Algorithms and Data Structures

Python Packages for Data Science

Agenda

- What are Python packages
- Packages specifically for Data Science
  - Numpy
  - Pandas
  - Matplotlib
  - Seaborn

What Is Pandas

- Pandas is a Python library, the name is derived from "panel data"
- Provides data structures and operations for manipulating tables and time series.
- Solves five tasks in data science:
  - Load
  - Prepare
  - Manipulate
  - Model
  - Analyze

Pandas Data Structures

- Pandas provides 3 data structures which all are built upon Numpy arrays:
  - Series (1 dimension)
    - Must contain same type of data; size immutable; value mutable
  - DataFrames (2 dimensions)
    - A container of series
    - May contain different types of data; size/value mutable
  - Panel (3 dimensions)
    - A container of DataFrames
    - May contain different types of data; size/value mutable

Series

- Creating Series:
  - From array object (one dimension only)
    - s1 = pd.Series(np.array())
  - From dictionary object
    - s2 = pd.Series({})
  - From a scalar value
    - s3 = pd.Series(x, index = [])

Series

- Access data from Series:
  - s = pd.Series({'a' : 1, 'b':2, 'c':3})
  - By integer index:
    - d1 = s[0]
  - By slicing:
    - d2 = s[:2]
  - By label index:
    - d3 = s['c']

Let's Do It

- Google CoLab

DataFrame

- A DataFrame is a two dimensional data structure - a tabular
  - Columns may contain different type of data
  - Size / Value are mutable
  - Rows / Columns are labeled
  - Can perform arithmetic operation (since it is based on Numpy)

DataFrame

- Create a DataFrame
  - From a list (default label for row / column will be np.arange(n))
    - df = pd.DataFrame([1, 2, 3, 4])
  - From a list of list
    - df = pd.DataFrame([['A', 1],['B', 2]])
    - df = pd.DataFrame([['A', 1, 101],['B', 2, 102]], columns = ['Char', 'Value1', 'Value2'])
  - From a dictionary
    - df = pd.DataFrame({'Char': ['A', 'B'], 'Value1': [1, 2], 'Value2':[101, 102]})

DataFrame

- Access value in a DataFrame
    - df = pd.DataFrame([['A', 1, 101],['B', 2, 102]], columns = ['Char', 'V1', 'V2'])
  - Column selection
    - df['Char']
  - Column addition
    - df['V3'] = df['V1'] + df['V2]
  - Column deletion
    - del df['V3'] or df.pop('V3')
  - Row selection
    - df.loc[0] or df.iloc[0]
  - Row addition
    - df.append()
  - Row deletion
    - df.drop()

Functionalities

- sum()
- mean(), median(), mode()
- count()
- std()
- min(), max()
- describe() and info()
- groupby()

Let's Do It

- Google CoLab

Visualization

- plot()
- plot.bar()
- plot.hist()
- plot.box()
- plot.area()
- plot.scatter()
- plot.pie()

Let's Do It

- Google CoLab

Practice

- Create a dataframe df that:
  - It has 6 columns: name, a01, a02, a03, a04, a05
  - It has 500 objects.
  - name is string data type, you can use student+a random id for it
  - a1 - 5 are numeric data type (float), should be in [0, 100], you can use a normal distribution np.random.normal(center, std, size)
- Print the stats of df
- Create a new column called: total= sum(a1, .. a5)
- Plot the stats of df

Let's Do It

- Google CoLab

Thank you!