COMS W4111: Introduction to Databases Section 002/V02, Spring, 2022

HW 1 Notebook

Introduction

This notebook has three top level sections:

1. *Setup* tests the environment setup, and should work assuming you completed HW0.
2. *Common Tasks* are the HW1 tasks for both the programming and non-programming track. All students complete this section.
3. *Non-Programing Track* contains the tasks that students in the non-programming track must complete.
4. *Programming Track* contains the tasks that students in the programming track must complete.

Submission format:

- All students (both tracks) submit a completed version of this notebook. Students need to complete the setup section, the common section, and the section specific to their track. The submission format is a PDF generated from the notebook. Students can generate the PDF by:
  - Choosing `File->Print Preview` in the notebook's menu bar. This will open a new browser tab.
  - In the new browser tab, select `File->Print` and choose to save as PDF.
  - **Make sure that everything renders properly in the generated PDF.** Troubleshoot/reach out if you have issues. Images/outputs that render incorrectly will not be graded.
- All students submit a zip file containing their cloned HW0/1 project, which they got by cloning the GitHub repository. Students can:
  - Open a command/terminal window in the root directory where they cloned the project.
  - Enter `git pull` to retrieve any updates to the project, including required data files.
- Students can edit the notebook using Anaconda Navigator to open Jupyter Notebook.
- Students on the programming track also create and modify Python files in the sub-folder `<UNI>_web_src` . Remember, you should be using a folder with your UNI. In my case, the folder would be `dff9_web_src.`
- The zip file you submit should contain **only** the following sub-folders/files:

- `<UNI>_src.` (All students) This folder must container your version of this notebook.
- `<UNI>_web_src.` (Only programming track)
- To be clear: the zipped directory for non-programming track submissions should contain **one** file. The corresponding `zip` for the programming track should contain **two** files.

- Make sure to submit your notebook in the PDF format separately from the zip file, based on your track as well. That is, you need to make **two** submissions in total like below:
  - Submit your notebook file in PDF format to `Homework 1: Non-programming or Programming` **(Make sure that you assigned pages properly).**
  - Submit your zip file to `Homework 1: Zip File Submission`

Setup

Note: You will have to put the correct user ID and password in the connection strings below, e.g. replace `dbuser` and `dbuserdbuser`.

iPython-SQL

In [1]:
```
%load_ext sql
```

In [2]:
```
%sql mysql+pymysql://root:dvuserdvuser@localhost
```

Out[2]:
```
'Connected: root@None'
```

In [3]:
```
%sql select * from db_book.student where name like "z%" or name like "sh%"
```

```
 * mysql+pymysql://root:***@localhost
2 rows affected.
```

Out[3]:

ID	name	dept_name	tot_cred
00128	Zhang	Comp. Sci.	102
12345	Shankar	Comp. Sci.	32

PyMySQL

In [4]:
```
import pymysql
```

In [5]:
```
conn = pymysql.connect(host="localhost", user="root", password="dvuserdvuser")
```

In [6]:
```
sql = """
    select * from db_book.student where
        name like %s or name like %s
"""
```

In [7]:
```
pattern_1 = "z%"
pattern_2 = "sh%"
```

In [8]:
```
cur = conn.cursor()
res = cur.execute(
    sql, args=(pattern_1, pattern_2)
)
res = cur.fetchall()
```

In [9]:
```
res
```

Out[9]:
```
(('00128', 'Zhang', 'Comp. Sci.', Decimal('102')),
 ('12345', 'Shankar', 'Comp. Sci.', Decimal('32')))
```

Pandas

In [10]:
```
import pandas as pd
```

In [11]:
```
#
#  Replace the path below with the path of your project directory.
#  Use // instead of / if you're on Windows.
#
project_root = "/Users/litinghuang/Desktop/Database/S22-W4111-HW-1-0"
```

In [12]:
```
people_df = pd.read_csv(project_root + "/data/People.csv")
```

In [13]:
```
people_df.loc[
    (people_df['nameLast'] == "Williams") & (people_df['birthCity'] == 'San Dieg
    ["playerID", "nameLast", "nameFirst", "birthYear", 'birthCity', 'bats', 'thr
]
```

Out[13]:

	playerID	nameLast	nameFirst	birthYear	birthCity	bats	throws
19773	willite01	Williams	Ted	1918.0	San Diego	L	R
19776	willitr01	Williams	Trevor	1992.0	San Diego	R	R

