



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 "키워드 추천 뉴스 알리미"를 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	키워드 추천 뉴스 알리미
팀 명	해보자
문서 제목	키워드 추천 뉴스 알리미 설계 보고서


Version	1.1
Date	2018-MAR-09

이름	이 가영
----	------

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	키워드 추천 뉴스 알리미	
	<b>팀 명</b>	해보자	
	Confidential Restricted	Version 1.1	2018-MAR-09


## 문서 정보 / 수정 내역

수정날짜	대표수정 자	Revision	추가/수정 항 목	내 용
2018-03-05	이가영	1.0	최초 작성	
2018-03-08	이가영	1.1	내용 수정	기능 추가

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	키워드 추천 뉴스 알리미	
	<b>팀 명</b>	해보자	
	Confidential Restricted	Version 1.1	2018-MAR-09

## 목 차

1	개요 .....	4
1.1	프로젝트 개요 .....	4
1.2	추진 배경 및 필요성 .....	4
1.2.1	키워드 알리미 기술의 시장 현황 .....	4
1.2.2	키워드 알리미 개발된 시스템의 문제점, 개발할 시스템의 필요성 .....	4
2	개발 목표 및 내용 .....	4
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	웹 서버 제작 .....	5
2.2.4	키워드가 포함된 뉴스 기사 분류 방식 .....	6
2.3	개발 결과 .....	6
2.3.1	시스템 기능 요구 사항 .....	오류! 책갈피가 정의되어 있지 않습니다.
2.3.2	시스템 구조 .....	오류! 책갈피가 정의되어 있지 않습니다.
2.4	기대효과 및 활용방안 .....	7
3	배경 기술 .....	8
3.1	기술적 요구사항 .....	8
3.2	현실적 제한 요소 및 그 해결 방안 .....	8
3.2.1	소프트웨어 .....	8
3.2.2	기타 .....	8
4	개발 일정 및 자원 관리 .....	9
4.1	개발 일정 .....	9
4.2	일정별 주요 산출물 .....	9
5	참고 문헌 .....	10

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	키워드 추천 뉴스 알리미	
	<b>팀 명</b>	해보자	
	Confidential Restricted	Version 1.1	2018-MAR-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. 인터넷 뉴스 기사는 매일매일 새롭게 업데이트가 되고 있다는 특징을 가지고 있기 때문에 하루에도 신문사별로 100개가 넘는 정보들이 업데이트 되고 있다. 이제는 이렇게 많은 정보들 속에서 사용자가 원하는 것을 찾아서 보여주는 것 또한 하나의 기술로 자리잡고 있다. 따라서, 우리는 매일 새롭게 업데이트가 되고 있는 뉴스 기사들 사이에서 사용자들이 원하는 정보들을 구별해서 보여주는 웹 어플리케이션을 개발하기로 했다.

## 1.2 추진 배경 및 필요성

### 1.2.1 키워드 알림 기술의 시장 현황

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

첫 번째로 ‘SNEK’은 특수하게 주식에 대해서 특화가 되어 키워드 알림이 적용이 되어 있다. 관심 있는 키워드를 저장하면 웹 홈페이지에서는 콘텐츠가 올라올 때 바로 푸시 알림으로 알려주고 있으며, 하루에 한 번씩 특정 시간에 알림을 메일로 해당 콘텐츠의 링크, 헤드라인, 출처와 함께 알려주고 있다. 해당 콘텐츠에 들어가면, 이것과 관련된 콘텐츠를 지속적으 받고 싶을 경우 설정하면 효율적인 키워드를 추천해주었다.


두 번째로 구글에서의 ‘알리미’라는 기능을 사용해 보았다. 구글에서는 수신 빈도와 출처, 언어, 지역, 검색결과와 개수, 수신 방법 등을 옵션으로 선택할 수 있었다.

세 번째로 키워드 뉴스라는 앱 어플리케이션을 사용해 보았다. 내가 관심이 있는 뉴스 키워드를 설정하면 해당하는 뉴스 콘텐츠를 알림을 주어 알려주고 있었다.

