Web Scraping Tool Kits Report

Provided by

Mohsen Nabian

PhD student in Mechanical Engineering Department

Northeastern University

By extraction and analysis of information from businesses centers, one can draw a safe and reliable path for his/her own business plan. In this homework we try 3 famous web scraper toolkits to extract useful data from Yelp website to get some information about restaurants located in Boston. Here is the link:

http://www.yelp.com/search?find_desc=Restaurants&find_loc=Boston%2C+MA&ns=1

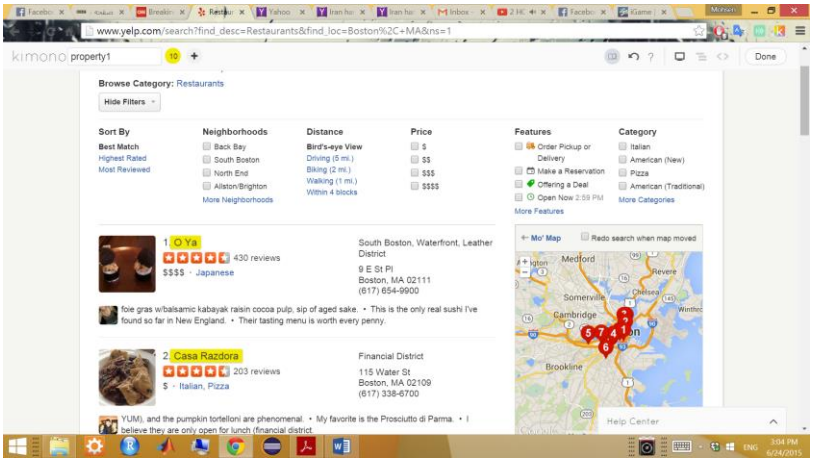
The information that are being extracted are as follows:

1)Restaurant_Name 2) Price level 3)zip_code 4)Tell Number

a) KIMONO

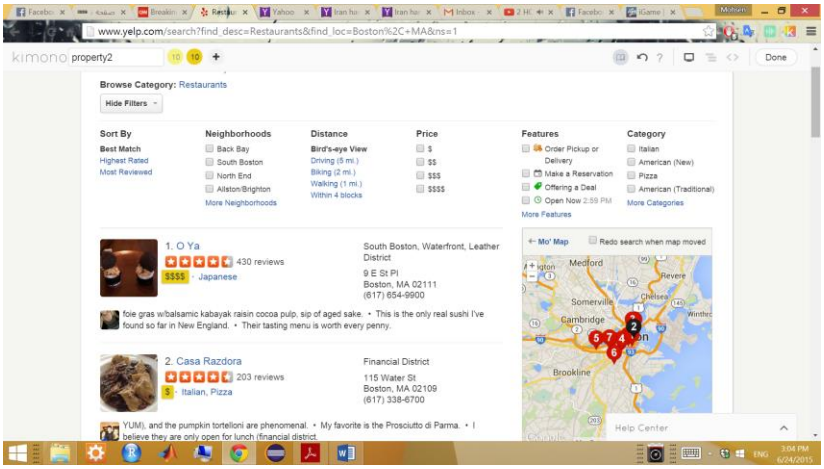
Kimono can be installed as a google chrome extension. Accuiraing data with Kimono is illustrated with the following figures:

1) Choosing one data from the vector of data that we are looking for. If any element is wrongly selected we can remove them. On the other hand if any data is missing, we can select them. In these two ways we can improve the filtering method.

