



국민대학교  
전자정보통신대학  
컴퓨터공학부


# 캡스톤 디자인 I

## 종합설계 프로젝트

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

Version	2.0
Date	2018-APR-09

이름	이 가영
----	------


 <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	
	Confidential Restricted	Version 2.0	2018-Apr-09

#### 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.


## 문서 정보 / 수정 내역

수정날짜	대표수정 자	Revision	추가/수정 항목	내 용
2018-03-05	이가영	1.0	최초 작성	
2018-03-08	이가영	1.1	내용 수정	기능 추가
2018-04-09	이가영	2.0	내용 수정	리뷰 확인 후 내용 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	
	Confidential Restricted	Version 2.0	2018-Apr-09

## 목 차

<b>1</b>	<b>개요</b>	<b>4</b>
1.1	프로젝트 개요	4
1.2	추진 배경 및 필요성	4
1.2.1	키워드 뉴스의 시장 현황	4
1.2.2	키워드 설정 기능이 개발된 시스템의 문제점, 개발할 시스템의 필요성	4
<b>2</b>	<b>개발 목표 및 내용</b>	<b>4</b>
2.1	목표	5
2.2	연구/개발 내용	5
2.2.1	뉴스 기사 크롤링	5
2.2.2	연관관계 분석	5
2.2.3	HTML 및 테스트 서버 제작	5
2.2.4	키워드가 포함된 뉴스 기사 분류 방식	6
2.3	개발 결과	6
2.3.1	시스템 기능 요구사항	6
2.3.2	시스템 구조	6
2.4	기대효과 및 활용방안	7
<b>3</b>	<b>배경 기술</b>	<b>8</b>
3.1	기술적 요구사항	8
3.2	현실적 제한 요소 및 그 해결 방안	8
3.2.1	기타	8
<b>4</b>	<b>개발 일정 및 자원 관리</b>	<b>9</b>
4.1	개발 일정	9
4.2	일정별 주요 산출물	9
<b>5</b>	<b>참고 문헌</b>	<b>10</b>

 <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	
	Confidential Restricted	Version 2.0	2018-Apr-09

# 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.

단어 간의 연관 관계를 분석한 벡터 모델을 만들고, 사용자가 입력한 키워드와 연관이 되어 있는 키워드들을 추천해줄 것이다. 이후에 사용자는 그중에서 원하는 키워드를 선택하여 그것이 포함된 뉴스 기사들을 보여주는 소프트웨어를 개발할 것이다.

## 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"> <li>- 주식 관련 정보에 특화되어 있음</li> <li>- 웹 애플리케이션</li> </ul>	<ul style="list-style-type: none"> <li>- 웹 홈페이지에서의 실시간 푸시 알림과 하루에 한 번 메일을 보내주기 때문에 편리함</li> <li>- 관심 콘텐츠에 들어가면, 관련된 콘텐츠를 지속해서 받고 싶은 경우 설정하면 효율적인 키워드를 추천해줌</li> </ul>	<ul style="list-style-type: none"> <li>- 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</li> </ul>
구글 ‘알리미’	<ul style="list-style-type: none"> <li>- 구글에서 지원하는 서비스</li> <li>- 웹 애플리케이션</li> </ul>	<ul style="list-style-type: none"> <li>- 매우 많은 데이터를 가지고 있어서 다양한 콘텐츠를 볼 수 있음</li> <li>- 수신 빈도, 출처, 언어, 지역, 검색결과와 개수, 수신 방법 등을 옵션으로 선택 가능</li> </ul>	

### 1.2.2 키워드 설정 기능이 개발된 시스템의 문제점, 개발할 시스템의 필요성

키워드 뉴스의 시장 현황에서 나타나 있는 단점들과같이, 우리는 기존에 있는 키워드 알림 시스템의 문제점을 파악했다. 따라서 산학 과제로 모아놓았던 정제된 뉴스 기사 데이터들을 사용하여 연관 관계 분석을 적용하기로 했다. 사용자가 키워드 단어 하나를 설정하면 그것과 연관된 키워드들을 추천해 주고, 사용자가 선택한 키워드가 포함된 기사 중 이를 이내로 업데이트가 된 기사들의 목록을 보여주는 소프트웨어를 개발하기로 했다.

