Programming Test Assignment

Alka Bharti

Approach:

- Defined two functions- Extract_Title_and_Link and JSON_Format
- Extract_Title_and_Link will extract the data using regex pattern matching techniques. Firstly we need to find the <section> in the URL which shows the 5 latest news. Then finding the <headers> which store the information of both title and link. Iterating over the header list we can extract the title and link, and store it into a dictionary called news.
- After successfully extracting data we can convert it into JSON Format by using json.dumps function.

Python Code:

```
from urllib.request import urlopen
import re
import json

latest_news = []

def Extract_Title_and_Link(url):
  page = urlopen(url)
  html_bytes = page.read()
  html = html_bytes.decode("utf-8")

  pattern = '<section class="homepage-module
latest".*?>(.|\n)*?<\/section.*?>'
  match_results = re.search(pattern, html, re.IGNORECASE)
  section = match_results.group()

  h2_list = re.findall('<h2.*?>.*?</h2>',section)

  for h2 in h2_list:
    news = {}
```

```
    link = re.findall('href=.*?/>',h2)
    extracted_link = re.sub('href=',"",link[0])
    extracted_link = re.sub('/>',"",extracted_link)
    extracted_link = "https://time.com" + extracted_link

    title = re.findall('/>.*?</',h2)
    extracted_title = re.sub('/>',"",title[0])
    extracted_title = re.sub('</',"",extracted_title)

    news["title"] = extracted_title
    news["link"] = extracted_link
    latest_news.append(news)

  return latest_news

def JSON_Format(latest_news):
  response = json.dumps(latest_news)
  print(response)

def main():
  Extract_Title_and_Link("https://time.com/")
  JSON_Format(latest_news)

if __name__=="__main__":
    main()
```

Note: Code and Output Screenshot is attached in the mail.

Screen Shot of the Output: