Py4SHS 2024

PARTICIPANT GUIDE

CONTENTS

PRACTICAL INFORMATION

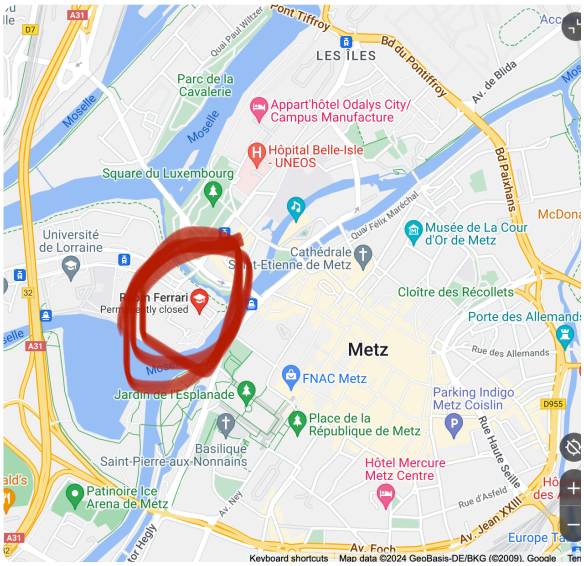
Daily schedule:

- 9-9:30 AM: Daily breakfast;
- 9:30-12:30 AM: Lecture;
- 2-5:30 PM: Lab.

Contact in case of issues: sophie.robert@univ-lorraine.fr // +33 06 09 64 69 81

Accessing the summer school:

All the event will take place in the *Salle Ferrari* (see map below, also available on Google Maps).



INSTALLING PYTHON

On Windows

Step 1: Download Python

1. Open your web browser and go to the official Python website: `https://www.python.org/downloads/`.
2. Click the button that says **Download Python 3.x.x**, where `x.x` is the latest version.

Step 2: Run the Installer

1. Once the download is complete, open the installer.
2. Before clicking **Install Now**, check the box that says **Add Python 3.x to PATH**.
3. Click **Install Now** and follow the on-screen instructions.

Step 3: Verify the Installation

1. Open the Command Prompt by searching for `cmd` in the Start menu.
2. Type the following command and press `Enter`:

```
python --version
```

3. You should see the version of Python that you installed. This confirms that Python is successfully installed.

On macOS

Python is often installed by default on MacOS, first try checking the install by trying:

```
python[3] --version
```

Step 1: Download Python

1. Open your web browser and go to the official Python website: `https://www.python.org/downloads/mac-osx/`.
2. Click the button that says **Download Python 3.x.x**, where `x.x` is the latest version.

Step 2: Run the Installer

1. Open the downloaded `.pkg` file to launch the Python installer.
2. Follow the instructions in the installer, clicking **Continue** and **Agree** when prompted.
3. The installer will also offer to install the Python Launcher and IDLE, the integrated development environment for Python. It's recommended to install these as well.
4. Once the installation is complete, the installer will display a summary page. Click **Close**.

Step 3: Verify the Installation

1. Open the Terminal application.
2. Type the following command and press `Enter`:

```
python3 --version
```

3. You should see the version of Python that you installed. This confirms that Python is successfully installed on your system.

On Linux

Python is often installed by default on most linux install, first try checking the install by trying:

```
python[3] --version
```

Step 1: Update Package Lists

1. Open your Terminal.
2. Update your package lists with the following command:

```
sudo apt-get update
```

Step 2: Install Python

1. Run the following command to install Python:

```
sudo apt-get install python3
```

Step 3: Verify the Installation

1. Open Terminal.
2. Type the following command and press `Enter`:

```
python3 --version
```

3. You should see the version of Python that you installed.

DAY 1

INTRODUCTION TO DIGITAL HUMANITIES AND PYTHON

Speakers: Sophie Robert-Hayek, Céline Lemarinier.

Lecture - 9:30-12:30 AM

- Understand this week's goals and requirements;
- Understand the possible classification of digital humanities project;
- Discover a wide range of existing Digital Humanities project;
- Have an understanding of what is the Python language.

Lab - 02:00-5:30 PM

- Have a working installation of Python and Jupyter Notebook;
- Be able to install packages within a Python environment;
- Manipulate basic data structures: list, dictionaries, sets;
- Manipulate basic operators: for loops, if/else conditions . . . ;
- Understand the concept of a function;
- Understand the basics of OOP and write your own class.

Bibliography

- Python documentation: `https://docs.python.org/3/`;
- OSullivan, James , ed. *The Bloomsbury Handbook to the Digital Humanities*. London,: Bloomsbury Academic, 2022. Bloomsbury Handbooks. Bloomsbury Collections.

