RIDI

자료분석특론II

기말 프로젝트 중간발표

RIDINOVEL 작품의 연결 관계 예측 연구

- 로맨스 웹소설을 중심으로 -

222STG10 김희숙

목차

01 RAW 데이터 소개

	cate	title	write	upload	tag	ro	detail
0	로맨스 웹소설 현대물	대척점	임은성 저\n에클라 출판\n총 82화완결	혜택\n무료이용권 1장 증정\n사용기한 30일, 1인 1회 발급 가능, 90일 대여	이 책의 키워드\n#기다리면무료\n#리뷰1000개이상\n#별점3000개이상\n#현대...	로맨스 가이드\n* 배경/분야: 현대물, 외국배경\n* 작품 키워드: 액션로맨스, ...	작품 소개\n착실한 모범생의 길을 걸어온 차혜주,\n30년 인생의 첫 일탈은 퇴사 ...
1	로맨스 웹소설 현대물	스프링 피버(Spring Fever)	백민아 저\n이지콘텐츠 출판\n총 65화미완결	연재\n매주 목,금,토,일,월 18시 연재	이 책의 키워드\n#기다리면무료\n#리뷰1000개이상\n#별점3000개이상\n#현대...	로맨스 가이드\n*배경/분야: 현대로맨스\n*작품 키워드: 현대물, 직진남, 능글남...	작품 소개\n*본 작품은 리디 웹소설에서 동일한 작품명으로 19세이용가와 15세이용...
2	로맨스 웹소설 현대물	폐쇄 병동의 주의 사항	이정운 저\n로즈엔 출판\n총 100화완결	혜택\n무료이용권 3장 증정\n사용기한 30일, 1인 1회 발급 가능, 90일 대여	이 책의 키워드\n#기다리면무료\n#리뷰1000개이상\n#별점3000개이상\n#현대...	로맨스 가이드\n*배경/분야: 현대물\n*작품 키워드:\n현대물, 동거, 오해, 복...	작품 소개\n“정신 병원에 잠입해서, 거기에 억울하게 갇힌 이분을 데리고 탈출해 주...

ro

로맨스 가이드
* 배경/분야: 현대물, 외국배경
* 작품 키워드: 액션로맨스, 생존로맨스, 조직/암흑가, 군대물, 친구>연인, , 운명적사랑, 뇌섹남, 능력남, 사이다남, 직진남, 계략남, 능글남, 다정남,
* 남자 주인공:
이도영(30) - 매력적인 얼굴 뒤로 수많은 비밀을 숨긴 남자. ‘심부름’을 하러 아르헨티나에 갔다가, 그곳에서 하룻밤을 함께 보낸 차혜주에게 공항까지 안전하게 데려
* 여자 주인공:
차혜주(30) - 통제광인 부모님 아래서 스스로를 억누르고 살아온 여자. 첫 일탈을 감행한 아르헨티나에서 살인 사건에 휩싸이고, 얼떨결에 이도영과 동행하게 된다.
* 이럴 때 보세요: 광활한 대륙을 배경으로 한 긴장감 넘치는 추격전과 그 속에서 피어나는 섹슈얼 로맨스를 함께 즐기고 싶을 때
* 공감 글귀: “친구 하자는 건 다 개수작이었어, 혜주야.”
펼쳐보기

detail

작품 소개
착실한 모범생의 길을 걸어온 차혜주,
30년 인생의 첫 일탈은 퇴사 후 가장 먼 나라로 여행을 떠나는 것이었다.

그렇게 도착한 이국에서 만난 한 남자.
원색의 도시를 배경으로, 온통 검은 옷을 입은 남자는 처음부터 시선을 잡아끌었다.

타앙, 고막이 멀 듯한 총성.
피를 뒤집어쓴 차혜주.
그녀를 끌고 도망치는 남자.

“나, 이 방 같이 쓰게 해 줘요.”

정체를 알 수 없는 남자, 이도영의 입가가 근사한 호선을 그렸다.

