



국민대학교  
전자정보통신대학  
컴퓨터공학부

# 캡스톤 디자인 I

## 종합설계 프로젝트

프로젝트 명	뉴스 키워드 추천
팀 명	해보자 팀
문서 제목	계획서

Version	2.0
Date	2018-04-09

이름	이승언
----	-----


### 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.

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	뉴스 키워드 추천	
	<b>팀 명</b>	해보자 팀 (23 조)	
	Confidential Restricted	Version 2.0	2018-04-09


## 문서 정보 / 수정 내역

수정날짜	대표수정 자	Revision	추가/수정 항목	내 용
2018-03-05	이승언	1.0	최초 작성	일정 및 역할분담
2018-03-08	이승언	1.1	추가	기능 수정 및 추가
2018-04-09	이승언	2.0	수정	개별 분담 내용 추가 및 수정

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	뉴스 키워드 추천	
	<b>팀 명</b>	해보자 팀 (23 조)	
	Confidential Restricted	Version 2.0	2018-04-09

## 목 차

<b>1</b>	<b>개요</b>	<b>4</b>
1.1	프로젝트 개요	4
1.2	추진 배경 및 필요성	4
<b>2</b>	<b>개발 목표 및 내용</b>	<b>5</b>
2.1	목표	5
2.2	연구/개발 내용	5
2.3	개발 결과	6
2.4	기대효과 및 활용방안	7
<b>3</b>	<b>배경 기술</b>	<b>7</b>
3.1	기술적 요구사항	7
3.2	현실적 제한 요소 및 그 해결 방안	7
3.2.1	소프트웨어	<b>오류! 책갈피가 정의되지 않았습니다.</b>
3.2.2	기타	7
<b>4</b>	<b>개발 일정 및 자원 관리</b>	<b>7</b>
4.1	개발 일정	8
4.2	일정별 주요 산출물	8

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	뉴스 키워드 추천	
	<b>팀 명</b>	해보자 팀 (23 조)	
	Confidential Restricted	Version 2.0	2018-04-09

# 1 개요

## 1.1 프로젝트 개요


스마트폰이나 태블릿 PC가 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 인터넷 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 목록화하여 보여준다.

## 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.

### 1.2.1 키워드 뉴스 시장 현황

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

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	뉴스 키워드 추천	
	<b>팀 명</b>	해보자 팀 (23 조)	
	Confidential Restricted	Version 2.0	2018-04-09

### 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 일일이 모든 키워드를 설정해야 하는 번거로움이 있다.

## 2 개발 목표 및 내용

### 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

#### 2.2.1 데이터 형태소 분석 및 Train 데이터 구성

BeautifulSoup 모듈을 이용하여 크롤링한 데이터를 한 라인에 한 기사를 적는 형태로 파일을 만든다. 키워드 분석 및 추출을 목표로 하기 때문에 불필요한 데이터를 제거해야 데이터 훈련 시간도 줄어 들 수 있으므로 Konlpy 모듈을 이용하여 형태소 분석을 하고 명사 추출을 한다.

#### 2.2.2 키워드 추천 서비스

사용자가 설정한 키워드를 중심으로 Gensim 모듈을 이용하여 연관 관계 분석한 결과를 통해서 사용자가 흥미를 가질만한 키워드를 추천해준다. 키워드 추천 형태는 지정한 키워드를 중심으로 연관 관계를 적용한 워드 클라우드로 보여준다. 워드 클라우드는 파이썬의 Wordcloud 모듈을 사용한다.

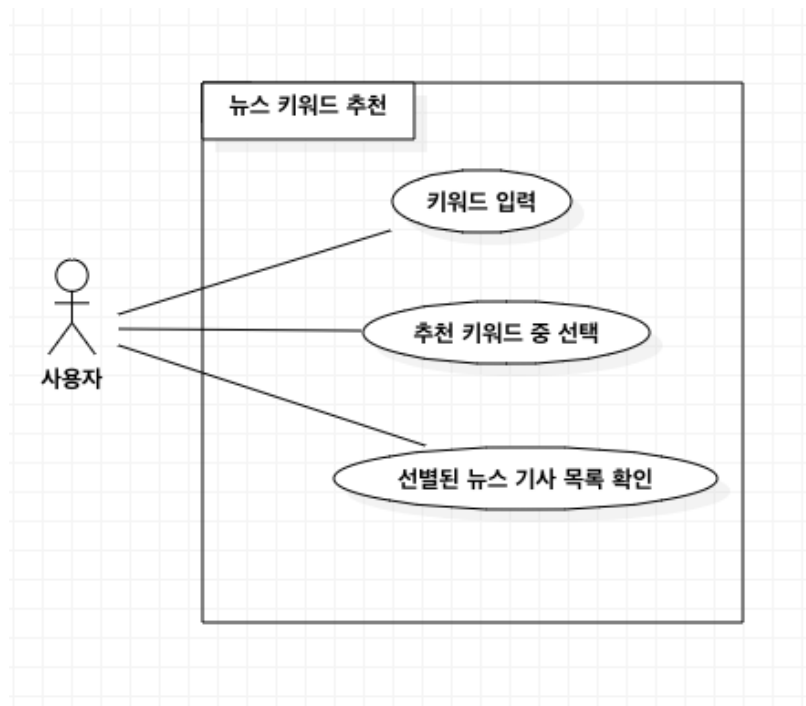
#### 2.2.3 선별한 뉴스 기사 목록화

크롤링한 뉴스 기사를 헤드라인과 본문 내용, 링크를 데이터베이스에 저장한 후 사용자의 키워드 선정이 끝나면 선정된 키워드가 포함된 뉴스 기사를 목록화하여 보여준다.

