[DM 2024] Lab 1 Environment settings

Hi everybody.

We will have our first lab session on September 23 (Monday) 9:00 AM on our Youtube Stream / Classroom. Please be on time.

We highly recommend you to attend the session with your personal laptop (that way you'll also have your environment set for the homework). These are some instructions for you to set up the environment:

1.A. Install Python:

We are using Python 3.9.6. You can use newer version, but use at your own risk. Follow the installation instructions below:

https://www.python.org/downloads/release/python-396/

Once you install the python, you can check whether your python is ready by opening a "terminal" windows (Linux/MacOS) or a "Command Prompt" window.

1.B. Set up a Virtual Environment:

If you use Python a lot for other projects, you might need to isolate different installation packages for compatibility purposes. If this is the first time you use Python or if you don't encounter any compatibility issues, you can skip this step.

To have virtual environments, you need to install Anaconda or virtualenv. Here are the instructions for how to install and use each one:

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

https://virtualenv.pypa.io/en/latest/installation.html

2. Install libraries:

We will use the following Python libraries for the lab: Jupyter Notebook, Scikit learn, Numpy, Pandas, NLTK, Matplotlib, Plotly and PAML

Once you have installed Python 3 (and optionally Anaconda), open a "terminal" windows (Linux/MacOS) or a "Command Prompt" window and type the following commands followed by "Enter":

```
pip3 install jupyter
pip3 install numpy
pip3 install pandas
pip3 install matplotlib
pip3 install plotly
pip3 install nltk
pip3 install scikit-learn
pip3 install seaborn
pip3 install paml
pip3 install umap-learn
```

Open a "Terminal" window (Linux/MacOs) or a "Command Prompt" window (windows) and type the following line (followed by Enter): python (or python3)

Then type the following lines (followed by Enter): import nltk nltk.download('punkt')

Note: you might need to add "Sudo" at the beginning if you use Linux, and you will be asked to input your password.

3. Run Jupyter Python and check your environment:

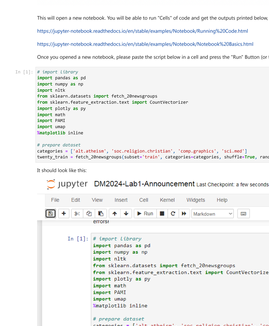
Open a new Jupyter notebook server. In order to do this, open a "terminal" windows (Linux/MacOS) or a "Command Prompt" window and type the following commands followed by "Enter":

```
jupyter notebook
```

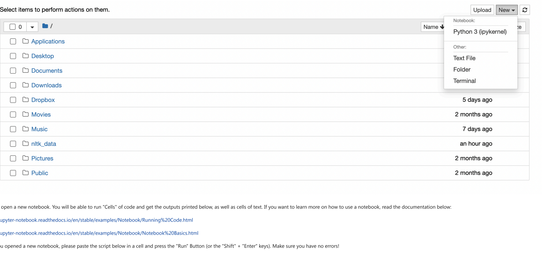
If you receive an error message, zsh: command not found: jupyter, type the following commands instead.

```
python3 -m notebook
```

Just like the image below:



A window like the one below should open in your browser. Please go to the "New" button on the top right cornered select "Python 3":



This will open a new notebook. You will be able to run "Cells" of code and get the outputs printed below, as well as cells of text. If you want to learn more on how to use a notebook, read the documentation below:

https://jupyter-notebook.readthedocs.io/en/stable/examples/Notebook/Running%20Code.html

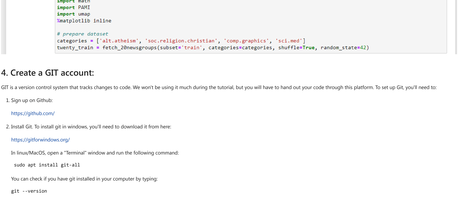
https://jupyter-notebook.readthedocs.io/en/stable/examples/Notebook/Notebook%20Basics.html

Once you opened a new notebook, please paste the script below in a cell and press the "Run" Button (or the "Shift" + "Enter" keys). Make sure you have no errors!

```
In [1]:
# import library
import pandas as pd
import numpy as np
import nltk
from sklearn.datasets import fetch_20newsgroups
from sklearn.feature_extraction.text import CountVectorizer
import plotly as py
import math
import PAML
import umap
%matplotlib inline

# prepare dataset
categories = ['alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.med']
twenty_train = fetch_20newsgroups(subset='train', categories=categories, shuffle=True, random_state=42)
```

It should look like this:



4. Create a GIT account:

GIT is a version control system that tracks changes to code. We won't be using it much during the tutorial, but you will have to hand out your code through this platform. To set up Git, you'll need to:

1. Sign up on Github:

   https://github.com/

2. Install Git: To install git in windows, you'll need to download it from here:

   https://gitforwindows.org/

   In linux/MacOS, open a "Terminal" window and run the following command:

   ```
   sudo apt install git-all
   ```

   You can check if you have git installed in your computer by typing:

   ```
   git --version
   ```

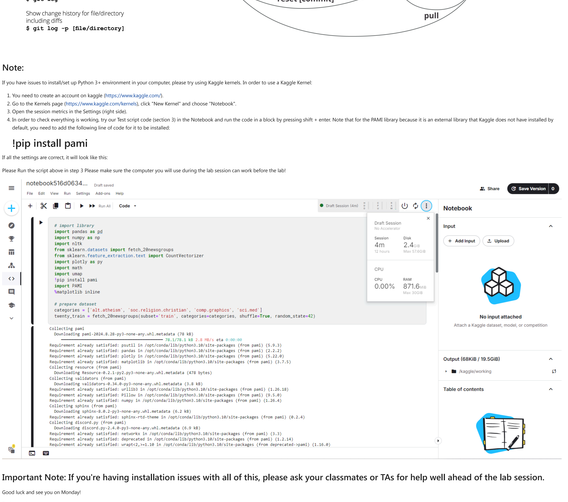
3. Add your "credentials" to git. Open a "Terminal" window and run the following commands(please use the same email you used to register for your Git account):

   ```
   git config --global user.name "YOUR_USERNAME"
   git config --global user.email "your_email@gmail.com"
   ```

It should look like this:

Note: If you're not comfortable with the command line, you may also download the Sourcetree application: https://desktop.github.com/

Here's a cheatsheet of commands(for those who won't use the command line and not the app):



Note:

If you have issues to install/set up Python 3+ environment in your computer, please try using Kaggle kernels. In order to use a Kaggle Kernel:

1. You need to create an account on kaggle (https://www.kaggle.com/).
2. Go to the Kernels page (https://www.kaggle.com/kernels), click "New Kernel" and choose "Notebook".
3. Open the session metrics in the Settings (right side).
4. In order to check everything is working, try our Test script code (section 3) in the Notebook and run the code in a block by pressing shift + enter. Note that for the PAML library because it is an external library that Kaggle does not have installed by default, you need to add the following line of code for it to be installed.

!pip install pami

If all the settings are correct, it will look like this:

Please Run the script above in step 3 Please make sure the computer you will use during the lab session can work before the lab!

Important Note: If you're having installation issues with all of this, please ask your classmates or TAs for help well ahead of the lab session.

Good luck and see you on Monday!

Best regards,
The TAs