NLP TECHNIQUES SIMILAR BUT DIFFERENT

Understanding Commonly Confused
Natural Language Processing Concepts

TOKENIZATION VS CHUNKING

Tokenization

What is it? Breaking text into individual meaningful units (tokens) like words, punctuation, or subwords.

Purpose: Create the smallest processing units for analysis.

Example:

- Input: "Hello, world!"
- Output: ["Hello", ",", "world", "!"]

Chunking

What is it? Grouping tokens into meaningful phrases or grammatical units.

Purpose: Identify syntactic structures and extract meaningful phrases.

Example:

- Input: ["The", "big", "red", "car", "is", "fast"]
- Output: [("The big red car", "NOUN_PHRASE"), ("is fast", "VERB_PHRASE")]

Key Difference: Tokenization breaks DOWN text, Chunking groups UP tokens.

STEMMING VS LEMMATIZATION

Stemming

What is it? Crude removal of word endings to find the root form using simple rules.

Purpose: Reduce words to their basic form quickly but roughly.

Example:

- "running" → "run"
- "studies" → "studi"
- "better" → "better"

LEMMATIZATION

What is it? Intelligent reduction of words to their dictionary base form using linguistic knowledge.

Purpose: Find the actual dictionary form (lemma) of words.

Example:

- "running" →"run"
- "studies" →"study"
- "better" → "good"

Key Difference: Stemming is FAST but CRUDE, Lemmatization is ACCURATE but SLOW.

STOP WORD REMOVAL VS NOISE FILTERING

Stop Word Removal

What is it? Removing common words that don't carry significant meaning (the, is, at, which).

Purpose: Focus on content-bearing words for analysis.

Example:

- Input: "The cat is on the mat"
- Output: "cat mat"

Noise Filtering

What is it? Removing various types of irrelevant content (URLs, special characters, HTML tags).

Purpose: Clean text from technical artifacts and formatting.

Example:

- Input: "Check out https://example.com for more info! #NLP"
- Output: "Check out for more info"

Key Difference: Stop words are LINGUISTIC noise, Noise filtering removes TECHNICAL artifacts.

PART-OF-SPEECH TAGGING VS NAMED ENTITY RECOGNITION

POS Tagging

What is it? Assigning grammatical categories (noun, verb, adjective) to each word.

Purpose: Understand grammatical structure and word functions.

Example:

- "John/NOUN runs/VERB quickly/ADVERB"

Named Entity Recognition (NER)

What is it? Identifying and classifying named entities (people, places, organizations).

Purpose: Extract specific real-world entities from text.

Example:

- "John/PERSON lives in Paris/LOCATION and works at Google/ORGANIZATION"

Key Difference: POS focuses on GRAMMAR roles, NER identifies REAL-WORLD entities.

WORD EMBEDDINGS VS TF-IDF

Word Embeddings

What is it? Dense vector representations that capture semantic relationships between words.

Purpose: Represent words in continuous vector space where similar words are close.

Example:

- "king" - "man" + "woman" ≈ "queen"
- Words with similar meanings have similar vectors

TF-IDF (Term Frequency-Inverse Document Frequency)

What is it? Sparse numerical representation based on word frequency and rarity across documents.

Purpose: Measure word importance in a document relative to a corpus.

Example:

- Common words get low scores
- Rare but frequent-in-document words get high scores

Key Difference: Embeddings capture SEMANTIC meaning, TF-IDF measures STATISTICAL importance.

DEPENDENCY PARSING VS CONSTITUENCY PARSING

Dependency Parsing

What is it? Analyzing grammatical relationships between words (who does what to whom).

Purpose: Understand how words relate to each other functionally.

Example:

- "John eats apples"
- eats John (subject)
- eats apples (object)

Constituency Parsing

What is it? Breaking sentences into nested grammatical phrases and clauses.

Purpose: Understand hierarchical sentence structure.

Example:

- "John eats apples"
- [S [NP John] [VP [V eats] [NP apples]]]

Key Difference: Dependency shows RELATIONSHIPS, Constituency shows STRUCTURE.

SENTIMENT ANALYSIS VS EMOTION DETECTION

Sentiment Analysis

What is it? Determining overall polarity (positive, negative, neutral) of text.

Purpose: Understand general attitude or opinion.

Example:

- "I love this movie!" Positive
- "This is terrible" Negative

Emotion Detection

What is it? Identifying specific emotions (joy, anger, fear, sadness, surprise).

Purpose: Recognize detailed emotional states.

Example:

- "I'm so excited!" Joy
- "This makes me furious!" Anger

Key Difference: Sentiment is BINARY/TERNARY, Emotions are MULTI-CATEGORICAL.

TEXT NORMALIZATION VS TEXT PREPROCESSING

Text Normalization

What is it? Converting text to a standard, consistent format (lowercasing, expanding contractions).

Purpose: Reduce variations of the same content.

Example:

- "Don't" "do not"
- "USA" "united states america"

Text Preprocessing

What is it? Comprehensive cleaning pipeline including normalization, tokenization, cleaning, etc.

Purpose: Prepare text for specific NLP tasks through multiple steps.

Example:

Complete pipeline: normalization tokenization stop word removal lemmatization

Key Difference: Normalization is ONE STEP, Preprocessing is the ENTIRE PIPELINE.

Remember the Key Patterns

- Tokenization splits, Chunking groups
- Stemming cuts roughly, Lemmatization reduces intelligently
- Stop words are linguistic, Noise is technical
- POS tags grammar, NER finds entities
- Embeddings capture meaning, TF-IDF measures importance
- Dependency shows relations, Constituency shows structure
- Sentiment judges polarity, Emotions identify feelings
- Normalization standardizes, Preprocessing is the full process

STATFUSION AI