import requests

from bs4 import BeautifulSoup

import re

import csv

import time

import urllib3

from datetime import date

from ais\_utilities import date\_format

from ais\_utilities import check\_keywords\_in\_title

from ais\_utilities import keywords

from ais\_utilities import getPage

from selenium.webdriver import Chrome

from selenium.webdriver import ChromeOptions

from selenium.webdriver.chrome import options

from selenium.webdriver.common.by import By

from selenium.webdriver.support.ui import WebDriverWait

# -\*- coding: utf-8 -\*-

import dataiku

import pandas as pd, numpy as np

from dataiku import pandasutils as pdu

# Read recipe inputs

driver\_path = '/data/dss\_data/code-envs/python/ASTELLAS\_INTELLIGENT\_SURVEILLANCE\_AIS/lib64/python3.6/site-packages/chromedriver'

def getContent(url):

print("Getting content for:- ", url)

cookie\_jar=requests.cookies.RequestsCookieJar()

options = ChromeOptions()

driver = Chrome(executable\_path=driver\_path, options = options)

driver.implicitly\_wait(5)

wait = WebDriverWait(driver, 10)

driver.get(url)

table = driver.find\_elements\_by\_class\_name("xmsonormal")

for t in table:

print(t.text)

def pacific(url):

cookie\_jar=requests.cookies.RequestsCookieJar()

options = ChromeOptions()

options.add\_argument('--headless')

options.add\_argument('--disable-gpu')

driver = Chrome(executable\_path=driver\_path, options = options)

driver.implicitly\_wait(0.5)

driver.get('https://www.pacificbridgemedical.com/resource-center/?cat\_resource\_type%5B%5D=17&cat\_resource\_market=0&\_type=')

table = driver.find\_elements\_by\_tag\_name("article")

all\_news = []

for t in table:

news = {}

title = t.find\_element\_by\_class\_name("entry-title").text.lower()

url\_content = t.find\_element\_by\_tag\_name("a").get\_attribute('href')

date1 = date\_format(t.find\_element\_by\_tag\_name("time ").text)

ret = check\_keywords\_in\_title(title, keywords)

print("Checking for:- ", title)

if ret:

print("Keyword matched.........")

match = ""

match = " | ".join(word for word in ret)

print(match)

news['identifier'] = url

news['title'] = title

news['url'] = url\_content

news['date'] = date1

news['keywords'] = match

all\_news.append(news)

else:

print("\t No keywords matched.......")

for news in all\_news:

options = ChromeOptions()

options.add\_argument('--headless')

options.add\_argument('--disable-gpu')

driver = Chrome(executable\_path = driver\_path, options = options)

driver.implicitly\_wait(5)

wait = WebDriverWait(driver, 10)

driver.get(news['url'])

table = driver.find\_elements\_by\_class\_name("entry-content")

content = ''

print("Content pouplated............ for ",news['title'])

for t in table:

content += t.text

news['content'] = content

driver.close()

return all\_news