SQLAlchemy

In [14]:
```
from sqlalchemy import create_engine
```

In [15]:
```
engine = create_engine("mysql+pymysql://root:dvuserdvuser@localhost")
```

In [16]:
```
sql = """
```

```
    select * from db_book.student where
        name like %s or name like %s
"""
pattern_1 = "z%"
pattern_2 = "sh%"
```

In [17]:

```
another_df = pd.read_sql(sql, params=(pattern_1, pattern_2), con=engine)
another_df
```

Out[17]:

	ID	name	dept_name	tot_cred
0	00128	Zhang	Comp. Sci.	102.0
1	12345	Shankar	Comp. Sci.	32.0

Common Tasks

Schema and Data Modeling

- There are three entity types:
  1. Employee with attributes:
     - employee_no
     - last_name
     - first_name
  2. Department with attributes
     - department_id
     - department_name
  3. Applicant with attributes:
     - email
     - last_name
     - first_name

Relational Schema

- Using the notation from the textbook slides and lecture notes, define the relation definitions for each of the entity types. That is, the schema definition for the relations. You will need to choose a primary key.

- The snippet below shows how to use under-bar.

$$This\ is\ a\ sentence\ with\ someting\_in\_parentheses(\underline{something}, another\_thing)\ and$$

You can double click on the cell above to see the source, which is

```
\begin{equation}
This\ is\ a\ sentence\ with\ someting\_in\_parentheses(
    \underline{something}, another\_thing)\  and\ something\ with\
```

```
underbar.
\end{equation}
```

Put your relation definitions below between the horizontal lines.

