PRÁCTICA 1: ¿CÓMO PODEMOS CAPTURAR LOS DATOS DE LA WEB?

AUTOR:

OSCAR JAVIER VÁSQUEZ CASALLAS

DOCENTE:

JOSE MOREIRA SANCHEZ

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

MESTRIA EN CIENCIA DE DATOS

UNIVERSITAT OBERTA DE CATALUNYA

19 DE NOVIEMBRE DE 2022

Contenido

1. Contexto

Se está creando un repositorio de información de investigación en la Corporación Colombiana de Investigación Agropecuaria AGROSAVIA el cuál servirá como núcleo y fuente de datos para el sistema de información de Sistema de Gestión Para La Información Científica Y Tecnológica CRIS, una de las entidades que conforman este sistema de información es la de Grupos de Investigación, que corresponde a un conjunto de personas que interactúan para discutir e intercambiar información técnico-científica y divulgativa en forma disciplinar e interdisciplinar, con el fin de investigar y generar productos y servicios de conocimiento en uno o varios temas (tendiente a la generación de oferta tecnológica que aporte al cambio técnico del sector agropecuario), de acuerdo con un plan de trabajo de corto, mediano o largo plazo. Se quiere cargar esta información al repositorio de datos, sin embargo, no se cuenta con un API que me permita recuperar los datos, ni tampoco se puede suministrar a través de algún sistema de información; una muy buena alternativa es el uso de web scraping para obtener estos datos ya que se encuentran en el sitio Web de la Corporación. La dirección del sitio Web es https://www.agrosavia.co/nosotros/grupos-de-investigacion.

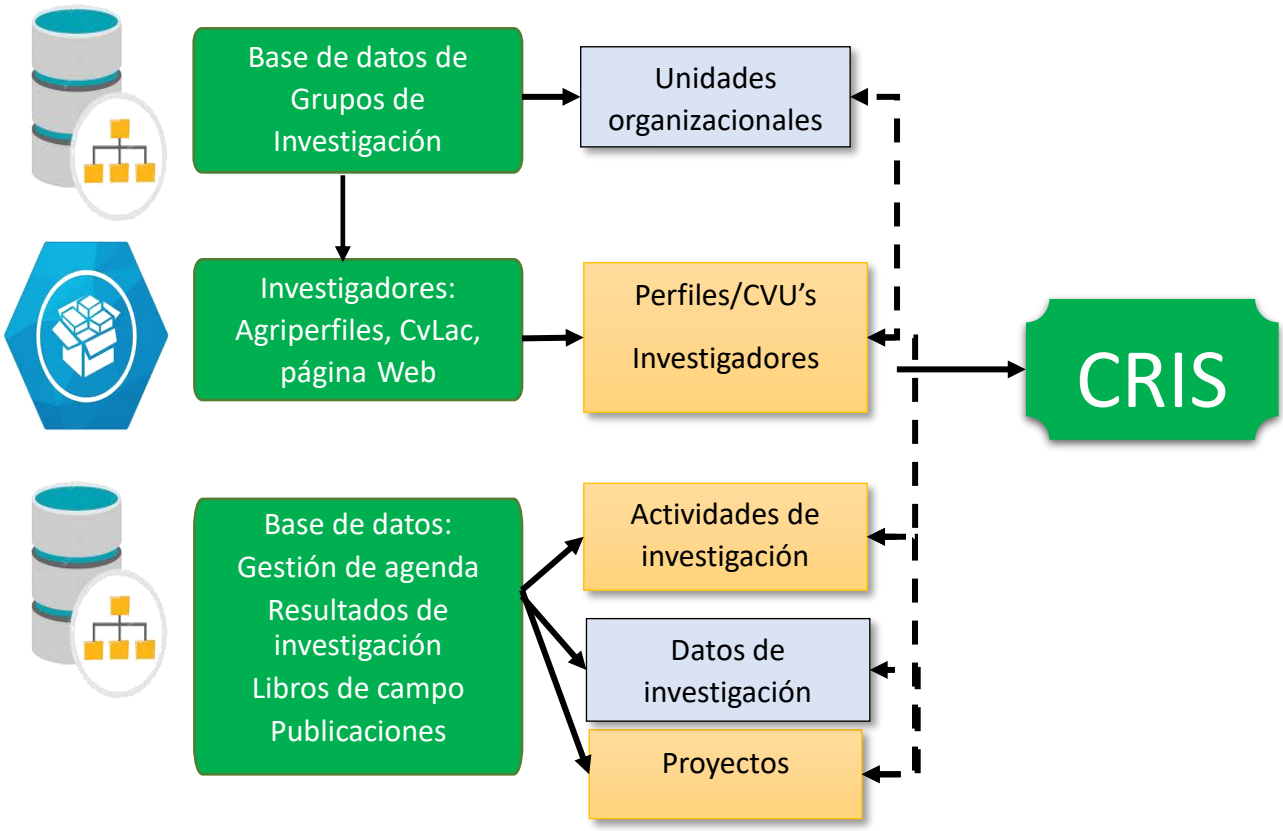
2. Título

Grupos de Investigación de la Corporación Colombiana de Investigación Agropecuaria AGROSAVIA.

3. Descripción del dataset

El dataset elegido para la práctica de Web Scraping corresponde a la campos generales de un grupo de investigación de la Corporación Colombiana de Investigación Agropecuaria AGROSAVIA. Cada una de las filas obtenidas traen los datos básicos propios del grupo y que también sirven de base para la categorización realizada por el Ministerio de Ciencia y Tecnología, entre ellos encontramos el nombre del grupo, líder, categoría, correo de contacto, áreas de conocimiento, año de creación y líneas de especialización

4. Representación gráfica



En la gráfica se puede observar que la entidad Grupos de Investigación hacen parte de un conjunto de información de diferentes fuentes que harán parte del repositorio de datos de investigación y este a su vez será el núcleo para el Sistema de Gestión Para La Información Científica Y Tecnológica CRIS AGROSAVIA.

5. Contenido

Estos son los campos de grupos de investigación que conforman el dataset:

- **titulo_grupo:** Nombre del grupo de investigación.
- **url_grupo:** URL de enlace a la página Web del grupo en el sitio de la Corporación.

- **lider_grupo:** Nombre de Investigador que cumple la función de liderar el grupo.

- **categoria_grupo:** Medición de grupos de investigación que propone Minciencias dentro del modelo, dependiendo de la cantidad y la calidad de los productos que genere un grupo se establece un puntaje para cada indicador entonces, cada grupo obtiene un puntaje determinado que establece la categoría del grupo, siendo A1 la más alta, las siguientes en su orden son A, B, C y grupo reconocido. El obtener la más alta categoría da cuenta de la calidad de la investigación de un grupo y sus integrantes, lo cual es un insumo importante para la acreditación en alta calidad de programas o instituciones académicas ante el Ministerio de Educación Nacional de Colombia.

- **correo_grupo:** Correo electrónico de atención de solicitudes y contacto del líder del grupo de investigación.

- **anocrea_grupo:** Año de creación del grupo

- **areaconoce_grupo:** Áreas de conocimiento en las que se especializa el grupo de investigación

- **prognal_grupo:** Programa nacional de ciencia y tecnología principal, clasificación dada por el Ministerio de Ciencia y Tecnología de acuerdo con el énfasis de conocimiento del grupo de investigación

- **prognalsec_grupo:** Programa nacional de ciencia y tecnología secundario, clasificación dada por el Ministerio de Ciencia y Tecnología de acuerdo con el énfasis de conocimiento del grupo de investigación.

- **lineas_grupo:** Líneas de investigación que trabaja el Grupo.

Esta información lleva un periodo de tiempo aproximado de 30 años, coincide con la creación de la Corporación Agrosavia, desde este momento se comenzaron a formar grupos de investigación. La información es actualizada constantemente con la inclusión de miembros y para la medición y categorización realizada por el Ministerio de Ciencia y Tecnología.

6. Propietario

El propietario de los datos obtenidos de grupos de investigación es la Corporación Colombiana de Investigación Agropecuaria. Esta información se encuentra en el sitio web de la Corporación, a la cual le agradecemos por su publicación, ya que es de carácter pública e informativa. Para obtenerla se realizó a través del lenguaje de programación Python técnicas de Web Scraping para extraer la información contenida en las páginas HTML

Se realizaron búsqueda de análisis previos a este conjunto de datos, pero no fueron encontradas. Sin embargo, si se encontró un tipo de Análisis similar a este, realizado también por estudiantes de la Universidad de Cataluña para la asignatura M2.851 Tipología y ciclo de vida de los datos. Los estudiantes Alejandro Medina y Federico Clavijo obtuvieron a través de Web Scraping el ***"Dataset de los grupos de investigación colombianos con categoría A y A1 y que hacen parte de los Programas Nacionales de Ciencia, Tecnología e Innovación en Salud e Ingeniería"***, el cual se obtuvo a través de las bases de datos de plataformas sobre currículos de investigadores (CvLAC) y hojas de vida de grupos de investigación (GrupLAC) colombianos pertenecientes al Ministerio de Ciencia Tecnología e Innovación (Minciencias). El dataset se puede encontrar en https://zenodo.org/record/6441181#.Y3T103bMKUm.

