



국민대학교
전자정보통신대학
컴퓨터공학부

캡스톤 디자인 I

종합설계 프로젝트


프로젝트 명	뉴스 키워드 추천
팀 명	해보자
문서 제목	중간보고서

Version	1.0
Date	2018-04-10

팀원	이 승언
	이 가영
지도교수	윤 성혜 교수

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	중간보고서		
	프로젝트 명		
	팀 명		
	Confidential Restricted	Version 1.2	20xx-APR-30

문서 정보 / 수정 내역

Filename	중간보고서-뉴스 키워드 추천.doc
원안작성자	이승언
수정작업자	이승언

수정날짜	대표수정자	Revision	추가/수정 항목	내 용
2018-04-10	이승언	1.0	최초 작성	

본 양식은 컴퓨터공학부 캡스톤 디자인 I 과목의 프로젝트 중간보고서 작성을 위한 기본 양식입니다. 문서의 필수 항목을 제시하는 것이니 폰트, 문단 구조 등의 디자인 부분은 자유롭게 설정하기 바랍니다. 양식 내에 붉은 색으로 기술한 부분은 지우고 작성하기 바랍니다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명		
	팀 명		
	Confidential Restricted	Version 1.2	20xx-APR-30

목 차

1	프로젝트 목표	4
2	수행 내용 및 중간결과	5
2.1	계획서 상의 연구내용	5
2.1.1	‘일간스포츠’ 크롤러 개발	5
2.1.2	데이터 형태소 분석 및 Train 데이터 구성	5
2.1.3	키워드 추천 시각화	5
2.1.4	선별한 뉴스 기사 목록화	5
2.2	수행내용	6
2.2.1	크롤러 개발	6
2.2.2	데이터 형태소 분석 및 명사 추출	6
2.2.3	데모 웹페이지 수정	7
3	수정된 연구내용 및 추진 방향	8
3.1	수정사항	8
4	향후 추진계획	9
4.1	향후 계획의 세부 내용	9
4.1.1	데이터베이스 구축	9
4.1.2	Word cloud 생성	9
4.1.3	선정된 뉴스기사 목록화	9

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명		
	팀 명		
	Confidential Restricted	Version 1.2	20xx-APR-30

1 프로젝트 목표

인터넷 때 뉴스 기사를 크롤링한 데이터와 새롭게 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 목록화하여 보여주는 것을 목표로 한다.

BeautifulSoup 모듈로 크롤링한 데이터를 키워드 간 연관 관계 분석의 훈련 데이터로 이용한다. Konlpy 의 Twitter 모듈을 이용하여 형태소 분석한다. 형태소 분석으로 명사 추출을 하여 훈련 데이터 형식을 만든다. 이 데이터를 통해서 Gensim의 Doc2vec 모듈을 이용하여 모델을 생성하고 연관 관계 분석을 한다. Flask를 사용해서 사용자가 키워드를 입력했을 때 이벤트를 발생시킨다. 연관 관계를 Wordcloud 모듈을 이용하여 시각화하여 보여준다. Mysql을 사용하여 데이터베이스에 뉴스 기사들을 헤드라인과 본문 내용과 링크를 저장해둔다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명		
	팀 명		
	Confidential Restricted	Version 1.2	20xx-APR-30

2 수행 내용 및 중간결과

2.1 계획서 상의 연구내용

팀 내에서 맡은 부분에 대한 내용은 다음과 같다.

2.1.1 '일간스포츠' 크롤러 개발

인터넷 상에 크롤링한 데이터는 대부분 정치, 사회 관련 분야이다. 프로젝트의 현실적 제한 요소로 인해 카테고리를 연예, 스포츠로 정하여서 이 카테고리에 대해 부족한 데이터를 크롤링하기 위해 연예, 스포츠 카테고리의 기사를 집중적으로 가지고 있는 '일간스포츠'를 크롤링할 수 있는 크롤러를 개발한다.

2.1.2 데이터 형태소 분석 및 Train 데이터 구성

키워드 분석 및 추출을 목표로 하기 때문에 불필요한 데이터를 제거해야 데이터 훈련 시간도 줄어 들 수 있으므로 Konlpy 모듈의 Twitter를 이용하여 형태소 분석을 하고 명사 추출을 한다. 한 라인에 한 기사에서 추출된 명사들을 순서대로 나열해서 Train 데이터를 구성한다.

2.1.3 키워드 추천 시각화

Gensim 의 Doc2Vec 을 통해 추천된 키워드들을 워드 클라우드로 보여주기 위해 추천 키워드와 연관 관계 정도를 나타내는 벡터 값을 추출한다. 추출된 데이터들은 파이썬의 Wordcloud 모듈을 사용하여 워드 클라우드로 나타낼 수 있도록 한다.

2.1.4 선별한 뉴스 기사 목록화

뉴스 기사를 사용자에게 보여주기 위해 필요한 데이터들을 Mysql을 사용해서 데이터베이스에 저장한다. 사용자의 입력 키워드가 포함된 뉴스 기사들을 데이터베이스에서 쿼리문으로 찾아서 웹 페이지에 목록화하여 보여준다.



