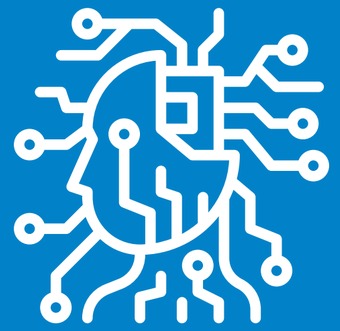
edquest

Day 4 Agenda

NLP & Embeddings

- Recap
- Assignment
- Bag-of-words vs word embeddings
- TF-IDF, Word2Vec, BERT (overview)
- Hands-on with vectorized NLP classification
- Quiz



edquest

Bag-of-words vs word embeddings

What is NLP?

What is NLP?

Natural Language Processing (NLP) is a part of AI that helps computers understand and work with human language, like English or Hindi, etc. The goal is to teach machines how to read, understand, and make sense of text or speech, just like humans do.

Why NLP matters?

Understanding Natural Language is essential for intelligent systems because most human knowledge is stored in unstructured text

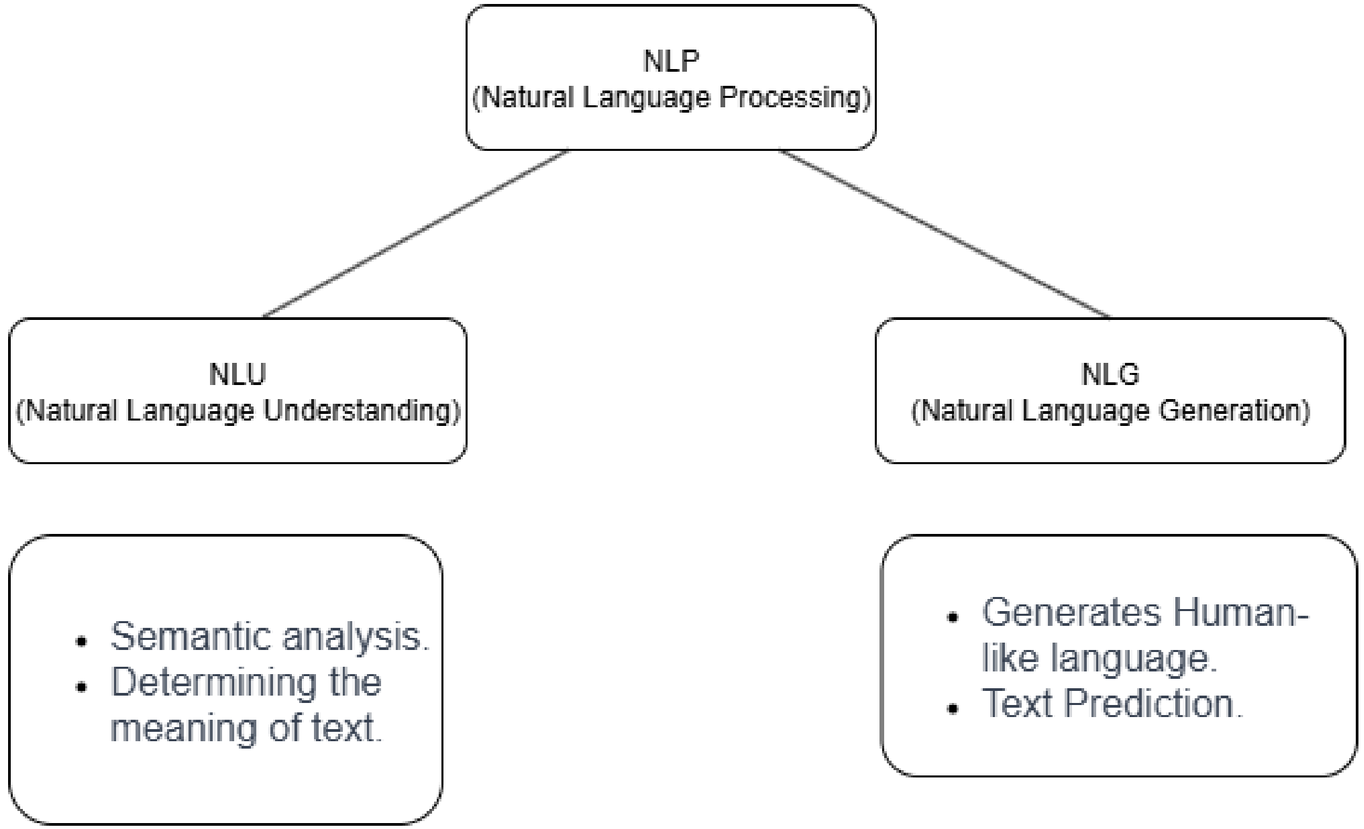
- **Natural communication:** Language is how humans express thoughts, needs, and emotions.
- **Data volume:** ~80% of enterprise data is unstructured (emails, reports, Policy documents).
- **NLP enables machines to:**
  - Interact with users (e.g., Chatbots, Virtual Assistants)
  - Extract insights (e.g., trends from customer reviews)
  - Automate tasks (e.g., spam filtering, document classification)

