

Natural Language Processing for Economists

Course Handbook

Instructor: Dr. Sowmya Vajjala

- National Research Council, Canada
- *Email:* sowmya.vajjala@nrc-cnrc.gc.ca

Course Overview: Natural Language Processing (NLP) is an area of research focused on analyzing human languages computationally and making computers understand and interact with humans, in their language(s). NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email provider. Considering that many areas of study also rely on a lot of textual data for knowledge dissemination and communication, NLP is now being used as a method to explore discipline specific research questions in many areas. Economics too is one of them. From measuring news sentiment and connecting it to the nation's economy to extracting key information from policy documents, there are many uses of NLP in economics research. In this course, we will introduce the methods of working on Natural Language Processing problems, and explore how it can be used to address some research questions in economics that require analysis of textual data.

Learning Outcomes: Upon successful completion of this course, students are expected to:

- be familiar with some foundational concepts in modern day NLP
- understand where NLP is useful for economists, and what are some of the common methods used, and
- given a problem description in economics research involving textual data, know what methods from NLP can be used to solve the problem, and design a step by step process to solve it.

Some concrete expectations are that the students will be able to:

- write small programs that involve reading, processing and writing textual content into files on the computer
- use existing NLP solutions and evaluate their applicability to economics data
- articulate how NLP is useful for their own research

Pre-requisites: Some familiarity with python programming, and being comfortable with installing software, python libraries etc on their won.

Timings and expectations: This course is graduate level guest course introducing NLP methods to economics students, and will happen before your semester begins in November. Primary mode of instruction are live video lectures and discussions, which will be recorded and stored for the duration of the course.

The course is divided into 6 sessions of 3 hours each. We will meet via zoom and lectures can be recorded for future reference, if needed. You are expected to read the recommended readings before attending the lecture.

Here is the schedule:

1. Dates: 5th, 6th, 7th, 11th, 12th, 13th October
2. Time: 3-6 pm CEST (9am-12 noon in EST, where I am based)
3. In the week of 17th October (on 17th or 18th), we can optionally have additional informal sessions with students who want to meet, if needed.

Course Registration: Through university website.

Credits: Pass or Fail (Pass is completing assigned activities - group presentation, individual term paper)

Resources/Reading Materials Books: There is no single textbook. We will try to rely on publicly accessible resources for as much as possible.

1. "Speech and Language Processing" by Jurafsky and Martin (2nd Edition: <https://github.com/rain1024/slp2-pdf>. 3rd Edition: <https://web.stanford.edu/~jurafsky/slp3/>)
2. "Python for Everybody" Charles Severance <https://www.py4e.com/html3/> (For Python)
3. "Practical Natural Language Processing" by Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta and Harshit Surana. <https://www.amazon.de/Practical-Natural-Language-Processing-Pragmatic/dp/1492054054/>. The book is also available on O'Reilly's online learning platform, where you can access it for free for a few days.
4. NLTK book -<https://nltk.org/book>
5. <https://transformersbook.com/>

Course Website: <https://econnlpcourse.github.io/>

List of Topics (one topic per session) Format will be a combination of lectures and hands-on exercises.

1. Introduction
 - Course overview
 - Introduction to NLP
 - NLP, Machine Learning, and Economics: an overview
 - Python: some basics.

Readings: (Note that you are not obligated to read everything thoroughly).

- Chapter 1 from "Speech and Language Processing" by Jurafsky and Martin (available online)
 - Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535-74.
 - <https://www.py4e.com/>
2. Python Overview (overview of python's syntax and features, some basic text analysis with Python and NLTK library)
 3. NLP and Machine Learning methods: an overview. Hands on activity, working with existing NLP tools.
 - Corpus collection (e.g., social media text, ethical issues etc)
 - Corpus analysis (basic analysis - e.g., frequent words/phrases etc)
 - Text classification
 - Information extraction (regular expressions, key phrase extraction, named entity recognition/linking etc)
 - Topic modeling
 - Text summarization
 4. Diving deeper: Text classification Ref: Chapters 4 and 5 in Jurafsky and Martin, more will be added later.
 5. Working with small (or no datasets) - different approaches such as weak-supervision, transfer learning etc.
 6. NLP and Economics: selected readings + Group discussion (Details on the website)

Recap (Optional sessions lasting 3-4 hours in total)

- Discussion on topics covered
- Review of exercises
- Resources for the future
- Any other one-one support

Final exercise: Student term papers

- Briefly summarize what you learnt about the intersection of NLP and Economics by taking this course, and note down some thoughts on how it is useful for your own research topics.

graded Assignments A group presentation, and a short write up submission.

Accommodation Requests If you need any accommodation (e.g., extending deadline due to serious circumstances etc.) contact me as soon as possible.