➢ Build a simple python web crawler

Python is a high level programming language including object-oriented, imperative, functional programming and a large standard library. For the web crawler two standard library are used - requests and BeautfulSoup4. requests provides a easy way to connect to world wide web and BeautifulSoup4 is used for some particular string operations.

```
import requests

from bs4 import BeautifulSoup

def web(page,WebUrl):

   if(page>0):

      url = WebUrl

      code = requests.get(url)

      plain = code.text

      s = BeautifulSoup(plain, "html.parser")

      for link in s.findAll('a', {'class':'s-access-detail-page'}):

         tet = link.get('title')

         print(tet)

         tet_2 = link.get('href')

         print(tet_2)
```

```
web(1,'http://www.amazon.in/s/ref=s9_acss_bw_cts_VodooFS_T4_w?rh=i%3Aelectronics%
2Cn%3A976419031%2Cn%3A%21976420031%2Cn%3A1389401031%2Cn%3A1389432031%
2Cn%3A1805560031%2Cp_98%3A10440597031%2Cp_36%3A1500000-
99999999&bbn=1805560031&rw_html_to_wsrp=1&pf_rd_m=A1K21FY43GMZF8&pf_rd_s=
merchandised-search-
3&pf_rd_r=2EKZMFFDEXJ5HE8RVV6E&pf_rd_t=101&pf_rd_p=c92c2f88-469b-4b56-936e-
0e65f92eebac&pf_rd_i=1389432031')
```

****

Second Example

```
from bs4 import BeautifulSoup

import requests

url = raw_input("Enter a website to extract the URL's from: ")

r  = requests.get("http://" +url)

data = r.text

soup = BeautifulSoup(data)

for link in soup.find_all('a'):

   print(link.get('href'))
```

****

But,

Scrapy is a fast, high-level screen scraping, and web crawling framework, it is completely written in Python and runs on Linux, Windows, Mac and BSD. Some basic features of Scrapy are given below

- Simple – Scrapy was designed with simplicity in mind, by providing the essential features
- Productive – We just have to write the rules to extract the data from web pages and Scrapy crawls the entire web site.
- Extensible – provides several mechanisms to plug new code without having to touch the framework core

Hence we combine Python with Scrapy for web crawling.

➢ Follow the below link and you can find detail documentation on Scrapy.

https://doc.scrapy.org/en/latest/intro/overview.html

Once you complete, solve this 2 following questions.

1. Implement single Topical Crawler(My Spider)
2. Then multiple crawlers Myspider1 first and then run the Myspider2 multiples times depending (sequentially) on some conditions.