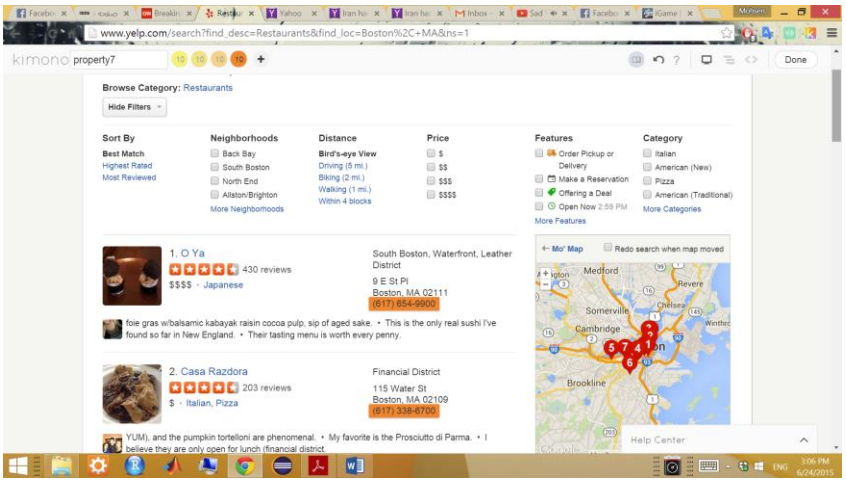


2) Adding next column

Now we can add another column simply by clicking on + sign and do the same as step1.

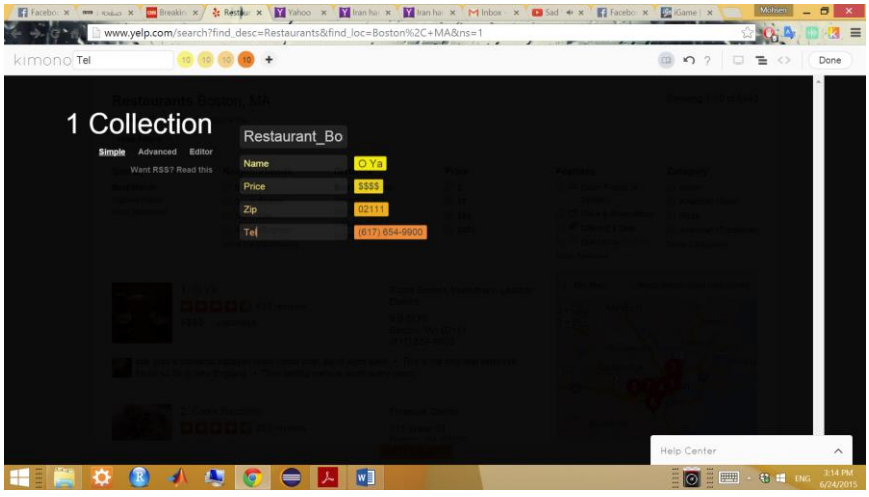


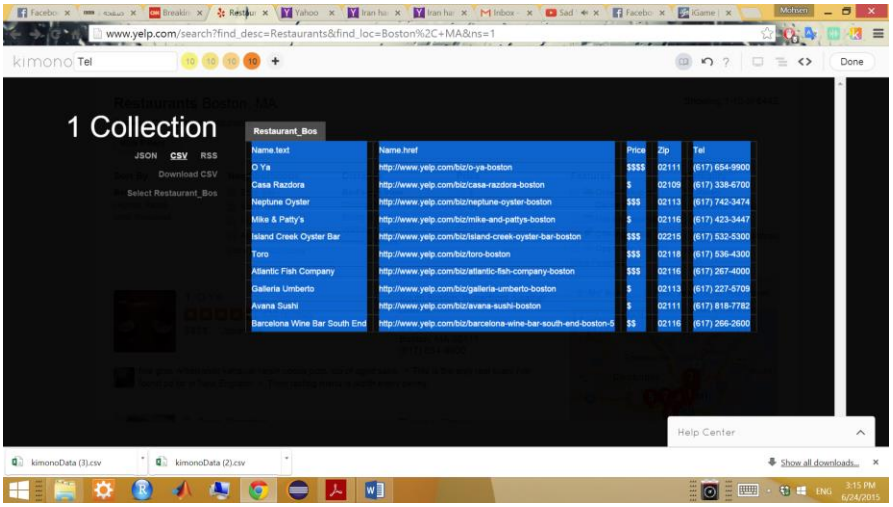
3) Getting all variables we want: Now all data is selected properly. Now we should press “Done”.

