LUND UNIVERSITY

Module 3: Analyzing text content with natural language processing

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

nils.holmberg@iko.lu.se

Functions of Texts in Sustainability Communication

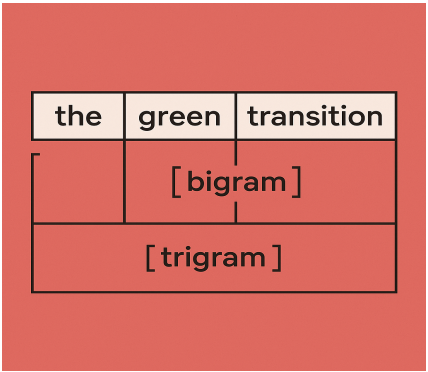
- Informational texts deliver facts
- Persuasive texts drive action
- Narratives build emotional ties
- Visual support boosts engagement
- Each text targets audiences

NLP and Challenges of Unstructured Text

- NLP covers sentiment, translation
- Unstructured text stays ambiguous
- Domain jargon complicates processing
- Data noise increases preprocessing
- Robust pipelines reduce challenges

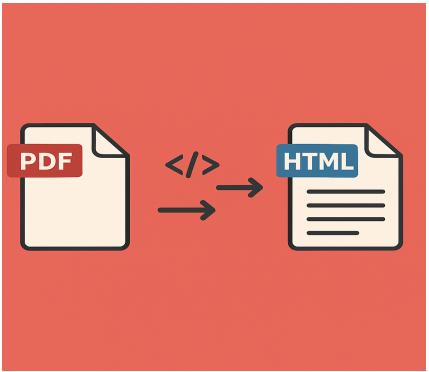
Basic Concepts: Units, Tokens, and N-grams

- Units span sentences, words
- Tokens capture smallest units
- N-grams model token context
- Common n-grams: bigrams, trigrams
- These basics support NLP



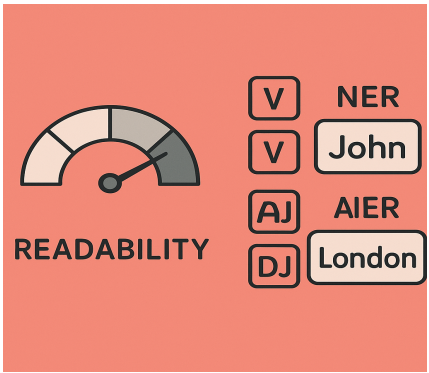
Formats and Conversion to Plain Text

- Text stored as PDF, HTML
- PDF artifacts hinder extraction
- HTML parsing strips markup
- BeautifulSoup, PDFMiner simplify conversion
- Plain text supports NLP



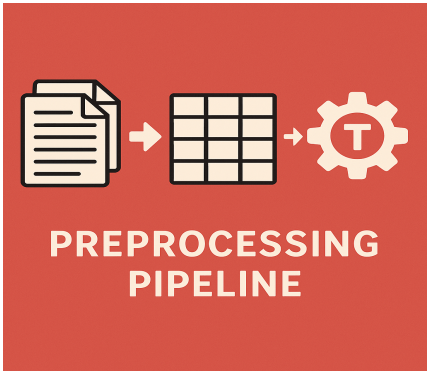
Text Features: Readability, POS, and NER

- Readability indices gauge complexity
- POS tagging labels grammar
- NER detects named entities
- Features reveal text structure
- Features enable contextual understanding



Reading Text into Dataframes and Preprocessing

- Dataframes organize text analysis
- Normalization standardizes casing
- Tokenization splits text units
- Pandas, NLTK drive preprocessing
- Clean tokens enable NLP



Manifest Text Content and Frequency Analysis

- Manifest content captures explicit text
- Sentence, word counts quantify
- Word frequencies surface themes
- Word clouds visualize insights
- Metrics support exploratory analysis