Mining Massive datasets – Lab introduction

The main goal for the laboratory is to gain familiarity with the different analysis showed during the course. We will implement different algorithms in Python. There are different possible ways to work with Python:

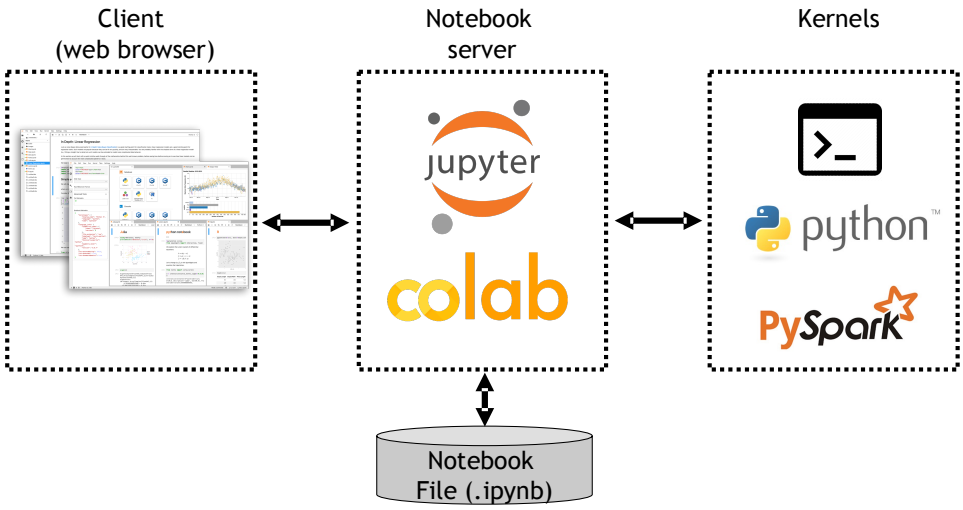
1. Through command line interface (CLI)
    - Just type "python" on a terminal (Linux) or on the command prompt (Windows)

2. Using an Integrated Development Environment (IDE)
    - For instance, PyCharm: https://www.jetbrains.com/pycharm/

3. Using an integrated environment based on ***Notebooks***

With options 1 and 2 it is possible to prepare a Python script and launch the script. While these options are probably the best way to manage large projects, the labs have been prepared using Notebooks. For an introduction to Jupyter Notebooks, please refer to https://jupyter.org/

The figure below summarizes the components of the system required to work with notebooks.



We will provide the notebook files (.ipynb), along with the sample (small) dataset. For the notebook **server** and **kernel**, there are two options:

1. Standalone version: you download and install all the required components on your computer.

2. Online version: you use online services, e.g., offered by Google.

In the following, we provide the instructions for both cases.

Standalone version

For the notebooks, the most used tool is **Jupyter**, which almost comes by default with any Python distribution. An example of a Python distribution, which contains NumPy, SciPy, Pandas, matplotlib, and Tensorflow is Anaconda

https://www.anaconda.com/distribution/

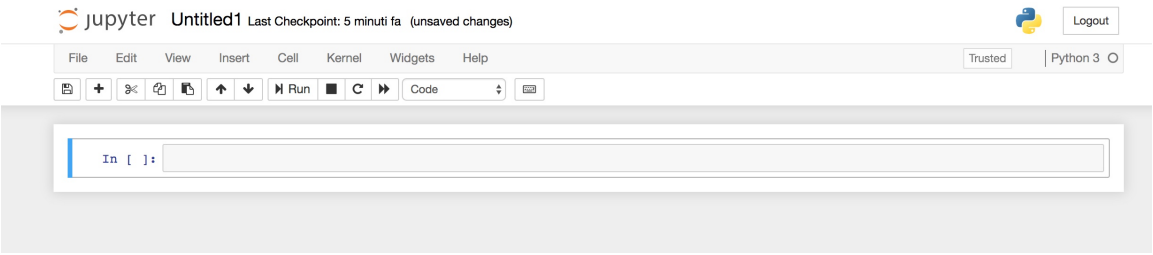
To launch a Jupyter notebook with Python kernel, type

```
jupyter notebook
```

This will open the browser pointing to the Jupyter server, e.g.:

If you want to check that everything is working correctly, create an empty notebook and test some basic code. In particular, in the top-right, select “New”, then, under “Notebook” select “Python3”

This will open a new notebook like this



You can put Pyhton code in the cell, and then hit the “Run” button that you can find in the top row with commands.

If everything works, close the notebook and shut it down. Then create folders with the notebooks with the lab instructions and the datasets, so you can open and run them.

Online version

If you have a Google account, you can use the Colab tool.

https://colab.research.google.com/notebooks/intro.ipynb

To open a new notebook, click on “File” then “New notebook”.

On Google Drive, create folders with the notebooks with the lab instructions and the datasets. To open an existing notebook, then click on “File” then “Open notebook”. The notebook will contain the instruction to mount the Google Drive, so you can read the dataset from within the notebook.