Assignment 2. Web Scraping with Google Spreadsheet and XPath

Due Wednesday, Feb. 20, at 11:59 EST

Worth 15 points

Using your existing teams, harvest Kijiji housing data. Map it. Note that there is a considerable variety in the types of housing data you can harvest. Create a story of why you're mapping what you're mapping.

Goals

Learn how to combine web-based tools to extract and then structure relatively 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. XPath operates in a similar way to SQL queries that you may be familiar with from the query builder in ArcGIS, although XPath uses a specific syntax that has similarities to html (http://www.w3schools.com/xpath/xpath_syntax.asp).

Become acquainted with W3 Schools (e.g.,http://www.w3schools.com/xpath/xpath_intro.asp). W3 is the standards body for the World Wide Web Consortium.

Individually reflect on the assignment

Tasks

1. Set up Google Spreadsheets to automatically harvest information from Kijiji. Choose a city. From that Kijiji home page, go to the 'Real Estate' listings (we'll entertain other categories if they include addresses). Copy the link from the URL and paste it in Google Spreadsheets. Remember that this choice will eventually be part of your smart city portal so you may wish to take that into consideration.

   Note: Ordinarily, scraping tools include what are called crawlers or spiders that search the Internet on your behalf to find websites and THEN scrape the pertinent content. We're simulating that part of the operation by pointing you to pages that already contain many URLs. Your goal is to automatically harvest those URLs and then search for the pertinent content within those URLS. This is a two-step process: use importXML to automatically obtain those URLs and then use importXML to scrape content from each of those URLs.

2. As you are seeing in class, a major challenge of web scraping is the structuring of unstructured information. Develop a strategy to format the housing 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. It’s very messy on a sheet. How can these be better displayed better? Is there common data between all listings that can be compared, or is each listing unique? What parts of unstructured data may be interesting? **Your report should serve as a guide for scraping this type of content. It also should serve as a manual so the reader can replicate the process from your text.**

   a. Pay special attention to the geographic information on the pages. Address data in Kijiji can vary in the level of detail (e.g., “21 and 23 The Ridgeway, London, ON, N6C 1A2”; “Guildwood Boulevard, London N6H5G2 ON”). Your goal is to capture as much address information as possible and geocode every single listing. Many listings already have lat/longs. The goal here is to harvest the address data and NOT the lat/longs. Why not do it the easy way and just map the lat/longs? Because not all sites supply lat/longs.
   b. Describe step-by-step the process that you propose to structure the data you are extracting. Address the above questions in this step-by-step process (1-2 pages/team, which could include a diagram).
   c. Extract the data using Xpath in Google Spreadsheets. You will need to make use of several operators, for example, the **|** operator or the **starts-with** operator. Good marks will depend on how automated the process is and how structured you can make the results.
   d. Every time that you open this spreadsheet, the list of rental units will likely change (why is that?). So, right before you submit your assignment, take a screenshot of the Spreadsheet that clearly shows the various XPath commands 📷.

3. Use Google Spreadsheets to extract and then classify the rental prices by value, as well as another variable of your choice (such as number of rooms). If you choose a variable other than price then there should be sufficient variation in this values to make several categories visually discernable on your map (i.e., Price range of High/Med/Low). Use XPath to extract the dollar amounts and Google spreadsheets to convert the text to numbers. Then use spreadsheet functionality to classify the values. 📷

4. Create a map in Google MyMaps of all apartment listings that you harvested. Do not import your Google Sheets directly into MyMaps, instead:

   a. Use an online batch geocoding tool (Google has one, but there are many others) to create the lat/longs. 📷
   b. Create a KML file that contains the points, whose symbols are colour coded by rental value. The actual rental price and other descriptive information should be stored in the info window in KML. Don’t forget to comment your code.
   c. Describe the process of creating the KML file from the data in Google Spreadsheets and the online geocoding tool (~0.5pg/team)
   d. Upload the KML to Google MyMaps. 📷

5. Take a section of HTML used in your most complicated XPATH command and create a DOM tree out of it (cf., http://www.w3schools.com/xml/dom_intro.asp).

6. Note: as much as possible these processes should be automated, which in this case means contained within the Google Sheets environment. You should minimize the amount of human intervention (manual data entry or manipulation of individual sites).

7. Individually (i.e., each person) reflect on the assignment (0.5 pg/person)

   a. What other sites (excluding Kijiji or Craigslist) could you have scraped that could aid a smart city initiative? Each of you should find three sites. Spend some time examining web pages and their code structure. Discuss what information the sites provide.
   b. Reflecting on the discussions of legal issues that we will cover on Tuesday, if a publicly accessible and free-to-listers site like Kijiji or Craigslist is used to advertise an apartment for rent, should anyone be able to gather and use that data for other purposes? Why/why not? Note the website(s) or articles that informed your response (not Wikipedia, please).

Submission

You have three submissions

1. As an individual, email us your report as a single .docx/team, including the screenshots (renee.sieber@mcgill.ca and sam.lumley@mail.mcgill.ca) and individual reflections.

2. As a group, share your Google Spreadsheet with resieber@gmail.com and zhengzhibin.allen@gmail.com.

3. As a group, send the KML to renee.sieber@mcgill.ca and sam.lumley@mail.mcgill.ca.