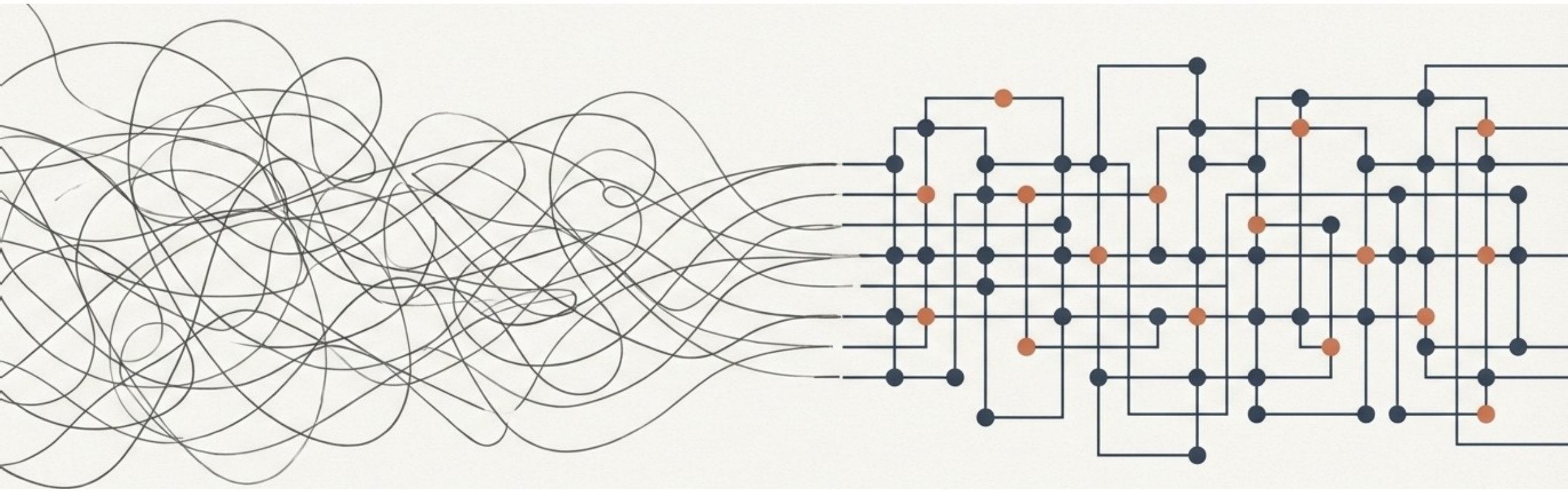
NLP Essentials

A Foundational Course in Analyzing and Understanding Human Language



M1S2 / 5 Credits

The Modern World is Written in Text.

Human language is the most common form of data. We are creating it constantly, from professional reports to casual conversations. The ability to process and understand this data at scale is a critical skill.

How do we give machines the ability to read, understand, and derive meaning from it all?

Your Starting Point

This course offers a foundational introduction to Natural Language Processing. We focus on the essential techniques and methodologies used to analyze and process human language data.

Foundational Theory

Master the classical methods that power modern NLP.

Hands-On Skills

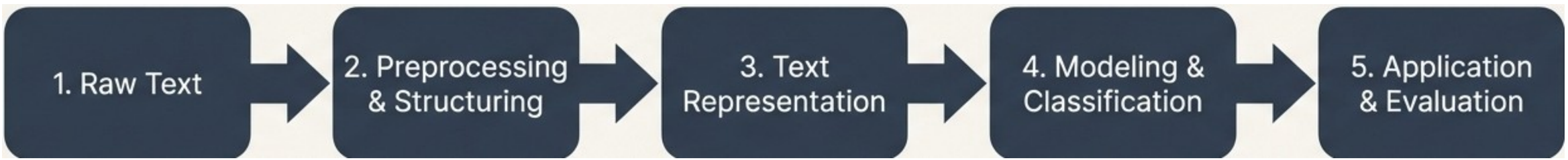
Move from concept to code with practical exercises on real-world text.