### 1.2.2 키워드 알림이 개발된 시스템의 문제점, 개발할 시스템의 필요성

실제로 현재 개발이 되어 있는 시스템을 사용해 본 결과 공통적으로 사용자가 지정한 키워드를 포함한 결과만을 보여주고 있다는 것을 알 수 있었다. 여기서 한 가지 단점을 발견했다. 한 가지의 단어를 표현하는 방법은 아주 많다는 것이다. 단어를 영어로 기술할 경우 등 같은 의미이지만 단어만 다르게 사용을 하고 있는 경우가 있기 때문에, 사용자가 원하는 정보를 최대한 많이 받기 위해서는 모든 경우의 수를 전부 키워드로 설정해 주어야 했다.

따라서 우리는 이러한 문제점을 판단하고, 산학 과제로 모아놓았던 정제된 뉴스 기사 데이터들을 사용하여 연관관계를 분석하여 키워드 알림에 적용할 계획이다. 사용자가 키워드 단어 하나를 설정하면 그것과 연관된 키워드들을 추천해 주고, 사용자가 선택한 키워드를 기반으로 해당하는 뉴스 콘텐츠가 업데이트되면 메일로 알림을 주는 시스템을 개발하기로 했다.

 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	키워드 추천 뉴스 알리미	
	<b>팀 명</b>	해보자	
	Confidential Restricted	Version 1.1	2018-MAR-09

## 2 개발 목표 및 내용

### 2.1 목표

기존에 있는 키워드 알림의 기능이 있는 웹, 앱 어플리케이션과 차별화 된 키워드 간의 연관 관계가 분석이 된 키워드 알림 어플리케이션을 개발한다.

즉, 사용자가 키워드 하나를 설정하면, 관련된 연관 단어들을 추천해주고, 사용자가 그 중에서 알림을 받길 원하는 키워드를 선택한다. 이후에 그 키워드가 포함되어있는 뉴스 기사를 사용자가 설정한 시간에 하루에 한 번씩 메일로 목록을 보여준다.

### 2.2 연구/개발 내용

#### 2.2.1 뉴스 기사 크롤링

기본적으로 파이썬 Beautiful Soup 모듈을 이용하여 개발한 크롤링 소프트웨어를 이용해 매일매일 업데이트가 되는 인터넷 신문 기사 데이터를 가져온다. 데이터베이스에 저장되어 있는 키워드들이 신문 기사의 본문에 포함되어 있다면 가져오는 헤드라인, 본문, 링크 url, 출처 데이터를 가져와서 사용자에게 결과물로 보여준다.

#### 2.2.2 연관관계 분석

산학 과제 업무 중 모아놓았던 신문 기사 데이터를 이용하여, 키워드 사이의 연관관계를 분석한다. 이 때 연관관계를 분석하는 방법은 단어를 벡터화하고, 벡터 사이의 거리를 0 부터 1 사이의 수치로 나타낸 후 1에 가까울수록 연관 관계가 높다고 판단할 것이다. 이 때 단어를 벡터화하는 방법으로는 word2vec 기술을 이용할 것이다.

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

#### 2.2.3 웹 서버 제작


서버로는 Amazon Web Service의 EC2를 사용한다.

사용자가 웹 홈페이지에 접속을 하면, 키워드를 설정할 수 있도록 한다. 이후에 미리 분석해 놓은 연관관계를 사용자가 지정한 키워드를 중심으로 보여주고, 연관된 단어들을 추천해준다. 그러면 사용자는 생각을 하지 못했던 관심 키워드에 대해서 추천을 받을 수 있고, 연관 키워드라 해도 관심이 없는 키워드는 삭제를 할 수 있다. 이후에 사용자가 선택한 키워드들은 Amazon RDS 중 MySQL을 사용하여 저장한다.

웹 홈페이지에서는 사용자가 실시간으로 알림을 받아 볼 수 있는 버튼을 만든다. 버튼을 누르면 사용자가 지정한 키워드와 관련된 콘텐츠들의 목록을 받아볼 수 있다.

