



국민대학교  
전자정보통신대학  
컴퓨터공학부

# 캡스톤 디자인 I

## 종합설계 프로젝트

|        |               |
|--------|---------------|
| 프로젝트 명 | 키워드 추천 뉴스 알리미 |
| 팀 명    | 해보자 팀         |
| 문서 제목  | 계획서           |

|         |             |
|---------|-------------|
| Version | 1.0         |
| Date    | 2018-MAR-05 |

|    |     |
|----|-----|
| 이름 | 이승언 |
|----|-----|

### 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.

|   |                         |               |             |
|---|-------------------------|---------------|-------------|
|  <b>국민대학교</b><br><b>컴퓨터공학부</b><br><b>캡스톤 디자인 I</b> | <b>계획서</b>              |               |             |
|   | <b>프로젝트 명</b>           | 키워드 추천 뉴스 알리미 |             |
|   | <b>팀 명</b>              | 해보자 팀 (23 조)  |             |
|   | Confidential Restricted | Version 1.0   | 2018-MAR-05 |


## 문서 정보 / 수정 내역

| 수정날짜       | 대표수정자 | Revision | 추가/수정 항목 | 내 용       |
|------------|-------|----------|----------|-----------|
| 2018-03-05 | 이승언   | 1.0      | 최초 작성    | 일정 및 역할분담 |
|            |       |          |          |           |
|            |       |          |          |           |
|            |       |          |          |           |
|            |       |          |          |           |
|            |       |          |          |           |
|            |       |          |          |           |

|   |                         |               |             |
|---|-------------------------|---------------|-------------|
|  <b>국민대학교</b><br><b>컴퓨터공학부</b><br><b>캡스톤 디자인 I</b> | <b>계획서</b>              |               |             |
|   | <b>프로젝트 명</b>           | 키워드 추천 뉴스 알리미 |             |
|   | <b>팀 명</b>              | 해보자 팀 (23 조)  |             |
|   | Confidential Restricted | Version 1.0   | 2018-MAR-05 |

## 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>8</b> |
| 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        |
| <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>국민대학교</b><br><b>컴퓨터공학부</b><br><b>캡스톤 디자인 I</b> | <b>계획서</b>              |               |             |
|   | <b>프로젝트 명</b>           | 키워드 추천 뉴스 알리미 |             |
|   | <b>팀 명</b>              | 해보자 팀 (23 조)  |             |
|   | Confidential Restricted | Version 1.0   | 2018-MAR-05 |

# 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 기사거리가 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 웹페이지와 메일을 통하여 쉽고 편리하게 볼 수 있는 것을 제공한다.

## 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 알림 서비스를 해주면 그 시간들을 절약할 수 있다.

### 1.2.1 키워드 알림 뉴스 시장 현황

- 키워드뉴스  
구글 플레이에서 제공하는 안드로이드 어플리케이션이다. 내가 관심 있어하는 키워드를 설정해서 키워드 뉴스만 모아서 볼 수 있다.
- Google 알리미  
구글에서 제공하는 서비스이다. 관심을 가지고 있는 키워드 검색어를 수신 빈도, 출처, 언어, 개수를 설정해서 이메일로 소식을 받을 수 있다.
- SNEK  
투자를 위한 금융 리서치 플랫폼으로 자신만의 키워드로 구성된 카테고리를 만들어 시장에서 발생하는 뉴스를 즉시 확인할 수 있다.

### 1.2.2 키워드 알림 기능이 개발된 프로그램의 한계

키워드 알림 기능이 있는 두가지 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 설정해야하는 번거로움이 있다.

|   |                         |               |             |
|---|-------------------------|---------------|-------------|
|  <b>국민대학교</b><br><b>컴퓨터공학부</b><br><b>캡스톤 디자인 I</b> | <b>계획서</b>              |               |             |
|   | <b>프로젝트 명</b>           | 키워드 추천 뉴스 알리미 |             |
|   | <b>팀 명</b>              | 해보자 팀 (23 조)  |             |
|   | Confidential Restricted | Version 1.0   | 2018-MAR-05 |

## 2 개발 목표 및 내용

### 2.1 목표

방대한 뉴스 기사들 속에서 관심있는 부분의 키워드를 별도로 일일이 설정하지 않고 하나의 키워드 설정만으로 키워드간의 연관 관계를 분석한 결과로 새롭게 선정된 여러개 키워드를 사용자에게 추천하고 흥미있는 키워드만 클릭해서 설정한다. 설정한 키워드를 포함한 뉴스기사들은 웹페이지에서 한눈에 볼 수 있고 메일로 알림을 받아볼 수 있는 서비스를 제공한다.