A Prerequisite for the Future

Build the base for advanced studies in AI and machine learning-based language models.

Our Roadmap: Building the NLP Pipeline

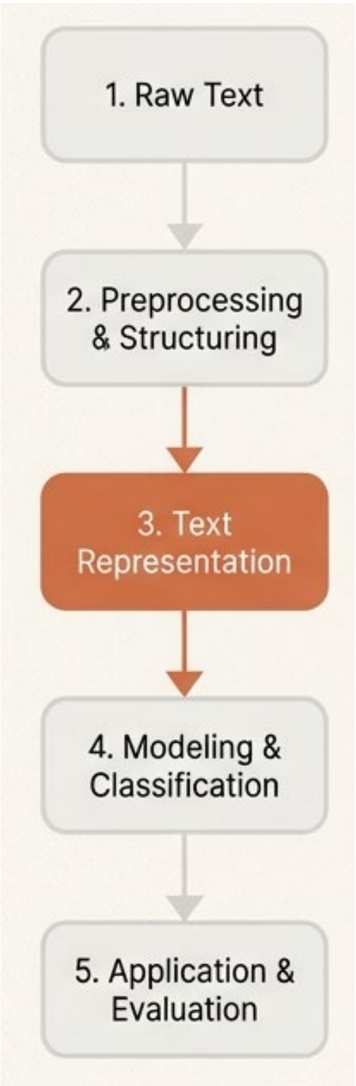
We will construct a complete processing pipeline, piece by piece. Each module adds a new, essential capability, taking us from raw text to actionable understanding.



Module 1: Deconstructing Language

Focus: Learning to see the fundamental structure in raw text.

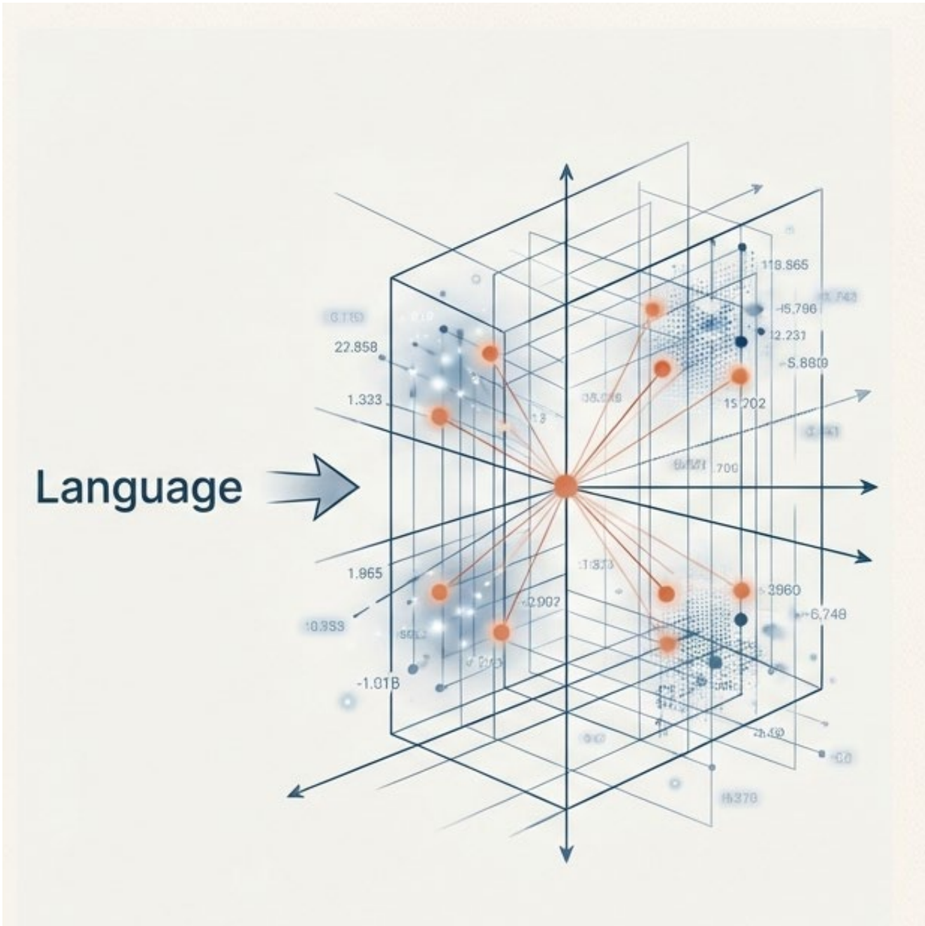
- **Text Preprocessing**: Normalization (lowercasing, punctuation removal), Stemming vs. Lemmatization.
- **Tokenization**: Rule-based and statistical strategies for sentence and word boundary detection.
- **Part-of-Speech (POS) Tagging**: Identifying nouns, verbs, adjectives using models like HMM.
- **Named Entity Recognition (NER)**: Finding people, places, and organizations.
- **Syntactic Parsing:** Understanding sentence structure through constituency and dependency parsing.



