EBOOK 추천 어플리케이션

문학의 숲

CODE 98

정현주, 박철훈, 신 궁

CONTENTS

01

개요

어플리케이션 소개

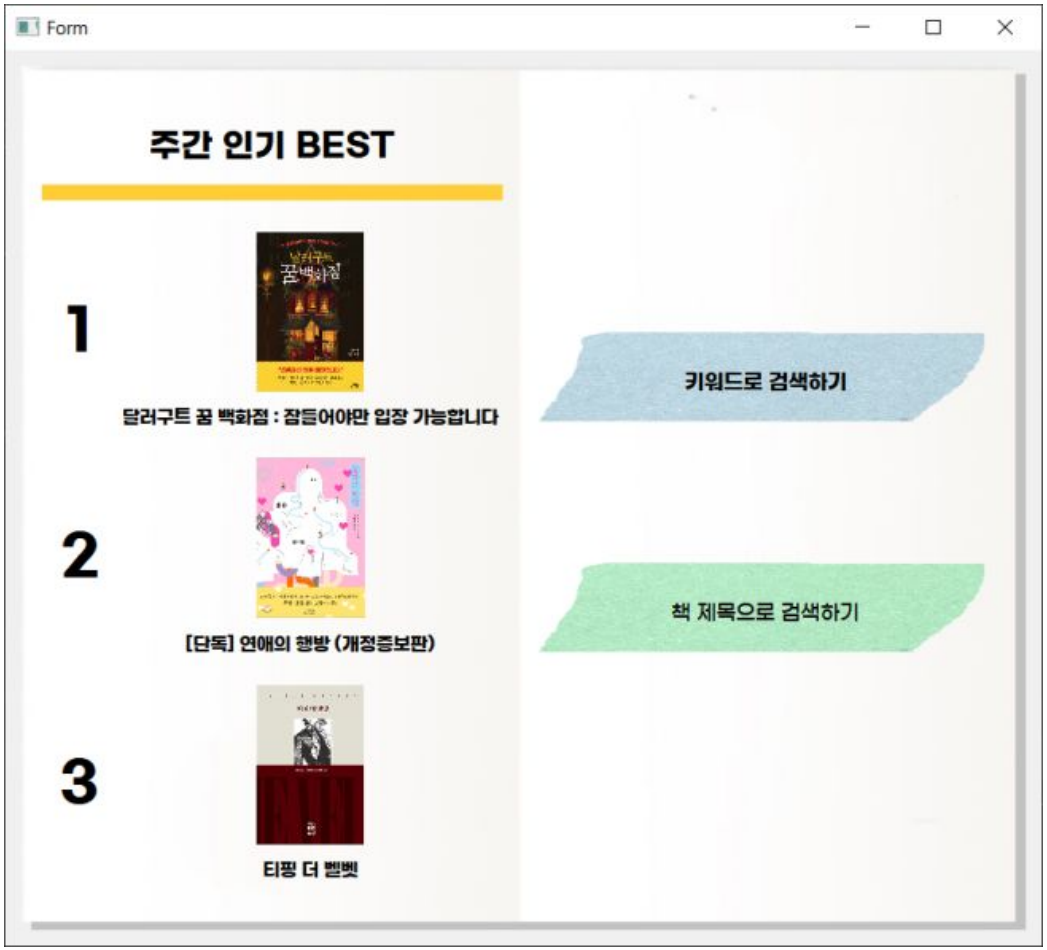
“문학의 숲”

