Hello everyone,

So I’m looking for some help on three fronts.

One, initial web scrpe of all URLs; two, text scraping of certain fields off the URLs organization in the table columns below; and lastly, suggestions for text analysis and colorations pattern recognition.

First, The text scrapping of all the URLs of federal departments “Forward Regulatory Plans”. Most are over the next two years. Here is the main URL in which all the trees fall under.

* <https://www.canada.ca/en/treasury-board-secretariat/services/federal-regulatory-management/government-wide-forward-regulatory-plans.html>

Normally, I would just scrape all the URLs on the page and then be able to pull the main text fields I’m looking to do some analysis on; however, the sub-trees seem to have no consistency and I’m hoping there are faster ways than manually collecting over 41 websites and their sub-trees. I have been using a free version of Import.io which has been helpful and has a decent auto-script for learning some of website formatting in order to scrape the text I’m trying to pull but still seem to be struggling.

There is a new common look and feel thanks to the Web standards by the federal government.

One scenario: Using Canadian Border Services Agency (hyperlinks are embedded in the below

The main fields I need to dig down into are the following: The Yellow items are for every proposed regulation.

|  |  |  |
| --- | --- | --- |
| **Field title** | **Information type** | **URL where applicable** |
| Federal Department | Text |  |
| Proposed Regulatory Initiative | Text |  |
| Regulations made under “X” Act | Text |  |
| Description of the objective | Text |  |
| Indication of business impacts | Text |  |
| Public consultation opportunities | Text |  |
| Departmental contact | Text |  |

There are not a significant number of fields but I hope to gather this all into a master CSV format for text analysis. If you have suggested to analysis technique are will to take a crack at this I would greatly appreciate it. The final will be the data visualization work assuming we see any patterns.

My goal is to identify common objectives, similar business impacts but also if there is any overlap/opportunities to collaborate when doing public consultations with the same stakeholders (citizens).

Very appreciative thanks to anyone willing to take this on with me,

Greg