프로젝트 공지

406.426B 데이터관리와 분석

2021년 1학기

본 과목의 프로젝트는 총 3회로 이루어져 있다. 1차와 2차 프로젝트에서는 주어진 requirement들을 만족하는 데이터베이스를 설계 및 구현, 그리고 이를 기반으로 한 DB 마이닝과 추천 시스템 구현을 목적으로 한다. 3차 프로젝트에서는 텍스트 데이터를 대상으로 정보 검색, 문서 분류 및 군집화를 구현하는 것을 목적으로 한다.

프로젝트는 다음과 같다.

Project #1) Conceptual DB design & DB implementation

팀 구성일: 3월 30일, 발표일: 4월 13일

Project #2) DB mining & Recommendation system

팀 구성일: 4월 27일, 발표일: 5월 11일

Project #3) Document search engine & Classification and Clustering

팀 구성일: 5월 27일, 발표일: 6월 10일

본 프로젝트는 팀별로 진행되며, 각 팀은 매 프로젝트마다 자율적으로 3~5명으로 구성한다. 단, 팀원을 구하지 못하는 경우 충원을 희망하는 팀에 임의 배정하거나 팀을 구성하지 못한 인원으로 팀을 임의로 구성한다.

비대면 강의가 기본이 됨에 따라 비대면 기간 중 프로젝트 발표는 프로젝트 결과 제출 시에 5분 이내의 발표 동영상을 제작하여 함께 제출하고, 프로젝트 발표 일시에 조교가 zoom 화상 강의를 통해 재생한 후 Q&A 시간을 가지는 것으로 한다. 결과물 제출은 발표일 전날 23시 59분까지 보고서와 발표 자료 및 발표 동영상을 ETL에 업로드해야 한다.

Project #1: Conceptual DB design & DB implementation

사이트 A는 온라인 소프트웨어 유통망으로, 개발자와 사용자가 해당 사이트를 통해 교류할 수 있는 서비스를 제공한다. 개발자들은 다양한 온라인 소프트웨어를 개발하여 사이트 A에 출시하고, 사용자들은 사이트 A를 통해 소프트웨어를 구매하고 이용한다. 본 프로젝트에서는 이러한 온라인 소프트웨어를 아이템으로 부르기로 한다. 본 프로젝트는 사이트 A가 사용하는 DB의 ER diagram 도식화와 DB 구현을 목적으로 하며, 크게 두 부분으로 나뉜다.

PART Ⅰ. ER diagram 도식화

PART Ⅱ. DB 구현 및 데이터 입력

PART Ⅰ. ER diagram 도식화

PART Ⅰ는 사이트 A가 사용하는 DB에 대한 ER diagram을 도식화하는 것을 목표로 한다. 사이트 A의 DB는 아래의 requirement들을 만족해야 한다.

(R1-1) 사이트 A는 사이트에 가입된 사용자에 대한 정보를 저장하고 있다. 사용자는 아이템을 이용하고 아이템에 리뷰를 남길 수 있다. 사용자에 대한 정보로는 사이트에서 부여한 고유번호, 사용자가 설정한 사용자 닉네임, 프로필 이미지 존재 여부, 사용자가 이용한 이력이 있는 아이템의 총 개수로 구성되어 있다. 이 때 사용자 닉네임은 모두 달라야 한다. 또한 사용자가 작성한 리뷰의 수, 작성한 리뷰들이 받은 총 추천 수, 가입 이래로 사용자가 이용한 아이템들의 이용 시간의 총합이 저장되어야 한다.

(R1-2) 사이트 A는 사용자들의 리뷰 정보를 가지고 있다. 사용자는 아이템에 리뷰를 작성할 때 해당 아이템을 추천하는지 여부를 기입해야 하며, 작성한 리뷰는 타 사용자들에 의해 추천 또는 비추천으로 평가받을 수 있다. 리뷰 정보에는 사이트에서 부여한 고유번호, 작성자 id, 아이템 id, 사용자의 아이템 추천 여부, 본문의 길이, 게시 일자, 리뷰가 받은 추천 수, 리뷰가 받은 평가 중 추천의 비율, 리뷰가 받은 총 평가 수가 저장되어야 한다.

(R1-3) 사이트 A는 사용자가 아이템을 이용할 때마다 이에 대한 정보를 저장한다. 따라서 사용자가 아이템을 이용한 이력 정보에는 사용자 id, 아이템 id, 사용자가 최근 2주 간 아이템을 이용한 시간, 사용자가 가입 이래로 아이템을 이용한 총 시간이 저장되어야 한다.

(R1-4) 사이트 A는 아이템에 대한 정보를 가지고 있다. 아이템에 대한 정보로는 사이트에서 부여한 고유번호, 아이템의 이름, 가격 정보, 정식 출시 전 베타 버전의 유무, 사용자들에게 받은 평가 점수, 외부에서 받은 평가 점수, 개발사 정보, 출시 일자, 스펙이 저장되어야 한다. 이 때 가격 정보, 개발사 정보, 출시 일자, 스펙은 빈 칸일 수 있다. 스펙의 경우 하나의 아이템이 여러 스펙을 가질 수 있다. 또한 아이템은 여러 장르에 속하거나 여러 가지 태그가 붙어 있을 수 있다. 따라서 아이템 정보에는

아이템이 받은 리뷰의 수, 아이템을 사용한 이력이 있는 사용자의 수, 아이템이 속해 있는 장르의 수, 아이템에 붙어 있는 태그의 수가 추가로 저장되어야 한다.

