

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Programme | : | **B.Tech** | Semester | : | **WIN Sem 22-23** |
| Course | : | **Web Mining Lab** | Code | : | **CSE3024** |
| Faculty | : | **Dr.Bhuvaneswari A** | Slot | : |  |
| Date | : | **12-12-2022** | Marks | : | **10 Points** |

**Exercise 1: Simple Web Crawlers/ Scrappers**

1. Given a seed/root URL, e.g., "Vit.ac.in", Design a simple crawler to return all pages (URLs) that contains a keyword "research" from this site. (25 pages)
2. Find documents that contain the word “admissions” and the word “international” within the URL “Vit.ac.in” using Python. (25 pages)
3. Find documents that contain the word “Programme” but not the word “programming” within the URL “Vit.ac.in” using Python. (5 pages)
4. Write a web crawler program which takes as input a url (Educational website) and a search key word and maximum number of pages (15-20 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.

[www.annauniv.edu](http://www.annauniv.edu)

hostel

16

1. Write a Python program to read the given website and extract the phone numbers and emails and contact addresses from Chennai, Amaravathi, Bhopal vit website. (5 Marks) (Challenging Task) Store it in a text file