Text Data in Economics

The goal of this course is to equip you with modern text data methods and integrate these techniques into your research. You should know what type of text methods are used in economic research and have some hands-on experience with text data methods. By the end of the course, you should have an actionable research plan.

The course begins with introductory lectures on text data methods and showcases examples of how these methods have been used in economics research. The topics covered include tokenisation, distance in text, vectorisation, and the use of large language models.

Following the introductory lectures, we start working on our own text data project and developing our own research ideas using text data. We work individually or in groups depending on the number of students. First, we develop research questions and consider potential data sources that would allow us to answer the question. We explore available data sources and design the analysis we would run on our data. We present the ideas to the group, write a research proposal, and give feedback to our peers. In the end, we write a term paper detailing the research question, contributions to the literature, data sources, analysis, and an empirical part using text data. If data is too difficult to acquire during the course, students can do a separate text data exercise. Ideally, the plan leads to a proper research project.

The course is for economics students in all fields interested in text methods and you will be encouraged to work on a project in your field of interest. However, most examples in the lectures will focus on my fields of expertise like political economy and economic development.

Prerequisites:

- Some Python programming experience is required (or excellent skills in another programming language and a willingness to acquire the necessary skills very fast).
- There are also prerequisite courses set by the administration: **Mathematics for Economists and Basic module Econometrics. I cannot give you credit if you have not done the prerequisite courses!**

Requirements:

- You develop and present a research idea to the group.
- You read your peers' proposals and provide short, written feedback with suggestions on how to improve them.
- You engage in the group discussion after each presentation.
- At the end of the course, you hand in a term paper on the final version of your research proposal. **(Deadline 28.2.)**

Grading: The final grade is a weighted average of your presentation and participation in the group discussions (40%) and your term paper (60%).

Syllabus

Lecture 1: Introduction to Text Data, Scraping and Tokenization

Reading:

Ash and Hansen, “Text Algorithms in Economics”

Gentzkow, Kelly, and Taddy, “Text as Data”

Lecture 2: Tokenization and Dictionaries

Reading:

Lecture 3: Vectorization and Document Distance

Reading:

Autor et al. “New Frontiers: The Origins and Content of New Work, 1940–2018”

Lecture 4: Large Language Models

Reading:

Schedule

Presentations: Text Data: Presentations

Date	**Content**
08.10.	Lecture 1: Overview
15.10.	Lecture 2: Tokenization Dictionaries
22.10.	Lecture 3: Vectorization and Document Distance
29.10.	Lecture 4: Language Models
05.11.	Paper Presentations
12.11.	Paper Presentations
19.11.	Paper Presentations
26.11.	No Lecture (option for Q&A)
03.12.	Paper Presentations
10.12.	Paper Presentations
17.12.	Q&A for Research Proposals (On Zoom): Message me if you want to talk your research ideas
07.01.	Research Proposal Presentations
14.01.	Research Proposal Presentations
21.01.	Research Proposal Presentations
28.02.	**Term Paper Deadline**

Notebooks

Data for the Notebooks: Link (Dropbox)
Notebook1: Corpora Matching Link (Dropbox)
Notebook2: Tokenization Link (Dropbox
Notebook3: Word Embeddings and Document Distance Link (Dropbox)
    Data for Notebook3: Link (Dropbox)
Notebook4: NNs and LLMs Link (Dropbox)

Paper Presentation Slides: Link (Dropbox)

Instructions for Paper Presentations:

Presentations should be **30** minutes with a focus on the text methods in the paper. The presentation should cover the following content:
- What is the authors' motivation and research question? Which gap in the literature do they address (short)?
- Which data do they use, how do they preprocess data? How did the research access data, Is it available for anyone?
- What text methods did they use?
- Are these methods complemented with other empirical methods or a research design to allow studying causal questions?
- What are the most important results?
- What are the limits of the study? Where do you see potential for future research?
- 2-3 questions or thoughts for discussion

Instructions for Research Proposal Presentations:

Presentations should be **20** minutes with a focus on the text methods in the paper. The presentation should cover the following content:

- Research Question
- Existing Literature
- What data is/will be used
- What text methods are being used?
- Is there a theoretical framework, other empirical methods, research design?
- Some preliminary results?

Instructions for the Term Paper:

The proposal has to include empirical research using text data.

The following content should be included:
- Introduction, including the motivation and research question
,p
- Suitable data source, how is it accessed?
- Description of text data method that will be used
- Some analysis on the text*

- Optional: discussion of other methods used in the paper
Length: maximum 10 pages. Also code and results should be submitted.

