



국민대학교
전자정보통신대학
컴퓨터공학부

캡스톤 디자인 I

종합설계 프로젝트


| | |
|--------|-----------|
| 프로젝트 명 | 뉴스 키워드 추천 |
| 팀 명 | 해보자 팀 |
| 문서 제목 | 결과보고서 |

| | |
|---------|-------------|
| Version | 1.0 |
| Date | 2018-MAR-23 |

| | |
|----|-----|
| 이름 | 이승연 |
|----|-----|

CONFIDENTIALITY/SECURITY WARNING


이 문서에 포함되어 있는 정보는 국민대학교 전자정보통신대학 컴퓨터공학부 및 컴퓨터공학부 개설 교과목 캡스톤 디자인I 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.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

문서 정보 / 수정 내역


| | |
|-----------------|---------------------|
| Filename | 수행결과보고서_23조_이승언.doc |
| 원안작성자 | 이승언 |
| 수정작업자 | 이승언 |

| 수정날짜 | 대표수정 자 | Revision | 추가/수정 항 목 | 내 용 |
|--------------|-----------|----------|--------------|--------|
| 2018. 5. 23. | 이승언 | 1.0 | 최초 작성 | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

목 차

| | | |
|-------|-------------------------------|----|
| 1 | 개요..... | 4 |
| 1.1 | 프로젝트 개요 | 4 |
| 1.2 | 추진 배경 및 필요성..... | 4 |
| 1.2.1 | 키워드 뉴스 시장 현황 | 5 |
| 1.2.2 | 키워드 설정 기능이 개발된 프로그램의 한계 | 5 |
| 2 | 개발 내용 및 결과물 | 6 |
| 2.1 | 목표..... | 6 |
| 2.2 | 연구/개발 내용 및 결과물..... | 6 |
| 2.2.1 | 연구/개발 내용..... | 6 |
| 2.2.2 | 활용/개발된 기술..... | 7 |
| 2.2.3 | 현실적 제한 요소 및 그 해결 방안 | 7 |
| 2.2.4 | 결과물 목록..... | 8 |
| 2.3 | 기대효과 및 활용방안 | 10 |
| 3 | 자기평가..... | 11 |
| 4 | 참고문헌..... | 11 |

| | | | |
|--|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |


1 개요

1.1 프로젝트 개요

스마트폰이나 태블릿 PC 가 개발되면서 누구나 지면 신문보다 인터넷 신문을 쉽게 접할 수 있게 되었다. 신속한 정보 제공이 가능하고 새로운 기삿거리가 생기면 바로 기사화되어서 뉴스를 읽을 수 있다. 이렇게 정해진 시간 없이 읽을 수 있는 인터넷 뉴스에 실시간으로 뉴스를 받아보는 기능을 필요로 하는 사람들이 많아졌다. 하지만 넘쳐나는 뉴스 기사 중 내가 원하는 내용의 기사는 적은 것에 비해 불필요한 기사가 더 많이 제공되는 불편한 점이 있다. 본 프로젝트는 뉴스 기사를 크롤링한 데이터를 이용하여 사용자가 설정한 키워드와 기사에 등장하는 단어들의 시맨틱 관계를 분석한다. 분석한 연관 관계를 적용하여 사용자가 필요로 하는, 필요로 할만한 키워드를 추천하고 관련 기사를 웹페이지에서 바로 찾아볼 수 있도록 해준다.

1.2 추진 배경 및 필요성

다양한 언론의 시각을 종합적으로 볼 수 있는 이유나 빠르게 많은 정보를 얻을 수 있는 장점이 있는 인터넷 뉴스이지만 정작 많은 기사 속에서 원하는 뉴스를 찾기 위해 카테고리별로 드래그하면서 시간을 많이 보내게 되는데 설정한 키워드와 관련된 키워드를 추천해서 관련된 기사를 보여주면 그 시간들을 절약할 수 있다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

