talent.com

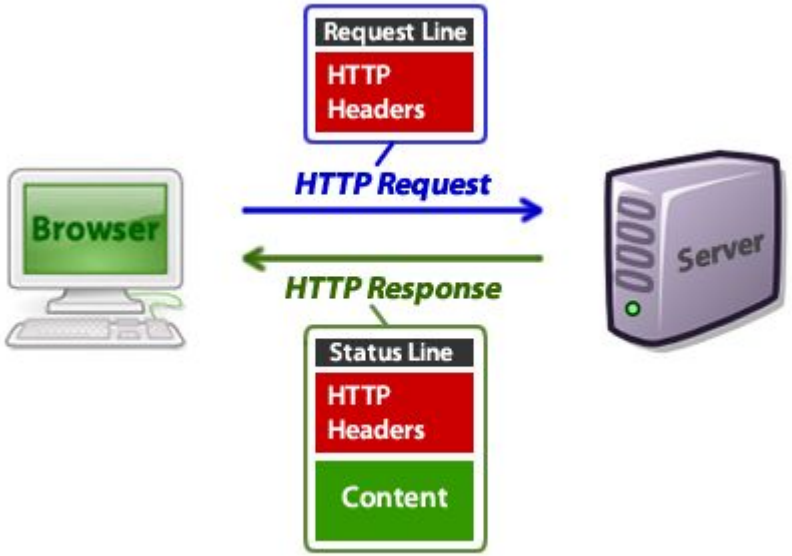
Index Training – Caso Json

Peticiones

HTTP= Protocolo de Transferencia de Hipertexto

Con este protocolo, la carga de una página web se basa en peticiones HTTP que envía el navegador al servidor y en respuestas HTTP que envía el servidor al navegador.

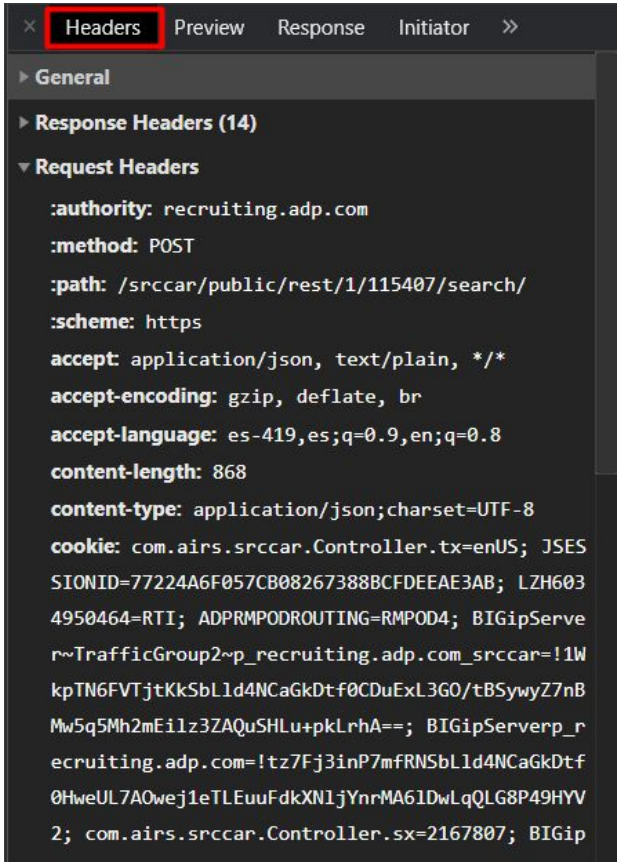
Al cargar una página web se producen decenas de peticiones y respuestas HTTP cada una de las cuales lleva cabeceras (headers).



¿Qué son las cabeceras HTTP?