### 2.2 연구/개발 내용

#### 2.2.1 뉴스 기사 크롤링

인턴 업무 중 파이썬 Beautiful soup 모듈을 이용해서 개발한 크롤러를 이용하여 각 신문사의 뉴스 기사들을 실시간으로 크롤링한다. 크롤링 데이터 형식은 헤드라인, 본문 내용, 해당 기사의 url 을 신문사별로 텍스트 파일로 저장한다. 각 신문사마다 기사가 업데이트 될때마다 크롤링하여 데이터를 쌓는다.

#### 2.2.2 연관 관계 분석

크롤링한 데이터의 단어들을 word2vec 방법으로 벡터화 시킨 후 벡터간 거리로 키워드와의 연관 관계를 설정한다. 연관 관계가 높으면 1 에 가까운 값을 가지고 연관 관계가 낮으면 0 에 가까운 값을 가지므로 단어 벡터간 거리가 1 에 가까운 값일 때 연관 관계 키워드로 포함시킨다. 그리고 하루에 한번 새로 크롤링된 데이터를 사용해서 위와 같은 방법으로 연관 관계 분석을 한다.

#### 2.2.3 키워드 추천 서비스

사용자가 설정한 키워드를 중심으로 연관 관계 분석한 결과를 이용해서 사용자가 흥미를 가질만한 키워드를 추천해준다. 키워드 추천 형태는 지정한 키워드를 중심으로 연관 관계를 적용한 워드 클라우드로 보여준다. 그리고 그 중 사용자가 관심있는 키워드만 선택해서 앞으로 그 키워드에 관한 뉴스 기사만 알림으로 서비스해준다.

#### 2.2.4 웹 서버 제작

계속해서 크롤링한 데이터를 통하여 뉴스 기사의 키워드 간의 연관 관계 분석을 한다. 이것을 위한 서버로 Amazon Web Service 의 EC2(클라우드 상의 가상 컴퓨터 - ubuntu)를 사용한다. 데이터베이스는 Amazon RDS 중 MySQL 을 사용하여 기존의 단어간 벡터와 새로 분석된 단어간 벡터를 통해 키워드를 포함하여 키워드와 연관된 새로운 키워드들을 담아둔다.

#### 2.2.5 웹 이메일 보내기 서비스

Amazon EC2 인스턴스에서 Amazon SES API 를 통해 이메일을 전송한다. 새로 업데이트된 기사를 사용자가 설정한 이메일로 헤드라인과 간단한 정보를 담은 형식의 템플릿으로 보내준다. 헤드라인을 클릭하면 해당 기사로 바로 연결할 수 있도록 한다.

