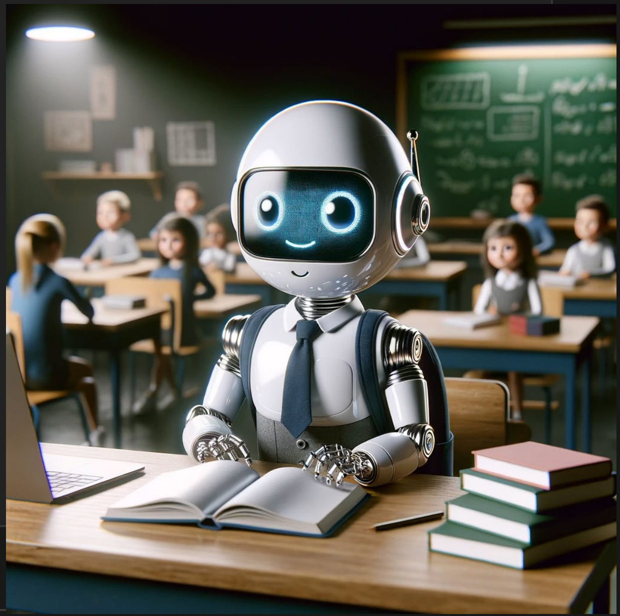
MODULE 9: TOPIC MODELING & ADVANCED TEXT ANALYSIS

ITAI 2373: Natural Language Processing

TODAY'S LEARNING OUTCOMES

- Distinguish between supervised and unsupervised NLP tasks
- Implement Topic Modeling using LDA and NMF
- Build automatic text summarization systems
- Apply document similarity measures for content analysis
- Evaluate the quality of unsupervised models



OUR ROADMAP FOR TODAY

- The Unsupervised World
- Deep Dive: Topic Modeling (LDA & NMF)
- Application: Text Summarization
- Application: Document Similarity
- Evaluation & Wrap-Up

BUILDING ON OUR FOUNDATION

MODULE 4 (VECTORIZATION)

We still need to turn text into numbers (TF- IDF is key here)

MODULE 8 (CLASSIFICATION)

What happens when you have no labels? That's where we are now

Building Connections

Cumulative Learning

SUPERVISED VS. UNSUPERVISED LEARNING

SUPERVISED

We have: (data, labels)

Goal: Predict the label for new data

> **Example:** Is this email spam or not spam?

UNSUPERVISED

We only have: (data)

Goal: Discover hidden structure in the data

> **Example:** What are the main topics in a collection of news articles?

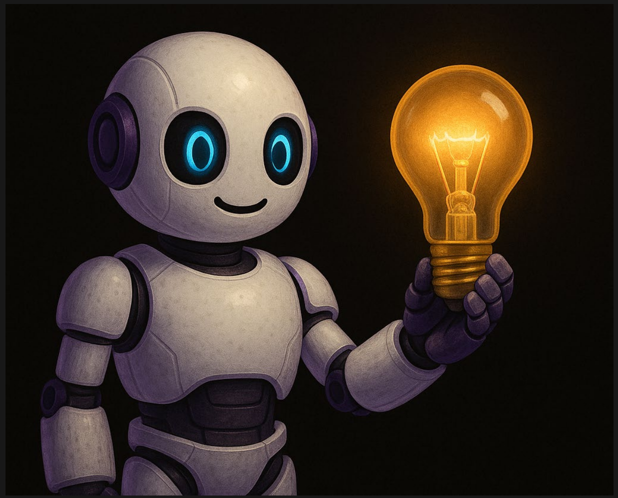
TOPIC MODELING: WHAT IS IT?

