

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 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 end of the course, participants will be able to:

- 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
- collect textual data from websites or APIs, respecting ethical and legal constraints
- use LLMs for simple text analysis tasks and critically assess their limitations

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 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 means 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: **Exclusion list**

Duration

16 academic hours

Teaching mode

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

Calendar

Lecture	Date	Time	Room
1	Tue 22/09/2026	18:15 - 19:45	202 (Sarfatti)
2	Wed 23/09/2026	18:15 - 19:45	202 (Sarfatti)
3	Tue 29/09/2026	18:15 - 19:45	202 (Sarfatti)
4	Wed 30/09/2026	18:15 - 19:45	202 (Sarfatti)
5	Tue 06/10/2026	18:15 - 19:45	202 (Sarfatti)
6	Wed 07/10/2026	18:15 - 19:45	202 (Sarfatti)
7	Tue 13/10/2026	18:15 - 19:45	201 (Sarfatti)
8	Wed 14/10/2026	18:15 - 19:45	201 (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 - Introduction to Jupyter Notebook - Brief recap of Python basics - Pandas: only the essentials <i>Exercises</i>
2	Textual data preparation - Mining or analytics? - Preprocessing - o Tokenization: sentences and words - o Regex/basic pattern cleaning - o Stop words - o Lexicon normalization: Stemming <i>versus</i> Lemmatization - o POS tagging - o N-grams <i>Exercises</i>
3	From preprocessing to text classification - Converting texts to numbers: - o Bag of words - o TF-IDF - o Word embedding: the basics - Classification methods applied to text - scikit-learn - Model evaluation <i>Exercises</i>
4	Sentiment analysis - Challenges in sentiment detection - Lexicon-based methods - Rule-based methods - Machine learning approach <i>Exercises</i>

Lecture	Topics
5	Collecting textual data from the web - What web scraping is - Ethics, legality, and good practices - HTTP requests - HTML structure - BeautifulSoup - Simple scraping workflow <i>Exercises</i>
6	From scraping to structured datasets - Handling multiple pages - Saving scraped data into structured formats - Selenium: when static scraping is not enough - APIs as a cleaner alternative to scraping <i>Exercises</i>
7	Text clustering - Clustering <i>versus</i> Classification - Text similarity - Topic detection - Mapping/visualization of textual data <i>Exercises</i>
8	LLM models for text analysis - What they are - Local LLMs - Prompting: - o Zero/few shot classification - o Sentiment analysis using GenAI - LLM limitations, privacy, and reproducibility <i>Exercises</i>

Software used

Jupyter Notebook inside Anaconda

Anaconda Distribution is a free version suited for students, available for Windows, Mac, and Linux operating systems. 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 instructor during the course and will be accessible on Blackboard.

Available seats

This activity is limited to **110** participants.

Registration will no longer be possible once this limit has been reached or after the registration period has closed.

Registration may be cancelled exclusively through the yoU@B Diary no later than the registration deadline for the course.