Work Integrated Learning Programmes

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Computer Science & Information Systems

Machine Learning - Lab sheet - Module 1

Exercise 1 - Data Frames in Pandas and Visualizations in Pandas

1 Objective

The objectives are

- set up the Python environment for Machine Learning projects.
- familiarize with data frames in pandas.
- familiarize with visualization of data frames.

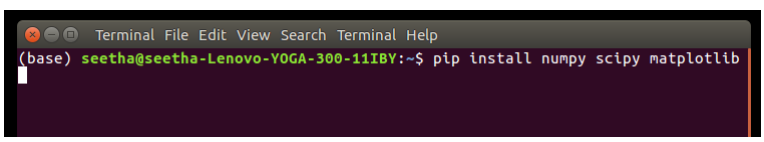
2 Steps to be performed

2.1 Install Anaconda or Miniconda

- To install Anaconda: https://docs.anaconda.com/anaconda/install/.
- To install Miniconda:https://docs.conda.io/en/latest/miniconda.html.

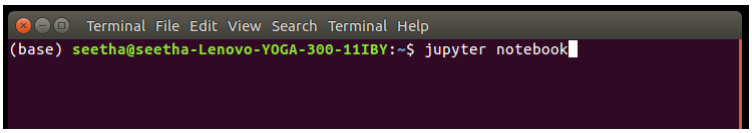
2.2 Install essential libraries

- Open a Linux terminal or Windows command prompt.
- Type the following commands at the prompt.
  - pip install numpy scipy matplotlib
  - pip install pandas seaborn jupyter
  - pip install scikit-learn scikit-image
- Let the installation complete.

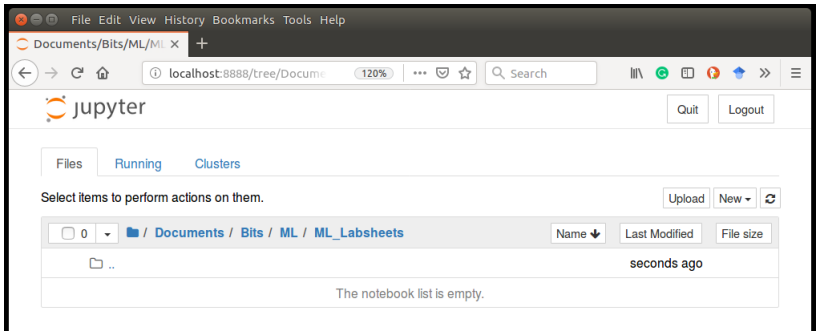


2.3 Check the installation

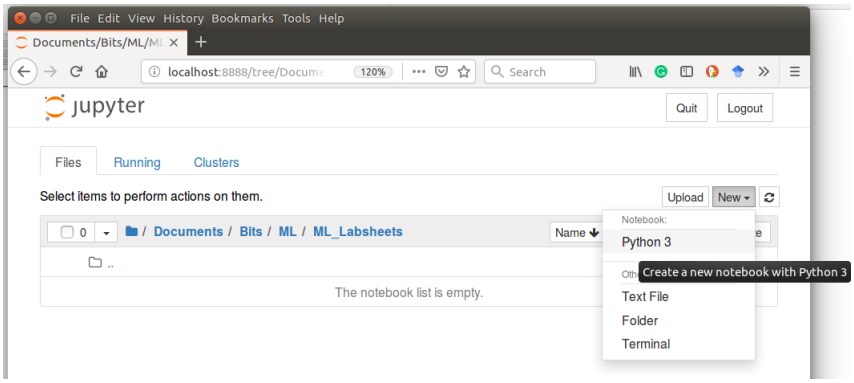
- Open a Linux terminal or Windows command prompt.
- Type jupyter notebook at the prompt.



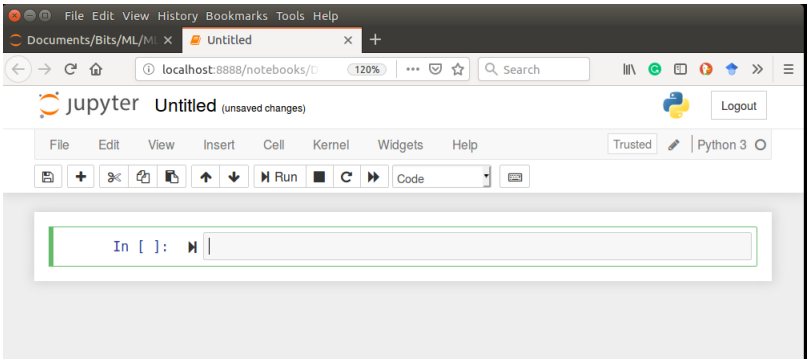
- This will open the default browser, showing the directory or folder structure.
- Navigate to the directory say ML-Labsheets.



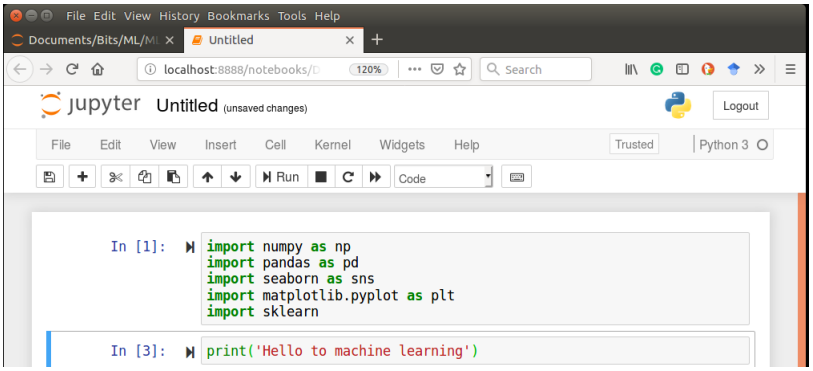
- On the right corner, click on New, choose Python 3.