Module 2: Translating Words into Vectors

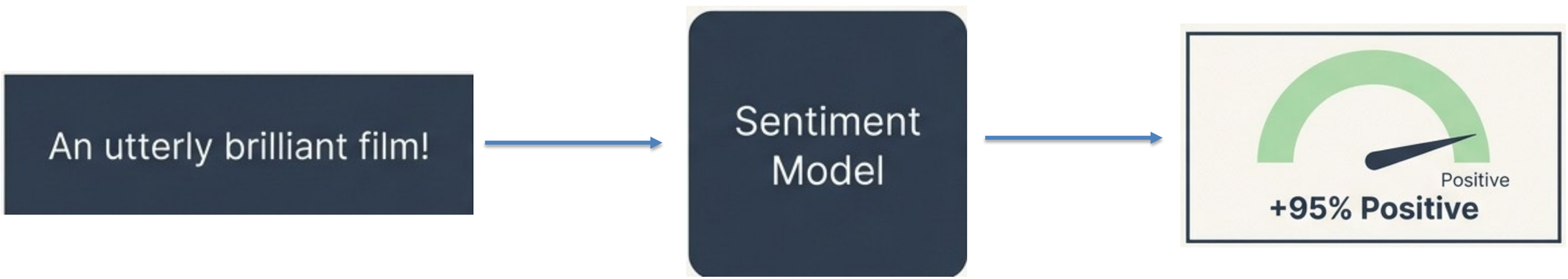
Focus: Moving from symbolic tokens to meaningful numerical representations.

- **Classical Representations:** Bag-of Words and TF-IDF-understanding their motivation, implementation, and limitations.
- **A Glimpse of the Modern:** A high-level introduction to the concepts behind dense word embeddings (e.g., word2vec, GloVe).



Module 3: Building Real-World Applications

Focus: Using our structured data and representations to perform classification and prediction.



- **Basic Sentiment Analysis:** Rule-based and supervised classifiers.
- **N-gram Language Modeling:** Sequence probability and smoothing.
- **Evaluation:** Metrics like perplexity for model assessment .