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

02

앱 소개

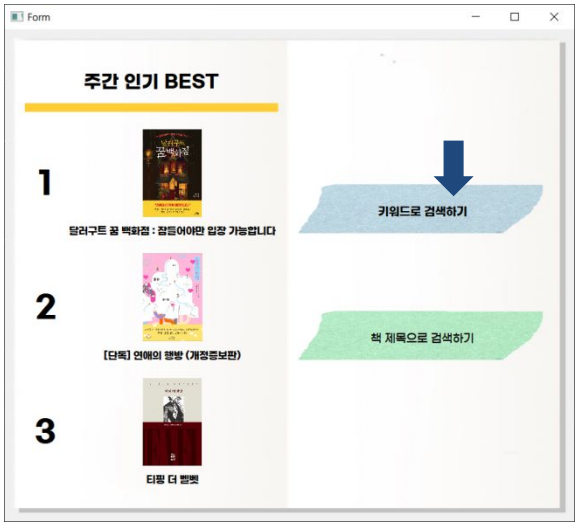
메인

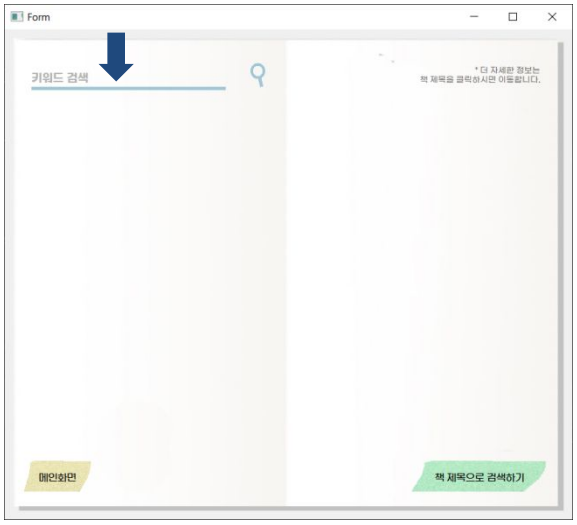

Form
주간 인기 BEST
1
달러구트 꿈 백화점 : 잠들어야만 입장 가능합니다
2
[단독] 연애의 행방 (개정증보판)
3
티핑 더 벨벳
키워드로 검색하기
책 제목으로 검색하기

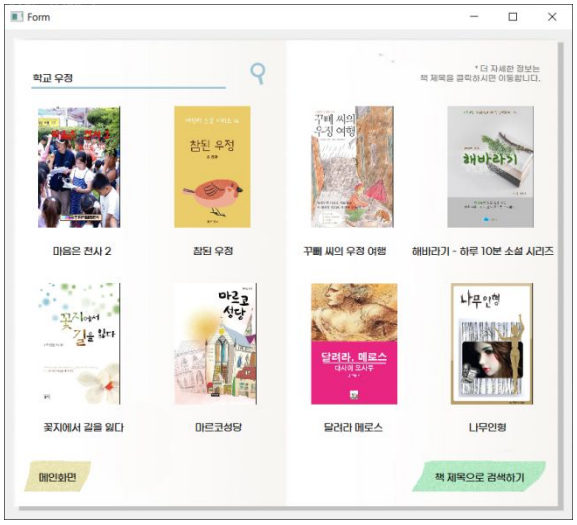
02

앱 소개

지정한 키워드를 기반으로 책 추천



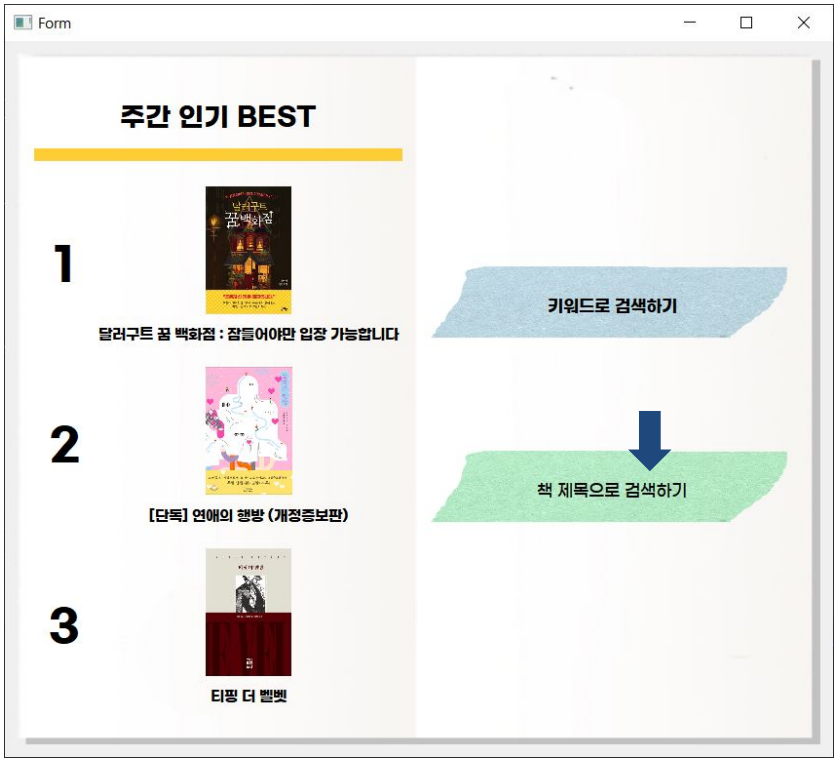


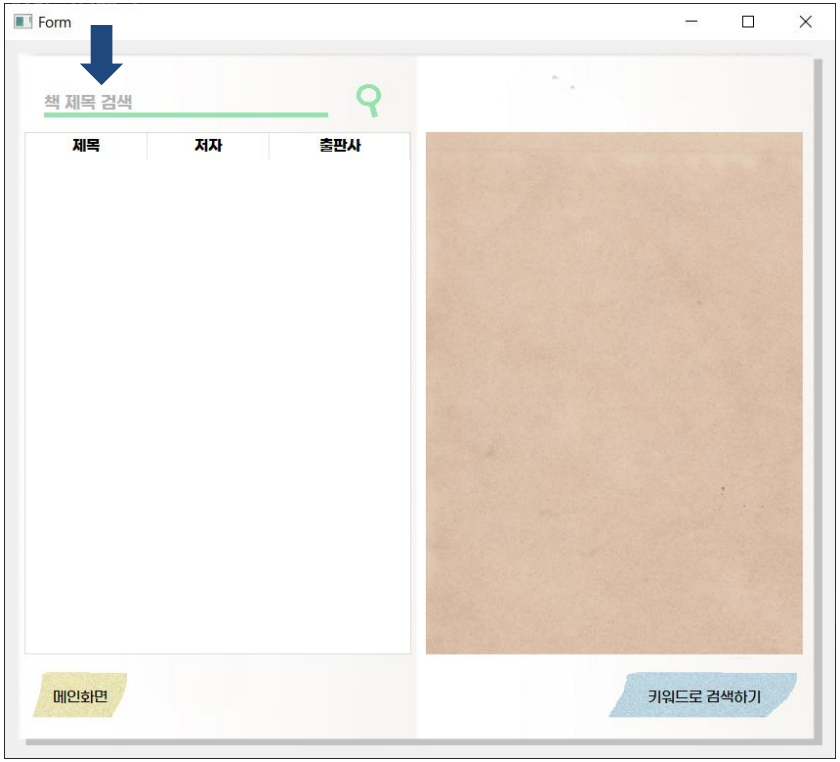


02

앱 소개

지정한 책과 유사한 책 추천