2.2 수행내용

2.2.1 크롤러 개발

연예, 스포츠 분야에 대한 데이터가 부족하다고 생각하여 '일간스포츠' 신문사를 크롤링하는 크롤러를 개발하였다.

2.2.2 데이터 형태소 분석 및 명사 추출

이전의 크롤링 데이터 형식 (한 라인에 한 기사씩 저장)

배우 심은경 주연의 스릴러 영화 '넬 기다리며'가 최근 촬영을 마쳤다고 이 영화 투자·배급을 맡은 뉴가 31일 밝혔다. '넬 기다리며'는 15년 전 연쇄살인범에 의해 아파를 잃은 소녀와 그녀를 보살피 온 형사, 15년 만에 세상에 나온 범인의 복잡하게 얽힌 운명을 그린 영화다. 이번 영화로 처음 스릴러에 도전한 심은경은 아파를 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 공송연과 여진규, 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 '언니네' 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.

형태소 분석 후 명사 추출을 마친 데이터 형식 (한 라인에 한 기사에 대한 명사 저장)

배우 심은경 주연 스릴러 영화 '넬' 최근 촬영 이 영화 투자 배급 뉴 일 넬 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 송연 여진규 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

이전에 크롤링해둔 데이터로 Train 데이터를 생성해보았다. . Twitter 모듈로 형태소 분석을 마친 뒤 nouns 함수로 명사 추출을 완료하였다. 이전에는 46.1MB였던 데이터가 명사 추출 후에 26.3MB로 크기가 줄었다. 키워드 분석 및 추천만을 목표로 하는 프로그램 이기때문에 훈련에 필요한 최소한의 데이터를 생성하기 위해 명사 추출을 하였다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명		
	팀 명		
	Confidential Restricted	Version 1.2	20xx-APR-30

2.2.3 데모 웹페이지 수정

팀원이 flask를 이용하여 키워드를 입력하면 결과를 나타나는 웹페이지를 만들었다. 입력 버튼을 누르면 새로운 창에 결과가 나타나는 것을 입력 창 바로 밑에 결과가 나타나도록 수정하였다. 결과 텍스트를 doc2vec.html에 템플릿으로 뿌려줄 때 넣어주었다.

[중간 결과물]

...

Doc2Vec Demo

Recommend Keyword

로딘%ㄹ 시고%ㄹ 고마츠%ㄹ 대성%ㄹ 철도우%ㄹ 산다라박
... ▼

Doc2Vec Demo

Recommend Keyword

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

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명		
	팀 명		
	Confidential Restricted	Version 1.2	20xx-APR-30

3 수정된 연구내용 및 추진 방향

3.1 수정사항

3.1.1 프로젝트 범위 수정

제안서에는 키워드를 추천하고 이메일 알림 서비스까지 제공하는 프로그램을 제작한다고 되어있으나 제작 기간 등에 적절하도록 키워드를 추천하고 관련 뉴스 기사를 보여주는 정도로 프로젝트 범위를 수정하였다. 따라서 아마존 웹 서비스를 사용하지 않는다.

매일 새롭게 뉴스 기사를 크롤링하여 좀 더 최신 정보를 담은 벡터 모델을 생성하려 했으나 서버 사용을 하지 않는 쪽으로 프로젝트 범위를 수정하여서 크롤링해둔 데이터만으로 모델을 생성해서 연관 관계를 분석한다.

3.1.2 크롤링 데이터 관련 수정사항

인턴십때 크롤링한 데이터만으로 모델을 생성하려 했으나 크롤링한 데이터가 대부분 정치, 경제 등 본 프로젝트에 필요 없는 데이터이다. 그래서 연예, 스포츠 관련 기사들을 더 수집하기 위해서 새로운 연예, 스포츠 관련 신문사를 크롤링할 수 있는 크롤러를 개발한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명		
	팀 명		
	Confidential Restricted	Version 1.2	20xx-APR-30

4 향후 추진계획

4.1 향후 계획의 세부 내용

4.1.1 데이터베이스 구축

사용자의 키워드 설정이 끝난 후 선정된 키워드들이 포함된 뉴스 기사를 데이터베이스에서 쿼리문으로 요청하여 목록화하도록 해야한다. Mysql과 파이썬, flask를 연동할 것이다. 데이터베이스에는 사용자에게 제공해야하는 헤드라인, 본문 내용 일부, 바로 그 기사로 이동할 수 있는 링크를 저장할 예정이다.

4.1.2 Word cloud 생성

사용자가 입력한 키워드에 따른 추천 키워드들을 시각화하여 보여주는 word cloud를 생성할 것이다. Word cloud는 파이썬의 Wordcloud 모듈을 사용할 것이다. 사용자가 입력하는 키워드가 들어올 때 flask에서 요청을 받으면 word cloud를 생성해서 다시 웹페이지로 보여주는 기능을 구현할 예정이다.

4.1.3 선정된 뉴스기사 목록화

데이터베이스에 쿼리문을 요청하여 Mysql과 html을 연동하여 선정된 뉴스 기사들을 사용자가 보기 편하게 UI를 설계할 예정이다.