Why NLP matters?

Understanding Natural Language is essential for intelligent systems because most human knowledge is stored in unstructured text

- **Natural communication:** Language is how humans express thoughts, needs, and emotions.
- **Data volume:** ~80% of enterprise data is unstructured (emails, reports, Policy documents).
- **NLP enables machines to:**
  - Interact with users (e.g., Chatbots, Virtual Assistants)
  - Extract insights (e.g., trends from customer reviews)
  - Automate tasks (e.g., spam filtering, document classification)

NLU &NLG


NLP
(Natural Language Processing)
NLU
(Natural Language Understanding)
NLG
(Natural Language Generation)
Semantic analysis.
Determining the meaning of text.
Generates Human-like language.
Text Prediction.

edquest

NLP in Action

Natural Language Processing is already integrated into many of the tools and technologies we use every day — often without us realizing it. From typing a search query to talking to a virtual assistant, NLP enables machines to process and respond to human language in real-time.

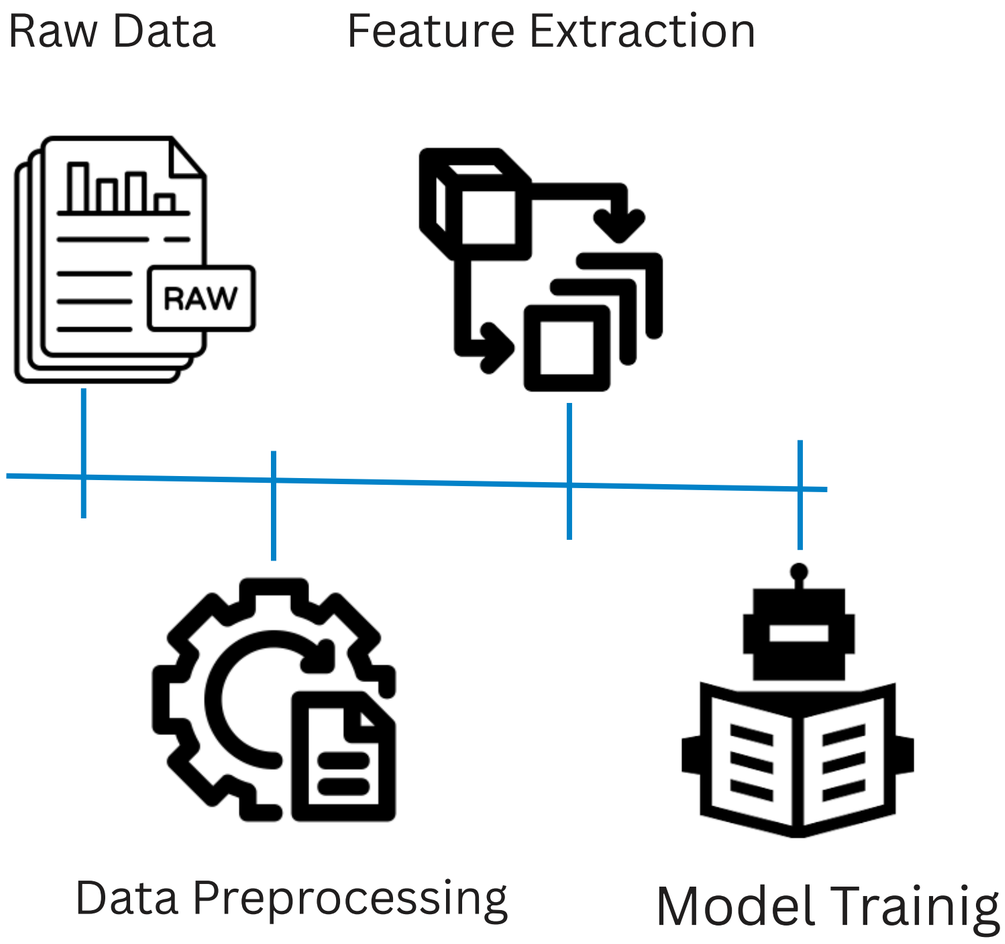
- Search suggestions (Google)
- Virtual assistants (Siri, Alexa)
- Sentiment analysis (social media, reviews)
- Spam detection (emails)
- Text summarization (news, reports)

