Component I: Crawler/Spider/Robot

- Building a “toy crawler” is easy
  - Start with a set of “seed pages” in a priority queue
  - Fetch pages from the web
  - Parse fetched pages for hyperlinks; add them to the queue
  - Follow the hyperlinks in the queue
- A real crawler is much more complicated…
  - Robustness (server failure, trap, etc.)
  - Crawling courtesy (server load balance, robot exclusion, etc.)
  - Handling file types (images, PDF files, etc.)
  - URL extensions (cgi script, internal references, etc.)
  - Recognize redundant pages (identical and duplicates)
  - Discover “hidden” URLs (e.g., truncating a long URL)