Al ser una información de carácter público y para acceder a ellas no fue necesario aceptar una serie de términos y condiciones, se asumió que se puede hacer un uso moderado de web scraping, adicionalmente los datos serán utilizados para la conformación de un repositorio de datos de investigación de la misma Institución, sin embargo, se tuvo en cuenta la siguiente:

- Se revisaron los términos y condiciones y la política de privacidad de la información, sin hallar ninguna restricción sobre el uso de esta información.
- Se trata de acceder al archivo robots.txt, pero no fue posible encontrarlo.
- Se realiza consulta del propietario de la información
- No sobrecargar el servidor con muchas peticiones
- No se intentó acceder a servidores o equipos sin permiso de acceso

7. Inspiración

Este conjunto de datos es interesante ya que permite conocer los diferentes grupos de investigación existentes en Agrosavia y que son reconocidos por el Ministerio de Ciencia y Tecnología, esto le podría brindar a los usuarios de la Corporación la posibilidad de saber el escalafón que cada grupo tiene, al igual las áreas del conocimiento en que son especializados y a que Programa nacional de ciencia y tecnología pertenece, con el fin de obtener asesoría sobre los sistemas productivos. Adicionalmente se tiene el contacto del líder del grupo para realizar cualquier tipo de solicitud con el mismo. Desde el punto de vista de análisis al tener un histórico de la evolución del grupo a través de las diferentes mediciones realizadas se puede llegar a determinar el crecimiento o decrecimiento de los mismos con el fin de fortalecer los recursos y las capacidades.

Comparando los datos que se obtienen con el dataset presentado en el punto 6 ***"Dataset de los grupos de investigación colombianos con categoría A y A1 y que hacen parte de los Programas***

Nacionales de Ciencia, Tecnología e Innovación en Salud e Ingeniería", se observa que los campos generados son muy similares, la diferencia radica en el filtro que su utilizo, ya que en mi proyecto extraigo los datos de todos los grupos de investigación de Agrosavia, el proyecto de mis compañeros obtienen los datos de los grupos categorizados como A1 y que pertenecen a Programas Nacionales de Ciencia, Tecnología e Innovación en Salud e Ingeniería. En ambos set de datos se recupera la información básica del grupo (nombre, líder, programa, área), el único dato relevante que observo que falta en el dataset de mi proyecto es el código de grupo asignado por MincienclaS, esto código considero que sería de gran ayuda si se quisiera realizar otros procesos de Web Scraping para por ejemplo conocer los integrantes del grupo de investigación.

8. Licencia

Considero que al dataset resultante se le puede asignar la licencia Atribución-NoComercial-Compartir Igual 4.0 Internacional (CC BY-NC-SA 4.0), a pesar de que los datos son de Dominio Público, se deben dar los créditos al momento de copiar, modificar, distribuir e interpretar y tiene las siguientes características según lo especificado en creative commons:

- Se debe dar crédito de manera adecuada, brindar un enlace a la licencia, e indicar si se han realizado cambios.
- No se puede hacer uso del conjunto de datos con propósitos comerciales.
- Si remezcla, transforma o crea a partir del material, debe distribuir su contribución bajo la misma licencia del original.

9. Código

El código con el que se ha obtenido el dataset se desarrolló en lenguaje Python y se encuentra en la carpeta **source** del repositorio de GitHub bajo el nombre **PracticaWebScraping.ipynb** (https://github.com/oscajvasquez/practica1_webscraping), en esta misma carpeta se adjunta el archivo **requirements.txt** donde se especifican las librerías y versiones utilizadas en el desarrollo del proyecto.

El código realiza Web Scraping a través de la utilización de las librerías Requests y BeautifulSoup, con las cuales se obtiene el contenido de la página principal de grupos de investigación y luego se crea una estructura anidada por los elementos que la componen. Posteriormente se encuentran todos las etiquetas de tipo <a>, para poder acceder a cada página especifica de grupo. Luego través

las manipulación de los diferentes objetos de BeutifulSoup (Tag, NavigableString, BeautifulSoup y Comment) y a través de la navegación vertical y horizontal del HTML se obtienen los valores de todas las columnas que conforman el dataset. Se aplica diferentes funciones para la limpieza y normalización de las cadenas de caracteres resultantes. Por último, se convierte el dataframe resultante a un archivo en formato .csv

La dificultad mas importante que se afronto en la realización del código y que presentaba el sitio web es que los diferentes elementos y etiquetas que conformaban la estructura HTML no se encontraban adecuadamente maquetados, no existan clases ni id para identificar los diferentes objetos; esta serie de inconvenientes se resolvieron utilizando la navegación vertical y horizontal de los objetos y usando condicionales para a través de los textos de las etiquetas poder obtener el valor deseado para cada columna.

10. Dataset

El dataset obtenido se incluyó en la carpeta /dataset del repositorio bajo el nombre **Dataset_Final.csv**.

Adicionalmente se publicó el dataset en Zenodo en el siguiente enlace se puede acceder: https://zenodo.org/record/7335188#.Y3ewLXbMKUk e igualmente fue asignado el siguiente DOI: https://doi.org/10.5281/zenodo.7335188.

11. Vídeo

Se realiza y sube el vídeo en el Google Drive de la Universidad en el siguiente enlace
https://drive.google.com/file/d/1MNo-66p1hUCVInlRcR3PQJ0P9yBGBjOc/view?usp=share_link

CONTRIBUCIONES	FIRMA
Investigación previa	OJVC
Redacción de las respuestas	OJVC
Desarrollo del código Integrante	OJVC
Participación en el vídeo	OJVC