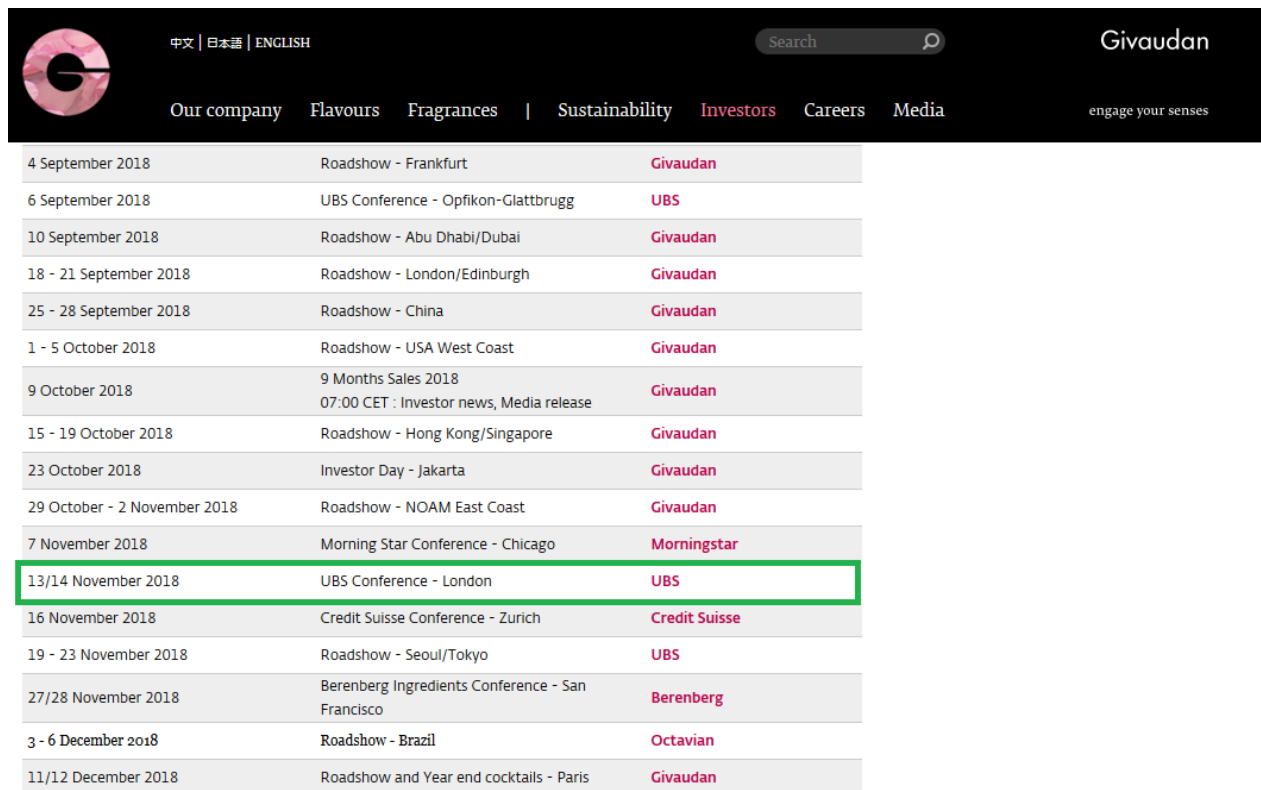


# Enunciado práctica WebMining

El objetivo de esta práctica es utilizar técnicas de *Web mining* para extraer información de una página web. En concreto, se desean encontrar todos los futuros eventos relevantes, y para cada evento identificar el lugar (City), la fecha (Date) y el patrocinador (Sponsor).

Por ejemplo, para la página web con código 1471, el *output* del programa debería incluir información sobre el congreso que tendrá lugar en Londres del 13 a 14 de noviembre de 2018, patrocinado por UBS, tal y como se muestra a continuación.

(Url: <https://www.givaudan.com/investors/shareholder-information/investor-calendar>)



4 September 2018	Roadshow - Frankfurt	Givaudan
6 September 2018	UBS Conference - Opfikon-Glattbrugg	UBS
10 September 2018	Roadshow - Abu Dhabi/Dubai	Givaudan
18 - 21 September 2018	Roadshow - London/Edinburgh	Givaudan
25 - 28 September 2018	Roadshow - China	Givaudan
1 - 5 October 2018	Roadshow - USA West Coast	Givaudan
9 October 2018	9 Months Sales 2018 07:00 CET : Investor news, Media release	Givaudan
15 - 19 October 2018	Roadshow - Hong Kong/Singapore	Givaudan
23 October 2018	Investor Day - Jakarta	Givaudan
29 October - 2 November 2018	Roadshow - NOAM East Coast	Givaudan
7 November 2018	Morning Star Conference - Chicago	Morningstar
13/14 November 2018	UBS Conference - London	UBS
16 November 2018	Credit Suisse Conference - Zurich	Credit Suisse
19 - 23 November 2018	Roadshow - Seoul/Tokyo	UBS
27/28 November 2018	Berenberg Ingredients Conference - San Francisco	Berenberg
3 - 6 December 2018	Roadshow - Brazil	Octavian
11/12 December 2018	Roadshow and Year end cocktails - Paris	Givaudan

...

Company: 1471

Type: 2

Date: 2018-11-13/14

Title: UBS Conference

City: 7

Sponsor: ff808081227ad9f7012298a8eb40031a

El programa dispondrá de un fichero con ciudades y sus códigos (`city.csv`) y otro fichero con patrocinadores y sus códigos (`sponsor.csv`). Si la ciudad o el patrocinador no están en la lista pero la página contiene esta información, se debe incluir el nombre en el apartado correspondiente (City o Sponsor). En caso contrario, se debería indicar “Not Available”. En el campo de título (Title) se debe incluir cualquier información que se considere oportuno (es un campo libre).

Los tipos de eventos que se quieren extraer son los siguientes (con su código correspondiente):

Código	Tipo de evento
1	Roadshow
2	Conference
3	Investor day(s)

**Nota importante:** Los eventos de tipo “Conference call” o “Press conference” no se deberían extraer.

Se recomienda el uso de los siguientes paquetes:

- Para leer páginas html, *urllib.request*  
<https://docs.python.org/3/library/urllib.request.html>
- Para navegar y buscar información, *BeautifulSoup* con el *lxml markup*  
<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- Para encontrar patrones, *re*  
<https://docs.python.org/3/library/re.html>
- Para el formato de las fechas, *dateutil.parser*  
<https://dateutil.readthedocs.io/en/stable/parser.html>