1.2.1 키워드 뉴스 시장 현황

| 종류 | 특징 | 장점 | 단점 |
|------------|--|--|-----------------------------|
| 키워드 뉴스 | <ul style="list-style-type: none"> 구글 플레이에서 제공하는 안드로이드 어플리케이션 사용자가 설정한 키워드가 포함된 뉴스 기사를 볼 수 있다. | 실시간으로 알림을 받아볼 수 있다. | 관심 있는 키워드를 일일이 입력해서 설정해야한다. |
| Google 알리미 | <ul style="list-style-type: none"> 구글에서 제공하는 서비스이다. | 수신 빈도, 출처, 언어, 개수 등 여러가지 옵션을 설정해서 이메일로 소식을 받아볼 수 있다. | |
| SNEK | <ul style="list-style-type: none"> 투자를 위한 금융 리서치 플랫폼이다. | 자신만의 키워드로 구성된 카테고리를 만들어 시장에서 발생하는 뉴스를 즉시 확인할 수 있다. | |

1.2.2 키워드 설정 기능이 개발된 프로그램의 한계

키워드 설정 기능이 있는 소프트웨어 모두 설정한 키워드가 포함되어 있어야만 검색과 알림을 받아 볼 수 있다. 즉 의미가 비슷하거나 관련성이 있는 기사라도 키워드 자체가 기사에 담겨있지 않으면 사용자는 알림을 받아볼 수 없고 받아보기 위해서는 하나하나 모든 키워드를 설정해야 하는 번거로움이 있다.

| | | | |
|--|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

2 개발 내용 및 결과물

2.1 목표

BeautifulSoup 모듈로 크롤링한 데이터를 키워드간 연관 관계 분석의 훈련 데이터로 이용한다. Konlpy 의 Twitter 모듈을 이용하여 형태소 분석한다. 형태소 분석으로 명사 추출을 하여 훈련 데이터 형식을 만들고 Gensim 의 Doc2vec 모듈을 이용하여 모델을 생성하고 연관 관계 분석을 한다. 사용자가 키워드를 입력했을때 연관 관계를 Wordcloud 모듈을 이용하여 시각화하여 보여준다. Javascript 로 사용자가 키워드를 선택하면 함수를 호출해서 선택한 키워드가 포함된 2 일이내의 뉴스 기사를 크롤링하여 목록화해서 보여준다.

2.2 연구/개발 내용 및 결과물

2.2.1 연구/개발 내용

2.2.1.1 데이터 크롤러 개발

BeautifulSoup 모듈을 이용하여 크롤링한 데이터를 한 라인에 한 기사(document)를 적는 형태로 파일을 만들었다. 익스포트에서 454.6MB, 스포츠 한국에서 142MB를 크롤링하였다.

2.2.1.2 형태소 분석 및 Train 데이터 구성

키워드 분석 및 추출을 목표로 하기 때문에 불필요한 데이터를 제거해야 데이터 훈련 시간도 줄어들 수 있으므로 . Twitter 모듈로 형태소 분석을 마친 뒤 nouns 함수로 명사 추출을 완료하였다. 총 596.6MB의 데이터가 명사 추출을 끝낸 후 406MB의 크기로 줄어들었다.

2.2.1.3 키워드 추천 시각화

사용자가 설정한 키워드를 중심으로 Gensim 모듈을 이용하여 연관 관계 분석한 모델을 생성하고 생성된 모델을 통해서 사용자가 흥미를 가질만한 키워드를 10가지를 추출하였다. 키워드 추천 형태는 지정한 키워드를 중심으로 연관 관계를 적용한 워드 클라우드로 보여준다. 워드 클라우드는 파이썬의 Wordcloud 모듈을 사용하여 만들어지고 키워드 추천의 갯수보다 10가지 더 많은 20가지의 키워드를 연관 관계의 가중치에 따른 폰트 크기로 보여주고 PNG파일 형식으로 저장된다.