02 전처리 진행 상황

	cate	title	write	upload	tag	ro	detail
0	로맨스 웹소설 현대물	대척점	임은성 저\n에클라 출판\n총 82화완결	혜택\n무료이용권 1장 증정\n사용기한 30일, 1인 1회 발급 가능, 90일 대여	이 책의 키워드\n#기다리면무료\n#리뷰1000개이상\n#별점3000개이상\n#현대...	로맨스 가이드\n* 배경/분야: 현대물, 외국배경\n* 작품 키워드: 액션로맨스, ...	작품 소개\n착실한 모범생의 길을 걸어온 차혜주,\n30년 인생의 첫 일탈은 퇴사 ...

- **cate**

```
ridi_re['cate'].unique()
```

```
array(['현대물', '역사/시대물'], dtype=object)
```

→ '로맨스 웹소설' 지움

- **title**

```
ridi_re[ridi_re['title'].str.contains('스프링 피버')]['title']
```

```
1                스프링 피버(Spring Fever)
35    스프링 피버(Spring Fever) (15세 개정판)
```

```
print(ridi_re[ridi_re['title'].str.contains('15세')]['title'])
ridi_re_up15 = ridi_re[~ridi_re['title'].str.contains('15세')]
ridi_re_up15.shape
```

(53, 7)

→ '15세 개정판'으로 중복 데이터 존재
→ drop

- **write**

임은성 저\n에클라 출판\n총 82화완결 →

writer	cnt
임은성	82

→ 작가명과 연재 횟수만 추출하여 새로운 변수에 저장(writer, cnt)

- **upload**

```
ridi_re['upload'].unique() #54개 변수에 대해
```

```
array(['혜택 무료이용권 1장 증정 사용기한 30일, 1인 1회 발급 가능, 90일 대여',
       '연재 매주 목,금,토,일,월 18시 연재',
       '혜택 무료이용권 3장 증정 사용기한 30일, 1인 1회 발급 가능, 90일 대여',
       '연재 본 작품은 매주 금, 토, 일, 월, 화 오전 7시 연재 됩니다.', '추천 소설 e북으로 보기',
       '공지 15세 이용가 안내', '연재 매주 월,화,수,목,금,토,일 18시 연재',
       '추천 이 작품 다양하게 즐기기 2 작품', '연재 매주 월,화,수,목,금 연재',
       '연재 매주 월,화,수,목,금,토 연재', '연재 매주 월, 화, 수, 목 연재', '공지 도서 파일 교체 안내',
       '연재 매주 월, 화, 수, 목, 금', '연재 매주 월,수,금 연재', '연재 매주 월, 화, 수, 목, 금 연재'],
      dtype=object)
```

→ 유의미한 text가 없다고 판단
→ drop

- **tag**

→ 불필요한 단어 및 tag 제거, 카테고리와 관련 있는 tag 제거 (ex.현대물, 가상시대물, 동양풍 등)

- **ro**

→ 불필요한 문장 및 단어 제거
→ 카테고리와 관련 있는 내용 제거
→ 공감글귀 따로 column생성

- **detail**

ro	detail
작품 소개\n한 번 취한 궁녀는 다시 찾지 않는 다는 세자, 광안.\n군호는 빛 광...	저자 프로필\n라혜\n작가 신간알림 소식\n2017.06.14. 업데이트 작가 프로...

