

Propuesta de Proyecto Web Scraping

Scraper de ofertas laborales desde LinkedIn



INTEGRANTES

Nombres**Usuario Git Hub**

Jorge Araya

GeorgeAdrock

Enrique salas

enriques76

Litzy Castro

LitzyCastro

OBJETIVO

Extraer data de ofertas laborales relacionadas con el campo del Data Science

Entender el comportamiento del mercado respecto de la necesidad de profesionales del área

elmostrador Noticias Mercados TV **Cultura** Revista Jengibre Agenda País Braga Avisos Legales

NOTICIAS | CULTURA+CIUDAD

CULTURA

Gobierno, sector privado y academia potencian proyecto para impulsar a Chile como referente en data science

por El Mostrador Cultura | 13 enero, 2020



El Observatorio ALMA en San Pedro de Atacama fue el lugar elegido para que los Ministerios de Ciencia y de Economía junto a Amazon Web Services y la Universidad Adolfo Ibáñez, presentaran los resultados preliminares del primer año de trayectoria del Data Observatory, una plataforma científica nacional de datos que busca contribuir al progreso en ciencia, tecnología, conocimiento e innovación.

#KingstonisWithYou
SSD de alto rendimiento
Kingston

#KingstonisWithYou
SSD de alto rendimiento
Kingston

Fuente: <https://www.elmostrador.cl/cultura/2020/01/13/gobierno-sector-privado-y-academia-potencian-proyecto-para-impulsar-a-chile-como-referente-en-data-science/>

Sitio a escapar :
<https://www.linkedin.com/>

The screenshot shows a LinkedIn search results page for the query 'data science' in Chile. The page displays four job listings. The first listing is for 'Data Scientist / Machine Learning Engineer' at Microsoft, which is highlighted in blue. The second listing is for 'Jefe Data Scientist' at Sodimac. The third listing is for 'Data Science' at Itaú Chile. The fourth listing is for 'Data Scientist' at Falabella Retail S.A. On the right side of the page, there is a detailed view of the 'Data Scientist / Machine Learning Engineer' job at Microsoft, including details about the location, experience level, and a description of the role.

LinkedIn search results for 'data science' in Chile. The page shows 312 results. The top results are:

- Data Scientist / Machine Learning Engineer** at Microsoft. Location: Gran Santiago, Región Metropolitana de Santiago, Chile (En remoto). 1 contacto trabaja aquí. Hace 1 día.
- Jefe Data Scientist** at Sodimac. Location: Gran Santiago, Región Metropolitana de Santiago, Chile (En remoto). Tu perfil se ajusta a este empleo. Hace 5 días. Solicitud sencilla.
- Data Science** at Itaú Chile. Location: Gran Santiago, Región Metropolitana de Santiago, Chile (Presencial). 1 contacto trabaja aquí. Hace 1 semana.
- Data Scientist** at Falabella Retail S.A. Location: Gran Santiago, Región Metropolitana de Santiago, Chile. 2 contactos trabajan aquí. Hace 1 semana.

On the right side, the details for the 'Data Scientist / Machine Learning Engineer' job at Microsoft are shown:

- Data Scientist / Machine Learning Engineer**
- Microsoft · Gran Santiago, Región Metropolitana de Santiago, día · 51 solicitudes
- Jornada completa · Sin experiencia
- Más de 10.001 empleados · Desarrollo de software
- 1 contacto · 35 antiguos alumnos
- Ve una comparación con los otros 54 solicitantes. [Prueba](#)
- En busca de personal
- [Solicitar](#) [Guardar](#)
- Preferred location: Argentina, Brazil, Chile, Colombia, Costa Rica
- We are seeking an experienced Data Scientist in our research to expand our cross-service, signal-based protections using state-spanning the areas of supervised and unsupervised machine learning. Our organization focuses on protecting customers against cyber engineering attacks such as phishing across our browsers and email. Machine learning (ML) is a fundamental part of how we protect you come in.
- At Microsoft, you'll have access to vast amounts of threat-related endpoints and other sources. You will have the opportunity to...

Datos a extraer



Fecha publicación



Cargo



Empresa



Descripción



Skills



Forma de postulación

Frame



Publicaciones últimas 24 horas



Keyword: data science



Estimación de datos: Promedio 20 publicaciones diarias (600 x mes)



Análisis acumulado de resultados



Base de conocimiento: Orientación de mercado



Feedback de la Base de Conocimiento: Mejorar Perfil Profesional o CV de acuerdo a descripción en publicaciones

