Text analysis with Python

Lecturer: Maria Chiara Debernardi

Language

English

Course description and objectives

In the contemporary digital ecosystem, textual communication remains a predominant modality of information exchange across diverse domains, including corporate documentation, legal and juridical texts, social media interactions, product evaluations, journalistic and media communications. The exponential growth of digital textual content necessitates sophisticated computational methodologies for efficient information extraction and knowledge generation.
This course provides a comprehensive introduction to Natural Language Processing (NLP) through exploration of computational linguistic strategies, textual data processing methodologies, and machine learning techniques. The course employs Python as programming language, complemented by a set of open-source NLP libraries.

At the course's conclusion, participants will:

- understand the steps of text analysis methodologies
- critically differentiate among diverse textual analysis techniques and their strategic applications
- develop proficiency in constructing basic text analysis pipelines utilizing Python programming
- demonstrate capabilities in extracting and processing textual data from web-based sources

Audience

The course is open to all students at Bocconi University. It is aimed at:

- those who want to approach the world of automated text analysis
- those who are interested in facing a Python hot topic in the AI and ML context

Prerequisites

Mandatory knowledge of Python basics, having attended either the curricular course 30424 Computer Science or one of the ITEC's courses "Python start" / "Programming with Python" (or having equivalent knowledge and skills).
Prior knowledge of Statistics and Python's Pandas library are highly welcome.

Guidelines

Registration:

You can sign up for the course only through the yoU@B student Diary, in the " **sign-up for various activities**" box (please note that the box appears only when registrations open. Before then it will not be visible).
You can only cancel your registration by Diary **no later** than the registration deadline for the course itself. No other ways of cancellation are allowed.
Registration will be confirmed a few days before the start of the course through a message posted in the yoU@B student Diary.

Attendance:

- Attendance of **75% or more** of class hours: obtainment of the **Open Badge**
- Attendance of **less than 25%** of class hours: **blacklisting**

Duration

16 hours

Teaching mode

This course will be only taught **in person**. Online mode will not be provided.

Calendar

Lecture	Date	Time	Room
1	Wed 26/03/2025	18.15 - 19.45	5 (Sarfatti)
2	Wed 02/04/2025	18.15 - 19.45	5 (Sarfatti)
4	Wed 09/04/2025	18.15 - 19.45	5 (Sarfatti)
3	Wed 16/04/2025	18.15 - 19.45	5 (Sarfatti)
5	Mon 05/05/2025	18.15 - 19.45	5 (Sarfatti)
6	Wed 07/05/2025	18.15 - 19.45	5 (Sarfatti)

7	Mon 12/05/2025	18.15 - 19.45	5 (Sarfatti)
8	Wed 14/05/2025	18.15 - 19.45	5 (Sarfatti)

Note: lessons will be held in the traditional room and **all the students must bring their own device**.

Syllabus of the course

Lecture	Topics
1	**Building a common ground** - Why NLP in today's world: its applications - Preliminaries - Introduction to Jupyter Notebook - Brief recap of Python basics - Pandas: the essentials *Exercises*
2	**Textual data preparation** - Tokenization: sentences and words - Stop words - Lexicon normalization: Stemming *versus* Lemmatization - POS tagging - N-grams *Exercises*
3	**Preprocessing and text classification** - Bag of words - TF-IDF - Word embedding - Classification methods applied to text *Exercises*
4	**Sentiment analysis** - Issues about sentiment detection - Lexicon-based methods - Rule-based analysis methods - Machine Learning based approach *Exercises*
5	**Web scraping - 1** - What it is - Legal issues - How to do it - Requests - BeautifulSoup *Exercises*

Lecture	Topics
6	**Web scraping - 2** - Selenium - Scrapy - Using APIs (*hints only*) *Exercises*
7	**Text clustering** - Clustering *versus* Classification - Topic detection - Mapping, textual data visualization *Exercises*
8	**What have we learnt?** - Recap - Doubts/issues? *Final exercise*

Software used

Jupyter Notebook inside Anaconda
Anaconda Distribution is a free version suited for students. Currently (February 2025) it supports Python 3.12. It is available for Windows, Mac, and Linux.
You can download it here (skipping the not mandatory registration, if you don't want to provide your email):

https://www.anaconda.com/download/success

Suggested bibliography

Materials, both about NLP theory and the Python packages used in classroom, will be provided by the teacher during the course and will be available on Blackboard.

Available seats

This activity is limited to **110** participants. Registration cannot be carried out once this number has been reached or after the registration period closes.