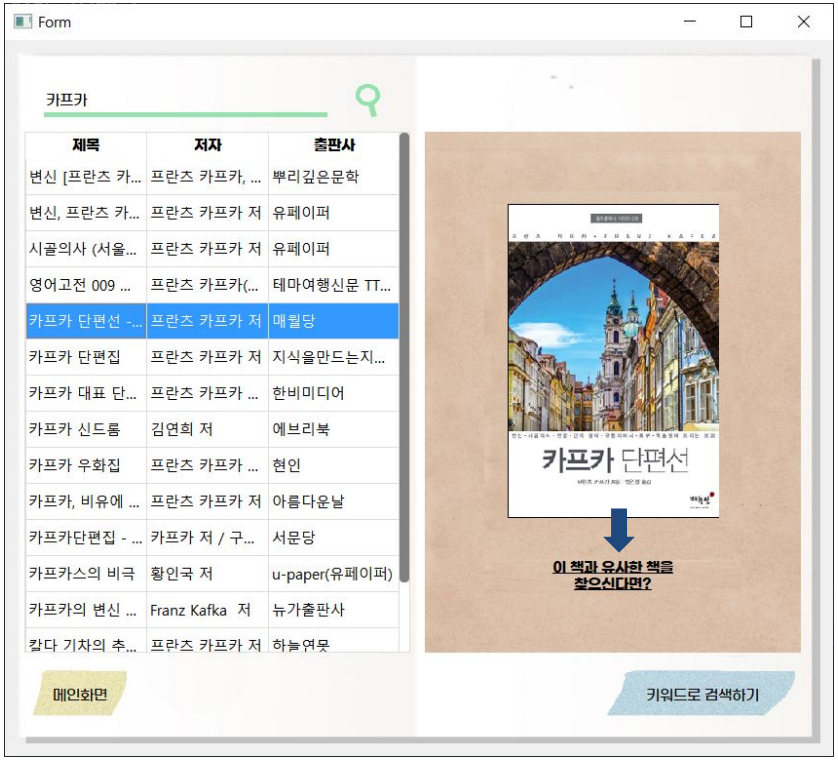


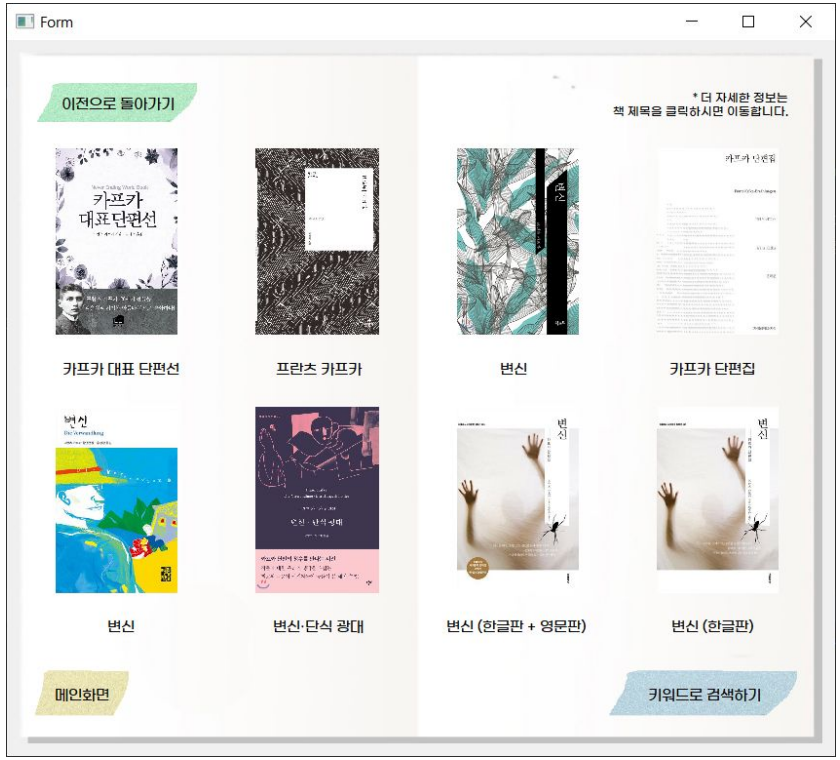


02

앱 소개

지정한 책과 유사한 책 추천





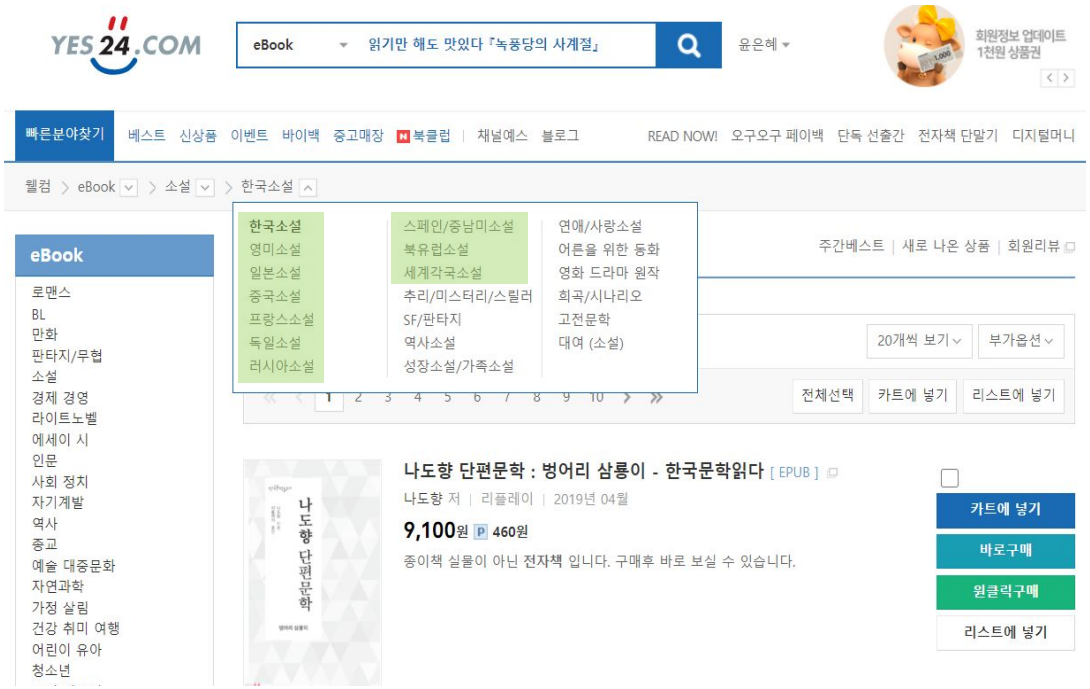
03

크롤링 (Crawling)

