Applied NLP

From Text to Intelligence

Roman Jurowetzki

Aalborg University / University of Strasbourg

February 10–12, 2026

About Me

Roman Jurowetzki

Associate Professor
Aalborg University Business School

Research:
NLP for innovation studies, science policy,
LLMs for social science research

Teaching:
Business Data Science, Applied NLP,
AI for Social Scientists

Find me:

@rjuro on GitHub
@rjuro on Twitter/X
rjuro@business.aau.dk

The NLP Task Landscape

What can we do with text?

Classification
sentiment, stance, spam

Extraction
NER, relations, JSON

Similarity
search, matching

Topic Modeling
BERTopic, LDA

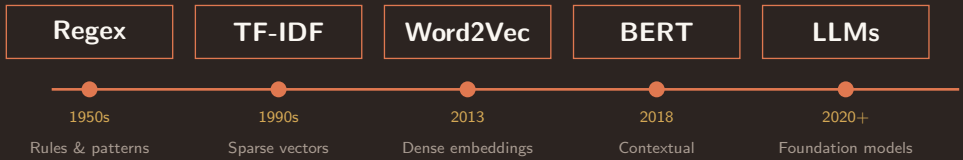
Generation
summaries, translation

Question Answering
RAG, chatbots

This workshop covers all of these — from simple baselines to LLM-powered pipelines.

The 5 Eras of Text Representation

A brief history of how computers read text



Each era didn't replace the previous one — it added a new tool to the toolkit.

TF-IDF still wins sometimes. The right tool depends on your task.

Where We Are Now

The foundation model era

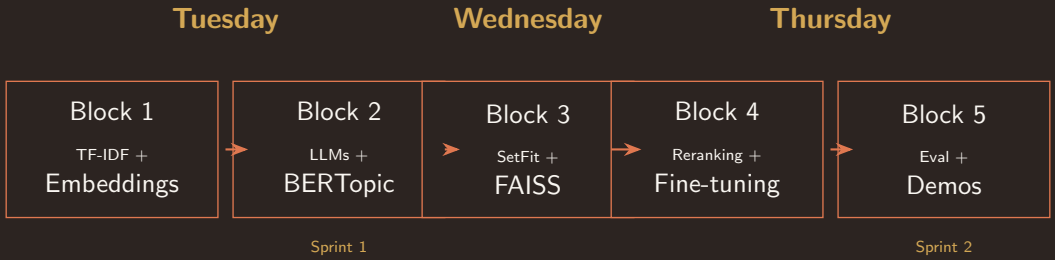
- Embeddings are everywhere — sentence, document, image, code
- Few-shot learning — classify with 8 examples, not 8,000
- Structured output — LLMs return JSON, not just text
- Knowledge distillation — big model teaches small model
- Open models — Llama, Qwen, Mistral — run locally

Key Insight

The cost of NLP went from "PhD + 6 months" to "API call + 5 minutes."
But evaluation and judgment matter more than ever.

Course Roadmap

20 hours across 3 days



Each block: guided notebooks + hands-on exercises + comparison to prior approaches.

The 11 Notebooks

Tuesday

NB01 TF-IDF + Linear Models
NB02 Sentence Embeddings
NB03 LLM Zero-shot
NB04 BERTopic

Wednesday

NB05 SetFit Few-shot
NB06 FAISS Retrieval

Thursday AM

NB07 Cross-encoder Reranking
NB08 LLM Distillation
NB09 Fine-tuning (Qwen3-4B)

Thursday PM

NB10 LLM Evaluation
NB11 Annotation & IRR

All notebooks run in Google Colab. No local setup required.

Tools & Setup

Everything is free

LLM Providers

- **Groq** (primary)
  Free, fast inference
  14,400 req/day
- **Together.AI** (backup)
  $5 free credit
- **Ollama** (local)
  Unlimited, runs in Colab

Infrastructure

- **Google Colab**
  Free T4 GPU
  All notebooks pre-configured
- **HuggingFace**
  Models + datasets
- **GitHub**
  All materials at
  `github.com/RJuro/unistra-nlp2026`

Project Tracks

Pick one, work across Sprint 1 + Sprint 2

A. Classification
Stance, sentiment, toxicity
Datasets: Moltbook, environmental claims

B. Topic Discovery
BERTopic + LLM annotation
Datasets: Moltbook, podcasts, Bluesky

C. Semantic Search
FAISS + cross-encoder reranking
Datasets: policy docs, academic papers

D. Structured Extraction
LLM $\rightarrow$ structured JSON $\rightarrow$ analysis
Datasets: SEC filings, podcast transcripts

Bring your own data — thesis-related encouraged!

Datasets We'll Use

Dataset	Size	Used In	Fun Factor
dk_posts	457 posts	NB01–03	Reddit advice posts
Moltbook	44K posts	NB04	AI agents on social media
Env. Claims	binary	NB05	Greenwashing detection
SciFact	300 docs	NB06–07	Scientific retrieval
SEC Filings	1000s	Project D	Real financial data
Bluesky	2M posts	Project B	Platform migration

All datasets available on HuggingFace or included in the repo.

What You'll Build

By Thursday afternoon

1. A baseline classifier that actually works (TF-IDF)
2. A zero-shot LLM classifier with structured output
3. A semantic search engine with reranking
4. A fine-tuned language model (Qwen3-4B)
5. A project pipeline with honest evaluation

Sprint Deliverables

Sprint 1 (Tue PM): Dataset + baseline + metric + 5 errors
Sprint 2 (Thu PM): Best pipeline + evaluation + model card

Ground Rules

- Ask questions — there are no stupid ones
- Help each other — peer learning is powerful
- Break things — that's how you learn
- Show your errors — error analysis ¿ accuracy chasing
- Use AI tools — Copilot, ChatGPT, Claude are all fair game

Breaks every 90 minutes. Coffee is essential.

Let's start.

Open NB01 in Colab

colab.research.google.com/github/RJuro/unistra-nlp2026/blob/main/notebooks/NB01_tfidf_baselines.ipynb