 <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	
	Confidential Restricted	Version 2.0	2018-Apr-09

## 2 개발 목표 및 내용

### 2.1 목표

기존에 있는 키워드 알림의 기능이 있는 웹, 앱 애플리케이션과 차별화된 키워드 간의 연관 관계가 분석된 뉴스 키워드를 추천해주는 소프트웨어를 개발한다.  
즉, 사용자가 키워드 하나를 설정하면, 관련된 연관 단어들을 추천해주고, 사용자가 그중에서 알림을 받길 원하는 키워드를 선택한다. 이후에 그 키워드가 포함되어있는 뉴스 기사 중 이를 내로 업데이트가 된 기사들의 목록을 보여줄 것이다.

### 2.2 연구/개발 내용

#### 2.2.1 뉴스 기사 크롤링

뉴스 키워드 추천을 위해 훈련할 데이터는 현실적 제한 요소로 인하여 스포츠, 연예 분야로 한정시켰다. 하지만, 인턴십 때 진행했던 크롤링 데이터는 스포츠, 연예 분야가 벡터 모델을 훈련하기에 부족하기 때문에 더 많은 데이터를 모을 것이다.

#### 2.2.2 연관 관계 분석

모아놓은 신문 기사 데이터를 이용하여, 키워드 사이의 연관 관계를 분석한다. 이때 연관 관계를 분석하는 방법은 단어를 벡터화하고, 벡터 사이의 거리를 0부터 1 사이의 수치로 나타낸 후 1에 가까울수록 연관 관계가 높다고 판단할 것이다. 이때 단어를 벡터화하는 방법으로는 word2vec 기술을 이용할 것이다.


최초로 분석한 이후에는 새롭게 업데이트되는 기사의 내용에 따라서 하루에 한 번씩(자정) 연관관계도 새롭게 분석하도록 한다.

#### 2.2.3 HTML 및 테스트 환경 구축

최종적으로 사용자가 키워드를 입력할 수 있는 텍스트 입력 창과 연관된 키워드들을 보여주고, 결과적으로 키워드가 포함된 기사들의 목록을 보여줄 수 있는 HTML UI를 제작할 것이다. 또한, 사용자가 입력한 키워드에 대한 결과들을 보여주는 방법으로는 python 라이브러리인 Flask를 이용할 것이다. 벡터 모델 구축을 python 언어를 사용하기 때문에 테스트용 서버를 python Flask로 사용한다면 편리할 것이라 예상된다.

