Step 1:

Parse the file to html format using “html. Parser”

Step 2:

Verify the tags under which required information is present (body tag considered here).

Step3:

Get the data and print

Code used:

**files =  glob.glob('/home/data/\*.txt')** // (for text documents) // and **glob.glob('/home/data/\*’)** for any generic search//

**for file in files:**

**print (file)**

**f = open(file, 'rb').read()**

**soup = BeautifulSoup(f,"html.parser")**

**ul = soup.find('body')**

**print (ul)**

**data = ul.get\_text()**

**print (data)**

(or)

Step1:

Open the file using Notepad++

Step2:

Find “**<[^>]+>**” and Replace with nothing (leave empty) with search mode set to Regular Expression

Step3:

Save the changes