MAESTRÍA EN CIENCIA DE DATOS

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Práctica 1: Web scraping

Cristhyan Leonardo Naranjo Puertas

Profesor
Xavier Vivancos Garcia

9 de abril de 2020

ÍNDICE

1. MARCO DE TRABAJO

La siguiente es una descripcion del trabajo realizado y su justificacion

1. **Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.**

   Solucion. El mercado laboral es un indicador muy importante en el estado económico de un país, por lo que con este proyecto se pretende obtener un registro administrativo a partir de las ofertas de empleo ofrecidas en la página "elempleo.com" en donde al realizar una toma continua de muestras se puede generar una base longitudinal que permita ver el ciclo de vida de una oferta laboral, los intervalos de salario con mayor oferta, los puestos laborales con mayor y menor oferta entre otros datos de interés para investigación. Además estamos frente a un momento histórico con las cuarentenas causadas por la pandemia del coronavirus que seguramente afectara la situación laboral de muchas personas, por lo que esta base de datos permitirá hacer un seguimiento a la evolución del mercado laboral durante y después de las cuarentenas, lo que permitirá realizar políticas laborales que conduzcan a un mejor manejo en situaciones futuras.

2. **Definir un título para el dataset. Elegir un título que sea descriptivo.**

   Recolección de información actualizada del mercado laboral usando web scripting sobre la página "el empleo.com "

3. **Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).**

Nombre	Tipo	Descripción	Ejemplo
Empleo	Character	Nombre del empleo ofrecido	Gerente general
Sueldo	Character	Intervalo de sueldo ofrecido	18 a 21 millones
Empresa	Character	Organización que ofrece el empleo	Empresa confidencial
Ciudad	Character	Ciudad en donde se ofrece el empleo	Cali
Vacante	Numeric	Cantidad de puestos ofrecidos para ese empleo	1
Fecha	Character	Fecha de publicación de la oferta	Publicado 16 Mar 2020
Experiencia	Character	Cantidad de años mínimos para poder aplicar a la oferta	10 años de experiencia
Contrato	Character	Tipo de contrato ofrecido	Contrato indefinido
Área	Character	segmento económico donde se ubica la empresa ofertante	Salud
Sub área	Character	sub area del segmento económico donde se ubica la empresa	Medicina
Nivel	Character	Nivel del cargo	Gerente
Sector	Character	Sección o departamento dentro de la empresa	Químicos
Nivel educativo	Character	Nivel educativo mínimo para poder aplicar	Especialización
Descripción	Character	requisitos necesarios para poder aplicar al empleo	Importante multinacional alemána requiere...
Serial	Numeric	Serial de la oferta dentro de la pagina	1884488870
Link	Character	Dirección web de la oferta	*https* : *//www.elemp*...

4. **Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente**

   La siguiente imagen muestra una la estructura de una publicación de empleo estándar en donde se pueden ver las diferentes variables que son capturadas

Figura 1.1: Ejemplo de una publicación

El siguiente diagrama ilustra el proceso que ser debe realizar para la captura de la información, dentro del programa se deberá seleccionar si se obtendrá la información de todo el país o de una ciudad especifica, luego se debe seleccionar si se van a capturar todos las o solo algún intervalo de sueldos especifico, después se inicia el ciclo según la cantidad de ofertas disponibles almacenando las 16 variables que se muestran.

