2022-2 Capstone-Design2 최종 발표

키워드 추출 및 확장을 통한 검색 시스템

2016104112 김영빈

Contents

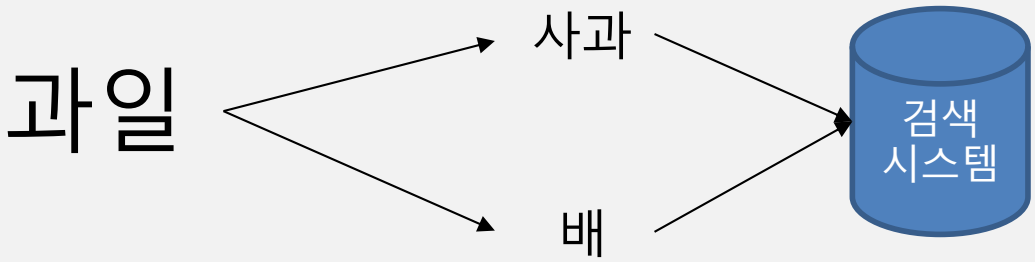
Introduction

- 텍스트 검색 도메인에서 검색어의 대한 확장을 통해 다양한 검색 결과를 얻고 있음
  - 연관어 온톨로지를 활용한 확장 검색 방식
  - 단어 임베딩을 활용한 확장 검색 방식

- 연관어 온톨로지를 활용한 방식의 경우 지속적으로 관리해줘야 한다는 문제 발생
- 단어 임베딩 등 자연어 처리를 활용한 확장 검색 방식이 연구되고 있음

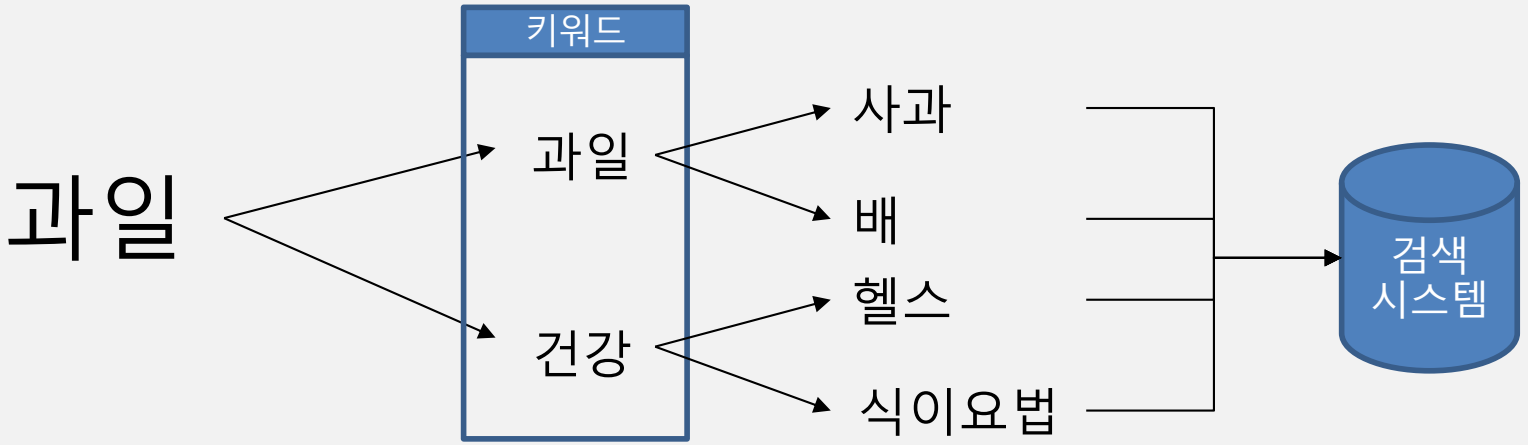
Problem Statement

- 기존 검색 방식
  - 사용자 질의어에 대해 확장한 단어들을 검색에 사용
  - ‘과일’이라는 질의어에 대한 확장 결과가 ‘사과’, ‘배’라면 이 두 가지 경우에 대한 검색 결과만 볼 수 있음
  - 검색 결과가 적다는 문제 발생



Approach

- 제안하는 검색 방식
  - 사용자 질의어에 대해 확장하지 않음
  - 질의어에 대한 최상위 검색 결과 문서의 ‘키워드’를 찾고 이를 확장하여 검색
  - ‘과일’이라는 질의어에 대한 검색 결과의 키워드가 ‘과일’, ‘건강’이라면 ‘사과’, ‘배’, ‘헬스’, ‘식이요법’ 으로 검색 가능



System

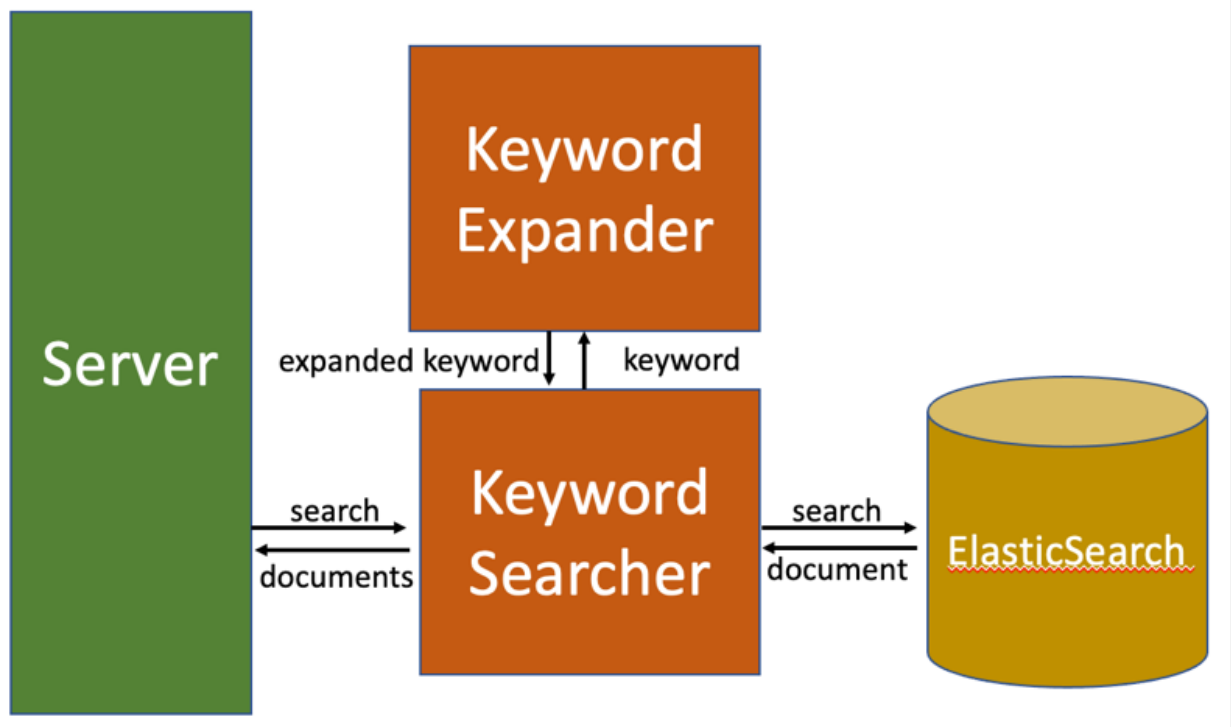
- 3가지 모듈로 구성
- Keyword Creator
  - 메타 데이터가 없는 문서들에 대해 메타 데이터로 문서의 키워드 추출
  - 키워드는 문서를 잘 나타내는 단어들
  - KeyBert 사용
- Keyword Expander
  - 문서의 키워드에 대해 연관어를 확장
  - 감성 단어인 경우 확장하지 않음 / 확장 결과는 가장 좋은 결과 1개만 제공
  - 한글 위키피디아 덤프 데이터를 학습한 Word2Vec 사용

System

- 3가지 모듈로 구성
- Keyword Searcher
  - ElasticSearch 검색 엔진을 사용하여 확장된 키워드들로 문서 검색
  - 단어의 일부분으로도 검색할 수 있도록 한글 형태소 분석기를 부착

