Dataset: Carreras populares en Galicia

Luis Antón López Sobrado

14 de Abril de 2019

Descripción

Los datos extraídos de la página web de la federación gallega de atletismo, contienen información de las carreras populares que se van a celebrar al año en cada lugar de Galicia. Las variables que se recogen son el nombre y tipo de la carrera, la fecha, y el lugar donde se celebrará. Este DataSet con los datos de las carreras de atletismo que se disputarán en Galicia durante el año 2019, se ha creado con el objetivo de analizar cuales son los meses en los que más carreras hay, que lugar ofrece un mayor número de carreras o de cuales hay más tipo. Está información salvo cancelación a última hora o error, los datos de cada carrera y el número de carreras no varía.

Imagen identificativa

Figura 1: V carrera nocturna en Santiago de Compostela

Contexto

Ante el boom que vivimos en estos momentos con la moda del *running* el número de corredores en Galicia aumenta cada año, con lo que cada vez más usuarios solicitan información de donde se realizaran las carreras más próximas, así como del tipo de carreras que se trata. Para obtener la información de las carreras populares se dispone de una página web https://atletismo.gal/competicions donde se muestra un listado con todas las carreras de Galicia. Esta página muestra la información de cada carrera, con la fecha, el nombre de la carrera, el lugar

donde se realizará y el tipo de carrera del que se trata. Para crear el conjunto de datos se ha cogido toda la información de las carreras de Galicia durante el año 2019.

Contenido

El contenido del dataset generado con los datos de las carreras fue extraído al realizar web scraping sobre la página de la federación gallega de atletismo mediante el uso de la librería BeautifulSoup. Antes de extraer la información de está forma se comprobó que no existiera una API que permitiera recuperar la información requerida sin limitaciones, se trató de no saturar el servidor para evitar ser bloqueados y se comprobó el fichero robots.txt para comprobar las restricciones de la página.
El programa para obtener la información de la web se implementó en Python, se instalaron las librerías correspondientes y se ejecutó mediante el IDE Spyder. Para ello fue necesario examinar la estructura de la página así como las etiquetas que contenían la información requerida, de tal forma que para cada registro del conjunto de datos, es decir, para cada carrera se recogieron las siguientes características:

- Fecha: día el que se va a realizar la carrera en el formato dd/mm/aaa.
- Nombre: edición y nombre de la carrera popular.
- Lugar: lugar donde se realizará la carrera.
- Tipo: los tipos de carreras que se muestran son:
  - Ruta → Ruta
  - Pista AL → Pista Aire Libre
  - PC → Pista Cubierta
  - Marcha → Marcha
  - CT' → Campo a Través
  - Montaña → Montaña
  - Trail → Trail

Agradecimientos

Los datos han sido recolectados desde la web de la Federación Gallega de Atletismo. Para extraer la información alojada en las páginas HTML, se ha hecho uso del lenguaje de programación Python mediante las librerías Requests y BeautifulSoup.

No se han encontrado ni investigación ni análisis previo de esta información. Los ayuntamientos o asociaciones que organizan las carreras informan a páginas como la de federación para publicitarse, pero no se ha encontrado hasta el momento ninguna investigación.

Inspiración

El conjunto de datos podría utilizarse para obtener la siguiente información. ¿Cuantas carreras se hacen al mes en Galicia? ¿Cuantas carreras realiza cada ciudad o cada pueblo? ¿De que tipo de carrera? ¿Cuantas realiza cada provincia? Estos son de gran utilidad para el ámbito periodístico y para todos los corredores.

Licencia

La licencia escogida para la publicación de este conjunto de datos ha sido **Released Under CCo: Licencia de Dominio Publico**. He seleccionado esta licencia porque así se puede copiar, modificar, distribuir los datos y hacer comunicación pública.

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/luilop/Web-scraping

Recursos

- David Masip Rodó (2019).El lenguaje Python. Editorial UOC
- Laia Subirats Maté, Mireia Calvo González (2019).Web scraping. Editorial UOC
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.