또 주기적으로 사용자가 알림을 받을 수 있도록 모아놓은 콘텐츠들의 목록을 메일로 보내주도록 한다. 이 주기는 사용자가 옵션으로 설정할 수 있도록 한다. 메일을 보낼 때에는 Amazon EC2 인스턴스에서 아마존 SES API를 통해 이메일을 전송한다. 이 때의 이메일에는 크롤링을 통해 가져온 데이터 중 헤드라인, 링크 URL, 출처 데이터를 보여준다.



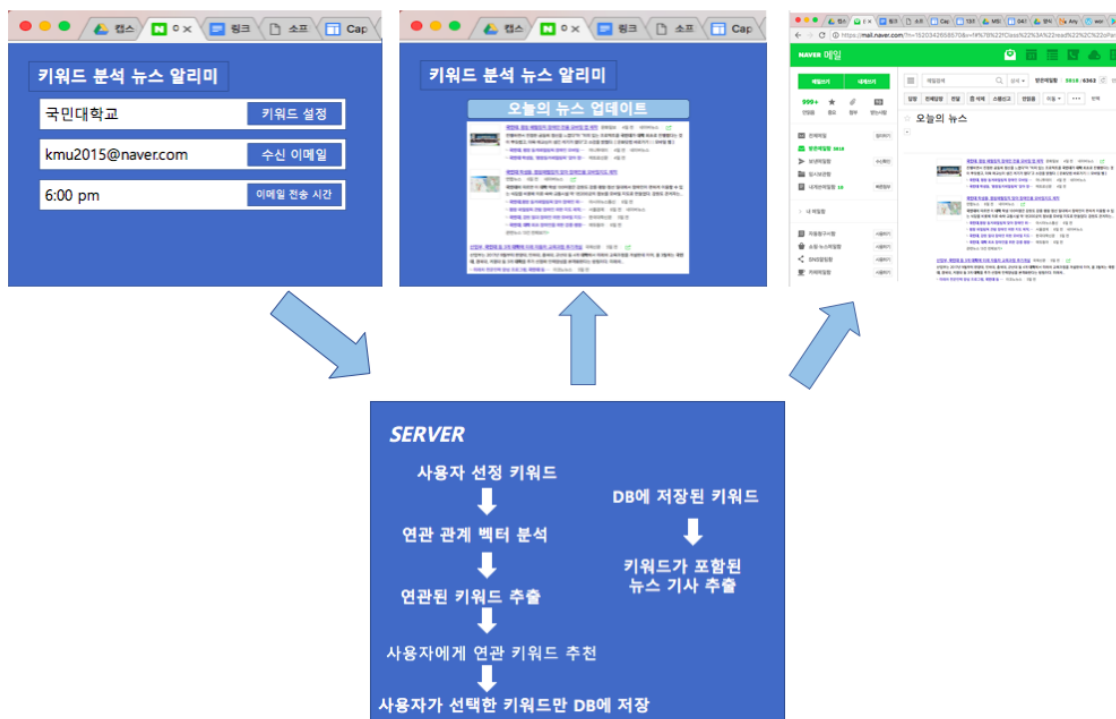
 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	키워드 추천 뉴스 알리미	
	<b>팀 명</b>	해보자	
	Confidential Restricted	Version 1.1	2018-MAR-09

하나의 단어 설정 만으로도 사용자가 원하는 최대한의 정보를 받을 수 있다. 그리고 하루에 한 번 사용자가 원하는 시간에 메일로 알림을 받을 수 있다.

- 유지보수성


연관 관계는 최초 한 번 분석을 하고 나면, 계속해서 변화한다는 특징이 있다. 특히, 우리가 조사할 연예, 스포츠 분야에서는 그 특징이 더욱 잘 드러날 것으로 예상된다. 따라서 하루에 한 번씩 모은 데이터를 통하여 새롭게 연관관계를 분석해서 업데이트를 할 계획이다.

