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 Buscar

NOTICIAS | CULTURA+CIUDAD

CULTURA

Cultura

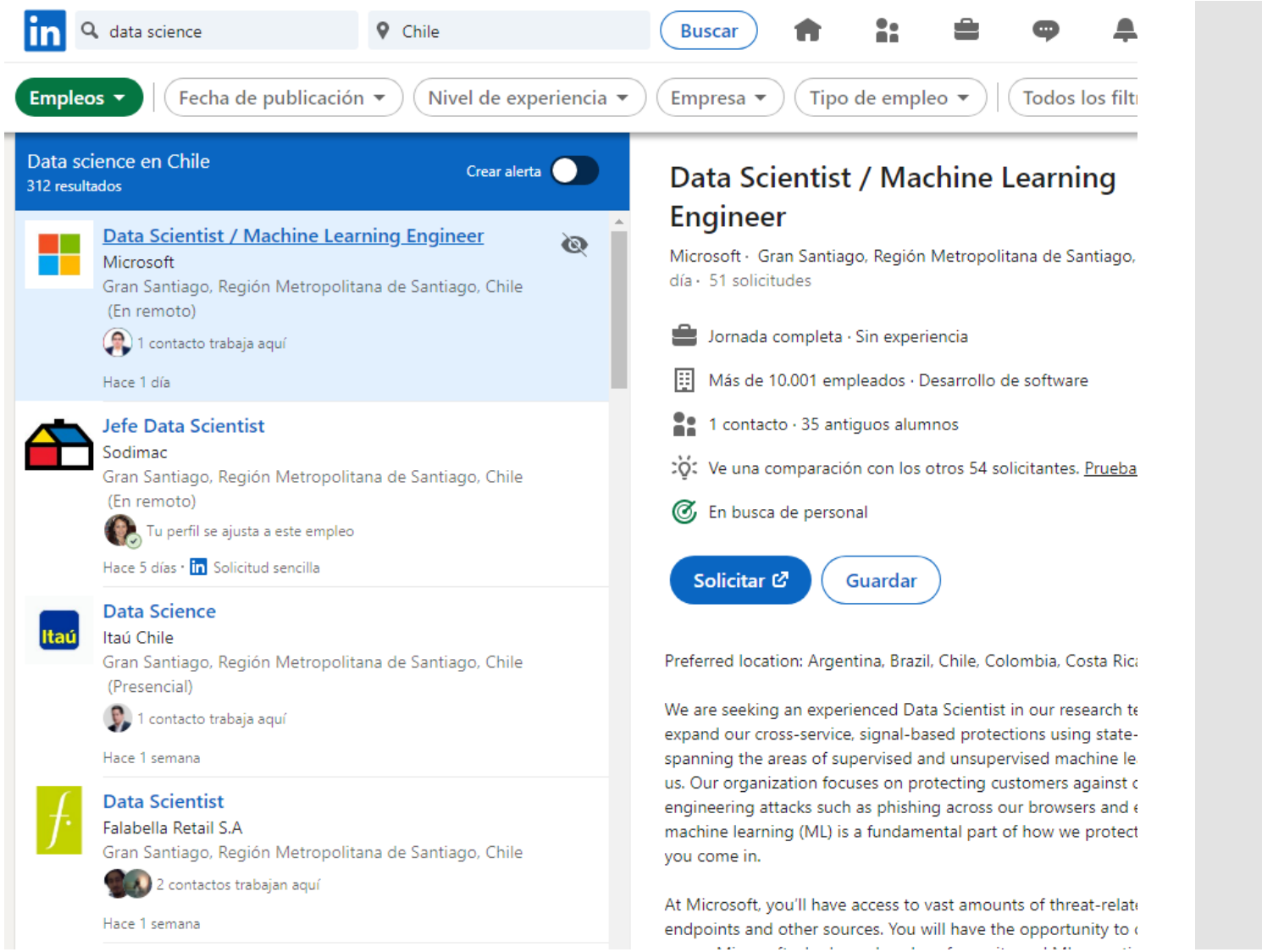
Gobierno, sector privado y academia potencian proyecto para impulsar a Chile como referente en data science

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.

