Module 3: Analyzing text content with natural language processing

Lesson 3.1: Natural language processing (NLP) in social science

AI-aided content analysis of sustainability communication

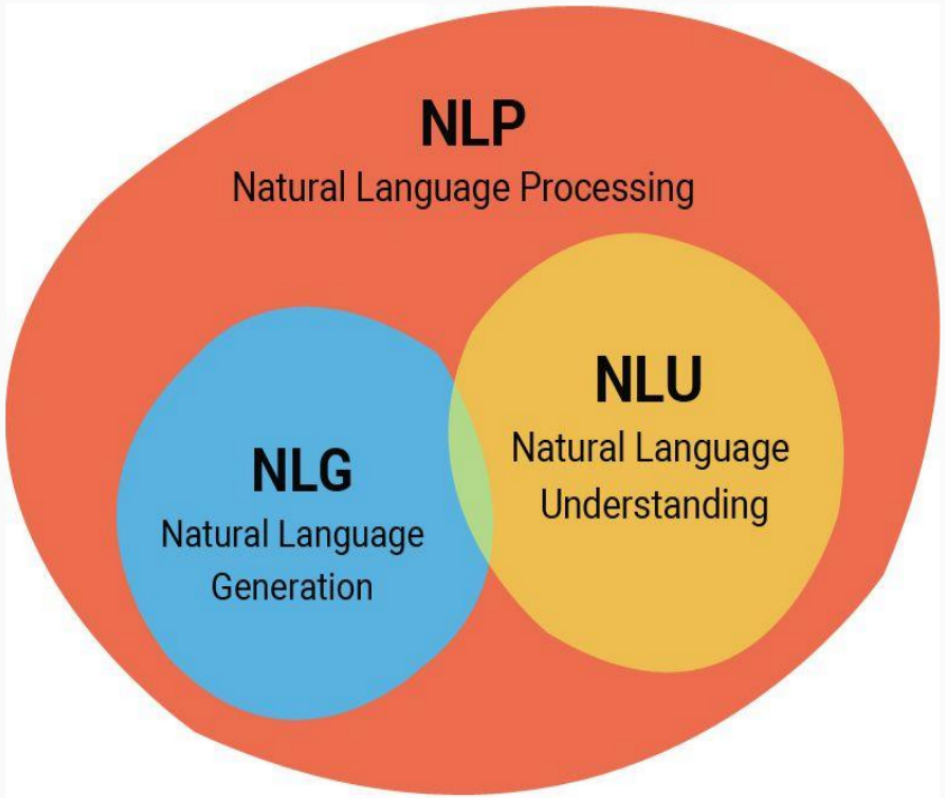
nils.holmberg@iko.lu.se

Functions of Texts in Sustainability Communication

- Informational texts provide data-driven insights and factual details.
- Persuasive texts motivate audiences toward action or change.
- Narrative texts create emotional connections to sustainability themes.
- Visual-supported texts enhance accessibility and engagement.
- Each text type targets specific audiences and communication goals.

NLP Areas and Challenges of Unstructured Text

- NLP applications include sentiment analysis, translation, and summarization.
- Unstructured text often contains ambiguous language and incomplete sentences.
- Domain-specific jargon and colloquialisms complicate processing.
- Noise in data sources like social media adds layers of preprocessing needs.
- Robust algorithms and preprocessing pipelines mitigate these challenges.



Basic Concepts: Units, Tokens, and N-grams

- Text units include sentences, words, and characters as analytical building blocks.
- Tokens are the smallest logical units, often derived from words.
- N-grams capture sequences of n tokens, revealing contextual patterns.
- Common n-grams include bigrams (two words) and trigrams (three words).
- These concepts underpin more advanced NLP tasks.

Formats and Conversion to Plain Text

- Text is often stored in formats like PDF, HTML, or Markdown.
- PDFs may contain layout artifacts, complicating text extraction.
- HTML requires parsing to remove tags and extract meaningful content.
- Tools like Beautiful Soup and Tika streamline these conversions.
- Converting to plain text ensures compatibility with NLP workflows.

Text Features: Readability, POS, and NER

- Readability indices assess the complexity of written content.
- POS tagging categorizes words by their grammatical function.
- NER identifies and classifies specific entities, such as names and dates.
- These features provide insights into the style and structure of text.
- They are essential for contextual and thematic understanding in NLP.

Reading Text into Dataframes and Preprocessing

- Dataframes structure text data for analysis and visualization.
- Normalization includes lowercasing and punctuation removal.
- Tokenization splits text into analyzable units like words or phrases.
- Pandas and NLTK are widely used for preprocessing workflows.
- Clean, tokenized data is a prerequisite for most NLP tasks.

Manifest Text Content and Frequency Analysis

- Manifest content refers to explicitly observable text elements.
- Sentence and word counts provide quantitative content metrics.
- Word frequency analysis highlights key terms and dominant themes.
- Visualizations like word clouds offer intuitive insights into text data.
- These metrics are foundational for exploratory text analysis.