<hr style="height: 1px";>

$$Employee(\underline{employee\_no}, last\_name, first\_name) \tag{2}$$

$$Department(\underline{department\_id}, department\_name) \tag{3}$$

$$Applicant(\underline{email}, last\_name, first\_name) \tag{4}$$

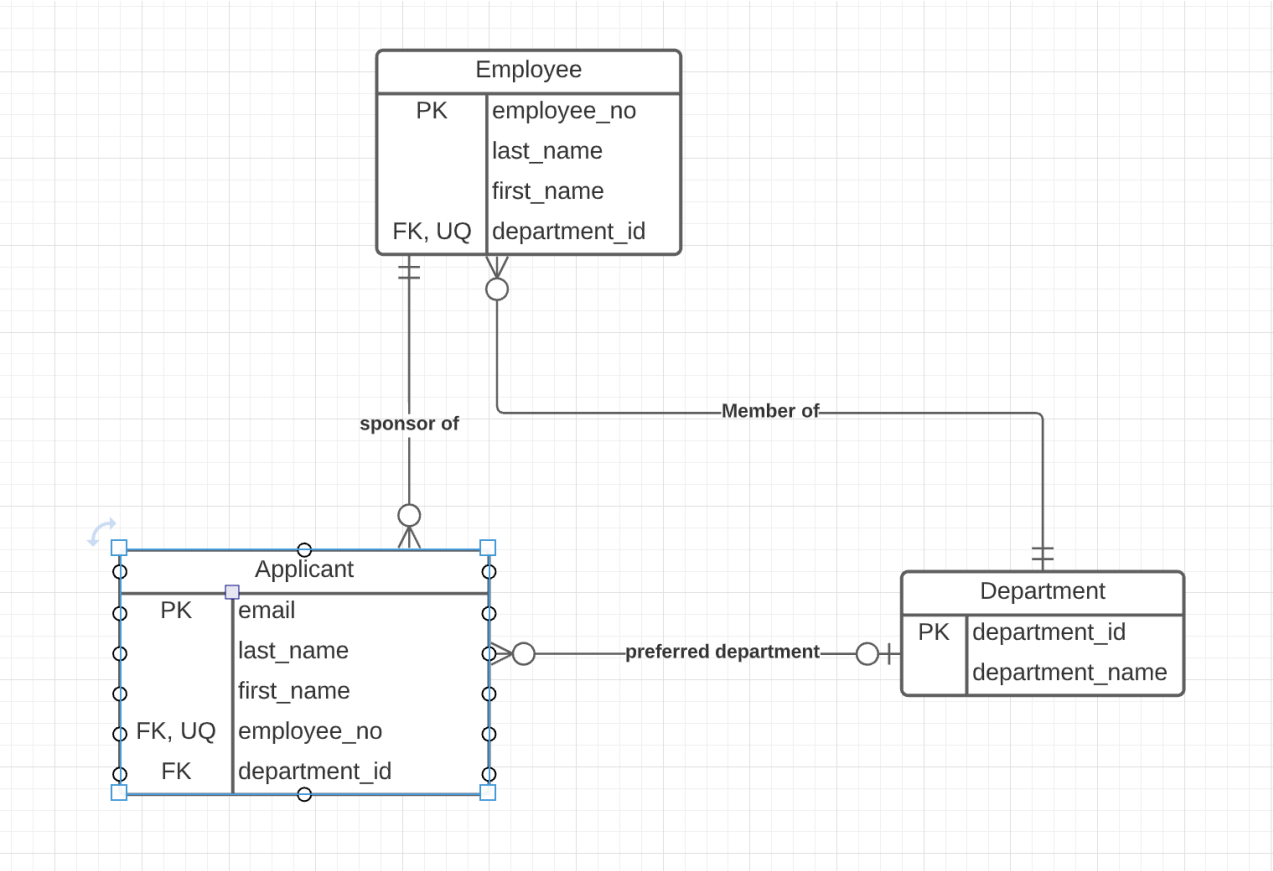
<hr style="height: 1px";>

ER Modeling

- Continuing the example above:
  - An *employee* is a _member_of_ exactly one department.
  - An *applicant* has exactly one *employee* who is _sponsor_of_ of the applicant.
  - An *applicant* may have specified a *department* that is the *applicant's* _preferred_dept._
- Use Lucidchart to draw the logical diagram.
- **Note:** You may have to add columns/attributes to some tables to implement the relationships.
- To submit the diagram, take a screen capture and modify the cell below to load your diagram from the file system. The following is an example for how to include the screenshot.

In [18]:
```
er_model_file_name = 'ER.png'

print("\n")
from IPython.display import Image
Image(filename=er_model_file_name)
```

Out[18]:



Relational Algebra

Instructions

- You will use the RelaX online relational algebra calculator.
- You must use the dataset `Silberschatz - UniversityDB.` I demonstrated how to select a dataset during a lecture.
- For submitting your answer, you must:
  - Cut and paste your relational expression in text.
  - Take a screenshot and include the image.
- The following is an example question and answer.

Example

Question: Produce a table of the form `(course_id, title, prereq_id, preqreq_title)` that lists courses and their prereqs.

```
π course_id, title, prereq_id, prereq_title
    (
        (π course_id, title, prereq_id (course ⋈ prereq))
        ⋈ prereq_id=x
```

```
        (π x←course_id, prereq_title←title (course))
    )
```

In [ ]:

Relational Algebra Q1

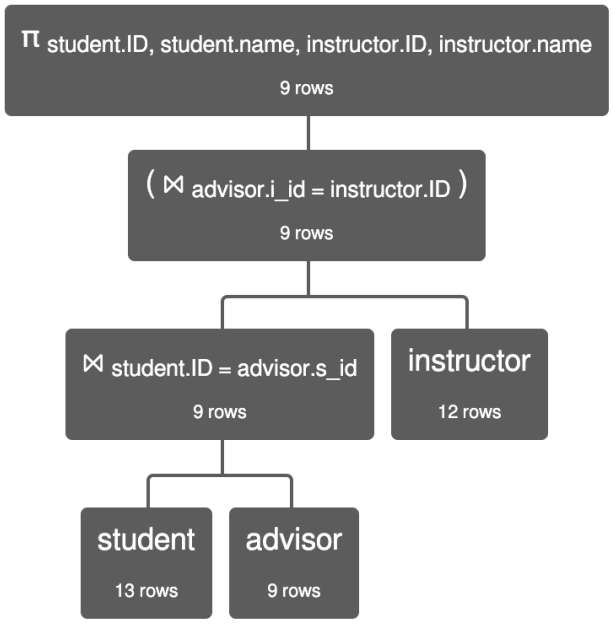
- Use `student, advisor` and `instructor` for this question.
- Produce a table of the form `(student.ID, student.name, instructor.ID, instructor.name)` that shows students and their advisors.

```
π student.ID, student.name, instructor.ID, instructor.name
(
 student
 ⋈ student.ID = advisor.s_id    advisor
 ⋈ advisor.i_id = instructor.ID instructor
)
```