Automatically finds hidden themes or "topics" within a large collection of text documents

No human labeling needed: The model discovers these themes on its own

Example: Grouping thousands of customer reviews into topics like "product quality," "shipping issues," or "customer service"

THE CORE IDEA BEHIND TOPIC MODELING



- Documents are made of multiple topics: Think of a news article covering both "politics" and "economy"
- Topics are defined by specific words: The "politics" topic might feature words like election, government, vote, while "economy" has market, stock, revenue
- The model learns both simultaneously: What topics are in each document, and what words define each topic

Core Concepts

Documents + Topics + Words

WHERE IS TOPIC MODELING USED?

NEWS ANALYSIS

Automatically categorizing articles and identifying major trends

CUSTOMER FEEDBACK

Sifting through thousands of reviews to find common complaints or praises

SCIENTIFIC RESEARCH

Discovering evolving themes and sub-fields in vast academic literature

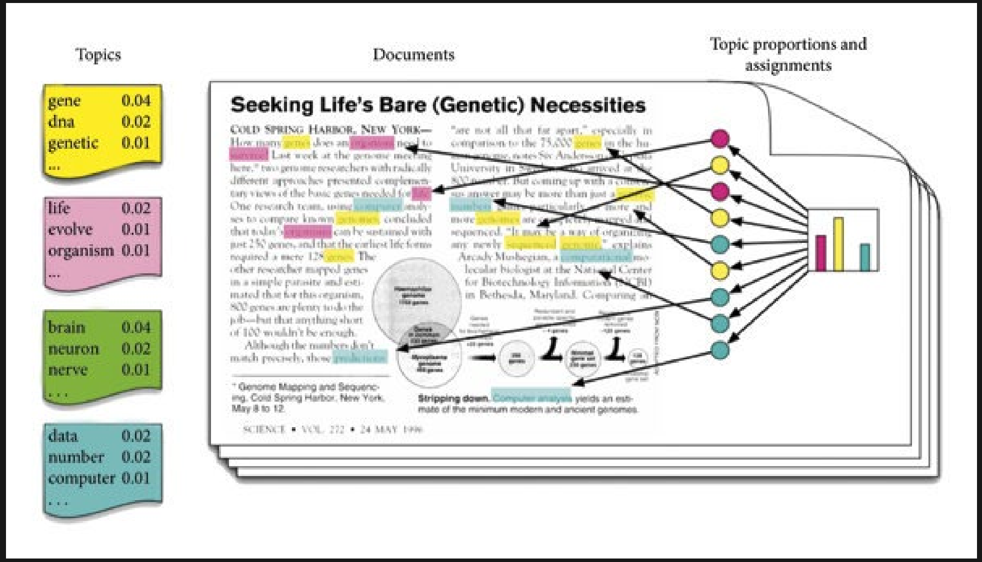
E-DISCOVERY (LEGAL)

