Web Scraping

If you are obtaining your corpus from the web (e.g., company and university mission statements, newspaper articles on specific events, book/music/film/restaurant reviews), you can copy and paste **by hand** documents, perhaps from different websites. However, **web scraping** may provide a more efficient solution.

Web scraping/crawling: A definition

Web scraping (web harvesting or web data extraction) is a computer software technique of extracting information from websites. This is accomplished by either directly implementing the Hypertext Transfer Protocol (on which the Web is based), or embedding a web browser.

Web indexing

Web indexing (or Internet indexing) refers to various methods for indexing the contents of a website or of the Internet as a whole so that it become available to search engines (e.g., Google). Individual websites or intranets typically use a back-of-the-book index; search engines prefer keywords and metadata as a more useful way of Internet or onsite searching.

Bots

A “bot” is short for robot, a computer software that crawls the web and performs automated tasks (e.g., data extraction).

Software options

There are a number of available freeware and commercial scrapers. A popular, **freeware** option is **OutWit Hub**. While the full version of OutWit Hub costs around $89, the freeware option will probably serve you well. You can download it at http://www.outwit.com/products/hub/.

Caveat!

Scraping requires knowledge of the data structure of each website where data are taken from. Scraping will be more efficient than human copy-and-paste if the documents to be scraped are stored under the same website (so that knowledge of only one type of data structure is required); otherwise, **you may be better off by copying and pasting by hand, depending upon how much data you are downloading and from how many websites.**

Legal issues

The legality of web scraping varies across the world. In general, web scraping may be against the terms of use of some websites, but the enforceability of these terms is unclear. Some websites require you to enter the content of an image, thus barring bots from crawling their data. Other websites limit the amount of data you can download in one go. **Make sure you understand the permission policy of each site you intend to scrape for data.**