In [19]:

```
er_model_file_name = 'RA1.png'

print("\n")
from IPython.display import Image
Image(filename=er_model_file_name)
```

Out[19]:



$$\pi \text{ student.ID, student.name, instructor.ID, instructor.name } ((\text{ student } \bowtie \text{ student.ID = advisor.s\_id advisor }) \bowtie \text{ advisor.i\_id = instructor.ID instructor })$$

student.ID	student.name	instructor.ID	instructor.name
128	'Zhang'	45565	'Katz'
12345	'Shankar'	10101	'Srinivasan'
23121	'Chavez'	76543	'Singh'

Relational Algebra Q2

- Use `student` and `takes` for this question.
- Produce a table of the form `(student.ID, student.name, student,tot_cred, student_dept_name)` for students that have not taken any course/section.

`student ▷ takes`

In [20]:

```
er_model_file_name = 'RA2.png'

print("\n")
```

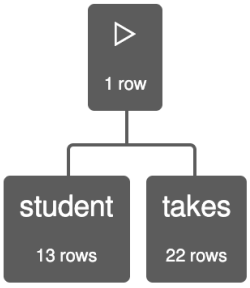
```
from IPython.display import Image
Image(filename=er_model_file_name)
```

Out[20]:

```
1 student ▷ takes
2
```

execute selection

download history



student ▷ takes

student.ID	student.name	student.dept_name	student.tot_cred
70557	'Snow'	'Physics'	0

SQL

Instructions

- The questions in this section ask you to write and execute SQL statements.
- Your answer should be a code cell with `%sql` and your query.
- You must execute the query.

Example

- This is the SQL version of the query from the relational algebra section above.

In [21]:
```
%%sql
use db_book;

select a.course_id as course_id,
       a.title as title,
       prereq_id,
       b.title as prereq_tiles
from
             (select course_id, title, prereq_id from course join prereq using(
join
    course as b on a.prereq_id=b.course_id
```

```
 * mysql+pymysql://root:***@localhost
0 rows affected.
7 rows affected.
```

Out[21]:

course_id	title	prereq_id	prereq_tiles
BIO-301	Genetics	BIO-101	Intro. to Biology
BIO-399	Computational Biology	BIO-101	Intro. to Biology
CS-190	Game Design	CS-101	Intro. to Computer Science
CS-315	Robotics	CS-101	Intro. to Computer Science
CS-319	Image Processing	CS-101	Intro. to Computer Science
CS-347	Database System Concepts	CS-101	Intro. to Computer Science
EE-181	Intro. to Digital Systems	PHY-101	Physical Principles

SQL Question 1

- Translate your answer from Relational Algebra Q1 into SQL.
- Do not worry about correctly naming the columns.

In [22]:
```
%%sql

select student.ID, student.name, instructor.ID, instructor.name
from
student, advisor, instructor
where
student.ID = advisor.s_id and instructor.ID = advisor.i_id
```

```
 * mysql+pymysql://root:***@localhost
9 rows affected.
```

Out[22]:

ID	name	ID_1	name_1
12345	Shankar	10101	Srinivasan
44553	Peltier	22222	Einstein
45678	Levy	22222	Einstein

ID	name	ID_1	name_1
00128	Zhang	45565	Katz
76543	Brown	45565	Katz
23121	Chavez	76543	Singh
98988	Tanaka	76766	Crick
76653	Aoi	98345	Kim
98765	Bourikas	98345	Kim

SQL Question 2

- You guessed it.
- Translate your answer from Relational Algebra Q2 into SQL.
- Do not worry about correctly naming the columns.

In [23]:
```
%%sql
select student.ID, student.name, student.tot_cred, student.dept_name
from student where
not student.ID in (select ID from takes)
```

```
 * mysql+pymysql://root:***@localhost
1 rows affected.
```

Out[23]:

ID	name	tot_cred	dept_name
70557	Snow	0	Physics

SQL Question 3

- The following query makes a copy of the `department` table.

In [24]:
```
%%sql

drop table if exists hw1_department;
create table hw1_department as select * from department
```

```
 * mysql+pymysql://root:***@localhost
0 rows affected.
7 rows affected.
```

Out[24]:
```
[]
```

- The next query shows the content.

