Final

Write a crawler to download all the pages discovered from seed URLs. You will be running this against a local repository of web pages so you do not need to worry about network programming or following a robots.txt file but you will have to hard code the directory separators for your particular OS. You will also not have to worry about dynamic pages with infinite link generation. Similarly, without access to the content-type header you will have to choose which types of files you do not wish to index based on file name extensions.

Your program should fully parse the links to crawl the pages. As each page is crawled it should be "downloaded" into another folder to be handed off to an indexer. You will not need to worry about scalability (as it should not take long to run an entire iteration) but be sure you are crawling as much of this web as you can automatically. It may be helpful to print out how many pages were crawled at the end of an iteration to give you a good idea of how well your code is following links and where to seed. Be careful to avoid following circular links and make your parsing as robust as possible in order to deal with all of the malformed HTML that exists.