Dataset: Datos de Productos de la competencia

Pablo A. Delgado

12/04/2021

Contexto

Como analista de datos de la empresa de productos de hogar "*AIQUIA*" se desea realizar un estudio de mercado de los productos ofrecidos por aquellas empresas que, a partir de notas en medios de información tradicionales, redes sociales y del boca a boca tienen cada vez más resonancia y parecen ser competidores relevantes para nuestro negocio. Con este estudio se desea confirmar o no esta hipótesis, y poder así plantear posibles estrategias de precios o campañas de marketing para mejorar nuestra posición. Es por eso que en primera instancia se deberá recolectar los datos públicos de sus productos para su posterior análisis.

Titulo

Dataset: competitors_raw_data.csv

Título: Datos de Productos de la Competencia

Descripción

El dataset obtenido será un listado de todos los productos ofrecidos por la competencia en su sitio web con los datos considerados más relevantes para su posterior análisis, como ser: precio, categorización, calidad percibida, características de los productos, etc.

Representación grafica

Contenido

El dataset resultante contiene una línea por cada producto publicado en el sitio web en cuestión al momento de la extracción. Por cada ítem se obtendrán los siguientes datos:

- ***Title***: Nombre del producto.
- ***Price***: Precio del producto.
- ***Category_path***: categoría a la cual pertenece el producto.
- ***Rating***: el rating es el promedio de las calificaciones recibidas por parte de los compradores del producto.
- ***Qty_califications***: cantidad de calificaciones recibidas.
- ***Features_JSON_format***: se guarda en un solo campo y en formato JSON las medidas y/o características del producto. Se elige este formato por dos razones principales:
    - Cada producto puede tener diferentes características de acuerdo a su categoría que hacen que tengan n posibles features y medidas, por ej, podemos tener una lata de barniz con una única característica como ser la capacidad en litros, hasta una cama donde podemos tener el ancho, largo, altura, el tipo de madera, etc.
    - Al ser esta info extraída via web scraping y al considerarla raw data creemos que como primer stage del dato es válida conservarla de esta manera. Podrán ser parseados estos en etapas posteriores según sea necesario y según la categoría del producto.
- ***Image_Url***: url de la primera imagen del producto en la publicación.
- ***Item_Url***: url del producto dentro del sitio web.

En una 2da iteración se podría incluir la descripción de cada producto, pero dado que solo es texto en prosa describiendo de manera general (en más del 92% de los casos) el producto, no creemos que sea necesario recolectar este campo por el momento.

Agradecimientos

Los datos han sido recolectados desde el sitio web Muebles Lufe. Para ello, se ha hecho uso del lenguaje de programación Python y de técnicas de Web Scraping para extraer la información pública contenida en las páginas HTML.

Inspiración

Conocer la cantidad y características de los productos publicados en internet por una empresa competidora pueden servir para diversos motivos, entre ellos podemos mencionar:

- Conocer el Competitor Ratio versus la empresa X, en este caso Muebles Lufe, para entender el nivel de competitividad que puede tener en cantidad de listings ***totales*** o ***por categoría***. El competitor ratio lo podemos definir como:
    - Cantidad de Publicaciones de la competencia / Cantidad de publicaciones de nuestra empresa.
- Conocer el pricing por categoría de la competencia y gestionar nuevas políticas de pricing si fuera necesario
- Conocer la calidad percibida de ciertos productos por parte de los clientes para productos similares o iguales a lo que tiene nuestra empresa.

Licencia

La licencia escogida para la publicación de este conjunto de datos ha sido CC BY-SA 4.0 License. Los motivos que han llevado a la elección de esta licencia tienen que ver con la idoneidad de las cláusulas que esta presenta en relación con el trabajo realizado:

- Se debe proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado. De esta manera, se reconoce el trabajo ajeno y en qué medida se han realizado aportaciones en relación con el trabajo original.
- Se permite un uso comercial. Esto haría que incrementen las probabilidades de que una empresa utilice los datos generados y realicen trabajos de calidad que reporten cierto reconocimiento al autor original.
- Las contribuciones realizadas a posteriori sobre el trabajo publicado bajo esta licencia deberán distribuirse bajo la misma. Esto hace que el trabajo del autor original continúe distribuyéndose bajo los términos que él mismo planteó

Código fuente.

Tanto el código fuente escrito para la extracción de datos como el dataset generado pueden ser accedidos desde este repositorio en GitHub:

https://github.com/PythonGreen/web_scrapers

Dataset

El DOI generado para el dataset es:

Recursos

- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
- Masip, D. (2010). El lenguaje Python. Editorial UOC.
- Tutorial de Github https://guides.github.com/activities/hello-world.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Get DOI for a github repo: https://guides.github.com/activities/citable-code/