**Explanation :**

Web Crawler is used to collect all the email addresses from the given web page. In this Lab we have programmed a Web Crawler to collect only distinct email addresses and remove duplicate email addresses.

Part1.) Here we have designed a Web Crawler to check distinct email links from ‘Lab6\_Webpage.html’ folder.

Part2.) Here we have designed a Web Crawler to check distinct email links from a sample URL – ‘http://e-mailid.blogspot.com/’. ([request@list.id.state.az.us](mailto:request@list.id.state.az.us) is found twice in the link, but by using WebCrawler and Regular Expression only unique id is displayed in our output)

**main() function:**

Here we get the two URL and stored it in url1 and url2 respectively.

urlopen() - Urllib package is the URL handling module for python. It is used to fetch URLs (Uniform Resource Locators). It uses the urlopen function and is able to fetch URLs using a variety of different protocols.

read() is used to read the web page consisting of urls in the entire file as a single string of text.

decode() is used to decodes the string using the codec registered for encoding. It defaults to the default string encoding.

document1 and document2 is used to store the decoded data for both part1 and part2 respectively.

emailListsHtml – is used to store unique email id’s from document1

emailListsUrl – is used to store unique email id’s from document2

**emails() function:**

This function is used to collect only unique emails from the given web page by using ‘regex’ as regular expression.