Efficiently grouping millions of legal documents by subject for review

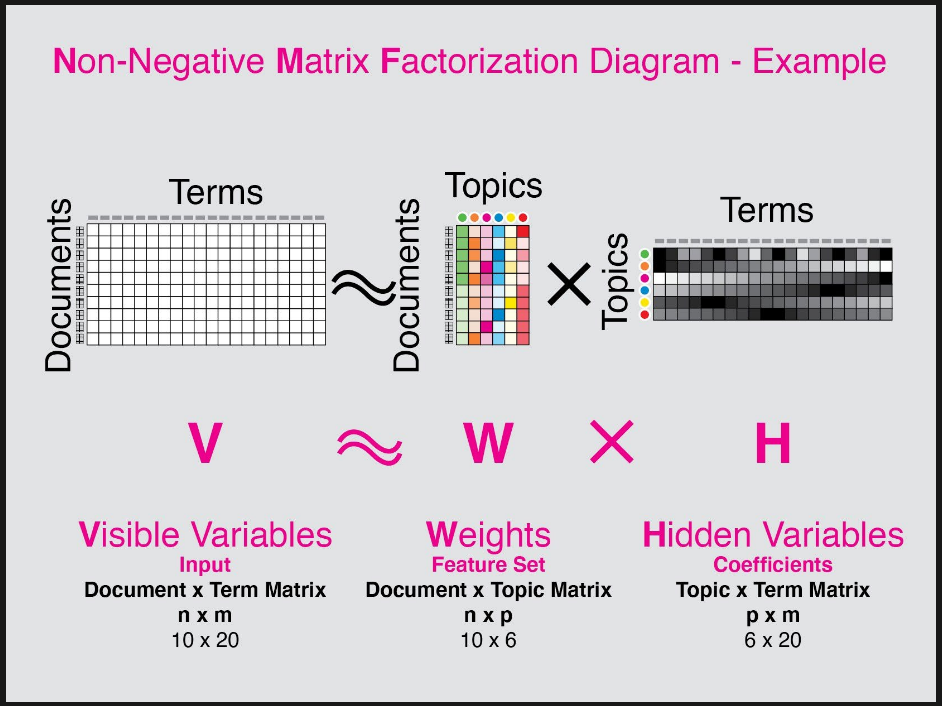
</> TWO MAIN ALGORITHMS

LATENT DIRICHLET ALLOCATION (LDA)

Probabilistic Approach



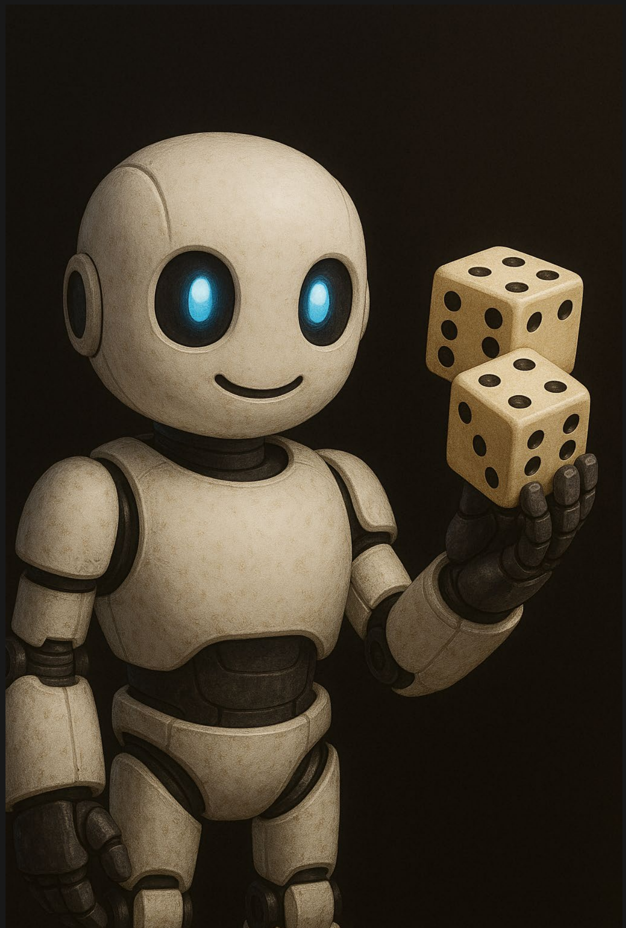
NON-NEGATIVE MATRIX FACTORIZATION (NMF)



LATENT DIRICHLET ALLOCATION (LDA)

- **Topic Modeling Algorithm** that automatically discovers **hidden topics** in large collections of text documents without prior knowledge of topic structure
- **Core Assumption**: Each document is a mixture of topics, and each topic is a mixture of words with different probabilities
- **Statistical Process**: Uses Bayesian inference to iteratively assign words to topics and documents to topic distributions until convergence
- **Applications**: Content recommendation, document clustering, information retrieval, and exploratory data analysis of text corpora

