Dataset: Conjuntos de datos gubernamentales abiertos de Colombia, uso y preferencias

Camilo Octavio Baez Ramos

11/04/2019

Contexto

La penetración de la tecnología, disminución del nivel de analfabetismo digital y mayor cobertura de Internet en las diferentes poblaciones del país ha ocasionado un aumento en la generación de datos tanto en entidades públicas como privadas.

El estado Colombiano dispuso de un sitio donde se publican los diferentes datasets generados por las entidades gubernamentales con el objetivo de que cualquier persona del territorio nacional los utilice de forma libre y sin restricciones para desarrollar aplicaciones o servicios de valor agregado, hacer análisis e investigación, ejercer labores de control o para cualquier tipo de actividad comercial o no comercial.

Descripción del dataset

Dataset del repositorio de datos abiertos gubernamentales de Colombia publicados en el sitio www.datos.gov.co para analizar el nivel de aceptación del proyecto open data en Colombia.

Contenido

La base de datos de datos de datasets abiertos de entidades públicas de Colombia se extrajo mediante un script desarrollado en lenguaje de programación Python utilizando las librerías BeautifulSoup y Selenium, la primera se utiliza para navegar el DOM del HTML de páginas estáticas y la segunda se implementó con un driver de Google Chrome para acceder al DOM del HTML generado dinámicamente con AJAX por el sitio.

Para cada registro presentado en el sitio web se recolectaran los siguientes datos:

- ***ID,*** identificador autoincremental generado por el scrapper para identificar los registros.
- ***Nombre,*** nombre del registro.
- ***Descripción,*** descripción del registro.
- ***URL,*** URL a la que direcciona el registro.
- ***Clase,*** clasificación del registro ej: Conjunto de datos, visualización, mapa, etc.
- ***Creado,*** fecha de creación del registro.
- ***# Visitas,*** cantidad de visitas al registro mediante el acceso a la URL.
- ***Temas,*** temáticas referenciadas por el registro.

Si el registro es de la categoría “Conjunto de datos” se accede a la página descrita en el campo URL y se extrae la siguiente información generada por Ajax:

- ***Área o Dependencia,*** Área de la entidad que genera el dataset.
- ***Nombre de la Entidad,*** entidad gubernamental dueño del dataset.
- ***Departamento,*** departamento nacional donde está ubicado la entidad.
- ***Municipio,*** municipio nacional donde está ubicado la entidad.

- ***Orden,*** orden al que pertenece el dataset. Ej.: Territorial.
- ***Sector,*** sector de la sociedad sobre el que trata el dataset.
- ***Idioma,*** idioma en el que se genera el dataset.
- ***Cobertura Geográfica,*** nivel de cobertura del dataset, ej.: municipal, regional, nacional.
- ***Frecuencia de Actualización,*** frecuencia con el que la entidad actualiza el dataset en el sitio.
- ***Categoría,*** categoría del dataset.
- ***Descargas,*** cantidad de descargas que ha tenido el dataset.
- ***Filas,*** cantidad de filas o registros que posee el dataset.
- ***Columnas,*** cantidad de columnas que posee el dataset.

Agradecimientos

Los datos han sido recolectados desde el sitio gubernamental de Colombia www.datos.gov.co. Para ello, se ha hecho uso del lenguaje de programación Python y de técnicas de Web Scraping para extraer la información alojada en las páginas HTML con las librerías BeautifulSoup (raspado de HTML estático) y Selenium (raspado de HTML generado dinámicamente con Ajax).

Inspiración del proyecto

Con el objetivo de analizar el nivel de aceptación del proyecto, se realizará rapado web de los datasets publicados en el sitio, así como las variables que faciliten el análisis (sitios geográficos, temas, sector) e identificación de preferencias del dataset como son cantidad de descargas y visualizaciones del mismo.

Datasets similares al recolectado han permitido la generación de estudios como el realizado en el país Español sobre el consumo de datos por parte del público general [1], sin embargo, este estudio se basó en encuesta realizada a través de internet la cual puede, o debe, contrastarse contra el uso real de los datos.

Licencia

Código fuente y dataset

Tanto el código fuente escrito para la extracción de datos como el dataset generado pueden ser accedidos a través del enlace: https://github.com/camilobaez1/ws_datos_gov_co

Tabla de contribuciones al trabajo

Contribuciones	Firma
Investigación previa	**COBR**
Redacción de las respuestas	**COBR**
Desarrollo código	**COBR**

Trabajos citados

[1] S. Álvarez-García, M. Gértrudix y M. D. C. Gertrudis Casado, Consumo de datos abiertos de las instituciones públicas por parte de los ciudadanos españoles, Madrid, 2016.