Crawling the Web

COMS10012 / COMSM0085

Software Tools

wget

Basic usage

Download a file/resource:

```
$ wget <url>
```

Does exactly what you ask.

Perhaps not what you intend?

Page requisites

Download *webpage* with requisites:

```
$ wget -p <url>
```

Include: stylesheets, images

Also: `robots.txt` ?

ROBOTS.TXT

Standard for websites to announce crawling preferences.

Simple text file stored in top-level directory of a site.

Specifies which user-agents are permitted to access which resources.

Recursive downloading

Download *webpage and linked pages*:

```
$ wget -r -l N <url>
```

recursively

By default, constrained to only pages on the same domain as that in the webpage URL.

`-l N:` The level of recursion to permit. (Default if omitted is 5).

Mirroring

Download *entire website*:

wget -m -w 1 http://www.google.com

```
$ wget -m <url>
```

Uses standard defaults for creating a local copy of a full website. *(Think about if you really want the **entire** website).*

`-w 1:` Wait 1 second between each request (to avoid annoying a server).

Ethics

Concerns

- Does the site permit you to crawl this resource? (robots.txt)
- Is there a better way to get a copy?
- Are you allowed to republish downloaded content? (e.g., copyright)
- Are you going to make something more public than it should be?