NLP Industry wise Use-Cases

- Finance
- Healthcare
- Insurance
- Legal
- Telecom
- Agriculture

edquest

NLP Workflow



Data preprocessing

Before a model processes text for a specific task, the text often needs to be preprocessed to improve model performance or to turn words and characters into a format the model can understand.

- **Stemming and lemmatization**
- **Sentence segmentation**
- **Tokenization**

edquest

Feature extraction

Feature extraction is a technique that reduces the dimensionality or complexity of data to improve the performance and efficiency of machine learning (ML) algorithms. This process facilitates ML tasks and improves data analysis by simplifying the dataset to include only its significant variables or attributes.

edquest

What is Bag-Of-Words

Bag-of-Words is a text representation technique that converts text into fixed-length numerical vectors by counting the frequency of each word in a document.

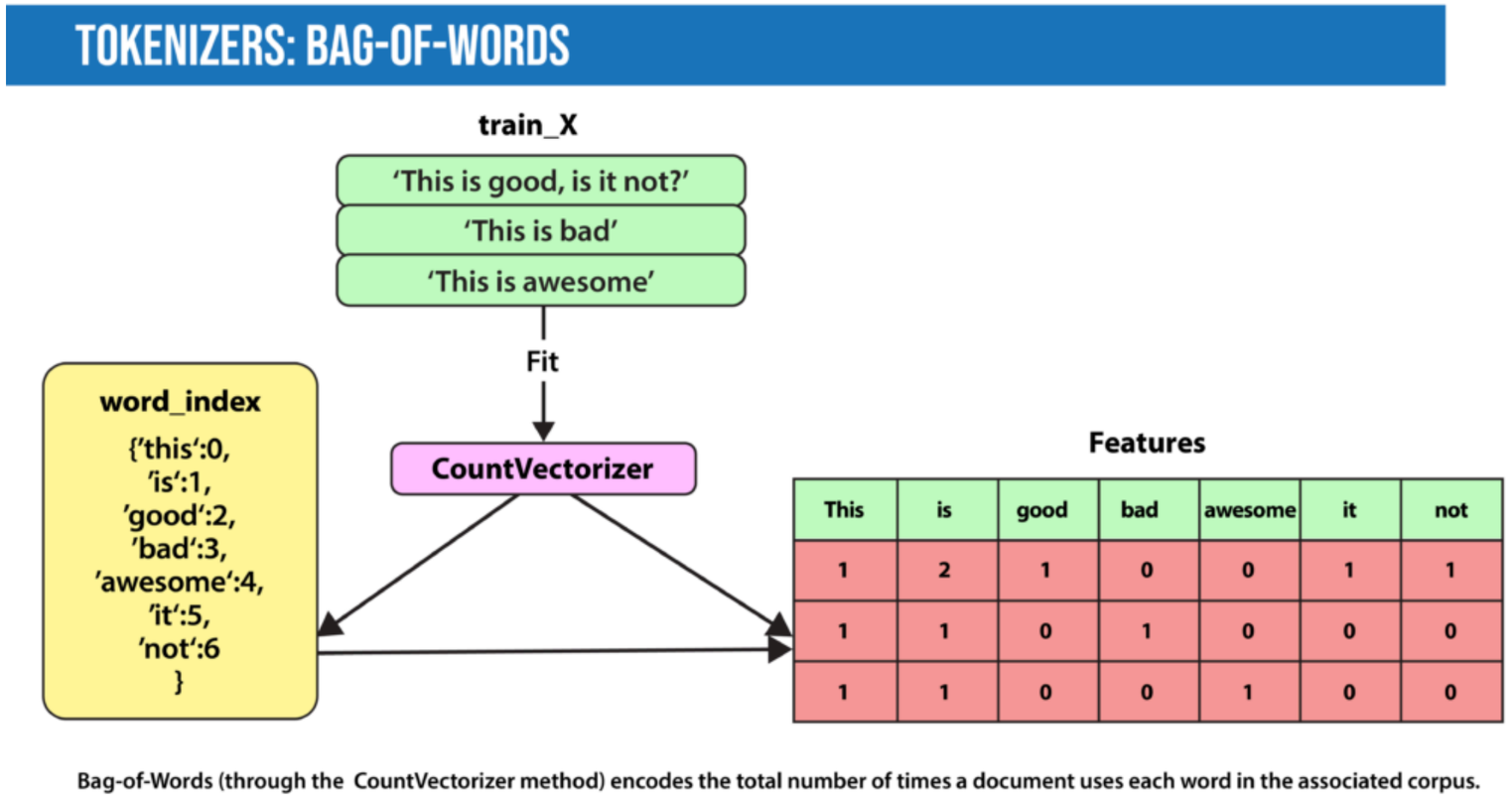
Key Points:

- Creates a vocabulary from all words in the dataset
- Represents each document as a vector of word counts
- Ignores grammar, word order, and semantics

Why “Bag”?

- It treats the words like a “bag” — only the presence and frequency matter, not the order.

Bag-of-Words Example



Bag-of-Words (through the CountVectorizer method) encodes the total number of times a document uses each word in the associated corpus.

Bag-of-Words Pros/Cons

Pros	Cons
Simple to implement	Ignores word order and context
Fast and computationally efficient	High-dimensional and sparse vectors
Good with basic ML models (e.g., SVM)	No understanding of semantics (e.g., synonyms)
No need for external pretraining	Fixed vocabulary — fails with unseen words

Example Limitation:
“Apple is tasty” vs. **“Apple releases new iPhone”** → Same **“Apple”** word, different meanings, but BoW treats them the same.

edquest

What Are Word Embeddings?

edquest

edquest

What Are Word Embeddings?

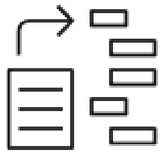
Word embeddings are dense, continuous vectors that represent words in a multi-dimensional space, where distance and direction encode meaning and relationships.

Why They're Important:

- Enable machines to understand semantic relationships between words
- Overcome limitations of one-hot encoding (sparse, high-dimensional, no semantics)
- Foundational to modern NLP and ML applications

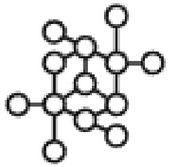
How Word Embeddings Are Used

Word embeddings enhance NLP models by capturing semantic meaning and contextual relationships. They are widely used in tasks such as:



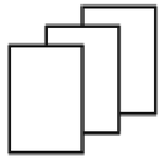
Text classification

Word embeddings are often used as features in text classification tasks, such as sentiment analysis, spam detection and topic categorization.



Named Entity Recognition (NER)

To accurately identify and classify entities (e.g., names of people, organizations, locations) in text, word embeddings help the model understand the context and relationships between words.



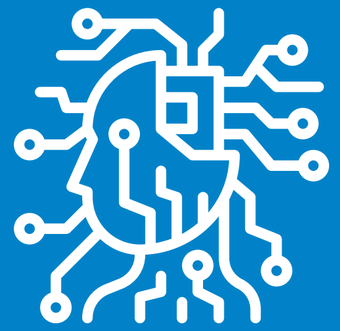
Semantic similarity and clustering

Word embeddings enable measuring semantic similarity between words or documents for tasks like clustering related articles, finding similar documents or recommending similar items based on their textual content.

