

Pandas для анализа данных

Вызов библиотеки pandas

Вызов библиотеки pandas

```
In import pandas
import pandas as pd
```

Конструктор DataFrame() для создания таблицы

```
In pd.DataFrame(data = data, columns =
columns)
# аргумент data – список с данными,
# аргумент columns – список с
# названиями столбцов
```

Метод tail() для вывода последних строк таблицы

```
In df.tail() # первые 5 строк
df.tail(15) # первые 15 строк
```

Метод read_csv() для чтения файлов формата CSV

```
In df = pd.read_csv('путь к файлу')
```

Метод head() для вывода первых строк таблицы

```
In df.head() # первые 5 строк
df.head(10) # первые 10 строк
```

Атрибут columns для вывода названий столбцов

```
In df.columns
```

Атрибут shape для вывода размера таблицы

```
In df.shape
```

Атрибут dtypes для получения информации о типах данных в таблице

```
In df.dtypes
```

Метод info() для просмотра сводной информации о таблице

```
In df.info()
```

Атрибут loc[строка, столбец] даёт доступ к элементу в DataFrame по строке и столбцу

```
In df.loc[:, 'column']
```

Out Вид	Реализация
Одна ячейка	<code>.loc[7, 'column']</code>
Один столбец	<code>.loc[:, 'column']</code>
Несколько столбцов	<code>.loc[:, ['column_1', 'column_4']]</code>
Несколько столбцов подряд (срез)	<code>.loc[:, 'column_5': 'column_8']</code>
Одна строка	<code>.loc[1]</code>
Все строки, начиная с заданной	<code>.loc[1:]</code>
Все строки до заданной	<code>.loc[:3]</code>
Несколько строк подряд (срез)	<code>.loc[2:5]</code>

Логическая индексация для получения элементов по определённому условию

Out Вид	Реализация	Сокращённая запись
Все строки, удовлетворяющие условию	<code>'df.loc[df.loc[:, 'column'] == 'X']</code>	<code>'df[df['column'] == 'X']</code>
Столбец, удовлетворяющий условию	<code>'df.loc[df.loc[:, 'column'] == 'X']['column']</code>	<code>'df[df['column'] == 'X']['column']</code>
Применение метода	<code>'df.loc[df.loc[:, 'column'] == 'X']['column'].count()</code>	<code>'df[df['column'] == 'X']['column'].count()</code>

Индексация в Series

Out Вид	Реализация	Сокращённая запись
Один элемент	<code>`df.loc[7]`</code>	<code>`df[7]`</code>
Несколько элементов	<code>`df.loc[[5, 7, 10]]`</code>	<code>`df[[5, 7, 10]]`</code>
Несколько элементов подряд (срез)	<code>`df.loc[5:10]`</code> включая 10	<code>`df[5:10]`</code> не включая 10
Все элементы, начиная с заданного	<code>`df.loc[1:]`</code>	<code>`df[1:]`</code>
Все элементы до заданного	<code>`df.loc[:3]`</code> включая 3	<code>`df[:3]`</code> не включая 3

Словарь

Библиотека

Это набор готовых методов для решения распространенных задач

CSV

Формат файла (от англ. Comma-Separated Values, «значения, разделённые запятой»). Каждая строка представляет собой одну строку таблицы, где данные разделены запятыми. В первой строке собраны заголовки столбцов (если они есть)

Кортеж

Одномерная неизменяемая последовательность данных. Она похожа на список, её тоже можно сохранять в переменной.

Series

Одномерная структура данных Pandas, её элементы можно получить по индексу. Каждый индекс представляет собой номер отдельного наблюдения, и поэтому несколько различных Series вместе составляют DataFrame.

- В Series хранятся данные одного типа.
- У Series есть имя (Name), информация о количестве данных в столбце (Length) и тип данных, которые хранятся в ней (dtype).
- Индексация в Series аналогична индексации элементов столбца в DataFrame.

DataFrame

Это двумерная структура данных Pandas, где у каждого элемента есть два индекса: по строке и по столбцу.

- Каждая строка — это одно наблюдение, запись об объекте исследования. А столбцы — признаки этого объекта.
- DataFrame() — это конструктор библиотеки Pandas, который используется для создания **DataFrame**. Перед именем конструктора стоит обращение к переменной, в которой библиотека хранится: `pd.DataFrame()`.
- У DataFrame есть неотъемлемые свойства, значения которых можно запросить. Они называются атрибуты. Например, это размер таблицы `df.shape` или количество столбцов `df.columns`.
- К каждой ячейке с данными в DataFrame можно обратиться по её индексу и названию столбца. Этот процесс называется индексация и для DataFrame его проводят разными способами.