

## Assignment 2. Web Scraping with Google Spreadsheet and XPath

Due Monday, October 16th, at 23:59 EST

Worth 15 points

Using your existing teams, harvest Resto MTL restaurant data. Map it. Note that there is a considerable variety in the types of data you can harvest. Decide on what it is you want to show on your maps and create a story of why you chose it.

To give a sense of the variety and volatility of scrapable data, it is possible that restomtl may begin to reject our attempts to scrape it (as we discussed in class) if this occurs, please contact Prof. Sieber and Ian ASAP and we will provide an alternate site to scrape.

### **Goals**

Learn that geospatial data can come from anywhere, even the content of an HTML page.

Learn how to combine web-based tools to extract and then structure unstructured content

Learn how to perform web-based geocoding

Gain experience in XPath, which is a standard for the W3 to navigate XML, HTML, and (even KML!) documents


Become acquainted with the XPATH content in W3 Schools (e.g., [http://www.w3schools.com/xpath/xpath\\_intro.asp](http://www.w3schools.com/xpath/xpath_intro.asp)).

Individually reflect on the assignment

### **Tasks**

1. Set up Google Spreadsheets to automatically harvest information from <https://restomontreal.ca> . Go to <https://www.restomontreal.ca/restaurants/> and choose either **a neighborhood (Quartier)** or **a Cuisine** from (cuisines are the second section on this page; neighborhoods are further down the page so just keep scrolling). Copy the URL from your chosen quartier or cuisine page and paste it in Google Spreadsheets.
2. As you will see in class, a major challenge of web scraping is the structuring of unstructured information. Develop a strategy to format the restaurant data into a usable and comparable form. For example, one of your XPath extractions will produce multiple rows or columns of


data for each listing. How can these be standardized? Is there common data between all listings that can be compared, or is each listing unique? What parts of unstructured data (e.g., in the description and bullet points) may be interesting? **Your report should serve as a manual for scraping and the reader should be able to replicate the process from your text.**

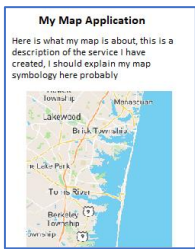
- a. Pay special attention to the geographic information on the pages. Address data in RestoMTL may vary can display as mashed together with cities and postal codes (e.g., “21 and 23 The Ridgeway, London, ON, N6C 1A2”; “Guildwood Boulevard, London N6H5G2 ON”). How would will you best structure this data in a Google Spreadsheet?
- b. Describe step-by-step the process that you propose to structure the data (one/team).
- c. Extract the data using XPath in Google Spreadsheets. You will need to make use of more operators than the ones shown in class, for example, the | operator or the **starts-with** operator.
- d. The results in your Google Spreadsheet with its importXML will change (refresh) each time it is opened. So, right before you submit your assignment, take a snapshot of the spreadsheet that clearly shows the various XPath commands .
- e. Describe the step-by-step process of extracting the data from the HTML pages (1-2 pages/team, which could include a diagram).
- f. Your steps should be structured and explained alongside your data. See below as an example:

	D	E
	Step 1:	Describe this step
	Step 2:	Describe this step
	Step 3:	Describe this step

3. Use Google Spreadsheets to extract and then classify the restaurant ratings and average prices (e.g., \$-\$\$\$) by value. Use XPath to extract the rating and price figures and Google spreadsheets to convert them into numbers. Then use standard spreadsheet functionality to classify the values.

**NOTE:** It is possible that you may find Lat and Long coordinates of the items on the pages you are scraping. You are not permitted to scrape these. The purpose of this assignment is to geocode the addresses you are scraping into coordinates within the spreadsheet.

4. Create a map using Mapbox Studio to display your data as a points file. (Think back to Assignment #1.)
  - a. Use an online batch geocoding tool (Google has one, but there are many others) to generate the lat/longs. 
  - b. Download your spreadsheet as csv and use a tool to convert it to geojson. That file should contain the points, whose map symbols will later be coded by their determined values. The actual values and other descriptive information should be stored using custom variables as **properties** of your geojson. This structure will enable you to customize your points in Mapbox Studio.
  - c. Upload the geojson to Mapbox studio using the same method as Assignment 1.
  - d. In the icon panel of Mapbox Studio, use the option “style across data range” and your price or rating points as the variable. You should use a scale of icons that are easy to interpret.
  - e. A title must be displayed for each location, using the text panel.
  - f. Describe the process of creating the geojson file from the data in Google Spreadsheets and the online geocoding tool (~0.5pg/team)
  
5. Using mapbox GL JS examples (<https://docs.mapbox.com/mapbox-gl-js/example/simple-map/>) and your knowledge of how to use custom styles in a Mapbox webmap from Assignment 1, create a simple HTML webpage that embeds your Mapbox studio style from Step 4. Yes, you’re mashing up Mapbox into HTML.



- a. Your HTML page must include a custom title, and some plain text on the page that explains what the user is looking at. This means that the map cannot take up the full screen of the website. You will have to change the size of the map div using css!
  - b. Upload the HTML and additional content to your group server on neogeoweb.ca. It would be a good idea to put the content inside a folder for assignment 2 and move your files from assignment 1 to another folder. Your result will look something like: <https://neogeoweb.ca/group2/assignment2/index.html>
  
6. Take a section of HTML code from your data source page on restomtl as referenced by your most complicated XPATH command and create a DOM tree out of it. Include the DOM tree along with a brief description in your report.
  
7. Note: as much as possible these processes should be automated. You should minimize the amount of manual data entry or manipulation.
  
8. Individually (i.e., each person) reflect on the assignment (0.5 pg/person)

- Reflect on the discussions of legal issues in the class, from the 2 articles, and from any other useful material. If a publicly accessible and free-to-view site like Zillow, RestoMontreal or Kijiji is used to for advertising/promoting items, should anyone be able to gather and use that data for other purposes? Why/why not? Cite as necessary and provide a reference list (any citation style can be used as long as it is consistent across all submissions in your group).
- Reflect on the accuracy of the geoparser you used, did it put the restaurants where they were supposed to go?

### ***Submission***

You have three submissions to make. Please read carefully.

Email [renee.sieber@mcgill.ca](mailto:renee.sieber@mcgill.ca) and [sichen.wan@mail.mcgill.ca](mailto:sichen.wan@mail.mcgill.ca) your report as a single .docx per team, including the screenshots and individual reflections.

Share your Google Spreadsheet with [resieber@gmail.com](mailto:resieber@gmail.com) and

Send the GeoJSON to [renee.sieber@mcgill.ca](mailto:renee.sieber@mcgill.ca) and [sichen.wann@gmail.com](mailto:sichen.wann@gmail.com)