Assigenement Embeddings?

Generate vector embeddings from the given sentence

e.g: I like to work on MLOps



edquest

TF-IDF, Word2Vec, BERT (overview)

Introduction to TF-IDF

In Natural Language Processing (NLP), TF-IDF (Term Frequency–Inverse Document Frequency) is a key technique used to measure the importance of words within documents relative to an entire corpus.

It helps identify terms that are important in individual documents but not too common across all documents, making it essential for tasks like:

- **Text mining**
- **Information retrieval**
- **Document classification**

What is TF-IDF?

TF-IDF is a numerical score that reflects how important a word is in a document compared to a larger collection of documents (called a corpus).

It combines two ideas:

- **Term Frequency (TF): How often a word appears in a document**
- **Inverse Document Frequency (IDF): How rare the word is across the entire corpus**

Together, TF-IDF highlights important but uncommon words in a document.

edquest

edquest

Term Frequency (TF)

TF measures how often a word appears in a document, relative to the total number of words in that document.

It helps highlight terms that are more important within a specific document.

Formula:

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

What is Inverse Document Frequency (IDF)?

edquest

Inverse Document Frequency (IDF) measures how rare or unique a word is across a set of documents.

- It is calculated using the logarithm of the ratio between the total number of documents and the number of documents containing the word.
- The more documents a word appears in, the lower its IDF score.

Purpose: To penalize common words (like "the", "is", "and") and emphasize words that help differentiate documents.

$$IDF(t, D) = \log\left(\frac{\text{Total number of documents in the corpus } N}{\text{Number of documents containing term } t}\right)$$

What is Word2Vec?

Word2Vec is an NLP technique that learns vector representations of words based on their context in a large corpus of text.

- Developed by Google (2013)
- Captures semantic meaning — similar words have similar vectors
- Uses cosine similarity to measure closeness between word meanings
    - e.g., “walk” ≈ “ran”, “Berlin” ≈ “Germany”

Once trained, Word2Vec can detect synonyms, suggest related words, and power tasks like semantic search or recommendation.

Word2Vec Architectures

Two main approaches:

- CBOW (Continuous Bag-of-Words):
- Predicts the target word from surrounding context words
- Skip-Gram:
- Predicts surrounding words from the target word

