

Pandas для анализа данных

1/2

Синтаксис

Импорт библиотеки

```
import pandas as pd
```

Конструктор **DataFrame()** для создания таблицы

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

Метод **head()** для вывода первых строк таблицы

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

Метод **tail()** для вывода последних строк таблицы

```
df.tail() # последние 5 строк
df.tail(15) # последние 15 строк
```

Чтение файлов формата **.CSV**

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

Атрибуты и методы датафрейма

```
df.columns # выводит названия столбцов
df.shape # выводит размер таблицы
df.dtypes # выводит информацию о типах данных в таблице
df.info() # метод выводит сводную информацию о таблице
```

Pandas для анализа данных

Полная и сокращённая записи при индексации датафрейма

	Полная запись	Сокращённая запись
Один столбец	<code>.loc[:, 'genre']</code>	<code>df['genre']</code>
Несколько столбцов	<code>.loc[:, ['genre', 'Artist']]</code>	<code>df[['genre', 'Artist']]</code>
Все строки, начиная с заданной	<code>.loc[0:]</code>	<code>df[0:]</code>
Все строки до заданной	<code>.loc[:3]</code> (включая 3)	<code>df[:3]</code> (не включая 3)
Несколько строк подряд (срез)	<code>.loc[2:5]</code> (включая 5)	<code>df[2:5]</code> (не включая 5)
Одна ячейка	<code>.loc[7, 'genre']</code>	-
Одна строка	<code>.loc[1]</code>	-
Несколько столбцов подряд (срез)	<code>.loc[:, 'total_play': 'genre']</code>	-

Индексация в Series

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

Глоссарий

Библиотека — набор готовых методов для решения распространённых задач.

.csv — формат файла (от англ. Comma-Separated Values, «значения, разделённые запятой»). Каждая строка представляет собой одну строку таблицы, где данные разделены запятыми.

Датафрейм — двумерная структура данных pandas, где у каждого элемента есть два индекса: по строке и по столбцу.

Объект Series — одномерная структура данных pandas, её элементы можно получить по индексу.

Конструктор — это метод, который создаёт новые объекты. Например, конструктор `DataFrame()` в pandas создаёт датафреймы.