키워드 기반 도서 추천시스템

김수정 김용우 노시영 박솔희 박준민

Contents

01. 프로젝트 목표

02. 데이터 셋

03. 모델링

04. 웹 구현

PART

01. 프로젝트 목표

장르별 구분, 유저 기반 추천이 대부분!

- 종합
- 가정/요리/뷰티
- 건강/취미/레저
- 경제경영
- 고전
- 과학
- 만화
- 달력/기타
- 사회과학
- 소설/시/희곡
- 어린이
- **에세이**
- 여행
- 역사
- 예술/대중문화
- 유아
- 인문학
- 자기계발
- 종교/역학
- 청소년
- 컴퓨터/모바일

마법사의 선택 : 국내도서 > 에세이

- 최근 알라딘에서 가장 많이 추천되는 신간 도서입니다.
- 로그인 하시면 고객님의 독서 취향에 맞는 도서를 추천해드립니다. 로그인

상품명순 | 판매량순 | 출간일순 | **등록일순** | 저가격순 | 고가격순

전체선택 | 장바구니 담기 | 보관함 담기 | 마이리스트 담기 | 관심없어요 | 구매했어요

새창열기 미리보기

[뱀주사위놀이 세트+변색유리컵(대상도서 포함, 에세이/여행 2만원 이상)]

오래 준비해온 대답 - 김영하의 시칠리아 *Choice*

김영하 (지은이) | 복복서가 | 2020년 4월

16,500원 → **14,850원** (10%할인), 마일리지 820원 (5% 적립)

★★★★½ (24) | 세일즈포인트 : 142,720

지금 **택배**로 주문하면 **오늘 (17~21시)** 수령

최근 1주 88.5% (중구 중림동) 지역변경

장바구니

바로구매

보관함

보관함
마이리스트

PART

01. 프로젝트 목표

책 내용에 기반한 추천시스템을 만들자!

책 전체 -> 현실적으로 불가능

책 소개 -> 책의 내용을 요약적으로 볼 수 있음

책소개

힘껏 달려야 하는 산과 의사의 일상,
더없이 특별한 탄생의 이야기

아기 울음소리를 듣기 어려워지는 저출산 시대, 생과 사의 경계에 위태롭게 선 수많은 고위험 임산부와 아기를 구하기 위해 날마다 분투하는 의사가 있다. 『태어나줘서 고마워』는 바로 그 의사, 성균관의대 삼성서울병원 산부인과 오수영 교수의 이야기다. 오수영 교수는 스무 해가 지나도록 산부인과 의사로 일하며 만나온 수많은 고위험 임산부와 손끝으로 받아낸 아기들을 마음에 품고, 기억하고, 기록했다.

강남역 한복판에서 애걸복걸하며 택시를 타고 달려가 응급수술을 했던 날, 생후 채 몇 시간을 살 수 없을지라도 끝까지 최선을 다해 아이를 낳고 싶다는 임산부의 수술을 집도한 날, 여섯 번의 유산 끝에 아기를 품에 안고 울었던 산모의 배를 봉합한 날... 저자가 거쳐온 이 모든 날의 이야기에는 의료진의 가쁜 숨과 더없이 애틋한 부모의 마음, 갓난아기의 어여쁜 첫울음이 깊게 배어 있다.

PART

02. 데이터 셋 - Crawling

Yes24 도서 약 40,000권 Crawling
Request와 Beautiful Soup 활용

책 제목 | 책 소개 | 작가

	Name	Story	Author
0	외눈박이 물고기의 사랑	\n\r\n시인이자 명상가, 번역가로 활동중인 류시화의 두번째 시집. 일상 언어들을...	\n류시화 저\r\n
1	지금 알고 있는 걸 그때도 알았더라면	\n\r\n20여 년간 명상과 인간의식 진화에 대한 번역서를 소개하면서 시를 써온 ...	\n류시화 저\r\n
2	연인	\n\r\n우리들 가슴 어딘가에 감추어져 있는 진정한 사랑의 풍경소리를 찾아가는 푸...	\n정호승 저\r\n
3	연인	\n\r\n우리들 가슴 어딘가에 감추어져 있는 진정한 사랑의 풍경소리를 찾아가는 푸...	\n정호승 저\r\n
4	섬진강 이야기1	\n\r\n남도 5백 리 길 세 개의 도와 열두 개의 군을 거쳐 지나가며 사람들의 ...	\n김용택 저\r\n

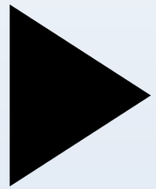
PART

02. 데이터 셋 - Tokenize

KoNLPy 활용
Tokenize 및 품사 Tagging

N

Adj



✓ 많은 의미를 품고 있다.

