

Data Scraping Exercise

Overview

The purpose of this exercise is to assess the candidate's automated data capturing knowledge. The exercise involves building a "proof of concept". The exercise will be evaluated on overall task execution, coding style, understanding of programming concepts and choice of techniques.

Description

As the data companies have available to them continues to grow in both amount and complexity, so does the need for an effective and efficient process by which to harness the value of that data.

Automated data capture reduces the need for manual data entry. Instead, it uses automated data entry software to collect data from data forms, surveys, invoices, docs, email, fax, and more. Then, it converts it into meaningful insights as a readable digital format with more accuracy and less cost.

The task is to create a POC for automatic data capturing from different online sources and generate automated excel or csv files from it which can be further uploaded to a DB.

Guidelines

1. Use Python.
2. Can code in any environment.
3. Output file should be in excel or csv only.
4. Proper headings and data cleaning should be included in the program.
5. A simple analysis of the data in hand would be a plus.
6. Attempt atleast 2 of the links given below:
 - a. NGO- <https://socialjustice.nic.in/UserView/index?mid=73590>
List of all the NGO's and the action taken should be as shown on the website.
(Plus point if you extract the links of the documents of the actions as well)

Name of the NGO	Action Taken by the Ministry
Association of Moral Guide and Service to Poor [AMGALAS], At/P.O.Distt.Nayagarh, Orissa	Blacklisted
Asha Bai Mandir Shiksha Samiti, Plot No.1, B.Krishnapuri, Jaipur-1	Blacklisted
Social Welfare Society, 2-A, Main Road, Tittagudi, Tamil Nadu	Blacklisted
Bhartiya Samajothan Sewa Sansthan, Nehru Nagar, Chakiyawa, Deoria, Uttar Pradesh	Blacklisted on 27.08.2002 (No.36-19(4)/2001-DD-II)
Shaheed Abdul Hameed Education Institute, Dherwaha, Khatipura, Ward No.60, Dist Yuvatmal, Maharashtra	Blacklisted
Shradha, Pakyong, East Sikkim	The state Government asked to recover grants and sieze the assets created out of Government Funds

- b. Lok_Sabha- <http://164.100.47.194/Loksabha/Members/MemberSearch.aspx>
List of all the members of the Lok sabha is to be extracted.
(Plus point if you extract the links of their individual pages as well)

Total Records : 18 (Page 1 of 1)








S.No.	Name of Member	Party Name	Constituency(State)
1	Wadiwa , Shri Narayanrao Maniram	Congress	Seoni ,Madhya Pradesh,
2	Wadiyar , Shri Srikanta Datta Narasimharaja	Indian National Congress	Mysore ,Karnataka,
3	Wagh , Dr, Pratap	Congress (I)	Nashik ,Maharashtra,
4	Waghmare , Shri Narayan Rao	Peasants And Workers' Party of India	Parbhani ,Hyderabad,
5	Wagmare , Shri Suresh Ganpatrao	Bharatiya Janata Party	Wardha ,Maharashtra,

- c. MCA- <https://www.mca.gov.in/content/mca/global/en/contact-us/official-liquidators.html>

List of all the official liquidators as well as all the information about them should be extracted.

(Plus point if you extract the Geo-Location links as well)

Official Liquidators

Name	Designation	Contact Details	Address
Sh Sitaram Sharan Gupta	Official Liquidator, Ministry of Corporate Affairs.	 0731-2710051  0731-2710568  ol[dot]indore[at]mca[dot]gov[dot]in	 1st Floor, Old CIA Building, Opp. GPO Comps Residency Area, Indore - 452001, Madhya Pradesh View on Map
Sh Prahlad Meena	ROC-cum- Official Liquidator, Ministry Of Corporate Affairs.	 0651-2531811,1401  ol-ranchi-mca[at]nic[dot]in	 Mangal Tower, 4th Floor, Old Hazaribagh Road, Near Kanta Toli Chowk, Ranchi - 834001

Please make assumptions wherever needed.

Deliverables:

1. The Source code of the program.
2. The data file generated.
3. Assumptions you took.

Evaluation:

1. Accuracy of the data
2. Clean code
3. Code structure
4. Code style
5. Completion of the task
6. Business Logic
7. Libraries used

Hint : Use selenium or BeautifulSoup libraries.