2단계의 추천질문 시스템

1.(타겟)질문 : 질문 (단순히 두문장에 대한 유사도 측면) <이건이미 코드로 구현해봄 tf-idf)> [단순히 이 질문을 했는데 질문자체에서 궁금한점이 생긴다 -> 던진질문에서 가장 유사한 질문을 추천]

구현내용

=🡺 konlpy 로 질문데이터셋 토크나이징

Tf-idf 방식(문장/문장 유사도를 구하는 가장 best방법) 으로 임베딩 벡터를 설정

유사도를 비교하는 두문장의 벡터거리(벡터 내적)를 score로 환산

가장 높은 score를 보이는 상위 n개의 문장을 select

2.타겟질문의 답변의 키워드중 이 타겟질문을 제시한사람이 답변을 보았을 때 답변에서 추가적인 궁금증 에 대한 질문 ->어 이 질문에 대해선 알겠어, 근데 이거를 이해했는데 이거에 답변에서 질문이랑 뭔가 관련이 있는거 같은데 이부분은 내가 지식이 없네?

->모를만한 키워드/토픽을 포함한 다른 질문을 추천해주거나, 그 키워드/토픽 을 설명하고 있는 답변의 질문을 추천

🡺질문을 한 사용자의 지식수준을 고려하기위해 질문의 문맥을 고려한 핵심키워드가 곧 사용자가 이 질문에서 묻고 싶은것이고 이게 사용자의 질문의도 이자 지식 수준이라고 판단하였다. 따라서 질문의 핵심키워드를 추출하고, 이 질문에 대한 답변(문장/문당)의 핵심 키워드와 비교해서 키워드를 1:1 비교 해서 겹치는 부분 ex) Jeus : Jeus client 이런것 들은 제거하고[답변의 키워드와 질문의 키워드가 겹치는 부분에 있어서는 답변내용에서 그 질문부분의 키워드에 대한 내용의 설명이 끝났다고 가정], 남은 답변의 키워드 중에서, 질문의 키워드와 최상의 유사도를 가지는 답변 키워드 n개를 select 한후, 질답DB에서 이 키워드를 포함하고 있는(1.retrieve) , 혹은 가장 유사한(2.tf-idf 나 다른 유사도 score) 답변과 매칭되는 질문을 추천해준다 혹은 3.gpt api를 사용하여, 프롬프트를 짜서 예)prompt= keyword\_list[0]+’context’+’이 답변 문맥을 참고해서 이 키워드를 설명해줄수있는 질문을 생성해줘’

코드구현관련 내용

🡺 타겟 질문과 그에 대한 답변에서 키워드를 추출해야 하기 때문에, pagerank 알고리즘을 구현한 textrank 라이브러리와, krwordrank 라이브러리를 활용해서, 한 문장의 키워드와 문단이 있을 때 핵심 문장을 뽑는 테스트를 진행을 해보았는데, 어느정도는 잘 나오는 부분이 있지만, 질문과 답변 문장의 맥락상 핵심이 되는 키워드는 뽑지 못하는 것 같아서, gpt api를 사용해서 프롬프트를 prompt=’context’+’answer’+” 다음 문장의 핵심 키워드들을 문맥을 참고하여 중요한 순서대로 dict자료형으로 뽑아줘” 이렇게 구성해서 테스트를 해보았는데, 문장을 설명할 수 있는 키워드를 잘 뽑아주는 것같아 일단 keyword를 뽑는 것은 이렇게 진행하였다.

이렇게 해서 질문과 답변의 키워드를 뽑은뒤 반복문과 in조건을 사용하여서, 겹치는 키워드를 제거하고, 질답 DB의 질문답변들을 토크나이징해 놓은 데이터를 word2vec 모델에 학습시켜놓은 모델을 가지고, 키워드:키워드 유사도 score를 구해서 질문 키워드와 비교했을때, score가 가장 높은 상위n개의 답변 키워드를 selelct한다.

1,2,3 에 대한 구현을 전부 진행해 볼지 아니면 여태까지 결과들로 보았을떄, 3->gpt api를 사용하는 것이 훨씬 깔끔하고 성능이 좋아보여서 3번으로 갈지 생각중이다.