Universitat Oberta de Catalunya

uoc.edu

M2.955 Models avançats de mineria de dades

Preparació de l'entorn: Instal·lació de Python i IPython

1. Accés a un entorn de pràctiques remot

En l'actualitat hi ha multitud de serveis o plataformes per a executar codi Python en remot. Qualsevol d'aquestes plataformes pot ser una bona solució per executar les activitats d'aquesta assignatura. L'únic requisit és que han de tenir instal·lades les principals llibreries que utilitzarem per a aquestes activitats (en general, es tracta de llibreries àmpliament utilitzades i molt populars, que solen estar presents per defecte en la majoria d'aquestes plataformes).

D'entre totes les opcions, en destaquem dues:

Jupyter Hub

És una eina basada en Jupyter notebooks per a plataformes multiusuaris, coneguda com Jupyter Hub (https://jupyter.org/hub). En aquest cas té incorporats nuclis per treballar amb sintaxi en Python i R.

Aquesta plataforma és pròpia de la UOC, i per accedir només hauries de demanar al professor responsable de l'assignatura (PRA) que us faciliti les credencials d'accés. La URL per accedir-hi és:

https://jpthub.eimt.uoc.edu/hub/login

Google Research Colab

Google ha posat a disposició del públic en general un servei de notebook en Python, amb capacitat per a connectar amb TensorFlow i amb possibilitat d'ús de GPU / TPU, encara que en les versions gratuïtes del servei no es garanteix la disponibilitat d'aquestes.

La URL per accedir-hi és:

https://colab.research.google.com

2. Preparació de l'entorn de pràctiques

Aquest document ens guiarà en la instal·lació de tot el programari necessari per poder desenvolupar les activitats pràctiques d'aquesta assignatura. Com es veurà més endavant en aquest mateix document, l'entorn es basa en el llenguatge de programació Python i les llibreries necessàries per poder manipular, analitzar i visualitzar dades.

Per a la realització de les activitats d'aquesta assignatura cal que cada estudiant instal·li i configuri al seu ordinador el llenguatge Python, les llibreries necessàries i (opcionalment) un entorn de desenvolupament integrat (IDE).

Instal·lació de Python i IPython

Per a aquesta i les següents activitats, utilitzarem Python 3.6, que està disponible per a sistemes Linux, Mac OS X i Windows en el lloc web oficial de Python:
https://www.python.org/downloads/

Hi ha dues versions principals per triar, Python 3.x i Python 2.x. La versió 2.x, tot i que encara molt utilitzada es manté per motius de compatibilitat amb programari existent. En el nostre cas optarem per l'última versió disponible de la sèrie 3.x.

Després de la instal·lació de Python, podem comprovar el seu correcte funcionament accedint a la terminal de Python com es mostra a continuació:

```
$ python3
Python 3.6.2 (v3.6.2:5fd33b5926, Jul 16 2017, 20:11:06)
[GCC 4.8.2] on Linux
Type ``help'', ``copyright'', ``credits'' or ``license'' for
more information.
>>> print(``Hello, world!'')
Hello, world!
>>> exit()
```

IPython és una plataforma per al desenvolupament de Python que conté una sèrie d'eines i entorns per executar Python i conté, a més, funcions addicionals a l'intèrpret estàndard. Entre d'altres, aquesta plataforma conté el potent IPython Notebook[1], que et permet escriure programes en un navegador web. També formata el teu codi,

[1] Més informació: https://ipython.org/notebook.html

mostra la sortida i et permet anotar els teus *scripts*. És una eina útil i interessant per explorar conjunts de dades i realitzar anàlisis més o menys senzills. En el cas d'executar *scripts* complexos que puguin portar hores d'execució, aconsellem utilitzar l'intèrpret estàndard de Python.

Per instal·lar IPython en el nostre sistema, hem d'escriure el següent en la línia d'ordres del sistema operatiu (no en l'intèrpret de Python):

```
$ pip3 install ipython[all]
```

Es necessiten privilegis d'administrador per instal·lar-lo en el sistema i que sigui accessible per a tots els usuaris. Si no volem (o no podem) fer canvis en tot el sistema, podem instal·lar-lo només per a l'usuari actual executant la següent comanda:

```
$ pip3 install --user ipython[all]
```

Això instal·larà el paquet IPython en una ubicació específica de l'usuari; podrem utilitzar-la, però ningú més en el nostre equip podrà executar-lo.

Per a una informació detallada sobre la instal·lació, podem referir-nos a la documentació oficial[2] per obtenir instruccions d'instal·lació més detallades.

Un cop la plataforma ha estat instal·lada en el nostre sistema, podem executar-la mitjançant l'ordre:

```
$ ipython3 notebook
```

Això farà dues coses. En primer lloc, crearà una instància de IPython Notebook que s'executarà en l'intèrpret de comandes que acaba d'utilitzar. En segon lloc, llançarà el seu navegador web i es connectarà a aquesta instància, el que li permetrà crear un nou projecte de treball, com es pot veure a la figura 1.

