AIX30011 Spring 2025

Handong Honor Code

- It is your responsibility to understand and adhere to the Handong Honor Code.
- Unauthorized copying of coding assignments, homework, or creative work will result in a failing grade (F) for both the individual who copies and the individual who shares the work.

Jaeyoung Chun | School of Applied Artificial Intelligence | Handong Global University

Midterm Exam

King James Version (KJV) 영어 성경을 사용하여 간단한 ETL(Extract, Transform, Load) 작업을 수행해보려고 합니다. KJV 성경구절들을 텍스트 파일에서 추출(Extract)하고 변형(Transform)한 후, MySQL 데이터베이스에 적재(Load)하여 궁극적으로는 성경을 검색할 수 있는 시스템을 만들려고 합니다.

1. Create Database and Table

1. SQL 명령어를 사용하여 `Bible`이라는 database를 만듭니다.
2. SQL 명령어를 사용하여 `Bible`이라는 table를 만듭니다.
    - Column 1:
        - Column Name: `id`
        - Data Type: 정수형 숫자(Integer)
        - PRIMARY KEY 설정
        - 정수형 숫자가 1부터 차례대로 자동 증가하도록 설정
    - Column 2:
        - Column Name: `book`
        - Data Type: 최대 30 글자를 담을 수 있도록 설정
    - Column 3:
        - Column Name: `chapter`
        - Data Type: 정수형 숫자(Integer)
    - Column 4:
        - Column Name: `verse`
        - Data Type: 정수형 숫자(Integer)
    - Column 5:
        - Column Name: `text`
        - Data Type: 무제한 글자를 저장할 수 있는 `TEXT`

3. 위 `Bible`이라는 table를 만들때, UNIQUE CONSTRAINT를 설정합니다.

- 즉, `book`, `chapter`, `verse` 이 세 개의 column **조합**이 고유(unique) 하도록 설정.
- 동일한 `book`, `chapter`, `verse` 조합이 한 번 이상 INSERT 될 수 없어야 함.
- 예를 들어, 창세기 1장 1절이 `Bible` table에 한 번 이상 INSERT 될 수 없어야 함. 한 번 이상 INSERT 하려고 시도하면 MySQL이 오류를 발생시켜야 함.

Schema directory must be removed manualy.

```
rm -rf
/Users/chunj/projects/bmp/workspace/mysql/data/Bible
```

2. Extract and Transform

주어진 `kjvdat.txt` 파일은 King James Version 영어 성경 전체를 담은 text 파일 입니다 (총 31,102 구절). 각 성경구절을 SQL INSERT 명령어로 변경하는 Python 코드를 작성해서 실행하고 코드를 제출합니다.

예를 들어, `kjvdat.txt` 파일의 첫 줄은 다음과 같은데,

```
Gen|1|1| In the beginning God created the heaven and the
earth.~
```

이 구절은 다음과 같이 SQL INSERT 명령어로 변경되어야 합니다. 특수문자들이 다 없어진다는 것 유의하셔야합니다. `Gen`이 `Genesis`로 변경된다는 것, 유의하셔야합니다. 마찬가지로 `Mat`은 `Matthew`로 변경되어야겠습니다. 과제와 함께 주어진 `bible-fullname.txt`, `bible-3letter.txt` 파일을 사용하셔야 합니다.

```
INSERT INTO Bible (book, chapter, verse, text)
VALUES ("Genesis", 1, 1, "In the beginning God created the
heaven and the earth.");
```

총 31,102 구절을 모두 SQL INSERT 명령어로 변경한 후, 이를 하나의 파일에 다 모아서 `import.sql`이라는 파일 이름으로 저장하고 제출합니다. 저장된 파일은 다음과 같은 형태가 되어야 합니다:

```
USE Bible;

INSERT INTO Bible ...(생략)...
...(생략)...
...(생략)...
...(생략)...
```

데이터 분석 시 정확한 결과를 도출하기 위해 필요한 경우 데이터 정제 작업을 수행해 주시기 바랍니

다. 단, 성경 말씀은 원본 그대로 보존되어야 합니다.

