

Data Science Techniques and Real-World Applications

Presentation Topics

Zahra Sharafi

WS 2025

Data Acquisition and Web Scraping

Web scraping is the process of automatically extracting data from websites to create structured datasets for analysis.

Things to cover:

- Basics: HTML, CSS, DOM, and how data is structured on web pages.
- Tools & Libraries: BeautifulSoup, Scrapy, Selenium, requests.
- Use cases: collecting product prices, reviews, financial data, research datasets.
- Ethical & legal considerations (robots.txt, rate limits, GDPR).

Text Analysis & NLP Latest Developments

Natural Language Processing (NLP) helps computers understand, interpret, and generate human language.

Things to cover:

- Basics: tokenization, stop words, stemming/lemmatization, sentiment analysis.
- Latest developments: Transformer-based models (BERT, GPT), zero-shot/few-shot learning.
- Applications: sentiment analysis on reviews, topic modeling, chatbots.

Recommendation Systems – A Case Study

Recommendation systems suggest items to users based on their preferences and behavior.

Things to cover:

- Types: Content-based filtering, collaborative filtering, hybrid approaches.
- Popular examples: Netflix movie recommendations, Amazon product suggestions, Spotify playlists.
- Case study: walk through how Netflix/YouTube personalizes recommendations.
- Metrics: Precision, Recall, RMSE, MAP, user engagement.
- Challenges: cold-start problem, scalability, bias.

Language Models

Language models predict and generate text, forming the backbone of modern NLP applications.

Things to cover:

- Evolution: n-gram models → RNNs/LSTMs → Transformers.
- How they work: predicting next word, embeddings, attention mechanism.
- Small Language models and Large Language Models
- Applications: machine translation, summarization, question answering.
- Recent trends: GPT, LLaMA, Mistral, Claude.
- Ethical aspects: bias, misuse, hallucinations.

Hugging Face

Hugging Face is an open-source platform providing pre-trained models, datasets, and tools for NLP and beyond.

Things to cover:

- What Hugging Face offers: transformers library, datasets, spaces.
- How to use pre-trained models in just a few lines of code.
 - Bias and Limitations
- Demo: text classification or sentiment analysis with Hugging Face.
- Community & ecosystem: Model Hub, fine-tuning, sharing models.
- Impact: making state-of-the-art NLP accessible to everyone.
 - Case study