DSCI 551 – Fall 2021

Homework #2: XML/XPath

Deadline: 10/3, Sunday (100 points)

In this homework assignment, you will use the same data set as in hw1 and the same two access patterns:

- Pattern 1: Given a film category (e.g., family), find titles, release_year, rating (e.g., PG-13), rental_rate, and rental_duration about all the films in that category.
- Pattern 2: Given an actor name (its full name, e.g., "ed chase"), find the title and release years of all films the actor has acted in.

For all your scripts, please keep your runtime under 10 seconds! Otherwise all points will be deducted.

1. [50 points] Write a Python program load.py that converts all the data in the data set into a single XML document. Requirement: each table in the data set should be represented by a separate XML element in the document. Save your xml file as 'main.xml'
   Execution format: python load.py
   Here's an example output template for reference (You may create your own schema).

```
<root>
  <actor_table>
    <actor actor_id="1">
      <first_name>PENELOPE</first_name>
      <last_name>GUINESS</last_name>
      <last_update>2006-02-15 04:34:33</last_update>
    </actor>
    <actor actor_id="2">
      <first_name>NICK</first_name>
      <last_name>WAHLBERG</last_name>
      <last_update>2006-02-15 04:34:33</last_update>
    </actor>
```

2. [30 points] Write a Python program film.py for pattern 1 from the xml file you created in Q1. Execution format:
   python film.py <category>
   e.g., <category> may be family, comedy, sci-fi, etc. (case insensitive)
   Output format (print your results, one film per line as tuple; results are case insensitive):
   ('AMADEUS HOLY', '2006', 'PG', '0.99', '6')
   ('AMERICAN CIRCUS', '2006', 'R', '4.99', '3')
   …

If no films in a given category, output:

No results found.

3. [20 points] Write a Python script "actor.py" for pattern 2 from the xml file you created in Q1. Execution format:

   python actor.py <actor full name>

   e.g., <actor full name> may be "ed chase". (case insensitive)

   Output format (print your results, one film per line as tuple; results are case insensitive):

   ('ALONE TRIP', '2006')
   ('ARMY FLINTSTONES', '2006')
   ('ARTIST COLDBLOODED', '2006')
   ('BOONDOCK BALLROOM', '2006')
   …

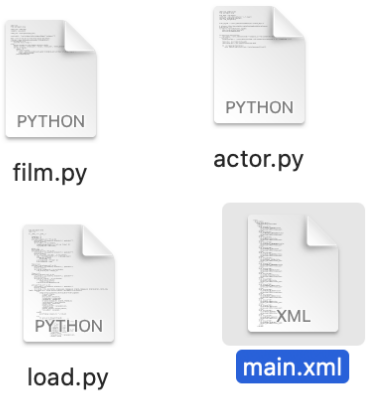
   If the actor did not play in any films, output:

   No results found.

Permitted libraries: lxml, sys, pandas

Submission

1. Three scripts described above
2. Your `main.xml`
3. Inside your submission folder should be exactly like



4. Put above required files in the same directory and compress it into a zip file.

   Zip file name format: LASTNAME_FIRSTNAME_HW2.zip