Developing Automated Internet Data-Collection Agents

A one-day seminar for academic researchers

Environment: Either Windows and VBScript or Linux and Perl

Prerequisites: Attendees should have some experience in computer programming and understand variables, loops, conditional statements, etc. Attendees need not have experience in the environment chosen for the course. Some basic understanding of HTML tags is helpful but not required.

Topics:

1. Hardware considerations: What kind of computer should be used? What will be the requirements for memory, processor speed, storage, etc.?
2. Legal implications: Become familiar with the current legal environment regarding automated online data collection. How should academic researchers behave and what should they avoid? Learn from the mistakes of others (and by the way, we've made some pretty big mistakes).
3. Writing a simple data collection agent to request a webpage: The first exercise is to develop a simple agent that requests a single webpage.
4. Parsing a document for data and links: Learn the procedures and techniques to scan and collect data from an HTML document. See how to extract URLs from hypertext links to govern additional data collection.
5. Writing the multi-page agent: Build an example agent that is able to retrieve a webpage, collect its data and then follow a link on the page to and repeat the process until a particular data set is collected.
6. Two-stage processing: Understand the techniques and methods for this approach to research data collection. Learn why it is important to academic researchers.
7. Simulating form submission: Following links is one thing, but getting an agent to submit a form requires a different skill set. Learn how to submit forms using both POST and GET methods.
8. Tips and tricks for understanding web servers: Reading the source of an HTML documents to determine the requirements for interacting with the scripts on a web server can be very challenging. Learn shortcuts that can dramatically simplify the process.
9. Automating longitudinal agents: Learn how to configure agents to collect data over an extended period (such as once a day for three months), without needing to leave the agent program running all the time.
10. Monitoring progress: When agents collect data over a long period of time, it is important to monitor their progress regularly to detect when changes in a website affect performance. Learn some techniques for making this task easy.