키워드 관계를 이용한 E-book 추천

- 팀명: 마음의 양식
- 팀원: 김소라, 안재하, 이영송, 주원진, 황경서
- 훈련과정명: 빅데이터 핀테크 전문가 양성과정
- 운영기관명: AI대학원

목차

I. 프로젝트 배경

A. 프로젝트 개요

- 컨셉
  - E-book 텍스트에 특화된 텍스트 분석 모델을 구축
- 훈련 내용과의 관련성
  - NLP분석 방법론을 사용해 E-book 텍스트의 키워드를 도출
- 개발 환경
  - 사내 전용 서버를 이용하여 GPU 사용

I. 프로젝트 배경

B. 프로젝트 주제

- AI 기반 추천 알고리즘
  - 챗봇 대화 기반의 E-book AI 추천 시스템 구현

- 신문기사 기반 구조화 알고리즘
  - 신문기사 텍스트 기반의 키워드로 E-book 추천 시스템 구현

E-book에 대한 키워드 도출 후
챗봇 대화문과 신문기사에 따라 서로 다른 서비스 제공

I. 프로젝트 배경

C. 프로젝트 목적

- AI 기반 추천 알고리즘
  - E-book 텍스트를 활용하여 E-book 전용 검색 기반 콘텐츠 추천을 위한 문맥 기반 딥러닝 모델 알고리즘 설계

- 신문기사 기반 구조화 알고리즘
  - 신문기사 텍스트 문맥과 유사한 E-book 키워드를 추출하여 추천하는 알고리즘 설계

I. 프로젝트 배경

D. 프로젝트 구조

- E-book 데이터에 대한 EDA
- 데이터 구조 전처리 (표준화, 성인 컨텐츠, 데이터 프레임 비대칭)
- 텍스트 전처리 (불용어, 금칙어, 지도 학습을 위한 카테고리 및 해시태그 라벨링)
- 키워드 추출 알고리즘 적용 (그래프 랭킹 기반, 통계 모델 기반, 딥러닝 기반)
- 인풋 모델링 (챗봇 학습)
- 아웃풋 모델링 (매칭 및 UI)

I. 프로젝트 배경

E. 프로젝트 기대효과

- 훈련생
  - 자연어 처리 알고리즘 학습 및 실습
  - 데이터 전처리 능력 신장

- 기업
  - E-book 컨텐츠 기반의 추천 알고리즘 구축
  - 키워드 간의 계층 구조화

II. 프로젝트 팀 구성 및 역활

훈련생 명	역활
주원진 (팀장)	• 매 주차 회의록을 작성 • 기업과의 커뮤니케이션을 담당 • 데이터 전처리 작업(성인 콘텐츠 금칙어 처리)을 담당함 • 형태소 분석기(konlpy)를 연구함
김소라 (팀원)	• 데이터 공유 환경을 세팅함 • 데이터 전처리 작업(데이터 노이즈 제거)을 담당함 • 텍스트 요약 알고리즘을 연구함
안재하 (팀원)	• EDA를 통해 데이터의 노이즈를 파악함 • 데이터 전처리 작업(데이터 노이즈 제거)을 담당함 • 대화체 내용을 발췌하는 코드를 작성함

II. 프로젝트 팀 구성 및 역활

훈련생 명	역활
이영송 (팀원)	• PPT 제작 • EDA를 통해 데이터의 노이즈를 파악함 • 형태소 분석기(soynlp)를 연구함
황경서 (팀원)	• KJ 메소드 적용 등 팀 회의의 퍼실리테이터 역할을 담당함 • 데이터 전처리 작업(외국어 카테고리)을 담당함 • 형태소 분석기(카이)를 연구함

III. 프로젝트 수행절차 및 방법

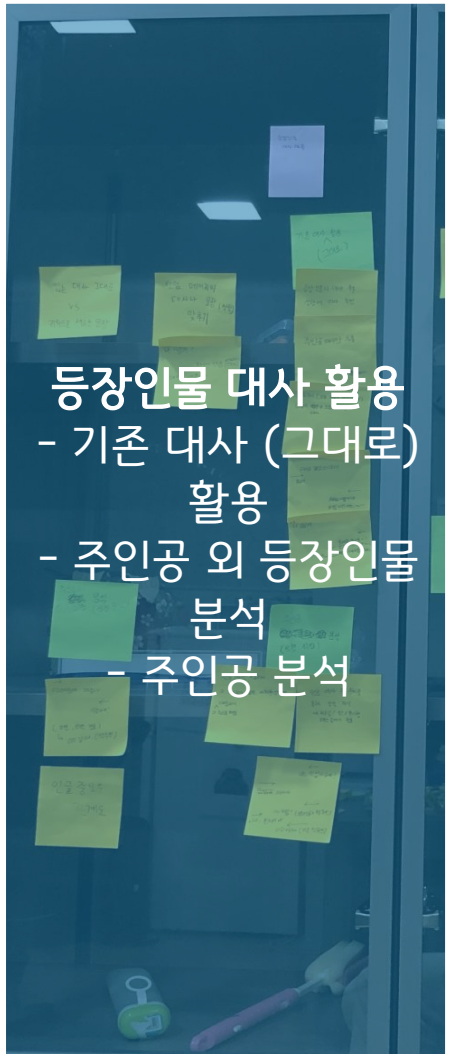
구분	기간	활동	비고
사전 기획	5/11(월)~5/13(수)	오리엔테이션 과제 개발 및 데이터 공유	주제 구체화
	5/14(목) ~ 5/15(금)	KJ메소드를 통해 아이디어 회의 데이터 EDA	타임라인 작성
개발	5/18(월) ~ 5/22(금)	데이터 전처리 모델 선정	EDA & Research
	5/25(월) ~ 6/12(금)	데이터 전처리 완료 텍스트 키워드 추출	알고리즘 개발
	6/15(월) ~ 6/19(금)	챗봇 학습 뉴스 키워드 추출	매칭
	6/22(월) ~ 6/26(금)	UI 구축	완성
수정/보완	6/29(월) ~ 7/3(금)	피드백 반영	최적화, 오류 수정
총 개발 기간	5/11(월) ~ 7/3(금) (8주)		

IV. 프로젝트 수행 결과

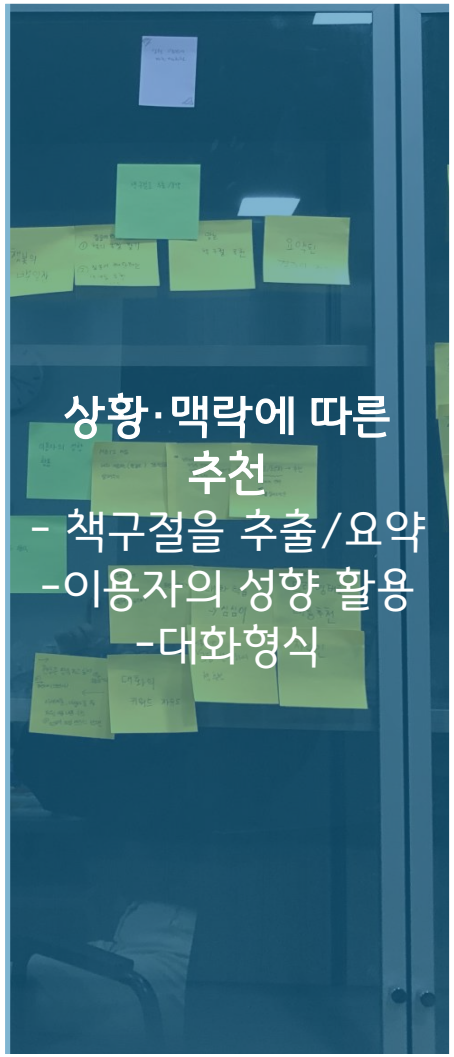
사용자의 만족도
- 도서내용 이외의 데이터 활용

책의 유용성(실용성)
- 제한된 분량 정보제공(요약)
- 전문용어·지식 설명 (like WiKi)
- 특정정보(키워드) 추출

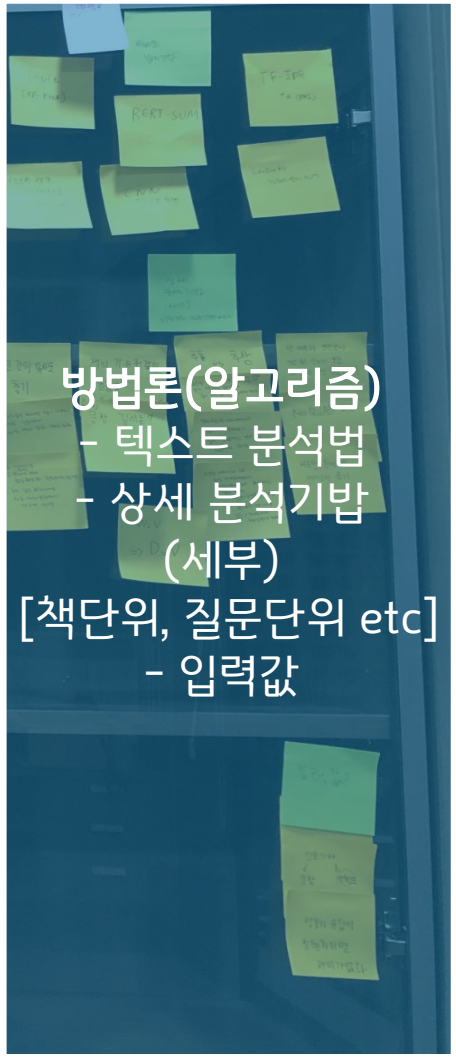
Output 결과 (서술방식)
- 차이점 강조



등장인물 대사 활용
- 기존 대사 (그대로) 활용
- 주인공 외 등장인물 분석
- 주인공 분석



상황·맥락에 따른 추천
- 책구절을 추출/요약
-이용자의 성향 활용
-대화형식



방법론(알고리즘)
- 텍스트 분석법
- 상세 분석기법 (세부)
[책단위, 질문단위 etc]
- 입력값

IV. 프로젝트 수행 결과

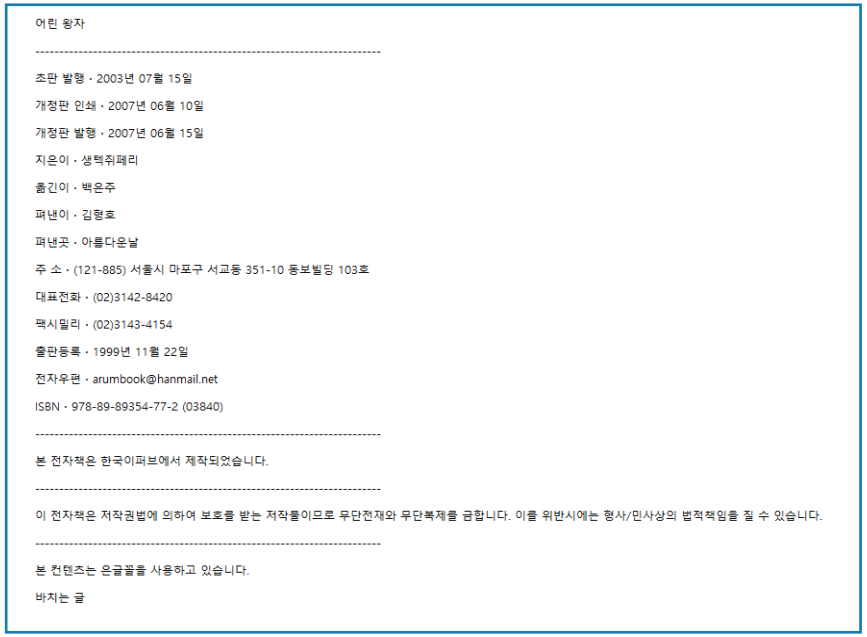
e-book_target

E-book에 대한 정보를 담고 있는 텍스트 파일

serial	title	cat_id	Cat	date	File_ID
3765732	와인 읽는 CEO	004013002	성공학/경력관리	20100329	92
3765733	육일약국 갑시다	004002010	CEO/비즈니스맨	20100329	84
3772985	이탈리아 오래된 도시로 미술여행을 떠나다	004011001	예술기행	20100405	181
...	...	...	...	...	...
90184260	지리 교재 연구 및 교수법	004009012	기타 자격증	20170421	1165862
90184261	읽고 배우고 익히는 고사성어	004010002002	동요/동시	20170421	1165863
90184262	하늘에서 읽는 대한민국	004004002001	여행/관광	20170421	1165864

e-pub files

E-book에 대한 내용을 담고 있는 텍스트 파일



어린 왕자

초판 발행 · 2003년 07월 15일

개정판 인쇄 · 2007년 06월 10일

개정판 발행 · 2007년 06월 15일

지은이 · 생텍쥐페리

옮긴이 · 백은주

펴낸이 · 김형호

펴낸곳 · 아름다운날

주 소 · (121-885) 서울시 마포구 서교동 351-10 동보빌딩 103호

대표전화 · (02)3142-8420

팩시밀리 · (02)3143-4154

출판등록 · 1999년 11월 22일

전자우편 · arumbook@hanmail.net

ISBN · 978-89-89354-77-2 (03840)

본 전자책은 한국이퍼브에서 제작되었습니다.

본 컨텐츠는 은글꼴을 사용하고 있습니다.

