

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Programme | : | **B.Tech** | Semester | : | **Win Sem 22-23** |
| Course | : | **Web Mining Lab** | Code | : | **CSE3024** |
| Faculty | : | **Dr.Bhuvaneswari A** | Slot | : | **L3+L4** |
| Date | : | **19-12-2023** | Marks | : | **10 Points** |

**Exercise 2: Custom Search Engines / Crawlers**

|  |
| --- |
| 1. Implement crawlers which take as input a url, a search word and maximum number of pages to be searched and returns as output all the web pages it searched till it found the search word on a web page or return failure.    1. Breadth-First-Search    2. Depth-First-Search    3. Twitter Topic Focused Crawler – on Topic “FIFA 2022”(Challenging Task)    4. Identify the Robot exclusion protocol file in your preferred website. Robots Exclusion protocol is used to tell search engine crawlers which URLs it should NOT request when crawling a Web site. |

References

1. <https://www.geeksforgeeks.org/web-crawling-using-breadth-first-search-at-a-specified-depth/>
2. <https://github.com/Dineshs91/crawler>