Generative AI

Course Glossary: AI models for NLP

Welcome! This alphabetized glossary contains many of the terms in this course. This comprehensive glossary also includes additional industry-recognized terms not used in course videos. These terms are essential for you to recognize when working in the industry, participating in user groups, and in other certificate programs.

Estimated reading time: 4 minutes

Term	Definition
Bag-of-words	A representation that portrays a document as the aggregate or average of one-hot encoded vectors. It represents documents as a set of words and considers the frequency of a word's occurrence within the document.
Bi-gram model	A conditional probability model with context size one, which means that you consider only the adjacent words in the sequence to predict the next one.
Context vector	Product of the context size and the size of the vocabulary. Typically, this vector is not computed directly but is constructed by concatenating the embedding vectors.
Continuous bag of words (CBOW)	A model that utilizes context words to predict a target word and generate its embedding.
Cross-entropy loss	A metric used to measure the performance of a classification model. The output is a number between 0 and 1. The smaller the number, the better the model.
Data loader	Application component that enables efficient batching and shuffling of data, which is essential for training neural networks. It allows for on-the-fly preprocessing, which optimizes memory usage. Data loaders are important for managing large data sets efficiently during model training.
Data set	A collection of data samples and their labels.
Embedding layer	A layer that accepts token indices and produces embedding vectors.
Fine-tuning	Adjusting a pretrained model to improve performance for a specific task or data set. This makes the model generate more accurate and contextually relevant content.
Gated recurrent units (or GRUs)	A popular recurrent neural network (RNN) enhancements with a gating mechanism to control information flow within the network. They are similar to long short-term memory (LSTM) but can be trained quickly.
Hyperparameters	Configuration settings of a neural network that are external to a model and define aspects such as behavior during training.
Large language models (LLMs)	Foundation models that use AI and deep learning with vast data sets to generate text, translate languages, and create various types of content. They are called large language models due to the size of the training data set and the number of parameters.
Learnable parameters	The weights and biases in a neural network that are optimized during the training of a model.
Learning rate	A hyperparameter that determines how quickly or slowly the neural network learns from the data. It regulates the step size in the optimization process.
Logits	Raw, unnormalized outputs of a neural network before the activation function is applied.
Long short-term memory (or	A popular recurrent neural network (RNN) enhancements effective for tasks involving extensive

LSTMs)	time-series data, such as natural language processing (NLP).
Loss function	A measure that represents the difference between the values predicted by a model and the actual values in the training data
Monte Carlo sampling	A statistical technique that involves generating random samples from a probability distribution. It is specifically beneficial when dealing with systems that involve uncertainty.
Natural language processing (NLP)	The subfield of artificial intelligence (AI) that deals with the interaction of computers and humans in human language. It involves creating algorithms and models that will help computers understand and comprehend human language and generate contextually relevant text in human language.
Neural networks	Computational models inspired by the structure of the human brain. A neural network model consists of an input layer, one or more hidden layers, and an output layer.
N-gram model	Language model that analyzes sequences of 'n' consecutive items, often words, to predict patterns or phrases occurring in a text. The n-gram model allows for an arbitrary context size.
NLTK	A Python library used in natural language processing (NLP) for tasks, such as tokenization and text processing.
One-hot encoding	The method used to convert categorical data into feature vectors that a neural network can understand
Perplexity	Metric for evaluating the efficiency of large language models (LLMs) and generative AI models. In language modeling, perplexity can be seen as a measure of how surprised or uncertain the model is when predicting the next word in a sequence. Lower perplexity values indicate better performance of language models.
PyTorch	A dynamic deep learning framework developed by Facebook's AI Research lab. It is a Python-based library well-known for its ease of use, flexibility, and dynamic computation graphs.
Recurrent neural networks (or RNNs)	Artificial neural networks that use sequential or time series data. You can use RNNs to solve data-related problems with a natural order or time-based dependencies. They have loops in their architecture, allowing information to persist over time, making them suitable for sequential data processing.
Sequence-to-sequence model	Neural network architecture, where both input and output are sequences of data. It is used in machine translation, such as converting English phrases into French.
Skip-gram model	A word embedding model that predicts surrounding context words from a specific target word. A skip-gram model is a reverse of the continuous bag of words (CBOW) model.
Word embedding	Representation of words as dense vectors, capturing their relationship based on the context
Word2vec	The group of models that produce word embeddings or vectors, which are numerical representations capturing the essence of words. It is the short form for "word to vector."