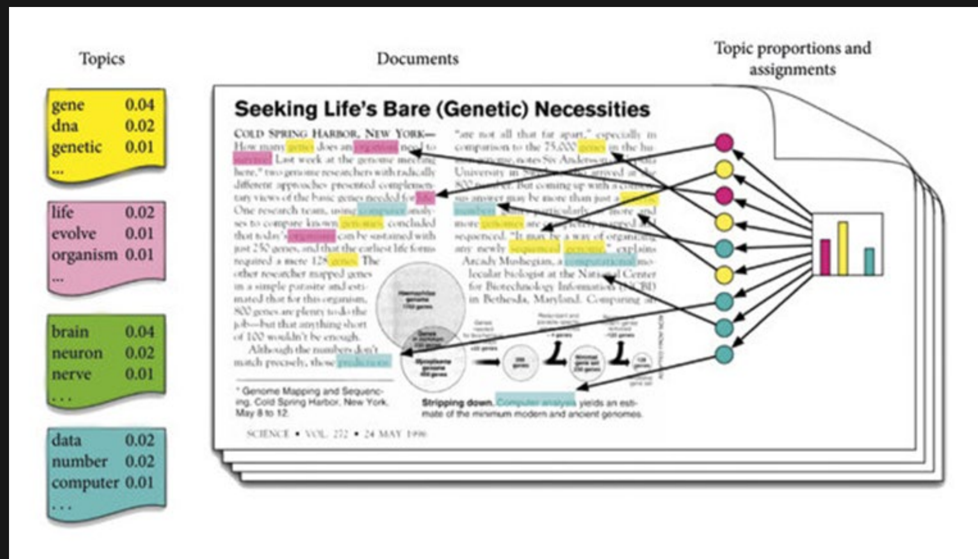
🎲 LDA: THE INTUITION (DICE ANALOGY)



- Topics are like "**loaded dice**": Each die is biased towards certain words (e.g., a "Sports" die rolls "game," "team" more often)
- Documents are "rolls" from a mix of dice: To create a document, you pick a few dice (topics) and roll them repeatedly to get words
- LDA's job: To figure out the dice (topics) and their biases (word distributions) by looking at the words in the documents

HOW LDA WORKS (SIMPLIFIED)

- **Input:** A collection of documents
- **You specify:** The number of topics (k)
- **LDA outputs:** A topic mixture for each document A word distribution for each topic



LDA Process

→ Input →Process →Output

EXAMPLE: LDA ON NEWS ARTICLES (K=4 TOPICS DISCOVERED)

TOPICS DISCOVERED:

Topic 1 (Politics): politics, election, government, vote, president

Topic 2 (Sports): sports, game, team, player, score

Topic 3 (Business): business, stock, market, company, economy

Topic 4 (Technology): tech, software, apple, data, ai

Document Example: "The presidential election is heating up."

LDA Output:

- 90% Topic 1 (Politics)
- 5% Topic 3 (Business)
- 5% Topic 4 (Technology)

Interpretation: LDA successfully identified distinct themes and assigned the document primarily to the "Politics" topic

👓 HOW TO INTERPRET LDA TOPICS

- Look at the top N words for each topic
- Give each topic a human-readable name
- This is a qualitative, human-in-the-loop process

⚖ LDA: THE GOOD AND THE BAD

PROS

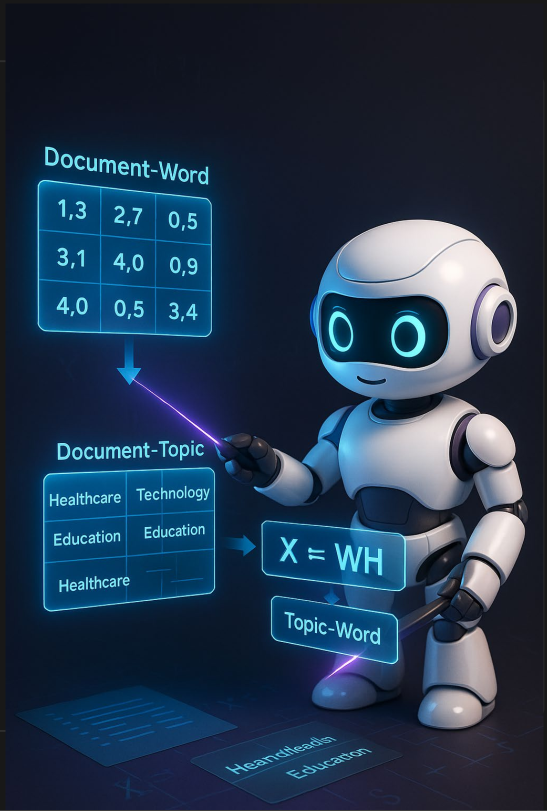
- Produces topic mixtures (soft clustering)
- Strong theoretical foundation
- Works very well in practice

