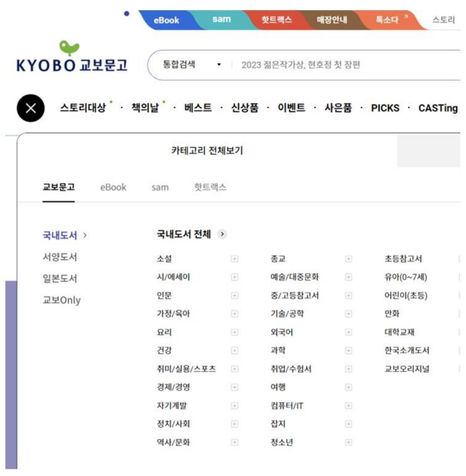
빅데이터 분석을 통한 트렌드 파악 및 사용자 맞춤 도서 추천

윤경서, 강승식
국민대학교 AI빅데이터융합경영학과, 국민대학교 소프트웨어학부

한국정보처리학회 Korea Information Processing Society

학부생 경진대회 논문

02. 데이터 수집 및 전처리



데이터 수집

교보문고에서 국내도서 카테고리 중 20개를 선정하여
해당 카테고리별 **베스트셀러 240권씩**

제목, 저자, 내용요약, 카테고리를

Selenium을 사용한 크롤링으로 데이터 수집을 진행

= 총 4800권의 데이터 수집

전처리

re 라이브러리의 정규표현식을 사용하여 전처리 진행

03. 형태소 분석

KoNLTK	KoNLPy	Stanza
KLT2023	OKT	nouns 함수
4800권의 도서에 대해	4800권의 도서에 대해	4800권의 도서에 대해
소요시간 : 31초	**소요시간** : 1분 33초	**소요시간** : 22분 29초
약 45만개의 명사 추출	약 54만개의 명사 추출	약 30만개의 명사 추출

KoNLTK로 명사만 추출한 위키피디아 데이터 (ko_wiki_KMA_utf8.txt) +

전처리

1. re라이브러리 정규표현식

2. 불용어 처리 →

약 3006만개의 명사로 Word2Vec model & FastText model 생성

04. 책 내용 벡터 생성

책 내용 추출
크롤링 원본 데이터에서 내용요약본에 해당하는 부분만 추출 →

책 내용 전처리
re라이브러리 정규표현식 사용 + 불용어 처리 →

책 내용 명사 추출
klt2023 형태소 분석기를 사용하여 명사만 추출 →

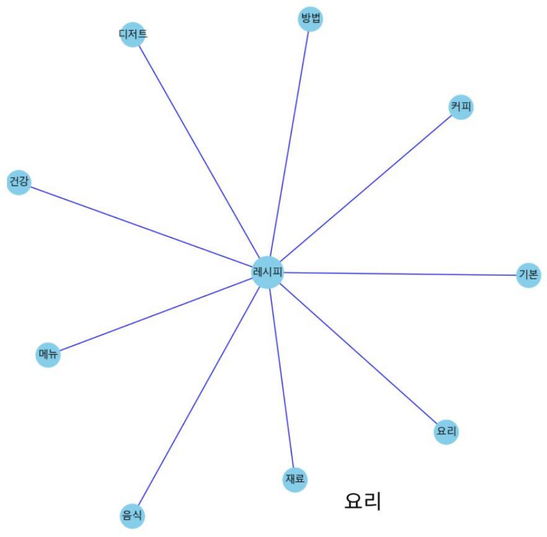
명사 중복제거
앞서 추출한 명사들에 대해 각 책 내용별 중복 제거 →

명사 벡터화
위키피디아 데이터로 만든 FastText model로 각 책 내용별 명사를 벡터화한 후 더함 →

모든 책에 대해 책 내용 벡터 생성

05. 워드클라우드 CountVectorizer

06. 키워드 시각화 plot4kcc_keyword



건강
건강
환자, 운동, 소개, 사람, 제시

자기계발
자기계발
인생, 성공, 사람, 방법, 출간, 저자

종교
종교
교회, 사람, 하나님, 이야기, 성경

예술/대중문화
작품
음악, 그림, 영화, 이야기, 사랑, 설명

외국어
외국어
영어, 학습, 단어, 문제, 문법, 표현

여행
여행
지역, 여행자, 추천, 가이드북, 정보, 소개

07. 임베딩 모델 활용 도서 추천

책 내용으로 추천 사용자가 문장을 입력하면 해당 문장과 책 내용 벡터의 유사도를 구해 유사도가 가장 높은 책 상위n권을 추천해준다.