→ 'ro'로 크롤링된 작품소개를 detail로 옮김
→ 불필요한 단어 및 엔터 삭제(작품소개, ₩n 등)
→ 한문장으로 변환

03 현재 최종 데이터

```
ridi_re_up15.head(3)
```

	cate	title	tag	ro	detail	writer	cnt
0	현대물	대척점	#리뷰1000개이상 #별점3000개이상 #현대물 #친구>연인 #다정남 #절륜남 #계...	로맨스 가이드 * 배경/분야: 현대물, 외국배경 * 작품 키워드: 액션로맨스, 생존...	착실한 모범생의 길을 걸어온 차혜주, 30년 인생의 첫 일탈은 퇴사 후 가장 먼 나...	임은성	82
1	현대물	스프링 피버 (Spring Fever)	#리뷰1000개이상 #별점3000개이상 #현대물 #로맨틱코미디 #상처남 #순정남 #...	로맨스 가이드 *배경/분야: 현대로맨스 *작품 키워드: 현대물, 직진남, 능글남, ...	*본 작품은 리디 웹소설에서 동일한 작품명으로 19세이용가와 15세이용가로 동시 서...	백민아	65
2	현대물	폐쇄 병동의 주의 사항	#리뷰1000개이상 #별점3000개이상 #현대물 #동거 #복수 #순정남 #다정남 #...	로맨스 가이드 *배경/분야: 현대물 *작품 키워드: 현대물, 동거, 오해, 복수, ...	“정신 병원에 잠입해서, 거기에 억울하게 갇힌 이 분을 데리고 탈출해 주십시오.” ...	이정운	100

04 문제점

- **데이터**
  - 크롤링을 통한 데이터 수집 시, 인기도를 기반으로 데이터가 실시간으로 바뀌어 중복되거나 누락되는 데이터가 존재.
  - 현대물과 역사/시대물 데이터 불균형. (현대물:역사/시대물 = 5:1 정도)

- **전처리**
  - tag: '평점4점이상', '리뷰1000개이상', '별점1000개이상' 등 태그 처리에 대해 고민중. Drop 고민.
  - ro: 주인공에 대한 정보가 랜덤하게 규칙없이 포함되어 있어, 이를 전처리 하는 것에 문제가 있음.
    (지난주에 인지했던 상황보다 더 심각함.)
  - detail: 불필요한 문장이 랜덤하게 규칙없이 포함되어 있어, 이를 제거하는 것에 문제가 있음.

```
ridi_re_up15[ridi_re_up15['detail'].str.contains('본 작품')]['detail']
```

```
1      *본 작품은 리디 웹소설에서
8      *본 작품은 리디 웹소설에서
15     ※본 작품은 관계 중 호불호
18     ※본 작품은 작가의 상상에
26     ※본 작품은 강압적 관계 및
29     ※ 본 작품은 리디 웹소설에
30     ※본 작품은 주인공의 서사가
43     *본 작품은 역사적 사실을
45       ※ 본 작품에는 다소 강압
```

05 모델링

KoBERT

- KoBERT
  - Korean BERT pre-trained cased (KoBERT)
    - Why'?'
    - Training Environment
    - Requirements
    - How to install
  - How to use
    - Using with PyTorch
    - Using with ONNX
    - Using with MXNet-Gluon
    - Tokenizer
  - Subtasks
    - Naver Sentiment Analysis
    - KoBERT와 CRF로 만든 한국어 객체명인식기
    - Korean Sentence BERT
  - Release
  - Contacts
  - License

CharBERT: Character-aware Pre-trained Language Model

Wentao Ma[†], **Yiming Cui**[‡†], **Chenglei Si**[¶†], **Ting Liu**[‡], **Shijin Wang**[†§], **Guoping Hu**[†]
[†]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[‡]Research Center for Social Computing and Information Retrieval (SCIR),
Harbin Institute of Technology, Harbin, China
[§]iFLYTEK AI Research (Hebei), Langfang, China
[¶]University of Maryland, College Park, MD, USA
[†§]{wtma,ymcui,clsi,sjwang3,gphu}@iflytek.com
[‡]{ymcui,tliu}@ir.hit.edu.cn

Abstract

Most pre-trained language models (PLMs) construct word representations at subword level with Byte-Pair Encoding (BPE) or its variations, by which OOV (out-of-vocab) words are almost avoidable. However, those methods split a word into subword units and make the representation incomplete and fragile. In this paper, we propose a character-aware pre-trained language model named **CharBERT** improving on the previous methods (such as BERT, RoBERTa) to tackle these problems. We first construct the contextual word embedding for each token from the sequential character representations, then fuse the representations of characters and the subword representations by a novel heterogeneous interaction module. We also propose a new pre-training task named NLM (Noisy LM) for unsupervised character representation learning. We evaluate our method on question answering, sequence labeling, and text classification tasks, both on the original datasets and adversarial misspelling test sets. The experimental results show that our method can significantly improve the performance and robustness of PLMs simultaneously.[1]