Focused Crawling

Implement your own web crawler, with the following properties:

- Be polite and use a delay of at least one second between requests to the web server.
- You should start from the seed document `http://en.wikipedia.org/wiki/Gerard_Salton`, the Wikipedia article on Gerald Salton, an important early researcher in information retrieval.
- You should only follow links with the prefix `http://en.wikipedia.org/wiki/`. In other words, do not follow links to non-English articles or to non-Wikipedia pages.
- Do not follow links with a colon (:) in the rest of the URL.
- Do not follow links to the main page `http://en.wikipedia.org/wiki/Main_Page`.
- You may use existing libraries to request documents over HTTP.
- Otherwise, you should implement your own code to extract links, keep track of what you've crawled, and decide what to crawl next.
- Crawl to at most depth 3 from the seed page. In other words, you should retrieve the seed page, pages it links to, and pages those pages link to.
- Your crawler should take two arguments: the seed page and an optional "keyphrase" that must be present, in any combination of upper and lower case, on any page you crawl. If the keyphrase is not present, stop crawling. This is a very simple version of focused crawling, where the presence or absence of a single feature is used to determine whether a document is relevant.

Hand in your code and instructions on how to (compile and) run it. In addition, hand in two lists of URLs:

1. the pages crawled when the crawler is run with no keyphrase, in other words all Wikipedia pages meeting the requirements above to a depth of 3 from the starting seed; and
2. the pages crawled when the keyphrase is "information retrieval".