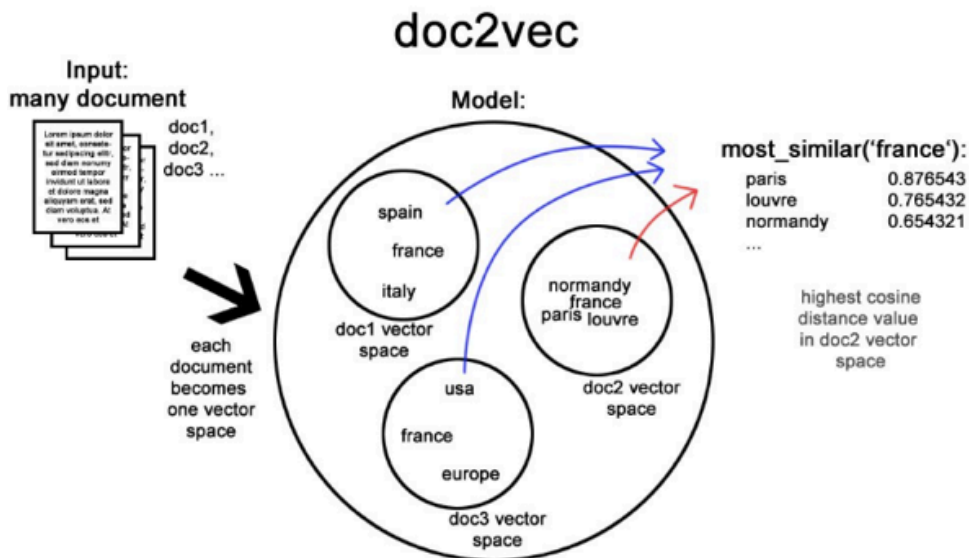
| | | | |
|--|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

2.2.1.4 웹 페이지 구성

저장된 워드클라우드를 Flask로 local 서버를 사용하여 사용자의 입력이 들어오면 워드클라우드를 구성하고 이미지를 업데이트하였다. 추천된 키워드를 해당 신문사의 기사와 연결되도록 링크를 설정해주었다.

2.2.2 활용/개발된 기술

- Doc2vec



연관 관계 분석을 위해 Doc2vec 기술이 활용되었다. Doc2vec이란 Word2vec의 확장된 기술로 뉴스 기사 본문과 같은 큰 텍스트 블록을 하나의 벡터값을 가지도록 하고 문서의 의미가 반영된 유사도에 기반하여 고정 크기의 문서 벡터 표현을 생성하는 알고리즘이다.

2.2.3 현실적 제한 요소 및 그 해결 방안

- (1) 연예와 같이 관계나 소속 집단이 있는 분야에 상하관계를 적용하기가 쉽다고 생각해서 카테고리를 연예로 한정한다.
- (2) 신문사가 매우 많아서 연예에 대한 기사가 많은 '엑스포츠'와 '스포츠한국'의 기사를 크롤링하였다.

| | | | |
|---|-------------------------|-------------|-------------|
|  | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

2.2.4 결과물 목록

2.2.4.1 크롤링 데이터

배우 심은경 주연의 스릴러 영화 '넌 기다리며'가 최근 촬영을 마쳤다고 이 영화 투자·배급을 맡은 뉴가 31일 밝혔다. '넌 기다리며'는 15년 전 연쇄살인범에 의해 아버지를 잃은 소녀와 그녀를 보살피며 온 형사, 15년 만에 세상에 나온 범인의 복잡하게 얽힌 운명을 그린 영화다. 이번 영화로 처음 스릴러에 도전한 심은경은 아버지를 죽인 진짜 범인을 밝혀내고자 15년을 기다린 소녀 '희주' 역을 맡았다. 심은경은 "시나리오를 보자마자 너무나 욕심이 났고 생애 첫 스릴러 영화라 어느 때보다 뜻깊은 작품이었다"고 촬영을 마친 소감을 밝혔다. 김성오가 15년간 간절히 출소만을 기다린 연쇄살인범 '기범' 역을 맡아 체중 감량을 감행하며 살기 어린 눈빛의 다층적인 인물을 표현했다. 오랜 시간 희주를 보살피며 기범을 기다려 온 형사 대영 역은 윤제문이 맡아 열연했다. 영화는 후반 작업을 거쳐 올 하반기 개봉할 예정이다.

17일 오후 서울 영등포 타임스퀘어 아모리스홀에서 열린 tvN 새 월화드라마 '써클 : 이어진 두 세계' 제작발표회에서 공승연과 여진구, 김강우, 이기광이 포즈를 취하고 있다. 드라마 '써클'은 현재와 미래를 배경으로 벌어진 미스터리한 사건을 추적해가는 과정을 그린 드라마로 오는 22일 첫 방송될 예정이다.

아들과 함께 포즈를 취한 한채영 배우 한채영이 지난 3일 인스타그램에 아들과 함께 찍은 '언니들의 슬램덩크 2' 출연 당시 '반전 매력'으로 시청자들의 호응을 얻었던 한채영은 '아들바보'인 모습으로도 화제를 얻었다. 한채영은 지난 2007년 일반인 남성과 결혼해 2013년 득남했다. '언니들의 슬램덩크' 당시 멤버들과 함께한 한채영. '송은이 김숙의 언니네 라디오' 개편맞이 특별 초대석, '언니가 돌아왔다'에 진지희와 함께 출연할 예정이다. 이번 '언니네' 출연은 김숙과의 의리로 인한 것이기에 더 의미가 크다. 지난 5월 31일 '언니쓰' 멤버 공민지와 함께 출연했을 당시 '다음에 꼭 다시 오겠다'고 한 약속을 지키게 된 것이다. 이에 따라 '절친' 김숙과 함께할 '언니네 라디오'에서 들려줄 한채영의 활약이 기대를 모으고 있다.

2.2.4.2 형태소 분석 후 데이터 (명사 추출)


배우 심은경 주연 스릴러 영화 '넌 기다리며'가 최근 촬영 이 영화 투자 배급 뉴 일 넌 년 전 연쇄살인범 아파 소녀 그녀 온 형사 년 세상 범인 얽힌 운명 그린 영화 이번 영화로 처음 스릴러 도전 심은경 아파 진짜 범인 년 소녀 역 심은경 시나리오 너무나 욕심 생애 첫 스릴러 영화 때 뜻 작품 고 촬영 소감 김성오 년 간절 소만 연쇄살인범 기범 역 체중 감량 감행 살기 눈빛 총 인물 시간 기범 온 형사 영 역 윤제문 열연 영화 후반 작업 하반기 개봉 예정

일 오후 서울 영등포 타임스퀘어 아모 리스 홀 새 월화드라마 써클 이어진 두 세계 제작발표회 공 승 연 여진구 김강우 이기광 포즈 드라마 써클 은 현재 미래 배경 미스터리 사건 추적 과정 그린 드라마 일 첫 방송 예정

아들 포즈 취한 한채영 배우 한채영 지난 일 인스타그램 아들 언니 슬램덩크 출연 당시 반전 매력 시청자 호응 한채영 아들 바보 인 모습 화제 한채영 지난 년 일반인 남성 결혼 년 득남 언니 슬램덩크 당시 멤버 한채영 송은이 김숙 언니네 라디오 개편 맞이 특별 초대 석 언니 진지희 예정 이번 언니네 출연 김숙 의리 것 이기 더 의미 지난 월 일 언니 멤버 공민지 당시 다음 꼭 다시 고 약속 것 이 절친 김숙 언니네 라디오 에서 한채영 활약 기대

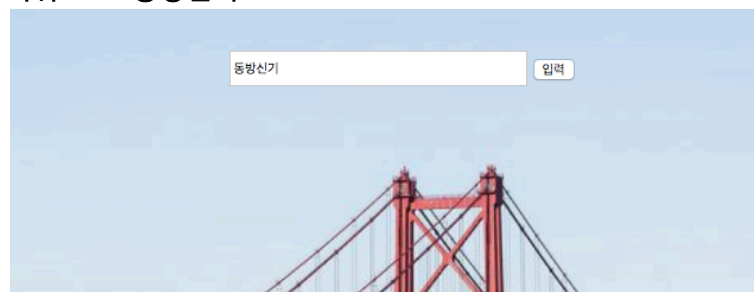
2.2.4.3 Doc2vec 모델 결과

```
>>>
>>> model = gensim.models.Doc2Vec.load("/home/nlpserver/Desktop/Test/model_10epoch/doc2vec.model")
>>> print(model.most_similar(u'동방신기', topn=10))
[('유노윤호', 0.35943925380706787), ('최강창민', 0.354735791683197), ('빅뱅', 0.34390050172805786), ('방탄소년단', 0.3332982659339905), ('위너', 0.33249378204345703), ('보아', 0.3282546401023865), ('김재중', 0.31260737776756287), ('엑소', 0.3112281560897827), ('신화', 0.3110308051109314), ('샤이니', 0.3098337650299072)]
```


