

Machine Learning for Text Classification

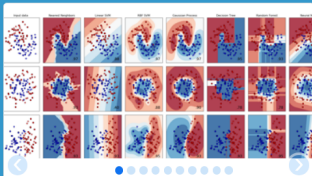
Alejandro Moreo and Fabrizio Sebastiani
`first.last@isti.cnr.it`

ISTI-CNR, Pisa, Italy

9 June 2022

Machine Learning for automatic text analysis: tasks, methods, and tools:

- The Language: Python
- The Toolkit: `scikit-learn`
- The Environment: Jupyter, Google Colab
- Structure:
 - 1st block (1h): Text processing with NLTK
 - 2nd block (1h): From raw text to hyperplanes
 - 3rd block (1h): Authorship Attribution
 - Hands-on exercises (1h): Sentiment Classification
- Concluding Remarks



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.

Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning

Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.

Modules: preprocessing, feature extraction. — Examples

The Environment: Jupyter

jupyter example Last Checkpoint: hace 2 minutos (autosaved)



Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted

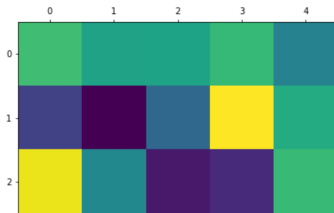
Python 3

Run

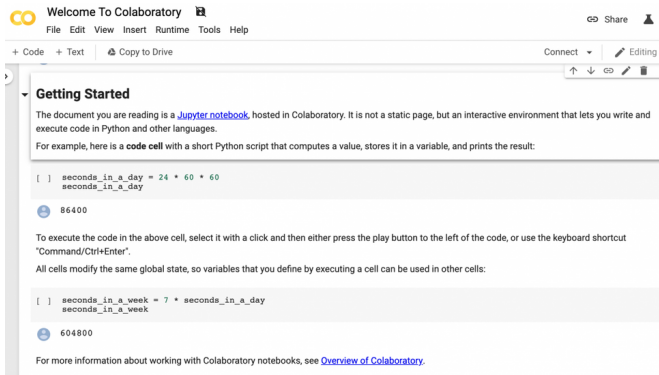
```
In [3]: import numpy as np
X = np.random.rand(3,5)
print(X)
```

```
[[0.66006012 0.55826627 0.56148918 0.63816535 0.44175898]
 [0.22338811 0.05267085 0.34929095 0.92795877 0.59288656]
 [0.90333882 0.46056606 0.11392294 0.15629945 0.64221199]]
```

```
In [5]: import matplotlib.pyplot as plt
plt.matshow(X)
plt.show()
```



The Installation: Google Colab



Welcome To Colaboratory

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

Connect Editing

Getting Started

The document you are reading is a [Jupyter notebook](#), hosted in Colaboratory. It is not a static page, but an interactive environment that lets you write and execute code in Python and other languages.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
[ ] seconds_in_a_day = 24 * 60 * 60
seconds_in_a_day
```

86400

To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command/Ctrl+Enter".

All cells modify the same global state, so variables that you define by executing a cell can be used in other cells:

```
[ ] seconds_in_a_week = 7 * seconds_in_a_day
seconds_in_a_week
```

604800

For more information about working with Colaboratory notebooks, see [Overview of Colaboratory](#).

• The code is accessible through:

- <https://drive.google.com/drive/folders/1KWLGYkckKHJaEo4J-FtaKBdrmAZRJBuN?usp=sharing>

Plan of the Hands-on activities

- We will learn some basic routines for **text processing using NLTK**, an open-access suite of text analytic tools.
- We will explore `scikit-learn`'s tools for **text classification** that instantiate the most important methods described in the lectures. We will create a classifier for the **topic** of a document.
- We will later explore the field of **Authorship Analysis**. We will concentrate on medieval Latin and we will try to figure out if the *Epistle XIII* (one of the most disputed works of Dante) was actually the work of Dante or not.
- Finally, you will try to solve some **exercices** by applying all the concepts and techniques that you will learn today.

... let's get started!