마음을 편안하게 해주는 위로의 글

Word2Vec
```
[(('행복을 담아줄게', '나란'), 0.35825759059687917),
 (('있는 그대로', '김지훈'), 0.3403578870621218),
 (('보이지 않는 곳에서 애쓰고 있는 너에게', '최대호'), 0.33851055987823103),
 (('당신과 아침에 싸우면 밤에는 입맞출 겁니다', '유래혁'), 0.3367892482086697),
 (('엄마의 말 공부 일력 365', '이임숙'), 0.3345570815563324)]
```

FastText
```
[(('당신이 좋아지면, 밤이 깊어지면', '안희연'), 0.7639079225989486),
 (('나로서 충분히 괜찮은 사람', '김재식'), 0.7610791092550505),
 (('당신과 아침에 싸우면 밤에는 입맞출 겁니다', '유래혁'), 0.7595994689581971),
 (('나는 당신이 행복했으면 좋겠습니다', '박찬위'), 0.7592670581220978),
 (('그대 늙어가는 것이 아니라 익어가는 것이다', '오평선'), 0.7533357641240681)]
```

단어 조합으로 추천 사용자가 positive word와 negative word를 입력하면 해당 단어들의 조합으로 유사 단어를 추출한 뒤, 유사 단어들의 벡터 합과 책 내용 벡터간의 유사도를 구해 유사도가 가장 높은 상위 1권의 책을 추천해준다.

Word2Vec
```
positive_word (여러 단어를 쓸 경우, 띄어쓰기로 구분해주세요.):모험 전쟁 성공
negative_word (여러 단어를 쓸 경우, 띄어쓰기로 구분해주세요.):공주 연애

[(('이미 시작된 전쟁', '이철', '정치/사회'), 0.27360490246192454)]
```

FastText
```
positive_word (여러 단어를 쓸 경우, 띄어쓰기로 구분해주세요.):모험 전쟁 성공
negative_word (여러 단어를 쓸 경우, 띄어쓰기로 구분해주세요.):공주 연애

[(('로마인 이야기 2: 한니발 전쟁', '시오노 나나미', '역사/문화'), 0.5823006581838892)]
```

07. 임베딩 모델 활용 도서 추천

카테고리로 추천

사용자가 단어를 입력하면 해당 단어와 각 카테고리에 해당하는 단어간의 유사도를 구한 뒤, 가장 유사도가 높은 카테고리의 책들의 내용벡터와 입력 단어 벡터와의 유사도를 구해 유사도가 높은 상위 n권을 추천해준다.

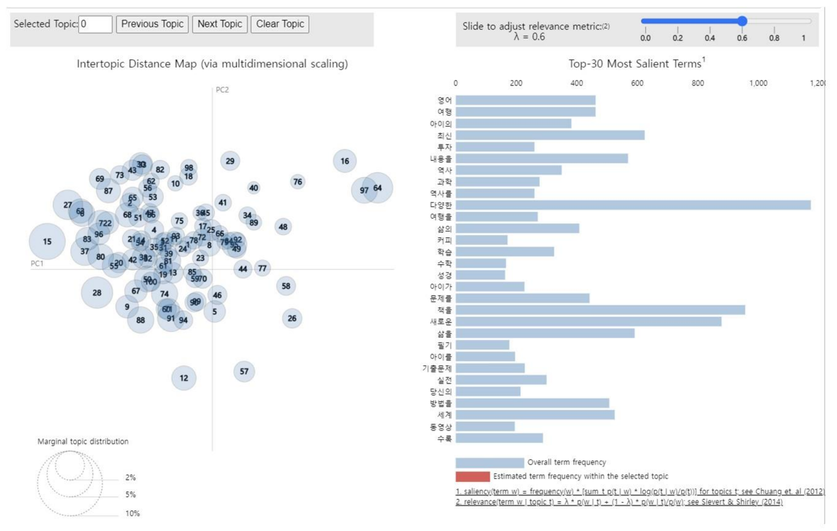
성공 (상위 5개)
```
(0.04292066,
 '대중문화',
 [(('SAVE THE CAT!: 흥행하는 영화 시나리오의 8가지 법칙', '블레이크 스나이더'), 0.44993079561155186),
  (('우리는 왜 임영웅을 사랑하는가', '조위'), 0.34890712733005674),
  (('시나리오 어떻게 쓸 것인가 세트', '로버트 맥키'), 0.343034047968187),
  (('데이비드 호크니', '마르코 리빙스턴'), 0.327195837798465),
  (('연기하지 않는 연기', '해럴드 거스킨'), 0.31482199930558946)])
```

