Práctica 1: Web scraping

Contexto

Usualmente, cuando vamos a algún tipo de evento cultural, en los días previos al mismo verificamos las condiciones meteorológicas previstas en diferentes webs o en app disponibles en nuestro smartphone. Esta información nos permite resolver preguntas del tipo: ¿qué medio de transporte uso? ¿qué tipo de ropa he de llevar?

Además, para los que vivimos en regiones con climas muy cambiantes como en las Islas Canarias, esta información se vuelve más necesaria. Es por ello, que consideramos de gran utilidad tener toda esta información en el mismo lugar.

Las webs usadas permiten construir este data set, centrándonos en los eventos que se anuncian en la web La Agenda y en las previsiones meteorológicas aportadas por la DarkSky. También se han obtenido datos a Google Maps para relacionar los lugares de La Agenda con los datos de DarkSky.

Título para el data set

Datos de eventos y condiciones meteorológicas esperadas

Descripción del data set

Para dar respuesta a esta necesidad, el data set está compuesto de las variables principales referentes al evento: fecha, descripción, hora, municipio donde se celebra,...; y datos referentes a las condiciones meteorológicas esperadas: temperatura, estado del cielo, probabilidad de lluvia.

Representación gráfica

Imagen real de uno de los eventos(TENERIFEANDO POR EL MUNDO | CICLO VIAJEROS)

Imagen real de uno de los eventos (TENERIFEANDO POR EL MUNDO | CICLO VIAJEROS)

Imagen de cómo se vería el evento con los datos de predicción meteorológica

Imagen de cómo se vería el evento con los datos de predicción meteorológica (TENERIFEANDO POR EL MUNDO | CICLO VIAJEROS)

Contenido

Para cada evento, el cual se corresponde con un registro en el conjunto de datos, se recogen las siguientes variables:

- **title**: título del evento. Str.
- **date**: día en el que se celebra el evento, en el formato dd/mm/aaa.
- **location**: lugar donde se celebra el evento. Str.
- **description**: descripción del evento. Str.
- **catrgory**: categoría en la que se engloba el evento. Str.
- **url**: url en la que se publica el evento. Str.
- **coordenadasLocalidad**: coordenadas centrales de la localidad donde se celebra el evento. Int.
- **estadoCielo**: porcentaje de cielo cubierto por nubes. Int.
- **probPrecipitacion**: recoge la probabilidad de que produzcan precipitaciones. Int.
- **sensTermMax**: recoge la sensación térmica máxima el día del evento. Int.
- **sensTermMin**: recoge la sensación térmica mínima el día del evento. Int.
- **temperaturaMax**: recoge la temperatura máxima el día del evento. Int.
- **temperaturaMin**: recoge la temperatura mínima el día del evento. Int.

Los datos del evento (**title, date, location, catrgory, url**) se obtienen con los paquetes *BeautifulSoup* y *Request*. Los datos de predicción meteorológica (**estadoCielo, probPrecipitacion, sensTermMax, sensTermMin, temperaturaMax, temperaturaMin**) se obtienen a través de la API de la **DarkSky**, ayudados de los paquetes *darksky.api* y *darksky.types*.

Además de los paquetes anteriores, se utilizan otras librerías habituales como *Pandas* o *Datetime*, entre otras.

En cuanto al periodo de tiempo, sólo se toman datos del 12 al 13 de noviembre. Los datos a recolectar no puede superar a una semana vista del momento en el que se ejecute, puesto que es el alcance de predicción disponible.

Agredecimientos

La información ha sido recopilada de diferentes fuentes:

La Agenda

Gracias a esta web hemos podido obtener los datos referidos a los eventos, utilizando técnicas de *Web Scraping* para extraer la información alojada en las páginas HTML.

DarkSky y Google Maps Platform

La disponibilidad de esta web nos ha posibilitado, junto con la API de Google, extraer predicciones meteorológicas en función de las coordenadas centrales de la localidad en la que se celebra el evento. Tenemos que tener en cuenta que, al igual que la API de Google, necesita un usuario y su uso tiene cargos asociados, si bien, esto se aplican a partir de un número bastante elevado de consultas.

AEMET

A pesar de que finalmente se toma la decisión de no usar esta API para obtener datos meteorológicos, debido a que se dieron problemas para relacionar las localizaciones de los eventos con uno de los parámetros de entrada de la API (el código del municipio), queremos agradecer la posibilidad que brinda esta fuente de datos abiertos. Un futuro desarrollo, podría incluir esta API, dada su gratuidad y el carácter oficial.

Otros

También han sido de vital ayuda otras webs y repositorios para poder realizar este proyecto. A modo enunciativo: * Repositorio DarkSky * DS Códigos Postales INE * https://python-para-impacientes.blogspot.com/2014/02/operaciones-con-fechas-y-horas.html

Inspiración

Como comentado en el contexto, la intención es poder cubrir la necesidad que tienen los espectadores de los eventos publicados en la web La Agenda de conocer con anticipación las condiciones meteorológicas esperadas en la fecha de dicho evento.

Sin embargo, esta no es la única utilidad que se le podría dar a este proyecto. Tanto los cuerpos de seguridad como los medios de transporte público podrían estar interesados en esta información. Por ejemplo, si se trata de un evento de exterior y se esperan condiciones de tormenta, esto incrementa la probabilidad de que se produzcan inundaciones, con lo que tendrán que prestar más atención a este evento, debido a la concentración de público.

Otra aplicabilidad de este data set sería poder hacer un estudio de la relación que existe entre la venta de entradas para un evento y las condiciones meteorológicas previstas. Este análisis permitiría estimar a los organizadores del evento la afluencia de público y, en consecuencia, dimensionar los servicios acordes con el público esperado.

Se desconoce si han habido iniciativas similares a esta, pero si son numerosos los estudios en los que se relaciona una variable endógena con variables exógenas, como las condiciones meteorológicas.

Licencia

Para ver la licencia de uso más adecuada, hemos de analizar nuestras 3 fuentes de datos. En cuanto a **Lagenda**, no nos ha sido posible ver su política de privacidad, pero es evidente el carácter divulgativo de la misma. En cuanto a la API de **Google**, sus términos de privacidad están muy desarrollados. Por tanto, cualquier uso de los datos de esta API debe hacerse siempre indicando que los datos provienen de esta organización. Para los datos de las predicciones meteorológicas, **DarkSky** permite, explícitamente, la difusión del contenido de la misma siempre que se les haga mención.

Por tanto, parece que lo adecuado es tener una licencia tipo **Released Under CC BY-SA 4.0 License**, puesto que vemos que podemos reproducir los datos de las bases de datos originales y queremos que, si se considerase de interés, pudiera seguir mejorándose.

Código y Dataset

El código se puede encontrar en el siguiente repositorio de Github PRAC-1

En cuanto al data set, está incluido en el repositorio con el nombre Datoslagenda.csv

Contribuciones

Investigación previa: Luis Cobiella y Jonay Velázquez

Redacción de las respuestas: Luis Cobiella y Jonay Velázquez

Desarrollo código: Luis Cobiella y Jonay Velázquez