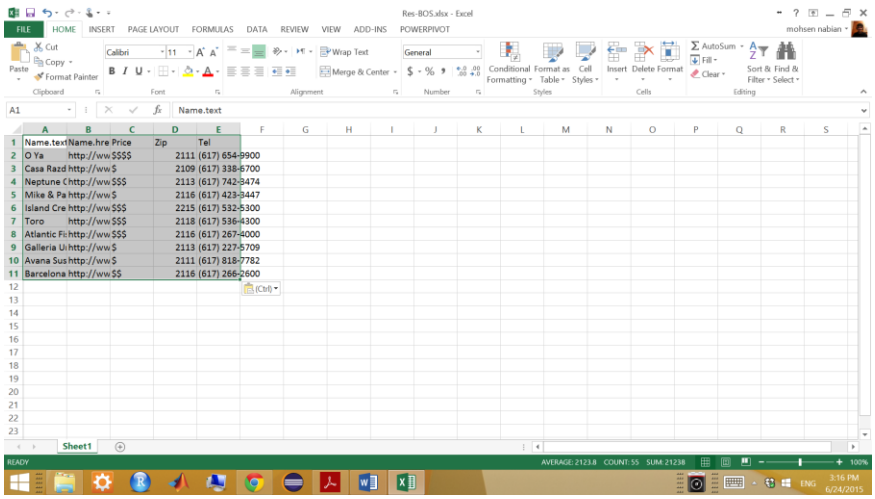


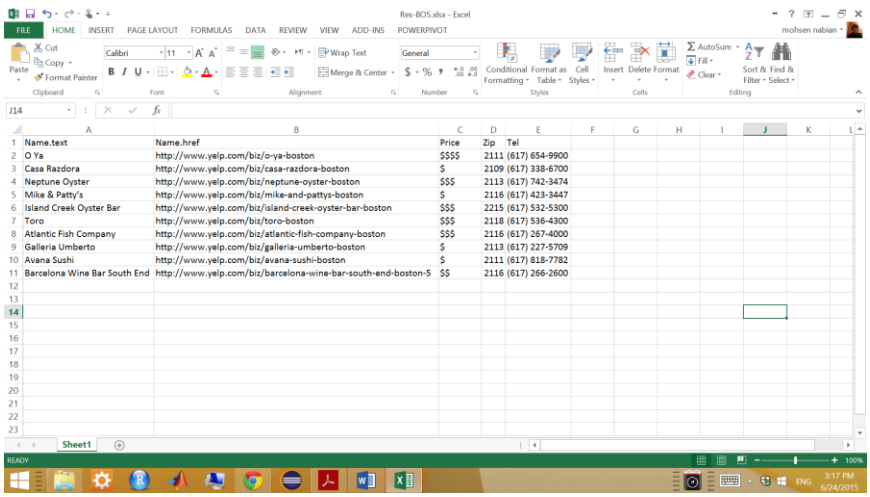
4) Sorting and Naming and Publishing:

In this stage, we can specify names for each vector. And By asking for CSV format you can simply copy the data and then paste is in a excel file. The following 4 pictures demonstrate the steps visually.





