Crawler precios de vinos

Inspiración

Muchas bodegas y cooperativas vinícolas españolas todavía en 2022 no tienen página web o no son se han digitalizado lo suficiente para aprovechar todo el potencial que las últimas tecnologías pueden ofrecer para el crecimiento de su negocio. Esto hace que como factor subyacente, no existan datos de suficiente calidad sobre cuáles son los vinos a nivel nacional clasificados por denominación de origen.

Se desean elaborar varios crawlers para extraer diversa información acerca de los vinos para la creación de varios datasets y su posterior análisis y extracción de valor a lo largo del tiempo. Para ello, se utilizará la tecnología de web scrapping con el framework de Scrapy en Python y un API. Para ello se ha escogido una web de un supermercado que destaca por su variedad de vinos en sus pasillos, Carrefour, y el sitio web de referencia en España cuando se trata de consultar información acerca de un vino, Vivino.

En el futuro, se pretende monitorizar los precios de los vinos para aumentar la competitividad de las bodegas y obtener datos y analíticas de calidad que los gobiernos puedan utilizar para poder tomar decisiones acertadas.

Título y descripción de los datasets

Habrá dos datasets con un nombre compuesto por wines_ seguido de la fuente de datos de donde proviene. De cada fuente se pueden obtener atributos diferentes para cada vino por lo que, aun intentando normalizar lo máximo posible para que resulten similares, los datasets contendrán pequeñas diferencias. El dataset de Vivino será bastante mayor al de Carrefour siendo 2.3Mb y 180Kb sus tamaños respectivamente.

Representación gráfica


Contenido

wines_carrefour.csv

Campos extraídos:

- wine: el nombre del vino.
- winery: la bodega del vino.
- origin: la Denominación de Origen del vino.
- country: en este caso siempre será España.

- variety: el color del vino (tinto, blanco o rosado).
- price: el precio del vino en euros.
- acidity: la acidez del vino en g/l.
- alcohol_percentage: el volumen de alcohol.
- date: la fecha de extracción de los datos del crawler.
- source: en este caso siempre será carrefour.

wines_vivino.csv

Campos extraídos:

- wine: el nombre del vino.
- winery: la bodega del vino.
- origin: la Denominación de Origen del vino.
- country: en este caso siempre será España.
- variety: el color del vino (tinto, blanco o rosado).
- grape: el tipo de uva del vino.
- price: el precio del vino en euros.
- rating: la puntuación del 1 al 5 que ha recibido.
- body: el cuerpo del vino en una escala del 1 al 5.
- acidity: la acidez del vino en una escala del 1 al 5.
- description: Un pequeño texto con los aspectos más destacables sobre el vino.
- food: Una lista de comidas con las que se sugiere acompañar el vino.
- date: la fecha de extracción de los datos del crawler.
- source: en este caso siempre será vivino.

Agradecimientos

Vivino es un mercado de vinos en línea y una aplicación de vinos. Fue fundada en 2010 por Heini Zachariassen y Theis Søndergaard. En 2021, Vivino contaba con una base de datos de vinos que contenía más de 12,5 millones de vinos diferentes y tenía 50 millones de usuarios. La sede de Vivino se encuentra en San Francisco (California, Estados Unidos), pero la empresa tiene varias filiales, incluso en Dinamarca, donde se fundó la empresa.

Carrefour es una cadena multinacional de distribución de origen francés. Es considerado el primer grupo europeo, a poca distancia en ingresos netos de la compañía alemana Schwarz Gruppe (matriz de Lidl y Makro), y el tercero del sector a nivel mundial. Contiene una división especializada en vinos, Carrefour Bodega, con una amplia selección, no solo de productos, si no a demás de información acerca de los mismos, facilitando su compra a los clientes más entendidos.

Por lo que he podido observar mientras surfeaba por internet, ya se han hecho proyectos open source de web scrapping de páginas web de supermercados (en especial a nivel académico) aunque no de Carrefour Bodegas. De Vivino sí, aunque los que encontré no funcionaban y me resultó más efectivo empezar de 0 ambos scrappers.

Contexto

Cabe mencionar que no se han extraído todos los campos disponibles de las diferentes fuentes. Se ha pretendido extraer solamente los atributos que en principio pudieran resultar de más interés para un análisis posterior en el futuro.

Licencia

La licencia CC BY-NC 4.0 permite que el dataset sea compartido, sin embargo, al contrario que la CC BY-SA 4.0, no permite un uso comercial. Por este mismo motivo, a su vez, se ha descartado ofrecer una licencia de dominio público. A demás, de acuerdo con la encuesta disponible en la web de Creative Commons, esta licencia es la que más se ajusta a nuestras necesidades.