How can I read HTML tables using Pandas, and could you provide an example?

Authored by

stats writer

July 1, 2024

Pandas is a popular Python library used for data analysis and manipulation. It also has the ability to read and analyze HTML tables, making it a useful tool for extracting data from websites. To read HTML tables using Pandas, the pd.read_html() function is used. This function takes the URL of the webpage as its input and returns a list of dataframes containing the table data. An example of using this function would be:

df_list = pd.read_html("https://www.examplewebsite.com/table.html")

This would return a list of dataframes, with each dataframe containing the data from a different HTML table found on the webpage. These dataframes can then be further analyzed and manipulated using Pandas' various functions and methods. Overall, Pandas provides a convenient and efficient way to extract and work with data from HTML tables.

Read HTML Tables with Pandas (Including Example)

You can use the pandas function to read HTML tables into a pandas DataFrame.

This function uses the following basic syntax:

df = pd.read_html('https://en.wikipedia.org/wiki/National_Basketball_Association')

The following example shows how to use this function to read in a table of NBA team names from .

Example: Read HTML Table with Pandas

Before using the read_html() function, you'll likely have to install lxml:

pip install lxml

Note: If you're using a Jupyter notebook, you need to restart the kernel after performing this installation.

Next, we can use the read_html() function to read every HTML table on :

```
import pandas as pd
import numpy as np
import matplotlib.pyplotas plt
from unicodedata import normalize

#read all HTML tables from specific URL
tabs = pd.read_html('https://en.wikipedia.org/wiki/National_Basketball_Association')

#display total number of tables read
len(tabs)
```

44

We can see that a total of 44 HTML tables were found on this page.

I know that the table I'm interested in has the word "Division" in it, so I can use the match argument to only retrieve HTML tables that contain this word:

```
#read HTML tables from specific URL with the word "Division" in them
tabs = pd.read_html('https://en.wikipedia.org/wiki/National_Basketball_Association',
match='Division')

#display total number of tables read
len(tabs)
```

1

I can then of the columns of the table:

```
#define table
df = tabs

#list all column names of table
list(df)
```

I'm only interested in the first two columns, so I can the DataFrame to only contain these columns:

#filter DataFrame to only contain first two columns
df_final = df.iloc

#rename columns
df_final.columns =

#view first few rows of final DataFrame
print(df_final.head())

Division Team
0 Atlantic Boston Celtics
1 Atlantic Brooklyn Nets
2 Atlantic New York Knicks
3 Atlantic Philadelphia 76ers
4 Atlantic Toronto Raptors

Additional Resources

The following tutorials explain how to read other types of files in pandas: