Text analysis with Python

Lecturer: Maria Chiara Debernardi

Language

English

Course description and objectives

The written word is still one of the main means of communication (e.g. business documents - including legal or juridical -, posts on social media, product reviews on the web, press reviews...), therefore the automated processing and analysis of natural language (NLP) has become a fundamental tool for quickly extracting information and knowledge from textual documents.
This course aims to provide students with an introduction to the main statistical techniques for conducting textual analyzes, using Python as programming language and its appropriate libraries of Text Mining (only the ones available for free).

At the end of the course, participants will be able to:

- Understand the text analysis steps
- Distinguish among the diverse types of analysis and their purposes
- Create simple text analysis pipelines using Python
- Understand how to extract textual data from web pages

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.

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	Mon 05/06/2023	14.45 - 16.15	D (Sarfatti)
2	Wed 07/06/2023	14.45 - 16.15	D (Sarfatti)
4	Mon 19/06/2023	14.45 - 16.15	D (Sarfatti)
3	Wed 21/06/2023	14.45 - 16.15	D (Sarfatti)
5	Mon 26/06/2023	14.45 - 16.15	D (Sarfatti)
6	Wed 28/06/2023	14.45 - 16.15	D (Sarfatti)
7	Mon 03/07/2023	14.45 - 16.15	D (Sarfatti)
8	Wed 05/07/2023	14.45 - 16.15	D (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*

Lecture	Topics
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*
6	**Web scraping - 2** - Selenium - Scrapy - Using APIs (*hints*) *Exercises*
7	**Text clustering** - Clustering versus Classification - Topic detection - Mapping/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 (April 2023) it supports Python 3.10. It is available for Windows, Mac, and Linux.

You can download it here:

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

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. Registrations cannot be carried out once this number has been reached or after closing of the registration period.
Please remember that you may unsubscribe from ITEC courses only before the registration deadline.