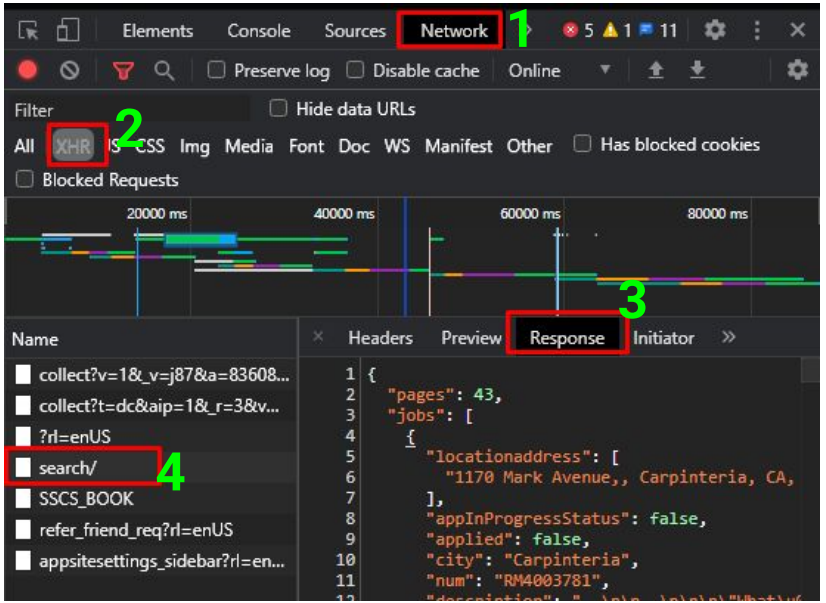
Las cabeceras = información necesaria para la comunicación

Ejemplo: tipo de navegador que realiza la petición, dirección de la página solicitada, juego de caracteres utilizado, etc.



Indexación Paso a Paso - ¿Dónde está el Json?

- Primero encuentra el jobsite que deseas indexar
- Cuando ya veamos los jobs **abrimos el inspector**
- Seleccionar la **pestaña network** (1)
- Seleccionar como interfaz la opción de filtrar por **XHR** (XMLHttpRequest) (2)
- Ahora selecciona la **pestaña Response** en el lado derecho (3)
- **Revisa en la lista** cada respuesta **buscando el json** con los jobs (4)



Indexación Paso a Paso

- Selecciona todo el json y copialo
- Para estudiar el json abre json parser online y pega el contenido.

En este momento tienes toda la información que necesitas extraer para la indexación.

y ahora, ¿desde el spider cómo haces la petición para obtener la información?

Información en Headers - Request Payload: Data

1- Necesitas copiar la data: información que se encuentra en el último header, el **Request Payload**. Importante: **debes hacer click en view source antes de copiar.**


```
//paso 1: La data: buscar en los headers request Payload (el último header)
var data ={"filters":[{"name":"country","label":"Country"},{"name":"state","label":"State/Province"},{"name":"city","label":"Town/City"},
```

Dentro de la data busca el contador de la s páginas y el valor lo debes sustituir por el nombre de la variable de tu contador de páginas:

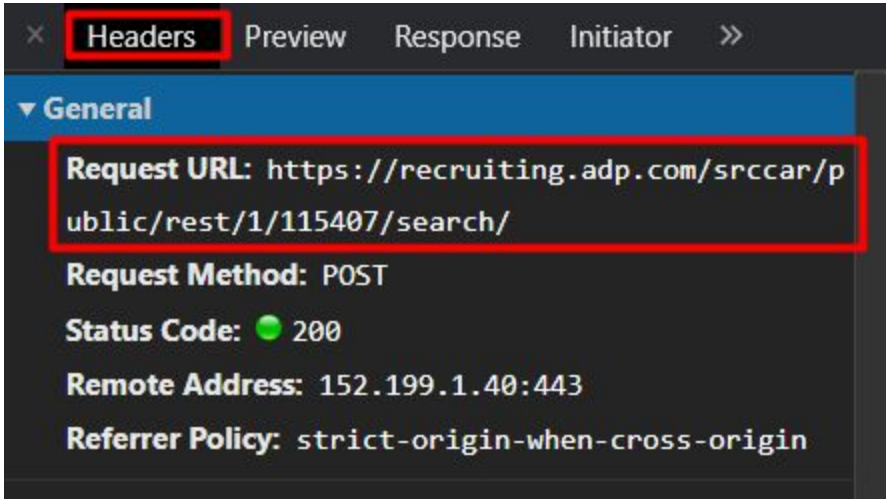
```
"pagefilter":{"page":cont},
```

A veces, la paginación esta en el url y no en la data.

Nota: hay casos donde no existe esta data, así que ignoras el paso.

Información en Headers - URL

2- Necesitas el url :

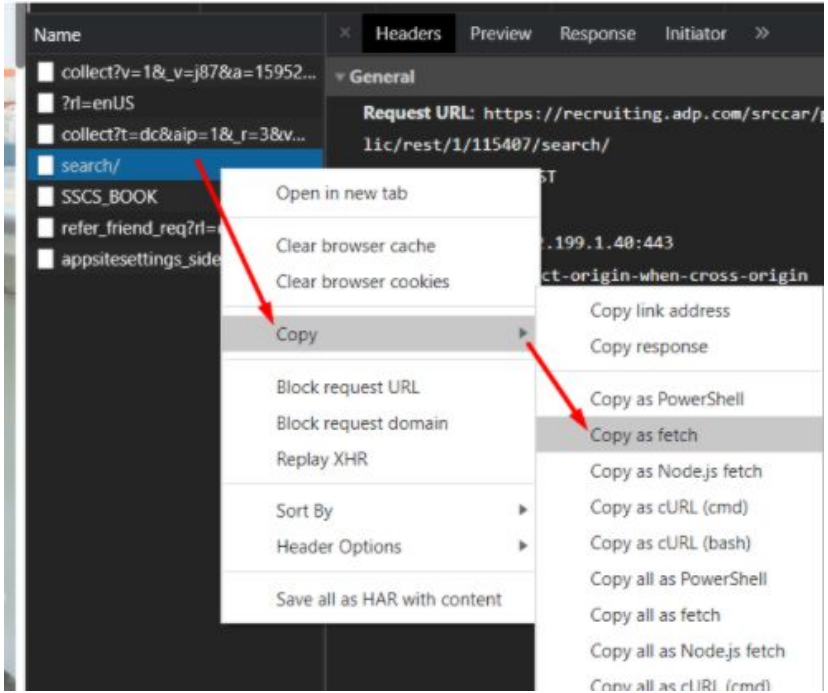


Lo debes copiar y pegar en el **paso 2 indicado en el template** (únicamente el url sin copiar las palabras “Request URL:”)

```
$.ajax({
    url: 'https://recruiting.adp.com/srccar/public/rest/1/115407/search/',  // paso 2) url
```

Información en Headers

3- Ahora vamos con la información del header **Request Headers**, para copiarla sigue los siguientes pasos haciendo click derecho en la petición que contiene los jobs



Información en Headers

4- En cualquier editor pega la información obtenida y solo nos quedaremos con lo que contiene el rectángulo verde y todo lo demás lo eliminaremos:

```
fetch("https://recruiting.adp.com/srccar/public/rest/1/115407/search/", {
  "headers": {
    "accept": "application/json, text/plain, */*",
    "accept-language": "es-419,es;q=0.9,en;q=0.8",
    "content-type": "application/json;charset=UTF-8",
    "sec-ch-ua": "\"Google Chrome\";v=\"87\", \" Not;A Brand\";v=\"99\", \"Chromium\";v=\"87\"",
    "sec-ch-ua-mobile": "?0",
    "sec-fetch-dest": "empty",
    "sec-fetch-mode": "cors",
    "sec-fetch-site": "same-origin"
  },
  "referrer": "https://recruiting.adp.com/srccar/public/nghome.guid?c=2167807&d=ExternalCareerSite",
  "referrerPolicy": "strict-origin-when-cross-origin",
  "body": "{\"filters\":[{\"name\":\"country\",\"label\":\"Country\"},{\"name\":\"state\",\"label\":\"State/Province\"},{\"name\":\"city\",\"label\":\"Town/City\"},{\
  "method": "POST",
  "mode": "cors",
  "credentials": "include"
});
```