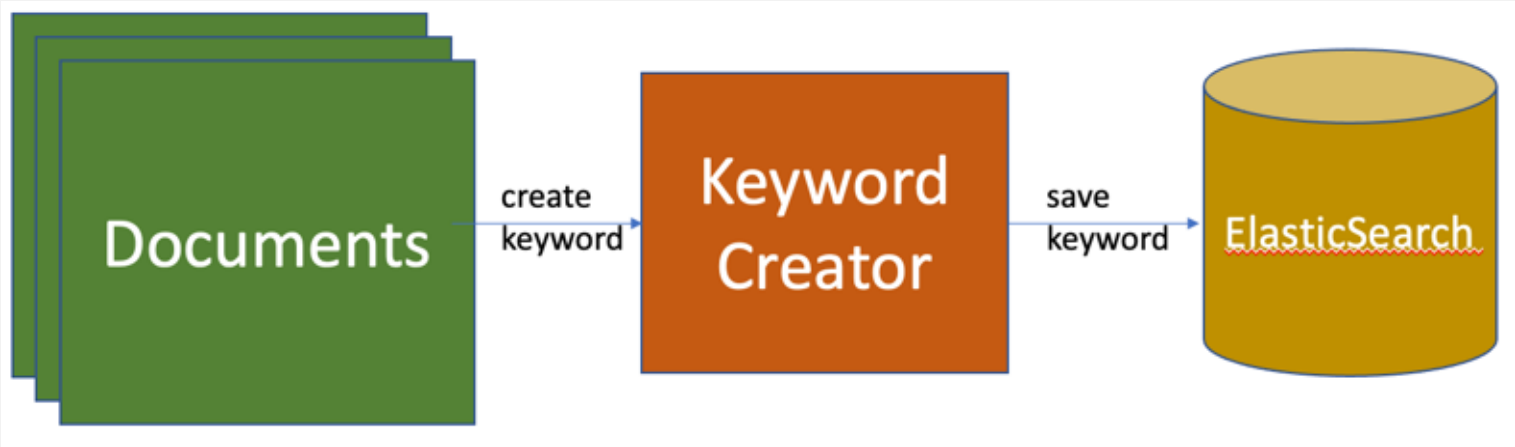
System Architecture

- 검색 단계 개념도



System Architecture

- 키워드 추출 단계 개념도



Restrictions & Solutions

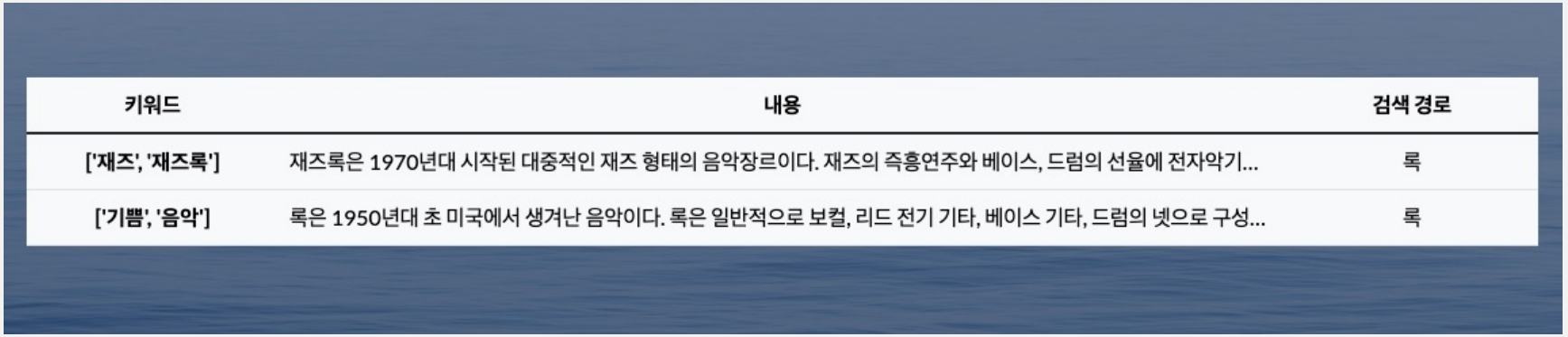
- 블로그 글 같이 메타 데이터가 없는 글은 키워드가 없으므로 검색할 수 없음
  - KeyBert 자연어 처리 모델을 사용해 키워드 추출
  - 비슷한 의미의 키워드가 없도록 추출(Maximal Marginal Relevance 알고리즘 활용)
- 검색어가 문서의 단어와 완벽히 일치해야 검색이 되는 문제
  - 검색 시스템에서 한국어 토크나이저를 통해 형태소 분석 사용

Experiment

- 메타 데이터가 없는 글들을 저장해 두고 ‘록’(Rock) 질의어를 검색
- 검색 대상 글에는 음악 관련된 글들을 저장해두었음
- 평가 방법
  - 제안하는 시스템에서의 검색 결과와 기존 질의어 확장 방식의 검색 결과 비교

