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.

M2.851 - Tipología y ciclo de vida de los datos aula 1 Práctica 1

Marzo 2022

Realizado por:

- FEDERICO CLAVIJO LÓPEZ
- ALEJANDRO MEDINA UICAB

DESARROLLO

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

En Colombia, el Ministerio de Ciencia Tecnología e Innovación (Minciencias) regula, gestiona y financia la mayor parte de las actividades de innovación, investigación y desarrollo de tecnología, a través de diferentes convocatorias públicas y programas gubernamentales. En el marco de esta gestión Minciencias ha generado varias bases de datos para gestionar y divulgar de forma estructurada, la información concerniente a entidades que investigan, grupos de investigación, investigadores y convocatorias vigentes, a través de plataformas en línea de acceso público.

Una de estas plataformas fue escogida para el desarrollo de esta práctica (https://scienti.minciencias.gov.co/ciencia-war/) en ésta se puede acceder a las bases de datos que recogen toda la información sobre currículos de investigadores (CvLAC) y hojas de vida de grupos de investigación (GrupLAC) colombianos.

Si bien en la plataforma se puede consultar la información a través varios criterios, hay búsquedas que no se pueden realizar de forma sencilla e implica un proceso tedioso de filtrado manual. A partir de estas dificultades se decidió centrar esta práctica en: recaudar la información concerniente a los grupos de investigación que tengan categoía A y A1 y que hagan parte de los Programas Nacionales de Ciencia, Tecnología e Innovación en Salud y en Ingeniería, sin importar si estas son la temática principal o secundaria del grupo. Información consultada en esta pagina web: https://scienti.minciencias.gov.co/ciencia-war/busquedaConteoGrupoXProgramaNacional.do?by=primario&codPrograma=2

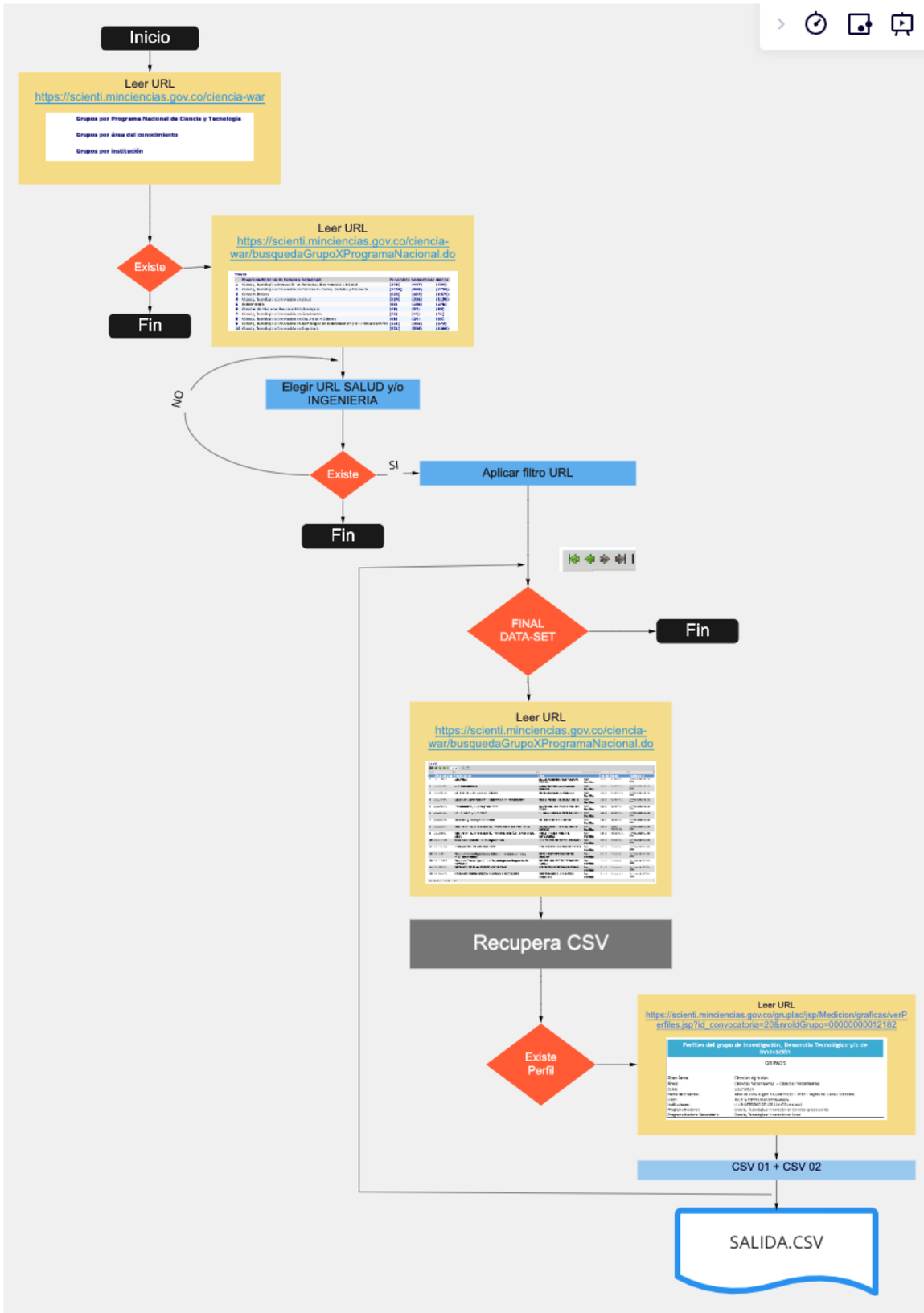
2. Título. Definir un título que sea descriptivo para el dataset:

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.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

El dataset obtenido con el web scraping corresponde al listado de los grupos de investigación categoría A y A1 registrados por el Ministerio de Ciencia Tecnoloagía e Innovación de Colombia, que hacen parte de los Programas Nacionales de Salud e Ingeniería, sin importar si estas son la temática principal o secundaria del grupo. Esta información es consultada en la pagina web oficial del Ministerio: https://scienti.minciencias.gov.co/ciencia-war/busquedaGrupoXProgramaNacional.do la cual contiene cerca de 5751 registros.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido.

El dataset obtenido con el web scraping corresponde a los grupos de investigación categoría A y A1 de las temáticas de salud e ingeniería. Los campos de información de cada grupo son:

- id: Corresponde al número de identificación del listado de grupo en cada b+usqueda realizada
- codigo_grupo: Código de identificación de cada grupo de investigación. Primary key del registro de grupos
- nombre_grupo: Nombre del grupo
- lider: Investigador principal del grupo
- avalado: Corresponde al número de instituciones que avalan la existencia del grupo de investigación sobre el total reportado
- estado: Indica la categoría que tiene el grupo de investigación, que para el presente data set corresponde a categoría A y A1
- calificado_en: refleja la convocatoria y el año en el cual fue evaluado y otorgado esa categoría reportada.
- Programa nacional: Corresponde a la temática principal de estudio del grupo de investigación
- Programa nacional secundario: Corresponde a la temática secundaria de estudio

- Programa tipo: Corresponde al área de conocimeinto sobre la cual trabaja el grupo de investigación, que puede ser Ingeniería o Salud.

Este dataset se ha recogido a partir de esta pagina web: https://scienti.minciencias.gov.co/ciencia-war/busquedaGrupoXProgramaNacional.do y a partir de ella se accedió a los grupos de investigación de Ciencia, Tecnología e Innovación en Salud e Ingeniería en la opción de "Ambos", ya que esto corresponde a que estas temáticas pueden ser tanto tema principal como segundario. Con lo cual, se extrajo la información de los siguientes links:

Salud: https://scienti.minciencias.gov.co/ciencia-war/busquedaConteoGrupoXProgramaNacional.do?by=union&codPrograma=2 Ingeniería: https://scienti.minciencias.gov.co/ciencia-war/busquedaConteoGrupoXProgramaNacional.do?by=union&codPrograma=15

Posteriormente se filtraron las tablas para unicamente tomar en consideración los grupos que tuvieran categoría A y A1. Para el caso de salud se tuvieron de base 1196 regisros y para ingeniería 1009, y la búsqueda se configuró para mostrar el máximo de 100 registros por página y que tuvieran Categoría A o A1, lo cual requirió de una función que permitiera acceder a los siguientes resultados hasta completar el total.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

La plataforma Scienti de Minciencias tiene un carácter público de consulta y restringe el uso comercial de información transformada a partir de sus plataformas. Según consulta del sitio web y dadas las características encontradas frente al tipo de información pública se decidió realizar el web scraping. Contemplando que está prohibida la reproducción total o parcial, traducción, inclusión, transmisión, almacenamiento o acceso a través de medios analógicos, digitales o de cualquier otro sistema o tecnología creada o por crearse, sin autorización previa y escrita de MINCIENCIAS. No obstante, es posible descargar material de las plataformas para uso personal y no comercial, siempre y cuando se haga expresa mención de la propiedad en cabeza MINCIENCIAS.

