INF 551 – Spring 2018

Homework #2: XML, HTML, and XPath

Due: February 26, Monday (end of day)
100 points

1. [50 points] In this question, you are provided with a (sample) of data set about the parks and recreation facilities in Los Angeles. The data set is in the TSV (tab-separated values) format.
   a. [20 points] Write a Python script "convert.py" that coverts the data set into an XML file with pretty print. The root tag shoud be <*parks*> and the tag for each park should be <*park*>.
   Example execution:

   ```
   python covert.py parks.tsv parks.xml
   ```

   where parks.tsv is the provided data set and parks.xml is the output XML file name.

   Note that your script will be tested using similar data sets (with the same format).

   b. [30 points] Write a Python script "stats.py" that takes the XML file you produced in the previous question and outputs the number of parks and recreation facilities by their types (Parks, Tennis Courts, etc. See LocationType column) that appears in the file. Sort the output by the ascending order of location types, one line per type. Use tab to separate type and count.

   Example execution:

   ```
   python stats.py parks.xml
   ```

   Example output:

   ```
    Parks	25
    Tennis Courts	3
   …
   ```

2. [50 points] In this question, we consider using **XPath** to extract data from Web pages. In particular, we consider search result pages from Amazon about books. For example, the screenshot below shows a list of books returned by Amazon on keyword search "data mining".

   Write a script "extract.py" that takes a search result page (e.g., "result.html" below) that contains a list of books and extracts the title, publication date, and authors of the books. The script should return an XML file with the extracted information. The file should follow this format:

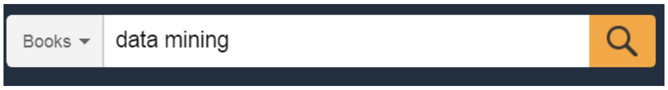
   ```
   <books>
         <book>
               <title>…</title>
               <publication_date>…</publication_date>
               <author>…</author>
   ```

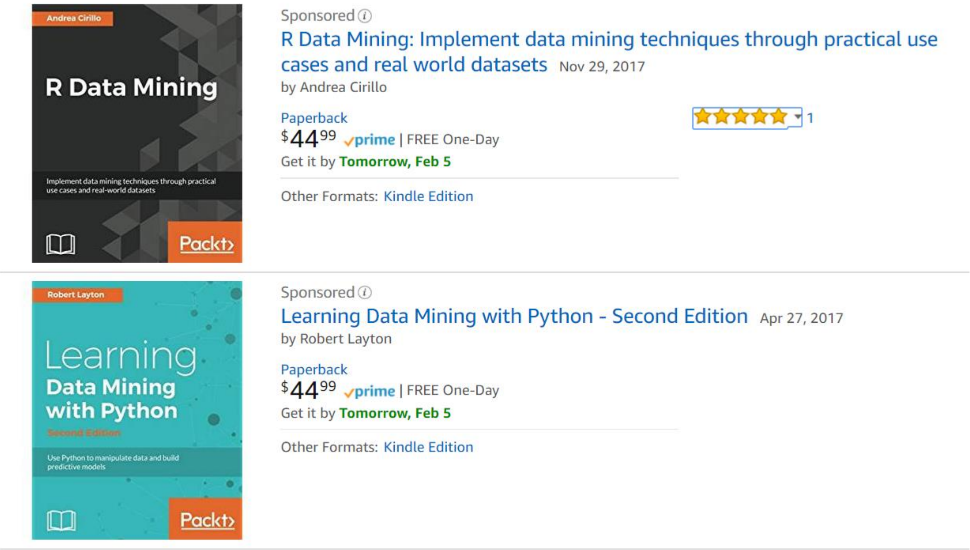
```
                <author>…</author>
                …

            </book>
            …

        </books>
```

Example execution:

python extract.py result1.html result1.xml

Make sure you select “Books” as the category when you submit the keyword searches to obtain some additional result pages for your own testing.




Andrea Cirillo
R Data Mining
Implement data mining techniques through practical use cases and real-world datasets
Packt>
Robert Layton
Learning Data Mining with Python
Second Edition
Use Python to manipulate data and build predictive models
Packt>

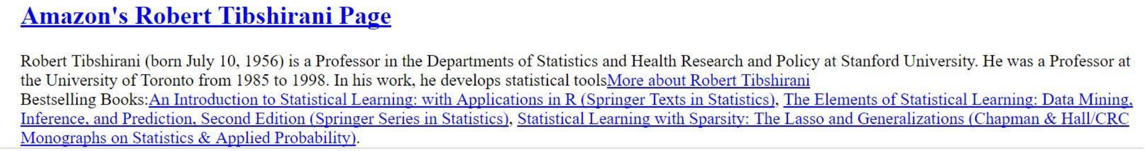
Sponsored ⓘ
R Data Mining: Implement data mining techniques through practical use cases and real world datasets Nov 29, 2017
by Andrea Cirillo
Paperback
$\$44^{99}$ ✓prime | FREE One-Day
Get it by Tomorrow, Feb 5
Other Formats: Kindle Edition
★★★★★ 1

Sponsored ⓘ
Learning Data Mining with Python - Second Edition Apr 27, 2017
by Robert Layton
Paperback
$\$44^{99}$ ✓prime | FREE One-Day
Get it by Tomorrow, Feb 5
Other Formats: Kindle Edition

<u>Hints for q2</u>:

Set *encoding* option as *“utf8”* in open() if you encounter *UnicodeDecodeError* when opening html file. Use etree.HTML() instead of fromstring() for html files to avoid *lxml.etree.XMLSyntaxError*.

The non-book items should not be in the result, e.g., author home page as shown below.



Amazon's Robert Tibshirani Page

Robert Tibshirani (born July 10, 1956) is a Professor in the Departments of Statistics and Health Research and Policy at Stanford University. He was a Professor at the University of Toronto from 1985 to 1998. In his work, he develops statistical toolsMore about Robert Tibshirani
Bestselling Books:An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics), Statistical Learning with Sparsity: The Lasso and Generalizations (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).

Notes: You may use library lxml and Python Standard Library in this homework. But no other libraries may be used.

Submissions: Name your scripts as above and then prepend your name into the script name. For example, John_Smith_xyz.py. Submit your work to Blackboard by the due time. **DO NOT** place them in a folder or zip file.

Note: Please use Python 2.7 (installed by default on EC2) for the coursework.