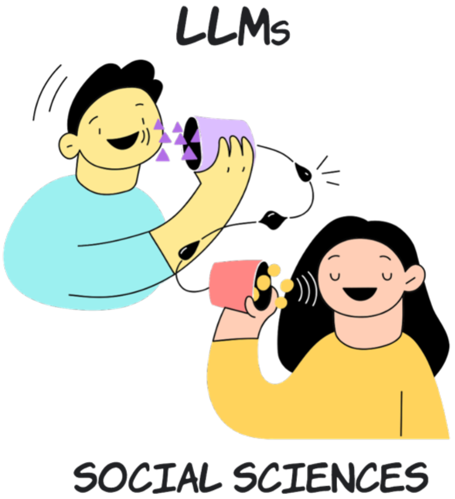
Introduction to Large Language Models (LLMs) in Social Sciences: Online (3 days)



Working with language and text can be challenging, but now we have a new tool, Large Language Models (LLMs), that offers a new way to not only analyse but also interact with text at an unprecedented scale. This masterclass is an introduction to LLMs in social sciences. You will learn about the basics of Natural Language Processing (NLP) and their applications, including text preprocessing, sentiment analysis, topic modeling, and text generation.

The course uses Python and Google Colab but does not require prior coding experience. Our focus is on practical hands-on experience that you can use to reach your research goals.

This masterclass is part of the ACSPRI suite of courses in social data science.

This course will be run over 3 days using the following timetable:

Day 1

- 9.30 am - 10.00 am – Introductions and setup check
- 10.00 am - 11.30 am - Session 1
- 12.30 pm - 2.00 pm - Session 2
- 3.00 pm - 5.00pm - Session 3 + exercises

Days 2 and 3

- 9.00 am - 10.30 am - Session 1
- 11.30 pm - 1.00 pm - Session 2
- 2.00 pm - 4.00pm - Session 3 exercises and consultation

Master Class - runs over 3 days

Instructor:

Dr. Maria Prokofieva is a Lead data scientist at the Mitchell Institute, Vic, where her expertise in cyberpsychology and business analytics informs policy development. As a machine learning engineer with a deep passion for the responsible application of AI, Maria's work deciphers complex online behaviours to inform consumer and business strategies. She also chairs the CPA Australia Business Analytics Group and spearhead R Business Software Development Group, driving innovation in data analysis tools. Maria's contributions to both research and practical applications are shaping the integration of AI in business and policy on a global scale.

Course next offered: Introduction to Large Language Models (LLMs) in Social Sciences: Online (3 days) - Master-class August 2024: Introduction to Large Language Models: Online (3 days)

About this course:

Have you heard about ChatGPT and probably used it yourself?

But do you know that the technology behind it can be used for many more applications?

This course will look into this and will equip you with some basic understanding (and appreciation) of Large Language Models (LLMs) which are AI models designed to understand, interpret, and generate human text. While the OpenAI's GPT model has gained significant attention already, there are other notable models available for free. This course will walk you through diverse options available for researchers looking to explore natural language processing (NLP) tasks, such as text preprocessing, sentiment analysis, classification, topic modeling, and many more.

This course is based on Python and uses TensorFlow libraries in Google Colab. The course does not assume prior coding experience or knowledge of Python, and one of the sessions will be dedicated to the basics of working with data in Python, including using the NumPy library for numerical operations, and Pandas for data manipulation.

This course is tailored for social scientists, PhD students, and researchers who aim to use NLP techniques in their work. Additionally, this course offers a great opportunity for marketing and media professionals and public policymakers to explore how LLMs can enhance language-related tasks, such as text generation, and analysis of complex datasets, including political speeches and media. The course does not expect prior programming experience and is for a wide audience keen on exploring recent advances in NLPs for decision-making.

Course syllabus:

Day 1: Foundations of Python and Introduction to LLMs

- Morning Session: Introduction to Python for Social Sciences
  - Overview of Python as a programming language
  - Introduction to Google Colab and basic Python syntax and concepts
  - Introduction to data preprocessing with Numpy and Pandas
- Case Demonstration:
  - Analysing a simple dataset (e.g., a CSV file containg survey responses) using pandas and drawing basic inferences
  - Key takeaway: Understanding how Python can be used to manipulate and analyse social science data
- Afternoon Session: Understanding Large Language Models (LLMs)

  - What are LLMs and how do they work?
  - Overview of the capabilities of LLMs
  - Ethics and considerations in using LLMs in Social Science Research
- Case Demonstration:
  - Using a pre-trained LLM to analyse text data (e.g., political speeches or social media posts) to extract themes and sentiments.
  - Key takeaway: intro to LLMs in qualitative data analysis.

Day 2: Hands-On with LLMs in Social Science Research

- Morning Session: Python Libraries for Working with LLMs
  - Introduction to Python libraries for LLMs (eg., transformers, OpenAI's GPT)
  - Simple text generation and text completion tasks using LLMs
- Case Demonstration:
  - Data augmentation in social sciences research: e.g. generating synthetic interview reponses based on a provided dataset.
  - Key takeaway: How LLMs can be used for data augmentation in social science research.
- Afternoon Session: Data Collection and Preprocessing for LLMs
  - Methods for collecting text data relevant to social science research.
  - Preprocessing text data for LLMs: tokenization, handling missing data, and batch processing.
- Case Demonstration:
  - Collection and preprocessing news articles for sentiment analysis using a LLM.
  - Key takeaway: Preparing real-world data for analysis wit LLMs.

Day 3: Advanced Aplications of LLMs in Social Sciences

- Morning Session: Fine-Tuning LLMs for Custom Use-Cases
  - The concept of model fine-tuning and transfer learning.
  - Preparing a dataset for fine-tuning an LLM on a social science-specific task.
  - Initiating a fine-tuning process on a subset of data.
- Case Demonstration:
  - Fine-tuning an LLM to recognize and classify academic articles into social science subfields.
  - Key takeaway: Tailoring LLMs to understand and categorize domain-specific content.
- Afternoon Session: Project Development and Ethical Implications
  - How to design a social science research project using LLMs.
  - Discussion on the ethical implications and potential biases in LLM use.
  - Sharing results responsibly and transparently.
- Case Demonstration:
  - Developing a project outline that uses an LLM to study social narratives in historical newspaper archives.
  - Key takeaway: Constructing a responsible and informative social science research project using LLMs.
- Final Activity: Workshop Wrap-up and Next Steps
  - Participants share their project ideas and receive feedback

  - Resources for further learning and exploration in Python, LLMs and social science research
  - Discuss potential collaborations and future research projects.

Course format:

This workshop will take place online.

BYO Laptop + Zoom. Both PC and MAC are great

The course uses **Google Colab** and requires a **Google account** (please make sure you have one or please register one before the session)

All course materials will be provided

Recommended Background:

The course requires understanding of a basic of statistical concepts and text analysis tasks, exposure to machine learning foundations is beneficial as well, such as **Machine Learning for Data Science: Supervised Learning Techniques**

The course assumes no prior knowledge of Python, though some programming experience (e.g. using R) is beneficial.

Recommended Texts:

HuggingFace official Getting Started Guide

https://huggingface.co/learn/

Tunstall, L., Von Werra, L., & Wolf, T. (2022). Natural language processing with transformers. " O'Reilly Media, Inc.".

https://learning.oreilly.com/library/view/natural-language-processing/9781098136789/