#### 2.2.4 키워드가 포함된 뉴스 기사 분류 방식

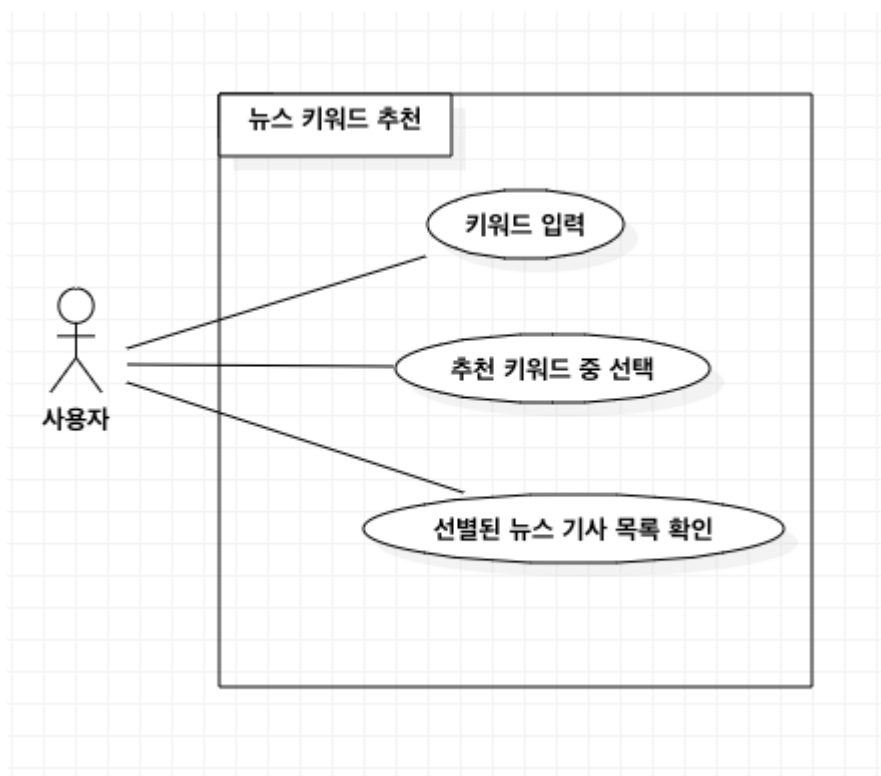
기본적으로 어떤 단어에 대해 인지할 수 있게 하기 위해서는 수치적인 방식으로 단어를 표현할 수 있어야 하므로 one-hot encoding 방식을 사용한다. 이 방법은 어떠한 단어를 표현하기 위해 길이 n 짜리 벡터를 하나 만들고, 그 단어가 해당하는 자리에 1을 넣고 나

 <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	
	Confidential Restricted	Version 2.0	2018-Apr-09

머지 자리들에는 0을 넣는 방식이다. 우리는 이 one-hot encoding 이론을 이용하여 Naïve Bayes 방식을 통해 키워드가 포함되어 있는지를 판별할 것이다. 뉴스 기사를 크롤링한 본문 데이터 전체에서 특정 키워드가 있으면 1로, 없으면 0으로 나타내는 식으로 뉴스 기사에 대한 벡터를 만든다. 따라서 1로 표시가 된 기사는 사용자가 원하는 문서라고 판단이 되므로 데이터베이스에 따로 저장해서 사용자에게 알림을 주도록 한다.

## 2.3 개발 결과


### 2.3.1 시스템 기능 요구사항

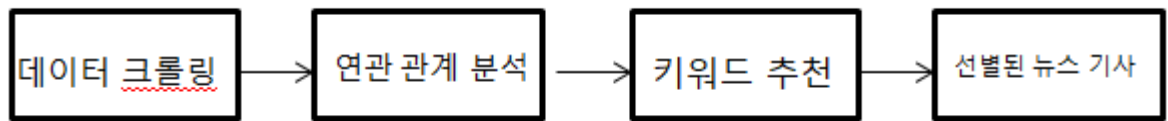


### 2.3.2 비기능(품질) 요구사항

- Performance
  - 사용자가 키워드 입력 시, word cloud 의 형태로 연관 관계가 분석된 결과를 2 초 이내로 보여주어야 한다.

### 2.3.3 시스템 구조


 <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	
	Confidential Restricted	Version 2.0	2018-Apr-09



1. 데이터 크롤링의 결과 벡터 모델이 완성된다. 사용자가 원하는 키워드를 입력하면 벡터 모델과 키워드 사이의 연관 관계 분석을 시작한다.
2. 연관 관계 분석의 결과로 사용자가 입력한 키워드와 관련이 되어 있는 키워드를 추천해준다.
3. 키워드 추천을 받은 사용자는 그중에 알람을 받길 원하는 키워드를 선택하고, 크롤러는 해당하는 키워드가 포함된 뉴스 기사를 선별하여 사용자에게 목록을 보여준다.