In [25]:
```
%sql select * from db_book.hw1_department
```

```
 * mysql+pymysql://root:***@localhost
7 rows affected.
```

Out[25]:

dept_name	building	budget
Biology	Watson	90000.00
Comp. Sci.	Taylor	100000.00
Elec. Eng.	Taylor	85000.00
Finance	Painter	120000.00
History	Painter	50000.00
Music	Packard	80000.00
Physics	Watson	70000.00

- You have two tasks for this question.
  1. Create a new table `db_book.hw1_schools` that has columns `school_id` and `school_name.`
  2. Modify table `db_boot.hw1_department` to contain a columns `school_id.`
- **Notes:**
  - You do not have to worry about foreign keys.
  - You do not need to populate any data or link `school_id` to the `hw1_schools.`
  - You can use DataGrip or another tool to produce the SQL DDL, but you must show successful execution on the code cells below.

In [26]:

```
%%sql

drop table if exists hw1_schools;
create table hw1_schools
(
    school_id   varchar(4)  null,
    school_name varchar(64) null
);

alter table hw1_department
    add school_id varchar(4) null;
```

```
 * mysql+pymysql://root:***@localhost
0 rows affected.
0 rows affected.
0 rows affected.
```

Out[26]:

```
[]
```

Non-Programming Track

Tasks

- There is a subdirectory in the project `data/GoT` that contains three CSV files:
  - `characters.csv`
  - `episodes.csv`
  - `character_relationships.csv`

- Your first task is to create tables to hold the data.
  - This means you must create three tables. Use a new schema and create the three tables:
    - `S22_W4111_HW1.characters`
    - `S22_W4111_HW1.episodes`
    - `S22_W4111_HW1.character_relationships.`
  - The table must have a column for each of the columns in the CSV.
  - You can use DataGrip or another tool to produce the create table statements, but you must execute the DDL statements in the code cells.

- Your second task is to load the data from the CSV files into the newly created tables. Do do this, you use a `LOAD` statement.

- Finally, you should examine the data and change column types to better reflect the actual values in the columns.

- To make the instruction more clear, I do an example of the tasks for another table. This is `got_imdb_names.csv.` You will do similar steps for the files above.

Example

- Manual examining the CSV file shows that the data has the following attributes.
  - `nconst`
  - `primaryName`
  - `birthYear`
  - `deathYear`
  - `primaryProfession`
  - `knownForTitles`

- So, my first step is to create a table to hold the information.

- **Note:** I have dozens of schema. So, I am prefixing this one with `aaaa_` to make it easy for me to find. You can drop this prefix.

- The following are the statements for creating the schema and table.

In [27]:
```
# Create the schema if it does not exist.
%sql create schema if not exists aaaa_S22_W4111_HW1;
```

```
 * mysql+pymysql://root:***@localhost
1 rows affected.
```

Out[27]:
```
[]
```

In [28]:
```
# Drop the table if it exists.
%sql drop table if exists aaaa_S22_W4111_HW1.got_imdb_actors;
```

```
 * mysql+pymysql://root:***@localhost
0 rows affected.
```

Out[28]:
```
[]
```

- Now create the table.

In [29]:
```
%%sql
create table if not exists aaaa_S22_W4111_HW1.got_imdb_actors
(
        nconst text null,
        primaryName text null,
        birthYear text null,
        deathYear text null,
        primaryProfession text null,
        knownForTitles text null
);
```

```
 * mysql+pymysql://root:***@localhost
0 rows affected.
```

Out[29]:
```
[]
```

- This is where it gets real and you do some wizard stuff.

In [30]:
```
# This command allows loading CSV files from the local disk.
# This is set of OFF by default.
# You should only have to run this once, that is if you execute the example, you
#
%sql SET GLOBAL local_infile = 'ON';
```

```
 * mysql+pymysql://root:***@localhost
0 rows affected.
```