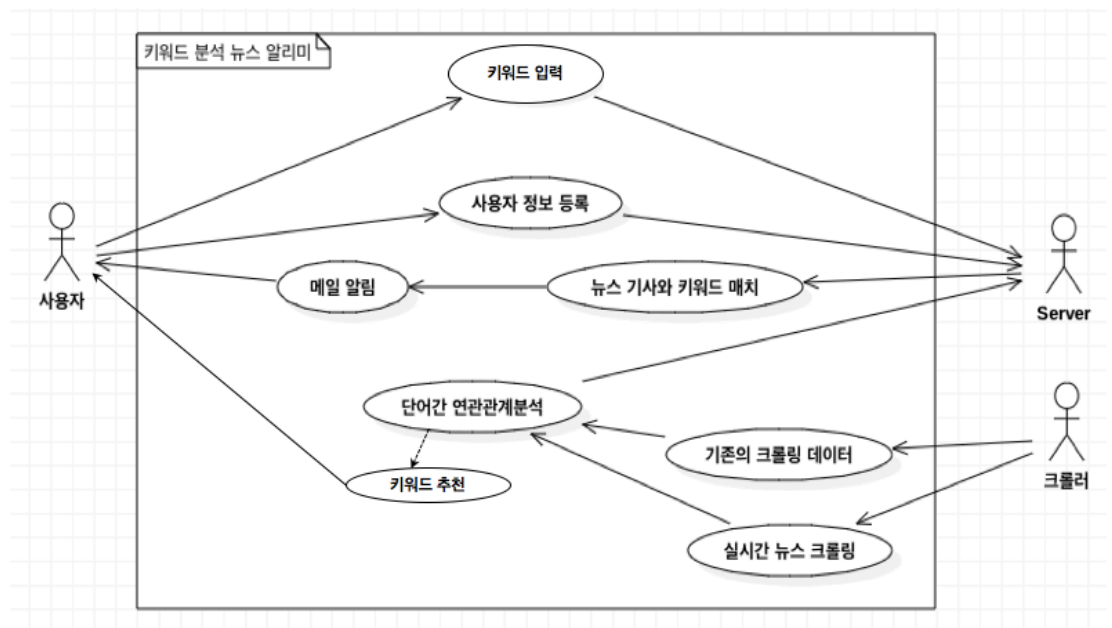
|  |                         |               |             |
|--|-------------------------|---------------|-------------|
|  <div> <b>국민대학교</b><br/> <b>컴퓨터공학부</b><br/> <b>캡스톤 디자인 I</b> </div> | <b>계획서</b>              |               |             |
|  | <b>프로젝트 명</b>           | 키워드 추천 뉴스 알리미 |             |
|  | <b>팀 명</b>              | 해보자 팀 (23 조)  |             |
|  | Confidential Restricted | Version 1.0   | 2018-MAR-05 |

### 2.2.6 키워드가 포함된 뉴스 기사 분류 방식

컴퓨터가 어떤 단어에 대해 인지할 수 있게 하기위해서 수치적인 방식으로 단어를 표현하는데 ‘one-hot encoding’ 방식으로 어떤 단어를 표현하기 위해 만들어진 벡터에 해당 단어는 1로 표현하고 나머지는 0으로 표현한다. Naive Bayes 는 ‘one-hot encoding’ 방식을 기반으로 벡터를 만드는 방법이다. 기존의 스팸 분류기가 이 방법을 이용하여 이메일 전체를 보면서 어떤 단어가 있으면 1, 없으면 0으로 나타내는 식으로 하나에 대한 벡터를 만들었다. 이 방식을 키워드 알리미 뉴스에 적용시켜서 설정한 키워드 포함해서 연관 관계가 있는 키워드가 있는 뉴스기사에 대한 벡터를 만들어서 알리미 서비스를 해준다.

## 2.3 개발 결과


### 2.3.1 시스템 기능 요구사항



### 2.3.2 비기능(품질) 요구사항

3 가지의 비기능 요구사항이 있으며 아래에 나열된 순서대로 우선순위를 정하였다.

- 안정성
  - 뉴스 기사 크롤링할때 신문사 홈페이지에 짧은 시간 동안 반복적으로 너무 많은 요청을 보내면 악의적인 행위로 인식해서 서버의 IP 를 차단당해서 서비스에 문제가 생길 수 있으므로 크롤링 요청 시간 간격을 랜덤으로 0~10 초로 설정한다.
- 용이성
  - 하루에 한번 사용자가 원하는 시간에 메일을 받아볼 수 있고 그 외에도 사용자가 원할때마다 홈페이지 업데이트 버튼을 통해서 기사를 볼 수 있다.

|   |                         |               |             |
|---|-------------------------|---------------|-------------|
|  <b>국민대학교</b><br><b>컴퓨터공학부</b><br><b>캡스톤 디자인 I</b> | <b>계획서</b>              |               |             |
|   | <b>프로젝트 명</b>           | 키워드 추천 뉴스 알리미 |             |
|   | <b>팀 명</b>              | 해보자 팀 (23 조)  |             |
|   | Confidential Restricted | Version 1.0   | 2018-MAR-05 |