모험 (상위 5개)
```
(0.1325179,
 '소설',
 [(('해리 포터 시리즈 1~4권 세트(해리포터 20주년 개정판)', 'J. K. 롤링'), 0.6852860696465063),
  (('해리 포터와 마법사의 돌 1(해리포터 20주년 개정판)', 'J. K. 롤링'), 0.6755744803449041),
  (('해리 포터와 마법사의 돌 2(해리포터 20주년 개정판)', 'J. K. 롤링'), 0.6755744803449041),
  (('해리 포터와 마법사의 돌(해리포터 20주년 개정판)', 'J. K. 롤링'), 0.6755744803449041),
  (('해리 포터와 아즈카반의 죄수 1(해리포터 20주년 개정판)', 'J. K. 롤링'), 0.6655232528503321)])
```

08. LDA 주제 분석

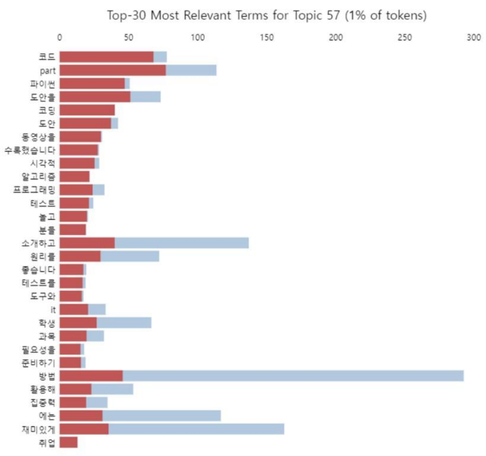
$\lambda = 0.6$

1에 가까울수록 토픽에서 가장 많이 나온 단어 기준
0에 가까울수록 전체적인 비율에 비해서 해당토픽에서 자주 나오는 횟수 기준

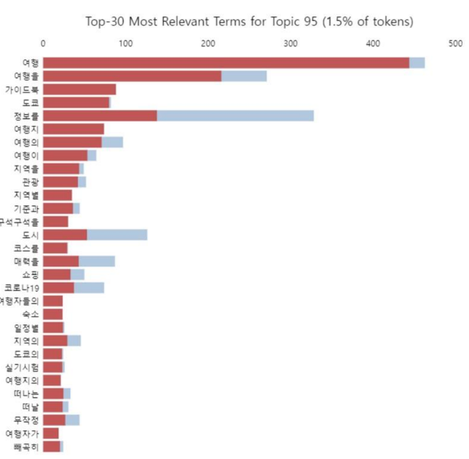


08. LDA 주제 분석

컴퓨터/IT



여행



결론

- 사용자에게 맞춤형 도서 추천을 위한 도서 데이터 분석 연구
- 도서 데이터를 수집하여 카테고리별 트렌드 파악
- 도서 데이터의 카테고리별 특징 파악
- 사용자 취향과 일치하는 도서 추천 연구 수행
- 워드임베딩 기법을 이용하여 도서 추천의 효율성 향상
- 도서 데이터 분석을 통해 트렌드 및 키워드 파악
- 독자들의 요구사항에 적합한 도서 추천 방법 연구