바치는 글

파일 ID로 연결

IV. 프로젝트 수행 결과

EDA

데이터 개요

e-book target 총 행 개수: 905,965개

e-pub files 폴더 개수: 346개

e-pub files 파일 개수 : 344,635개

카테고리 수 : 245개

출판일: 1900년 1월 2일 ~ 2017년 9월 15일

Null: 카테고리 정보, 출판일 정보 1개

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 905965 entries, 0 to 905964
Data columns (total 7 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   도서id    905965 non-null  object
 1   도서명     905965 non-null  object
 2   카테고리id  905964 non-null  float64
 3   카테고리명   905964 non-null  object
 4   출판일     905964 non-null  float64
 5   파일id    905965 non-null  int64
 6   폴더id    905965 non-null  int64
dtypes: float64(2), int64(2), object(3)
memory usage: 48.4+ MB
```

IV. 프로젝트 수행 결과

교집합 추출

Txt 파일이 존재하는 정보만 남긴 데이터 프레임

	도서id	도서명	카테고리id	카테고리명	출판일	파일id	폴더id
0	4280896	러시아 문화예술의 천년	4.011003e+06	건축	20101011.0	0	0
1	3808843	2000만원으로 연봉버는 경매투자	4.002009e+06	투자/재테크	20100502.0	1	0
2	10998438	황석영 삼국지 세트(1~10권)	4.000000e+00	ebook	20130906.0	2	0
3	10998441	황석영 삼국지 세트(1~10권)	4.000000e+00	ebook	20130906.0	2	0
4	10998440	황석영 삼국지 세트(1~10권)	4.000000e+00	ebook	20130906.0	2	0
...	...	...	...	...	...	...	...
905960	4029599	카라반의 전설	4.010002e+09	그림책/동화책	20100718.0	99026262	0
905961	4025427	떨메와 달궁이	4.010002e+09	그림책/동화책	20100715.0	99027004	0
905962	4029124	이야기가 있는 경복궁 나들이	4.001002e+09	풍속/문화이야기	20100718.0	99027502	0
905963	4029120	극장경영과 공연제작	4.010002e+09	문화/예술	20100718.0	99034035	0
905964	4028587	소설 21세기 - 제7호 MOOK	4.003001e+09	한국소설	20100718.0	99034748	0

IV. 프로젝트 수행 결과

전처리

표준화 문제

도서 메타 정보 삭제
상, 하위 50 line 추출
추출된 토큰의 공통 단어 삭제

외국어 카테고리

외국어 비율 칼럼
일정 비율 이상 외국어 도서 > 분석 제외
일정 비율 이하 외국어 도서 > 띄어쓰기 토큰 처리

성인 컨텐츠

금칙어 처리
BL/로맨스/판타지/무협/라이트노벨/만화/성인
형태소 분석 후 다른 장르와 비교하여 추출
내부 금칙어, 나무위키, SAFE SEARCH

데이터 간 불일치

Info (90만) >Text (36만)
1) Info의 ID 중복 case
> 텍스트 ID와 Info 정보의 매칭 문제
2) Info의 제목 중복 case
> 동일 내용 및 시리즈 처리 문제

IV. 프로젝트 수행 결과

표준화 문제

서식 및 용어 통일성 문제

용어의 통일성 문제

목차, 차례, Contents …

지은이, 저자, 작가…

프롤로그, 추천사, 머리말…

출판사, 편집자, 편집인, 책임 편집 …

서식의 통일성 문제

메타 정보 없음

글의 머리, 말미 등장

앞, 뒤 50줄 씩 100줄의 공통 단어를 찾고

해당 단어가 나타나는 문장을 삭제하기

[COVER_PAGE]

한국문학평론가협회 | 한길사 공동기획

김종회 지음

이 땅에서 글을 써 가지고 살림을 차려본다는 것은 거의, 절망에 가까운 일이 아닐 수 없건만, 그러나 나에게는

..... 박태원, 『여인성장』

머리말

구보 박태원은 한국 현대문학, 더 나아가 남북한 현대문학에 있어 간과할 수 없는 중요성을 지닌 작가이다. 19

그러기에 남북한 문화 및 문학의 교류를 넘어서 양자의 통합을 바라보는 다양한 시도들이 이루어지고 있는 오

지은이 Garryowen J. Rhee
펴낸곳 G English 연구소
펴낸이 이진호

초판 발행 2010년 12월 15일
등 록 1999. 1. 28 No 4-120
전 화 031) 285-3908
팩 스 031) 285-3909
도메인 www.genglish.com / www.genglish.co.kr
ISBN 978- 89-950338-5-2

First published 2010
in South Korea

IV. 프로젝트 수행 결과

표준화 문제

책 메타 정보 토큰 추출 (전, 후 50줄 씩 100줄)

```
('것은', 353),
('그리고', 349),
('것이', 331),
('하는', 316),
('있습니다 ', 301),
('하지만', 284),
('목차', 261),
('그의', 258),
('지은이', 249),
('없는', 247),
('같은', 241),
('그런', 241),
('다른', 240),
('것을', 238),
('했다 ', 233),
('나는', 232),
('1 ', 232),
('그는', 229),
('내가', 219),
('대한', 218),
('않았다 ', 215),
('없었다 ', 208),
('자신의', 199),
('하고', 190),
('다시', 190),
('그러나', 189),
```



