

EBOOK 추천 어플리케이션

# 문학의 숲

CODE 98

정현주, 박철훈, 신 궁



# CONTENTS

01

개요

- 어플리케이션  
소개

02

앱 소개

- UI 디자인  
- 기능

03

크롤링

- 책 제목 코드  
- 책 소개 및  
출판사 리뷰  
- 책 표지 이미지

04

전처리

- 토큰화

05

보완점

- 개선사항

01

## 개요

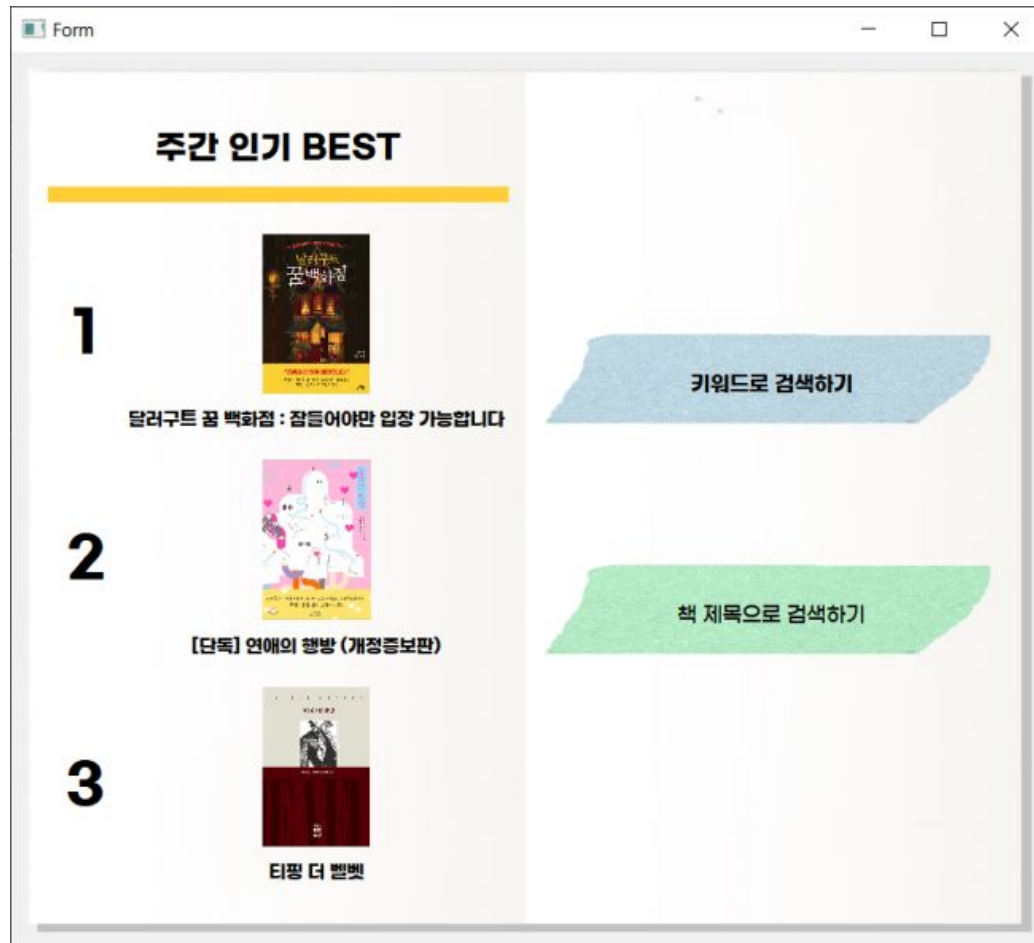