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	뉴스 키워드 추천	
	<b>팀 명</b>	해보자 팀 (23 조)	
	Confidential Restricted	Version 2.0	2018-04-09

## 2.3 개발 결과

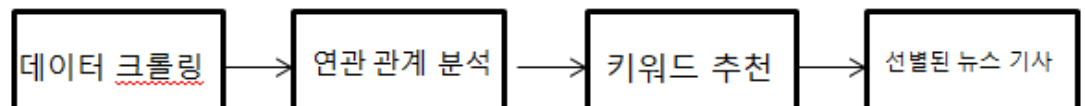
### 2.3.1 시스템 기능 요구사항




### 2.3.2 비기능(품질) 요구사항

- Performance
  - 키워드를 입력창에 입력하였을 때 추천 키워드와 워드클라우드는 2 초 이내로 나타날 수 있도록 한다.

### 2.3.3 시스템 구조(알고리즘)



- 데이터 크롤링의 output 으로는 한 기사가 한 라인에 적힌 텍스트 파일이다.
- 연관 관계 분석은 Gensim 의 Doc2Vec 모듈을 사용하여 크롤링 텍스트를 형태소 분석후 명사 추출한 데이터로 모델을 생성한다.
- 사용자의 키워드를 모델과 분석하여서 유사한 단어들을 output 으로 한다.
- 데이터베이스에 저장된 기사 중 사용자가 선정한 키워드를 포함한

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	뉴스 키워드 추천	
	<b>팀 명</b>	해보자 팀 (23 조)	
	Confidential Restricted	Version 2.0	2018-04-09

뉴스 기사를 output 으로 한다.

## 2.4 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 활용할 수 있다.

## 3 배경 기술

### 3.1 기술적 요구사항

#### - 결과물 확인

클라이언트의 PC/노트북의 window 환경

#### - 개발 IDE

Pycharm

#### - 라이브러리

Gensim, Konlpy, Wordcloud, BeautifulSoup 4.4.0

#### - 개발 언어

Python3, Html+CSS, Javascript, MySQL

#### - 기타


Git Hub

### 3.2 현실적 제한 요소 및 그 해결 방안

#### 3.2.1 기타

연예나 스포츠와 같이 관계나 소속 집단이 있는 분야에 상하관계를 적용하기가 쉽다고 생각해서 카테고리들 연예와 스포츠로 한정한다.

신문사가 매우 많아서 연예와 스포츠에 대한 기사가 많은 ‘스포츠조선’과 ‘스포츠경향’의 기사를 크롤링한다.

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	뉴스 키워드 추천	
	<b>팀 명</b>	해보자 팀 (23 조)	
	Confidential Restricted	Version 2.0	2018-04-09

## 4 개발 일정 및 자원 관리


### 4.1 개발 일정

항목	세부내용	1 월	2 월	3 월	4 월	5 월	비고
요구사항분석	아이디어 구상						
	정보 수집						
관련분야연구	주요 기술 연구						
	관련 시스템 분석						
설계	시스템 설계						
구현	코딩 및 모듈 테스트						
테스트	시스템 테스트						

### 4.2 일정별 주요 산출물

마일스톤	개요	시작일	종료일
계획서 발표	개발 환경 설치(Gensim, Konlpy, Wordcloud, Python3 설치) <b>산출물 :</b> 1. 프로젝트 수행 계획서	2018-03-02	2018-03-09
설계 완료	시스템 설계 완료 형태소 분석 관련 논문 공부 <b>산출물 :</b> 1. 시스템 설계 사양서	2018-03-10	2018-03-15
1 차 중간 보고	형태소 분석 후 명사 추출 Train data 형식 만들기 <b>산출물 :</b> 1. 프로젝트 1 차 중간 보고서 2. 프로젝트 수행 계획서 수정본 3. 1 차분 구현 소스 코드 4. Train data 텍스트 파일	2018-03-16	2018-04-13



 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	뉴스 키워드 추천	
	<b>팀 명</b>	해보자 팀 (23 조)	
	Confidential Restricted	Version 2.0	2018-04-09

2 차 중간 보고	데이터베이스에 뉴스 기사를 저장하고 키워드 포함된 뉴스 기사를 목록화하기 <b>산출물 :</b> 1. 프로젝트 2 차 중간 보고서 2. 2 차분 구현 소스 코드	2018-04-14	2018-05-18
구현 완료	시스템 구현 완료 <b>산출물:</b> 1. 최종 소스 코드 및 웹 페이지	2018-03-02	2018-05-18
테스트	시스템 통합 테스트 <b>산출물:</b> 1. 최종 소스 코드 및 웹 페이지	2018-05-19	2018-05-25
최종 보고서	최종 보고 <b>산출물:</b> 1. 최종 보고서	2018-05-26	2018-05-31