kakaobank 인재영입

AI Search Engineer

과제전형 안내

안녕하세요, 지원자님.

귀한 시간 내어 카카오뱅크 과제전형에 참여해주셔서 감사합니다.

아래와 같이 AI Search Engineer 과제를 안내드립니다.

과제 관련 문의가 있으시다면 recruit@kakaobank.com 으로 문의해주세요!

---

1. 과제 내용

1) 평가 항목

1. 요구 사항을 분석하고 명확한 아키텍처를 의사 결정하고 설계하는 능력
2. 요구 사항을 고려하여 효율적이고 확장 가능하며 유지보수하기 쉬운 코드를 작성하는 능력
3. 개발한 과제의 성능 분석 및 최적화 능력
4. 컨테이너 (Docker) 기반 개발환경에 대한 이해와 활용 능력
5. 본 과제는 지원자의 문제 해결 과정과 의사결정을 평가하기 위한 목적으로, 모든 문제를 구현하지 않아도 되며, 각 단계에서의 접근 방법과 선택 이유를 명확히 설명하는 것이 중요합니다.

2) 문제 관련 공통 안내 사항

2-1. 문제 요약

텍스트 검색 서비스에 필요한 기능들을 제공하는 프로그램을 개발하고자 합니다. [문제#1]제공된 문서들을 순차적으로 수집하여 검색할 수 있는 형태로 가공하여 [문제#2]Opensearch에 적재하고, [문제#3]적재된 문서들을 바탕으로 검색할 수 있는 API 서버와, [문제#4]검색 결과 Report를 제공하는 프로그램을 Python으로 개발합니다.
검색 Input/Output은 report_app/resources/data/report.tsv 에 '질문'과 '필수 포함 text'를 참고하세요.

2-2. 적재 문서 설명

수집 날짜

- 적재된 문서를 수집 날짜에 맞춰 순차적으로 수집합니다.
- day_1 : 첫 번째 수집된 데이터 디렉토리
- day_2 : 첫 번째 수집 이후, 변경된 데이터 디렉토리
- day_3 : 두 번째 수집 이후, 변경된 데이터 디렉토리

html 문서

첨부파일 프로젝트의 **api_server/resources/data/html** 문서

- 출처 : https://ko.wikipedia.org/wiki(위키백과)

tsv 문서

첨부파일 프로젝트의 **api_server/resources/data/tsv** 문서

컬럼명	속성명	컬럼유형	설명
id	primary key	String	문서를 식별할 수 있는 unique 값
question	질문	String	고객 질문
answer	답변	String	고객 질문에 대한 답변
published	공개 여부	String	질문/답변 공개 여부
user_id	질문자	String	질문한 사람의 ID

2-3. 과제 제출 참고 사항

1. 첨부된 프로젝트 구조를 참고하여 과제물을 제출해주세요.
2. opensearch 2.17.1, python 3.11 버전을 사용하여 과제를 구성해주세요.
3. 과제 실행 환경은 외부와 격리된 환경에서 실행되어야 합니다. (외부 api 혹은 서비스 등의 사용이 불가함을 참고해주세요.)
4. 첨부된 프로젝트의 README.md에 과제의 실행 방법과 과제 구성, 풀이 방법과 의사 결정의 근거 등을 포함하여 주세요.
5. 아래의 사항들을 고려하여 과제를 개발해주세요
    a. 절차적 프로그래밍(PP)을 지양하고 객체지향 프로그래밍으로 작성해주세요. (OOP)
    b. 프로그램 코드에 대한 신뢰성을 높이기 위해 단위/통합 테스트 코드를 작성하고, 그 결과를 확인할 수 있어야 합니다. (Testing)
    c. docstring, typing, comment 를 최대한 활용해주세요. (Documentation)
    d. RuntimeError 등 프로그램에서 발생할 수 있는 예외 상황을 고려해 작성해주세요. (Exception Handling)

3) 과제

3-1. 문제#1

원천 데이터를 추출하여 검색에 적합한 형태로 가공하고, 그 데이터를 JSON 형태로 정해진 로컬 파일 시스템에 저장하는 API 서버를 개발합니다. API 호출 시 다음 요구 사항을 충족해야 합니다.

