0:13
In this exercise we look at scraping data from IMDB.
0:18
Our goal is to convert the top two of the list of movies and IMBD into tabular form
0:23
using python.
0:25
This data can then be used for further analysis.
0:28
So, the first step is to import necessary libraries.
0:33
We first have the BeautifulSoup library.
0:36
This is the most important library for this tutorial.
0:40
This helps us create data from the web.
0:44
Then we have the request library that is used to access websites.
0:48
We have the pandas library that is a very popular library in python for data manipulation.
0:55
Let us run this module next we load the webpage.
0:58
So, for that we need to know the URL.
1:02
What we need is the top 50 list of movies.
1:06
Let us see where to find it?
1:08
I go to Google type top 50 movies and then this takes you to the first link.
1:22
Let us see what that is.
1:24
So, here we have the top rated movies and IMDB showing the top three titles.
1:29
This is what we want.
1:31
We want the movie title the year the rating and we want this data in a tabular form.
1:39
So, this is the data.
1:41
So, let us take this link and copy it.
1:46
It is a URL copy that goes back to our jupyter notebook and this is where we paste the URL.
1:56
Let us paste then we convert this content to a BeautifulSoup object.
2:03

And then we can print out the content and see what it is pretty slowly and here we have
2:16
the country.
2:17
So, in order to get the information from this page we need this information and this Id
2:27
of the year and the ratings are actually located within html tags.
2:31
For example if you rightly this tag and go to inspect, it will take you directly to this
2:41
tag.
2:42
So, this title the Shortage and Predominance located within this tag called a href.
2:46
So, our code is to extract this title and this title is located within this tag.
2:53
So, similarly if you just hover your mouse over this that particular IT will get highlighted.
3:04
Similarly if you want to know which tag here actually has this information 1994.
3:10
Just to go below an hour, see highlights of that information.
3:15
So, where is the rating located?
3:18
Just below and here you see the new hover your mouse over this line it actually highlights
3:23
the rating there.
3:25
So, what we want is we need these tags so that we can extract the information that is
3:31
located within these tags.
3:34
So, we saw that it was very easy.
3:38
All it had to do was, if you want this year, highlight the year and inspect it.
3:47
This will directly take you to the tag that has been given information that required information.
3:55
The other way that we saw him was to print the contents of the html page.
4:00
And here you get very detailed information.
4:02
So, it is very difficult to locate the tags here.

4:04
So, the easier way is as I showed you here go the information rightly inspect it will
4:11
take you to me respective rags.
4:15
So, now that we are done here, let us create an empty list.
4:19
We need the title, the year rating , and some temporary variables.
4:22
Let us run the module and next let us extract the estimated tag contents.
4:28
This is the most important step in this exercise.
4:32
So, what we need let us see first we are getting the class name called chart full width.
4:39
Let us see where this is located.
4:41
Why are we doing this?
4:43
Let us see.
4:44
Let us go to the web page.
4:46
So, we need the title, the year and the rating.
4:48
Let us see where the title is located to find the tag right like inspecting it will directly
4:57
take you to the tag.
4:59
So, all you need is this href tag name.
5:03
So, that you can extract this title but this href tag is located within a higher level
5:10
class called title column.
5:12
And if you see the title column this higher level td class the a href which contains the
5:18
title and span class which contains 1994.
5:22
So, we can just use this tag to extract both the title as well as the year information.
5:33
So, you see the title column we are using here in the next line for the movie year and
5:44

for the movie title.
5:48
But then what the start through is to let us go back.
5:52
So, if you move up we see that here start from with this located here it is a much higher
5:59
level tag.
6:00
So, what we saw here is it is just for one movie the Shortest Redemption the year this
6:07
is just for one movie.
6:08
But what if you want the information for all the movies for that you need to locate a higher
6:13
level tag and this was the most suitable that we could find to extract all the movies.
6:19
So, chart full width this is the highest level tag that we are using chart full width.
6:24
Next for the movie title and here we use title column as I showed you title column let us
6:34
run this let us run this and then we also need the rating.
6:40
So, this is also a higher level tag.
6:42
Let us see where this is located.
6:45
So, if you are not able to locate just go to the rating highlighted right click inspect
6:55
takes you here see here the rating is located within this rating main point is located within
7:02
strong title.
7:03
And this is inside the td class rating column IMDB rating column I am deteriorating.
7:11
We are using this higher level.
7:14
So, once we are done with the higher level class we need to extract this information:
7:21
the title, the year and rating.
7:23

So, let us first run this.
7:27
Now the movie title, how do we extract the movie title?
7:31
We see that this title is equal to row dot text.
7:35
So, row is an inbuilt function: a dot text a refers to the tag name dot text extracts
7:43
the text within the tag.
7:45
So, what is a dot tag?
7:47
So, we see the packing the packing the shortage redemption is located within a href all we
7:53
are using is a dot tag this will extract the movie title let us run this.
8:00
So, we get all the movie titles.
8:03
How do we get the movie here?
8:06
Let us see which tag it is located in.
8:09
Let us go to year, the year is located within the span class.
8:18
So, row dot span dot text.
8:23
Span dot text gets the text within the span tag.
8:26
Run this we get all the years.
8:28
Now where are the ratings located?
8:32
So, let us look for the ratings?
8:36
The ratings are located within the tag here 9.0 is located within the tag strong dot title.
8:42
So, we need strong dot text to extract the text within the strong tag that is to run
8:49
it.
8:50
There you go we get all the ratings.
8:52
Now that we have the movie title the movie year in the movie ratings, let us combine
8:58
all this and we can do that using the panda data frame.

9:03
It gives you a tabular form for your data and then we print out the top thirty rows.
9:08
Let us run this.
9:11
There you go all the data that we needed the movie typing the year and the ratings are
9:16
available in tabular form.