# INF 558 : BUILDING KNOWLEDGE GRAPH

## Homework 2: CRAWLING

**Name**: KAVYA SETHURAMAN
**USC-ID:** 7852999061

**Task 1:**

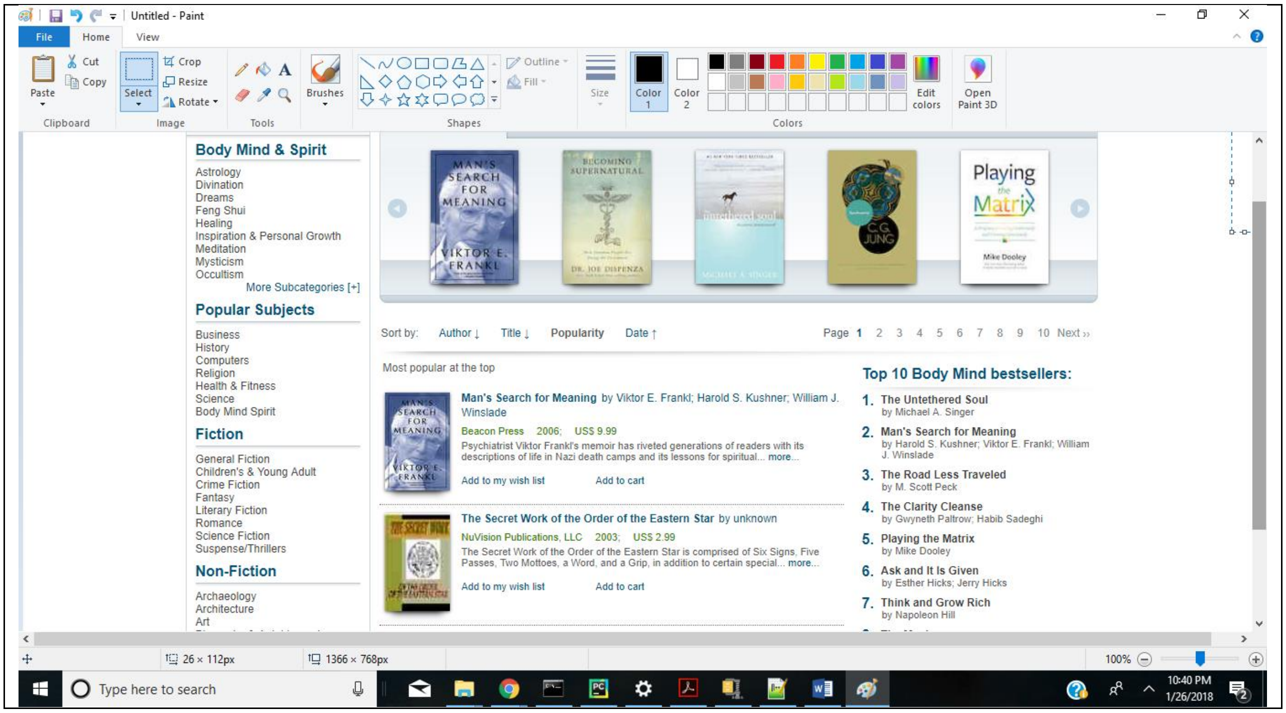
**-Website URL**: www.ebooks.com

**-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:



**Body Mind & Spirit**

Astrology
Divination
Dreams
Feng Shui
Healing
Inspiration & Personal Growth
Meditation
Mysticism
Occultism
More Subcategories [+]

**Popular Subjects**

Business
History
Computers
Religion
Health & Fitness
Science
Body Mind Spirit

**Fiction**

General Fiction
Children's & Young Adult
Crime Fiction
Fantasy
Literary Fiction
Romance
Science Fiction
Suspense/Thrillers

**Non-Fiction**

Archaeology
Architecture
Art

Sort by: Author ↓ Title ↓ Popularity Date ↑ Page 1 2 3 4 5 6 7 8 9 10 Next ››

Most popular at the top

**Man's Search for Meaning** by Viktor E. Frankl; Harold S. Kushner; William J. Winslade
Beacon Press 2006; US$ 9.99
Psychiatrist Viktor Frankl's memoir has riveted generations of readers with its descriptions of life in Nazi death camps and its lessons for spiritual... more...
Add to my wish list Add to cart

**The Secret Work of the Order of the Eastern Star** by unknown
NuVision Publications, LLC 2003; US$ 2.99
The Secret Work of the Order of the Eastern Star is comprised of Six Signs, Five Passes, Two Mottoes, a Word, and a Grip, in addition to certain special... more...
Add to my wish list Add to cart

**Top 10 Body Mind bestsellers:**

1. The Untethered Soul
by Michael A. Singer
2. Man's Search for Meaning
by Harold S. Kushner; Viktor E. Frankl; William J. Winslade
3. The Road Less Traveled
by M. Scott Peck
4. The Clarity Cleanse
by Gwyneth Paltrow; Habib Sadeghi
5. Playing the Matrix
by Mike Dooley
6. Ask and It Is Given
by Esther Hicks; Jerry Hicks
7. Think and Grow Rich
by Napoleon Hill

**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()