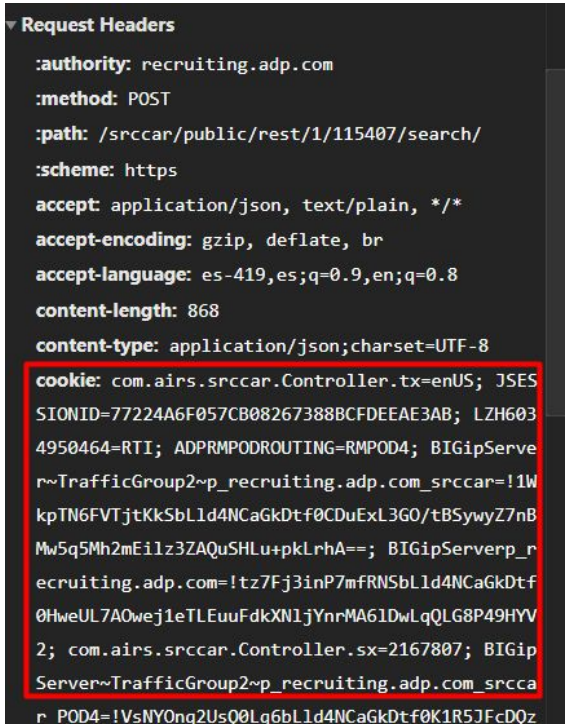
Configuración de Headers

5- Ahora pega la información en el template:

```
headers: { // paso 3) headers
  "accept": "application/json, text/plain, */*",
  "accept-language": "es-419,es;q=0.9,en;q=0.8",
  "content-type": "application/json;charset=UTF-8",
  "sec-ch-ua": "\"Google Chrome\";v=\"87\", \" Not;A Brand\";v=\"99\", \"Chromium\";v=\"87\"",
  "sec-ch-ua-mobile": "?0",
  "sec-fetch-dest": "empty",
  "sec-fetch-mode": "cors",
  "sec-fetch-site": "same-origin"
```

Cookies

6- Nota importante: con el método de copiar como fetch no se copian automáticamente las cookies así que debemos agregarlas de forma manual a nuestro header en el spider.




```
headers: { // paso 3) headers
  "accept": "application/json, text/plain, */*",
  "accept-language": "es-419,es;q=0.9,en;q=0.8",
  "content-type": "application/json;charset=UTF-8",
  "sec-ch-ua": "\"Google Chrome\";v=\"87\", \" Not;A Brand\";v=\"99\", \"Chromium\";v=\"87\"",
  "sec-ch-ua-mobile": "?0",
  "sec-fetch-dest": "empty",
  "sec-fetch-mode": "cors",
  "sec-fetch-site": "same-origin",
  "cookie": 'com.airs.srccar.Controller.tx=enUS; JSESSIONID=77224A6F057CB08267388BCFDEEAE3AB; LZH6034950464=RTI; ADPRMPODROUTING=
```

Method

7- Ahora en el header Request Headers veremos si el type es GET ó POST para configurar la información en nuestro template

En headers:

En el spider:

```
type: 'POST',    // 4) tipo
dataType: "json", // 5) data que retorna
```

Nota: la respuesta no solo puede ser un json, en ocasiones puede ser html

Ruta de los jobs

8- En el spider debes especificar la ruta de los empleos, esto lo podras hacer de manera correcta observando el json y preguntando: cuál es el tag que contiene todos los jobs?


```
{
  "pages":44,
  "jobs":[
    {
      "locationaddress":[
        "5050 East Holmes Road,Suite 105, Memphis, TN, US
        38118"
      ],
      "appInProgressStatus":false,
      "applied":false,
      "city":"Memphis",
      "num":"RM4003560",
      "description":"\"What\u0027s it like to work at
      Agilent in Manufacturing? Watch the video\"
      \n\nAgilent inspires and supports discoveries that
      advance the quality of life. We provide life science,
      diagnostic, and applied market laboratories worldwide
      with instruments, services, c",
      "type":"non-evergreen",
```


En el spider:

```
json = result.jobs;  //6) ruta de los trabajos
msg(json.length);
```

Extract

En el extract debemos obtener toda la información disponible en el json.