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 escrapear : https://www.linkedin.com/



Datos a extraer

- Fecha publicación
- Cargo
- Empresa
- Descripción
- Skills
- Forma de postulación

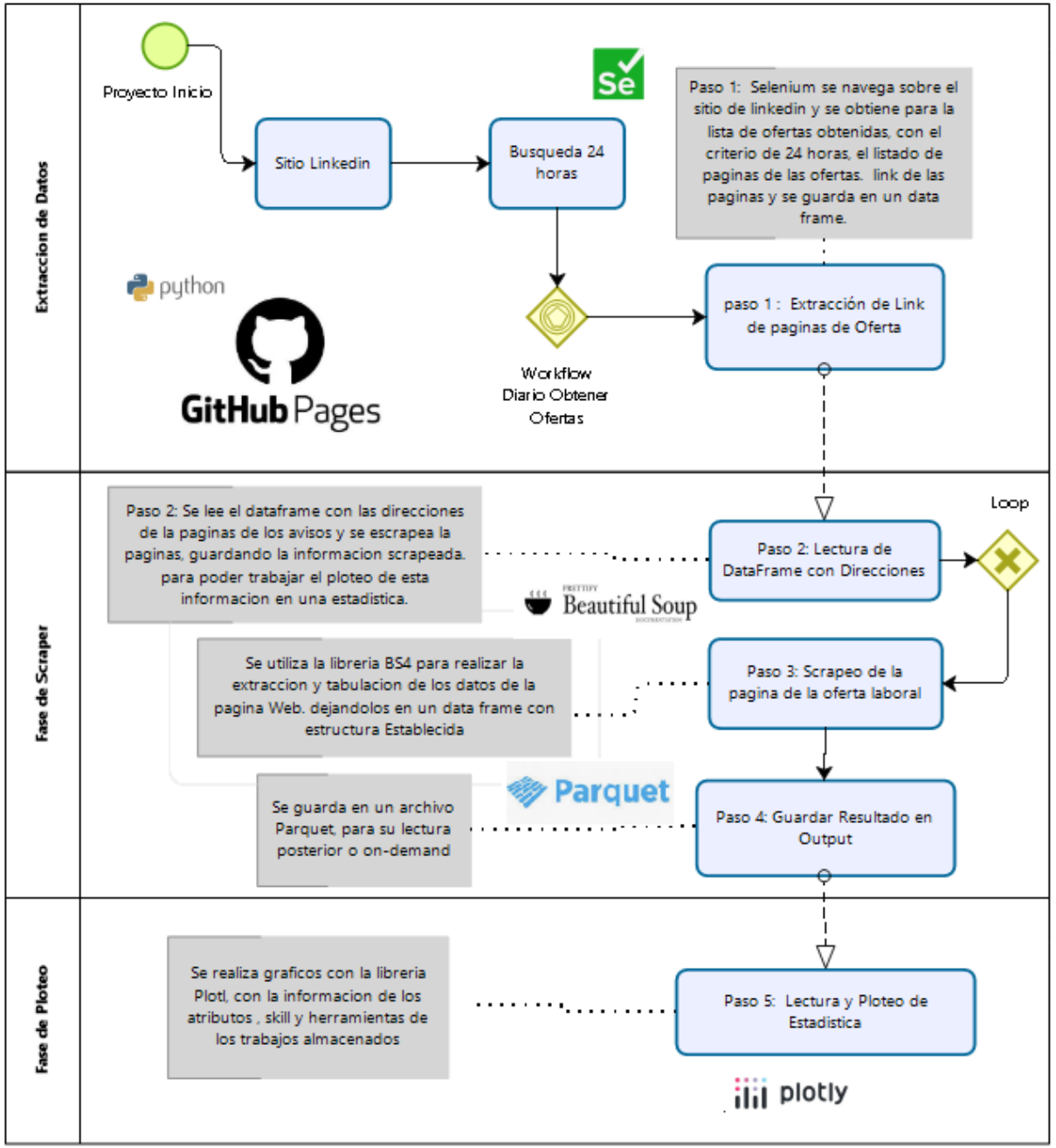
Frame

- 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

```
# Tratamiento de datos
# ==============================================================================
import numpy as np
import pandas as pd
import re
import time
from datetime import date
import pyarrow

# Manejo Web, paginas y webScrapping
# ==============================================================================
import urllib.request
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium import webdriver
from bs4 import BeautifulSoup as bs

```

Funciones

```
def ExtraerLink(linkPage,patron):
    lista = []
    for tag in linkPage:
        valor = tag.get('href')
        if(str(valor).find(patron) != -1):
            lista.append(valor)
    df = pd.DataFrame (lista, columns = ['url'])
    df = df.drop_duplicates()
    return df
def leerUrl(pagina):
    soup = bs(urllib.request.urlopen(pagina).read().decode())
    #print(str(soup) )
    time.sleep(3)

    return  soup
```

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()
```

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.