Féidearthachtaí as Cuimse
Infinite Possibilities

Programming for Analytics

Lecture 6: Introduction to Pandas

Bojan Božić
School of Computer Science
TU Dublin, Grangegorman

bojan.bozic@tudublin.ie

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Overview

- What is an Pandas?
- DataFrames
- Loading Data from CSV
- Selecting and Filtering
- Sorting, Grouping and Aggregating
- Missing Values

What is Pandas?

- Pandas is a Python library for data analysis
- Built on top of NumPy
- Enables working with tabular data:

  `DataFrames`

Installing Pandas

- Use pip to install:
  - `pip install pandas`
- Import using
  - `import pandas as pd`

What is a DataFrame?

- A 2D table with labelled rows and columns
- Each column can have a different data type
- Like a spreadsheet or SQL table in Python

Example

```
data = {“key1”: “value1”, …}

df = pd.DataFrame(data)
```

Activity: Create a DataFrame

- Use a Python dictionary to create a small DataFrame

Loading Data from CSV

- Use `pd.read_csv('filename.csv')`
- DataFrame is created automatically
- Pandas infers column names and data types

Activity: Load Sample Dataset

- Load the pov-data.csv
- Print the first few rows using `.head()`

Inspecting Data

- `df.head(), df.tail()` – first and last rows
- `df.info()` – structure and data types
- `df.describe()` – statistics of numbers columns
- `df.shape` – dimensions of DataFrame

Activity: Explore a Dataset

- Call `.info()`, `.describe()`, `.shape` on your loaded data
- Explain what these methods (and property) do

Selecting Columns

- Access single column: `df[‘column’]` or `df.column`
- Access multiple columns:

  `f[[‘col1’, ‘col2’]]`
- Result is `Series` or `DataFrame`

Filtering Rows

- Use Boolean expressions:

```
df[df[‘Age’] > 30]
```

- Combine conditions with `&` (and), `|` (or), `~` (not)
- Enclose conditions in parentheses!

Activity: Filter Data

- Find the names of characters who appear in the first chapter of any of the books.

Sorting Data

- Use `df.sort_values(by='column')`
- Use `asending=False` for descending order
- Can sort by multiple columns

Activity: Sort DataFrame

- Sort dataset by book in descending order.

Adding New Columns

- Create columns using assignment: `df['new'] = …`
- Can use existing columns or functions
- E.g. `df['BMI'] = df['Weight'] / df['Height']**2`

Activity: Create a Calculated Column

- The events in books 1, 3 and 7 of the Malazan Book of the Fallen series, mostly occur on Genabackis.
- Create a new column which shows ‘Yes’ if the POV is (most likely) located on Genabackis, and ‘No’ if they aren’t.

Dropping Columns or Rows

- `df.drop(‘col’, axis=1)` - drop column
- `df.drop(index)` – drop row by index
- Set `inplace=True` to modify directly

Grouping and Aggregating

- Use `df.groupby(‘col’)` to group data
- Use `.agg(), .mean(), .sum(),` etc. to summarise
- Great for summarising by category or type

Activity: Group Data

- Group dataset by book and calculate the number of POVs in each book

Handling Missing Data

- Check for missing data: `df.isna().sum()`
- Remove with `df.dropna()`
- Fill with `df.fillna(value)`

Changing Data Types

- Use `df[‘col’].astype(type)` to change type
- Convert object to int, float, string, category, etc.

Renaming Columns

- Use

```
df.rename(columns={‘old’: ‘new’})
```

- Useful for standardising column names

Saving to CSV

- Use

```
df.to_csv(‘output.csv’,
index=False)
```

- Save your cleaned or processed data to file

Activity: Export Modified Data

- Save filtered or modified dataset to a new file

Final Challenge

- Go to kaggle.com
- Create an account if you don’t have one already
- Find and download a dataset you’re interested in
- Try out some of the techniques discussed in this lecture

Best Practices

- Always inspect your data before analysis
- Use meaningful column names
- Avoid modifying DataFrames in place unless needed
- Use chaining and functions to keep code clean

Recap and Next Week

- You’ve learned to load, explore, and clean tabular data
- Next: deeper analysis and basic visualisation with pandas and matplotlib

Questions?

OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL
UNIVERSITY DUBLIN