```
for line in text_list2:
    if '출판사' in line:
        print(line)
```

```
출판사 신고번호 : 제1081-1-89호
출판사 : 도서출판 청어람
출판사 신고번호 : 제1081-1-89호
출판사 : 러브홀릭
펴낸곳 | (주)반디출판사
출판사 유페이퍼
출판사 : 케이너스(CAINNUS)
출판사 신고번호 : 제319-2014-23호
출판사 : 도서출판 청어람
출판사 : ㈜북큐브네트웍스 (전자책브랜드 : 북큐브, 판무스토리, 환상북스, 로맨스토리)
출판사 신고번호 : 제25100-2009-000058호
출판사 : 도서출판 청어람
출판사 등록일 2010년 8월 10일
출판사 : 프로무림
* 이 책은 1990년 출간된 『나는 왜 작은 일에만 분개하는가』(햇빛출판사)를 재편집하였습니다.
어수룩한 제 글을 좋게 봐주셔서 두 번째 작가의 말을 쓰는 기회를 주신 출판사 분들께도 감사드립니다.
출판사 AP 북스
출판사 도디드
출판사 유페이퍼
출판사 오연출판사
```

IV. 프로젝트 수행 결과

표준화 문제

단편 및 시리즈

단편집 문제

책 한 권의 텍스트 안에

독립된 내용이 묶음으로 존재하는 경우

시리즈 문제

개별 텍스트 간

동일한 주제가 분산되어 존재하는 경우

단편집의 경우 1권 > 개별 n권으로 분화

시리즈의 경우 1/2/…/m권 > 개별 m권으로 고려

단편집

제3부 함박눈 내리는 날
첫눈
함박눈
함박눈 내리는 날
산개의 추억
최후의 만찬

제4부 100년 만의 정월대보름
노동귀족의 종말을 위한 협주곡
전면적 직선제 쟁취를 위한 공동투쟁본부 출범선언
출범 다음날 새벽에 쓴 출범선언
철도노동자총파업투쟁선언
순식간에 써내려간 파업선언
100년 만의 정월대보름
기차를 세울 수밖에 없었던 이유

시리즈

	unknown	title	id_cat	cat	date	id
530	3909264	카론의 창 1/5	004027	판타지	20100607	3465
531	3909265	카론의 창 2/5	004027	판타지	20100607	3464
532	3909266	카론의 창 3/5	004027	판타지	20100607	3463
533	3909267	카론의 창 4/5	004027	판타지	20100607	3462
534	3909268	카론의 창 5/5	004027	판타지	20100607	3461
...	...	...	...	...	...	...
905831	43882482	[대여] 무한공포 4권	004027	판타지	20170608	651684
905832	43882483	[대여] 무한공포 5권	004027	판타지	20170608	651685
905833	43882484	[대여] 무한공포 6권	004027	판타지	20170608	651686
905834	43882485	[대여] 무한공포 7권	004027	판타지	20170608	651688
905835	43882486	[대여] 무한공포 8권	004027	판타지	20170608	651690

IV. 프로젝트 수행 결과

카테고리 체계 (3자리)

중 카테고리

Lcat	Lcat Name	Mcat	Mcate Name	
004	ebook	027	판타지	[25.44%] 79,232
		026	로맨스	[20.23%] 63,026
		028	무협	[12.28%] 38,261
		003	문학	[11.40%] 35,508
		001	인문/사회	[5.76%] 17,950
		Null	Null	[5.60%] 17,432
		010	어린이/청소년	[3.08%] 9,592
		029	BL	[2.91%] 9,065
		017	만화	[2.38%] 7,419
		013	자기관리	[1.98%] 6,156
		004	가정과생활	[1.92%] 5,993
		002	비즈니스와경제	[1.79%] 5,562
		008	국어와외국어	[1.66%] 5,180
		019	개인출판물	[0.91%] 2,835
		030	라이트노벨	[0.65%] 2,023
		006	잡지	[0.50%] 1,563
		011	예술/대중문화	[0.44%] 1,373
		009	수험서/자격증	[0.38%] 1,185
		014	자연과과학	[0.34%] 1,068
		012	대학교재	[0.20%] 618
		015	컴퓨터와인터넷	[0.14%] 437

소 카테고리

Mcate N..	Scat	Dcat	Cat	
문학	001	001	한국소설	9,343
		002	외국소설	9,330
		004	한국근대소설	5,539
		005	어른을위한동화	227
		003	감성소설	178
		Null	소설	7
	002	001	고전	1,329
		002	희곡	544
	003	001	시	2,649
		002	이론과평론	

0K 2K 4K 6K 8K 10K

Ex) 004 003 001 001 : 한국소설

IV. 프로젝트 수행 결과

카테고리 재분류

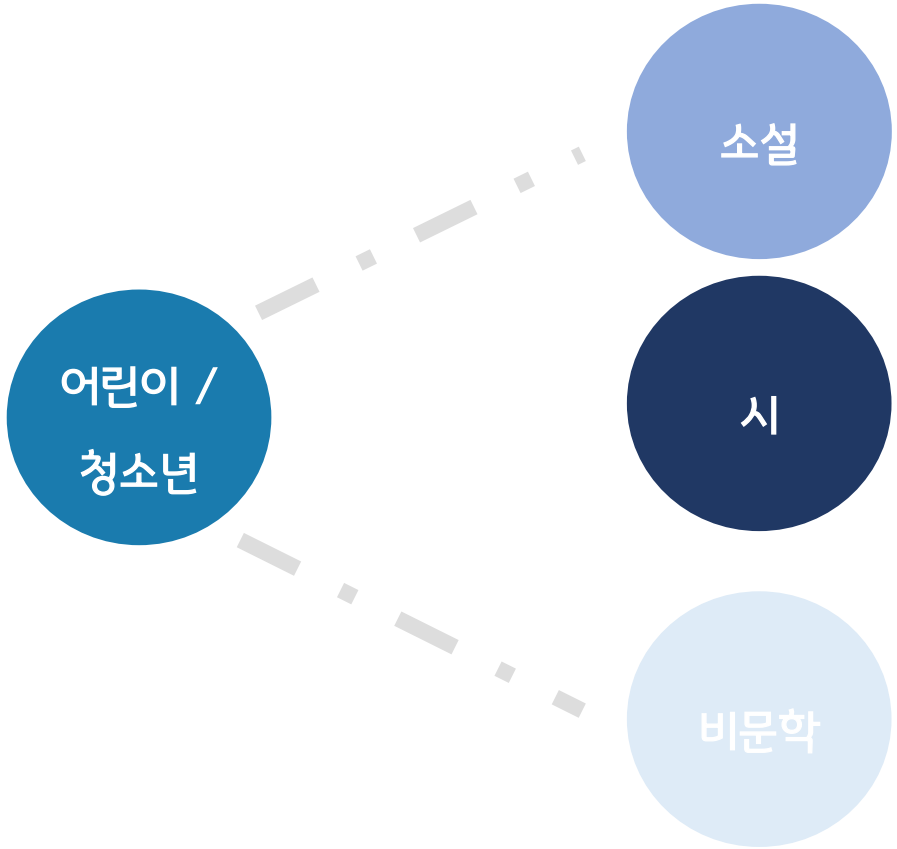
형식 기반 메타 분류

소설 시 비문학

도메인 리서치 후 메타 분류

메타 분류에 해당하는 분류 체계의 데이터를 취사선택

K-means와 같은 비지도 학습법 사용



IV. 프로젝트 수행 결과

외국어 처리 문제 영어, 독일어, 한자 등 다수의 외국어 사용

	도서id	도서명	카테고리id	카테고리명	출판일	파일id
29768	7001812	The Wonderful Wizard of Oz	4020001.0	구텐베르크 프로젝트	20120430.0	50670
29769	7001813	NREN	4020001.0	구텐베르크 프로젝트	20120430.0	50671
29770	7001814	Aladdin and the Magic Lamp	4020001.0	구텐베르크 프로젝트	20120430.0	50672
29771	7001815	Paradise Regained	4020001.0	구텐베르크 프로젝트	20120430.0	50673
29772	7001816	Discourse on the Method of Rightly Conducting ...	4020001.0	구텐베르크 프로젝트	20120430.0	50674
...	...	...	...	...	...	...
46773	7122835	De roman van den schaatsenrijder	4020001.0	구텐베르크 프로젝트	20120523.0	69779
46774	7122838	Vandiemenin maa	4020001.0	구텐베르크 프로젝트	20120523.0	69782
46775	7122839	The Stock-Feeder's Manual / the chemistry of f...	4020001.0	구텐베르크 프로젝트	20120523.0	69783
46776	7122840	玉梨魂	4020001.0	구텐베르크 프로젝트	20120523.0	69784
46777	7122841	Freedom In Service / Six Essays on Matters Con...	4020001.0	구텐베르크 프로젝트	20120523.0	69785

0. 카테고리 삭제

외국 소설, 해외 e-book 등

(번역서 등의 경우 일반적으로 다른 카테고리 존재)

1. 외국어 비율 칼럼 생성

정규식 사용하여 토큰의 비율

2. 외국어 도서 제외

일정 비율 이상 도서

분석 대상 제외

3. 외국어 형태소 분석이 필요한 경우

일정 비율 이하 도서

외국어 덩어리를 하나의 토큰으로 분석

(전문서적)

IV. 프로젝트 수행 결과

성인 콘텐츠

성인 콘텐츠 Token

```
Out[16]: ['것',
          '나',
          '그녀',
          '그',
          '내',
          '생각',
          '그것',
          '여자',
          '친구',
          '수',
          '이름',
          '몸',
          '서로',
          '안',
          '가슴',
          '때',
          '우리',
          '원나잇',
          '맘'
```

−

아동 콘텐츠 Token