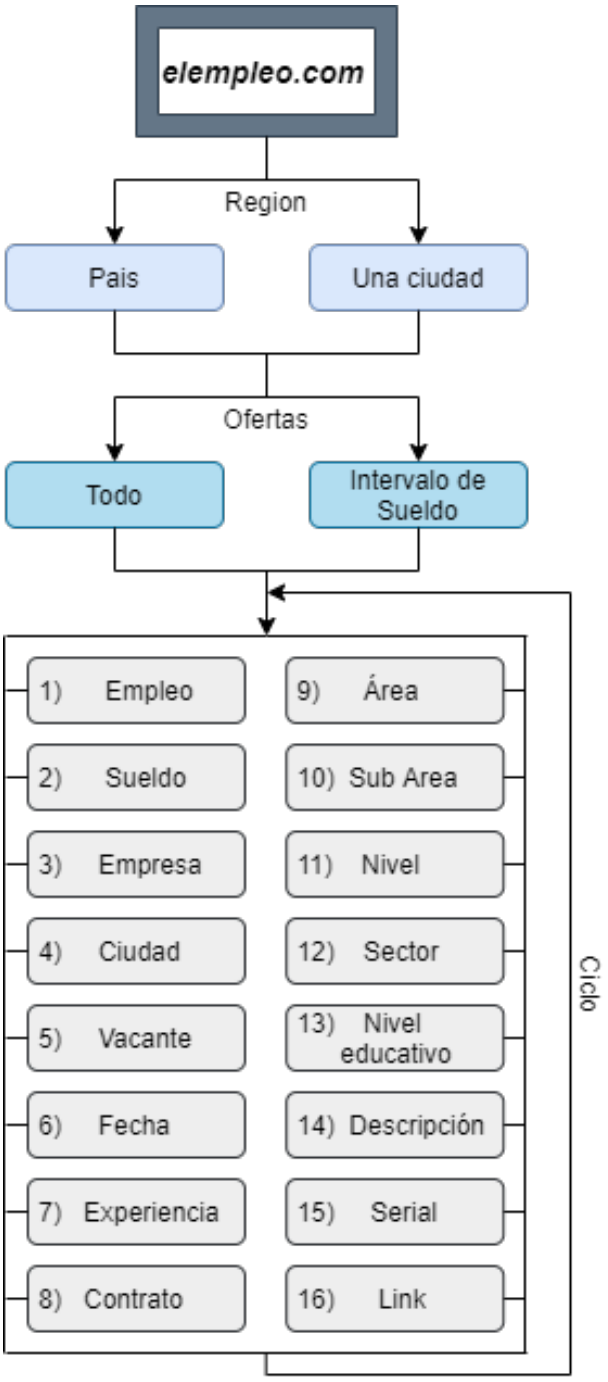

Figura 1.2: Diagrama

5. **Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.**

   El data set llamado *Empleo_2020-04-05_.csv* fue realizado en una única toma el 5 de abril del 2020, seleccionando todas las ofertas de empleo ofrecidas para la ciudad de Cartagena Colombia. El data set contiene los siguientes datos: *Empleo, Sueldo, Empresa, Ciudad, Vacante, Fecha, Experiencia, Contrato, Área, Sub área, Nivel, Sector, Nivel, educativo, Descripción, Serial y Link*, con las características que se describieron en la tabla del tercer punto. En general los datos contienen la información detallada de la oferta de empleo ofrecida, destacando el serial que permite un seguimiento de la oferta desde su creación hasta su desaparición.

6. **Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).**

   Agradezco a la empresa Leadearsearch S.A.S propietaria del producto elempleo.com (COPYRIGHT © 2019)

   Esta investigación no tiene fines comerciales y se realiza acorde a los términos y condiciones de la pagina, cito (la siguiente cita no busca un respaldo por el creador del contenido, solo se usa para ilustrar el punto)

   > USO DEL MATERIAL[1]
   > La Compañía lo autoriza a usted a consultar, revisar y usar el material que se encuentra en el Web Site, únicamente para su uso personal y no comercial y según lo establecido en estos Términos y Condiciones. El contenido de este Web Site, incluyendo pero sin limitarse a los textos, gráficas, imágenes, logotipos, iconos, software y cualquier otro material (el 'Material) están protegidos bajo las leyes colombianas de derechos de autor, leyes de propiedad industrial y otras leyes aplicables. Todo el Material es de propiedad de la Compañía o de sus proveedores o clientes. El uso no autorizado del material puede constituir una violación de las leyes colombianas o extranjeras sobre derechos de autor, leyes de propiedad industrial u otras leyes. Usted no podrá vender o modificar el Material en manera alguna ni ejecutar o anunciar públicamente el Material ni distribuirlo para propósitos comerciales. Usted no podrá copiar o adaptar el código HTML que la Compañía crea para generar sus páginas, ya que el mismo está protegido por los derechos de autor de la Compañía.

[1] https://www.elempleo.com/co/terminos-condiciones

7. **Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.**

   Actualmente trabajo en el instituto de estadísticas de mi país (DANE) en un proyecto para crear una base de datos longitudinal a partir de la información recopilada por el sistema de seguridad social, para esto se han buscado mas registros administrativos que enriquezcan la base producidos por otras entidades publicas. Con esta actividad pude reconocer el potencial que tiene el web scraping de la pagina «El empleo» para recopilar información directamente del mercado laboral, con una vigencia única y un coste muy bajo, aun no se si se puede incluir esta información directamente en la base, pero si estoy seguro que va a ser útil para realizar estadística que permitan comparar con los datos que produzca la base.

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:**

   Al elegir la licencia adecuada estoy teniendo en cuenta que no soy el productor ni el propietario de la información que esta en el dataset, y que no realice ninguna alteración a los datos, solo me limite a recolectarlos y organizarlos. Por esto no me es posible autorizar un uso que valla en contra de los limites que impuso el productor y dueño de los datos. Por estos elijo la licencia **Released Under CC BY-NC-SA 4.0 License** ya que prmite copiar,adaptar,transformar y redistribuir en cualquier medio o formato la información, pero se debe realizar un reconocimiento adecuado a los autores y prohíbe el uso comercial, por lo que considero que va acorde a los términos y condiciones del creador del contenido.

