Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

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

Laboratorio 2: web scraping y 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 **L2**.
- **Entrega: jueves 18 de noviembre a las 23:59 hrs.**
- **Formato de entrega:**
  - Archivo Python Notebook (**L2.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 (**L2.py**) con el mismo código disponible en su archivo ipynb.
  - Todos los archivos deben estar ubicados en la carpeta **L2**. No se debe subir ningún otro archivo a la carpeta. **No suba las bases de datos a su repositorio 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 19 de noviembre 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 del los problemas no podrán ser utilizadas. Solo se podrán utilizar las que han sido aprobadas en las *issues* de GitHub.

Introducción

En este laboratorio deberá utilizar técnicas de bases de datos y *web scraping* con el fin de: i) obtener los datos para construir una base de datos relacional y 2) responder consultas sobre los datos utilizando SQL.

Descripción de los datos

Considere la base de datos *Laureates*, que consiste en información sobre los ganadores históricos del premio Nobel en sus distintas categorías. Los datos se encuentran en un archivo en formato *json*, que puede ser abierto y manipulado utilizando la librería **json** de Python. Si decide utilizarla para extraer la información del archivo, debe usar las siguientes líneas de código:

```
import json
with open('laureates.json', encoding = 'utf8') as laureates_file:
    laureates = json.load(laureates_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

M1. Extracción de información web (3.0 ptos.): Utilizando la librería *Beautiful Soup*, extraiga desde Wikipedia toda la información necesaria para reconstruir el contenido de la base de datos *Laureates* y luego cree una base de datos relacional con esta información, utilizando como referencia de la estructura

y contenido de la base de datos resultante de ejecutar la pauta del Taller 3a. El puntaje obtenido en esta misión será proporcional a la similitud entre la base de datos de referencia y la creada por ud. a partir de la información de Wikipedia.

M2. Consultas (3.0 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 países con más ganadores de premios Nobel.
2. Encuentre los 5 ganadores más longevos, que hayan nacido y muerto en el mismo país.
3. Construya un ranking descendente de las afiliaciones con más ganadores, donde para cada una se muestre la cantidad de premios logrados en cada una de las categorías.

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.