Recitation 06

01:198:210 Data management for Data Science

Xinxi (Chris) Zhang

Recap

- GPT’s new feature
  - ✦DALL E 3
- Numpy
  - ✦Broadcasting
  - ✦Itertool
- Pandas Intro
  - ✦What and why
  - ✦Data Frame, Serires
  - ✦Basic Operations

Pandas 02

Some Highlights

In-Place vs Copy Operations

- Object.func()
- **In-Place** Operation
  - Modify the Original Object
- **Copy** Operation
  - Create a new one by copying the original one, then modify the new one

Iloc vs loc

- **Indexing**
- **iloc**
  - Inter-location based indexing
- Loc
  - Label-based indexing
- **Advanced indexing?**

Iloc vs loc

- **Indexing**
- **iloc**
  - Inter-location based indexing
- Loc
  - Label-based indexing
- **Advanced indexing?**

Apply

- Takes a function as input and applies this function to each element in a Series or along a particular axis in a DataFrame
- Lambda
- Def a Function
- Parallel Computing?

Before Data Cleaning

```
df1 = pd.DataFrame(np.arange(9).reshape((3, 3)),
                   columns=["a", "b", "c"])
df2 = pd.DataFrame(np.arange(12).reshape((3, 4)),
                   columns=["a", "b", "c", “d"])
df1+df2?
```

- **Robustness**
- **Why Jupyternotebook**
- **How are we gonna further deal with this**

Data Cleaning 01

Intro

What Kind of Problems are We Facing

What Kind of Problems are We Facing

- Missing Data (NaN)
- Duplicate Data
- Outliers
- Inconsistent Feature Num
- Inconsistent Data Types
- Mixed

Missing Data

Missing Data

- Drop
- Fill with a specific value (0), why 0?
- Forward fill or backward fill
- Numerical Fill

Inconsistant Data

Inconsistant Data

- Extra rolls
  - Keep it or delete it
- Inconsistent data type
  - Convert Data type
  - Deal with NaN

Mixed

- Hierarchy
- A general Solution?
- Why do we still need to learn?

Generator

- produce a sequence of values lazily
- Memory efficient
- A Clean and Powerful Logic
  - DataLoader in PyTorch