MGSE Guest Course:

Natural Language Processing for Economists

Course Handbook

Instructor: 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 Natural Language Processing (and Machine Learning) and explore it can be used to address some research questions in economics that require analysis of textual data. Before starting with this, we will do a Python crash course in the first few sessions.

Learning Outcomes: Upon successful completion of this course, students are expected to be able to do the following:

- Understand how text is represented on computers, and how to work with it by writing Python programs.
- Understand where NLP is useful for economists, and what are some of the common methods used
- Given a problem description in economics research involving textual data, understand 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, and
- articulate how NLP is useful for their own research

Pre-requisites: Familiarity with using computers and an interest in learning how they work and how to make them do what you want. Knowledge of some programming language is useful but not mandatory.

Timing of the course 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 pre-recorded video lectures, followed by live discussions. The course is divided into four broad topics (See List of Topics heading below). Each topic will have 3-4 lecture videos, followed by one (or two) live discussions (via zoom, 1-1.5 hours long). This is followed by a couple of live meetings for group discussions, and perhaps student term paper presentations. Lecture and exercises for each session will be uploaded in advance and you are expected to read the recommended readings before listening to the lecture. You are expected to have tried the hands on programming exercises before attending the live sessions. I will try to upload the lecture videos for all the four main topics before October 15th as much as possible, and we can schedule the zoom sessions (perhaps 7-8 of them in total) between 15th to 31st October, preferably around 3pm in your time during the day. Exact schedule will be mailed later.

Course Registration: Through MGSE internal process

Credits: Pass or Fail (Pass is completing all assigned exercises as much as possible)

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-P dp/1492054054/. The book is also available for free on O'Reilly's online learning platform (if your university is subscribed). A 30 day trial code to read the book online is here: https://learning.oreilly.com/get-learning/?code=PNLP20.
4. NLTK book -https://nltk.org/book

Course Website: https://econnlpcourse.github.io/ Link to lecture videos (uploaded to youtube or some other password protected website) will be posted here and lecture slides will also be uploaded. While we don't have a setup for a discussion forum yet, we can figure out some way to interact, such as a google group, if needed.

List of Topics (tentative)

Note: The following syllabus is tentative assuming only basic previous knowledge about python programming.

1. Introduction

- Course overview
- Introduction to NLP
- NLP, Machine Learning, and Economics: an overview

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.

2. Python fundamentals crash course

   - Installing python on your personal machines/lab machines. Writing a hello world program
     See this link for instructions: https://www.py4e.com/install
   - Installing Jupyter notebook
   - Writing basic variable declarations, performing arithmetic operations
   - Basic data structures (strings, lists, dictionaries)
   - Basic programming: conditional statements, loops, functions, error handling
   - Reading and writing files
     Readings: "Python for Everybody" by Charles Severence. https://www.py4e.com/html3/. The content covered in this Chapter is taken from the first 10 chapters in the book.
     One live discussion session.

3. Python & textual data

   - How to install various libraries
   - Reading and writing files in different formats (e.g., pdf, html, text, doc etc)
   - Pre-processing text (e.g., sentence splitting, removing punctuation/digits etc if needed)
   - Representing text as a numeric vector (e.g., bag of words, TF-IDF, embeddings)

   Readings: Chapters 2 and 3 in "Practical Natural Language Processing"
   One live discussion session

4. NLP and Machine Learning methods
   (with econ specific datasets where possible)

   - 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

Readings: Chapter 1-2 from "NLTK book" (https://nltk.org/book) and Chapters 4-7 in "Practical Natural Language Processing"
1-2 discussion sessions.

5. NLP and Economics: selected readings + Group discussion - perhaps working in groups of 2-3 people? (3-4 sessions) You can choose from some of these papers (list will be updated by the time course starts - I am trying to collect a list of papers from NLP/AI conferences as well as Economics Journals. Some of these papers may go away and news ones be added.

   1.1 Shapiro, A., Sudhof, M., & Wilson, D. J. (2017). Measuring News Sentiment, Federal Reserve Bank of San Francisco Working Paper 2017-01. Accessed, 17, 51.

   1.2 Bholat, David, Stephen Hansen, Pedro Santos, and Cheryl Schonhardt-Bailey. "Text mining for central banks: handbook." Centre for Central Banking Studies 33 (2015): 1-19.

   1.3 Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications, 42(3), 1314-1324.

   1.4 Lawrence, A. (2013). Individual investors and financial disclosure. Journal of Accounting and Economics, 56(1), 130-147.

   1.5 Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics, 45(2-3), 221-247.

   1.6 Li, F. (2010). Textual Analysis of Corporate Disclosures: A Survey of the Literature. Journal of accounting literature, 29, 143-165.

   1.7 Iaria, A., Schwarz, C., & Waldinger, F. (2018). Frontier knowledge and scientific production: Evidence from the collapse of international science. The Quarterly Journal of Economics, 133(2), 927-991.

   1.8 Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: a computational linguistics approach. The Quarterly Journal of Economics, 133(2), 801-870.

   1.9 Händschke, S. G., Buechel, S., Goldenstein, J., Poschmann, P., Duan, T., Walgenbach, P., & Hahn, U. (2018, July). A corpus of corporate annual and social responsibility reports: 280 million tokens of balanced organizational writing. In Proceedings of the First Workshop on Economics and Natural Language Processing (pp. 20-31).

   1.10 Jelveh, Z., Kogut, B., & Naidu, S. (2014, October). Detecting latent ideology in expert text: Evidence from academic papers in economics. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1804-1809).

6. 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.

- Depending on the time and interest, we can decide whether we want to have a presentation session or just writeup submissions.

7. Recap (1 live session)
    - Discussion on topics covered
    - Review of exercises
    - Resources for the future

Assignments/Exercises Each topic has an associated exercise, as shown below:

1. Topic 1 - Introduction
2. Topic 2 - Python Fundamentals
3. Topic 3 - Python and Text
4. Topic 4 - NLP and Machine Learning Methods
5. NLP and Economics Group Discussion (choose one paper from the list and form into teams of 2-3 people)
6. Individual Reports live presentation or written submission (depending on number of students)

Academic Conduct: Generally, you are encouraged to work in (socially distanced) groups, discuss, and exchange ideas. At the same time, you are expected to do your assignments by yourself and with honesty. For the text you write, you always have to provide explicit references for any ideas or passages you reuse from somewhere else. Note that this includes text taken from the web. You should cite the url of the web site in case no more official publication is available.

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