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 for 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 purpose
- 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.

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

Distance learning. Lessons will take place **exclusively** in **synchronous remote mode**.

Calendar

Lecture	Date	Time	Room
1	Mon 03/06/2024	16.30 - 18.00	Virtual room
2	Wed 05/06/2024	16.30 - 18.00	Virtual room
4	Wed 12/06/2024	16.30 - 18.00	Virtual room
3	Mon 17/06/2024	16.30 - 18.00	Virtual room
5	Mon 24/06/2024	16.30 - 18.00	Virtual room
6	Wed 26/06/2024	16.30 - 18.00	Virtual room
7	Mon 01/07/2024	16.30 - 18.00	Virtual room
8	Wed 03/07/2024	16.30 - 18.00	Virtual room

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*
6	**Web scraping - 2** - Selenium - Scrapy - Using APIs (*hints only*) *Exercises*

Lecture	Topics
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 (April 2024) it supports Python 3.11. 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. Registrations cannot be carried out once this number has been reached or after closing of the registration period.