## 2.4 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 알람이나 중고 거래 사이트, 혹은 많은 광고 메일들 속에서 원하는 정보를 찾고 싶을 때, 등 이 기술을 다양한 목적으로 활용할 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	
	Confidential Restricted	Version 2.0	2018-Apr-09

## 3 배경 기술

### 3.1 기술적 요구사항


- 결과물 확인  
pc 및 노트북의 window 환경
- 개발 언어  
python 3 버전  
javascript, html, css  
Mysql
- 개발 IDE  
Pycharm
- 라이브러리  
BeautifulSoup4.4.0  
Konlpy  
gensim  
Word Cloud  
Flask
- 기타  
Git hub

### 3.2 현실적 제한 요소 및 그 해결 방안

#### 3.2.1 기타

- 연관 관계 분석을 할 때 그 관계가 뚜렷하지 않을 수 있다는 문제점이 있다. 따라서 그 관계가 비교적 뚜렷하게 나타나는 연예, 스포츠 분야만을 대상으로 프로젝트를 진행할 계획이다.
- 부족한 연예, 스포츠 분야의 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 '스포츠 경향', '일간 스포츠'이다.



 <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	
	Confidential Restricted	Version 2.0	2018-Apr-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	Python 3 과 gensim, Konlpy 라이브러리 설치 산출물 : 1. 프로젝트 수행 계획서	2018-03-02	2018-03-09
설계 완료	시스템 설계 완료, Doc2Vec 관련 논문 공부 산출물 : 1. 시스템 설계 사양서	2018-03-09	2018-03-15
1 차 중간 보고	Gensim Doc2Vec 라이브러리를 이용한 알고리즘 작성, 테스트 진행 Html 웹 페이지 UI 구성 산출물 : 1. 프로젝트 1 차 중간 보고서 2. 연관 관계 분석 벡터 모델 3. Html UI 4. 1 차분 구현 소스 코드 5. 프로젝트 수행 계획서 수정본	2018-03-15	2018-04-13
2 차 중간 보고	크롤링 대용량 데이터 수집, 뉴스 기사에서 키워드 포함 여부 판단 알고리즘 작성, 실제 벡터 모델 구축 산출물 : 1. 프로젝트 2 차 중간 보고서 2. 2 차분 구현 소스 코드 3. 스포츠, 연예 분야 대용량 뉴스 데이터	2018-04-13	2018-05-18
구현 완료	시스템 구현 완료 산출물 : 최종 소스 코드 및 웹 홈페이지	2018-03-02	2018-05-18
테스트	시스템 통합 테스트 산출물 : 최종 소스 코드 및 웹 홈페이지	2018-05-18	2018-05-25
최종 보고서	최종 보고 산출물: 1. 최종 보고서 2. 최종 소스코드 및 웹 홈페이지	2018-05-25	2018-05-31

 <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	
	Confidential Restricted	Version 2.0	2018-Apr-09

## 5 참고 문헌

- 문서 : word2vec 관련 이론 정리  
<https://shuuki4.wordpress.com/2016/01/27/word2vec-%EA%B4%80%EB%A0%A8-%EC%9D%B4%EB%A1%A0-%EC%A0%95%EB%A6%AC/>
- 문서 : 텐서플로를 이용해 자연어를 처리하기 – word Embedding(Word2vec)  
<http://solarisailab.com/archives/374>
- AWS 개발자 안내서 : Amazon SES API를 사용하여 이메일 보내기  
[https://docs.aws.amazon.com/ko\\_kr/ses/latest/DeveloperGuide/send-email-api.html](https://docs.aws.amazon.com/ko_kr/ses/latest/DeveloperGuide/send-email-api.html)
- 논문 : 한국어 텍스트 내 용어연관성 분석을 위한 기초연구  
<http://www.dbpia.co.kr/Journal/ArticleDetail/NODE00170700>
- 논문 : 의미 정보를 활용한 관계 추출 시스템 개발 및 성능 평가  
[http://semantics.kisti.re.kr/technical\\_reports/pdf/isbn/005.pdf](http://semantics.kisti.re.kr/technical_reports/pdf/isbn/005.pdf)
- 논문 : Model-based Measurement for Text Similarity  
<http://www.itdaily.kr/conference/image/Model-based.pdf>