Test 1 – Web scraping

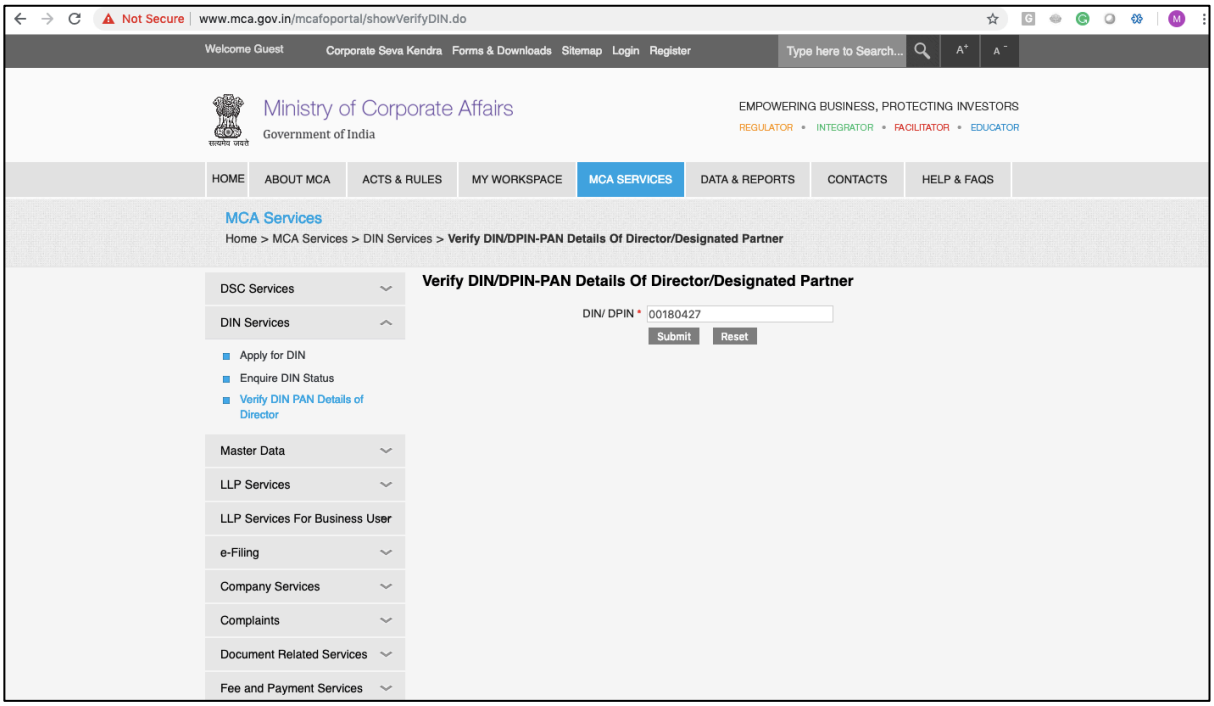
Requirement – To insert the director identification number (DIN) values shared in the xls workbook in the below mentioned field and URL and extract the output in an csv or txt file. This module should work on any given number of DIN values provided in an input csv file and post-extraction from the web the output should be available in an cvs or txt file with the name of the director and DIN value.