✓ 키워드로 활용하기 좋다.

PART

02. 데이터 셋 - Tokenize

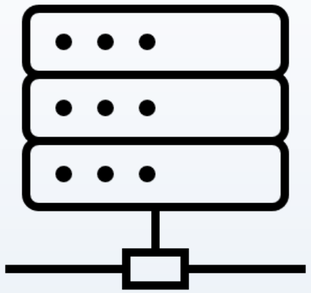
명사 + 형용사

	Name	Story	NounAdj
0	외눈박이 물고기의 사랑	\n\r\n시인이자 명상가, 번역가로 활동중인 류시화의 두번째 시집. 일상 언어들을...	시인 이자 명상 번역가 활동 류시화 두번째 시집 일상 언어 사용 세계 빛 그 시 마...
1	지금 알고 있는 걸 그때도 알았더라면	\n\r\n20여 년간 명상과 인간의식 진화에 대한 번역서를 소개하면서 시를 써온 ...	여 년 명상 인간 의식 진화 대한 번역 를 소개 온 류시화 사랑 잠언 시집 삶 대한...
2	연인	\n\r\n우리들 가슴 어딘가에 감추어져 있는 진정한 사랑의 풍경소리를 찾아가는 푸...	우리 가슴 어딘가 사랑 풍경 소리 푸른 툭눈 이야기 운주사 풍경 삶 현재 사랑 못 ...
3	연인	\n\r\n우리들 가슴 어딘가에 감추어져 있는 진정한 사랑의 풍경소리를 찾아가는 푸...	우리 가슴 어딘가 사랑 풍경 소리 푸른 툭눈 이야기 운주사 풍경 삶 현재 사랑 못 ...
4	섬진강 이야기1	\n\r\n남도 5백 리 길 세 개의 도와 열두 개의 군을 거쳐 지나가며 사람들의 ...	남도 백 리 길 세 개 도 개 군 사람 삶 강 우리 뇌리 가장 강 기억 섬진강 사계...

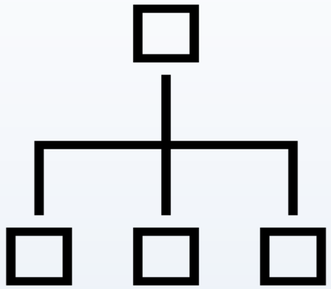
PART

03. 모델링 - 구현하고자 하는 시스템

1) 책 이름을 검색한다



2) 검색한 책의 핵심 키워드 5개가 뜬다



3) 마음에 드는 키워드 3개를 고른다

4) 해당 키워드와 유사한 내용의 책이 추천된다

PART

03. 모델링 - 구현 과정

①

Rake

핵심 키워드
뽑아내기

②

TF-IDF vectorizer

유사도 검사를
위한 벡터화

③

Cosine similarity

키워드간 유사도

03. 모델링 - Rake

RAKE(Rapid Automatic Keyword Extraction)

- Nltk의 키워드 추출 모델
- 문서 자체만의 맥락을 고려하여 중요한 키워드 추출
- Stop word 주변에 있는 단어들을 활용하여 어구를 만들고, 추후에 stop word 제외

stopwords

[is, not, that, there, are, can, you, with, of, those, after, all, one]

03. 모델링 - Rake

	keyword	extraction	difficult	many	libraries	help	rapid	automatic
keyword	3	3	0	0	0	0	1	1
extraction	3	3	0	0	0	0	1	1
difficult	0	0	1	0	0	0	0	0
many	0	0	0	1	1	0	0	0
libraries	0	0	0	1	1	0	0	0
help	0	0	0	0	0	1	0	0
rapid	1	1	0	0	0	0	1	1
automatic	1	1	0	0	0	0	1	1

☞ Content word가 함께 등장하는 횟수를 Matrix로 생성

PART

03. 모델링 - TF-IDF vectorizer

TF * IDF

TF : Term Frequency

특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값

IDF : Inverse Document Frequency

여러 문서에서 등장한 단어의 가중치를 낮추는 역할

03. 모델링 - cosine similarity

Cosine Similarity

코사인 유사도: 두 벡터 간의 각도를 이용하여 유사도를 계산
방향이 동일한 경우 : 1
방향이 반대인 경우 : -1

PART

04. 웹 구현

Thank You!

YBigta