1. Introduction to Python and Spark

Welcome to the first laboratory of Internet Analytics! This lab is an introductory tutorial to Python and Spark. We will cover the (very) basics of the language, as well as starting to use it for Data Science. Spark, a distributed processing framework which will allow us to operate on large datasets, will be also introduced. Finally, you will start using your newly acquired skills to extract information from networks.

1.1 General Instructions

Our labs are designed to be performed on the provided cluster. However you can use your own machine to run the notebooks that do not require spark. To run on the provided cluster, log in to the server `iccluster031.iccluster.epfl.ch` using your Gaspar credentials, and upload the assignments. To run the notebooks locally you will need the following:

- Python 3.10 or later installed with Anaconda
- Install the packages matplotlib, numpy, scipy, gensim, networkx, scikit-learn, and bokeh.

1.2 Python

The goal of this laboratory is for you to discover the language that we are going to use throughout this class: Python. However, in order to make it more interactive and to introduce you tools that will help you in all sort of tasks later, we are using the concept of Jupyter Notebook. A Jupyter Notebook is an HTML-based notebook which allows you to create and share documents that contain live code, equations, visualizations and explanatory text. It allows a clean presentation of computational results as HTML or PDF reports and is well suited for interactive tasks such as data cleaning, transformation and exploration, numerical simulation, statistical modeling, machine learning and more. It runs everywhere (Window, Mac, Linux, Cloud) and supports multiple languages through various kernels, e.g. *Python*, R, Julia, Matlab.

For all these reasons, data scientists use this tool everyday to explore and exploit data. In this class, you will use them to do your lab exercises. The reports you will have to hand in will be in the form of a notebook. So let us discover this environment!

Exercise 1.1 Python and Jupyter Notebook.
Open the file `assignment/1-python.ipynb` and follow the instructions.

Hint: Do not hesitate to ask the teaching assistants if you need help! ■

1.3 Data Science

The technique that you are going to learn in this class are widely used in Data Science. We introduce in this exercise the basic libraries to move around and process data in Python.

Exercise 1.2 Python for Data Science.
Open the file `assignment/2-data-science.ipynb` and follow the instructions. ■

1.4 Spark (requires cluster)

When dealing with large datasets, it usually becomes impossible to process data on a single machine. Spark provides an interface for distributed programming.

Exercise 1.3 Distributed programming with Spark.
Open the file `assignment/3-spark.ipynb` *on the cluster* and follow the instructions.
■

1.5 Networks

As seen in class, networks are a natural approach to model numerous problems and systems. Their mathematical abstraction, the so-called *graphs*, have been well studied and provide a lot of interesting and useful theoretical properties. The next laboratory will focus on networks, so let us start analyzing them.

Exercise 1.4 Introduction to networks.
Open the file `assignment/4-networks.ipynb` and follow the instructions. ■