UTBM, Spring 21

Lab session 7

DS54

Gathering Web Data

1. Data gathering web data using cURL

One of the most important tools for gathering data from the web is cURL. It's a command line utility that allows you to request and download a webpage or other data over HTTP, like images or documents. cURL is often scripted to scrape web data from various sources or used as part of a web crawler to index pages.

Bring up the webpage, top5ofanything.com and look at a list of data that contains the Top 5 Most Reliable Car Brands for 2021 : https://top5ofanything.com/list/5b6d4671/The-Top-5-Most-Reliable-Car-Brands-for-2021

In a folder called curl-demo, run the command curl (Use sudo apt install curl to install it if required).

direct it into a file and save it (which file type should you use?)

use curl --head and then the URL that shows header information. The important thing is that you shuld get a status code 200 as a valid request

to save on bandwidth if you are downloading a lot of data (a lot of pages), you can use the compressed option : use the command, curl --compressed –head and then the URL. the content encoding should come back as gzip

to search an element, you can use the following command: curl –data ‘’q=France’’ and then the url/search that it's posting to (in double quotes, you give the form element q that is the search element). This is similar to top5ofanything.com/search on the browser, you will get the same results from that posted data (to check).

executes the command, curl --data-urlencode "q=highest mountains" https://top5ofanything.com/search/ and comments the option urlencode.

cURL provides cookie handling. It’s information that gets passed back and forth between the browser and the web server. To store the cookies in file cookies.txt, check the following command:

curl --cookie-jar cookies.txt --output output.html https://www.google.com

examines the file cookies.txt (using command cat or more).

2. Extracting Spreadsheet Data with in2csv

Having data in Microsoft Excel format is very common, but this is not always a good format for data processing. We'd like tabular data like CSV.

use in2csv to convert an Excel spreadsheet data to CSV format. In a folder in2csv-demo, get an excel file , e.g., https://www.itu.int/ITU-D/ict/statistics/material/excel/EstimatedInternetUsers00-09.xls, which is Estimated based on urban/rural distribution of Internet users. Then use in2csv which is a

Python utility that comes part of the CSV kit library from Python (install csvkit first using the pip package manager).

pipe the command into grep, to do grep for France.

Save in a csv file and check it.

3. Extracting Spreadsheet Data with Agate

If we use Python for scripting, then extracting spreadsheet data is best done with the Agate library. Agate is a general-purpose data analysis library that can be used for data wrangling and other data science tasks.

Write out a script xls2csv.py on extracting the spreadsheet data with the Python Agate library.

4. Extracting HTML Data using Python and BeautifulSoup

If your data science needs require that you extract web data from HTML documents, you'll need to be able to parse and extract HTML tags from documents.

Write out a script using the Python package BeautifulSoup to download and extract HTML tags from the web.

5. work with metadata in email headers

write out a script parse-email.py to parse a source file from your email server using a Python library for email message parsing.

6. Connecting to Remote Data

to connect directly to a remote server that hosts your data, the most frequently used utility to connect to remote systems is Secure Shell, or SSH.

On a trusted computer, you can copy your SSH keys to the remote computer so that it trusts you implicitly (without password prompt except when you install it first time). Check the usage of the command ssh-copy-id to have a passwordless connection that uses a key-based authentication.

7. Copying Remote Data

To copy remote data using a secure copy, the command secure copy or SCP for short is used to copy a remote file securely using key-based authentication.

Explain how scp can be used with key-based authentication

8. Synchronizing Remote Data

Keeping remote data synchronized between devices, for example from a server to a single workstation can be complex. You can use rsync for synchronizing remote data (with ssh).

You could be using rsync between data in directories on a single computer. For instance, you can synchronize copies across different mapped drives or storage media.

Check rsync usage and answer this question: What is difference between scp and rsync?