```
Out[19]: ['것',
          '단단',
          '사람',
          '꽃',
          '수',
          '나라',
          '말',
          '우리',
          '그',
          '남북',
          '통',
          '나',
          '속',
          '세상',
          '분단',
          '사랑',
          '문화',
          '거',
          '생각'
```

=

금칙어 Token

```
Out[14]: ['원나잇',
          '파티',
          '섹스',
          '택시',
          '공허',
          '레드홀릭스',
          '키스',
          '애무',
          '페니스',
          '침대',
          '모텔',
          '상위',
          '토크명전 | 글쓴',
          '샤워',
          '오빠',
          '화장실',
          '혀',
          '아무튼,',
          '하악',
```

파일 ID에 대한 성인 콘텐츠 정보로 카테고리 처리 예정

IV. 프로젝트 수행 결과

데이터 간 불일치

제목과 파일 아이디 중복(3,131개)

모두 중복되는 경우

가장 먼저 중복 제거

	도서Id	도서명	카테고리Id	카테고리명		출판일	파일Id
3337	3910477	중독 2/2	4026	로맨스	2.01006e+07	7204	210
3338	3910483	중독 2/2	4026	로맨스	2.01006e+07	7204	210

IV. 프로젝트 수행 결과

데이터 간 불일치

파일 ID 중복 (25,587개)

① 내용이 같은 경우

E.g. 대여, 세트, 시리즈

대여 등의 경우가 다수 해당하며

소장본이 보통 존재하기 때문에 대여본을 삭제

	file_id	title
0	867398	벤허
1	867398	[대여] 벤허
2	840689	마가리타
3	840689	[대여] 마가리타
4	994161	바다 이야기 2권 (완결)
...	...	...
397928	968066	나의 아름다운 죽음을 위하여
397929	841668	INNOVA 7권
397930	841666	INNOVA 5권
397931	841677	가벼움의 시대
397932	841677	[대여] 가벼움의 시대

IV. 프로젝트 수행 결과

데이터 간 불일치

파일 ID 중복 (25,587개)

② 내용이 다른 경우

내용과 확인하여 매칭하거나

데이터 노이즈로 취급하여 삭제하는 경우 고려

	도서id	도서명	카테고리id	카테고리명	출판일	파일id	폴더id
2	3774126	사진 읽는 CEO	4.013001e+06	처세술/삶의자세	20100405.0	70	313
3	11104858	[세트] 레 미제라블(한글+영문)(전5권)	4.003001e+09	외국소설	20131001.0	70	313

IV. 프로젝트 수행 결과

데이터 간 불일치

제목 중복(58,053개)

① 내용이 같은 경우

중복을 제거

출판일이 최근이거나 도서ID가 높은 순 선택

	도서id	도서명	카테고리id	카테고리명	출판일	파일id	폴더id
282127	38951910	(곽선희 설교01집) 한 청년의 고민 1	.001e+09	기독교	2.01703e+07	630497	155
332583	51949213	(곽선희 설교01집) 한 청년의 고민 1	.001e+09	기독교	2.01703e+07	765526	118
282128	38951909	(곽선희 설교01집) 한 청년의 고민 2	.001e+09	기독교	2.01703e+07	630498	69
332584	51949226	(곽선희 설교01집) 한 청년의 고민 2	.001e+09	기독교	2.01703e+07	765527	153
332585	51949238	(곽선희 설교01집) 한 청년의 고민 3	4.001e+09	기독교	2.01703e+07	765528	181

IV. 프로젝트 수행 결과

데이터 간 불일치

제목 중복(58,053개)

② 내용의 차이가 있는 경우

출판사, 개정판 등의 차이 존재 가능성

같은 제목 다른 내용일 가능성

제목과 카테고리 동일한 경우 중복 제거

	unknown	title	id_cat	cat	date	id
489	3907823	어린 왕자	004003002001	고전	20100607	652
6813	4020599	어린 왕자	004003001002	외국소설	20100715	5030118
18959	4577221	어린 왕자	004003001005	어른을위한동화	20110114	24922
19454	4629432	어린 왕자	004003001002	외국소설	20110131	23632
66664	7197277	어린 왕자	004003001002	외국소설	20120612	76771
68137	7290567	어린 왕자	004010002001	그림책/동화책	20120709	78685
83995	8229313	어린 왕자	004003001002	외국소설	20121229	102761
101170	8778640	어린 왕자	004003001002	외국소설	20130419	120515
102253	8845534	어린 왕자	004008002	영어	20130503	121832
139406	11262385	어린 왕자	004003001005	어른을위한동화	20131023	176833

129521	13338091	환멸	4029001	BL소설	20140602	212382	16
220579	27801978	환멸	4E+09	한국소설	20160520	461911	176
220611	27802014	환멸	4E+09	한국소설	20160520	461943	326

V. 프로젝트 향후 고민

Step 01	전처리	카테고리 라벨링 (지도학습 – 해쉬태그, target 정보(제목/카테고리/목차…)) 형태소 분석기(Konlpy, Khaiii, Mecab, etri tokenizer – 말뭉치 사전 기반 지도 학습 방식) 형태소 분석기(Soynlp, SentencePiece(BPE) - 통계 빈도 기반 비지도 학습 방식)
Step 02	텍스트 키워드 알고리즘	그래프 랭킹 기반: 텍스트 랭크(gensim, 일반), Word-rank (비지도) 통계 학습 기반: LDA, HDP (비지도) 딥러닝 기반: BERT (Extractive, Abstractive) ,CNN (지도)
Step 03	챗봇 학습	korQuad 알고리즘 대화체 데이터 학습 (Scene 구분)

V. 프로젝트 향후 고민

Step 01. 전처리

딥러닝 지도학습

해쉬태그 정보 사용

텍스트 문장에 대한 라벨 값 부재

텍스트 각 문장에 대한 키워드의 벡터 값을 계산

문서의 해쉬 태그 단어와 가장 높은 유사도를 매칭

103075	로맨틱코미디,다정남,엉뚱녀,갑을관계,달달물,현대물,상처녀,까칠남,쾌활발랄녀,라이벌/앙숙,연예인,
163887	소유욕/독점욕/질투,현대물,나쁜남자
164328	재회물,현대물,복수,첫사랑,군대물
233639	영화드라마원작
274440	신파,현대물,연예인,카리스마남,불치병/장애
28232	신분차이,강공,미인수,동양풍,집착공,계략수,나이차이,짝사랑공,능력수,감금,사랑꾼공,절륜공,사건물,
45484	순정남,정략결혼,상처남,다정남,잔잔물,나이차커플,외유내강,현대물
120844	현대물,까칠남,운명적사랑,친구>연인,순정남,소유욕/독점욕/질투,로맨틱코미디,신데렐라,상처남,시
121170	영화드라마원작
1708	판타지물,삼각관계,외유내강,순정남,까칠남,왕족/귀족
219097	좋은글의비밀
219113	삼각관계,현대물,상처녀,바람둥이,다정남
261203	재회물,현대물
261363	로맨틱코미디,현대물,까칠남,능글남,상처녀,능력녀,친구>연인
261365	현대물,운명적사랑,상처남,쾌활발랄녀,순진녀,엉뚱녀,연하남,달달물
250375	외국인/혼혈,현대물,동거,재벌남,평범녀

V. 프로젝트 향후 고민

01. 텍스트 전처리

분석 방법	패키지 종류	분석기 이름	사용 여부
사전 기반 지도 학습	Konlpy	한나눔, 코모란, Okt, kkma	O
	Mecab	Mecab	O
	Khaii	Khaii	O
	Etri	Etri	X
통계 기반 비지도 학습	Soynlp	LR Extract, Word Extract	O

V. 프로젝트 향후 고민

Step 01. 전처리

Komoran
한나눔
Okt
kkma
soynlp

```
[감정, 노동자, 보호, 법, 은, 사업주, 로, 하여금, 감정노동, 으로부터, 근로자, 를, 보
호, 하, 는, 예방, 조치, 를, 이행, 하, 도록, 강제, 하, ㄴ다, ., 다만, 현장, 근로자,
를, 을, 중심, 으로, ㄴ, 이, 같, 은, 법안, 이, 현장, 에, 제대로, 적용, 되, 기, 위하,
아서, 는, 회사, 의, 수직, 적, 위계, 구조, 와, 인력, 부족, 문제, 등, 구조, 적, 문제,
가, 우선, 해결, 되, 어야, 하, ㄴ다는, 지적, 도, 나오, ㄴ다, .]
----------------------------------------------------------
[감정노동자, 보호법, 은, 사업주, 로, 하여금, 감정노동, 으로부터, 근로자, 를, 보호,
하, 는, 예방, 조치, 를, 이행, 하, 도록, 강제, 하, ㄴ다, ., 다만, 현장, 근로자를, 을,
중심, 으론, 이, 같, 은, 법안, 이, 현장, 에, 제대로, 적용, 되, 기, 위하, 어서, 는, 회
사, 의, 수직적, 위계, 구조, 와, 인력, 부족, 문제, 등구조적, 문제, 가, 우선, 해결, 되,
어야, 하, ㄴ다는, 지적, 도, 나, 아, 오, ㄴ다, .]
----------------------------------------------------------
[감정노동자, 보호, 법, 은, 사업, 주로, 하여금, 감정노동, 으로부터, 근로자, 를, 보호,
하는, 예방, 조치, 를, 이행, 하도록, 강제, 한다, ., 다만, 현장, 근로자, 를, 을, 중심,
으론, 이, 같은, 법안, 이, 현장, 에, 제대로, 적용, 되기, 위, 해서는, 회사, 의, 수직,
적, 위계, 구조, 와, 인력, 부족, 문제, 등, 구조, 적, 문제, 가, 우선, 해결, 돼야, 한다
는, 지적도, 나온다, .]
----------------------------------------------------------
[감정, 노동자, 보호법, 은, 사업주, 로, 하여금, 감정, 노동, 으로, 부터, 근로자, 를, 보
호, 하, 는, 예방, 조치, 를, 이행, 하, 도록, 강제, 하, ㄴ다, ., 다만, 현장, 근로자,
를, 을, 중심, 으로, ㄴ, 이, 같, 은, 법안, 이, 현장, 에, 제대, 로, 적용, 되, 기, 위하,
어서, 는, 회사, 의, 수직적, 위계, 구조, 와, 인력, 부족, 문제, 등, 구조적, 문제, 가,
우선, 해결, 되, 어야, 하, ㄴ다는, 지적도, 나오, ㄴ다, .]
```

['이', '전자책은', '저작권법에', '의하여', '보호', '를', '받는', '저작물이므로', '무단전재와', '무단복제를', '금합니다', '이를', '위반시에는', '형사', '민사상의', '법적책임을', '질', '수', '있습니다'],
['보아구렁이', '는', '먹이를', '씹지도', '않고', '통째로', '삼킨다', '그러고는', '먹이가', '다', '소화', '될', '때까지', '여섯', '달', '동안', '잠을', '잔다'],

MECAB

```
import MeCab
m = MeCab.Tagger()
out= m.parse("미캅이 잘 설치되었는지 확인중입니다.")
print(out)
```

```
미      NNP,인명,F,미,*,*,*,*
캅      NNP,인명,T,캅,*,*,*,*
이      JKS,*,F,이,*,*,*,*
잘      MAG,*,T,잘,*,*,*,*
설치    NNG,행위,F,설치,*,*,*,*
되      XSV,*,F,되,*,*,*,*
었      EP,*,T,었,*,*,*,*
는지    EC,*,F,는지,*,*,*,*
확인    NNG,행위,T,확인,*,*,*,*
중      NNB,*,T,중,*,*,*,*
입니다  VCP+EF,*,F,입니다,Inflect,VCP,EF,이/VCP/*+ㅂ니다/EF/*
.       SF,*,*,*,*,*,*,*
EOS
```

V. 프로젝트 향후 고민

02. 텍스트 키워드 알고리즘

2.1 그래프 랭킹 기반 비지도 학습 방식

TEXTRANK

```
summary
```

```
['개인의 노력과 능력만을 성공의 조건으로 삼는 것은 부당 하다고 말하는 이도 있을 것이다. 맞는 말이다.',
 '시간이 흘러 드디어 선택의 순간이 찾아왔다.',
 '피땅 흘려 마련한 집을 헐값에 처분하고 빚을 갚았더니 수중에는 2000만 원만 달랑 남았다.',
 '돈을 벌려고 하는 사람보다 돈과 상관없이 자기 일을 좋아하는 사람이 부자가 될 확률이 더 높다는 것이다.',
 '차 동 엽 신부가 쓴 책 『 잊혀 진 질문 』에는 이런 글이 나온다.',
 '" 무슨 일을 하든지 그 자체를 즐겨 라. 배를 곯을지언정 의미 없는 일은 하지 마라. 돈만을 위하여 일하는 사람은 영혼을 잃기 쉽다.',
 '그래서 그들의 머릿속에는 ' 어떻게 하면 더 많은 돈을 벌 것인가 ' 에 대한 궁리로 가득 차 있다.',
 '당시 리포트는 성적에 영향을 주는 것이 아니어서 대부분 형식적인 답변을 써냈는데 그중 3% 의 학생들 만이 성의 있는 리포트를 제출했다.',
 '그녀는 성공을 원하는 사람들에게 이렇게 말한다.',
 '이 소식을 들은 원은 즉시 투항했다.']
```

```
keywords
```

['사람', '때문', '은행', '생각', '당사', '직원', '시간', '모두', '자신', '정도', 미국']

TEXTRANK(gensim.summarization)

```
summ = summarize(text=data, ratio=1,word_count=10, split=True)
```

```
summ
```

```
['하지만 아무리 어려운 환경이라고 해도 긍정적으로 세상을 바라보고 자신을 변화시키는 긍정의 턴어라운드를 해나간다면 분명 답은 있을 것이다.']
```

```
gensim.summarization.keywords(data,split=True, words=10,deacc=True, scores=True)
```

```
[('것이다', 0.48377804680117914),
 ('하지만', 0.23396110537070064),
 ('때문이다', 0.21045461088196094),
 ('있었다', 0.14689738208886677),
 ('그러나', 0.1427712750786367),
 ('그래서', 0.11942384937136219),
 ('때문에', 0.11472574344562035),
 ('이렇게', 0.10674269700512687),
 ('아니라', 0.10404199448610353),
 ('사람을', 0.09522981848539334)]
```

V. 프로젝트 향후 고민

02. 텍스트 키워드 알고리즘

2.1 그래프 랭킹 기반 비지도 학습 방식

WORDRANK

sents

```
['자신이 좋아하고 잘하는 일을 하니까 일을 즐기게 되고 그 일이 남까지 기쁘게 하다 보니 자연스레 많은 돈을 벌게 된다는 것이다',
 '내가 옳은 방향으로 가고 있는지 내 생각이 과연 맞는 건지 자신을 스스로 점검하고 반성하는 시간의 중요성을 그들 덕분에 배울 수 있었던 것
이다',
 '책 빌 게이츠는 왜 생각 주간을 만들었을 까 중에서 우리는 지금까지 남다른 성공을 위해 많은 것을 투자해 왔다',
 '당시 내가 몸담고 있던 은행은 서울에 지점이 하나밖에 없었는데 경쟁이 치열하다 보니 몇 년이 지나면 다시 지방 지점으로 내려가야 했다',
 '하지만 나의 결정으로 단 한 명이라도 자리를 지킬 수 있다면 조금이라도 위안을 받는 직원이 있다면 그것으로 충분하다고 생각했다',
 '하나는 내가 원하는 사람들을 모아 팀을 구성할 수 있다는 것이고 다른 하나는 함께 일하는 직원들의 급여를 내 마음대로 정할 수 있다는 것이
었다',
 '은행원이 되어 돈을 많이 벌어야 한다고 생각하면서도 한편으로는 대학에 들어가고 싶다는 생각이 나를 괴롭힌 것이다',
 '사실 그때만 해도 나는 초급 행원이었기 때문에 맡은 업무가 적었고 할 수 있는 일이 별로 없어 야근을 하는 경우가 드물었다',
 '그 결과 우리 지점은 당시 전국 공항 지점을 통틀어 1만 명이라는 가장 많은 급여 이체 고객을 확보할 수 있었다',
 '주변 사람들이 나를 믿어 주는 마음과 나의 성공을 바라는 마음이 하나둘 모여서 혼자서는 결코 이룰 수 없는 성공이 나에게로 돌아오는 것이
다']
```

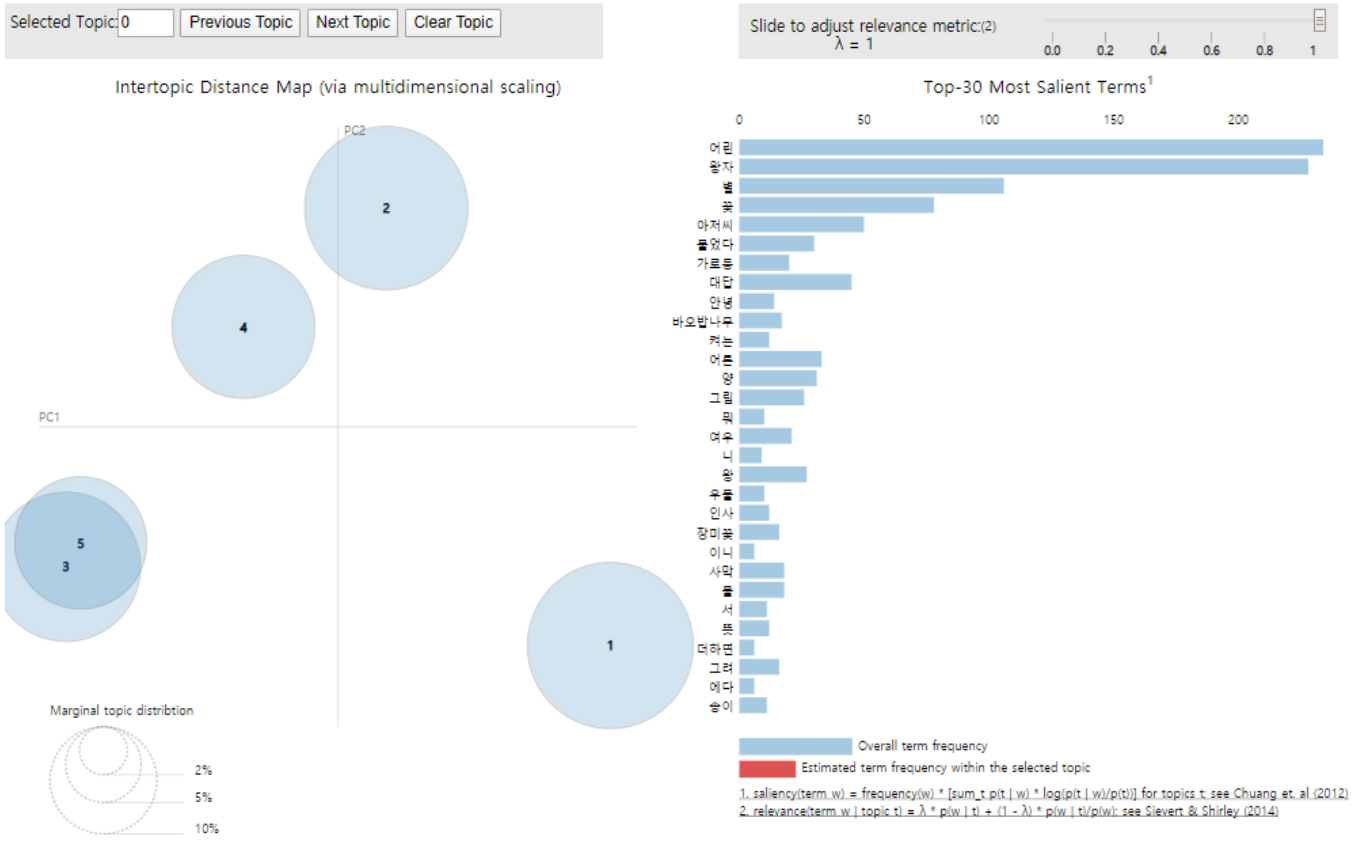
keywords