| | | | |
|--|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

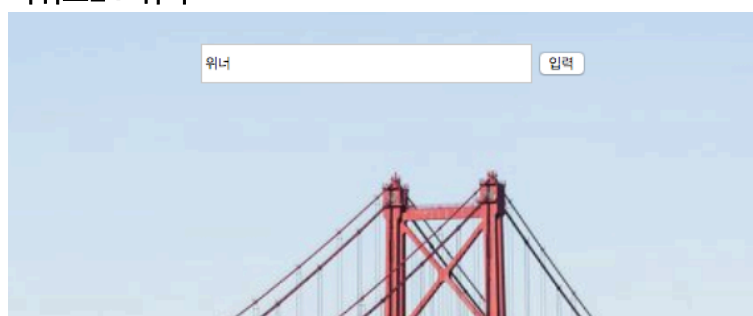
2.2.4.4 Wordcloud

- 키워드1 : '동방신기'




슈퍼주니어 위너 최강창민 엑소 김재중
 소녀시대 유영진 니아스 유노 윤호
 빅뱅 방탄소년단
 보아 신화 더보이즈

- 키워드2 : '위너'



이승훈 아이콘 쇼미더머니 현아
 젝스키스 동방신기 컴백 빅뱅
 트와이스 위너 강승윤
 송민호 야생유엔비 럽미럽미
 엑소 스트레 방탄소년단 에픽하이

두 키워드를 검색 결과 해당 그룹의 멤버이름과 동료 아이돌 그룹의 이름이 연관되어 검색된다는 것을 알 수 있다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

2.2.4.5 추천 키워드 목록

Recommand Keyword

| | | | | | | | | | | |
|---|-----|-----|------|------|------|-----|----|------|-----|-------|
|  | 강승윤 | 송민호 | 젝스키스 | 동방신기 | 럽미럽미 | 아이콘 | 워너 | 트와이스 | 이승훈 | 방탄소년단 |
|  | 강승윤 | 송민호 | 젝스키스 | 동방신기 | 럽미럽미 | 아이콘 | 워너 | 트와이스 | 이승훈 | 방탄소년단 |
|  | 강승윤 | 송민호 | 젝스키스 | 동방신기 | 럽미럽미 | 아이콘 | 워너 | 트와이스 | 이승훈 | 방탄소년단 |
|  | 강승윤 | 송민호 | 젝스키스 | 동방신기 | 럽미럽미 | 아이콘 | 워너 | 트와이스 | 이승훈 | 방탄소년단 |

10개의 추천 키워드를 신문사별로 나열하고 키워드를 클릭하면 해당 키워드와 관련된 기사들이 링크로 연결되어 손쉽게 기사를 읽어볼 수 있다.

2.3 기대효과 및 활용방안

- (1) 키워드들을 하나하나 입력하지 않고 추천 키워드로 관련 있는 키워드를 선택할 수 있으므로 키워드를 설정하는 시간과 비용을 절약할 수 있다.
- (2) 키워드 분석 알림 서비스를 뉴스 기사 이외에도 일이 확인하지 않으면 정보를 놓칠 수 있는 학교의 공지사항이나 게시판, 중고 거래 사이트에 이 기술을 이용한 서비스를 활용 할 수 있다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

3 자기평가

3.1 차별성

기존의 키워드 알림서비스는 키워드를 하나씩 설정해야해서 연관된 키워드도 따로 설정해야하는 번거로움이 있었는데 연관된 키워드를 분석하여 키워드를 추천해주어서 여러번 검색하고 설정해야 하는 번거로움을 줄여주었다. 또한 검색만으로 연관된 키워드들을 시각적으로 잘 보여주는 점이 기존의 서비스와는 다른 점이다.

