Tipología y ciclo de vida de datos
PRA1 - Web Scraping

Integrantes:
Giovanny Eduardo Caluña Chicaiza - Ivan Dario Ovalle Benavides
Fecha:
07/04/2021

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

Un usuario busca la mejor opción para la compra de un computador personal (macbook pro). Se entiende por mejor opción a: el mejor descuento o precio. Debido al poco conocimiento, al usuario también le gustaría tener recomendaciones de los mejores productos a pesar de no contar con descuentos. Toda la búsqueda de información se realizará en el sitio web Amazon, ya que este es es el mayor marketplace del mundo donde la variedad de productos encontrados allí es muchísimo mayor que cualquier competidor y los precios son más asequibles.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

El dataset se llamará: TopAmazonPromos.csv. Este nombre trata de describir el objetivo del dataset, encontrar los mejores descuentos y/o precios para un artículo (computadora personal en este proyecto).

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)

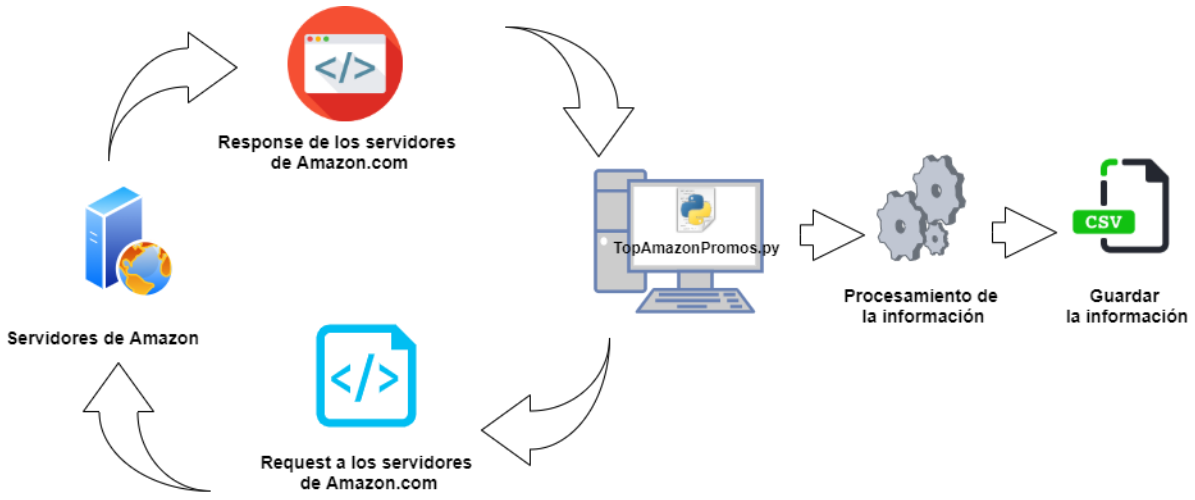
El data set cuenta con 10 atributos que nos ayudarán a definirlo. Los atributos son :

- **Título:** Atributo de tipo string**.** Aquí se almacena el nombre del artículo.
- **Título abreviado:** Atributo de tipo string**.** En este campo se almacena el nombre acotado a una longitud (30 caracteres).
- **Link:** Atributo de tipo string**.** URL específico para un producto en la web de amazon.
- **Precio actual:** Atributo de tipo float**.** Este atributo es el precio al cual se está comerciando el producto al momento de la captura de datos.

- **Precio sin descuento:** Atributo de tipo float**.** Este atributo es el precio anterior al cual se está comerciando el producto al momento de la captura de datos. (si no hay precio anterior, se toma el actual)
- **Críticas:** Atributo de tipo entero**.** El número de críticas realizadas a los productos por los compradores.
- **Calificación:** Atributo de tipo float**.** La calificación de los compradores a un producto está dada en un rango de 1 a 5 donde 5 es la mejor calificación, para este campo se guarda el promedio de todas las calificaciones.
- **Amazon Choice:** Atributo de tipo booleano**.** Este valor será verdadero cuando el producto esté marcado por la etiqueta “Amazon Choice”.
- **Fecha:** Atributo de tipo date**.** Aquí se guarda la fecha en que se realizó la extracción de datos en el formato dd/mm/aa.
- **Porcentaje de descuento:** Atributo de tipo float**.** Aquí se registrará el descuento al momento de la extracción de datos.

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

A continuación se presenta un pequeño esquema que representa el proceso que se seguirá para obtener el dataset. El proceso empieza ejecutando el script “TopAmazonPromos.py” de python. El script arma la solicitud y la envía hacia los servidores de Amazon. Una vez que los servidores envían una respuesta, el script toma la información donde la filtra, ordena, calcula y añade los datos necesarios. Finalmente el script convierte toda la información a un archivo plano de tipo .csv y la guarda en la computadora local.



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

Como se mencionó en el apartado anterior, el dataset cuenta con 10 campos. Los datos de cada registro son recolectados desde la página web del marketplace

Amazon, para un producto en particular a lo largo de un determinado tiempo. Para esto, se crea un URL con el sitios web de amazon y el producto del cual queremos extraer la información. Esta URL nos devolverá una larga lista de artículos que coinciden con el producto que buscamos. El programa va a tomar cada uno de los artículos de la lista y va a extraer la siguiente información de cada uno:

- **Título:** El título contiene el nombre del artículo que generalmente se encuentra junto a varias características. Se extrae directamente de la página web.
- **Título abreviado:** En este campo se almacena el nombre acotado a una determinada longitud (30 caracteres). Se obtiene a través de una operación sobre el campo anterior. Este campo se crea para mayor facilidad durante el análisis de los datos.
- **Link:** Para cada producto, Amazon crea un URL directo donde se encuentra toda la información de cada producto. El link se extrae directamente de la página web. En caso de que el usuario quiera mirar a detalle cierto producto, podrá acceder a través de este link.
- **Precio actual:** Este atributo toma el precio al cual se está comerciando el producto al momento de la captura de datos. Se extrae directamente de la página web.
- **Precio sin descuento:** Amazon muestra el precio original o al que se solía vender el producto previo a un descuento u alguna otra rebaja debido a diferentes factores. Para analizar estos cambios y “aprovechar” la oportunidad, este valor se almacena. Este valor se extrae directamente de la página web.
- **Críticas:** Las críticas a los productos son muy útiles al momento de realizar una compra, ya que nos ayudan como referencia para la compra de un producto. Por esta razón, se almacena el número total de críticas realizadas al producto. Se extrae directamente de la página web.
- **Calificación:** La calificación de un producto está dada en un rango de 1 a 5 donde 5 es la mejor calificación. Este valor es fundamental para una compra, sin embargo este debe ser contrastado con el numero de criticas para una mejor interpretación y selección. La calificación se extrae directamente de la página web.
- **Amazon Choice:** Amazon se caracteriza por dar opciones llamadas “amazon choice”, que generalmente se las asigna a productos que han mostrado una buena relación entre calidad-precio y por la alta acogida de los usuarios. Este atributo se marca con un booleano si el artículo es o no una opción recomendada de amazon (amazon choice). Ese valor se define cuando se verifica si existe el atributo en la página web.
- **Fecha:** Aquí se guarda la fecha en que se realizó la extracción de datos en dia mes y año. Este valor se obtiene con la función Datetime de python.
- **Porcentaje de descuento:** El porcentaje de descuento se calcula con el precio actual y precio sin descuento utilizando una función simple implementado en python. El dataset se ordenará de manera decente basado

en este campo ya que buscamos el mejor descuento y de ahí partimos a analizar diferentes factores.

BUENAS PRÁCTICAS:
Los datos fueron recolectados por medio de web scraping usando el lenguaje de programación Python, para lo cual se realizó la consulta del fichero robots.txt para respetar aquellas rutas y subdominios en los cuales se debe evitar acceder. Para realizar la extracción, se crean tres funciones principales que se usan para obtener la data, buscar los datos específicos del producto que son de interés y realizar la navegación dentro de varias páginas del producto. Posteriormente se guarda la información obtenida en una lista, se realiza el cálculo de descuento del producto si es que lo hay y se exporta la información a un archivo csv llamado ¨TopAmazonPromos.csv¨

Para el caso de los bloqueos se ha utilizado un delay de 5 segundos y adicionalmente se ha condicionado las búsqueda las primeras 20 páginas del producto. Información más detallada y comentarios se encuentran disponibles dentro el archivo de python llamado: TopAmazonPromos.py.

Es muy importante aclarar que los datos se han extraído en formato .csv como raw data. Esto quiere decir que cualquier proceso de preprocesado de datos o limpieza se tendrá que hacer luego de cargar el fichero y de acuerdo a requerimientos específicos posteriores que se necesiten para el estudio

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.

Por medio de este apartado expresamos formalmente nuestros agradecimientos a la compañía Amazon cuyo accionista mayoritario es Jeff Bezos, por permitir extraer la data para nuestro proyecto educativo.

7. Inspiración. Explique 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 conjunto de datos obtenido es bastante interesante porque nos permite tener agrupados precios, descuentos, número de críticas, calificaciones y otros datos de manera muchísimo más rápida que hacer la búsqueda de forma manual. Lo interesante es que la necesidad de este estudio nace desde un problema personal, en específico, tener una idea de lo que se quiere comprar pero no tener el tiempo

para revisar cual proveedor entre los miles que tiene este marketplace ofrece el producto o uno similar a mejor precio y con buen feedback.

Específicamente, luego de recolectar la información del sitio web se podrán contestar preguntas del tipo: ¿El producto que busco tiene rebaja con algún proveedor?¿Cuáles son los mejores proveedores con mejor calificación para el producto que busco? ¿Cuántas críticas ha recibido el proveedor sobre el producto en cuestión? vale la pena mantener la decisión de compra sobre el producto o puedo escoger uno similar recomendado por Amazon? entre otras que surgirán a partir de la transformación de los datos.

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:**
   Released Under CC0: Public Domain License
   Released Under CC BY-NC-SA 4.0 License
   Released Under CC BY-SA 4.0 License
   Database released under Open Database License, individual contents under Database Contents License
   Other (specified above)
   Unknown License

   Para nuestro dataset se ha elegido el tipo de licencia libre o Open Database License. Se toma esta opción ya que los datos guardados son de dominio público y que se encuentran publicados en la web (Amazon). Además, en caso de que algún usuario utilice la información y la modifique, no generará ningún inconveniente ya que el propósito específico es inmediato. De hecho, que sea libre ayudaría a que el dataset siga aumentando en tamaño con futuras actualizaciones hechas por los usuarios.

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

   El código se encuentra disponible en el siguiente repositorio:
   https://github.com/Giiovhanny/PRA1_WebScrapping_TipologiayCicloDeVidaDeDatos

10. **Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción**

    Una vez publicado el data set en Zenodo, obtenemos el siguiente identificador (DOI) y el link donde se encuentra publicado.

    Dataset DOI: 10.5281/zenodo.4676394
    Link Dataset: https://zenodo.org/record/4676394

11. **Contribuciones del trabajo:**

Contribuciones	Firma
Investigación Previa	G.E.C.C. - I.D.O.B.
Redacción de las respuestas	G.E.C.C. - I.D.O.B.
Desarrollo código	G.E.C.C. - I.D.O.B.