PLEASE READ THIS DOCUMENT CAREFULLY. IF YOU ASK QUESTIONS ANSWERED IN THIS DOCUMENT, YOU WILL NOT BE CONTACTED.

Project description

This is a **data scraping** project requested on fiverr . You are expected to use a toolchain with libraries such as Selenium, Scrapy, BeautifulSoup, python etc. If you do not know how to use such tools, please do not make an offer.

All deadlines agreed upon must be met, entries (row data) must adhere to the requirements described further down in this document. Under no circumstances should you change the supplied files or instructions. If you fail to deliver as described here OR if you fail to deliver in the format described OR if you do not deliver a complete dataset, the payment will be reclaimed — this is non-negotiable.

Prerequisites

Before you send an offer, please keep in mind the following;

- Payment is subject to your analysis of the website in question and the deal made before the project is accepted.
- You need to have a toolchain already. You will not be paid for any setup.
- It is expected that you already have written tools to manipulate, process and recalculate datatypes required. Examples are;
  - Dates & times must be converted to the required format.
  - Addresses must be converted to the required format.
  - Latitudes and longitudes must be converted to the required format and length.
  - Phone numbers must be converted to the required format.
  - Email addresses must be converted to the requried format.
  - Characters such as Æ æ ø Ø å Å must not be corrupted.
- You need to write a script to populate the required data in the same way as the sample file, **in the order it is presented in the file**.
- Keep in mind that several people will be paid to do this project. And you will be measured against them. If you deliver faulty data, bogus or otherwise non-usable data you will be reported. If you use an unfair amount of time, deliver considerable lower quality or number of entries than others, you run the risk of not getting paid.

THERE IS NO EXCUSE IF YOU DIDN'T UNDERSTAND THE PROJECT DESCRIPTION.
Ask if something is unclear.

Requirements

This section explains the requirements for the dataset. It covers three things; **data to be collected**, **where to collect the data** and **how to deliver the data collected**.

Data to be collected

This section explains what data should be collected, and requirements for this data.

Below is a complete list of requirements

- event data entries must be in Norwegian.
- if the request page specifies cities, please collect data from the cities described below.
- all entries must be unique, no duplicate events are allowed.
- events must be scraped from 1st of march 2018 forward as long as possible.
- if an event lacks an address, latitude and longitude **must** be provided to the nearest place possible.
- the data must be delivered as CSV *sample is provided.*
- data must be as described in sample, please read additional information on each column below in descriptions.
- give them file names `events_requiredFields_dataset.csv`
- or `events_requiredFieldsNotCompleted_dataset.csv` if some fields are missing.
- files must have the header row as in the sample file.
- leave non-retrievable data fields empty.
- if ticket information is available (a link to a purchase page) it should be provided in ticket_link column.

Cities to to use when searching for event data on sites where this is required.

If the request on **fiverr** specifies that you need to search using city keywords, please scrape using the following list of cities:

Åkrehamn	Ålesund	Alta	Åndalsnes	Arendal	Askim
Bergen	Bodø	Brekstad	Brevik	Brønnøysund	Brumunddal
Bryne	Drammen	Drøbak	Egersund	Elverum	Fagernes
Farsund	Fauske	Finnsnes	Flekkefjord	Florø	Førde
Fosnavåg	Fredrikstad	Gjøvik	Grimstad	Halden	Hamar
Hammerfest	Harstad	Haugesund	Hokksund	Holmestrand	Hønefoss
Honningsvåg	Horten	Jessheim	Jørpeland	Kirkenes	Kolvereid
Kongsberg	Kongsvinger	Kopervik	Kragerø	Kristiansand	Kristiansund
Langesund	Larvik	Leirvik	Leknes	Levanger	Lillehammer
Lillesand	Lillestrøm	Lyngdal	Måløy	Mandal	Mo i Rana
Molde	Mosjøen	Moss	Mysen	Namsos	Narvik
Notodden	Odda	Orkanger	Oslo	Otta	Porsgrunn
Risør	Rjukan	Sandefjord	Sandnes	Sandnessjøen	Sandvika
Sarpsborg	Sauda	Ski	Skien	Skudeneshavn	Sortland
Stathelle	Stavanger	Stavern	Steinkjer	Stjørdalshalsen	Tananger
Tønsberg	Tromsø	Trondheim	Tvedestrand	Ulsteinvik	Vadsø
Vardø	Verdalsøra	Vinstra			

Required fields

These are the required data fields to be populated in the file **events_requiredFields_dataset.csv** if any of these fields are empty in a row from your scripts output, those rows should be moved to the file **events_requiredFieldsNotCompleted_dataset.csv** instead.