요구 사항

1. json 파일 경로 api_server/resources/data/json/(html/tsv)/day_(1,2,3)
2. 순차적 데이터 수집
    - API 호출 파라미터에 수집 날짜를 추가하여 순차적으로 수집을 진행합니다.
    - 첫 번째 데이터 수집은 day 1 디렉토리의 데이터 입니다.
    - 첫 번째 데이터 수집 이후, day 2, 3이 지나며 주어진 데이터 파일에 변경 사항(추가, 수정, 삭제)이 발생합니다.
        - 추가: 기존 데이터에 없던 새로운 데이터가 추가됨.
        - 수정: 기존 데이터가 업데이트됨.
        - 삭제: 기존 데이터가 삭제됨.
3. 원천 데이터 추출 및 JSON 형식으로 저장
    - 원천 데이터를 추출하여 검색하기 적합한 형태로 데이터를 가공한 후, JSON 형태로 지정된 파일 경로에 저장합니다.
4. 문서의 분할 저장
    - 길이가 긴 문서의 경우, 검색 속도 및 품질을 고려하여 다양한 청킹 방식을 통해 문서를 나누어 저장합니다.
5. 저장되는 데이터의 필수 정보
    - 공통 정보:
        - 원천 데이터 고유 아이디 (source_id)
        - 원천 데이터 경로 (source_path)
        - 원천 데이터 파일 타입 (file_type)
    - HTML 문서의 추가 정보:
        - 문서 제목 (title)
    - TSV 문서의 추가 정보:
        - 질문 (question)
        - 답변 (answer)
        - 공개 여부 (published)
    - 위 외에 검색 속도 및 품질을 고려한 기타 추가 정보는 자유롭게 구성합니다.

3-2. 문제#2

문제#1에서 가공한 데이터를 OpenSearch에 적재하는 API를 추가 개발합니다. API 호출시 다음 요구 사항을 충족해야 합니다.

요구사항

- 검색을 위해 필요한 opensearch index를 생성합니다.
- local filesystem에 저장된 데이터를 생성한 opensearch index에 적재합니다.
- API 호출 파라미터에 수집 날짜를 추가하여 순차적으로 데이터를 적재 및 수정/삭제 합니다.

3-3. 문제#3

OpenSearch index에 적재한 데이터를 텍스트로 검색하는 API를 추가 개발합니다. API 호출시 다음 요구사항을 충족해야 합니다.

요구사항

- 검색 속도와 품질을 고려하여 제공된 질문에 맞는 문서를 검색할 수 있는 API를 구현합니다.
- 검색 결과 문서의 개수(topK)는 3개로 고정합니다.

3-4. 문제#4

구현한 API서버를 활용하여 데이터 적재 API 호출부터 검색 API까지 각각의 API를 호출하여 최종 검색 결과 report를 생성하는 프로그램을 구현합니다.

요구사항

- report 경로 : report_app/resources/data/report.tsv
- report 내용
    - 검색 품질

        - 품질 측정 방식
            - report파일의 모든 “질문” 대상으로 품질을 측정합니다.
            - report파일에는 총 100개의 “질문”과 “필수 포함 text”가 제공됩니다.
            - 각 질문을 검색 API에 요청하여 검색된 순서대로 report파일에 “검색된 text”에 작성되도록 합니다.(문서 1,2,3은 score가 높은 순서대로 작성)
            - report파일에 문서 id, type 등은 제거하고 실제 문서의 내용만 작성되면 됩니다.
            - “검색된 text 1”에 필수 포함 text가 포함된 경우 true, 포함되지 않으면 false로 “정답 포함 여부” 셀에 작성되도록 합니다.
    - 검색 속도
        - 속도 측정 방식
            - 질문 하나 당 단일 검색 쿼리 속도를 각각 측정합니다.
            - 최종 결과(평균 속도) : 단일 검색 쿼리 속도의 합(초)/100(쿼리 개수)
    - 상세 결과 예시
        - 문서 가장 하단에 “최종 결과”에 작성
        - “정답 포함”셀에서 true인 결과 값의 합, “속도”셀의 평균 속도가 작성되도록 합니다.

	질문	필수 포함 text	검색된 text 1	…	정답 포함 여부	속도
1	카카오뱅크 채용 과정	서류전형, 과제전형	카카오뱅크 채용 과정: 서류 전형, 과제 전형, 면접 전형		true	0.1s
…						
100	카카오뱅크 문화	자기 주도 마일리지	안식 휴가에 대한 설명		false	0.3s
최종 결과					77	0.4s

- API호출 순서
    - 데이터 추출 API(day_1) -> 데이터 변환 API(day_1) -> 데이터 적재 API(day_1) -> 데이터 추출 API(day_2) -> 데이터 변환 API(day_2) -> 데이터 적재 API(day_2) -> 데이터 추출 API(day_3) -> 데이터 변환 API(day_3) -> 데이터 적재 API(day_3) -> 데이터 검색 API -> 최종 결과 저장

2. 유의사항

카카오뱅크 채용 과제 내용을 제3자에게 제공하거나, SNS 등 제3자가 볼 수 있는 서비스에 공개하는 행위는 엄격히 금지됩니다.

3. 과제 진행 기간

주말 포함 7일