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/Hugh_of_Saint-Cher, the Wikipedia article on Hugh of Saint-Cher, one of the originators of a key information retrieval technology.
- 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. This will help filter out Wikipedia help and administration pages.
- 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, including following redirects that handle alternate article titles in Wikipedia.
- 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 5 from the seed page. In other words, you should retrieve the seed page, pages it links to, the pages those pages link to, the pages they link to, and the pages they link to. The seed page is thus not depth 0, but depth 1.
- Wikipedia has a lot of links, so you should also stop when you reach 1000 unique URLs.
- Your crawler should take two arguments: the seed page and an optional "key phrase" that must be present, in any combination of upper and lower case, on any page you crawl (after the seed). Don't worry about tokenization: just match the characters ignoring case. If the key phrase 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, each with at most 1000 entries:

1. The pages crawled when the crawler is run with no key phrase, in other words all Wikipedia pages meeting the requirements above to a depth of 5 from the starting seed; and
2. The pages crawled when the key phrase is ‘concordance’. (If you already did the crawl with ‘index’, that's OK, too.)