Laboratory Proposal

"Data Science for Economics: Job Postings Analysis and Measuring AI Usage"
Duration: **20 hours, 5 sessions of 4 hours each**

Objectives of the Laboratory

The laboratory aims to introduce students to advanced Data Science methods applied to labor economics, with a specific focus on measuring the **use of Artificial Intelligence in the writing of job postings**.

Students will learn to:

1. Retrieve real-world data from web platforms using both traditional scraping techniques (Selenium) and AI-based scraping (ScrapeGraphAI).
2. Clean, transform, and organize complex datasets in a cloud environment (AWS Academy).
3. Build a classification model to identify whether a job posting has been written (or assisted) by an AI system.
4. Apply topic modelling techniques to analyze content, linguistic patterns, and pre/post-AI changes.
5. Present results through interactive dashboards (Tableau).

Students will work both with a **dataset provided by Lightcast** (job postings pre-2020 and post-2020) and with data collected live during the sessions via scraping.

Laboratory Content

Session 1: Introduction and Web Scraping (4 hours)

Objective: acquire operational skills in retrieving data from websites.

Topics covered:

- What job postings are (advantages, limitations, etc.)
- What can be legally scraped and under which constraints
- **Traditional scraping** with *Selenium*:
    - dynamic navigation
    - HTML parsing
    - handling infinite scroll pages
- AI-based scraping using ScrapeGraphAI:
    - defining objectives

    - automatic script generation
    - advanced parsing of textual content
- Collection of a small live dataset during the session

Tools: Google Colab, Selenium, ScrapeGraphAI

Session 2: Data Cleaning, Organization, and Cloud (4 hours)

Objective: prepare data for advanced analysis.

Topics covered:

- Removing duplicates, normalization, handling special characters
- Language analysis of job postings: tokenization, stopwords, lemmatization
- Structuring the dataset for longitudinal pre/post-2020 analysis
- Introduction to AWS services (via **AWS Academy**):
    - S3 for storage
    - AWS Glue for cleaning, ETL, and data preparation
    - the concept of a "data lake" for research projects
    - AWS Athena
- Saving datasets in the cloud and managing workflows

Logistical note: student email addresses will be required 5 days in advance to activate AWS Academy accounts.

Session 3: Classification of AI Usage in Job Postings (4 hours)

Objective: develop a binary model to predict whether a job posting is written (or assisted) by AI.

Topics covered:

- How AI-generated language differs in job postings (linguistic patterns, style, repetition, structure)
- Creation of a training dataset using pre-2020 and post-2020 job postings
- Development of an **NLP classification model**:
    - baseline Bag-of-Words / TF-IDF
    - transformer-based models (BERT or similar)
- Evaluation metrics: accuracy, precision, recall, ROC curve
- Discussion of results: how identifiable are AI-written job postings?

Tools: Scikit-learn, HuggingFace Transformers, Google Colab

Session 4: Topic Modelling and Advanced Analysis (4 hours)

Objective: extract recurring themes and compare pre/post AI.

Proposed activities:

- Introduction to topic modelling:
  - LDA (Latent Dirichlet Allocation) for interpretability
  - BERTopic for more semantically coherent results
- Comparison of topics in job postings pre-2020 and post-2020:
  - variation in linguistic complexity
  - emphasis on soft skills vs. hard skills
  - possible influence of AI on content structure

Session 5: Dashboarding and Presentation of Results (4 hours)

Objective: operationalize results and build an analytical narrative.

Topics covered:

- Introduction to Tableau:
  - connecting to CSV files
  - creating interactive visualizations
  - temporal and sector/language filters
- Preparation of a short final presentation of results by student groups

Required Tools

- Google Colab
- AWS Academy
- Tableau Public or university license

Tentative calendar (18h):

April 16th, 12:30–16:30

April 23rd, 12:30–15:30

April 30th, 12:30–15:30

May 7th, 12:30–16:30

May 21st, 12:30–16:30