Washington State University Vancouver

Web Data - CS 454

Assignment 1

Professor:
Ben McCamish

Overall Assignment

The goal of this assignment is to familiarize yourself with web scraping using Python 3. To ensure that all students are scraping available data and not spamming the servers, always practice safe scraping techniques. You may use Wikipedia's API if you wish. Once you have collected the data, it must be stored in a database. You can store the information retrieved in flat files such as CSV, XML, JSON, etc.

Feel free to come show me some data partway through the assignment to make sure that the amount collected is acceptable. Wikipedia is not the only option. You are encouraged to look into other data sources that you can scrape. Some might be behind a deep web and have an API you can use. However, these often contain the most interesting data points. Pick something that is of personal interest to you and you might find the assignment enjoyable!

Some Helpful Wikipedia API Lines

- **Wikipedia API Help Pages:** https://www.mediawiki.org/wiki/MediaWiki
- **Base Wikipedia API:** https://en.wikipedia.org/w/api.php?
- **Search:** action=opensearch, search=yousearchterms, limit=limitofitems
- **Format of Returned Results:** format=xml
- **Query For Specific Information:** action=query, titles=title1—title2—title3
- **Properties:** prop=extracts (extracts the text on the page), many others
- **Example:** https://en.wikipedia.org/w/api.php?action=opensearch&format=xml&search=Computer&limit=5

What you need to do

First you need to create a program that will search the Wikipedia (or your choice of source) pages for information relevant to your type of database. For example, if I wanted to create a database containing information about famous Computer Scientists, I may use the search API and find the list of computer scientists and then visit each of their pages to collect information about them. This can include text, images, links, etc.

I encourage you to explore the Wikipedia API and find out different ways you can collect information from Wikipedia Pages. You will need a diverse database for your final project.

To Receive full marks you must:

- (10%) Specify what topics you are searching for in Wikipedia.
- (40%) Write a Python program to collect information on a specific topic. Your code needs to be well commented. Minimum of 2000 tuples required. Your data must contain some descriptive text on each tuple, an image (if applicable), a link to the original source (if applicable), and some unique identifier.
- (20%) Parse the XML/JSON/HTML files and store the information in a database of your choosing. This can be CSV files, a SQL database, XML files, etc. Again, your code for this should be well commented. **Note:** You cannot just store the returned XML files, some parsing needs to take place such that only the information you want from the page is being stored.
- (20%) A writeup, written in LaTeX, explaining why you chose the topic and what the goal of the database is. For example, if I had a database on coffee, I would explain that the database created from the crawled wikipedia pages contains information on different types of coffee beans, where they are grown, how they are roasted, some pictures to go along with each of those topics, etc. Be detailed so that you have your database contents well documented. Be sure to include how many tuples are in your database. There needs to be at least 2000.

What to turn in (in a zip on Canvas):

- All code (well commented)
- A sample of your database (an XML document, your SQL schema and a few tuples, a snippet of your CSV file, etc)
- A summary of your homework, both the LaTeXsource and a PDF.
- README.txt on how to run your code (detailed if required).