Razón por la cual se aclara que esta práctca corresponde a una actividad académica, sin fines de lucro. Y que acepta los términos y condiciones de la página web rastreada.

Frente a las recomendaciones para el buen uso de web scraping, es importante considerar proteger la integridad de la página web evitando sobrecargarla con consultas, para lo cual se puede limitar el número de consultas a la web desde la misma IP. De igual forma, se puede programas el proceso de web scraping para que se lleve a cabo en las horas de menor uso. Otras recomendaciones corresponden a hacer revisión de la existencia de los link presentes en el proceso de web scraping, para evitar que futuras actualizaciones en la página web, afecte la operatividad del código.

Así mismo es importante estan atentos ante cambios en la política de Minciencias que puedan llegar a afectar la legalidad del dataset resultante. Y hacer revisión de la existencia de Robots.txt, que si bien actualmente en la página web seleccionada no existe, si puede llegar a estar presente en un futuro. En este caso, es importante estar analizando con regularidad el contenido de robots.txt para identificar posibles restricciones a tener en cuenta en el web scraping. Finalmente y la más importante de las recomendaciones serí abogar por el uso responsable tanto de los datos como del proceso de web scraping, respetando el objetivo en la existencia de esta información determinado por sus creadores, en este caso Mincienias.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El presente listado corresponde a los grupos de investigación con mejor calificación que trabajan en temas de Salud e Ingeniería en Colombia. Este listado permite reconocer aquellos actores relevantes en la generación de conocimiento, facilitando la búsqueda de entidades para establecer alianzas al rededor de estos temas, ya sea para investigar e innovar. Esta información consolidada facilita a las empresas y universidades con quienes pueden trabajar en estas temáticas específicas.

Este se convierte es un instrumento para el libre acceso a la información sobre la oferta nacional de investigación y desarrollo tecnológico en términos de capacidad científica y tecnológica de Colombia y de los resultados y productos de la investigación existentes.

Esta información cambia una vez al año, sin embargo, la página web suele no ser muy consultada por las empresas debido a su complejidad y visualización poco amena. Lo cual, lleva a esta base de datos, a perder valor frente a su información gestionada.

8. Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

Se escoge la licencia CC BY-NC-SA 4.0 License (Reconocimiento-NoComercial-CompartirIgual). Debido a que la obra original tiene esta misma condición. Esta licencia no permite un uso comercial, siendo este el objetivo de la Minciencias, ya que el interés es dar esta información al público en general.

Según políticas del Ministerio, en el marco de este trabajo damos reconocimiento de su autoría, proporcionamos el enlace a la fuente original de los datos e indicamos los cambios realizados. Dejando claro un uso no comercial del resultado. De igual forma, nos restringimos al criterio "compartirIgual", el cual nos invita a difundir bajo la misma licencia la remezcla, transformación y/o creación a partir del material de base.

9. Código. Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

Se adjunta en el repositorio de Github el código desarrollado en python en el archivo denominado Codigo_practica.py

10. Dataset. Publicar el dataset obtenido()* en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

Se adjunta en el repositorio de Github el dataset de salida como producto del web scraping en el archivo Salida.csv. De igual forma se publica el dataset en Zenodo y el código DOI es: 10.5281/zenodo.6441181 Link: https://doi.org/10.5281/zenodo.6441181

11. Vídeo. Se debe hacer entrega de un vídeo explicativo de la práctica en donde cada uno de los integrantes del grupo explique con sus propias palabras tanto las respuestas del proyecto como el código utilizado para llevar a cabo la extracción. El vídeo debe ser enviado a través de un enlace a Google Drive que deben proporcionar, junto con el enlace al repositorio Git, al momento de entregar la práctica.

El video fue cargado en el drive y se puede acceder a través del link en el README

Contribuciones	Firma
Investigación previa	AM - Alejandro Medina Uicab y FC - Federico Clavijo López
Redacción de las respuestas	AM - Alejandro Medina Uicab y FC - Federico Clavijo López
Desarrollo del código	AM - Alejandro Medina Uicab y FC - Federico Clavijo López