CSE205: Introduction to Networking

Project 1: Imagecrawler

Due date: Nov 27, 11:59 p.m.
Name: Kai-Yu Lu
Student ID: 1614649
Major: Computer Science and Technology (CST)

1. Implementation

- **a.** TCP sockets should be used to construct http requests and obtain replies form the website. Folders will be created correspondingly to the website.
- **b.** Regular expression for image and href will be built so that the sentence of image and href could be grasped
- **c.** All available images will be downloaded in to the relevant folders. Additionally, the images have the same name are considered as the duplicate images which will exist in the folder once.
- **d.** All href links will be accessed and the folders will be created again whose name is the same as the current website and the images inside the current href link will be downloaded as well.

2. Functions

1. graspData(currentHost, currentPath): Obtain all information of html.
2. imageFormat(myData): Use regular expression to get the image format in html.
3. getHref(myData): Use regular expression to get the href format in href.
4. storeImage(currentHost, currentPath, imageList, myData): This function is to get the format of image and then the image will be downloaded from the website to the relevant path.
5. createFolder(hrefList, takePath, currentPath): Create the paths for saving images.
6. createdepth(depth): It is the key to realize the depth function so that the images could be downloaded in the right folders (path).
7.

3. Limitations (bugs)

1. https cannot be realized. A function should be created for detecting format of https.
2. Thread has not been realized and speed could be improved.