STRING MANIPULATION

PYTHON for DATA SCIENCE

Jaume Boguñá

Dive into Python

String Manipulation in Pandas

In Pandas, **string manipulation** is an important task when working with **textual data**.

Let's explore some common **string manipulation functions**:

1. `str.contains()`
2. `str.replace()`
3. `str.split()`
4. `str.upper()` / `str.lower()` / `str.title()`
5. `str.startswith()` / `str.endswith()`
6. `str.len()`
7. `str.strip()` / `str.lstrip()` / `str.rstrip()`
8. `str.pad()`

1. *str.contains()*

Checks if a substring is present in a string or Series of strings

Parameters

```
Series.str.contains(
    pat: str,
    case=True,
    flags=0,
    na=None,
    regex=True
    )
```

1. *str.contains()*

Creating the Pandas DataFrame

```
import pandas as pd

df = pd.DataFrame({'emails': ['jaume@gmail.com',
                              'sara@hotmail.com',
                              'sam@yahoo.com',
                              'mike@gmail.com']})

df
              emails
0    jaume@gmail.com
1   sara@hotmail.com
2      sam@yahoo.com
3     mike@gmail.com
```

Jaume Boguñá

Dive into Python

1. str.contains()

Returning a Series of booleans using only a literal pattern

```
df
              emails
0   jaume@gmail.com
1  sara@hotmail.com
2     sam@yahoo.com
3    mike@gmail.com


# Check if the word 'gmail' is in the 'emails' column

df['is_gmail'] = df['emails'].str.contains('gmail')

              emails  is_gmail
0   jaume@gmail.com      True
1  sara@hotmail.com     False
2     sam@yahoo.com     False
3    mike@gmail.com      True
```

Jaume Boguñá

Dive into Python

2. *str.replace()*

Replaces occurrences of a substring with another substring

Parameters

```
Series.str.replace(
    pat,
    repl,
    n=-1,
    case=None,
    flags=0,
    regex=False
    )
```

Jaume Boguñá

Dive into Python

2. *str.replace()*

Using *str.replace* to Clean the Data

```
df
   Stock     Price    Cap
0   AAPL   $145.09  2.41T
1  GOOGL  $2730.81  1.82T
2   AMZN  $3401.46  1.73T
3  TSLA   $699.60  0.71T


# Removing the dollar sign from the Price column

df['Price'] = df['Price'].str.replace('$', '')

   Stock    Price    Cap
0   AAPL   145.09  2.41T
1  GOOGL  2730.81  1.82T
2   AMZN  3401.46  1.73T
3   TSLA   699.60  0.71T
```

Jaume Boguñá

Dive into Python

3. *str.split()*

Splits a string into a list of substrings based on a delimiter

Parameters

```
Series.str.split(
    pat=None,
    n=-1,
    expand=False,
    regex=None
    )
```

Jaume Boguñá

Dive into Python

3. *str.split()*

Using *str.replace* to Clean the Data

```
df
             Name       City Profession
0    Jaume Boguñá     Madrid   Engineer
1   Joan Pellicer  Tarragona      Actor

# Splitting the 'Name' column into 'First Name' and 'Surname'

df[['First Name', 'Surname']] = df['Name'].str.split(' ', n=1, expand=True)

# Dropping Name column

df = df.drop('Name', axis=1)

df = df[['First Name', 'Surname', 'City', 'Profession']]

  First Name    Surname       City Profession
0      Jaume     Boguñá     Madrid   Engineer
1       Joan   Pellicer  Tarragona      Actor
```

Jaume Boguñá

Dive into Python

4. str.title()

Converts the first letter of each word to uppercase (title case)

```
df
         Author                         Book  Year
0  george orwell                        1984  1949
1  jane austen           pride and prejudice  1813
2  mark twain   the adventures of tom sawyer  1876

# Applying str.title() to the 'Author' and 'Book' columns

df['Author'] = df['Author'].str.title()

df['Book'] = df['Book'].str.title()

          Author                          Book  Year
0  George Orwell                          1984  1949
1    Jane Austen           Pride And Prejudice  1813
2     Mark Twain  The Adventures Of Tom Sawyer  1876
```

Jaume Boguñá

Dive into Python

5. str.startswith()

Returns True if the string starts with the given substring

```
df
        Sport                  Team Country
0    Football     Manchester United      UK
1  Basketball    Los Angeles Lakers     USA
2    Baseball      New York Yankees     USA

# Use str.startswith() to Filter Teams Starting with 'L'

df['teams_starting_with_L'] = df['Team'].str.startswith('L')

        Sport                  Team Country  teams_starting_with_L
0    Football     Manchester United      UK                  False
1  Basketball    Los Angeles Lakers     USA                   True
2    Baseball      New York Yankees     USA                  False
```

Jaume Boguñá

Dive into Python

6. str.len()

Returns the length of each string (number of characters)

```
df
         Username  Language  Level
0         jbo1881   Catalan      4
1     cami2000_fr    French      8
2  apocrotte_pydf  Japanese      1

# Use str.len() to get number of characters of the Username

df['Username_len'] = df['Username'].str.len()

         Username  Language  Level  Username_len
0         jbo1881   Catalan      4             7
1     cami2000_fr    French      8            11
2  apocrotte_pydf  Japanese      1            14
```

Jaume Boguñá

Dive into Python

7. str.strip()

Removes leading and trailing whitespace or characters

```
df
      Name   Telephone   Location
0    Pedro  +346555555     Bilbao
1  Laurent  +336555559   Bordeaux
2  Massimo  +396555558       Rome

# Use str.strip() to remove '+' from the phone numbers

df['Telephone'] = df['Telephone'].str.strip('+')

      Name  Telephone   Location
0    Pedro  346555555     Bilbao
1  Laurent  336555559   Bordeaux
2  Massimo  396555558       Rome
```

Jaume Boguñá

Dive into Python

8. *str.pad()*

Adds padding (spaces or specified characters) to strings to a specified width

Parameters

```
Series.str.split(
    width,
    side='left',
    fillchar=' '
    )
```

8. str.pad()

Adds padding (spaces or specified characters) to strings to a specified width

```
df
     Player       Sport    Country
0   Messi     Football  Argentina
1  LeBron  Basketball        USA
2   Serena      Tennis        USA

# Applying str.pad to the 'Player' column

df['Player'] = df['Player'].str.pad(width=10, side='both', fillchar='.’)

        Player       Sport    Country
0   ..Messi...    Football  Argentina
1   ..LeBron..  Basketball        USA
2   ..Serena..      Tennis        USA
```

Jaume Boguñá

Dive into Python

Like

Comment

Share

Jaume Boguñá

Aerospace Engineer | Data Scientist