Robots.txt Tutorial

Excerpts from:
http://www.searchengineworld.com/robots/robots_tutorial.htm
http://www.robotstxt.org/wc/faq.html

Search engines will look in your root domain for a special file named "robots.txt" (http://www.mydomain.com/robots.txt). The file tells the robot (spider) which files it may spider (download). This system is called, The Robots Exclusion Standard.

The format for the robots.txt file is special. It consists of records. Each record consists of two fields : a User-agent line and one or more Disallow: lines. The format is:

```
<Field> ":" <value>
```

The robots.txt file should be created in Unix line ender mode! Most good text editors will have a Unix mode or your FTP client *should* do the conversion for you. Do not attempt to use an HTML editor that does not specifically have a text mode to create a robots.txt file.

User-agent

The User-agent line specifies the robot. For example:

```
User-agent: googlebot
```

You may also use the wildcard character "*" to specify all robots:

```
User-agent: *
```

You can find user agent names in your own logs by checking for requests to robots.txt. Most major search engines have short names for their spiders.

Disallow

The second part of a record consists of Disallow: directive lines. These lines specify files and/or directories. For example, the following line instructs spiders that it can not download email.htm:

```
Disallow: email.htm
```

You may also specify directories:

```
Disallow: /cgi-bin/
```

Which would block spiders from your cgi-bin directory.

There is a wildcard nature to the Disallow directive. The standard dictates that /bob would disallow /bob.html and /bob/indes.html (both the file bob and files in the bob directory will not be indexed).

If you leave the Disallow line blank, it indicates that ALL files may be retrieved. At least one disallow line must be present for each User-agent directive to be correct. A completely empty Robots.txt file is the same as if it were not present.

White Space & Comments

Any line in the robots.txt that begins with # is considered to be a comment only. The standard allows for comments at the end of directive lines, but this is really bad style:

```
Disallow: bob #comment
```

Some spider will not interpret the above line correctly and instead will attempt to disallow "bob#comment". The moral is to place comments on lines by themselves.

White space at the beginning of a line is allowed, but not recommended.

```
 Disallow: bob #comment
```

Examples

The following allows all robots to visit all files because the wildcard "*" specifies all robots.

```
User-agent: *
Disallow:
```

This one keeps all robots out.

```
User-agent: *
Disallow: /
```

The next one bars all robots from the cgi-bin and images directories:

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
```

This one bans Roverdog from all files on the server:

```
User-agent: Roverdog
Disallow: /
```

This one bans keeps googlebot from getting at the cheese.htm file:

```
   User-agent: googlebot
   Disallow: cheese.htm
```

For more complex examples, try retrieving some of the robots.txt files from the big sites like CNN, or Looksmart.

A listing of known bots can be found here http://www.robotstxt.org/wc/active.html