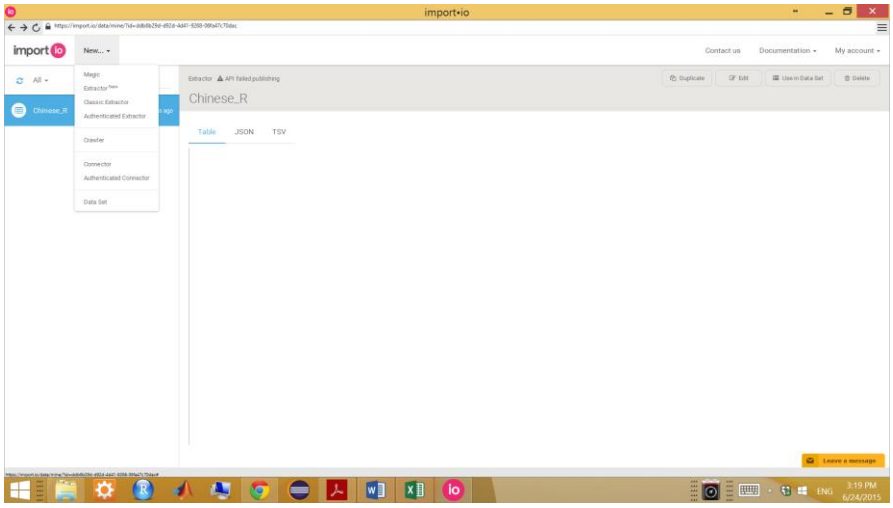




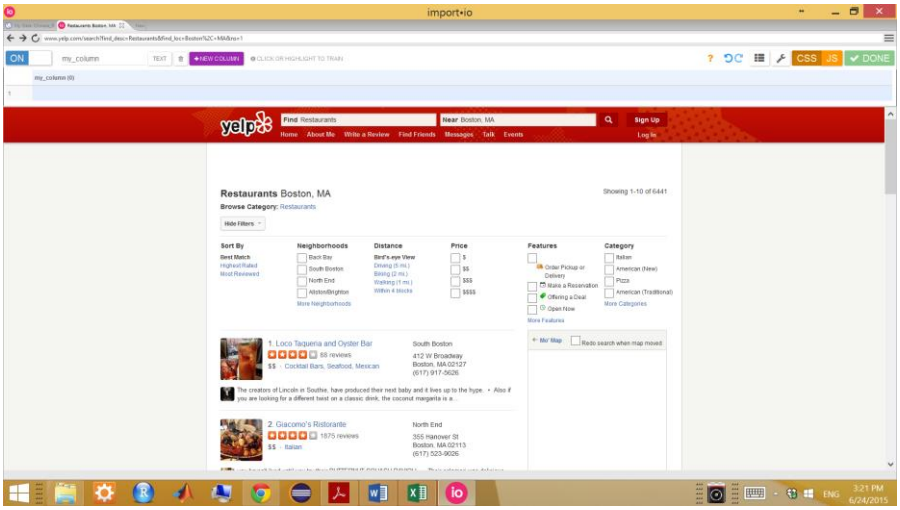
b) Import.io

This software should be downloaded and installed to the computer hard-drive. The following pictures shows the steps.

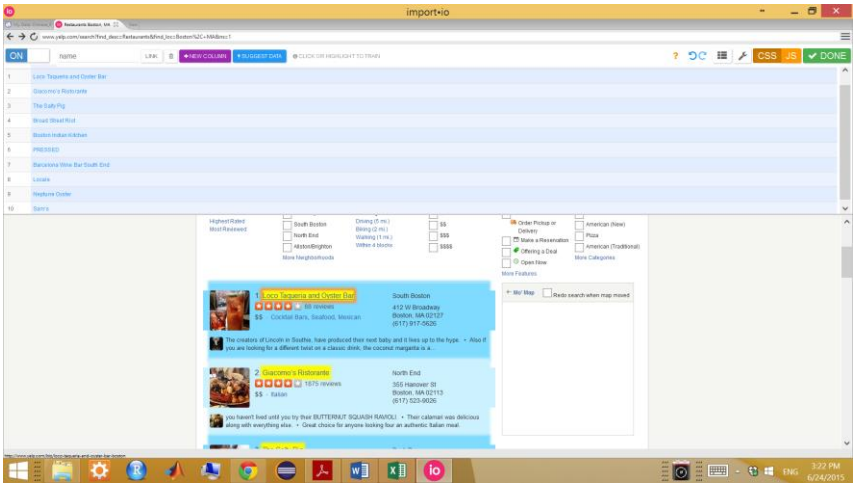
1) Starting new Extractor



2) Pasting and loading website
   It is needed to paste the website link in the URL bar and then the software loads the page as well as does analysis on its' html codes.

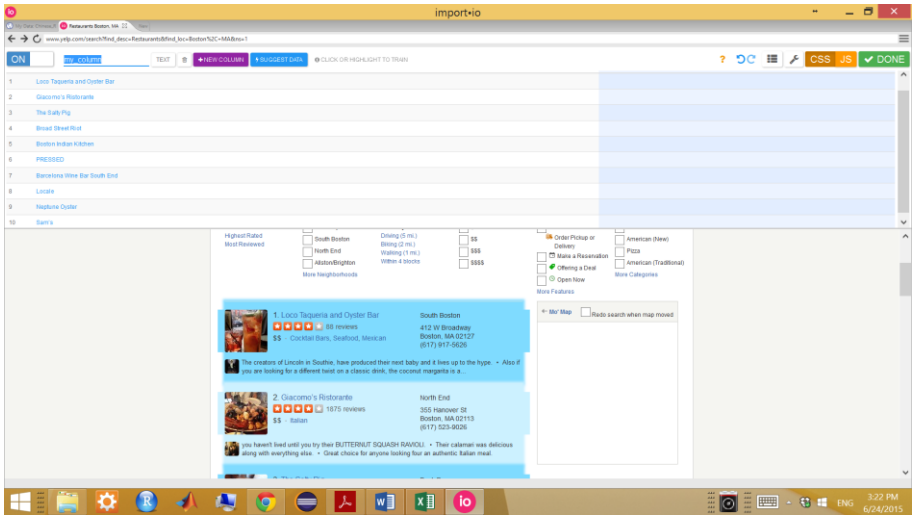


3) Data selection
   Now we can simply select the data that we are intending to extract. Selected data are being shown with yellow background.