The Toolkit

Language — PYTHON

Environment — JUPYTER NOTEBOOKS

Libraries — NLTK | spaCy | scikit learn | TextBlob

Data Sources

- IMDB Reviews
- News Articles
- Social Media Data

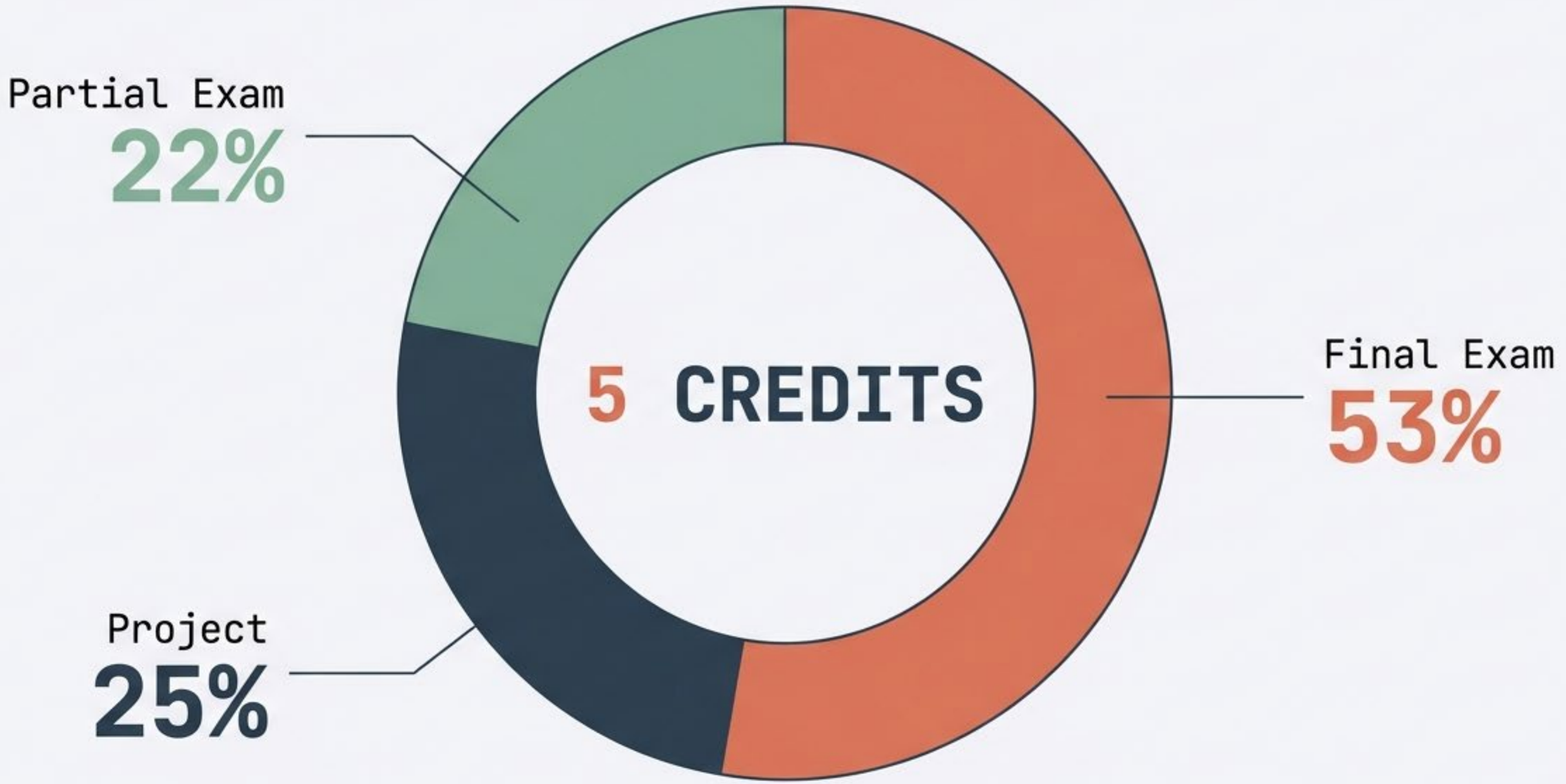
By the end of this course, you will be able to:

1. **Dissect and prepare** any raw text corpus for machine analysis.

2. **Implement algorithms** to extract structural features like entities and parts-of-speech.

3. **Build and evaluate** classic models for tasks like sentiment classification.

4. **Create n-gram models** to understand language sequence and probability.

5. **Establish a robust foundation** for tackling advanced NLP and LLM challenges.

Assessment Strategy