Team 7
추천모델 Prj.

소영이

소설과 영화를 추천해 드립니다

최규형 이정현 지정원

목차

프로젝트 개요

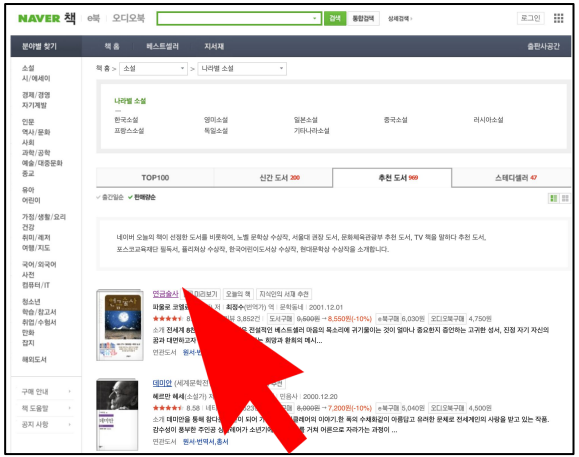
- NAVER 소설 리뷰 문자열 데이터 수집
- 기존 2017-2020 NAVER 영화 리뷰 문자열 데이터와 병합
- 병합된 리뷰 데이터 전처리하여 Word2Vec 추천 모델 생성
- 좋아하는 영화 선택 또는 입력 시, 유사도 높은 순으로 [영화], [책] 각각 10개씩 추천

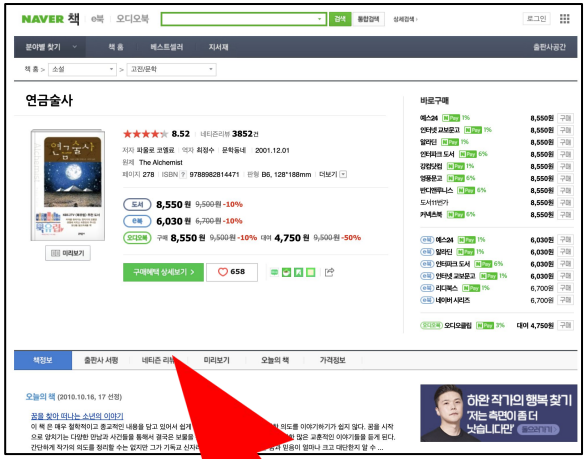
[데이터 수집 및 활용]

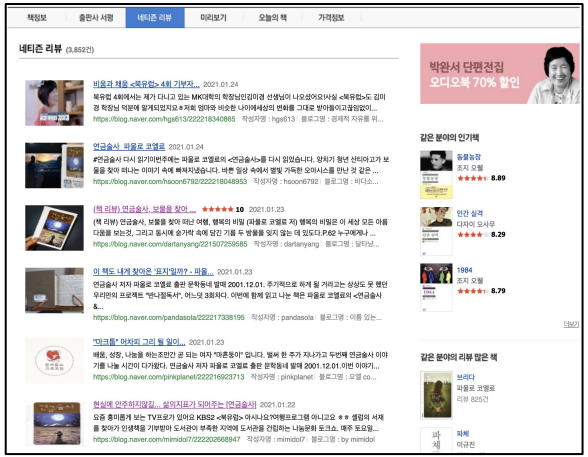
NAVER 책 > 소설 > 나라별 소설 > 추천도서 969권(판매량 순)에 대한 다음 항목

- book_title
- book_url
  - 책 소개 웹페이지 주소 > 추천 결과에 hyperlink 삽입
- book_review
  - 100건 초과할 경우 100건까지만 수집

구현기능 상세_웹크롤링







book_title

book_url

blog_url

네이버 블로그 주소에 접근하였으나,

웹크롤링 불가능한 문제 발생 (본문 요소가 숨겨져 있음)

구현기능 상세_웹크롤링

사용자에게 노출되는 블로그 URL = **Fake URL** / 웹크롤링 가능한 블로그 URL = **Real URL**

Fake URL vs. Real URL

- real_url = 'https://blog.naver.com' + '#mainFrame의 src 속성값
- **Real URL** BeautifulSoup 객체 생성 > 'div#postViewArea' 접근

