WEB MINING

LAB3

Name : Soumok Dutta

Reg No. : 16BCE1184

**import** requests  
**import** urllib3  
**from** bs4 **import** BeautifulSoup **as** bs  
**import** re  
  
  
page = requests.get(**'https://en.wikipedia.org/wiki/Student'**)  
  
soup = bs(page.text,**'html.parser'**)  
soup.prettify()  
  
**print**(soup.findAll(**"b"**,text=re.compile(**'^student\*'**)))  
  
stack=[]  
**for** a **in** soup.find\_all(**"a"**,attrs={**'href'** : re.compile(**"^http"**)}):  
 stack.append(a.get(**'href'**))  
  
**for** i **in** stack:  
 **print**(i)  
  
  
depth = 2  
Stackfinal=[]  
**for** k **in** range(depth-1):  
 stackin=[]  
 **for** i **in** stack:  
 **print**(i)  
 page2=requests.get(i)  
 soup=bs(page2.text,**'html.parser'**)  
 **for** j **in** soup.find\_all(**"a"**,attrs={**'href'** : re.compile(**"^http"**)}):  
 stackin.append(j.get(**'href'**))  
 Stackfinal.append(stackin)  
 k=k+1  
 **print**(**"\n\n"**)  
  
**print**(Stackfinal)   
  
  
stackstudent=[**'https://en.wikipedia.org/wiki/Student'**]  
**for** i **in** stack:  
 page3=requests.get(i)  
 soup2 = bs(page.text,**'html.parser'**)  
 **print**(soup2.findAll(**"b"**,text=re.compile(**'^<b>student\*'**)))  
 stackstudent.append(i) **if** soup2.findAll(**"b"**,text=re.compile(**'^student\*'**)) **else** None  
 **print**(stackstudent)