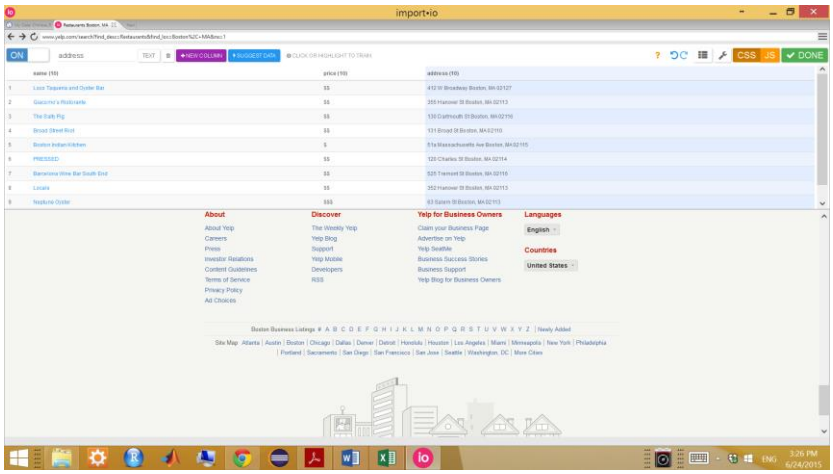
4) New Column

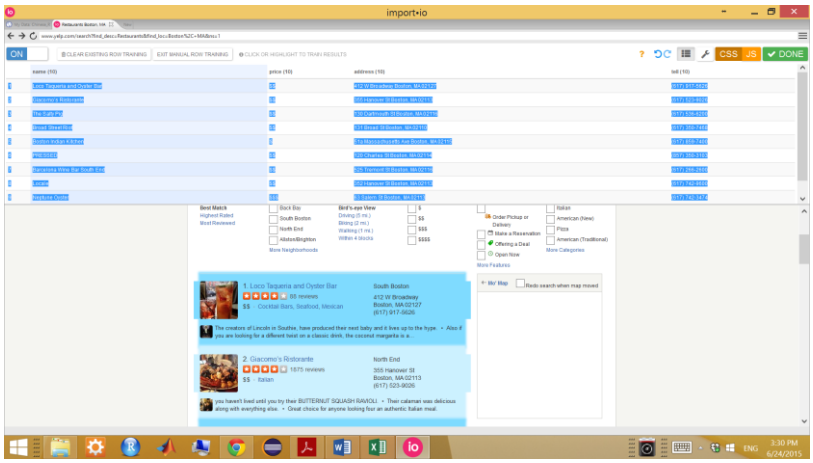
Here we can add new column by clicking on NEW COLUMN and specify the column name and then do the same procedure in 3.



5) Final Step:

Now the whole data is extracted as a table, we can simply copy and paste the data into an Excel file. Note: This software crashes a lot and copy and paste is the fast and more reliable than the automatic extraction that this software has provided.



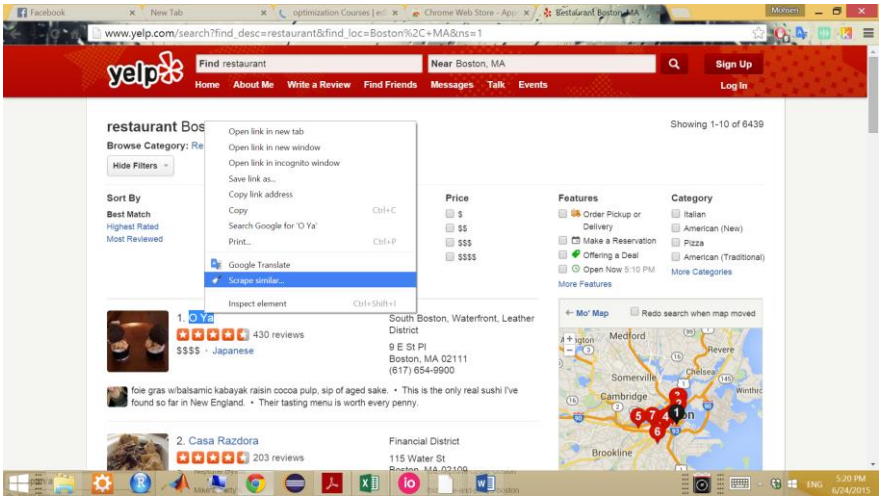


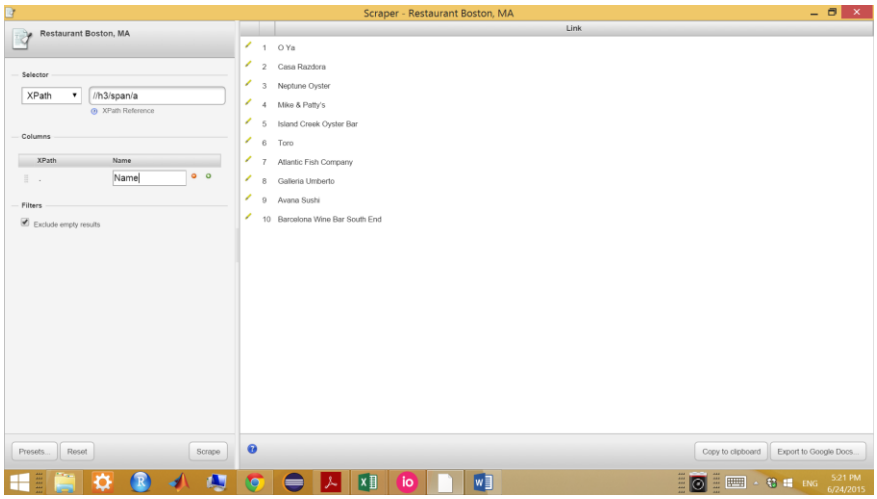
c) Scraper

Scraper is an embedded tool in Google Chrome. It is needed to download the Scraper extension to google chrome. Here are the steps for data parsing:

1) Choosing the data

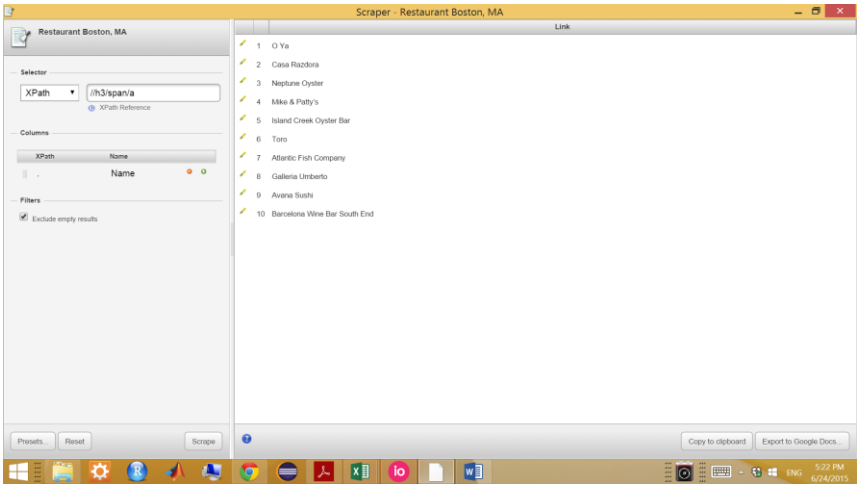
   Once you are in the webpage with chrome browser, just highlight the text you need to extract and “right-click” and then press “scraper”. Scraper will automatically finds a right pattern to capture similar data and will outputs a vector as shown below.





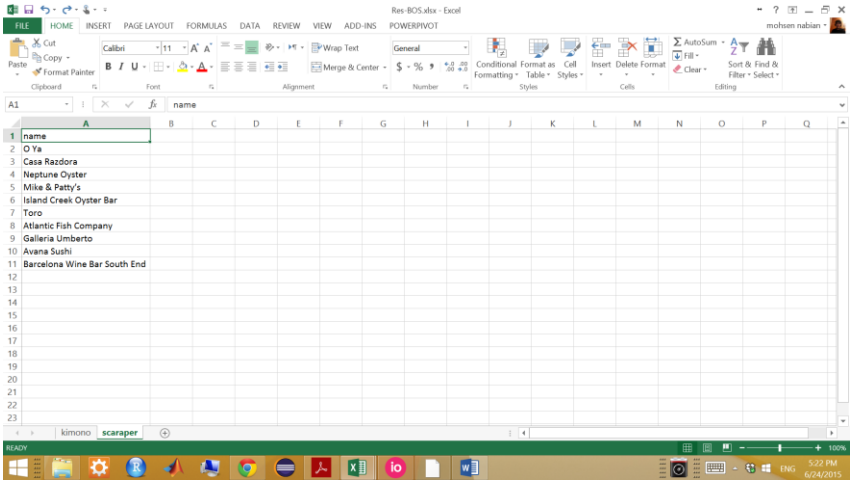
2) Modify extracted data:

In the next appearing page, the selected data is shown and user can remove trash data as well as to manipulate the xpath to obtain the right data. By clicking Scrape button, changes will be applied.



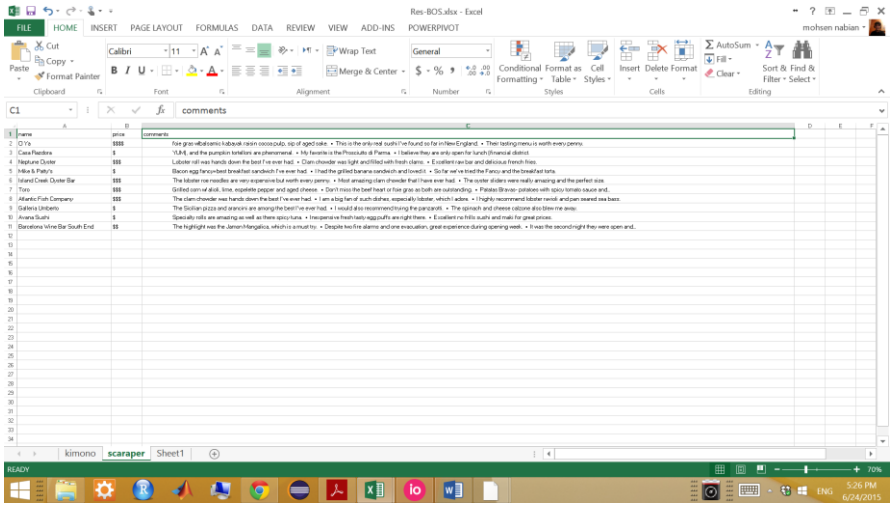
3) Extracting to Excel:

Finally we can copy and paste the data into an excel file for future data analysis.



4) Other Columns:

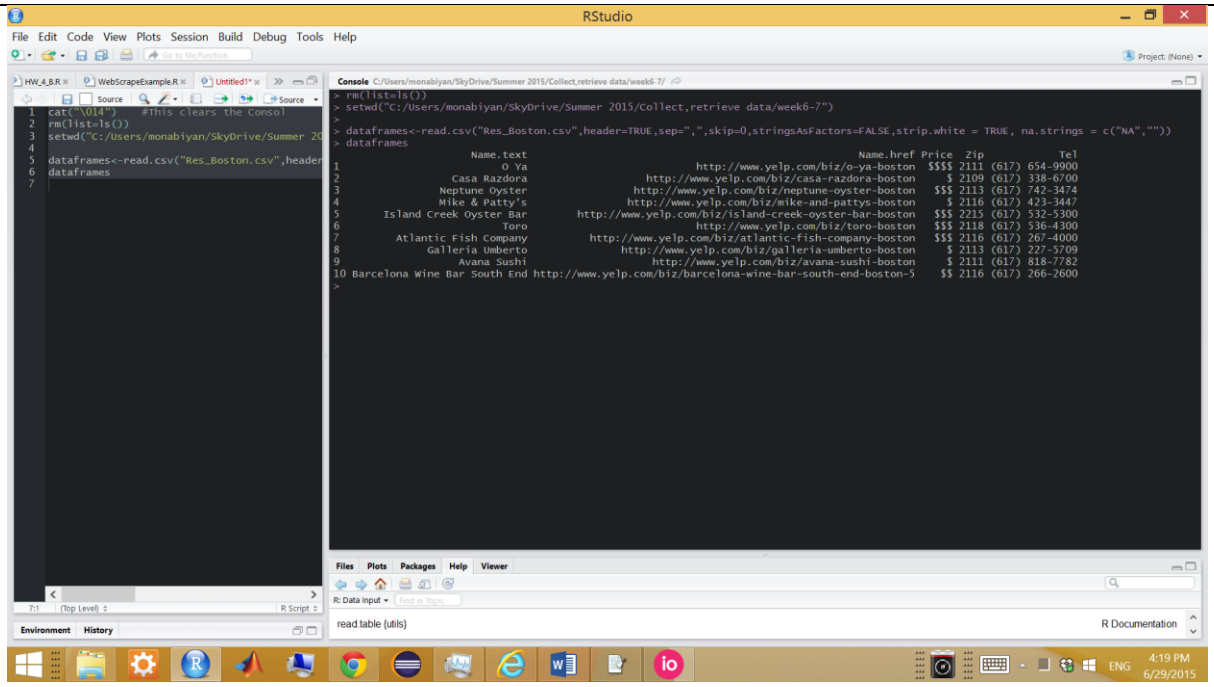
We will do the same steps for EACH column separately and will paste it in the same excel file to have them bind together for analysis.



Importing Data in R:

Now we can import the data in the excel file into R with some coding:

```
setwd("C:/Users/monabiyan/SkyDrive/Summer 2015/Collect,retrieve data/week6-7")
dataframes<-read.csv("Res_Boston.csv",header=TRUE,sep=",",skip=0,stringsAsFactors=FALSE,strip.white = TRUE, na.strings = c("NA",""))
dataframes
```



Discussion:

We extracted similar data from 3 web scraping toolkits. Here we intend to provide pros and cons of each:

I) Kimono:

Pros:

1) Is an extension to chrome browsers. Easy to install and very accessible everywhere if you have a signed in chrome.

2**) Visually Attractive**. It highlights all the similar elements.

3) **Very Easy to use** when for creating a column or adding another column

4) **Robustness:** User can improve the automatic searching by adding or deleting elements.

5) It has **free** plan as well as **enterprise** plan for more advanced features.

6) Updating Feature: You can update the data extraction some periods of time to have updated data.

Cons:

1) you do not have access to xpath as you have in Scrapy.

II) import.io

Pros:

1) It has little bit smarter parsing algorithms comparing to Kimono and captures data quite accurately.

It automatically can detect tables and simply extract them without choosing the data.

2) It has several features for different web sites to make data extraction easier. Some of the Features are "Magic" (automatically parse data without any user interference) , "Extractor" , "Classic Extractor" (later version) ,….

It is free for many features and **not** free for some more advanced features and high volume data.

Cons:

1) Should be installed as an external software on the hard drive.

2) It is very slow sometimes.

3) It crashes sometimes when you try to export **the data.**

4) Not able to modify searching pattern.

5) No access to xpath

II) Scrapy

Pros:

1) Is an extension to chrome browsers. Easy to install and very accessible everywhere if you have a signed in chrome..
2) Very easy to use: You will right-click on what you want to Scrape.
3) User has access to xpath filtering address.
4) It is Free.

Cons:

1) you are not able to choose all vectors and must do it one by one and store it one by one. However, it is very easy to use.

2) Not as visually attractive as Kimono

3) There is no scheduling for data updates.

Summary and conclusion:

My preference order is 1)Kimono 2)Scrapy 3)import.io

All these 3 software have very useful free features, however Kimono and import.io have some paid advanced features for enterprises. Kimono and Scrapy are working on chrome browser however import.io runs separately as individual software. Import.io is slow and crashes a lot. Kimono is the best since you can improve the filtering, You can make scheduling to update data. Kimono is fast, very easy to use and visually attractive. Scrapy is also very responsive, easy to use and provides access to xpath syntax. However, you can only choose one column at a time and also is not as visually attractive as Kimono.