DEPARTMENT OF POLITICAL AND SOCIAL SCIENCES

1st Term Workshop, Academic Year 2022-2023

Natural Language Processing and Machine Learning Methods for the Social Sciences

Organiser: Arnout van de Rijt (arnout.vanderijt@eui.eu)

Instructor: Steven Skiena (steven.skiena@eui.eu)

Contact: Charlotte Bufano (charlotte.bufano@eui.eu)

Syllabus

Workshop Description

Data Science is a rapidly emerging discipline at the intersection of statistics, machine learning, data visualization, and mathematical modeling. This course will provide an introduction to data science methods for social scientists, focusing on methods for networks and text-oriented data associated with social media. I will present modern approaches to natural language processing (NLP) based on machine learning, and vector representations (embeddings) for effectively capturing network data in predictive models. The final choice of topics will depend upon student interests.

This minicourse will be based on the instructor's book, "The Data Science Design Manual", Springer-Verlag, 2017.

Schedule

- Wednesday, 9 November: h. 9.00 – 13.00, *Theatre*
- Friday, 11 November: h. 13.00 – 16.00, *Seminar Room 2*
- Wednesday, 16 November: h. 9.00 – 12.00, *Sala del Capitolo*

Course Requirements

- Attending all three workshop sessions, and participating in discussions.
- Submitting a one-to-two page description of a data set/substantive question you are interested in, with a sketch of an analysis plan using analysis techniques discussed in the workshop. To be handed in by November 23

Course Objectives

- General introduction to data science and applied machine learning for social scientists.
- General understanding of the basics of natural language processing
- Discuss representations of network data for machine learning, i.e. graph embeddings.

Learning Outcomes

- Become more proficient at dealing with large datasets, particularly of text and network data.
- Understand the strengths and limitations of modern natural language (NLP) models.
- Appreciate the power of vector representations (word and graph embeddings) in diverse data analysis applications.

Recommended Readings

- S. Skiena, "*The Data Science Design Manual*" Springer-Nature, 2017. An e-link to the book will be made available by the EUI Library.
- Lecture notes from CSE 519 at Stony Brook University: https://www.cs.stonybrook.edu/~skiena/519

Course Outline (Tentative – subject to change based on student interests)

- Day 1: Data Science and Machine Learning
  - Class discussion of student datasets projects, and interests.
  - Overview of basic machine learning techniques, including
    - Regression by gradient descent search
    - Decision trees
    - Nearest neighbor methods
    - Neural networks / deep learning

- Day 2: Natural Language Processing
  - Basic approaches to working with social media and text documents
  - Classical NLP: POS tagging, entity recognition and sentiment analysis
  - Word embeddings and applications
  - Modern language models

- Day 3: Topics in Data Science (depending upon student interests)
  - Graph embeddings and network data
  - Scores and ranking methods
  - Geometric datasets
  - Visualization techniques