CONS

- Have to specify the number of topics (k)
- Can be slow on very large datasets
- Topics are not always easy to interpret

NON-NEGATIVE MATRIX FACTORIZATION (NMF)

- Another powerful topic modeling algorithm
- Based on Linear Algebra: It breaks down a large document-word table into two smaller, meaningful tables
- Often produces distinct topics: Can sometimes be easier to interpret than LDA topics

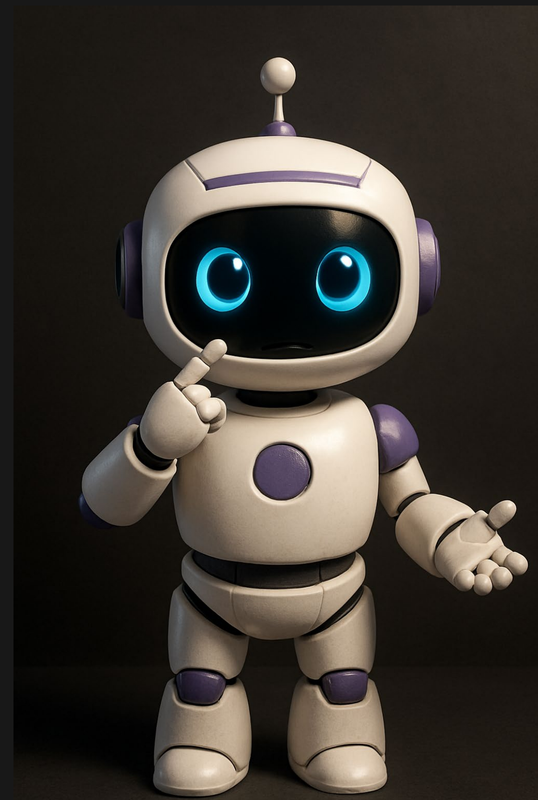


NMF: THE INTUITION (BREAKING DOWN DATA)

Imagine your documents and words as a big table (Document-Word Matrix) NMF "breaks" this table into two smaller, simpler tables:

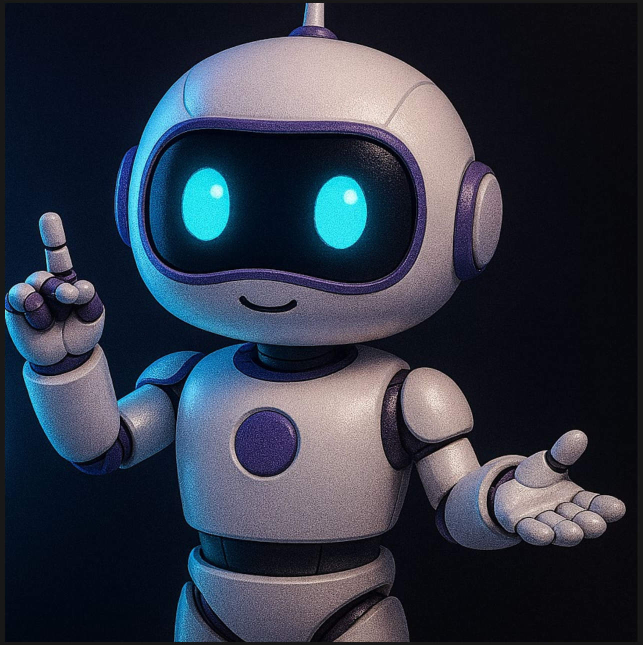
Document-Topic Table (shows how much each document relates to each topic) Topic-Word Table (shows which words are important for each topic)