### 어플리케이션 소개

# “문학의 숲”

YES24 Ebook 소설의 책 소개 및 출판사 리뷰를 기반으로  
키워드 혹은 책 제목으로 검색하여 유사한 책을 추천해주는  
어플리케이션

## 앱 소개

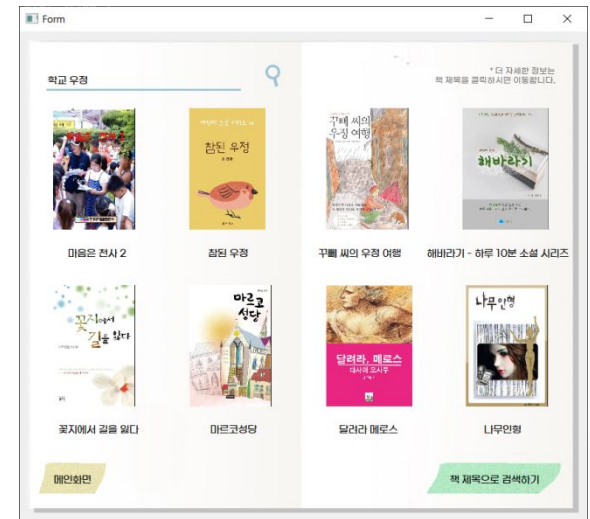
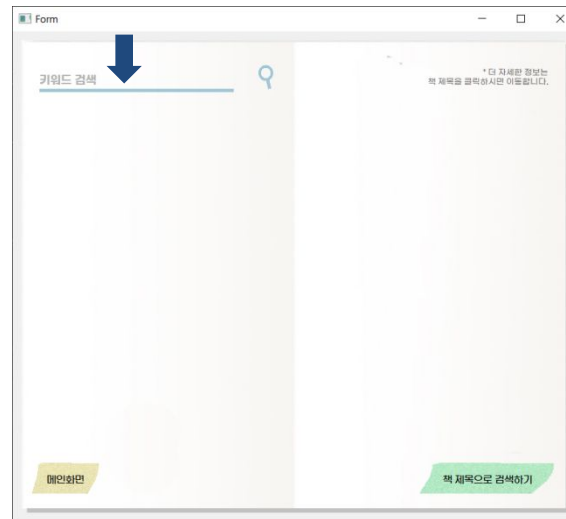
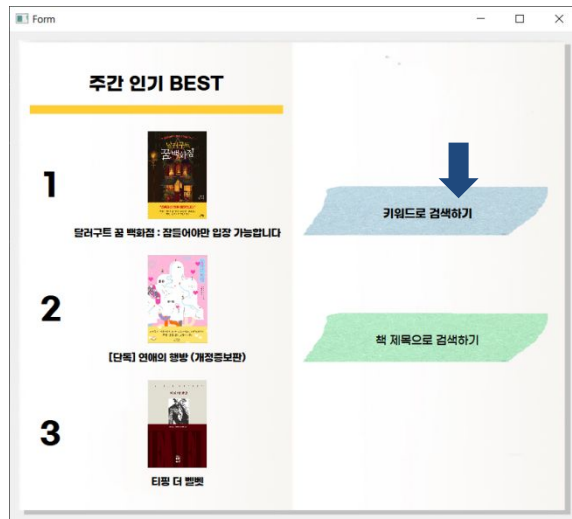
## 메인



