Program-1: Image crawler

Write a program that performs limited image crawling. The program must take the URL of a website as command-line argument and download the HTML page pointed to by the given URL. The program must parse the HTML page, extract all the hyper links to PNG images, download these images and save the images to disk.

The images may also be encoded as base-64 format if the page uses data URI scheme. Even these images need to be saved by your program. More information on data URI scheme is available on Wikipedia page: http://*en.wikipedia.org/wiki/Data_URI_scheme*.

The sample page containing data URI encoding for images is available at:
http://www-archive.mozilla.org/quality/networking/testing/datatests.html

Program-2: File Downloader

Write a program to implement peer-to-peer file sharing. The program takes list of available peers as input through a text file. The format of the peers list is as shown below.

peers_list.txt
10.12.3.40
10.12.5.43

All the shared files are kept in a subdirectory called **Share**. Each peer maintains a list of shared files in the following format.

shared_list.txt
<filename>,<filesize>,<index>

The fields filename and file size are self-explanatory. The peering program generates an index for each file and stores the index in **shared_list.txt**. The index starts at zero and increases by one for each next file.

The program uses a peer protocol called BPGCP2P to communicate with the peers. The protocol uses the following message PDUs for communication.

PDU Name	**Intent**
Query	Search for a keyword in a file name
Query Hit	Indicate availability of file(s) with the keyword of interest
Download	Download the file of interest from one of the peers who sent a query hit
ACK	Generated by receiver at the end of successful file download

The PDU formats are shown below.

Query: BPGC\tQuery\r\n
QueryID: <random_number>
QueryWord: <keyword>

This query shall be sent to all the peers in **peers_list.txt**.

Query Hit: BPGC\tQueryHit\r\n
QueryID: <random_number>
QueryWord: <keyword>
\r\n
filename,filesize,index\r\n

There must be one entry in the query hit message body for each matching file.

Download: BPGC\tDownload\r\n
Index: <file index no>\r\n
FileName: <file name>\r\n
TCPPort: <port number>\r\n

ACK: BPGC\tACK\r\n
Index: <file index no>\r\n
FileName: <file name>\r\n

The peers exchange Query, Query Hit, Download and ACK PDU messages using UDP protocol on ports 30000 and 30001. Port 30000 is to be used to receive incoming PDUs and 30001 is to be used for sending outgoing PDUs. The file download happens on top of TCP. The port number for connecting to peer to send the file is specified in Download PDU. The sending peer may use any TCP port to send the file.