Key Idea: It finds hidden patterns by simplifying complex data into its core components



⚙️ HOW NMF WORKS (SIMPLIFIED)

- **Input:** A Document-Word Matrix (TF-IDF)
- **You specify:** The number of topics (k)
- **NMF outputs:**
- A score for each document on each topic A score for each word on each topic



NMF Process

→ Matrix →Decomposition →Insights

LDA VS. NMF: WHICH TO CHOOSE?

LDA

Probabilistic model

Often better for understanding the nuances of document composition

NMF

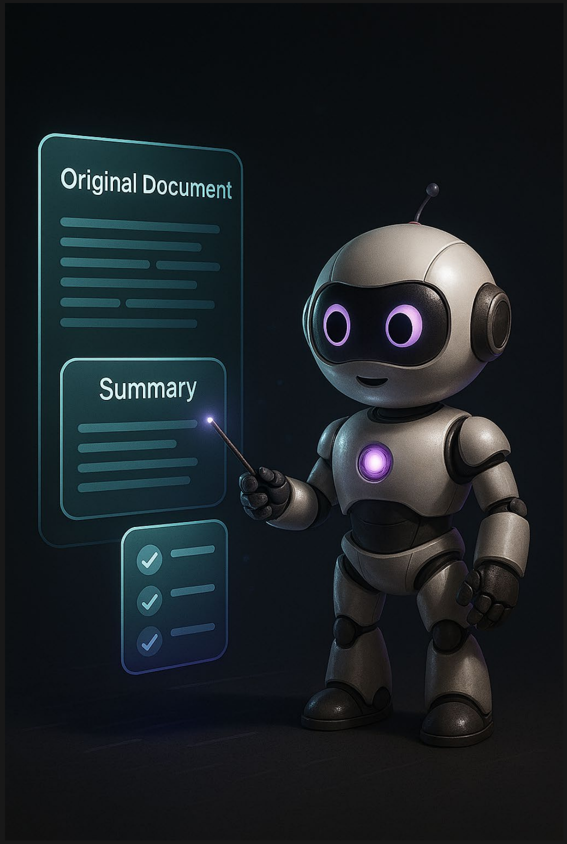
Linear algebra model

Often produces more distinct, interpretable topics

Can be faster than LDA

APPLICATION: TEXT SUMMARIZATION

- **Goal:** To automatically create a shorter, easy-to-read version of a longer document
- **Why it's useful:** Saves time, helps quickly grasp main points, reduces information overload
- **Example:** Turning a long news article into a brief paragraph or a few bullet points



TWO FLAVORS OF SUMMARIZATION

EXTRACTIVE SUMMARIZATION

Picks the most important sentences directly from the original text.

Think: Using a highlighter to mark key sentences.

Original: "The cat sat on the mat. It was sunny."
Summary: "The cat sat on the mat."

ABSTRACTIVE SUMMARIZATION

Generates new sentences to capture the main meaning.

Think: Writing a summary in your own words.

Original: "The cat sat on the mat. It was sunny."
Summary: "A cat enjoyed the sunshine indoors."

HOW EXTRACTIVE SUMMARIZATION WORKS (CONCEPTUALLY)

- **Sentence "Fingerprints":** Each sentence is converted into a numerical representation (like a unique fingerprint)
- **Find "Core" Sentences:** The system identifies sentences that are most similar to many other sentences in the document. These are considered the most important
- **Build Summary:** The top-scoring, most "central" sentences are selected and combined to form the summary

WHERE IS SUMMARIZATION USED?