Out[30]:
```
[]
```

In [31]:
```
# This is creating a connection to the database.
# You need to replace the user and passsword with your values for your installat
# Do not ask about the local_infile. That is Voldemort stuff.
#
con = pymysql.connect(host="localhost",
                      user="dbuser",
                      password="dbuserdbuser",
                      autocommit=True,
                      local_infile=1)
```

```
---------------------------------------------------------------------------
OperationalError                          Traceback (most recent call last)
/var/folders/g4/fks2s04x0jqf6w1j0cpy36xr0000gn/T/ipykernel_12213/3109354378.py i
n <module>
      3 # Do not ask about the local_infile. That is Voldemort stuff.
      4 #
```

```
----> 5 con = pymysql.connect(host="localhost",
      6                       user="dbuser",
      7                       password="dbuserdbuser",

~/opt/anaconda3/lib/python3.9/site-packages/pymysql/connections.py in __init__(s
elf, user, password, host, database, unix_socket, port, charset, sql_mode, read_
default_file, conv, use_unicode, client_flag, cursorclass, init_command, connect
_timeout, read_default_group, autocommit, local_infile, max_allowed_packet, defe
r_connect, auth_plugin_map, read_timeout, write_timeout, bind_address, binary_pr
efix, program_name, server_public_key, ssl, ssl_ca, ssl_cert, ssl_disabled, ssl_
key, ssl_verify_cert, ssl_verify_identity, compress, named_pipe, passwd, db)
    351             self._sock = None
    352         else:
--> 353             self.connect()
    354
    355     def __enter__(self):

~/opt/anaconda3/lib/python3.9/site-packages/pymysql/connections.py in connect(se
lf, sock)
    631
    632             self._get_server_information()
--> 633             self._request_authentication()
    634
    635             if self.sql_mode is not None:

~/opt/anaconda3/lib/python3.9/site-packages/pymysql/connections.py in _request_a
uthentication(self)
    919                 and plugin_name is not None
    920             ):
--> 921                 auth_packet = self._process_auth(plugin_name, auth_packe
t)
    922             else:
    923                 # send legacy handshake

~/opt/anaconda3/lib/python3.9/site-packages/pymysql/connections.py in _process_a
uth(self, plugin_name, auth_packet)
   1016
   1017         self.write_packet(data)
-> 1018         pkt = self._read_packet()
   1019         pkt.check_error()
   1020         return pkt

~/opt/anaconda3/lib/python3.9/site-packages/pymysql/connections.py in _read_pack
et(self, packet_type)
    723             if self._result is not None and self._result.unbuffered_acti
ve is True:
    724                 self._result.unbuffered_active = False
--> 725             packet.raise_for_error()
    726         return packet
    727

~/opt/anaconda3/lib/python3.9/site-packages/pymysql/protocol.py in raise_for_err
or(self)
    219         if DEBUG:
    220             print("errno =", errno)
--> 221         err.raise_mysql_exception(self._data)
    222
    223     def dump(self):

~/opt/anaconda3/lib/python3.9/site-packages/pymysql/err.py in raise_mysql_except
```

```
ion(data)
    141     if errorclass is None:
    142         errorclass = InternalError if errno < 1000 else OperationalError
--> 143     raise errorclass(errno, errval)

OperationalError: (1045, "Access denied for user 'dbuser'@'localhost' (using pas
sword: YES)")
```

In [ ]:
```
# This statement performs the load.
# You will need to change the TABLE name and the INFILE to the correct values.
#
sql = """
LOAD DATA LOCAL INFILE
'/Users/donaldferguson/Dropbox/Columbia/W4111-Intro-to-DB-S22/HWs/S22-W4111-HW-1
INTO TABLE aaaa_S22_W4111_HW1.got_imdb_actors
    FIELDS TERMINATED BY ','
    ENCLOSED BY '"'
    LINES TERMINATED BY '\n'
    IGNORE 1 LINES;
"""
```

In [ ]:
```
# Create a cursor. Again. Voldemort stuff, or maybe Sauron stuff.
#
cur = con.cursor()
```

In [ ]:
```
# Run the sql
cur.execute(sql)
```

In [ ]:
```
# Close the cursor. Sort of like the opposite of alohomora
cur.close()
```

In [ ]:
```
# Now test that your loading worked.
%sql select * from aaaa_S22_W4111_HW1.got_imdb_actors;
```

In [ ]:
```
%sql select * from aaaa_S22_W4111_HW1.characters;
```

- The final part of the task for each of the tables will be making some corrections.
- We would only ask you to do two or three corrections per table.
- Mine for this example would be in the following.

In [ ]:
```
%%sql

use aaaa_S22_W4111_HW1;

alter table got_imdb_actors modify nconst varchar(12) null;

alter table got_imdb_actors modify primaryName varchar(256) null;
```

```
alter table got_imdb_actors modify birthYear char(4) null;

alter table got_imdb_actors modify deathYear char(4) null;
```

Characters

- Perform the tasks for characters.