[2] Disponible a http://ipython.org/install.html

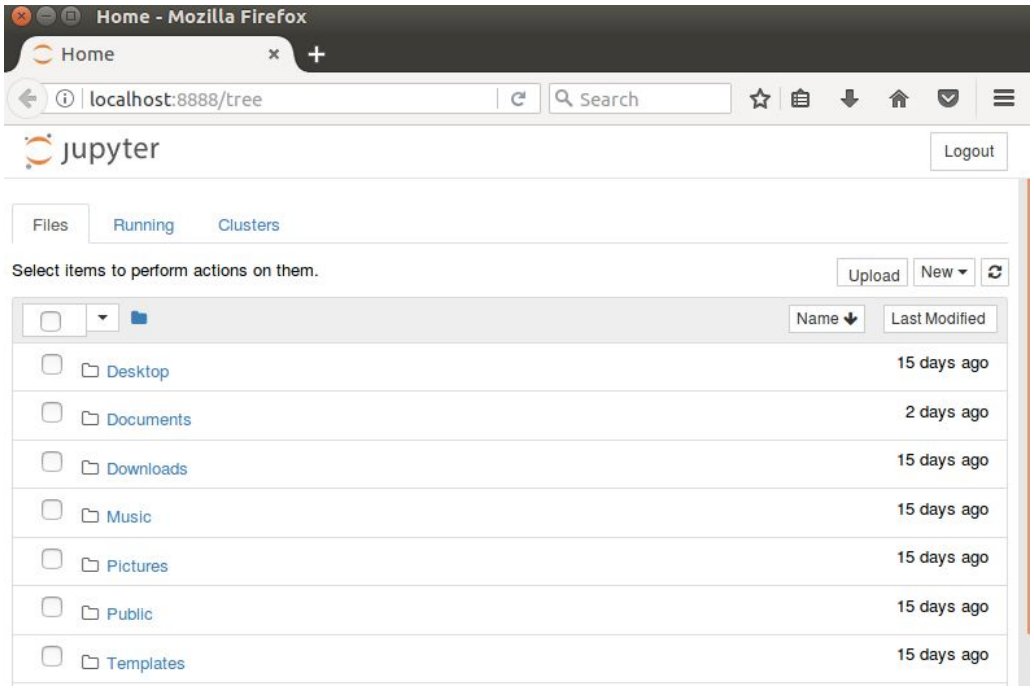

Figura 1. Interfície web de IPython notebook (Jupyter notebook)

Alternativament, es pot instal·lar una distribució de Python anomenada Anaconda[3]. Aquesta distribució inclou moltes de les llibreries utilitzades en ciència de dades i *machine learning*, a més de la interfície web basada en IPython i un entorn integrat per a la gestió dels serveis i paquets.

Instal·lació de les llibreries

Com hem comentat anteriorment, Python inclou un programa anomenat *pip*, que és un paquet extremadament útil que ens ajudarà a instal·lar noves llibreries en el nostre sistema. Es pot verificar que *pip* està instal·lat al nostre sistema executant la següent comanda:

```
$ pip3 freeze
```

Que ens indica que paquets tenim instal·lats al nostre sistema.

La instal·lació de qualsevol llibreria és un procés molt senzill emprant l'eina *pip*. Per instal·lar qualsevol llibreria només haurem d'indicar l'opció "install" i el nom de la llibreria que desitgem instal·lar.

[3] Disponible a https://www.anaconda.com/distribution/

Per exemple, per instal·lar la llibreria "NumPy" executarem la següent comanda des de la línia d'ordres del sistema operatiu:

```
$ pip3 install numpy
```

Les llibreries bàsiques i necessàries per a realitzar aquesta activitat es detallen a continuació:

- numpy
- pandas
- matplotlib
- scikit-learn

Entorn de desenvolupament integrat (IDE)

Podem utilitzar qualsevol editor de text per codificar els programes en Python. Per tant, aquest pas és opcional, i depèn molt de les preferències de cada usuari.

Una primera opció són els entorns "lleugers", com editors de text estàndard que reconeixen la sintaxi de Python i permeten identar i acolorir el codi (la qual cosa és de gran ajuda). En aquest sentit es poden trobar moltíssims editors disponibles i gratuïts (alguns) amb una simple recerca per Internet. No farem recomanacions en aquest sentit.

D'altra banda, hi ha entorns de desenvolupament integrats que van molt més enllà de la simple edició de codi. Aquests IDE permeten, a més de facilitar la codificació, l'ús de *debuggers* i altres eines que faciliten el desenvolupament de projectes i codi. Entre d'altres, hi ha l'entorn comercial PyCharm[4], que disposa d'una versió *community* i una versió *professional* (que incorpora funcions addicionals i pot ser utilitzada de forma gratuïta per estudiants i investigadors del món acadèmic).

Una altra opció interessant és Spyder[5], que ofereix una interfície similar al conegut RStudio, però per al llenguatge de programació Python. A més, en aquest cas s'ofereix sota una llicència d'ús lliure.

[4] Més informació: https://www.jetbrains.com/pycharm/
[5] Més informació: https://github.com/spyder-ide/spyder