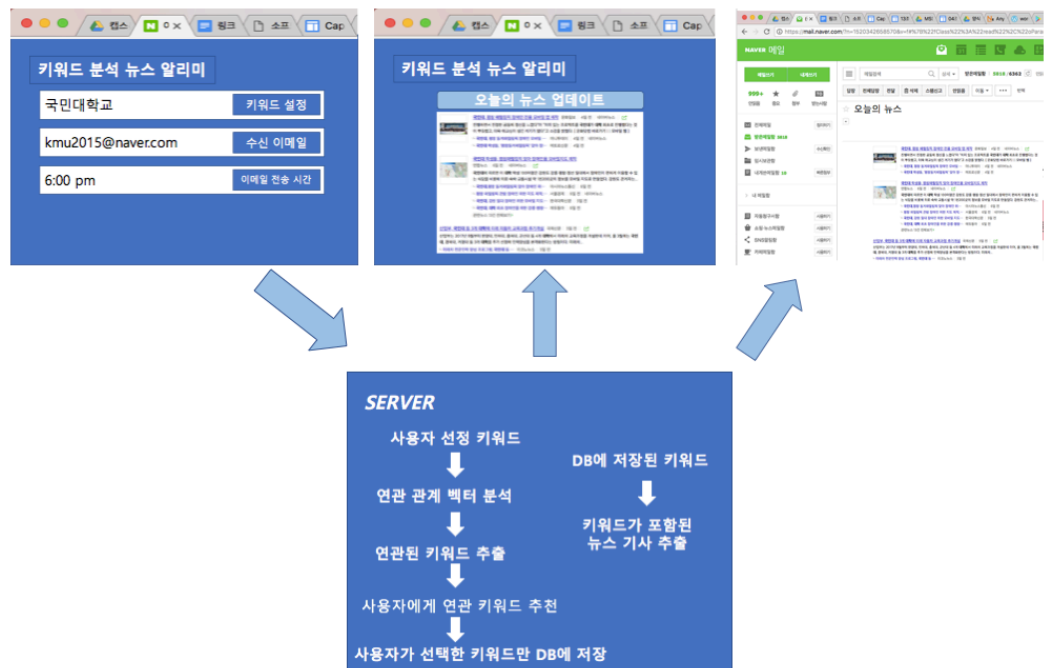
- 하나의 키워드 설정만으로 그와 관련된 흥미로운 기사들을 볼 수 있다.

- 유지보수성

- 하루동안 축적된 데이터를 이용해서 연관 관계를 반복적으로 분석하기때문에 새로 업데이트된 의미 관계로 키워드가 DB 에 업데이트되면서 새로 변화된 정보에 적응할 수 있다.

### 2.3.3

### 시스템 구조



## 2.4 기대효과 및 활용방안

1. 주식이나 비트코인 등 실시간 시세를 빠르게 알고 싶을 때마다 검색어를 검색 엔진에 타이핑하고 검색결과 중 다양한 주가나 비트코인 시세와 관련된 기사 속에서 원하는 정보를 찾지않아도 관심 있는 주가 시세 정보나 비트코인 키워드를 설정해서 시간과 비용을 절약할 수 있다.
2. 이메일 알림이 오는 시간을 설정할 수 있어서 출근 시간이나 퇴근 시간 등 스마트폰 이용하면서 잠잠이 뉴스 기사를 보고 싶을 때 기사를 받아볼 수 있다.
3. 키워드 분석 알림 서비스를 뉴스 기사 이외에도 일일이 확인하지 않으면 정보를 놓칠 수 있는 학교의 공지사항이나 게시판, 중고 거래 사이트에 이 기술을 이용한 서비스를 활용할 수 있다.
4. 광고 메일이 많이 와서 메일 알림을 켜두면 하루에도 수백 개의 알림을 받게 되어서 알림을 꺼두는 경우가 있다. 그러면 정작 필요한 정보가 있는 메일의 알림까지 켜두지 못하게 되는데 이 키워드 분석 알림 서비스를 이용하면 원하는 정보와 관련 있는 정보를 가진 메일까지 알림을 받을 수 있어서 편의성을 제공할 수 있다.

|   |                         |               |             |
|---|-------------------------|---------------|-------------|
|  <b>국민대학교</b><br><b>컴퓨터공학부</b><br><b>캡스톤 디자인 I</b> | <b>계획서</b>              |               |             |
|   | <b>프로젝트 명</b>           | 키워드 추천 뉴스 알리미 |             |
|   | <b>팀 명</b>              | 해보자 팀 (23 조)  |             |
|   | Confidential Restricted | Version 1.0   | 2018-MAR-05 |