Episodes

- Perform the tasks for episodes.

Characters Relatrionships

- Perform the tasks for character_relationships.

Programming Track

Note: If you have activated student license when installing Datagrip, you can also use Pycharm Professional version instead of Community edition.

Tasks

- You will create and modify files in the directory `<uni>_web_src.`
- You will use the database that comes with the book, e.g. `db_book,` that you previously installed.
- Your web application will support `GET` on the path `/api/db_book/students/<ID>.` This means you have to implement two things:
  1. A function in `application.py` that implements the path endpoint.
  2. A method on a class `Student` that connects to the database, runs the SQL and returns the result. The project has been updated to have implementation templates for where your code goes.
- For submission, you must copy your code from the Python file below to show your code.
- You must include a screenshot of calling your application from a browser.

Modified application.py

```
from flask import Flask, Response, request
import json
from datetime import datetime
import rest_utils
```

```
app = Flask(__name__)


##########################################################################


# DFF TODO A real service would have more robust health check
methods.
# This path simply echoes to check that the app is working.
# The path is /health and the only method is GETs
@app.route("/health", methods=["GET"])
def health_check():
    rsp_data = {"status": "healthy", "time": str(datetime.now())}
    rsp_str = json.dumps(rsp_data)
    rsp = Response(rsp_str, status=200,
content_type="application/json")
    return rsp


# TODO Remove later. Solely for explanatory purposes.
# The method take any REST request, and produces a response
indicating what
# the parameters, headers, etc. are. This is simply for education
purposes.
#
@app.route("/api/demo/<parameter1>", methods=["GET", "POST", "PUT",
"DELETE"])
@app.route("/api/demo/", methods=["GET", "POST", "PUT", "DELETE"])
def demo(parameter1=None):
    """
    Returns a JSON object containing a description of the received
request.

    :param parameter1: The first path parameter.
    :return: JSON document containing information about the
request.
    """

    # DFF TODO -- We should wrap with an exception pattern.
    #

    # Mostly for isolation. The rest of the method is isolated from
the specifics of Flask.
    inputs = rest_utils.RESTContext(request, {"parameter1":
parameter1})

    # DFF TODO -- We should replace with logging.
    r_json = inputs.to_json()
    msg = {
        "/demo received the following inputs": inputs.to_json()
    }
    print("/api/demo/<parameter> received/returned:\n", msg)
```

```
    rsp = Response(json.dumps(msg), status=200,
content_type="application/json")
    return rsp

##########################################################################



@app.route("/api/db_book/students/<ID>", methods=["GET"])
def get_student_by_id(ID):
    msg = student_resource.get_by_id(ID)
    rsp_str = json.dumps(msg, default=str)
    rsp = Response(rsp_str, status=200,
content_type="application/json")
    return rsp


if __name__ == '__main__':
    app.run(host="0.0.0.0", port=5000)
```

Modified student_resource.py

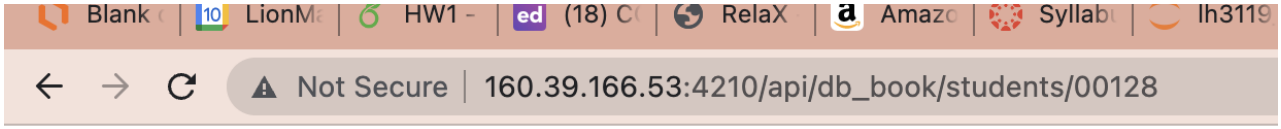
```
import pymysql


class Student:

    def __init__(self):
        # You may have to put code here.
        pass

    def get_by_id(self, ID):
        conn = pymysql.connect(host="localhost", user="root",
password="dvuserdvuser")
        sql = """
            select * from db_book.student where
                db_book.student.ID = %s
        """
        studentID = ID
        cur = conn.cursor()
        res = cur.execute(
            sql, args=studentID
        )
        res = cur.fetchall()
        return res
```

Screen Capture of Calling from Browser

In [1]:
```
screenshot = 'screenshot.png'
```

```
print("\n")
from IPython.display import Image
Image(filename=screenshot)
```

Out[1]:




```
[["00128", "Zhang", "Comp. Sci.", "102"]]
```