## 02

## 앱 소개

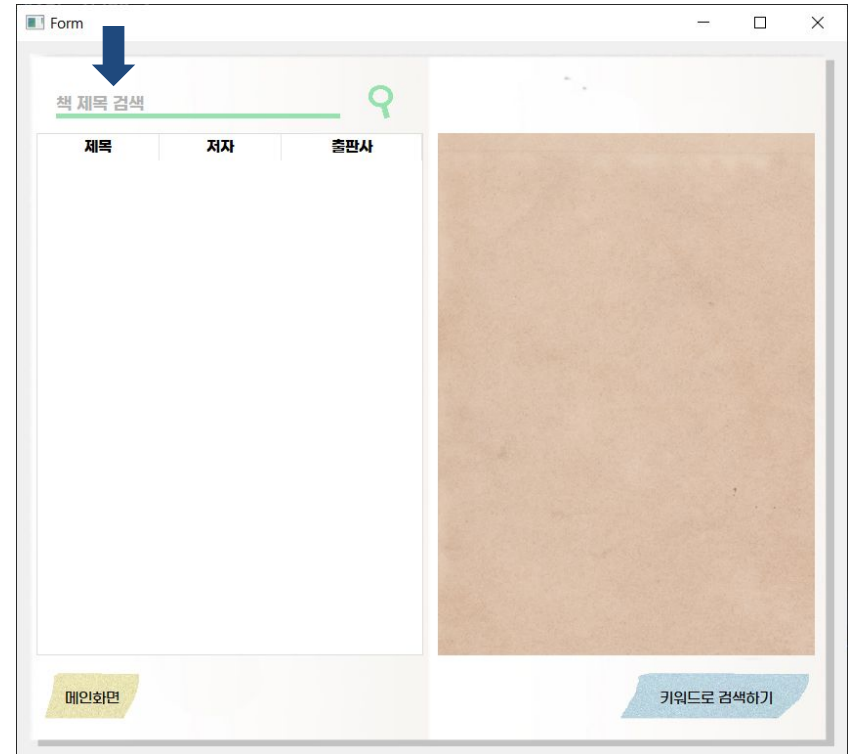
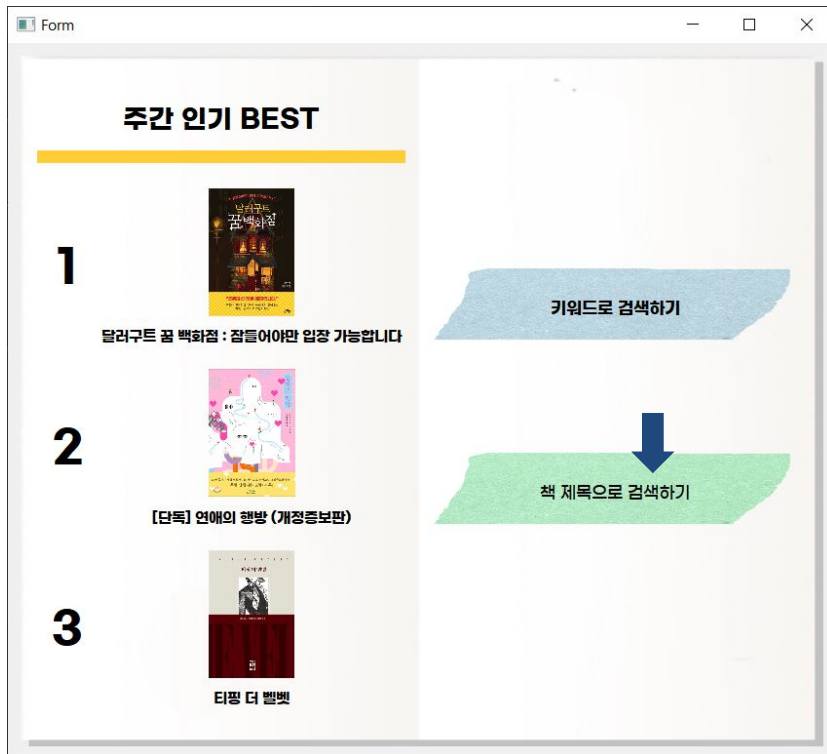
## 지정한 키워드를 기반으로 책 추천