Experiment

- 기존의 질의어 확장 기반 검색 방식



키워드	내용	검색 경로
['재즈', '재즈록']	재즈록은 1970년대 시작된 대중적인 재즈 형태의 음악장르이다. 재즈의 즉흥연주와 베이스, 드럼의 선율에 전자악기...	록
['기쁨', '음악']	록은 1950년대 초 미국에서 생겨난 음악이다. 록은 일반적으로 보컬, 리드 전기 기타, 베이스 기타, 드럼의 넷으로 구성...	록

- 기존 질의어였던 '록'에 대해 검색

```
expand_results ['록', '락']
```

- ' 록'의 확장 결과인 '락'에 대해서는 해당 내용을 가진 글이 없어서 검색되지 않았고 있었다면 검색되었을 것임

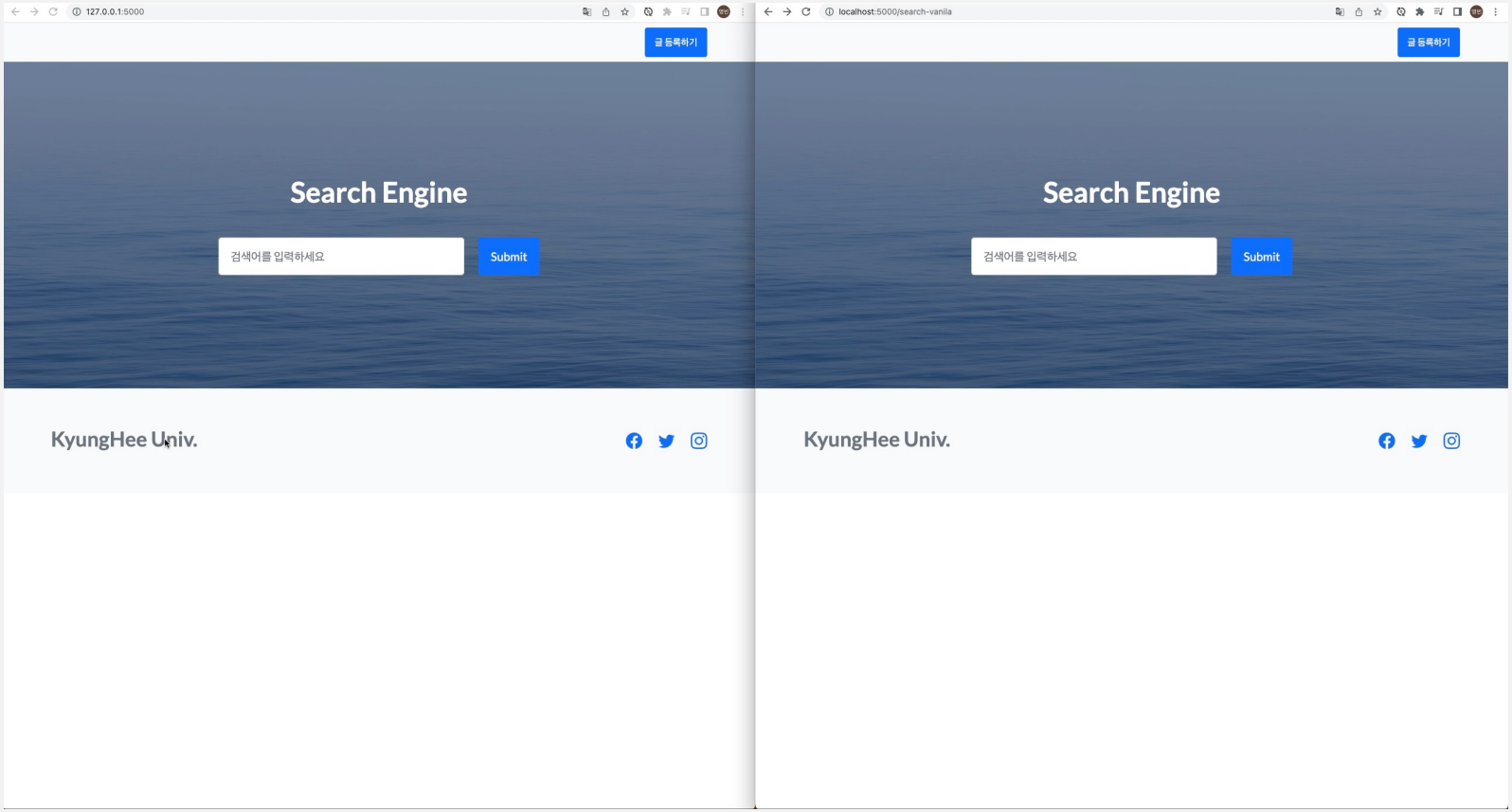
Experiment

- 키워드 확장 기반 검색 방식

키워드	내용	검색 경로
['기쁨', '음악']	록은 1950년대 초 미국에서 생겨난 음악이다. 록은 일반적으로 보컬, 리드 전기 기타, 베이스 기타, 드럼의 넷으로 구성...	록
['재즈', '재즈록']	재즈록은 1970년대 시작된 대중적인 재즈 형태의 음악장르이다. 재즈의 즉흥연주와 베이스, 드럼의 선율에 전자악기...	음악
['재즈', '대중음악']	재즈가 전 세계로 퍼지며 국가와 지역, 지역 음악 문화를 끌어들였고, 이로 다양한 음악 유형이 생겼다. 뉴올리언스 재즈...	음악
['기쁨', '음악']	록은 1950년대 초 미국에서 생겨난 음악이다. 록은 일반적으로 보컬, 리드 전기 기타, 베이스 기타, 드럼의 넷으로 구성...	음악
['음악', '동물']	음악이 역사상 언제부터 어떻게 발생되었는지는 확실하지 않다. 음악의 역사는 약 5만 년으로부터 1만 년쯤 전에 발생...	음악
['연주회', '음악사']	전 세계를 통하여 오늘날과 같이 음악이 보편화된 시기는 없다. 세계 각지에서 개최되고 있는 각종 음악 연주회는 말할 ...	음악
['재즈', '전통음악']	재즈는 19세기 말 ~ 20세기 초 미국 뉴올리언스 아프리카계 미국인 사회에서 유래된 음악 장르로 블루스와 래그타임에...	음악
['서양음악', '전통음악']	음악은 소리를 재료로 하는 시간예술이다. 그러나 그 보존 및 표기는 시각적인 매체인 악보를 사용한다. 인간의 고도의 ...	음악