Hidden Tag

- hidden_tag = 'div#post-view' + Real URL에 포함된 logNo 숫자값
- **Real URL** BeautifulSoup 객체 생성 > 'div#post-view{logNo}' 접근

Loaded Contents

- 네이버 블로그 본문을 네이버 책 페이지로 불러온 형태
- **Fake URL** BeautifulSoup 객체 생성 > 'div.rvw_cnt' 접근

*Naver Cafe 게시글 및 비공개/삭제 처리된 블로그 포스트 > return None (손실률 3-5% 수준)

구현기능 상세_데이터 전처리

전처리된 기존 영화 리뷰 데이터와 동일 방식으로 전처리 진행

- movie_review sample size = 영화 2279편
  - 영화 리뷰 하나당 평균 token 갯수 = 6496개
- novel_review sample size = 소설 835권
  - 소설 리뷰 하나당 평균 token 갯수 = 9511개

영화와 소설 아이템 구분을 위한 text tagging

- 영화/소설 각 데이터셋의 title 문자열 앞에 '영화' 또는 '소설' 문자열 삽입
- column name 통일 후 pd.concat

stopwords 추가

- Ranks.nl 등 3개 출처로부터 stopwords 추가 수집
- 기존 stopwords에 중복 제외하고 추가: (기존)619개 > (변경)921개
- 영화+소설 취합 데이터의 리뷰 하나당 평균 token 갯수 감소 확인
  - (기존 stopwords) 평균 7305개 > (추가 stopwords) 평균 7051개

구현기능 상세_모델링

학술저널 Top 10%

한국어 단어 임베딩을 위한 Word2vec 모델의 최적화

Optimization of Word2vec Models for Korean Word Embeddings

http://journal.dcs.or.kr/_common/do.php?a=current&b=11&bidx=1542&aidx=19540

한국어의 언어학적 특성을 반영한 유추 검사를 이용해서 하이퍼파라미터 최적화 시도

- 학습 알고리즘으로는 skip-gram 방식이 CBOW보다 우수
- 단어 벡터의 크기는 300 차원이 적절
- 문맥 윈도의 크기는 5에서 10 사이가 적절
- 최소 출현빈도 값은 총 어휘 수가 100만개 이하일 경우에는 1로 설정
  - 가급적 학습될 어휘 수를 적정 수준으로 유지하는 것이 중요

영화리뷰+소설리뷰 > 학습하고자 하는 말뭉치(token) 갯수 = 약 60만 개 수준

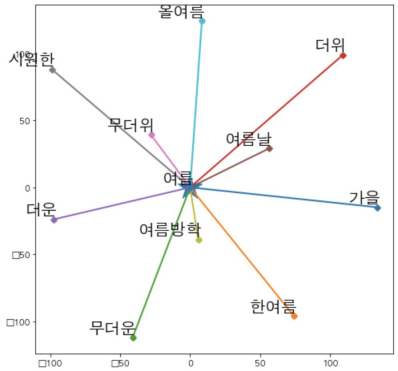
Word2Vec(cleaned_tokens, size=300, window=7, min_count=1, workers=4, iter=100, sg=1) > 1.5GB

구현기능 상세_시각화

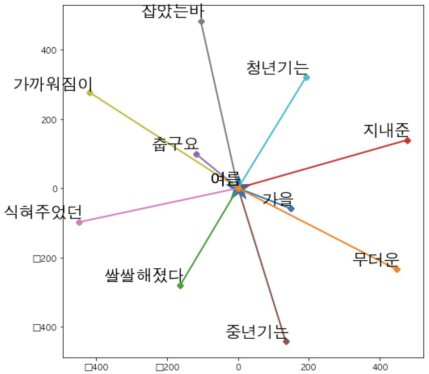
기존 영화 모델 대비 생활어/감성어에 더 가까운 similar word를 보여줌