Web URL – http://www.mca.gov.in/mcafoportal/showVerifyDIN.do

Input Field Name – DIN/DPIN

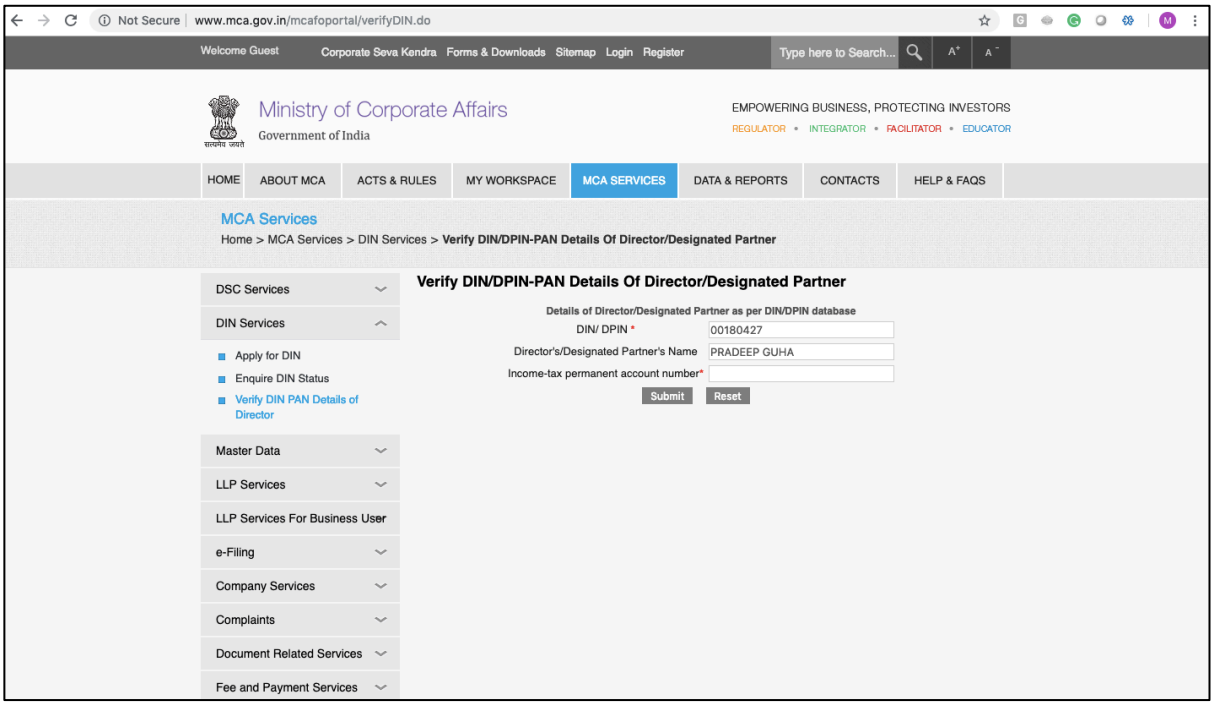
Input Screen – 1

1. To insert DIN value



2. Click on Submit

Output Screen – 1



Name of the Director is part of the output, which is to be captured in the output with the given input DIN value. Please note Income-tax permanent account number is not shared in the output so please ignore it.

Output Format Headers – DIN | Name of the Director

Other Points for Consideration:

- Input file has a list of 100 DIN values and the developed module should be capable to ping the above URL, insert the DIN value and fetch the name of the director and store it back in the local system in a sequential manner.
- No limitations on use of technology for scraping information and one can go with Python, Java, NodeJS or any other programming language.
- Candidate can make use of any open source database if needed as there is no compulsion for the same.

Test 2 – Name Entity Recognition (NER)

Requirement – To use publicly available open source libraries of NER and apply it on a given set of media articles provided and share an output on name of People, Place, and Organization name in an xls file. Each media article comprises of a full body media article and objective is use NER algorithms to extract names of people, product and organization. For each article file name, the xls file should have comma separate values under people, place, and organization headers.

Output Xls Structure

File Name	People	Place	Organization
In_12435.txt	Ramesh,Suresh,Amit B Saxena,Hari,Peter	Bhopal,USA,Texas,California,Singapore,Delhi, New Delhi,UP,Vizag	Google,Miccorsoft,Bain,Mckinsey&Co,Reliance Industries Limited,Atul Limited,Facebook Co., United Breweries
In_53638.txt	Max,Peters,Uno,Sachin Tendulkar	Bandra,Mumbai,Coimbatore,Chennai,Jaipur	CMS Infosystems,Xander Finance,RBI,SBI,JPMC

Other Points for Consideration:

- Output will not be accurate as the openly available libraries has to be trained during the course of time
- Test is to check analytical and logical capabilities to improve output and implementation capabilities.
- No restrictions on use of a programming language
- Each txt file the given zip file has a unique ID in the naming convention and that will comprise of full body media article in English language