['재즈', '대중음악']	재즈가 전 세계로 퍼지며 국가와 지역, 지역 음악 문화를 끌어들였고, 이로 다양한 음악 유형이 생겼다. 뉴올리언스 재즈...	재즈
['재즈', '전통음악']	재즈는 19세기 말 ~ 20세기 초 미국 뉴올리언스 아프리카계 미국인 사회에서 유래된 음악 장르로 블루스와 래그타임에...	재즈
['재즈', '재즈록']	재즈록은 1970년대 시작된 대중적인 재즈 형태의 음악장르이다. 재즈의 즉흥연주와 베이스, 드럼의 선율에 전자악기...	재즈
['기쁨', '음악']	록은 1950년대 초 미국에서 생겨난 음악이다. 록은 일반적으로 보컬, 리드 전기 기타, 베이스 기타, 드럼의 넷으로 구성...	기쁨

- '록'에 대한 최상위 글의 키워드 '기쁨', '음악'
- ' 음악'을 확장하여 '재즈'로 검색하고 '기쁨'은 감성 단어이므로 확장하지 않음
- 기존 방식보다 많은 검색 결과를 보여줌

Demo


127.0.0.1:5000
글 등록하기
Search Engine
검색어를 입력하세요
Submit
KyungHee Univ.
localhost:5000/search-vanila
글 등록하기
Search Engine
검색어를 입력하세요
Submit
KyungHee Univ.

Conclusion

- 사용자 질의 확장 방식과 제안하는 방식의 차이를 보였음
  - 기존 방식보다 더 다양한 검색 결과를 얻을 수 있음

Reference

[1] 정종진, 김경원, and 김구환. "데이터셋 검색 지원을 위한 메타데이터 자동 추출에 관한 연구." *한국통신학회 학술대회논문집* (2020): 867-868.

[2] 신동하, 김창복. "한글 워드임베딩과 아프리오리를 이용한 검색 시스템의 질의어 확장." 한국항행학회논문지, 20.6 (2016): 617-624.

Dong-ha Shin, Chang-bok Kim. "Query Extension of Retrieve System Using Hangul Word Embedding and Apriori." The Journal of Korea Navigation Institute, 20.6 (2016): 617-624.

[3] https://github.com/Kyubyong/wordvectors

[4] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

감사합니다