Tecnología a utilizar



GitHub



Python

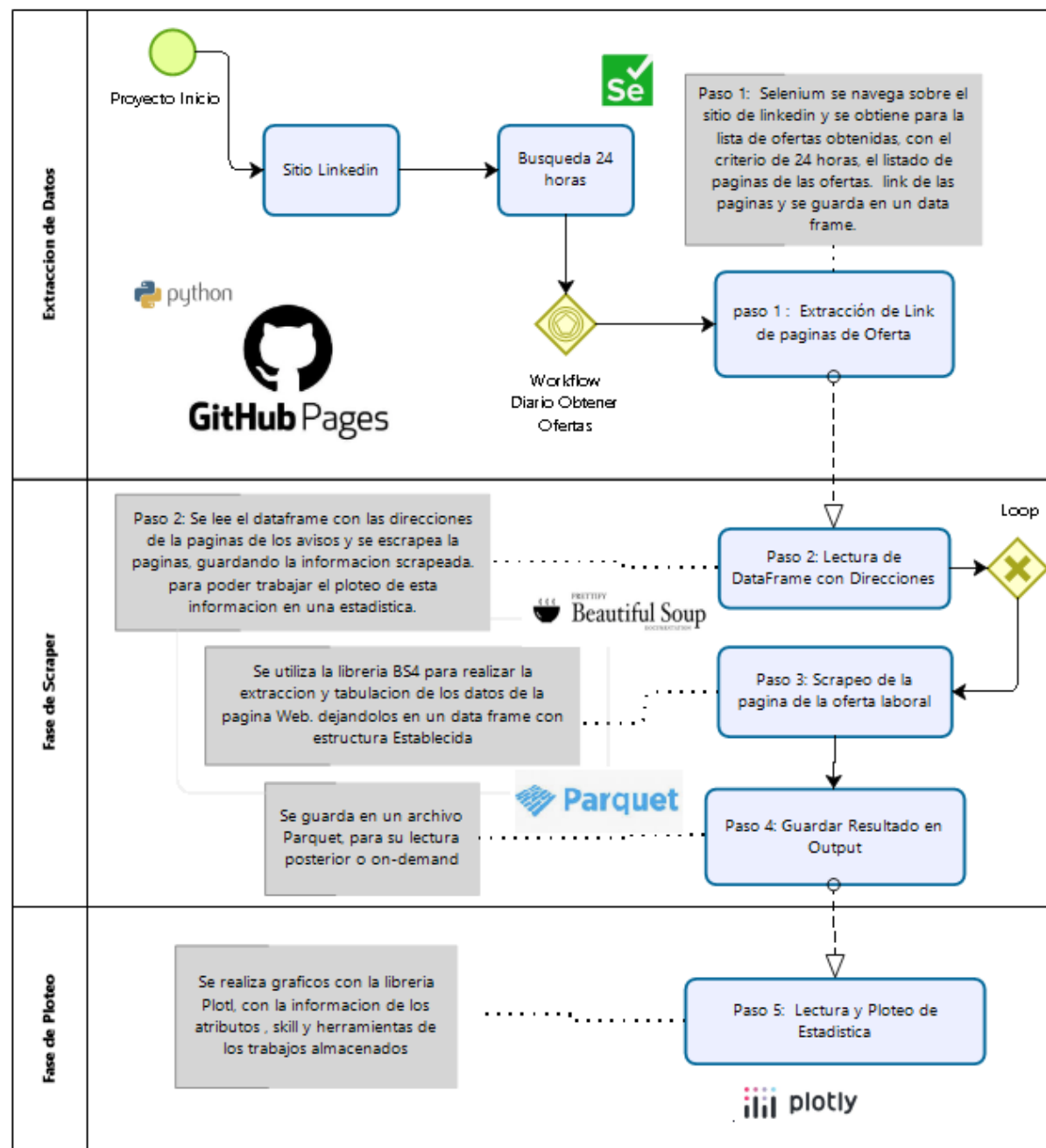


Beautiful Soup



Selenium

Estructura y estrategia de Implementación del scraper



Fragmento de Código

Importación de librerías

```
1 # Tratamiento de datos
2 # =====
3 import numpy as np
4 import pandas as pd
5 import re
6 import time
7 from datetime import date
8 import pyarrow
9
10 # Manejo Web, paginas y webScrapping
11 # =====
12 import urllib.request
13 from selenium import webdriver
14 from selenium.webdriver.common.keys import Keys
15 from selenium.webdriver.common.by import By
16 from selenium.webdriver.support.ui import WebDriverWait
17 from selenium.webdriver.support import expected_conditions as EC
18 from selenium import webdriver
19 from bs4 import BeautifulSoup as bs
20
```

Búsqueda de datos a extraer

```
# Instantiate the webdriver with the executable location of MS Edge
# Provide the full location of the path to recognise correctly
PATH = 'msedgedriver.exe'
options = webdriver.EdgeOptions()
options.add_argument('--start-maximized')
options.add_argument('--disable-extensions')
options.add_argument('disable-dev-shm-usage')
options.add_argument('--no-sandbox')
options.add_argument('--blink-settings=imagesEnabled=false')
options.add_argument('--headless')
driver = webdriver.Edge(PATH, options=options)

linkedin_soup = bs(driver.page_source.encode("utf-8"), "html")
#print(linkedin_soup)
patron = '/jobs/view/'
df = ExtraerLink(linkedin_soup['a'], patron)
df.info()
```

Funciones

```
1 def ExtraerLink(linkPage, patron):
2     lista = []
3     for tag in linkPage:
4         valor = tag.get('href')
5         if(str(valor).find(patron) != -1):
6             lista.append(valor)
7     df = pd.DataFrame(lista, columns = ['url'])
8     df = df.drop_duplicates()
9     return df
10 def leerUrl(pagina):
11     soup = bs(urllib.request.urlopen(pagina).read().decode())
12     #print(str(soup))
13     time.sleep(3)
14
15     return soup
```

Desarrollo de Wordcloud

```
wordcloud = WordCloud(width = 800, height = 800,
                       background_color = 'white',
                       stopwords = stop_words,
                       min_font_size = 10).generate(comment_words)

plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```

Resultados

- Scraper que permita construir una base de conocimiento de ofertas laborales de LinkedIn.
- Estadísticas respecto de la demanda del mercado de profesionales del área de Data Science.