## 02

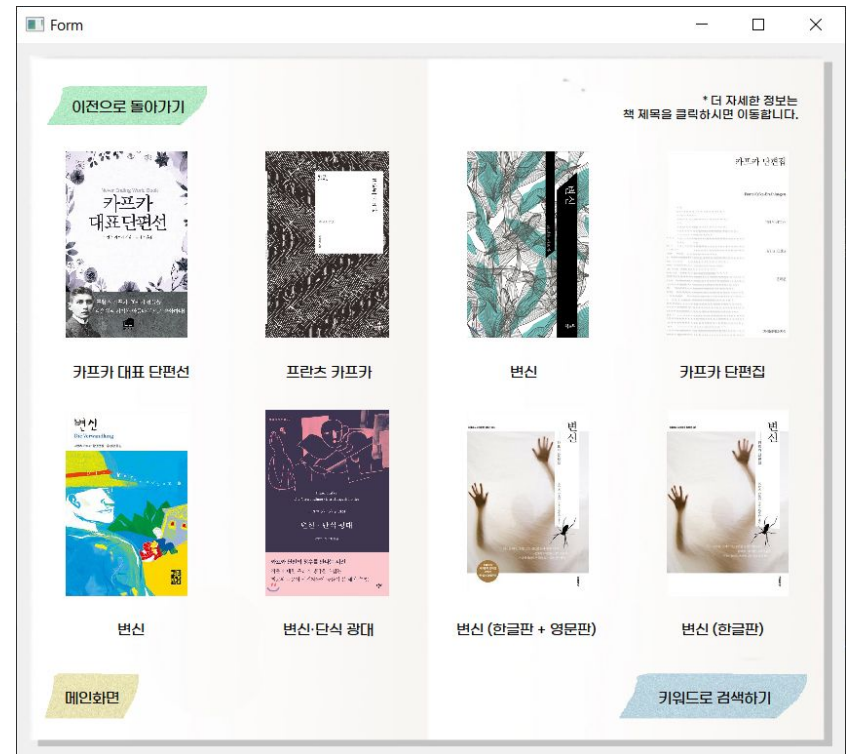
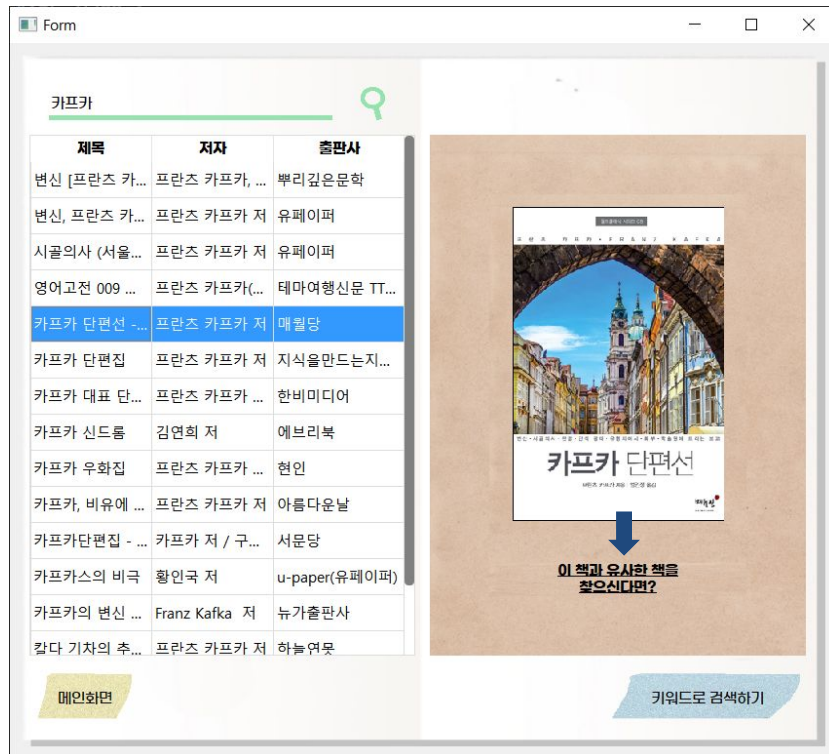
## 앱 소개

## 지정한 책과 유사한 책 추천



## 앱 소개

## 지정한 책과 유사한 책 추천



# 크롤링 (Crawling)

## 1) 데이터

YES24 > EBOOK > 소설 > 한국소설 ~ 세계각국소설

\* 성인인증 필요한 책 제외, 상품명 순으로 정렬 후 크롤링 작업

YES24.COM

eBook 읽기만 해도 맛있다 『독풍당의 사계절』

회원정보 업데이트 1천원 상품권

빠른분야찾기 베스트 신상품 이벤트 바이백 중고매장 북클럽 채널예스 블로그 READ NOW! 오구오구페이백 단독 선출간 전자책 단말기 디지털머니

월검 > eBook > 소설 > 한국소설

eBook

로맨스  
BL  
만화  
판타지/무협  
소설  
경제 경영  
라이트노벨  
에세이 시  
인문  
사회 정치  
자기계발  
역사  
종교  
예술 대중문화  
자연과학  
가정 살림  
건강 취미 여행  
어린이 유아  
청소년

한국소설  
영미소설  
일본소설  
중국소설  
프랑스소설  
독일소설  
러시아소설

스페인/중남미소설  
북유럽소설  
세계각국소설  
추리/미스터리/스릴러  
SF/판타지  
역사소설  
성장소설/가족소설

연애/사랑소설  
어른을 위한 동화  
영화 드라마 원작  
희곡/시나리오  
고전문학  
대여 (소설)

주간베스트 | 새로 나온 상품 | 회원리뷰

20개씩 보기 ▾ 부가옵션 ▾

전체선택 | 카트에 넣기 | 리스트에 넣기

나도향 단편문학 : 영어리 상품이 - 한국문학읽다 [ EPUB ]

나도향 저 | 리플레이 | 2019년 04월

9,100원 460원

종이책 실물이 아닌 전자책입니다. 구매후 바로 보실 수 있습니다.

