**INF 558 : BUILDING KNOWLEDGE GRAPH**

**Homework 2: CRAWLING**

**Name**: KAVYA SETHURAMAN

**USC-ID:** 7852999061

**Task 1:**

**-Website URL**: [www.ebooks.com](http://www.ebooks.com)

**-Description**: Lists the ebooks based of different genres, with a description for each of them.

**Task 2:**

- **Sample webpage**: Figure 1

- **Interesting information**:

* Book Title.
* Author
* Summary of the book.

Figure 1:

|  |
| --- |
|  |

**Task 3:**

**Crawler Used:**

**SCRAPY:** Scrapy is a free and open source web crawling framework, written in Python. Originally designed for web scraping, it can also be used to extract data using APIs or as a general purpose web crawler

**SEED URL: ‘https://www.ebooks.com/subjects/body-mind-spirit/’**

**How did you manage to only collect the webpages respecting the template(s) in Task 2? How did you discard irrelevant pages?**

* Websites are traversed based on pagination rules.
* Pagination rules are set by the div class as:

|  |
| --- |
| * Rule(SgmlLinkExtractor(allow=(), restrict\_xpaths=(**'//div[@class= "paging bottom floatLeft"]'**)), callback=**'parse\_url\_contents'**,  follow=True) |

* SgmlLinkExtractor is used to extract the links in every page.

**Task 4:**

Json object is created for every webpage that is crawled.

The json key value pairs consist of the below:

- **doc\_id**: printed using UUID

- **URL**: printed using response.url

- **raw\_content**: response.text

- **timestamp\_crawl**: when did you crawl. Extracted using datetime.now()