*I want every student to get some hands-on experience on analyzing text. However, if the data is not accessible fast enough for this course, you should not let that restrict you writing the Term

Paper on that idea. So, in this case you can do some analysis on different text data, write a short report and submit that analysis as part of your Term Paper.

Papers

Angelico, Cristina, Juri Marcucci, Marcello Miccoli, and Filippo Quarta. “Can We Measure Inflation Expectations Using Twitter?” *SSRN Electronic Journal*, 2021. https://doi.org/10.2139/ssrn.3827489.

Ash, Elliott, and Stephen Hansen. “Text Algorithms in Economics.” *Annual Review of Economics* 15, no. 1 (September 13, 2023): 659–88. https://doi.org/10.1146/annurev-economics-082222-074352.

Autor, David, Caroline Chin, Anna Salomons, and Bryan Seegmiller. “New Frontiers: The Origins and Content of New Work, 1940–2018.” *The Quarterly Journal of Economics*, March 15, 2024. https://doi.org/10.1093/qje/qjae008.

Ayyar, Sreevidya, Uta Bolt, Eric French, and Cormac O’Dea. “Imagine your life at 25: gender conformity and later-life outcomes ” *Working Paper*, 2024.

Bertrand, Marianne, Matilde Bombardini, Raymond Fisman, Brad Hackinen, and Francesco Trebbi. “Hall of Mirrors: Corporate Philanthropy and Strategic Advocacy.” *The Quarterly Journal of Economics* 136, no. 4 (October 11, 2021): 2413–65. https://doi.org/10.1093/qje/qjab023.

Cagé, Julia, Nicolas Hervé, and Marie-Luce Viaud. “The Production of Information in an Online World.” *The Review of Economic Studies* 87, no. 5 (October 1, 2020): 2126–64. https://doi.org/10.1093/restud/rdz061.

Enke, Benjamin. “Moral Values and Voting.” *Journal of Political Economy* 128, no. 10 (October 1, 2020): 3679–3729. https://doi.org/10.1086/708857.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. “Text as Data.” *Journal of Economic Literature* 57, no. 3 (September 1, 2019): 535–74. https://doi.org/10.1257/jel.20181020.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. “Measuring Group Differences in High‐Dimensional Choices: Method and Application to Congressional Speech.” *Econometrica* 87, no. 4 (2019): 1307–40. https://doi.org/10.3982/ECTA16566.

Hansen, Stephen, Peter John Lambert, Nicholas Bloom, Steven J. Davis, Raffaella Sadun, and Bledi Taska. “Remote Work Across Jobs, Companies, and Space.” *NBER Working Paper*, 2023. https://www.nber.org/system/files/working_papers/w31007/w31007.pdf.

Hansen, Stephen, Michael McMahon, and Andrea Prat. “Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*.” *The Quarterly Journal of Economics* 133, no. 2 (May 1, 2018): 801–70. https://doi.org/10.1093/qje/qjx045.

Hassan, Tarek A, Stephan Hollander, Laurence Van Lent, and Ahmed Tahoun. “Firm-Level Political Risk: Measurement and Effects*.” *The Quarterly Journal of Economics* 134, no. 4 (November 1, 2019): 2135–2202. https://doi.org/10.1093/qje/qjz021.

Hoberg, Gerard, and Gordon Phillips. “Text-Based Network Industries and Endogenous Product Differentiation.” *Journal of Political Economy* 124, no. 5 (October 2016): 1423–65. https://doi.org/10.1086/688176.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. “Measuring Technological Innovation over the Long Run.” *American Economic Review: Insights* 3, no. 3 (September 1, 2021): 303–20. https://doi.org/10.1257/aeri.20190499.

Michalopoulos, Stelios, and Melanie Meng Xue. “Folklore.” *The Quarterly Journal of Economics* 136, no. 4 (October 11, 2021): 1993–2046. https://doi.org/10.1093/qje/qjab003.

Truffa, Francesca, and Ashley Wong. “Undergraduate Gender Diversity and the Direction of Scientific Research.” *American Economic Review (Conditionally Accepted)*, 2024.https://www.dropbox.com/s/qpz64fh8cs6dyg3/coed_draft.pdf?e=1&dl=0.

“What Drives Media Slant? Evidence From U.S. Daily Newspapers.” *Econometrica* 78, no. 1 (2010): 35–71. https://doi.org/10.3982/ECTA7195.

Resources

Text Data Course (Elliott Ash / Benjamin Arold)

- https://github.com/elliottash/text_econ_2022
- https://github.com/BenjaminArold/Course-Text-Data-23
- Our course borrows a lot from this one

Coding for Economists:

- https://aeturrell.github.io/coding-for-economists/

Intro to Deep Learning:

- https://dl-intro.readthedocs.io/en/latest/
- A course on deep learning and NLP taught by Janoś Gabler at Bonn in 2023.

Melissa Dell’s course: Deep Learning for Economics (2023)

- https://econdl.github.io
- Review article: https://www.aeaweb.org/articles?id=10.1257/jel.20241733
- Very comprehensive course on deep learning, with an overview of CNNs, RNNs and Transformers
- Applications focus mostly on Computer Vision and Text Recognition

Data Sources:

- Wikipedia
- (Digitized) Archives
- Manifesto: https://manifesto-project.wzb.eu/

  - Court Documents
  - News data
  - https://www.oja-guide.de/ (a database of German Job Ads, suggested by Periklis)

Learning Materials

Books

- *Natural Language Processing in Python*, Third Edition ("NLTK Book").
  - Available at nltk.org/book.
  - Classic treatments of traditional NLP tools.
- Aurelien Geron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2019)
  - O'Reilly Book, should be available with an academic account using ETH email.
  - A great practical book for machine learning and deep learning in Python, but not NLP-focused. We will use material from Chapters 2-4, 7-11, 13, and 15-17.
  - The deep learning chapters use Keras + TensorFlow.
  - Jupyter notebooks
- Yoav Goldberg, *Neural Network Methods for Natural Language Processing* (2017)
  - ETH Library Online Access (email me if this doesn't work)
  - A more advanced theoretical treatment of neural networks with an NLP focus, but already somewhat dated. We will use material from Chapters 1-17 and 19.
- Jurafsky and Martin, *Speech and Language Processing* (3d Ed. 2019).
  - Available here.
  - The standard theory text on computational linguistics.

Programming

Python is probably the best option for NLP, used by most data scientists. All the sample code is in Python.

- New to Python?

    - Python installation instructions
    - Codecademy Online Python Course
    - numpy tutorial
    - pandas tutorial
    - Jupyter Notebooks Tutorial
    - Jupyter Notebook Keyboard Shortcuts
    - Google Colab Tips for Power Users
    - Dash Web Apps Tutorial
    - Other Resources
- New to Machine Learning?
    - Codecademy Machine Learning Course
    - Read the Geron Book, Chapters 1-7
    - fast.ai Practical Deep Learning for Coders Course
- New to Text Mining / NLP?
    - Codecademy Online NLP Course
    - Read the NLTK Book, Chapters 1-5
    - fast.ai Code-First Introduction to Natural Language Processing
- Papers with Code (NLP)
    - Lists of papers with replication repos.
- Other resources:
    - How to use the terminal
    - How to use Google Colab notebooks
    - Introduction to Git

Python Libraries

```
pip install pandas seaborn scikit-learn tensorflow nltk gensim flair spacy transformers
```

- Basics:
    - pandas: data loading and management
    - numpy: scientific computing
    - seaborn: visualization
    - sklearn: general purpose Python ML library
    - Keras + TensorFlow: deep learning library
- Web Scraping:
    - urllib: URL handling
    - beautifulsoup: HTML and XML parser
    - selenium: automating webdriver interactions
- NLP Necessities:
    - nltk: standard NLP tools
    - gensim: topic models and embeddings
    - spaCy: tokenization, NER, syntactic parsing, word vectors
    - flair: sentiment analysis and some other tools (tutorials)
    - huggingface transformers: transformer architectures
- Specialized tools:

- AllenNLP: library of models for semantic role labeling, entailment, question answering, etc
- fastText: library of embeddings
- spacy-transformers: interface from spaCy to huggingface

Word Embeddings

https://www.samyzaf.com/ML/nlp/nlp.html (Word2Vec)

https://rare-technologies.com/word2vec-tutorial/

Transformers

3Blue1Brown, Visual Intro

- Lecture 1 Transformers, https://www.youtube.com/watch?v=wjZofJX0v4M
- Lecture 2 Attention, https://www.youtube.com/watch?v=eMlx5fFNoYc

More visualisations

- Attention https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/
- Transformers http://jalammar.github.io/illustrated-transformer/