Homework #1

Analytic Queries on Movie

DS² NoSQL 12th-term

문제 상황

당신은 *Netflix* 에서 근무하는 데이터 분석가이다. 다른 부서에서 요청받은 질문이 있는데, 이를 처리하기 위한 질의를 작성하라.

질문

Question #1: 아래 3가지 question을 수행하는데 오랜 시간이 소요될 것이다. 이들을 더욱 빠르게 처리하게 위한 index를 생성하시오. 어떻게 index를 생성하면 좋을지 **타당한 근거를 들어 제시**하고, 실제로 **성능이 얼마나 개선이 되었는지 설명**하시오.

Question #2: 특정 사용자 tag가 입력된 영화를 추출하고자 한다. 사용자 tag가 주어졌을 때, 해당 tag가 설정된 영화의 title을 전부 출력하시오. 결과물은 중복이 없고 사전식 순서(lexicographical order)상 오름차순으로 출력하시오. (힌트: Python의 `set()`을 활용)

Question #3: 특정 영화의 평점평균을 계산하고자 한다. 영화 title이 주어졌을 때, 이 영화의 평점평균을 계산하시오.

Question #4: 특정 고객의 평점에 대한 성향(bias)이 어떤가 알아보고자 한다. 유저 ID가 주어졌을 때, 아래 식을 이용하여 해당 사용자의 성향(bias)를 계산하시오.

식 1: 성향(bias) $=\frac{\sum_{i \in I} R_{i,u} - R_{i,avg}}{N}$, I는 유저가 rating한 영화의 집합, $R_{i,u}$은 사용자 u가 영화 i에 대해 평가한 점수, $R_{i,avg}$는 영화 i에 대한 전체 사용자의 평점 평균, N은 사용자가 평가한 영화의 수.

데이터 설명

본 과제는 *MovieLens* 데이터셋을 사용한다. *MovieLens*는 MovieLens에서 평점과 태그들을 모아 만든 데이터셋이다.

3개의 JSON 파일이 제공되며, 각 파일은 개별적인 collection에 저장된다. 과제 구현 시 본인에게 필요한 데이터만 사용하면 된다. 각각의 collection에 대한 설명은 아래와 같다.

- `ml_movies`
  - 각 영화에 대한 `movieId`와 `genres`가 포함되어 있다. 예를 들어 `"Interstellar"`의 경우, `movieId`가 `109487`이며, `[ "Sci-Fi", "IMAX" ]` 장르로 분류되어 있다.
  - 필드: `movieId, title, genres`
    - `movieId`: 영화에 대한 고유 식별자.
    - `title`: 영화의 제목.
    - `genres`: 영화의 장르. 배열 형태로 저장되어 있다. 예를 들어, `[]` 나 `['Action', 'Animation']` 형태가 될 수 있다.
- `ml_ratings`
  - 사용자들이 영화에 대해 평가한 평점 기록들이 포함되어 있다.
  - 필드: `userId, movieId, rating, timestamp`
    - `userId`: 사용자에 대한 고유 식별자.
    - `rating`: 사용자가 남긴 평점.
    - `timestamp`: 평점을 남긴 시간.
- `ml_tags`
  - 사용자들이 영화에 태그를 붙인 기록들이 포함되어 있다.
  - 필드: `userId, movieId, tag, timestamp`
    - `tag`: 태그.

제출

제출 방법:

- Python py 파일이 뼈대 소스코드로 제공됩니다. 각 질문에 해당하는 코드를 뼈대 코드에 작성하여 제출하시기 바랍니다.
- 소스코드에 반드시 주석을 포함하시길 바랍니다. 특히 1번 문항에 대해서 자세하게 적어주시기 바랍니다.
- 소스코드를 작성한 뒤에는 `tar.gz`나 `zip` 확장자로 압축한 뒤, 이메일로 TA에게 제출하시기 바랍니다. TA 메일 주소는 `ds2_nosql@dbs.snu.ac.kr` 입니다. 팀(조) 당 한번만 제출하시면 됩니다.

제출 기한:

- 2024년 4월 4일 목요일 23:59. 이보다 늦을 경우 0점 처리됩니다.

실행 환경:

- MongoDB 4.2.x
- 입력값은 python의 매개변수를 통해 주어집니다.
  - 아래는 Question 4의 예시입니다. 주어진 값(유저 ID)을 매개변수를 이용하여 `q4.py`에 전달하면, `q4.py`는 이를 이용하여 고객 성향을 계산하여 출력해야 합니다.

```
python q4.py 8619
```

표절:

- 각 팀 별로 현재 및 과거 제출물에 대한 표절 체크를 진행할 예정입니다. **절대로 남의 소스코드를 표절하지 마시길 바랍니다.** 표절이 확인될 경우, 0점 처리됩니다.

제공 파일:

- `result.txt`: 테스트용 코드와 이에 대한 정답
- `ml-25m-json.zip`: 데이터셋
- `import.bat`: 제공된 데이터셋을 한번에 import 할 수 있는 스크립트
- `q1.py, q2.py, q3.py, q4.py`: 뼈대 소스 코드