UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Dottorato di Ricerca in Ingegneria dell'Informazione

Doctoral Course

"Neural Models and Techniques in Natural Language Processing and Information Retrieval"

Prof. Fabrizio Silvestri, Dr. Nicola Tonellotto
DIAG, Sapienza University of Roma (IT), DII, University of Pisa (IT)
fsilvestri@diag.uniroma1.it, nicola.tonellotto@unipi.it

Short Abstract: Advances from the natural language processing community have recently sparked a renaissance in the task of ad-hoc search. Particularly, large contextualized language modeling techniques, such as BERT, have equipped ranking models with a far deeper understanding of language than the capabilities of previous bag-of-words models. Applying these techniques to a new task is tricky, requiring knowledge of deep learning frameworks, and significant scripting and data munging. In this course, we provide background on classical (e.g., Bag of Words), modern (e.g., Learning to Rank). We introduce students to the Transformer architecture also showing how they are used in foundational aspects of modern large language models (e.g., BERT) and contemporary search ranking and re-ranking techniques. Going further, we detail and demonstrate how these can be easily experimentally applied to new search tasks in a new declarative style of conducting experiments exemplified by the PyTerrier search toolkit.

Course Contents in brief:

- PyTorch
- Language Models
- Self-attention
- Transformers
- BERT and beyond
- HuggingFace Transformers
- PyTerrier

- Classical IR: bag of words and probabilistic ranking
- Modern IR: learning to rank
- Contemporary IR: neural models and techniques

Total # of hours of lecture: 20

References:

[1] Vaswani et al. Attention is all you need, NIPS, 2017. Online: https://arxiv.org/pdf/1706.03762.pdf

[2] Yates, Nogueira, Lin. Pretrained Transformers for Text Ranking: BERT and Beyond, Morgan-Claypool, 2021. Online: https://arxiv.org/pdf/2010.06467.pdf

[3] Macdonald, Tonellotto. Declarative Experimentation in Information Retrieval using PyTerrier, ACM, 2020. Online: https://arxiv.org/pdf/2007.14271.pdf

CV of the Teacher

- Fabrizio Silvestri is a full professor at Dipartimento di Ingegneria informatica, automatica e gestionale (DIAG) of the University of Rome, La Sapienza. His research interests lie in the area of Artificial Intelligence and in particular, Fabrizio Silvestri deals with machine learning applied to web search problems, and natural language processing. He is the author of more than 150 papers in international journals and conference proceedings. It holds 9 industrial patents. He is the holder of the "test-of-time" award at the ECIR 2018 conference for an article published in 2007. He is the holder of three best paper awards and other international awards. Fabrizio Silvestri spent 8 years abroad in industrial research laboratories (Yahoo! and Facebook). At Facebook AI, Fabrizio Silvestri has directed research groups for the development of artificial intelligence techniques in order to combat malicious actors who used the Facebook platform for malicious purposes (hate speech, misinformation, terrorism, etc.)
- Nicola Tonellotto is assistant professor at the Information Engineering Department of the University of Pisa since 2019. From 2002 to 2019 he was a researcher at the Information Science and Technologies Institute of the National Research Council of Italy. His main research interests include Cloud Computing, Web Search, Information Retrieval and Deep Learning. He co-authored more than 70 papers on these topics in peer reviewed international journals and conferences. He was co-recipient of the ACM's SIGIR 2015 Best Paper Award, as well as two best paper awards in international workshops. He taught or teaches BSc, MSc and PhD courses on computer architectures, cloud computing, distributed enabling platforms and information retrieval.

Room and Schedule

Room: *Aula Riunioni del Dipartimento di Ingegneria dell'Informazione, Via G. Caruso 16, Pisa – Ground Floor*

Schedule:

Day 1 – 9 – 13. Intro to PyTorch, Language Models, Implementing Word2Vec in PyTorch. Examples in Google Colab.

Day 2 – 9 – 13. Self-attention, Transformers, BERT, and Beyond. HuggingFace Transformers. Examples in Google Colab.

Day 3 – 9 – 13. Intro to Information Retrieval. Classical models and limitations. PyTerrier. Examples in Google Colab.

Day 4 – 9 – 13. Neural Models for IR. Examples in Google Colab.
Day 5 – 9 – 13. Exam