Como se observa en el json iremos iterando con un ciclo for en todos los objetos del json.

Para acceder a cada uno usaremos la notación json[i].nombreDelTag

```
for (var i = 0; i < json.length; i++) {
    var job = {};

    job.title = json[i].ptitle;
    job.location = json[i].location;
```

Paginación

Una forma de paginar distinta a las que ya conoces es mediante una estructura de control do-while en el extract.

En el template, tanto el **do** como el **while** están comentados así que cuando llegue el momento de paginar debes eliminar el comentario en ambas líneas.

En el template se puede ver la siguiente condición de parada:

```
} while (json.length > 0);
```

La cual indica que paginara mientras el json tenga al menos un objeto.

Pero las opciones para paginar son muchas, puedes validar si json tiene el dato de cantidad de jobs totales o páginas totales para ir comparando con un contador que tengas con este propósito.

Job description

Escenario 1:

Si extraes el job.url de cada job puedes configurar en el info el GET JOBDATA en YES y sigues el método tradicional de extraer la descripción en su respectivo step de boo3.1

Escenario 2:

Si el json cuenta con la descripción completa de cada job puedes obtenerla desde el extract utilizando el tag que contenga dicha descripción. Recuerda que en el GET JOBDATA debes elegir la opción NO.

Configuración del spider

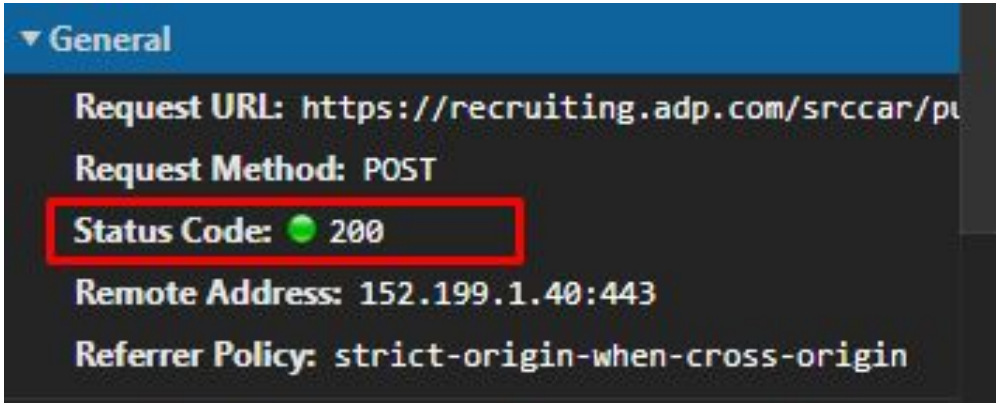
```
{
  "options": {
    "inactivateJQuery": false,
    "ignoreLoadErrors": false,
    "waitForResources":  true,
    "waitForPageLoadEvent": true
  },
  "noimage": true,
  "skipResources": false,
  "noUnnecessaryResources": false
}
```

Códigos de estado de respuesta HTTP

Son códigos que indican si se ha completado satisfactoriamente una solicitud HTTP específica.

Las respuestas se agrupan en cinco clases:

1. Respuestas informativas (100–199),
2. Respuestas satisfactorias (200–299),
3. Redirecciones (300–399),
4. Errores de los clientes (400–499),
5. y errores de los servidores (500–599).



Para mayor información visite este link

Casos para practicar

- https://recruiting2.ultipro.com/BLH1000/JobBoard/95c669d4-039f-422b-b64d-fc6f62387957/?q=&o=postedDateDesc
- https://recruiting.adp.com/srccar/public/nghome.guid?c=2167807&d=ExternalCareerSite