Dataset: Carreras deportivas en España para el año 2019

Eugenio Carmona Soriano - Antonio Ruiz Falco Rojas

14 de abril de 2019

Descripción

El conjunto de datos generado como parte de esta actividad práctica recoge los eventos de carreras deportivas que se realizan en España el año 2019. Las variables que se recogen en el conjunto de datos son la fecha, el nombre, el lugar, el tipo de carrera, la distancia, la página web de la carrera y si incluye categorías infantiles.

Para cada carrera, se crea un registro en el conjunto de datos que recogen los siguientes campos para todo el año 2019:

- **Fecha**: el día en el que se realiza la carrera en formato dd/mm/aaaa.
- **Carrera**: Nombre de la carrera.
- **Ciudad**: Ciudad en la que se realiza la carrera.
- **Provincia**: Provincia en la que se realiza la carrera.
- **Tipo**: Tipo de carrera. Ruta, Trail, Triatlón, Ciclismo, Duatlón, Obstáculos y Cross/Tierra.
- **Distancia**: La distancia a recorrer en la carrera:
- **Web**: Enlace a la página web de la carrera.
- **Infantil**: Si o No. Según incluya o no categorías infantiles.

Dependiendo del tipo de tratamiento que se desee hacer, los datos necesitarán pretratamiento. Por ejemplo, se puede quitar la hora del campo fecha. También se puede normalizar el campo distancia, pues algunas carreras expresan la distancia en metros y otras en kilómetros. Por su parte, los campos carrera, ciudad, provincia y tipo se encuentran normalizados. La columna infantil es una variable booleana, en la que el valor True implica que la carrera es infantil.

Imagen identificativa

Figura 1: Cartel de la media maratón de Barcelona

Contexto

El interés de la sociedad en general por el deporte y la actividad física ha crecido enormemente en los últimos años. La actividad deportiva básica es el *running* en sus diversas modalidades. Debido a ello, muchas instituciones organizan periódicamente competiciones en todos los niveles, desde carreras populares a alta competición. Sin embargo, si se desea realizar cualquier tratamiento de datos sobre las carreras que se celebran en España, se carece de la información necesaria. Para ello se ha creado el presente conjunto de datos, que incluye todas las carreras programadas para el año 2019.

El conjunto de datos se corresponde con las competiciones deportivas, más concretamente carreras deportivas, que se realizan en España el año 2019. Se incluyen varios tipos de carreras entre los que encuentran: Rutas, Trails, Triatlón, Ciclismo, Duatlón, Obstáculos y Cross/Tierra.

La página web elegida para obtener la información la ofrece a titulo informativo a todo aquel que quiera conocer las diferentes competiciones que se realizan, el lugar y la fecha. Además de ofrecer el enlace a la página web oficial de cada carrera.

Contenido

El conjunto de datos se ha construido mediante técnicas de Web Scraping rastreando la página https://www.corriendovoy.com/calendario-carreras. Los autores de la página corriendovoy recopilan información de carreras y ofrecen servicios de grabación de corredores, además de noticias, el calendario de carreras y sorteo de material deportivo.

Para la extracción de la página se ha construido un programa en Python, que debe ejecutarse en un entorno Python 3. La información rastreada está codificada en UTF-8, por lo que es necesario utilizar características propias de Python 3, que hace un tratamiento transparentes de *str's* y *unicode's.*

Precisamente, lo primero que hace el programa es comprobar la versión de Python. En caso de que no sea 3, muestra un mensaje y finaliza la ejecución. El programa también lee el contenido de robots.txt para comprobar si la página permite el rastreo de la url especificada. Por último, el programa hace uso de la librería fake-useragent para cambiar el User Agent de cada sesión HTTP, con el objetivo de evitar el bloqueo del servidor.

La tabla de carreras ocupa muchas páginas, por lo que es necesario leerlas todas. El programa va avanzando por el índice hasta llegar a la última página. Se da la circunstancia que los caracteres de avance de página son Unicode.

Agradecimientos

A los autores de la página web corriendo voy (https://www.corriendovoy.com/calendario-carreras) por poner los datos al alcance de los usuarios. No se han encontrado análisis anteriores de este tipo de datos, lo que incrementa el valor del conjunto presente.

Inspiración

No se han encontrado artículos realizando estudios sobre carreras deportivas, lo que muestra la importancia del conjunto de datos creado, que podrá utilizarse para propósitos muy diversos: desde realizar mapas visuales con la localización de carreras que cumplan unos determinados requisitos a proyectos clásicos de minería de datos, como agrupación, clustering y detección de reglas.

La versatilidad de los posibles tratamientos los hace enormemente útiles a diversos colectivos. Desde fabricantes y comercializadores de material deportivo, para la toma de decisiones de publicitar/vender productos, ayuntamientos o entidades que quieren organizar eventos y quieren analizar posibles competencias, entrenadores deportivos, etc.

Además, el conjunto de datos puede utilizarse como fuente de futuros conjuntos de datos de resultados. Es decir, una vez inventariadas las carreras existentes puede intentarse rastrear los resultados. Esto permitiría realizar nuevos proyectos de analítica de datos sobre los resultados, detectando patrones en los mismos por estación del año, ubicación, etc. De esta forma, podría establecerse un calendario de carreras óptimo para cada corredor.

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 y dataset

Tanto el código fuente escrito para la extracción de datos como el dataset generado pueden ser accedidos a través de este enlace: https://github.com/ujinuoc/runnings-ws.

Recursos

1. Masip, D. El lenguaje Python. Editorial UOC.
2. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data

Contribuciones

Contribuciones	Firma
Investigación previa	ECS – ARFR
Redacción de las respuestas	ECS – ARFR
Desarrollo código	ECS – ARFR