NOTE: these fields must have its data formated as displayed in the **Example** column show in the tables below. No other way.

Fieldname	Description	Example
event_name	the name of the event	NEON Summerfest
event_desc	a description of the event, usually the explanation stuff on the event page.	The fest is about [. . .]
start_date	what date it starts as `DD-MM-YYYY`.	22-12-2018
start_time	time it starts as `HH:MM`.	12:15
end_date	end date. On festivals and stuff last for days, if single day, just put the same as start_date.	22-12-2018
end_time	when it ends, if this is festival, just put last day time.	16:15
host_name	the name of the company or organization hosting the event.	Festivals Corp.
location_name	The name of the location.	Rock stadium
location_address	the address of the venue or place the event will take place as follows: Streetname, streetnumber, zipcode, city.	Merlon Street 21, 0101, Oslo
host_contact	email of host.	summerfest@festivals.no
city	city by it self for indexing.	Oslo
GEODATA: Fetched from maps.google.com using API by address lookup		
lat	latitude of location.	59.0555477
lng	longitude of location.	11.0555477

Additional fields

These datafields are not required, but should be collected where possible.

Fieldname	Description	Example
ticket_link	Link to the page where tickets can be purchased, such as ticketmaster link, ebillett link etc.	linkToTicketSite.com/event/29292929
facebookhost_id	link to facebook page of host	/nameOfPageOrGroup
website	Link to the hosts website or provided website for event (not Facebook or other social networks as below)	example.com
facebook_id	Link to facebook event page.	/events/6546545351351351231321
Other social networks: Least important fields		
linkedin_id	Link to host linkedIn page	/linkedInName
twitter_id	link to host/event on twitter.	/somename
instagram_id	link to host/event on this site.	/somename
pinterest_id	link to host/event on this site.	/somename
google_id	link to host/event on this site.	/somename
skype_id	link to host/event on this site.	/somename
youtube_id	link to host/event on this site.	/somename
discord_id	link to host/event on this site.	/somename
snapchat_id	link to host/event on this site.	/somename
ello_id	link to host/event on this site.	/somename
periscope_id	link to host/event on this site.	/somename
vimeo_id	link to host/event on this site.	/somename
meerkat_id	link to host/event on this site.	/somename
vine_id	link to host/event on this site.	/somename
flickr_id	link to host/event on this site.	/somename
tumblr_id	link to host/event on this site.	/somename
medium_id	link to host/event on this site.	/somename
tripadvisor_id	link to host/event on this site.	/somename
dribble_id	link to host/event on this site.	/somename
whatsapp_id	link to host/event on this site.	/somename

Where to collect the data

The data is to be collected from the website(s) as described in the request on **fiverr**. Under no circumstance should you include data from other requests of the same type.

How to deliver the data collected

When you are ready to deliver the file(s), please make sure the files have the header (top most row) as shown in the sample.csv file provided. This file should also give you an idea of the data that is asked for. **DO NOT MAKE CHANGES TO THE STRUCTURE OF THE SAMPLE FILE**

The csv header should look like this:

```
ai_id,category,event_name,desc,start_date,start_time,end_date,end_time,host_name,
location_name,location_address,location_desc,host_contact,lat,lng,city,ticket_link,
facebookhost_id,website,facebook_id,linkedin_id,twitter_id,instagram_id,pinterest_id,
google_id,skype_id,youtube_id,discord_id,snapchat_id,ello_id,periscope_id,vimeo_id,
meerkat_id,vine_id,flickr_id,tumblr_id,medium_id,tripadvisor_id,dribble_id,whatsapp_id
```

(**NOTE:** The fields ai_id, category and location_desc should be left empty):