DAY 2

MANIPULATING DATA USING PYTHON

Speakers: Céline Lemarinier.

Lecture - 9:30-12:30 AM

- Understand the concept of dataset and variables;
- Understand the principles behind univariate and multivariate analysis;
- Learn about the most adequate statistical estimators for univariate and multivariate analysis;
- Learn the different available plot types and select the most adequate per variable type;
- Understand the basics of data manipulation in Python using the PyData ecosystem: pandas, numpy and seaborn.

Lab - 02:00-5:30 PM

- Load two CSV datasets into RAM using Python containing Shakespeare's Romeo and Juliet;
- Perform data manipulation on the datasets using pandas;
- Compute statistical indicators on dataset;
- Perform univariate and multivariate analysis using seaborn plots to understand data relationship.

Bibliography

- *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*, Wes McKinney, 2022, O'Reilly Media, Inc.
- Pandas documentation: `https://pandas.pydata.org/`.
- Seaborn documentation: `https://seaborn.pydata.org/`.

DAY 3

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

Speakers: Sophie Robert-Hayek

Lecture - 9:30-12:30 AM

- Reminders on datasets and variables;
- Understand the basics of Machine Learning and the distinction between supervised and unsupervised learning;
- Understand the necessity of textual embedding;
- Discover popular text cleaning method: stop words removal, stemming and lemming;
- Discover term-frequency and inverse term-frequency for textual embedding;
- Understand the basics of language models (BERT, ROBERTA...) for textual embedding.

Lab - 02:00-5:30 PM

- Load into RAM a dataset containing the text of several Shakespeare's plays;
- Statistically analyze dataset;
- Lemmatize the dataset using spacy;
- Compute term-frequency and inverse time frequency to characterize plays;
- Clusterize similar plays and list most frequent term per cluster using *k-means*.
- Project the textual data into an embedding space using Doc2Vec;

6:00-6:30: Presentation of Romain Pierronnet on Open-Source software at the University of Lorraine.

Bibliography

- *Artificial Intelligence: A Modern Approach*, Russel and Norvig, Global Edition, 2014.
- *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python First Edition*, Hobson Lane, Hannes Hapke, Cole Howard, *Manning publishing*, 2019.

- Scikit-Learn documentation: https://scikit-learn.org/stable/index.html
- NLTK documentation: https://www.nltk.org/
- SpaCy documentation: https://spacy.io/

DAY 4

INTRODUCTION TO DEEP LEARNING FOR OCR

Speakers: Mathieu Pister, Sophie Robert-Hayek

Lecture - 9:30-12:30 AM

- Reminders on supervised learning and its application to OCR;
- Introduction to Deep Learning and Neural Networks: perceptron, neural networks, RNN and LSTM;
- Introduction to cross-entropy and accuracy to measure model performance;
- Introduction to Tesseract for OCR recognition.

Lab - 02:00-5:30 PM

- Use Python and Tesseract to numerize content;
- Learn how to preprocess images to improve OCR performances;
- Bonus: train a new tesseract model

End of day session: Mini consulting sessions by Mathieu Pister and Sophie Robert-Hayek.
Gala dinner @ the Teatris.

Bibliography

- *Deep Learning*, Ian Goodfellow and Yoshua Bengio and Aaron Courville, MIT Press, 2016.
- Tesseract documentation: https://github.com/tesseract-ocr/tesseract;
- Bishop, C.M. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer (India) Private Limited, 2013

DAY 5

INTRODUCTION TO STATISTICAL TESTING

Speakers: Sophie Robert-Hayek

Lecture - 9:30-12:30 AM

- Introduction to statistical testing and methodology;
- Introduction to popular univariate tests: Student's t-test, Kruskal-Wallis . . . ;
- Introduction to multivariate approaches and linear regression.

Lab - 02:00-5:30 PM

- Compute statistical estimators on the Pauline letters;
- Perform Student's test on the different use of stop words between Colossians and authentic letters;
- Perform linear regression regarding the likelihood of same stop words use across authentic epistles and Colossians.

Bibliography

- Ott, Lyman. *An introduction to statistical methods and data analysis*. United States: Duxbury Press, 1977.
- Haslwanter, Thomas. *An Introduction to Statistics with Python: With Applications in the Life Sciences*. Germany: Springer International Publishing, 2022.
- Seber, George A. F., Lee, Alan J. *Linear Regression Analysis*. Germany: Wiley, 2012.