## 3 배경 기술

### 3.1 기술적 요구사항

- **결과물 확인**  
클라이언트의 PC/노트북의 window 환경
- **개발 IDE**  
Pycharm
- **라이브러리**  
Tensorflow 1.4.1, BeautifulSoup 4.4.0
- **개발 언어**  
Python3, PHP, Html+CSS
- **DB**  
Mysql
- **Server**  
AWS EC2
- **기타**  
Git Hub

### 3.2 현실적 제한 요소 및 그 해결 방안

#### 3.2.1 소프트웨어

신문사마다 헤드라인과 본문 내용을 담고 있는 HTML 태그가 모두 달라서 HTML 태그를 이용해서 크롤링해오는데 각각의 패턴을 적용해서 통일된 크롤러를 만드는 것이 어려움이 있다. 그래서 신문사 개수를 한정적으로 4 개로 제한한다. 또 키워드를 단어로 설정했기 때문에 뉴스 기사에 나타난 단어가 동음이의어일 경우 의미를 인식할 수 없으므로 판별할 수 없어서 사용자가 원하는 키워드를 직접 지정할 수 있게 하였다.

#### 3.2.2 기타

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



|   |                         |               |             |
|---|-------------------------|---------------|-------------|
|  <b>국민대학교</b><br><b>컴퓨터공학부</b><br><b>캡스톤 디자인 I</b> | <b>계획서</b>              |               |             |
|   | <b>프로젝트 명</b>           | 키워드 추천 뉴스 알리미 |             |
|   | <b>팀 명</b>              | 해보자 팀 (23 조)  |             |
|   | Confidential Restricted | Version 1.0   | 2018-MAR-05 |


## 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(텐서플로우 및 Beautiful Soup 라이브러리 설치)<br><b>산출물 :</b><br>1. 프로젝트 수행 계획서<br>2. 프로젝트 기능 일람표                                     | 2018-03-02 | 2018-03-09 |
| 설계 완료     | 개발 환경 완성(AWS 서버 런치, 개발 툴 설정)<br>시스템 설계 완료<br><b>산출물 :</b><br>1. 시스템 설계 사양서  | 2018-03-10 | 2018-03-15 |
| 1 차 중간 보고 | 연관관계 분석 구현 완료, 서버 구축 및 웹 홈페이지 제작 완료<br><b>산출물 :</b><br>1. 프로젝트 1 차 중간 보고서<br>2. 프로젝트 진도 점검표<br>3. 1 차분 구현 소스 코드                   | 2018-03-16 | 2018-04-13 |
| 2 차 중간 보고 | 추천 시스템과 키워드 알림 구현 완료, 서버와 웹 홈페이지 안정화, 메일 알림 서비스 구현 완료<br><b>산출물 :</b><br>1. 프로젝트 2 차 중간 보고서<br>2. 프로젝트 성능 점검표<br>3. 2 차분 구현 소스 코드 | 2018-04-14 | 2018-05-18 |

|   |                         |               |             |
|---|-------------------------|---------------|-------------|
|  <b>국민대학교</b><br><b>컴퓨터공학부</b><br><b>캡스톤 디자인 I</b> | <b>계획서</b>              |               |             |
|   | <b>프로젝트 명</b>           | 키워드 추천 뉴스 알리미 |             |
|   | <b>팀 명</b>              | 해보자 팀 (23 조)  |             |
|   | Confidential Restricted | Version 1.0   | 2018-MAR-05 |

|        |  |            |            |
|--------|--|------------|------------|
| 구현 완료  | 시스템 구현 완료<br><b>산출물:</b><br>1. 최종 소스 코드 및 웹 페이지  | 2018-03-02 | 2018-05-18 |
| 테스트    | 시스템 통합 테스트<br><b>산출물:</b><br>1. 최종 소스 코드 및 웹 페이지 | 2018-05-19 | 2018-05-25 |
| 최종 보고서 | 최종 보고<br><b>산출물:</b><br>1. 최종 보고서                | 2018-05-26 | 2018-05-31 |