1.)

Import ulrlib2

Print urllib2.urlopen(“the url you want to extract the data”).read()

Now you we will get the file form the web in the markup language .

Add package in the notebook for data bricks (html2text) --.> import html2text

2.)

import urllib2

from html2text import html2text

print html2text(urllib2.urlopen(“the url of the page you want to extract”).read())

this will remove all the markup values and filter the data.

3) now we will iterate through each line

import ulrlib2

from html2text import html2text

for line in html2text(urllib2.urlopen(“the url of the page”).read()).split(“\n”):

print line

this gives us the ability to read line by line

4) now if you want to check a particular line # check for indentation

import urllib2

from html2text import html2text

for line in html2text(urllib2.urlopen(“your url of the page which you want open”).read()).split(“\n”):

if “IMDb” in line:

print line

5) now we want to clear the remaining text if any attached to the url link :

import urllib2

from html2text import html2text

for line in html2text(urllib2.urlopen(“your url of the page which you want open”).read()).split(“\n”):

if “IMDb” in line:

print line.split(“[IMDb]”)

now this will give two separate lines . beginning as ‘u’

This will be the output of the file . now we have to clear it more.

[u'![](dot.jpg) Links: ', u'(http://www.imdb.com/title/tt0112573/),']

6)now accessing the index [1]

import urllib2

from html2text import html2text

for line in html2text(urllib2.urlopen(“your url of the page which you want open”).read()).split(“\n”):

if “IMDb” in line:

print line.split(“[IMDb]”) [1]

(<http://www.imdb.com/title/tt0112573/)>,

7)now we will have to remove the brackets and commas from the line.

Import urllib2

from html2text import html2text

for line in html2text(urllib2.urlopen(“name of the url “).read()).split(“\n”):

if “IMDb” in line:

print line.split(“[IMDB]”)[1].strip(“(“).strip(“,”).strip(“)”):

output

<http://www.imdb.com/title/tt0112573/>

8) now if we want to access some thing from the page

Import urllib2

from html2text import html2text

for line in html2text(urllib2.urlopen(“name of the url “).read()).split(“\n”):

if “IMDb” in line:

print line.split(“[IMDB]”)[1].strip(“(“).strip(“,”).strip(“)”):

if “Film:” in line:

print line

if “count: in line:

print line

output

http://www.imdb.com/title/tt0112573/

![](dot.jpg) Entire Film: 184

![](dot.jpg) Wallace's (Gibson) count: 33

![](dot.jpg) Hamish (Gleeson) count: 10

now again split the lines and get the required data

from html2text import html2text

for line in html2text(urllib2.urlopen(“name of the url “).read()).split(“\n”):

if “IMDb” in line:

print line.split(“[IMDB]”)[1].strip(“(“).strip(“,”).strip(“)”):

if “Film:” in line:

print line.split(“Film:”)

if “count: in line:

print line.split(“count:”)

[u'![](dot.jpg) Entire ', u' 184 ']

[u"![](dot.jpg) Wallace's (Gibson) ", u' 33 ']

[u'![](dot.jpg) Hamish (Gleeson) ', u' 10 ']

and same we need the index positon 1 so

from html2text import html2text

for line in html2text(urllib2.urlopen(“name of the url “).read()).split(“\n”):

if “IMDb” in line:

print line.split(“[IMDB]”)[1].strip(“(“).strip(“,”).strip(“)”):

if “Film:” in line:

print line.split(“Film:”)[1]

if “count: in line:

print line.split(“count:”)[1]

final answer

http://www.imdb.com/title/tt0112573/

184

33

10