# "Foundation of NLP"

Nazarenko Elena, PhD

# Bio

**Elena Nazarenko, PhD**

- Senior Lecturer at HSLU/ Co-head of NLP/LLMs bootcamp
- Senior Data scientist/NLP developer and AI expert at "BIAS" project.
- Served as a Head of Data and AI at Witty Works.
- Built a core algorithm of Witty - inclusive writing assistant at Witty works (part of Hugging Face start-up accelerator, Finalist of Microsoft’s Entrepreneurship for Positive Impact Cup 2024)

**Background:**

PhD University Grenoble Alpes, Research institutes in France, Sweden, Switzerland (Paul Scherrer Institute - ETH Domain)

# Today's Agenda

### Foundational Principles

Named Entities Recognition concepts and applications

### Text Cleaning Pipeline

Understanding the essential preprocessing steps

### Hands-on: Text Cleaning

Practical implementation of cleaning techniques

### Hands-on: Named Entity Recognition

Applied entity extraction and classification

# Named Entity Recognition Fundamentals

### Definition

Text extraction technique that identifies and classifies named entities into predefined categories.

### Entity Types

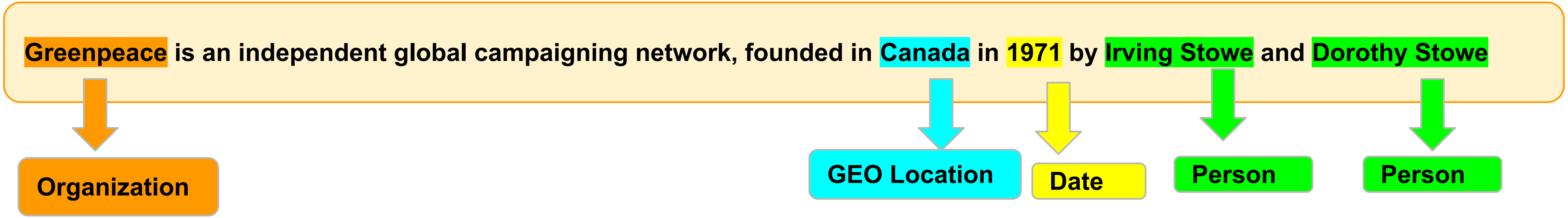
Persons, organizations, locations, dates, specific terminology, and numerical values.

### Applications

Information retrieval, question answering, content recommendation, and research analysis.

# Foundational Principles of NLP - II

**Named Entity Recognition:** information extraction from the text that classifies named entities into predefined categories such as the person names, organizations, locations, numbers, specific terms (medical, legal), percentages, etc.



Greenpeace is an independent global campaigning network, founded in Canada in 1971 by Irving Stowe and Dorothy Stowe

Organization

GEO Location

Date

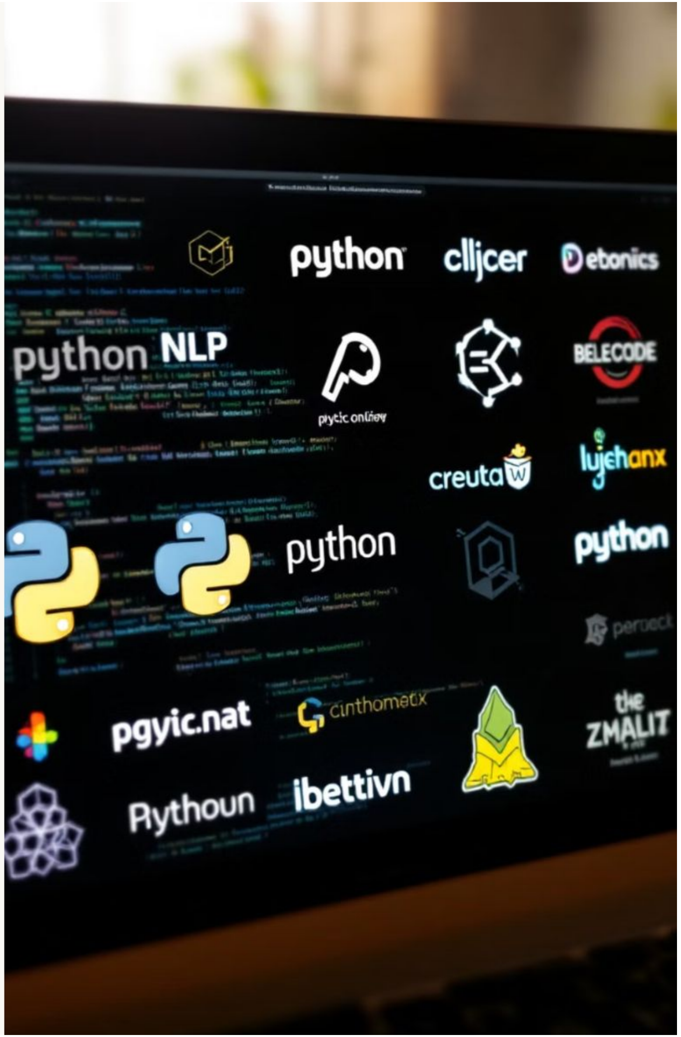
Person

Person

**Importance:**
**Information Retrieval:** Helps in improving the accuracy of search systems by focusing on key terms.
**Question Answering:** Enables systems to answer questions related to specific entities.
**Content Recommendation:** Helps in content personalization and recommendation by focusing on key entities.
**Research:** Useful in extracting structured information from massive datasets for academic or corporate research.

# NLP Toolkit Ecosystem

| Library | Primary Use | Best For |
|---|---|---|
| NLTK | Academic research | Learning, experimentation |
| SpaCy | Production applications | Speed, multilingual support (60+ languages) |
| Stanza | Research-grade analysis | Accuracy, linguistic detail |
| TextBlob | Simplified tasks | Quick prototyping, beginners |
| Gensim | Topic modeling | Document similarity, large corpora |



# Text Cleaning Pipeline: Basic Steps

### Remove HTML Tags

Strip markup language elements from text

### Remove URLs & Emails

Clean web addresses and contact information

### Normalize Text

Convert to lowercase and remove extra spaces

### Handle Contractions

Expand shortened forms like "don't" to "do not"

# Text Cleaning Pipeline: Advanced Cleaning Steps:

Remove special characters, digits, non-ASCII characters

Correct typos

Remove emojis

Perform spelling correction

# Text Cleaning Pipeline: Preprocessing

### Tokenization

Breaking text into words, phrases or symbols

### Stop Word Removal

Filtering common words with little semantic value

### Lemmatization

Reducing words to their base forms

### Custom Cleaning

Domain-specific filters and normalization

# Hands-on
# Text cleaning

# Hands-on
# Named Entity Recognition