Assignment 4 – Scraping Web Pages

StackOverflow's R Questions

STA141B Spring 2023

Professor Duncan Temple Lang

Due: 10pm Friday, June 2nd 2023

Submit via Canvas

As usual, please ask any questions on Piazza to clarify what you are expected to do for this assignment.

In assignment 3, we explored the stats.stackexchange database. These data came from dumps of the StackExchange databases as large XML files. In this assignment, we will see how, if necessary, we could programmatically extract similar data from the stackexchange/stackoverflow Web pages, if we had to.

We will programmatically scrape questions, answers, comments and metadata from StackOverflow, specifically questions about R, starting at the URL

`https://stackoverflow.com/questions/tagged/r`

- We start at this first page of the search results on stackoverflow.
- We get the links to the questions on this page.
- We then get the next page of results and the links to the questions on that.
- We will process the questions on the first 3 and last page of results of the search results, fetching 50 results/questions per page.

Explore the source of the HTML pages for the search results and one or more questions and find HTML structures identifying the elements of interest described below.

For each question in these 4 pages of results, you will get the following information:

- the number of views of the question
- the number of votes
- the text of the question
- the tags for the question
- when the question was posted
- the user/display name of the person posting the question
- their reputation
- how many gold, silver and bronze badges they have
- who edited the question and when
- each answer
  - the text
  - the person who posted
  - when they posted
  - their reputation and badge information
  - all of the comments on this answer
    - the text of the comment
    - who posted the comment
    - when they posted the comment
- and any other relevant information about the question, answers and comments.

You should write functions to

- read a page of results from the search query
  - find the links to the Web page for each question and answer
- find the URL for the next page of query results
- read a question page, extracting the question, answer and comments
- process the 4 pages of search results of interest
- assemble the data into one or more data.frames
- and many helper functions to extract the different elements.

Plan and describe the data structure(s) you plan to create.

Use XPath to find the relevant content in each page.

Feel free to use the XML, xml2 or any similar package to parse and process the HTML pages. And you can use R's own functions, curl2, httr, RCurl or any package to make any HTTP requests.