



IIC2115 – Programación como Herramienta para la Ingeniería (I/2021)

Laboratorio 1 Parte b: bases de datos

Objetivos

- Aplicar los conocimientos de bases de datos relacionales y adquisición y procesamiento de datos web para modelar y consultar datos.

Entrega

- **Lenguaje a utilizar:** Python 3.6 o superior
- **Lugar:** repositorio privado en GitHub. Recuerde incluir todo en una carpeta de nombre **L1b**.
- **Entrega:** jueves 27 de mayo a las 23:59 hrs.
- **Formato de entrega:**
 - Archivo Python Notebook (**L1b.ipynb**) con la solución de las misiones de este laboratorio. Utilice múltiples celdas de texto y código para facilitar la revisión de su laboratorio. **Deje todo ejecutado antes de realizar su commit**, se recomienda utilizar la opción de "restart and run all" disponible en jupyter notebook.
 - Archivo python (**L1b.py**) con el mismo código disponible en su archivo ipynb.
 - Todos los archivos deben estar ubicados en la carpeta **L1b**. No se debe subir ningún otro archivo a la carpeta. **No suba las bases de datos a GitHub o tendrá un descuento adicional inapelable de 5 décimas.**
- **Descuentos:** el descuento por atraso se realizará de acuerdo a lo definido en el programa del curso. Además de esto, tareas que no cumplan el formato de entrega tendrán un descuento de 0,5 ptos.

- **Laboratorios con errores de sintaxis y/o que generen excepciones serán calificadas con nota 1.0.**
- Si su laboratorio es entregado fuera de plazo, tiene hasta el **viernes 28 de mayo a las 11:59AM hrs** para responder el formulario de **entregas fuera de plazo** disponible en el Syllabus.
- Las discusiones en las *issues* del Syllabus en GitHub son parte de este enunciado.
- El uso de librerías externas que sean estructurales en la solución de los problemas no podrán ser utilizadas. Solo se podrán utilizar las que han sido aprobadas en las *issues* de GitHub.

Introducción

En esta parte del laboratorio deberá utilizar técnicas de bases de datos, *web scrapping* y expresiones regulares con el fin de: i) complementar los datos de una base de datos relacional y 2) responder consultas sobre los datos utilizando SQL.

Descripción de los datos

Considere la base de datos *Movies* utilizada en evaluaciones anteriores, que consiste en información básica de películas estadounidenses extraída de Wikipedia. Los datos se encuentran en un archivo en formato *json*, que puede ser abierto y manipulado utilizando la librería **json** de Python. Si lo prefiere, para extraer la información del archivo, debe utilizar las siguientes líneas de código:

```
import json  
with open('movies.json', encoding = 'utf8') as movies_file:  
    movies = json.load(movies_file)
```

Para cumplir las misiones de este laboratorio, es su responsabilidad explorar inicialmente el contenido del archivo y familiarizarse con el formato en que está almacenada la información.

Misiones

M0. Preparando la base de datos (0 ptos.): Esta misión no tiene puntaje y consiste en preparar los datos para las siguientes misiones. Para esto, replique lo desarrollado en las misiones 1 y 2 del Taller 3a, incluyendo un proceso de limpieza de registros con valores incorrectos o incompletos.

M1. Extracción e inserción de información complementaria (1.3 ptos.): Utilizando las librerías *Beautiful Soup* y *REMatch*, extraiga desde Wikipedia información complementaria sobre las entidades modeladas en la base de datos. Una vez extraída, inserte esta nueva información en la base de datos, actualizando las tablas donde corresponda. Se espera que agreguen a la base de datos al menos 2 nuevas columnas. Algunos ejemplos son: recaudación de cada película, fecha nacimiento de un actor, director de una película, duración de una película. Para evaluar esta misión, se considerará la cantidad de registros que son complementados con información adicional.

M2. Consultas (1.5 ptos.): Escriba en Python usando `sqlite`, consultas que permitan responder a las siguientes preguntas (todas las consultas tienen el mismo puntaje):

1. Encuentre los 3 años con más películas estrenadas.
2. Encuentre a los 5 actores con la trayectoria más larga, es decir, mayor cantidad de años actuando.
3. Construya un ranking descendente de popularidad de los géneros, donde para cada uno se muestren los 3 actores más populares.

Política de Integridad Académica

“Como miembro de la comunidad de la Pontificia Universidad Católica de Chile me comprometo a respetar los principios y normativas que la rigen. Asimismo, prometo actuar con rectitud y honestidad en las relaciones con los demás integrantes de la comunidad y en la realización de todo trabajo, particularmente en aquellas actividades vinculadas a la docencia, el aprendizaje y la creación, difusión y transferencia del conocimiento. Además, velaré por la integridad de las personas y cuidaré los bienes de la Universidad.”

En particular, se espera que mantengan altos estándares de honestidad académica. Cualquier acto deshonesto o fraude académico está prohibido; los alumnos que incurran en este tipo de acciones se exponen a un procedimiento sumario. Ejemplos de actos deshonestos son la copia, el uso de material o equipos no permitidos en las evaluaciones, el plagio, o la falsificación de identidad, entre otros. Específicamente, para los cursos del Departamento de Ciencia de la Computación, rige obligatoriamente la siguiente política de integridad académica en relación a copia y plagio: Todo trabajo presentado por un alumno (grupo) para los efectos de la evaluación de un curso debe ser hecho individualmente por el alumno (grupo), sin apoyo en material de terceros. Si un alumno (grupo) copia un trabajo, se le calificará con nota 1.0 en dicha evaluación y dependiendo de la gravedad de sus acciones podrá tener un 1.0 en todo ese ítem de evaluaciones o un 1.1 en el curso. Además, los antecedentes serán enviados a la Dirección de Docencia de la Escuela de Ingeniería para evaluar posteriores sanciones en conjunto con la Universidad, las que pueden incluir un procedimiento sumario. Por “copia” o “plagio” se entiende incluir en el trabajo presentado como propio, partes desarrolladas por otra persona. Está permitido usar material disponible públicamente, por ejemplo, libros o contenidos tomados de Internet, siempre y cuando se incluya la cita correspondiente.