CBOW → faster, better for frequent words

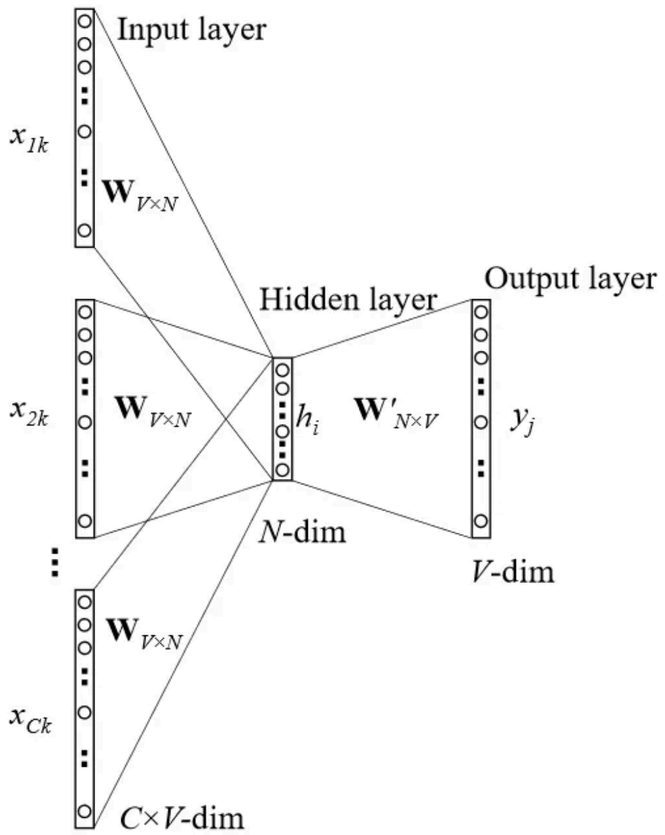
Skip-Gram → better for rare words

edquest

CBOW (Continuous Bag-of-Words) Model

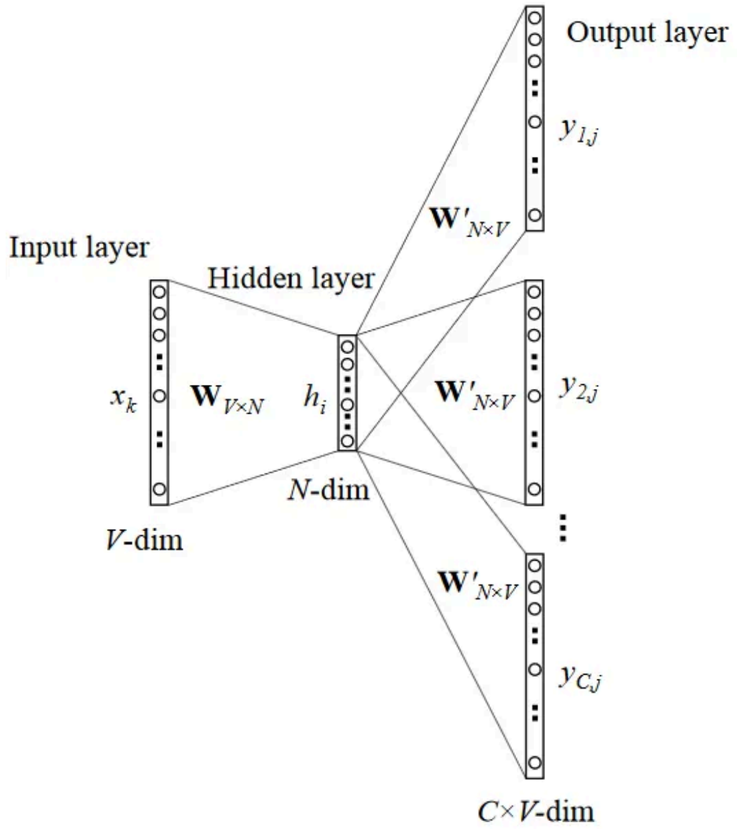
We take words surrounding a given word and try to predict the latter. Each word is a one-hot coded vector. Via an embedding matrix, this is transformed into a N-dimensional vector that's the average of C word vectors. From this vector, we compute probabilities for each word in the vocabulary. Word with highest probability is the predicted word.

CBOW (Continuous Bag-of-Words) Model



Continuous Skip-gram

We take one word and try to predict words that occur around it. At the output, we try to predict C different words.



BERT

BERT is a model for natural language processing developed by Google that learns bi-directional representations of text to significantly improve contextual understanding of unlabeled text across many different tasks.

It’s the basis for an entire family of BERT-like models such as RoBERTa, ALBERT, and DistilBERT.

What Makes BERT Different?

BERT (Bidirectional Encoder Representations from Transformers) — developed by Google in 2018 — is the first model to pre-train deep bidirectional representations using only unlabeled text.

Key Innovations:

- Truly bidirectional: Considers both left and right context simultaneously
- Transformer-based: Uses attention to process all words in parallel
- Context-aware embeddings: Captures meaning based on entire sentence

edquest

How BERT Differs from Earlier Models:

Model	Contextual?	Direction	Architecture
Word2Vec	❌	None	Shallow NN
GloVe	❌	None	Matrix factor.
ELMo	✅	Bi-directional (separate)	Bi-LSTM
BERT	✅	**Jointly bi-directional**	**Transformer**

Why BERT?

BERT is a breakthrough in NLP because it provides a single, pre-trained model that can be fine-tuned for many tasks — without requiring labeled data or complex task-specific architectures.