3.2 개선 필요성

학습에 불필요한 데이터를 제거하기 위해 명사 추출을 하였는데 의도한 바와 다르게 추출되어 결과에 좋지 않은 영향을 미쳤다. 한가지의 예로 '레드 벨벳'과 같이 그룹명에 띄어쓰기가 들어간 경우 고유명사로 인식하지 못하고 '레드'와 '벨벳'이라는 단어가 따로 추출되었다. 초반의 계획은 키워드를 추천해주고 추천 키워드를 선택하면 알림서비스를 해주는 것이었는데 현실적 제한 요소때문에 본 프로젝트를 추천만 하게 되었다. 연관 키워드를 추천한 후 알림서비스에 접목시킨다면 사용성이 증가할 것으로 예상된다.


4 참고문헌

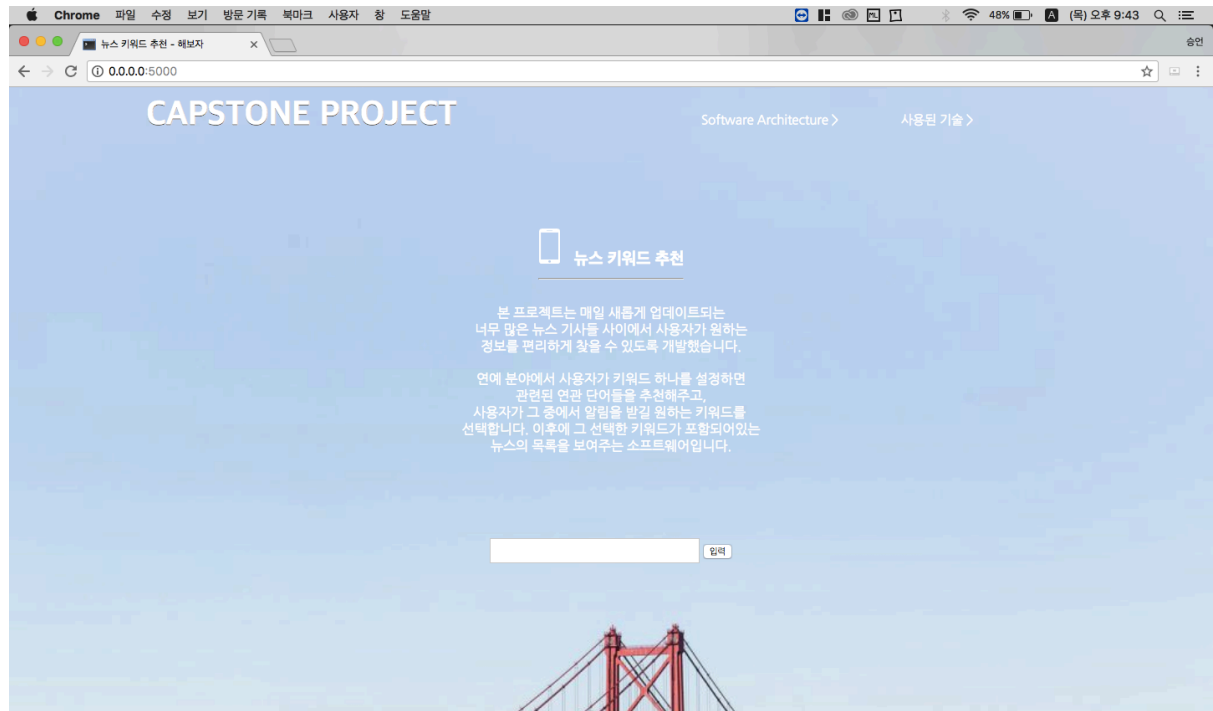
- (1) 김도우, 구명완. (2017). Doc2Vec과 Word2Vec을 활용한 Convolutional Neural Network 기반 한국어 신문 기사 분류. 정보과학회논문지, 44(7), 742-747.
- (2) 정영미, 이재윤. (1998). Preliminary on the Analysis of Term Associations in Korean Text. 한국정보관리학회 학술대회 논문집, , 243-246.
- (3) 구글 검색 알고리즘 원리 <http://markov.tistory.com/3>
- (4) Text mining(python) <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>