NEWS APPS

Creating headlines and snippets

SEARCH ENGINES

Generating summaries for search results

BUSINESS INTELLIGENCE

Summarizing long reports for executives

⚖ APPLICATION: DOCUMENT SIMILARITY

Goal: To measure how alike two different text documents are in their content **Output:** A "similarity score" (e.g., from 0 to 1), where higher means more similar **Example:** Comparing a job description to a resume, or finding duplicate articles

HOW TO MEASURE DOCUMENT SIMILARITY (CONCEPTUALLY)

- **Document "Fingerprints":** Each document is converted into a numerical representation (like a unique fingerprint or a topic profile)
- **Compare Fingerprints:** A mathematical method (like "cosine similarity") measures how "close" these fingerprints are.
  - **Score of 1:** Documents are very similar (fingerprints point in the same direction)
  - **Score of 0:** Documents are completely different
  - **Analogy:** Like comparing two people's interests to see how much they have in common

HOW GOOD IS MY TOPIC MODEL (EVALUATION)

The Challenge: Unlike classification, there's no single "correct" answer for unsupervised models

TWO MAIN APPROACHES:

QUANTITATIVE METRICS (E.G., TOPIC COHERENCE)

- Measures if the words within a topic frequently appear together in real documents. (Does the topic "make sense" statistically?)

QUALITATIVE EVALUATION (HUMAN JUDGMENT)

- A human reviews the topics and decides if they are meaningful, distinct, and useful for the task. (Do the topics "look good" to you?)

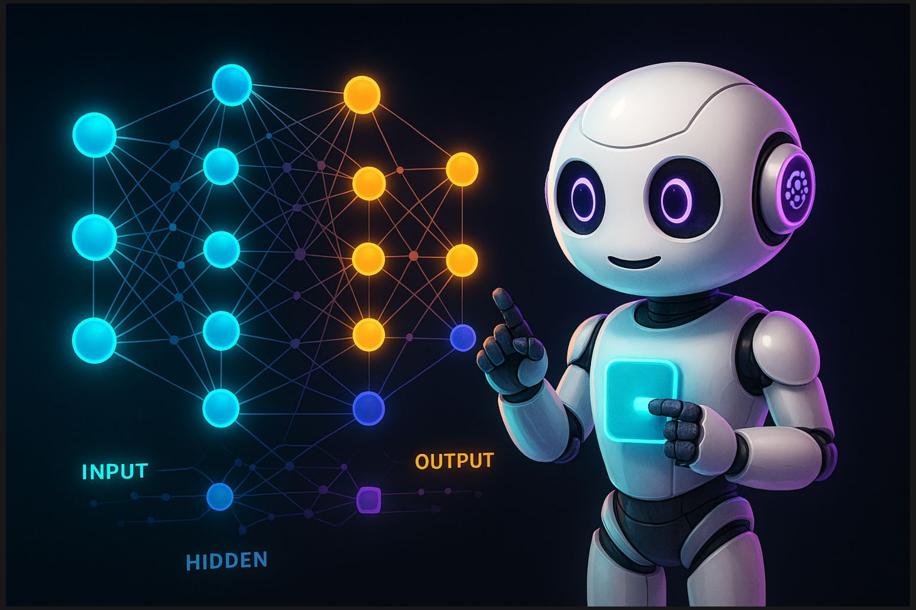
KEY TAKEAWAYS

- Unsupervised methods discover structure in unlabeled data Topic Modeling (LDA, NMF) finds hidden themes in text.
- These techniques power applications like summarization and document similarity Evaluation often requires human judgment

🚀 NEXT TIME: THE NEURAL REVOLUTION

We've mastered classical ML. Now, we enter the world of Deep Learning

Module 1 0 : Introduction to Neural Networks for NLP



Coming Next

⏩ Module 10 Preview