Contribuciones	Firma
Investigación previa	Cristhyan Naranjo
Redacción de las respuestas	Cristhyan Naranjo
Desarrollo código	Cristhyan Naranjo

2. OBSERVACIONES

Realice el programa en R porque ya había tenido experiencia manejando el paquete RSelenium , por lo que me pareció mas cómodo usar este lenguaje. Se realizaron varios ensayos para lograr un programa eficiente que capturara la mayor cantidad de información útil, pero hubo un problema que no se ha podido solucionar, resulta que de forma al parecer de manera aleatoria (al inicio pensé que se causaba por la inclusión de nuevas ofertas, luego vi que no) después de entrar a una oferta, al volver a la lista de empleos , estos aparecen ordenados de manera diferente,

luego como el programa recorre de forma ordenada las ofertas, terminan habiendo ofertas que no se consultan y otras que se consultan varias veces, sospecho que esto lo hace la página precisamente para evitar que robots consigan la información. Por lo anterior se obtuvo un dataset con registros repetidos, y algunos registros nulos, así que por el momento se está generando una base interesante para practicar técnicas de limpieza y obtener una muestra significativa de los empleos ofrecidos, Además queda abierto a que junto con la participación de la comunidad se puedan arreglar los problemas detectados. En las siguientes imagines se puede ver como después de realizar la primer consulta, al volver a la lista de ofertas, algunos empleos cambiaron de posición.

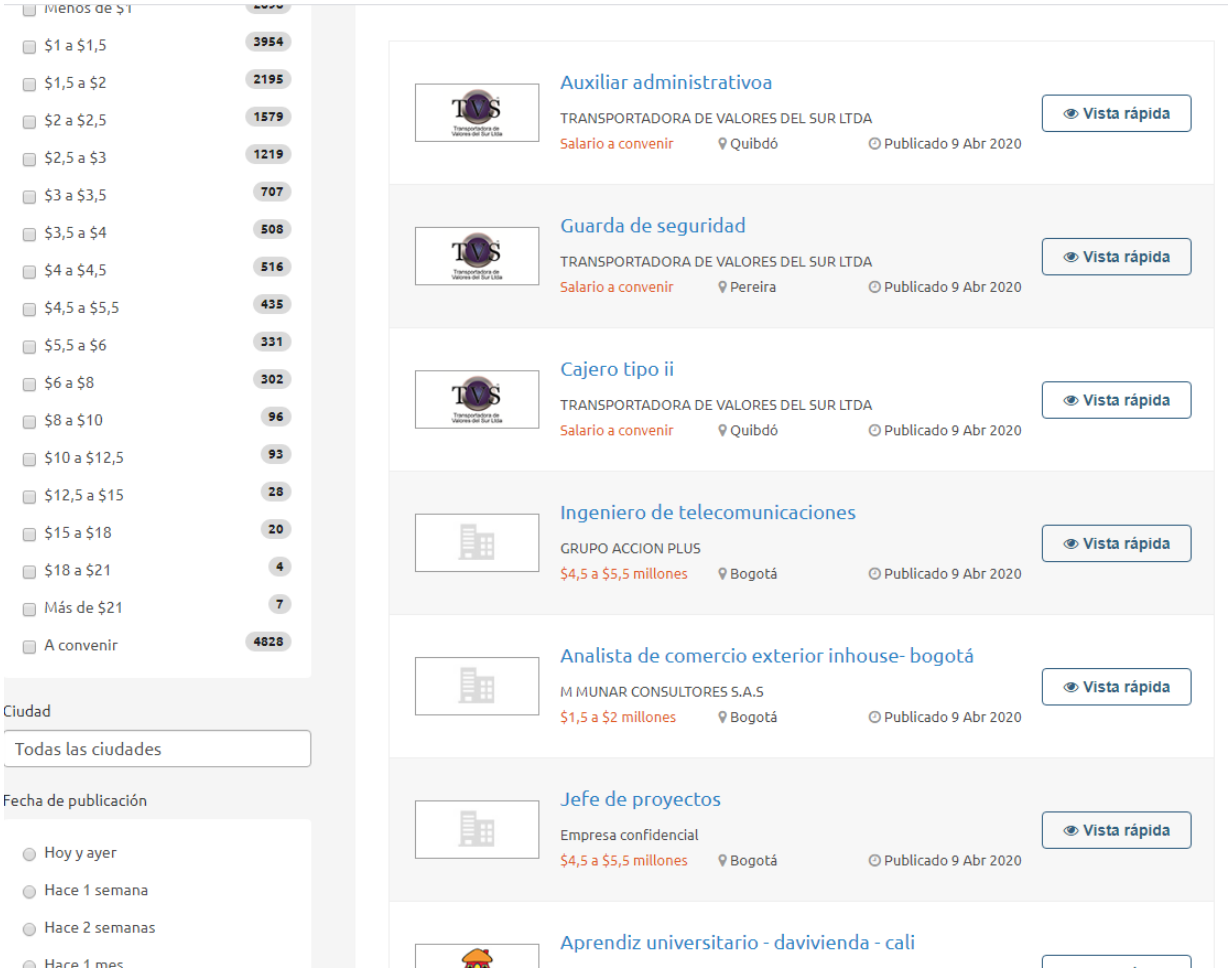

Figura 2.1: Lista de ofertas antes de la consulta

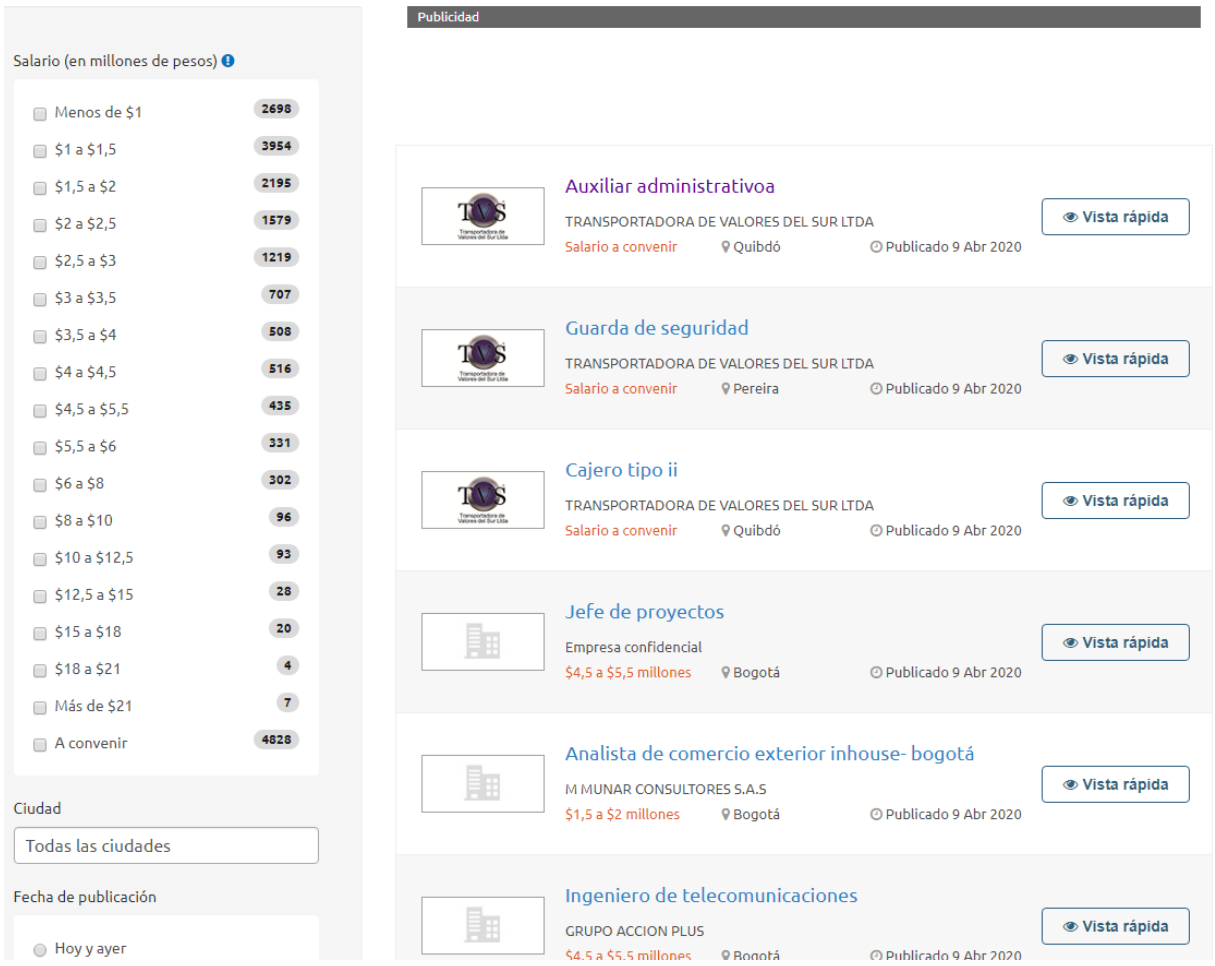

Figura 2.2: Lista de ofertas despues de la consulta