5 부록

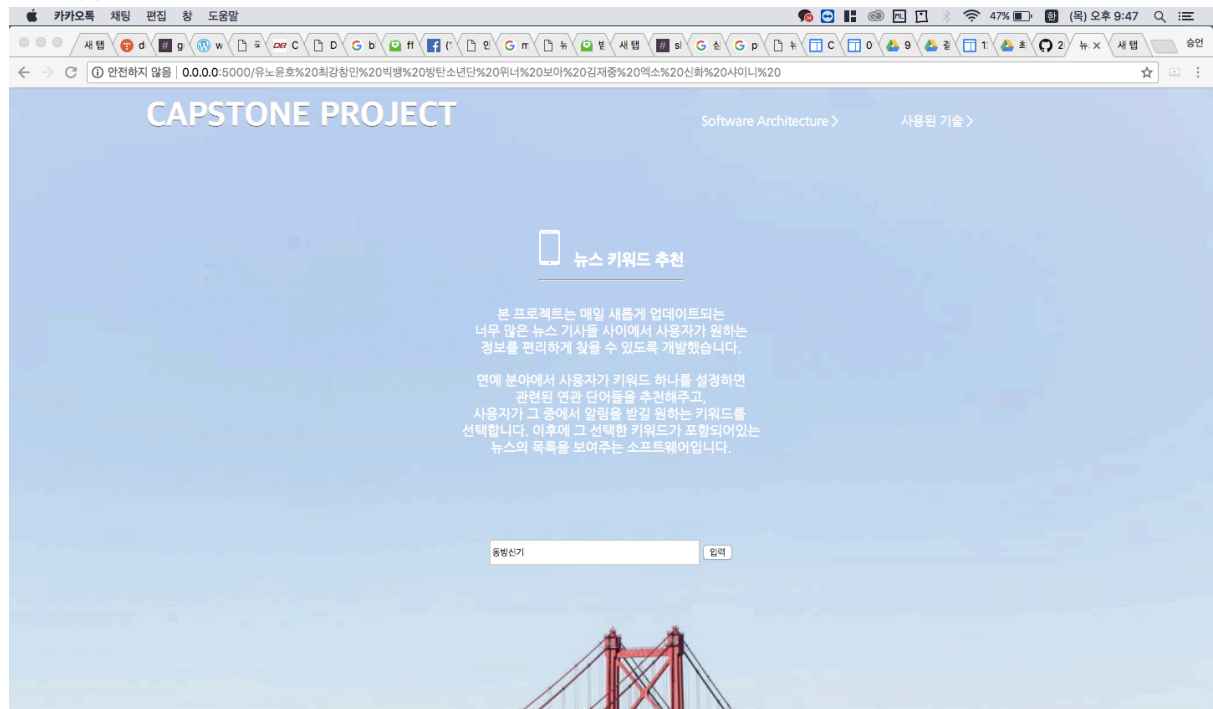
5.1 사용자 메뉴얼

- (1) 메인 화면

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |




(2) 입력 칸에 원하는 키워드를 입력하고 입력 버튼을 클릭한다.