### 2.3.3 시스템 구조



## 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. 웹 홈페이지에서의 실시간 알림과 메일로 주기적인 알림을 사용자가 선택적으로 받아 볼 수 있기 때문에 편의성을 높일 수 있다.
3. 뉴스 기사 이외에도 학교의 공지사항 알림이나 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>	키워드 추천 뉴스 알리미	
	<b>팀 명</b>	해보자	
	Confidential Restricted	Version 1.1	2018-MAR-09

## 3 배경 기술

### 3.1 기술적 요구사항

- 결과물 확인  
pc 및 노트북의 window 환경
- Server  
Apache
- 개발 언어  
python 3 버전  
php, html, css
- 개발 IDE  
Pycharm
- 라이브러리  
Tensorflow 1.4.1  
BeautifulSoup4.4.0
- DB  
Mysql
- 기타  
Git hub

### 3.2 현실적 제한 요소 및 그 해결 방안

#### 3.2.1 소프트웨어


다른 키워드 알리미 어플리케이션을 사용했을 때에도 발생했던 문제로, 제품 이름, 혹은 프로그램 이름 등의 대명사에 사용자가 원하는 키워드가 포함되어 있을 때, 혹은 동음이의어인 경우에는 그것이 사용자가 원하는 내용인지 아닌지 판별을 할 수가 없다. 이것은 우리가 계획한 연관관계 분석 알고리즘에서도 키워드 설정을 문장이 아닌 단어로 하기 때문에 문맥에서의 단어의 의미를 파악할 수 없어서 해결을 할 수가 없었다. 따라서 우리는 연관 관계를 분석한 것을 바탕으로 사용자가 지정한 키워드와 연관된 단어들을 추천해주는 것으로 사용자가 원하지 않은 정보들을 판별할 수 있도록 할 계획이다.

또한 이 전부터 인터넷 뉴스 기사를 크롤링하여 데이터를 모으는 데도 몇 가지 한계를 느꼈다. 첫 번째로 각 신문 기사마다 html 태그가 제외된 순수한 텍스트를 가져오기 위해서는 하나의 소스 코드를 사용하지만 그 안에서 html 태그의 패턴을 분석하여 일부 코드를 수정해야 하는 문제점이 있다. 두 번째로 기자의 이름, 문자를 제외한 기호 등 정제된 말뭉치를 획득하기 위해서도 신문사마다 다른 특징을 보이고 있기 때문에 패턴을 분석하고 코드를 수정해야 하는 문제점이 있다. 이러한 문제점으로 우리는 지금까지 패턴을 분석한 국민일보, 중앙일보, 이투데이, 아이뉴스이 4개의 신문사 데이터만으로 프로젝트를 진행할 계획이다.

#### 3.2.2 기타

연관관계 분석을 할 때 그 관계가 뚜렷하지 않을 수 있다는 문제점이 있다. 따라서 그 관계가 비교적 뚜렷하게 나타나는 연예, 스포츠 분야만을 대상으로 프로젝트를 진행할 계획이다.



 <b>국민대학교</b> <b>컴퓨터공학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
	<b>프로젝트 명</b>	키워드 추천 뉴스 알리미	
	<b>팀 명</b>	해보자	
	Confidential Restricted	Version 1.1	2018-MAR-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 Tensorflow 및 BeautifulSoup4 라이브러리 설치 산출물 : 1. 프로젝트 수행 계획서 2. 프로젝트 기능 일람표	2018-03-02	2018-03-09
설계 완료	개발 환경 완성 (AWS 서버 런치) 시스템 설계 완료 산출물 : 1. 시스템 설계 사양서	2018-03-09	2018-03-15
1 차 중간 보고	연관관계 분석 구현 완료, 서버 구축 및 웹 홈페이지 제작 완료 산출물 : 1. 프로젝트 1 차 중간 보고서 2. 프로젝트 진도 점검표 3. 1 차분 구현 소스 코드	2018-03-15	2018-04-13
2 차 중간 보고	추천 시스템과 키워드 알람 구현 완료, 서버와 웹 홈페이지 안정화, 메일 알람 서비스 구현 완료 산출물 : 1. 프로젝트 2 차 중간 보고서 2. 2 차분 구현 소스 코드 3. 프로젝트 성능 점검표	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>	키워드 추천 뉴스 알리미	
	<b>팀 명</b>	해보자	
	Confidential Restricted	Version 1.1	2018-MAR-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>