- 어떤 말뭉치가 1번이라도 출현했다면 모두 학습시킴 (min_count=1)
- 어떤 말뭉치와 유사한 말뭉치가 무엇인지 학습할 때 더 넓은 범위를 탐색 (window=7)

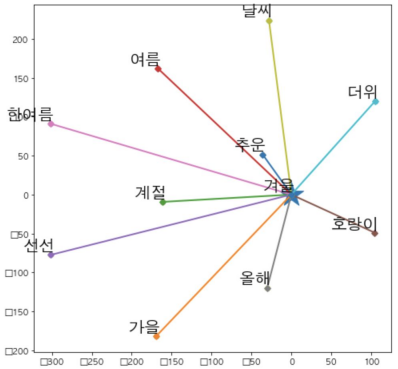
기존 Word2Vec "여름"



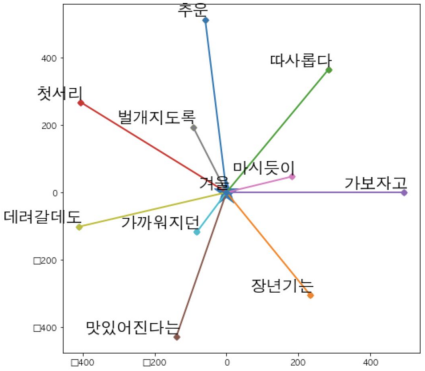
변경 Word2Vec "여름"



기존 Word2Vec "겨울"



변경 Word2Vec "겨울"

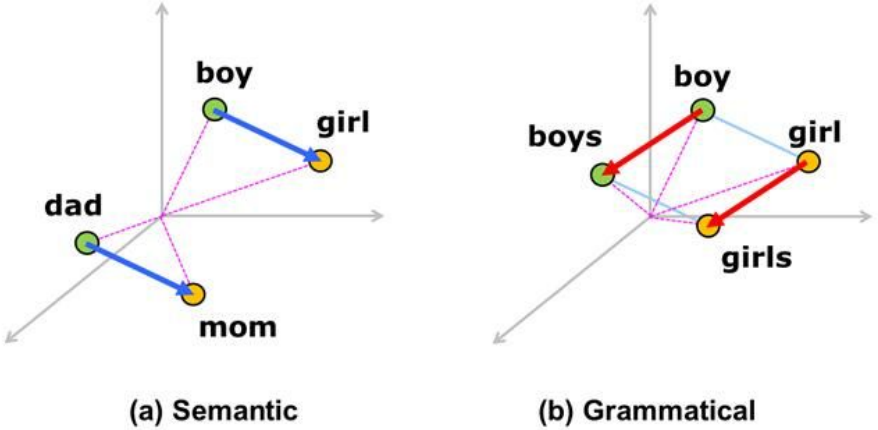


리뷰 문자열에 담긴 사용자 감성(sentiment) 요소를 추천 결과에 반영할 것으로 기대할 수 있음

구현기능 상세_성능평가

유추 검사

- Word2Vec 모델의 성능을 평가하는 방법 / 의미론적, 문법적 관계를 검사
  - 소년 : 소녀 = 남자 : 여자



(a) Semantic

(b) Grammatical

'아빠'에 대한 '엄마' 좌표 차이

- 43.713425 , 17.15274
- distance = 46.9582

'소년'에 대한 '소녀' 좌표 차이

- 31.804937, -241.26982
- distance=243.3571

구현기능 상세_어플리케이션

기존 영화추천 어플리케이션과 동일기능 구현

- 드랍다운 리스트 또는 텍스트 입력을 통하여 item 선택
  - 추천 영화 10개, 추천 소설 10개 title을 구분된 Label 영역에 출력
- 사용자에게 추천 title을 보여줄 때 > text tag를 삭제하고 본래 title만 출력
  - (예시) '소설 연금술사' >>> '연금술사'

사용자 편의기능 추가 구현

- 소설 추천 title의 경우, 해당 소설에 대한 Naver Book 웹페이지 하이퍼링크 삽입(하려고 했는데 못함)

작업 결과물 시연