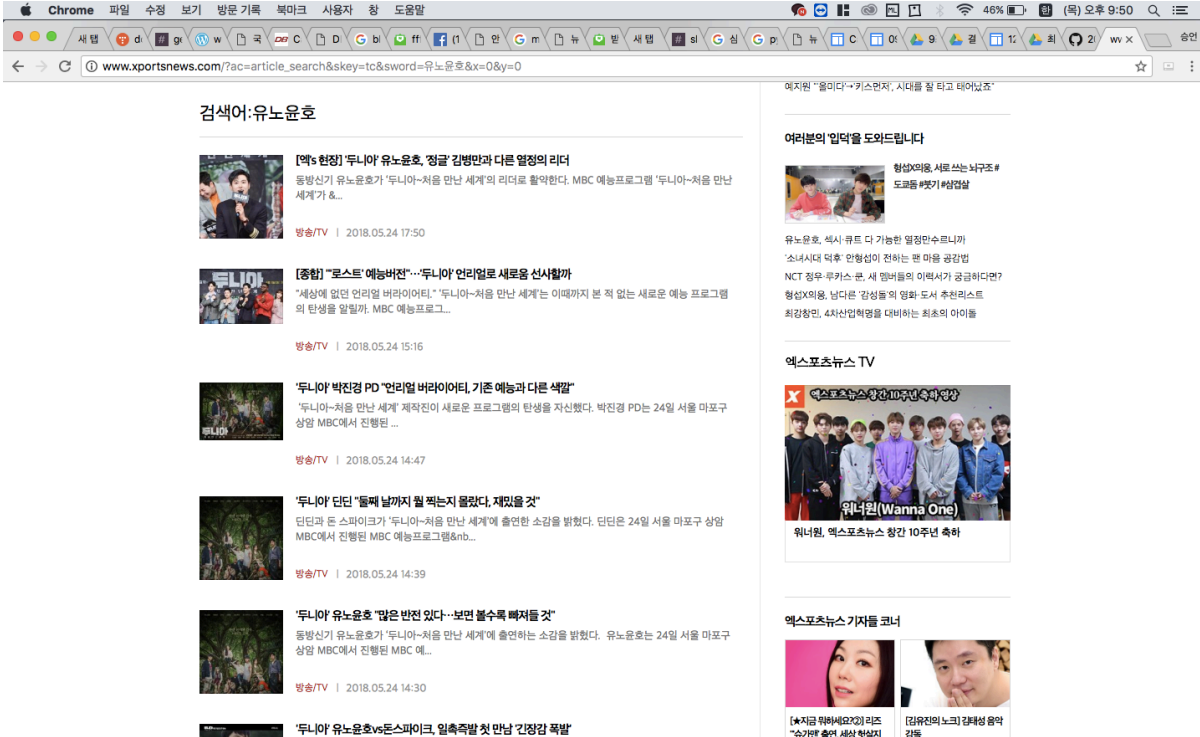
| | | | |
|--|-------------------------|-------------|-------------|
|  <div> <p>국민대학교</p> <p>컴퓨터공학부</p> <p>캡스톤 디자인 I</p> </div> | 결과보고서 | | |
| | 프로젝트 명 | 뉴스 키워드 추천 | |
| | 팀 명 | 해보자 팀 | |
| | Confidential Restricted | Version 1.0 | 2018-MAR-23 |

- (3) 워드클라우드로 연관 관계를 시각적으로 볼 수 있으며 신문사별로 키워드를 클릭하면 해당 신문사의 기사로 바로 이동할 수 있다.



The screenshot shows a web browser window with a word cloud in the center. The word cloud contains various terms related to the search, including '유노윤호', '최강창민', '빅뱅', '방탄소년단', '워너', '보아', '김재중', '엑소', '신화', '사이니', '유노', '엑스키스', '종현', '김재중', '유영진', '슈퍼주니어', '소녀시대', '완전체', '안방', '사이니', '보아', '방탄소년단', '대포미끼'. Below the word cloud, there is a section titled 'Recommend Keyword' with a list of keywords: 유노윤호, 최강창민, 빅뱅, 방탄소년단, 워너, 보아, 김재중, 엑소, 신화, 사이니. The browser's address bar shows a URL with search parameters for '유노윤호'.

- (4) 키워드를 클릭했을 때 나타나는 기사의 목록이다.



The screenshot shows a web browser window displaying search results for '유노윤호'. The results are listed on the left side of the page, each with a thumbnail image, a title, and a brief description. The titles include: '[엑스 한정] '두니아' 유노윤호, '정글' 김병민과 다른 열정의 리더', '[종합] '로스쿨' '엑스' '두니아' '언리얼'로 새로운 선상할까', '두니아' 박진경 PD '언리얼' 버라이어티, 기존 예능과 다른 색깔', '두니아' 단단 '둘째' 낳까지 될 뻔한지 몰랐다, 재밌을 것', '두니아' 유노윤호 '같은 반친 있다'--보인 불수록 빠져들 것', '두니아' 유노윤호vs트스파이크, 일족출발 첫 만난 긴장감 폭발'. The right side of the page shows a sidebar with additional content, including a section titled '여러분의 입덕을 도와드립니다' and a section titled '엑스포츠뉴스 TV'.