A Hands-on Approach to NLP Fundamentals for External Domain Experts in the LLM Era

Angel Daza
j.daza@esciencecenter.nl
Netherlands eScience Center
Amsterdam, The Netherlands

Abstract

With the advent of Large Language Models (LLMs) researchers outside the Natural Language Processing (NLP) field are interested in learning how to process textual data for their own domain research goals. They are particularly motivated to start experimenting directly with LLMs, implicitly neglecting the large amount of accumulated knowledge that NLP has to offer them. In this text, we briefly share our new lesson materials that aim to show aspiring practitioners the strong connection between NLP fundamentals and LLMs, in the form of a two-day workshop. Our training material is mainly aimed at graduate students outside the NLP sphere who have basic technical knowledge and wish to start working with text, is fully open source and available online[1].

1 Introduction

Researchers across fields regularly encounter large textual data sources. While many domain experts have long used NLP techniques in "computational" or "digital" sub-fields, LLMs have dramatically broadened the curiosity about automatic text processing, especially because these tools promise to bypass the steep learning curve of traditional Machine Learning and NLP concepts. However, this accessibility introduces risks: high-quality popular blogs and tutorials exist (HuggingFace, 2022; Alammar, 2022; Parcalabescu, 2024), but they focus on specific topics (hindering learning continuity) or prioritize quick application over theoretical foundations. This can obscure pre-made modeling choices and latent biases, and users who lack sufficient computational proficiency might stay unaware of those risks or simply feel too distant from more academic sources (Hovy, 2022; Jurafsky and Martin, 2023; Alammar and Grootendorst, 2024).

We aim to fill a critical gap that currently exists between NLP fundamentals (the basic concepts that practitioners inside the NLP field learn, regardless of the chosen models) and the hands-on LLM tutorials dominating online pedagogy (what *hands-on* practitioners see). Our teaching material pursues the following goals: i) to show practitioners that sometimes much simpler existing tools can help solve their problem while keeping more control over the experiments; ii) to emphasize that even when using an LLM, comprehension of NLP fundamentals encourages more robust practices, and, more importantly, iii) to encourage proper experimental setups and evaluation of outputs as mandatory, to avoid common pitfalls, misconceptions, and biases which have always been present in NLP-related tasks (Hovy and Prabhumoye, 2021) and persist even in the latest models (Resnik, 2025).

2 Our Approach

The eScience Center offers a range of training courses in the Netherlands, open to all researchers affiliated with Dutch research organizations [2]. We attract a very broad audience in terms of domain expertise (as an example, Table 1 shows the domains of expertise of the participants of our first workshop), but everyone has the goal of putting reproducible research into practice. Within this environment, we designed a two-day workshop called *Fundamentals of Natural Language Processing in Python*, aimed at graduate students, research software engineers, and early-career faculty from diverse domains who possess at least basic technical skills but require structured guidance to navigate NLP methodologies confidently. We follow the teaching style of *The Carpentries*[3], a non-profit whose goal is to teach foundational coding and data

[1] https://github.com/carpentries-incubator/Natural-language-processing

[2] https://www.esciencecenter.nl/digital-skills/

[3] http://carpentries.org/

Proceedings of the Seventh Workshop on Teaching Natural Language Processing (TeachNLP 2026), pages 37–40

science skills to researchers worldwide, through a learn-by-doing approach.

Domain Expertise	Responses
Social Sciences	5
Humanities	3
Research Software Engineering	3
Mathematics or Statistics	2
Economics or Business	2
Education	2
Environmental Sciences	1
Planetary Sciences	1
Computer Science	1
Physical Sciences	1
Space Sciences	1
Library & Information Science	1
Genetics, Bioinformatics	1
Life Sciences	1
Psychology or Neuroscience	1

Table 1: Domain Expertise of the Participants.

Our audience typically has basic experience using the terminal, writing simple python scripts, and sometimes dealing with tabular data, but is not familiar with NLP basic concepts (e.g., tokenization, stop words, part-of-speech tags, embeddings, encoders, attention, etc.). Our first challenge is audience heterogeneity, we cannot satisfy every learner's needs, but we aim to establish familiarity with the core concepts and encourage further informed exploration. For example, those from the social sciences tend to have a solid statistics background, research engineers ace coding skills but are not familiar with linguistic concepts, while those from the humanities often lack deeper coding skills, but understand better the nuances of human language. While this can compromise the teaching flow, it also gives space for discussions from different perspectives and allows to reflect on the fundamental aspects of NLP that are usually taken for granted in the material that people find online.

A second challenge is the time constraint, as we only have two days to cover the whole material. Our hands-on approach prioritizes building implementation confidence by coding along with the participants while discussing the theoretical foundations of processing linguistic data and reasoning behind designing NLP pipelines, highlighting the fact that the engineering steps rely on the linguistic properties of the problem we aim to solve (Opitz et al., 2025). This makes transparent the choices made during NLP model creation, including language modeling. By connecting fundamentals to LLM technologies, learners recognize that LLMs are not just magic black boxes, but still statistical models that can learn and amplify biases from their training data (Gallegos et al., 2024; Resnik, 2025) resulting in gender (Kaneko et al., 2022), cultural (Naous et al., 2024) and linguistic biases (Fleisig et al., 2024) among others. With this in mind, learners can later pursue their own goals independently through our online materials or recommended follow-up readings.

3 Teaching Material

To test our approach, we ran a pilot workshop on December 2nd and 3rd, 2025. As expected, we attracted a broad audience in terms of background (See Table 1). This material was taught in four episodes[4]. Here is a brief summary:

First episode: the first half defines NLP, the common resources and tasks in the field, starting from how to segment a text into sentences and words and including supervised learning, language modeling and text generation. The second half focuses on exercises that expose the different levels of language (morphology, syntax, semantics, pragmatics), and some of the difficulties of processing language: ambiguity, compositionality, discreteness, and sparsity (Goldberg, 2017).

Second episode: explains how to pre-process own text files and build NLP pipelines using spaCy (Honnibal and Montani, 2020) followed by the intuitions behind transitioning from words (as string representations) into vectors (as continuous numerical representations) to represent textual data. We dive into the concept of distributional semantics and how this inspired the Word2Vec algorithms (Mikolov et al., 2013), including code to show them to train and share their own models using gensim (Rehurek and Sojka, 2010).

Third episode: introduces the Transformer architecture (Vaswani et al., 2017), including intuitive descriptions of the Encoder, Attention Mechanism, and Decoder. We then move into using BERT-based WordPiece tokenizers and models (Devlin et al., 2019) with the transformers library (Wolf et al., 2020). We show how BERT incorporates context into word representation by experi-

[4]Full Event website, including agenda: https://esciencecenter-digital-skills.github.io/2025-12-02-ds-nlp/

menting with polysemic words. We then use the `pipeline()` function to tackle NLP tasks such as `'fill-mask'` and `'text-classification'` using Sentiment Analysis as an example. We close this episode with basic evaluation metrics for classification including precision, recall and F1 scores.

Fourth episode: dives into the differences between the vanilla Transformer and the enhancements that make *LLMs* more powerful [5] . We write scripts that use local LLMs with `Ollama`[6] and `langchain` (Chase, 2022) [7] . Practitioners learn to experiment basic prompting by calling portable models (e.g. `SmolLMv2`, `llama3.2:1b`) and also how to build multi-turn interactions. The exercises include showcasing common biases when using LLMs, where we can discuss the outputs together.

4 Conclusion

Our workshop emphasizes that practitioners can benefit from understanding the fundamentals to make informed decisions about which tools best suit their specific research questions and constraints, rather than defaulting to the latest trends. This shows that simpler methods can offer greater control, transparency, and effectiveness than LLMs. Understanding that LLMs remain statistical models subject to biases, and that proper experimental design and output evaluation are mandatory regardless of the tool chosen, enables researchers to navigate the NLP landscape more confidently.

Feedback received during the workshop was predominantly positive, confirming that we fulfilled our aim of providing valuable content and promoting interesting discussions for each segment of our heterogeneous audience. This response also encouraged us to develop different course variants with the same core materials: a more basic version for true technical beginners, and an advanced track that progresses more rapidly through technical details while covering fundamentals more deeply.

The final conclusion is that hundreds of ways of doing valid NLP-related software remain, and there is not necessarily a linear progression where the latest models will always be better for everything and for everyone.

[5]We also highlight the fact that "LLM" is an ill-defined term, but keep using it to avoid confusions.

[6]`https://ollama.ai`

[7]While this feels like an overkill for simple examples, it gives confidence to later explore the APIs on their own.

Limitations

We are aware that there are several limitations when designing a two-day workshop with such an ambitious coverage. Given this time constraint, we obviously just manage to scratch the surface of every topic and cannot aim at replacing deeper NLP theoretical courses. However, we also understand that there is strong demand for more pragmatic hands-on approaches, and we believe that with our course material and workshops, we can appeal to the audience who want to immediately implement software using NLP technologies.

By focusing the material on fundamentals, we also run the risk of not demonstrating the usage of NLP "at scale" immediately, which could demotivate participants; however, if discussions during workshops make them interested enough, this should at least promote a more responsible use of LLMs. We also hope that exposing participants to the relevant fundamentals will encourage critical usage of LLMs, instead of considering them as black-box oracles. Ideally, these foundations will empower researchers to develop their own solutions instead of falling for a consumer-only approach, where dealing with text data becomes a matter of making API calls and processing outputs without questioning their validity.

Acknowledgments

Special thanks to Carsten Schnober and Kody Moodley for their direct contributions, valuable feedback, and dedication to teaching the final version of the first workshop. Also thanks to Laura Ootes and Eva Viviani who came up with the idea of an NLP course for domain experts outside the field, and Thijs Vroegh who contributed to the material; all this planted the seed for what later evolved into the current online course. The development of this course was made possible through the support and guidance of the eScience Center training team, particularly Fenne Riemslagh.

References

Jay Alammar. 2022. The illustrated transformer. `https://jalammar.github.io/illustrated-transformer/`. Accessed: 2025-12-18.

Jay Alammar and Maarten Grootendorst. 2024. *Hands-On Large Language Models*. O'Reilly.

Harrison Chase. 2022. Langchain. https://github.com/langchain-ai/langchain. Accessed: 2025-12-18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Yoav Goldberg. 2017. *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Matthew Honnibal and Ines Montani. 2020. spacy: Industrial-strength natural language processing in python.

Dirk Hovy. 2022. *Text Analysis in Python for Social Scientists: Prediction and Classification*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

HuggingFace. 2022. The hugging face course, 2022. https://huggingface.co/course. [Online; accessed 2025-12-18].

Dan Jurafsky and James H. Martin. 2023. *Speech and Language Processing*, 3rd ed. draft edition. Accessed: 2025-12-18.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Juri Opitz, Shira Wein, and Nathan Schneider. 2025. Natural language processing relies on linguistics. *Computational Linguistics*, 51(3):1009–1032.

Letitia Parcalabescu. 2024. Ai coffee break with letitia. YouTube Channel. Educational videos on AI, NLP, computer vision, and multimodal learning. Accessed: 2025-12-18.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Philip Resnik. 2025. Large language models are biased because they are large language models. *Computational Linguistics*, 51(3):885–906.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.