카트에 넣기  
바로구매  
원클릭구매  
리스트에 넣기



## 03

## 크롤링 (Crawling)

## 2) 책 제목 코드 수집

Selenium과 css selector 사용하여 책 제목의 코드 수집

CSS : '#category\_layout .goods\_name > a:first-of-type'  
 attribute : 'href' (get\_attribute() 메소드 사용)  
 정규식 패턴 : 'Goods/(.+)\$'

## 한국소설

주간베스트 | 새로 나온 상품 | 회원리뷰

기본순 판매량순 신

ACCESSIBILITY

Contrast 12.63 ✓

Name 錦山, 錦江

Role link

Keyboard-focusable ✓

20개씩 보기 ▾ 부가옵션 ▾

전체선택 카트에 넣기 리스트에 넣기

錦山, 錦江 [ EPUB ]

최병진 저 | 좋은땅 | 2014년 08월

6,000원 P 300원

종이책 실물이 아닌 전자책 입니다. 구매후 바로 보실 수 있습니다.

전교생의 눈과 앞에 담임선생님의 눈이 예상치 못한 상황에 나에게 시선이 멈추어 있었다. 나도 내 자신에게 놀라 지켜보았다.교장선생님 연설처럼 앞산, 뒷산이 메아리쳤다.

카트에 넣기

바로구매

원클릭구매

리스트에 넣기

```
<div class="cCont_listArea" id="category_layout">
  <ul class="clearfix">
    <li>
      <div class="cCont_goodsSet">
        <p class="goods_img">...</p>
        <div class="goods_info">
          <div class="goods_name">
            <span class="gd_nameE">...</span>
            <a href="/Product/Goods/14148895">錦山, 錦江</a>
            <span class="gd_nameE">...</span>
            <span class="gd_feature"> [ EPUB ]</span>
            <a href="/Product/Goods/14148895" class="bgYUI ic
              o_nWin" target="_blank"></a>
          </div>
          <div class="goods_pubGrp">...</div>
          <!--특징2-->
          <div class="goods_price">...</div>
          <div class="goods deli">...</div>
        </div>
      </div>
    </li>
  </ul>
</div>
```

Styles Computed Layout Event Listeners DOM Breakpoints

## 크롤링 (Crawling)

### 3) 책 소개, 출판사 리뷰 수집

Beautifulsoup 사용

책 이미지 CSS : '.gd\_imgArea img' / url이 담긴 attribute : 'src' => get() 메소드 이용하여 해당 정보 추출

#### 이슈 사항

Parser로 'lxml'을 사용하는 경우, "<책 제목>" 이런 문구도 모두 삭제

=> Parser를 'html.parser'로 바꿔서 진행

.text 또는 get\_text()를 실행했을 때 줄바꿈 태그인 <br> 태그를 빈문자로 변환

=> get\_text('\n')을 사용하여 <br>을 줄 바꿈으로 대체하여 읽어오도록 진행

## 전처리 (Processing)

### 토큰화

Komoran 사용하여 직접 필요한 품사를 선택 (NA, NF, NNG, NNP, XR)

\* 필요한 데이터가 명사라서 비교적 명사를 잘 분류해주는 komoran 사용

'MM': '관형사',	'XPN': '체언 접두사',
'NA': '분석불능범주',	'XR': '어근'
'NF': '명사추정범주',	'XSA': '형용사 파생 접미사',
'NNB': '의존 명사',	'XSN': '명사파생 접미사',
'NNG': '일반 명사',	'XSV': '동사 파생 접미사'}
'NNP': '고유 명사',	
'NP': '대명사',	
'NR': '수사',	

숫자, 한글 제외 문자 제거

```
text = re.sub('[^가-힣0-9]', '', str(text))
```

문자가 없는 경우 토큰화 생략

```
text = re.sub(r'^\Ws*$', '', text)
if text == '':
    continue
```

## 전처리 (Processing)

### 이슈사항

문장부호 특히 ()나 , .이 남아 있는 경우 token 정확도가 높아지는 것을 확인

⇒ 기본적인 문장부호를 남기고 token화 진행 한 이후 삭제하도록 설정

토큰화한 단어가 한 단어라는 보장이 없음

예) 영화 "말할 수 없는 비밀"이나 특정 고유명사 등

⇒ 토큰화한 이후 토큰을 합치기 전에 공백을 제거하여 하나의 단어로 만든 후 토큰을 합치는 방식 사용

## 보완점

### 개선사항

- 1) 시리즈 별로 있거나 출판사만 다른 중복된 책 제목
- 2) 앱 실행 시 책 제목(라벨) 길이로 인한 어플 사이즈 크기 조절 문제
- 3) 검색 시, 특히 키워드로 검색할 때 유사도 0인 책 추천 문제

THANK  
YOU