Why It Matters:

- **Trained on massive text data (no labels required)**
- **Excels in language understanding tasks:**
  - **Q&A, Sentiment Analysis, Translation, Classification**
- **Enables faster development of NLP solutions**

Real-World Benefits of BERT

edquest

Smarter Search

- Understands intent better, even with poor grammar
- Reduces query experimentation and frustration
- Improves relevance → better user experience and ad targeting

Better Natural Language Interfaces

- Enhances voice assistants, chatbots, and helpdesks
- Powers accessible tech for users with disabilities
- (e.g., voice-controlled wheelchairs, web navigation)

Improved Business Intelligence

- Enables non-technical users to extract insights using natural queries
- Reduces errors in data access due to misphrased questions

edquest

Q1. Two different words appear in similar contexts and get mapped to similar vector representations. This behavior is most characteristic of:

A. Bag-of-Words.

B. TF-IDF.

C. One-Hot Encoding.

D. Word Embeddings.

edquest

Q1. Two different words appear in similar contexts and get mapped to similar vector representations. This behavior is most characteristic of:

A. Bag-of-Words.

B. TF-IDF.

C. One-Hot Encoding.

D. Word Embeddings.

Only word embeddings capture context-based similarity. BoW and TF-IDF don't.

edquest

Q2. Which of the following might give high importance to the word "the" in a document?

A. Bag-of-Words.

B. TF-IDF.

C. BERT .

D. Word Embeddings.

edquest

Q2. Which of the following might give high importance to the word "the" in a document?

A. Bag-of-Words.

B. TF-IDF.

C. BERT .

D. Word Embeddings.

BoW counts raw frequencies — so common stopwords like "the" may get high values.

edquest

Q3. Which of the following best defines Natural Language Processing (NLP)?

A. Programming computers to read binary code

B. Teaching machines to interpret and generate human language

C. Compressing human language into zip files

D. Translating HTML documents into JSON

edquest

Q3. Which of the following best defines Natural Language Processing (NLP)?

A. Programming computers to read binary code

B. Teaching machines to interpret and generate human language

C. Compressing human language into zip files

D. Translating HTML documents into JSON

edquest

Q4.Why is BERT considered a breakthrough in Natural Language Processing (NLP)?

A. It requires extensive labeled data for each new task

B. It generates one-hot vectors for sentence representation

C. It uses a single pre-trained model that can be fine-tuned for many NLP tasks

D. It only works for machine translation and not classification

edquest

Q4.Why is BERT considered a breakthrough in Natural Language Processing (NLP)?

A. It requires extensive labeled data for each new task

B. It generates one-hot vectors for sentence representation

C. It uses a single pre-trained model that can be fine-tuned for many NLP tasks

D. It only works for machine translation and not classification

edquest

Q5.What is a key feature of the Word2Vec model in NLP?

A. It represents each word as a unique one-hot vector

B. It predicts the next sentence in a paragraph

C. It learns word vectors based on surrounding context and captures semantic similarity

D. It requires labeled data for training

edquest

Q5.What is a key feature of the Word2Vec model in NLP?

A. It represents each word as a unique one-hot vector

B. It predicts the next sentence in a paragraph

C. It learns word vectors based on surrounding context and captures semantic similarity

D. It requires labeled data for training

edquest

Q6.What does the Inverse Document Frequency (IDF) component of TF-IDF help achieve?

A. Increases the weight of frequently occurring words across all documents

B. Penalizes common words and highlights unique terms that differentiate documents

C. Removes all stopwords from a document

D. Measures the number of characters in a word

edquest

Q6.What does the Inverse Document Frequency (IDF) component of TF-IDF help achieve?

A. Increases the weight of frequently occurring words across all documents

B. Penalizes common words and highlights unique terms that differentiate documents

C. Removes all stopwords from a document

D. Measures the number of characters in a word

edquest

Q7.Why is text preprocessing important before feeding text into an NLP model?

A. To randomly shuffle word order for better generalization

B. To reduce text complexity and convert it into a machine-readable format

C. To generate new vocabulary dynamically during inference

D. To ensure that only labeled data is used in training

edquest

Q7.Why is text preprocessing important before feeding text into an NLP model?

A. To randomly shuffle word order for better generalization

B. To reduce text complexity and convert it into a machine-readable format

C. To generate new vocabulary dynamically during inference

D. To ensure that only labeled data is used in training

edquest

Q7.Which of the following is not typically considered part of data preprocessing in NLP

A. To randomly shuffle word order for better generalization

B. To reduce text complexity and convert it into a machine-readable format

C. To generate new vocabulary dynamically during inference

D. To ensure that only labeled data is used in training