> 나는 이 책의 예언의 말씀들을 듣는 모든 사람에게 증거합니다. 누구든지 이 말씀들에 어떤 것을 더하면 하나님께서 이 책에 기록된 재앙들을 그에게 더하실 것입니다.
>
> 그리고 누구든지 이 예언의 책의 말씀들로부터 어떤 것이라도 없애 버리면 하나님께서는 이 책에 기록된 생명나무와 거룩한 도성에서 그의 몫을 없애 버리실 것입니다.
>
> 요한계시록 22:18-19 (우리말 성경)

앞서 언급한 추출/변형 및 기타 전처리 작업을 하는 코드를 작성해주세요:

3. Load

이 섹션에서는, 앞에서 작성한 `import.sql` 파일을 실행시켜서 MySQL Database에 성경 전문을 저장하는 작업을 수행하게 됩니다.

섹션 `3.1` 과 `3.2` 는 반드시 Terminal에서 작업하셔야 합니다. Big Data를 다루다 보면, GUI 환경(Graphic User Interface)에서 할 수 없는 작업들이 많습니다. 데이터가 너무 커서 GUI 환경에서 파일을 읽는 것 자체가 불가능한 경우가 대부분입니다. 그 한 예로, 1주차 때, Microsoft Excel의 한계를 보여드렸습니다. DataGrip과 그 밖에 많은 GUI 툴들에 비슷한 한계가 있습니다. 터미널에서 작업 수행하는 것에 익숙해지셔야합니다.

3.1. Step 1

MySQL이 실행되고 있는 container에서 `import.sql` 파일을 액세스 할 수 있도록 하기 위해서는, 우선 `import.sql` 파일을 MySQL Docker container를 실행시킨 폴더 아래에 존재하는 `scratch` 폴더에 복사해야합니다. 참고로 이 `scratch` 폴더는 이전에 `airportdb` 파일을 다운로드 받아서 저장했던 폴더와 동일합니다 (`~/projects/bmp/workspace/mysql/scratch`).

복사를 완료한 후, 아래 명령어를 실행해서 파일이 있어야 할 위치에 존재하는지 확인할 수 있습니다:

```
docker exec -it aix_mysql ls /scratch/import.sql
```

위 명령어의 출력물은 다음과 같아야 합니다:

```
/scratch/import.sql
```

3.2. Step 2

MySQL container로 들어갑니다:

```
docker exec -it aix_mysql bash
```

`mysql` 콘솔 프로그램을 사용해서 `import.sql`을 실행시키고, MySQL에 성경전문을 저장합니다:

```
mysql -p Bible < scratch/import.sql
```

만약 아래와 같은 오류가 발생했다면 Step 1을 제대로 수행하지 않았을 가능성이 큽니다.

```
bash: scratch/import.sql: No such file or directory
```

3.3. Step 3

DataGrip을 실행시킨 후, 요한복음 3장 16절을 검색/출력하는 SQL 명령어를 작성하세요 (모든 컬럼 출력). SQL 명령어를 실행한 후 얻은 결과물, 즉 DataGrip에 출력된 row(s)를 .csv 파일로 저장해서 제출하세요.

4. Question

신약 성경에서 `Joseph`이라는 단어가 들어가는 모든 구절을 출력하는 SQL 명령어를 작성하세요. 모든 컬럼을 출력하고, 출력 순서는 성경책 순서대로 출력되어야합니다 (마태, 마가, ..., 요한계시록). SQL 명령어를 실행한 후 얻은 결과물, 즉 DataGrip에 출력된 row(s)를 .csv 파일로 저장해서 제출하세요.

5. Question

앞 문제에서 `Joseph` (요셉)이 여러번 KJV 신약성경에 언급되는데, 동명이인이 있습니다. 총 몇 명의 요셉이 등장했습니까?

예를 들어, 마태복음 1:16과 마태복음 1:18에 등장하는 요셉이 같은 인물이면 한 번만 카운트하셔야 합니다 (두 번 카운트 하는 것이 아니라). 예를들어, 이렇게 계산해서, 요셉이라는 이름을 가진 서로 다른 사람이 10명 등장했으면 `10`이라고 적으면 됩니다.

답을 적은 후, 그 아래에 그렇게 생각하게 된 이유를 적어주세요.

(*) 참고로, 이 문제는 학기 후반부에 출제될 모델링 과제와 연결됩니다.