Information Retrieval and Web Analytics
Semester 2 - 2022

Practical Sheet 06

1. Implement a program to test if a given page is found or not on the server.

2. Python implementation to show the content of robot.txt for sliit.lk.

3. Python implementation to print first <h1> tag from
   https://en.wikipedia.org/wiki/Main_Page

4. How to print all the header tags from en.wikipedia.org/wiki/Main_Page.

5. How to display all the image links from https://www.ibm.com/lk-en.

6. Python implementation to retrieve all the links in a particular URL.

7. Python implementation to display section that contain a specified string in a given web page.

8. Python implementation to obtain html code of parent content of a particular html tag.

9. Python implementation to obtain html code of child nodes inside a particular html tag

10. Build a web scrapper using BeautifulSoup to extract (scrape) movie data(url, name, summary) from https://www.imdb.com/chart/top/ in to a excel file.