(R1-5) 사이트 A는 장르에 대한 정보를 가지고 있다. 여기에는 사이트에서 부여한 고유번호, 장르 이름, 장르에 속하는 아이템 id, 장르에 속하는 아이템의 수가 저장되어야 한다. 하나의 장르는 최소 하나 이상의 아이템을 포함해야 하며, 장르의 이름은 모두 다르다.

(R1-6) 사이트 A는 특성이 비슷한 여러 아이템들을 하나의 번들로 묶어서 할인 판매한다. 번들에 대한 정보로는 사이트에서 부여한 고유번호, 번들의 이름, 가격 정보, 할인율, 할인된 번들의 최종 가격, 번들에 포함되는 아이템 id, 번들에 포함되는 아이템의 수가 저장되어야 한다. 번들은 최소 하나 이상의 아이템을 포함해야 하고, 번들에 포함되는 아이템들이 어느 장르에 몇 개나 속해 있는지에 대한 정보도 저장되어야 한다.

(R1-7) 사이트 A는 태그에 대한 정보를 저장한다. 태그 정보는 태그가 붙어 있는 아이템 id, 태그의 이름과 태그가 아이템에 붙어 있는 순서를 포함한다. 하나의 아이템에 붙어 있는 태그는 모두 다른 이름을 가진다.

PART Ⅱ. DB 구현 및 데이터 입력

PART Ⅱ는 이 사이트 A의 데이터에 적합한 데이터베이스 스키마를 설계하여 데이터베이스 테이블을 실제로 생성한 후 데이터 입력까지를 목표로 한다. 해당 프로그램은 Python과 MySQL을 사용하여 구현하여야 하며, 다음의 요구 조건들을 만족하여야 한다. Python에서 mysql-connector-python 외의 **별도 라이브러리는 사용할 수 없다**.

(R2-1) 사이트 A의 데이터를 활용하기에 앞서 이를 MySQL 상에 저장해야 한다. 이를 위해 먼저 DMA_team##의 이름을 가지는 schema를 생성해야 한다. 예를 들면, 1조의 schema 명은 DMA_team01이다. 이 때 schema가 존재할 경우 생성 과정을 다시 수행하지 않아야 한다.

(R2-2) schema를 설계한 이후에는 데이터를 저장하기 위한 table을 생성해야 한다. 생성하는 table과 column 이름과 순서는 주어진 데이터셋의 table 및 column과 일치해야 한다. 0 또는 1의 값을 가지는 column은 TINYINT(1)로, INTEGER type는 'INT(11)'로, 범위가 큰 INTEGER type은 'BIGINT(20)'로, STRING type는 'VARCHAR(255)'를 이용하여 생성한다. 그 외 날짜는 'DATE'를 통해 생성한다. 이 때 table이 존재할 경우 생성 과정을 다시 수행하지 않아야 한다. (R2-2)에서는 foreign key 조건을 작성하지 않고 (R2-3)에서 데이터 입력 후 foreign key 조건을 추가한다.

(R2-3) 생성된 table에 데이터를 저장해야 한다. 데이터는 csv파일로 주어지며 이를 직접 변형해서는 안된다.

(R2-4) 해당 데이터베이스 schema에 foreign key 조건들을 반영해주어야 한다.

채점 기준(절대평가)

- PART Ⅰ ER diagram의 requirement 만족 여부(50 %)
- PART Ⅱ 설계한 데이터베이스 스키마와 constraints(25%), requirement 만족 여부(15%)
- 보고서 품질(5%), 발표(5%)

결과물들을 'DMA_project1_team##.zip' 파일로 압축하여 발표일 전날인 **4월 12일 23:59까지** ETL에 업로드해야 한다. ETL 상에 문제가 생겼을 경우 osa8361@snu.ac.kr 로 오류 증명 파일과 함께 제출 기한 전에 보내야 한다. 제출해야 할 결과물과 파일명, 파일 확장자는 다음과 같다.

- 보고서

  - 파일명: DMA_project1_team##_보고서.pdf
  - 보고서에는 PART Ⅰ에서의 문제 정의, 도식화한 ER diagram의 도식화 과정과 최종 ER diagram, PART Ⅱ에서의 문제 정의, 설계한 스키마와 코드에 대한 설명이 포함되어야 한다. 이 때 스키마 설계 시 constraints들과 이들의 설정 근거가 포함되어야 한다.

- 발표 자료 및 발표 동영상

  - 발표 자료 파일명: DMA_project1_team##_발표자료.pdf
  - 발표 동영상 파일명: DMA_project1_team##_발표동영상.mp4
  - 발표 동영상은 팀 당 5분 이내로 제작되어야 하며 powerpoint의 녹화 기능을 사용한다.

- Python 프로그램 코드

  - Python 코드 파일명: DMA_project1_team##.py
  - 함수들의 입력 값들의 의미는 다음과 같다.

    host, user, password: MySQL에 접근하기 위한 계정 정보

    directory: 데이터가 저장된 주소 (ex. C:/dir/user.txt → 'C:/dir/')

  - 뼈대 코드의 주석에 작성된 TODO들에 따라 팀의 번호, MySQL 계정 정보, 데이터(csv)들이 저장된 주소 등을 바꿔야 한다.
  - mysql.connector 외의 다른 패키지를 import하여 사용하는 것은 허용되지 않는다.
  - PART Ⅱ의 각 requirement(R2-1~R2-4)에 해당하는 Python 코드는 주어진 뼈대 코드의 requirement# 함수로 구현되어야 한다. 예를 들어 R2-1은 requirement1 함수에 구현되어야 한다.