

Laborator 1 Big Data – master anul 1

Tehnologii folosite în cadrul laboratorului de semestru

- Limbajul Python
- Spark
- Tensorflow

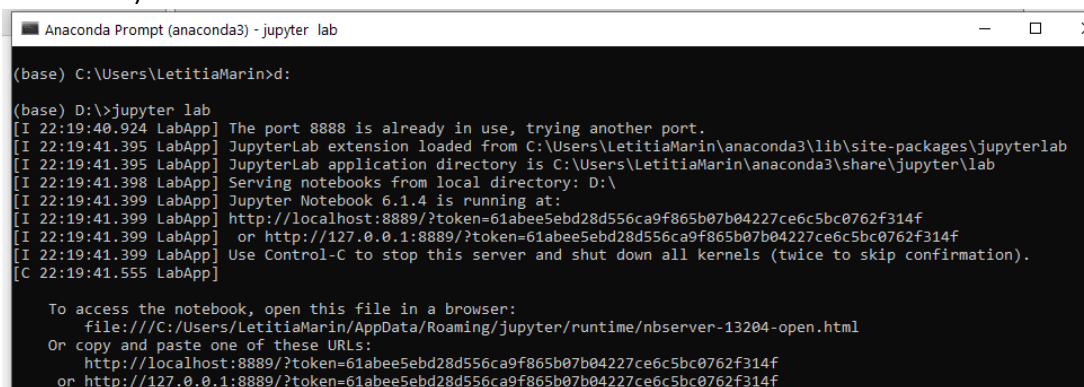
Planul laboratorului 1

- Prezentarea mediului de lucru și instalarea acestuia
- Exerciții – introducere în limbajul Python
- Exerciții – concepte de bază Spark, RDD-uri și operații

Mediul de lucru

Instalare locală

- Instalare Anaconda Navigator:
<https://www.anaconda.com/products/individual>
- Lansare *jupyter lab*
- Doar în cazul în care dorim să lucrăm pe altă partiție:
 - Lansăm Anaconda Prompt
 - Executăm comenzile de mai jos și lăsăm fereastra Anaconda Prompt deschisă (doar o minimizăm) :



```
Anaconda Prompt (anaconda3) - jupyter lab

(base) C:\Users\LetitiaMarin>d:
(base) D:\>jupyter lab
[I 22:19:40.924 LabApp] The port 8888 is already in use, trying another port.
[I 22:19:41.395 LabApp] JupyterLab extension loaded from C:\Users\LetitiaMarin\anaconda3\lib\site-packages\jupyterlab
[I 22:19:41.395 LabApp] JupyterLab application directory is C:\Users\LetitiaMarin\anaconda3\share\jupyter\lab
[I 22:19:41.398 LabApp] Serving notebooks from local directory: D:\
[I 22:19:41.399 LabApp] Jupyter Notebook 6.1.4 is running at:
[I 22:19:41.399 LabApp] http://localhost:8889/?token=61abee5ebd28d556ca9f865b07b04227ce6c5bc0762f314f
[I 22:19:41.399 LabApp] or http://127.0.0.1:8889/?token=61abee5ebd28d556ca9f865b07b04227ce6c5bc0762f314f
[I 22:19:41.399 LabApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 22:19:41.555 LabApp]

To access the notebook, open this file in a browser:
    file:///C:/Users/LetitiaMarin/AppData/Roaming/jupyter/runtime/nbserver-13204-open.html
Or copy and paste one of these URLs:
    http://localhost:8889/?token=61abee5ebd28d556ca9f865b07b04227ce6c5bc0762f314f
    or http://127.0.0.1:8889/?token=61abee5ebd28d556ca9f865b07b04227ce6c5bc0762f314f
```

Instalare pySpark

- Lansăm Anaconda Prompt (separat, dacă acesta a fost lansat deja și rulează, precum indicat mai sus)
- Executăm comenzile următoare, pentru upgrade și instalare pySpark :

```
Administrator: Anaconda Prompt (anaconda3)

(base) C:\WINDOWS\system32>pip install --upgrade pip
Requirement already satisfied: pip in c:\users\letitiamarin\anaconda3\lib\site-packages (21.0.1)

(base) C:\WINDOWS\system32>pip install pyspark
Collecting pyspark
  Downloading pyspark-3.0.1.tar.gz (204.2 MB)
    | 204.2 MB 6.4 MB/s
Collecting py4j==0.10.9
  Downloading py4j-0.10.9-py2.py3-none-any.whl (198 kB)
    | 198 kB ...
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.0.1-py2.py3-none-any.whl size=204612247 sha256=ef05656ea624e219dd810acc7998e3c7dc2163d831cd1bde02f373c35c34b76a
  Stored in directory: c:\users\letitiamarin\appdata\local\pip\cache\wheels\ea\21\84\970b03913d0d6a96ef51c34c878add0de9e4ecbb7c764ea21f
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9 pyspark-3.0.1

(base) C:\WINDOWS\system32>
```

Alternativa Google: <https://colab.research.google.com/>

```
+ Code + Text RAM Disk Editing ^

[1] from time import time
    from pyspark import SparkContext

-----
ModuleNotFoundError                               Traceback (most recent call last)
<ipython-input-1-e16718cb79db> in <module>()
      1 from time import time
----> 2 from pyspark import SparkContext

ModuleNotFoundError: No module named 'pyspark'

-----
NOTE: If your import is failing due to a missing package, you can
manually install dependencies using either !pip or !apt.

To view examples of installing some common dependencies, click the
"Open Examples" button below.

OPEN EXAMPLES SEARCH STACK OVERFLOW

pip install pyspark

Collecting pyspark
  Downloading https://files.pythonhosted.org/packages/f0/26/198fc8c0b98580f617cb03cb298c6056587b8f0447e20fa40c5b634ced77/pyspark-3.0.1.tar.gz (204.2MB)
    | 204.2MB 60kB/s
Collecting py4j==0.10.9
  Downloading https://files.pythonhosted.org/packages/9e/b6/6a4fb90cd235dc8e265a6a2067f2a2c99f0d91787f06aca4bcf7c23f3f80/py4j-0.10.9-py2.py3-none-any.whl (198kB)
    | 204kB 52.9MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.0.1-py2.py3-none-any.whl size=204612242 sha256=e67c7f0f9017e0c45cd882aa14353a6d37acb7990903b4fa411bb12487c36f9
  Stored in directory: /root/.cache/pip/wheels/5e/bd/07/031766ca628adec8435bb40feb0d83bb676ce65ff4007f8e73f
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9 pyspark-3.0.1
```