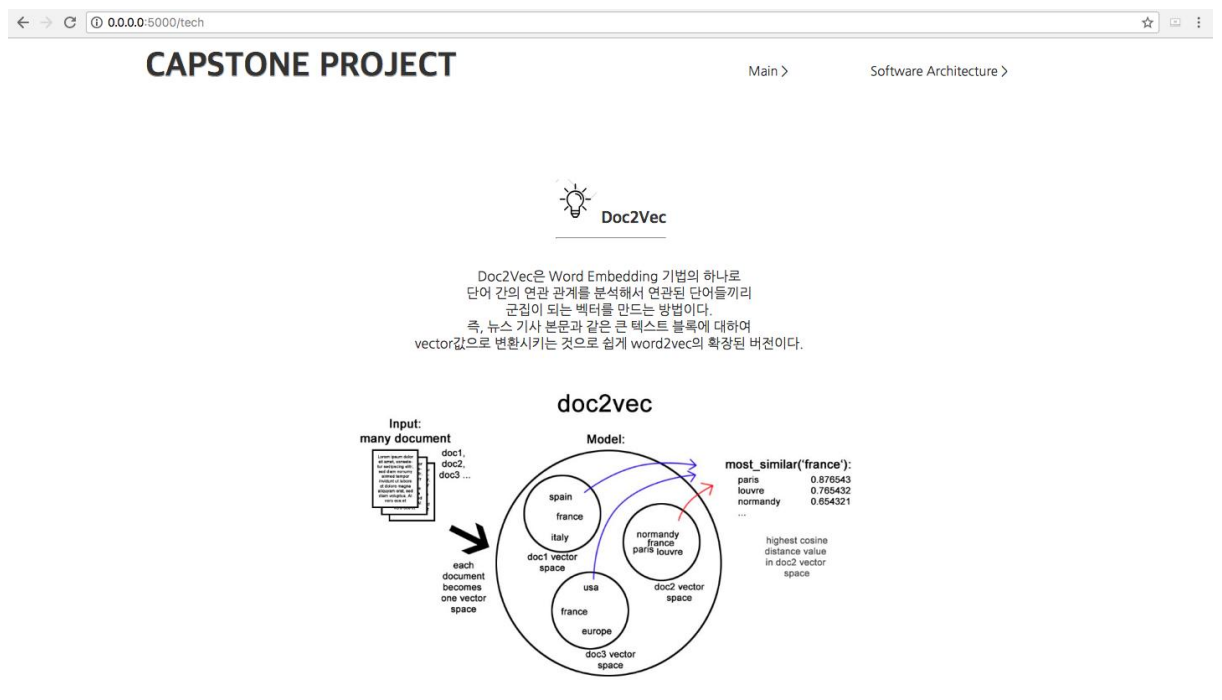
 <div> <p>국민대학교</p> <p>컴퓨터공학부</p> <p>캡스톤 디자인 I</p> </div>	결과보고서		
	프로젝트 명	뉴스 키워드 추천	
	팀 명	해보자	
	Confidential Restricted	Version 1.0	2018-MAY-24

software architecture을 누르면



다음과 같은 화면을 확인할 수 있다.

사용된 기술을 누르면



다음과 같은 화면을 볼 수 있다. 여기서 Main을 누르면 다시 메인 페이지로 돌아갈 수 있다.