Generative AI Applications with RAG and LangChain

Welcome! This alphabetized glossary contains many terms used in this course. Understanding these terms is essential when working in the industry, participating in user groups, and participating in other certificate programs.

Term	Definition
Chunk size	Refers to the maximum number of characters each text chunk can contain after being split by a text splitter.
Chroma DB	An open-source vector store supported by LangChain, used for storing and retrieving vector embeddings, particularly useful in semantic search engines over text data.
Cosine similarity	A measure used to calculate the similarity between two non-zero vectors of an inner product space, which measures the cosine of the angle between them.
Document loader	A component in LangChain that gathers information from various sources (like websites, files, and databases) and converts it into a format that can be processed by the LangChain framework.
Embedding	A numerical representation of data, typically in a high-dimensional space, that captures the semantic meaning of the data.
JQ schema	A schema used by the JSONLoader in LangChain to parse JSON files according to specific needs, particularly to extract particular values from a JSON structure.
LangChain	A framework that simplifies the development of applications using large language models (LLMs) by providing tools for loading, processing, and querying data from various sources.

Large language model (LLM)	A type of artificial intelligence model designed to understand and generate human language, often used in NLP tasks such as text generation, translation, and summarization.
Markdown header text splitter	A tool in LangChain that splits a markdown file by a specified set of headers, useful for maintaining document structure during text processing.
Maximum marginal relevance (MMR)	A retrieval technique used in vector stores to balance the relevance and diversity of the retrieved results, ensuring comprehensive coverage of different aspects of the query.
PyPDFLoader	A class in LangChain used to load PDF files into an array of document objects, each representing a page along with its metadata.
PyMuPDF loader	A tool in LangChain, known for its speed, that loads PDF files into document objects with detailed metadata about the PDF and its pages, providing one document object per page.
RecursiveCharacterTextSplitter	A text splitter in LangChain that employs recursion to split large texts into smaller chunks using a set of characters, suitable for general text processing.
Retrieval-augmented generation (RAG)	A method that combines retrieval-based and generative-based approaches to improve the quality of the generated responses, often used in question-answering systems.
Self-query retriever	A type of LangChain retriever that converts a query into two components: a string to look up semantically and a metadata filter, used to retrieve documents based on both text and metadata.

Separator	The character or set of characters used by a text splitter to divide the text into manageable chunks, such as a line break or a paragraph change.
Similarity search	A method used in vector databases to find and retrieve the most relevant content based on the similarity of vector embeddings to a given query vector.
Vector database	A specialized type of database designed to store and retrieve vector embeddings, allowing for efficient and effective information retrieval based on similarity calculations.
Vector store-based retriever	A retriever in LangChain that queries a vector database to retrieve the most similar chunks of data to a given query, without requiring an LLM.
WebBaseLoader	A component in LangChain that extracts all text from HTML webpages, converting it into a document format suitable for downstream processing, avoiding unnecessary HTML tags and links.