1) 데이터

YES24 > EBOOK > 소설 > 한국소설 ~ 세계각국소설

* 성인인증 필요한 책 제외, 상품명 순으로 정렬 후 크롤링 작업



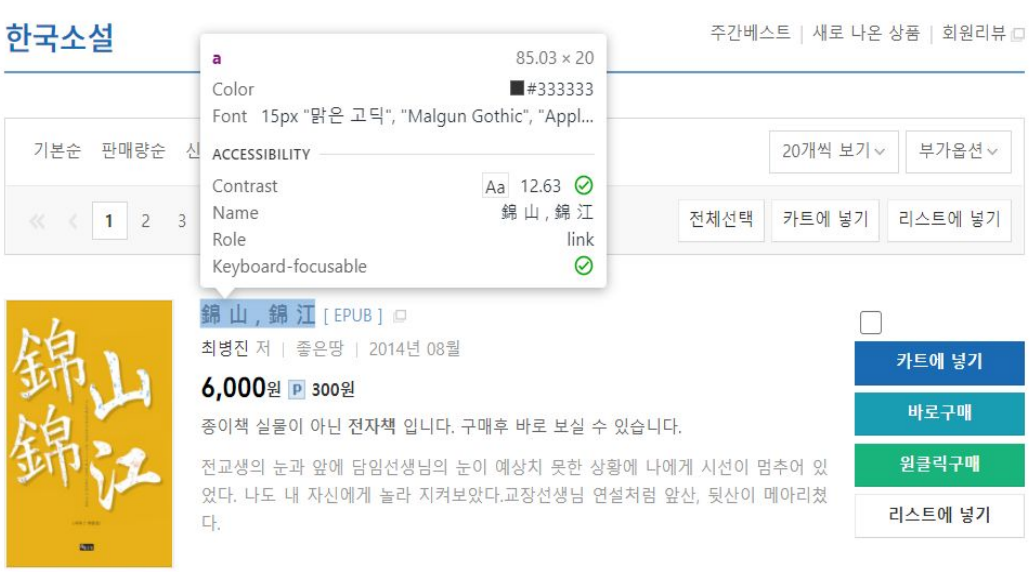
03

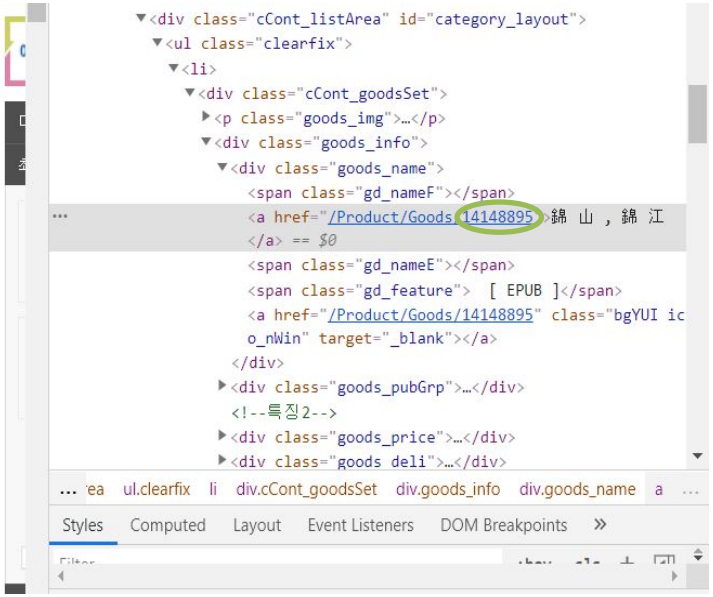
크롤링 (Crawling)

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/(.+)$'





03

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
 태그를 빈문자로 변환

=> get_text(‘₩n”)을 사용하여
을 줄 바꿈으로 대체하여 읽어오도록 진행

04

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

숫자, 한글 제외 문자 제거

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

문자가 없는 경우 토큰화 생략

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

04

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

05

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