```
{'것이': 18.072165009151657,
 '사람': 15.332817423560789,
 '돈을': 8.259640530529442,
 '많은': 6.45227679401119,
 '있는': 6.4269229744415,
 '한다': 6.05742430341285,
 '다른': 5.926830538893686,
 '하는': 5.907351526278407,
 '했다': 5.88241818777538,
 '것은': 5.464855340807871,
 '위해': 5.447340439787785,
 '있었': 5.257610932898015,
 '가장': 5.246820795364047,
 '우리': 5.056857486004656,
 '은행': 4.9070774078209825,
 '그러나': 4.837036255536234,
 '그래서': 4.777747986606224,
```

V. 프로젝트 향후 고민

02. 텍스트 키워드 알고리즘

2.2 통계 학습 기반 비지도 학습 방식(LDA)



V. 프로젝트 향후 고민

03. 챗봇 학습

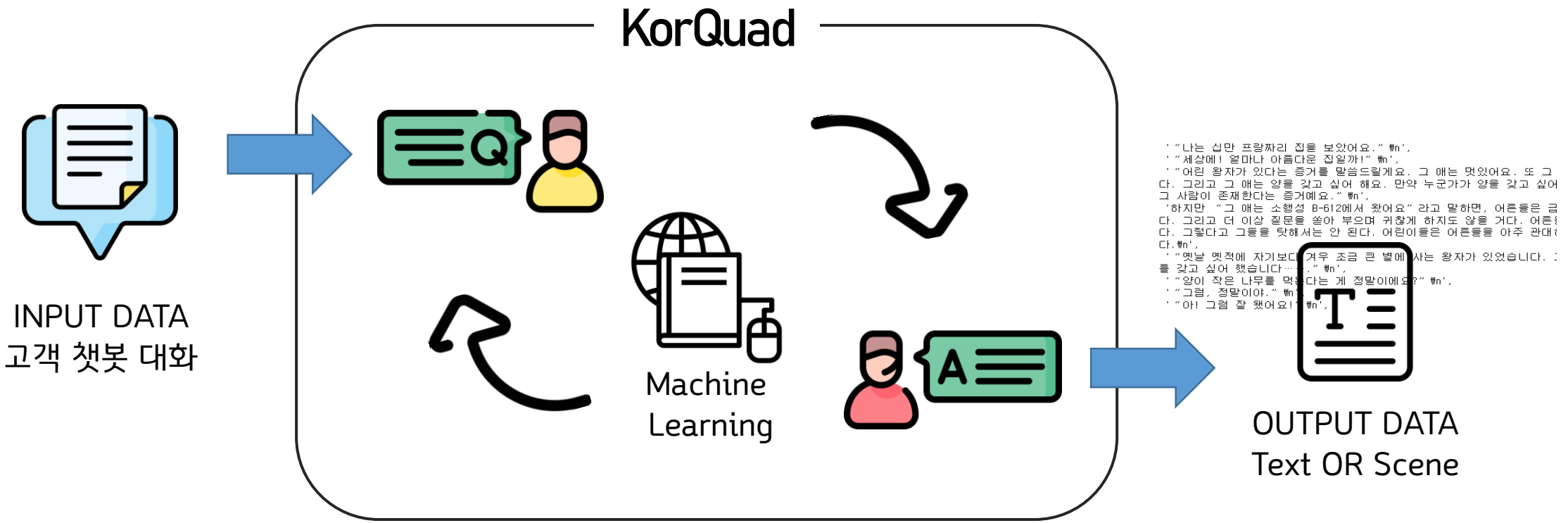
KorQuad 2.0: 웹문서 기계독해를 위한 한국어 질의응답 데이터셋

100,000+ 질의응답 쌍으로 구성
구조화된 문서를 모으기 위해 위키백과 문서들을 활용
2016/06/01 ~ 2019/05/31의 page view 상위 15만 문서를 선별
다양한 문서 도메인을 다루기 위해 임의 5만개 문서를 추가 HTML 문서 데이터를 수집

책 추천 챗봇 구현을 위해 사용
질문 [Input]에 대한 응답 [Output]의 keyword :: 책 텍스트 요약 keyword와 유사도 비교 후 추천

V. 프로젝트 향후 고민

03. 챗봇 학습



V. 프로젝트 향후 고민

03. 챗봇 학습

3. 챗봇 학습_데이터 대화체 학습

- 패턴 처리를 활용한 대화문 추출

- 대화문 사이의 문장 개수를 활용한 Scene 추출

고 해도 믿을 만한 꼬락서니였다.

"이번에도 삼 개월을 안 넘겼네"

"넘겼어. 하루."

성이가 정정했다. 상훈은 비닐봉지에 각 얼음을 담으며 장하다고 비웃었

[야... 일어나...벌써 해가 중천이여.]

시멘트 바닥에 매트리스만 깔린 창고 같은 곳에서 누워있는 교재를 발로 툭툭 친 남자가 억지로 교재를 깨웠다.
그대로 등을 '쳐' 먹고 사는 인간이었다.

「할매. 그런 말 말고 나 점이나 봐줘요.」

「공짜로는 안 돼. 복 나가.」

주머니를 뒤져봐도 돈이 있을 리가 만무했다. 하는 수 없이 근처에서 바

—그래도 우리를 상대로 이만큼 버틴 건 칭찬해 주지. 허나 그뿐이다. 네가 태초의 마법사라 불려도 우리 앞에선 어떤 상대도 되지 않는다.

'시끄러워.'

드래곤의 말에 대꾸를 하려 했지만 그럴 기력조차 남아 있지 않았다. 데

V. 프로젝트 향후 고민

03. 챗봇 학습

- 패턴 처리를 활용한 대화문 추출

```
#'ㅡ'은 예외로 처리
head_pattern = '[“ ‘₩(₩[₩<《 「 『]'
tail_pattern = '[” ’ ₩)₩]₩>》 」 』 ]'

#str=뽑아낼 원본문장 , df=전처리한문장(탐색용)
def spokenData(str,df):
```

- 대화문 사이의 문장 개수를 활용한 Scene 추출

```
#str=뽑아낼 원본문장 , df=전처리한문장(탐색용)
def SceneData(str,df):
```

```
 '“나는 십만 프랑짜리 집을 보았어요.” ₩n',
 '“세상에! 얼마나 아름다운 집일까!” ₩n',
 '“어린 왕자가 있다는 증거를 말씀드릴게요. 그 애는 멋있어요. 또 그 애는 웃었답니다. 그리고 그 애는 양을 갖고 싶어 해요. 만약 누군가가 양을 갖고 싶어 한다면, 그건 그 사람이 존재한다는 증거예요.” ₩n',
 '하지만 “그 애는 소행성 B-612에서 왔어요” 라고 말하면, 어른들은 금세 알아들을 거다. 그리고 더 이상 질문을 쏟아 부으며 귀찮게 하지도 않을 거다. 어른들은 이런 식이다. 그렇다고 그들을 탓해서는 안 된다. 어린이들은 어른들을 아주 관대하게 대해야 한다.₩n',
 '“옛날 옛적에 자기보다 겨우 조금 큰 별에 사는 왕자가 있었습니다. 그는 양 한 마리를 갖고 싶어 했습니다……,” ₩n',
 '“양이 작은 나무를 먹는다는 게 정말이에요?” ₩n',
 '“그럼, 정말이야.” ₩n',
 '“아! 그럼 잘 됐어요!” ₩n',

[씬시작]',
“이건 상자야. 네가 갖고 싶어 하는 양은 그 안에 들어 있어.” ',
그러자 놀랍게도 꼬마 심판의 얼굴이 갑자기 환하게 밝아지는 것이다.',
“내가 원한 게 바로 이거예요! 이 양을 먹이려면 풀이 많이 있어야겠죠?” ',
“왜?” ',
“내가 사는 곳은 너무 작으니까……,” ',
“그 정도면 충분할 거야. 내가 그려준 양은 아주 작거든.” ',
그는 고개를 숙이고 가만히 그림을 들여다보았다.',
“그렇게 작은 것 같진 않은데……. 이것 좀 봐요! 양이 잠이 들었어요……,” ',
[씬끝남]',
```