Phd seminar: “Machine Learning Techniques for Text”

In the context of *Doctoral Programme in Applied Linguistics: Managing Languages, Arguments and Narratives in the Datafied Society*

ECTS: 1

Program:

- 23. November, 2023, 09:00-17:15 (inclusive of breaks and hands-on sessions)
- 24. November, 2023, 09:00-17:15 (inclusive of breaks and hands-on sessions)

- Place: USI, Lugano
  Nov 23 - Campus Est Sector A. room A2.09 P2 (Via la Santa 1, 6962 Lugano)
  Nov 24 - Main Building, room 355
- Lecturer: Nikos Tsourakis (research associate at the faculty of Translation and Interpreting at the University of Geneva)

Topic, focus:
This introductory to intermediate-level seminar focuses on natural language processing and machine learning techniques for text using Python. Although some basic programming knowledge is desirable, the seminar will start with a Python crash course to onboard all participants. The exercises will be in the form of Jupyter notebooks to avoid the burden of installing and configuring the tools. Moreover, the focus will not solely be on the programming tasks; participants will benefit from the presentations of the methods incorporated in the exercises. Intuitive explanations and real-world analogies will be provided when possible. In this respect, the content of the different exercises will consist of a theoretical and a practical part.

Main value added from a theoretical and methodological perspective:
In recent years we are experiencing an astonishing paradigm shift in software programming for data processing. Traditionally, computers are given data and “recipes” on how to solve specific problems. Currently, however, the focus moves from the manual implementation of the recipes to some sort of automatic extraction of the instructions. This seminar builds on this new paradigm and focuses on human written language. Participants will be exposed to the basic theory behind each incorporated method or technique and how they can be combined in a pipeline. The exercises will follow the same pattern that consists of three main steps: (1) acquire some intuition on the data (exploratory data analysis), (2) incorporate a machine learning algorithm (for training and inference), and (3) evaluate its performance on the problem under study (metrics).

Main value added from a practical perspective:
The seminar will be solution-oriented; each exercise presents a specific machine-learning problem and the steps taken to solve it. This case study approach allows participants to engage more easily with the content and not just passively absorb information. Putting each method into action is an excellent way to identify at least one practical usage of a method, even though it can be used in many other contexts. The problem statement will be set from the beginning so that everybody will be aware of the challenge. Even if the discussion temporarily diverts from the principal aim (e.g., presenting some fundamental concept), participants will easily be reoriented on the problem under study. Finally, real-world datasets will be used in the exercises.

Learning objective: knowledge. Participants will learn 1) the fundamental machine learning concepts for text data, 2) the processing pipeline for solving pertinent machine learning problems, and 3) the evaluation metrics for accessing the performance of the implemented systems.

Learning objective: skills, practices. Participants will 1) familiarize themselves with the related Python libraries, 2) use Python to perform tasks like text processing, visualization, classification, clustering, and so forth, and 3) use Python to evaluate performance.

Learning objective: researcher attitudes. Participants will identify patterns and ways to transfer the knowledge from the presented case studies to their own work.

Evaluation: Assignment which will be assessed with a pass/ fail evaluation.

Bio notes/ Profiles

Nikos Tsourakis is a research associate at the faculty of Translation and Interpreting at the University of Geneva. He also teaches Text Mining and Python courses in the Business Analytics program of the International Institute in Geneva. He has over 20 years of experience designing, building, and evaluating intelligent systems using speech and language technologies. He has also co-authored over 50 research publications in the area and wrote the book “Machine Learning Techniques for Text.” In the past, he worked as a software engineer, developing products for major telecommunication vendors. He also served as an expert for the European Commission and is currently a certified educator at the Amazon Web Services Academy. He holds a degree in electronic and computer engineering, a master’s in management, and a PhD in multilingual information processing.