Text analysis with Python

Lecturer: Maria Chiara Debernardi

Course 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 process and analysis of natural language (NLP) is becoming 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 carrying out textual analyzes using Python as programming language and its appropriate libraries of Text Mining available for free.

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

- Understand the text analysis steps
- Create a simple text analysis pipeline using Python
- Extract textual data from web

Audience

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

- those who want to approach the world of automated text analysis
- students who want to acquire basic web scraping knowledge to develop by their own
- 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 the ITEC's courses Python start / Programming with Python (or having equivalent knowledge and skills).
Prior knowledge of Pandas and, at least, some basic Statistics are welcome.

Duration

12 hours

Teaching mode

The course will be held in distance learning mode. It will be possible to follow the live streaming (Live Session) of each lesson accessing through Blackboard to the corresponding virtual room.

Calendar

Lecture	Date	Time
1	Tue 08/06/2021	18.40 – 20.10
2	Thu 10/06/2021	18.40 – 20.10
4	Tue 15/06/2021	18.40 – 20.10
3	Thu 17/06/2021	18.40 – 20.10
5	Tue 22/06/2021	18.40 – 20.10
6	Thu 24/06/2021	18.40 – 20.10

Syllabus of the course

Lecture	Topics
1	**Building a common ground** - Preliminaries - Introduction to Jupyter Notebook - Brief recap of Python basics - Pandas: the essentials - Why NLP in today's world: its applications *Exercises*
2	**Text mining, Text analytics and NLP** - Tokenization: sentences and words - Stop words - Lexicon normalization: Stemming versus Lemmatization - POS tagging - N-grams *Exercises*

Lecture	Topics
3	**Text classification and clustering** - Bag of words - TF-IDF - Word embedding - Classification versus Clustering - Mapping/visualization *Exercises*
4	**Sentiment analysis** - Issues about sentiment detection - Lexicon-based methods - Rule-based analysis methods - Machine Learning based approach *Exercises*
5	**Web scraping** - How to do it - BeautifulSoup - Selenium *Exercises*
6	**Final Exercise**

Software used

Jupyter Notebook inside Anaconda
Anaconda Individual Edition is free version for solo practitioners, students and researchers, and currently supports Python version 3.8. It is available for Windows, Linux and OS X, for 32 bit or 64 bit systems, and can be downloaded from here: https://www.anaconda.com/products/individual

Suggested bibliography

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

Available seats

This activity is limited to **60** participants. Registrations cannot be carried out once this number has been reached or after closing of the registration period.