- This will open up another tab with an empty notebook.



- In the coding cell, type:
  import numpy as np
  import pandas as pd
  import seaborn as sns
  import matplotlib.pyplot as plt
  import sklearn
- To execute the cell, press Control + Enter or Shift + Enter (depending on the OS).



- Type print('Hello to machine learning') and execute.
- Both the statements should execute without giving any errors.

2.4 Import dataset and represent as dataframe

- In the coding cell, type the following and execute.
  import pandas as pd
  imdbDF = pd.read_csv('imdb.csv')

2.5 Perform basic operations / manipulations on dataframe

- Print the number of rows and columns.
  print (imdbDF.shape)
- Print the name of columns.
  print (imdbDF.columns)
- Print the first 9 rows data for inspection.
  print (imdbDF.head(9))
- Accessing a field 'wins' of imdbDF.
  imdbDF.wins
- Access only columns 'wins' and 'nominations'.
  imdbDF[['wins', 'nominations']]
- Access only a first 5 rows of columns 'wins' and 'nominations'.
  imdbDF[['wins', 'nominations']].head()
- Create a new dataframe from columns 'wins' and 'nominations'.
  myDF = imdbDF[['wins', 'nominations']]

- Access first row of myDF.
  myDF.iloc [0]
- Acess rows 1,3,5 of myDF.
  myDF.iloc [[1,3,5]]
- Access rows from 16 to 21 of myDF.
  myDF.iloc [16:22]
- Access rows from 16 to 21 only for 'wins' of myDF.
  myDF.iloc [16:22]['wins']
- Access a particular value in the myDF at [0,1].
  myDF.iloc [0,1]

2.6 Perform some statistical analysis using Pandas.

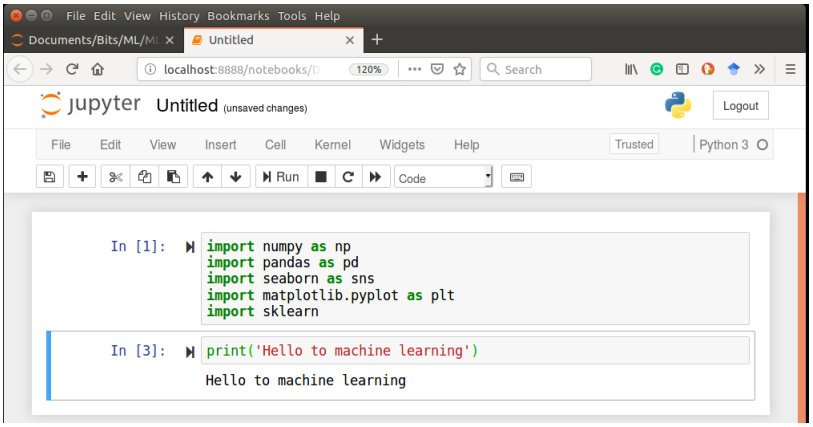
- Print the summary of data set.
  print (imdbDF.info())
- Print the statistical summary of all numerical attributes.
  print (imdbDF.describe())

2.7 Perform some visual analysis using Pandas.

- Plot the 'reviews' column.
  tempDf.plot()
- Plot histogram of 'reviews' column.
  tempDf.plot.hist()

3 Results

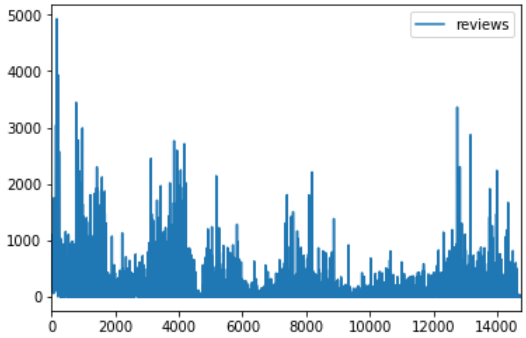
- The text "Hello to machine learning" should be printed successfully.



- Plot of the 'reviews' column.

```
tempDf.plot()
```

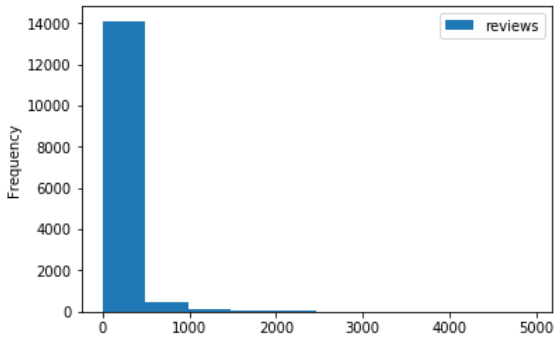
```
<matplotlib.axes._subplots.AxesSubplot at 0x1eb6790c5f8>
```



- Histogram of 'reviews' column.

```
tempDf.plot.hist()  #7.3 call hist under plot module.
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1eb663130f0>
```



4 Observation

- The Python environment is ready for Machine Learning projects.
- Knowledgeable about Python Pandas and Dataframe.

5 References

5.1 Jupyter Notebook References

- Running the Jupyter Notebook
- Jupyter Notebook for Beginners

5.2 Python Pandas References

- Python Pandas
- Python Pandas visualization