Web Scarping Toolkit Report

For this Module 6 assignment I am using 3 Web Scraper Toolkits to extract Super Bowl (Most valuable players') data from url

http://www.espn.com/nfl/superbowl/history/mvps

Information being extracted are as follows:

1)Number(Super Bowl Number)

2)Player Name(Most Valuable Player)

3)Highlight (Highlight of the match)

Web scraper tool kits used are

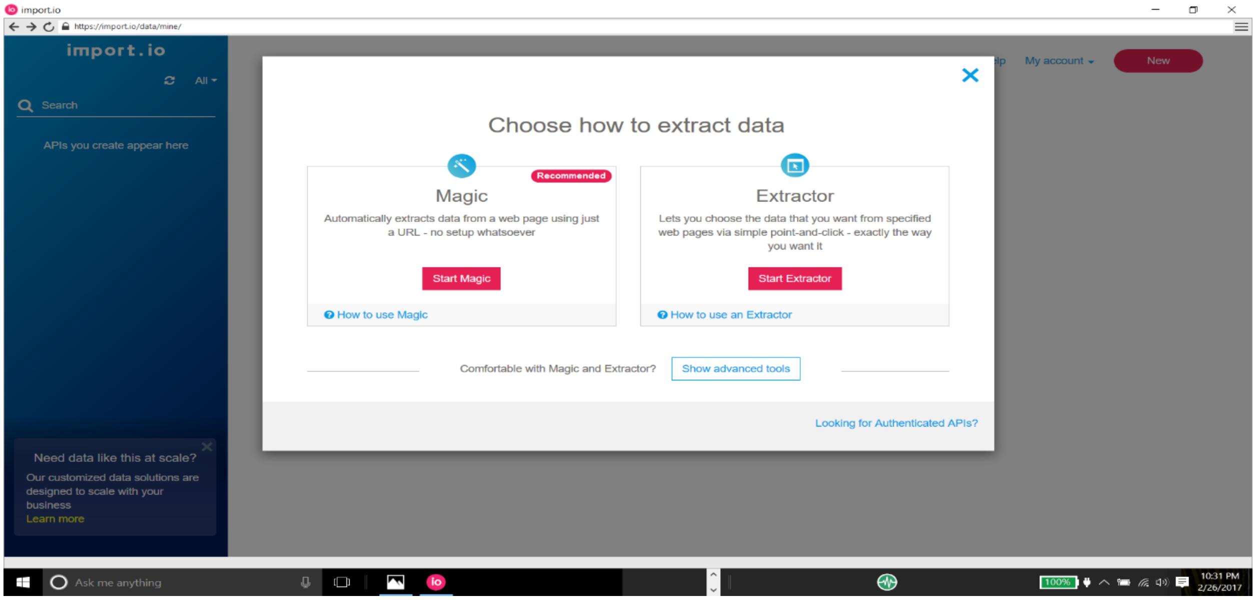
1) **Impot.io**

Pros :

1) It automatically can detect tables and simply extract data without choosing it. Smarter algorithm.
2) It has features like "Magic" (automatically extract data) and "Classic Extractor"(user has to choose which data to add)

Cons :

1) It is not free, though free trial version is available.
2) Slower in extraction sometimes.
3) Not able to modify searching pattern.
4) No access to Xpath.



Choose how to extract data

Magic

Automatically extracts data from a web page using just a URL - no setup whatsoever

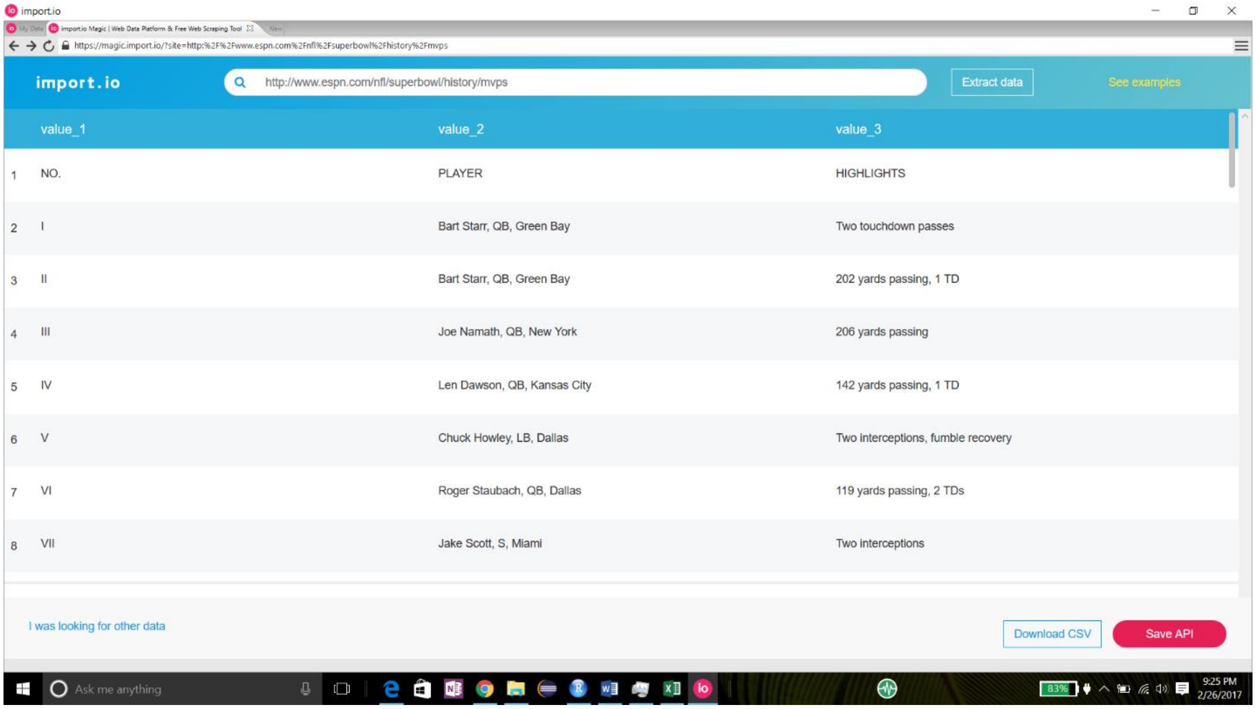
Extractor

Lets you choose the data that you want from specified web pages via simple point-and-click - exactly the way you want it

Comfortable with Magic and Extractor?

Need data like this at scale?

Our customized data solutions are designed to scale with your business


import.io
import.io Magic | Web Data Platform & Free Web Scraping Tool
https://magic.import.io/?site=http:%2F%2Fwww.espn.com%2Fnfl%2Fsuperbowl%2Fhistory%2Fmvps
import.io
http://www.espn.com/nfl/superbowl/history/mvps
Extract data
See examples
I was looking for other data
Download CSV
Save API
Ask me anything
83%
9:25 PM
2/26/2017

	value_1	value_2	value_3
1	NO.	PLAYER	HIGHLIGHTS
2	I	Bart Starr, QB, Green Bay	Two touchdown passes
3	II	Bart Starr, QB, Green Bay	202 yards passing, 1 TD
4	III	Joe Namath, QB, New York	206 yards passing
5	IV	Len Dawson, QB, Kansas City	142 yards passing, 1 TD
6	V	Chuck Howley, LB, Dallas	Two interceptions, fumble recovery
7	VI	Roger Staubach, QB, Dallas	119 yards passing, 2 TDs
8	VII	Jake Scott, S, Miami	Two interceptions

2)Instant Data Scraper (Google Extension)

Pros :

1)Easy to install as being google extension.

2)Easy to use : go to page you want scrape and click on the logo.

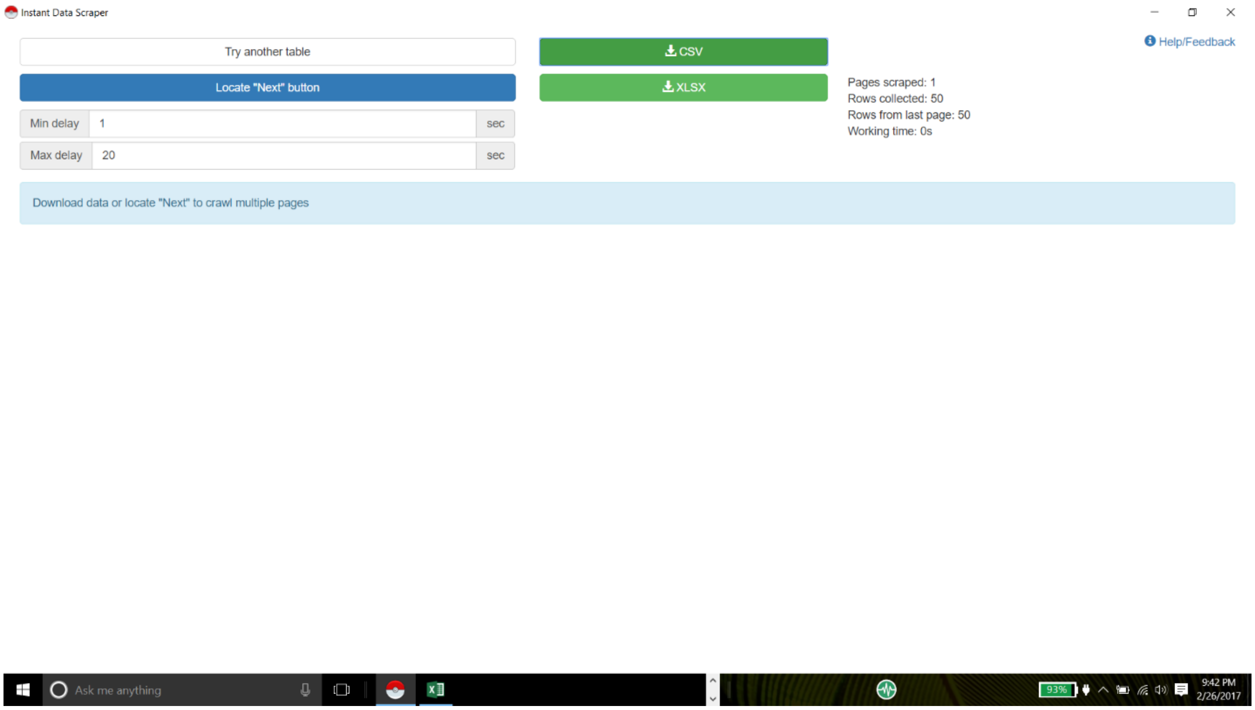
3)It is free and fast.

4)Extract tables automatically.

Cons:

1)Not able to choose all vectors and tables of the page together.

2)Data cleaning and tangling should be done manually.



Min delay 1 sec

Max delay 20 sec

Pages scraped: 1
Rows collected: 50
Rows from last page: 50
Working time: 0s

Download data or locate "Next" to crawl multiple pages

3)rvest

Pros:

1)Easily available as rvest package and its free.

2)Coding is easier than RCurl toolkit.

3)Reading data is easier.

4) Less cleaning and editing is required as we can select required data only and directly store into data frame.

Cons:

1)Requires source code of required web page .

2)Knowledge of HTML and nodes is required to extract.

3)Required coding to extract data instead of drag and drop data as import.io

Easiest to use -> import.io

Recommended to use ->rvest