

[ ![Logo](../../\_static/logo.png) ](../../index.html)

## Getting Started

- \* [Installation](../../docs/installation.html)

- \* [Install with pip](../../docs/installation.html#install-with-pip)

- \* [Install with Conda](../../docs/installation.html#install-with-conda)

- \* [Install from Source](../../docs/installation.html#install-from-source)

- \* [Editable Install](../../docs/installation.html#editable-install)

- \* [Install PyTorch with CUDA support](../../docs/installation.html#install-pytorch-with-cuda-support)

- \* [Quickstart](../../docs/quickstart.html)

- \* [Sentence Transformer](../../docs/quickstart.html#sentence-transformer)

- \* [Cross Encoder](../../docs/quickstart.html#cross-encoder)

- \* [Next Steps](../../docs/quickstart.html#next-steps)

## Sentence Transformer

- \* [Usage](../../docs/sentence\_transformer/usage/usage.html)

- \* [Computing Embeddings](../computing-embeddings/README.html)

- \* [Initializing a Sentence Transformer Model](../computing-embeddings/README.html#initializing-a-sentence-transformer-model)

- \* [Calculating Embeddings](../computing-embeddings/README.html#calculating-embeddings)

- \* [Prompt Templates](../computing-embeddings/README.html#prompt-templates)

- \* [Input Sequence Length](../computing-embeddings/README.html#id1)

- \* [Multi-Process / Multi-GPU Encoding](../computing-embeddings/README.html#multi-process-multi-gpu-encoding)

	*	[Semantic	Textual
Similarity](../../docs/sentence_transformer/usage/semantic_textual_similarity.html)			
	*		[Similarity
Calculation](../../docs/sentence_transformer/usage/semantic_textual_similarity.html#similarity-calculation)			
* Semantic Search			
* Background			
* Symmetric vs. Asymmetric Semantic Search			
* Manual Implementation			
* Optimized Implementation			
* Speed Optimization			
* Elasticsearch			
* Approximate Nearest Neighbor			
* Retrieve & Re-Rank			
* Examples			
* [Retrieve & Re-Rank](../retrieve_rerank/README.html)			
* [Retrieve & Re-Rank Pipeline](../retrieve_rerank/README.html#retrieve-re-rank-pipeline)			
* [Retrieval: Bi-Encoder](../retrieve_rerank/README.html#retrieval-bi-encoder)			
* [Re-Ranker: Cross-Encoder](../retrieve_rerank/README.html#re-ranker-cross-encoder)			
* [Example Scripts](../retrieve_rerank/README.html#example-scripts)			
	*	[Pre-trained	Bi-Encoders
(Retrieval)](../retrieve_rerank/README.html#pre-trained-bi-encoders-retrieval)			
	*	[Pre-trained	Cross-Encoders
(Re-Ranker)](../retrieve_rerank/README.html#pre-trained-cross-encoders-re-ranker)			
* [Clustering](../clustering/README.html)			
* [k-Means](../clustering/README.html#k-means)			
* [Agglomerative Clustering](../clustering/README.html#agglomerative-clustering)			

\* [Fast Clustering](../clustering/README.html#fast-clustering)

\* [Topic Modeling](../clustering/README.html#topic-modeling)

\* [Paraphrase Mining](../paraphrase-mining/README.html)

\*

[`paraphrase\_mining()`](../paraphrase-mining/README.html#sentence\_transformers.util.paraphrase\_mining)

\* [Translated Sentence Mining](../parallel-sentence-mining/README.html)

\* [Margin Based Mining](../parallel-sentence-mining/README.html#margin-based-mining)

\* [Examples](../parallel-sentence-mining/README.html#examples)

\* [Image Search](../image-search/README.html)

\* [Installation](../image-search/README.html#installation)

\* [Usage](../image-search/README.html#usage)

\* [Examples](../image-search/README.html#examples)

\* [Embedding Quantization](../embedding-quantization/README.html)

\* [Binary Quantization](../embedding-quantization/README.html#binary-quantization)

\* [Scalar (int8) Quantization](../embedding-quantization/README.html#scalar-int8-quantization)

\* [Additional extensions](../embedding-quantization/README.html#additional-extensions)

\* [Demo](../embedding-quantization/README.html#demo)

\* [Try it yourself](../embedding-quantization/README.html#try-it-yourself)

\* [Speeding up Inference](../../docs/sentence\_transformer/usage/efficiency.html)

\* [PyTorch](../../docs/sentence\_transformer/usage/efficiency.html#pytorch)

\* [ONNX](../../docs/sentence\_transformer/usage/efficiency.html#onnx)

\* [OpenVINO](../../docs/sentence\_transformer/usage/efficiency.html#openvino)

\* [Benchmarks](../../docs/sentence\_transformer/usage/efficiency.html#benchmarks)

\* [Creating Custom Models](../../docs/sentence\_transformer/usage/custom\_models.html)

\* [Structure of Sentence Transformer

Models](../../docs/sentence\_transformer/usage/custom\_models.html#structure-of-sentence-transfo

mer-models)

\* [Sentence Transformer Model from a Transformers Model](../../../../docs/sentence\_transformer/usage/custom\_models.html#sentence-transformer-model-from-a-transformers-model)

\* [Pretrained Models](../../../../docs/sentence\_transformer/pretrained\_models.html)

\* [Original Models](../../../../docs/sentence\_transformer/pretrained\_models.html#original-models)

\* [Semantic Search Models](../../../../docs/sentence\_transformer/pretrained\_models.html#semantic-search-models)

\* [Multi-QA Models](../../../../docs/sentence\_transformer/pretrained\_models.html#multi-qa-models)

\* [MSMARCO Passage Models](../../../../docs/sentence\_transformer/pretrained\_models.html#msmarco-passage-models)

\* [Multilingual Models](../../../../docs/sentence\_transformer/pretrained\_models.html#multilingual-models)

\* [Semantic Similarity Models](../../../../docs/sentence\_transformer/pretrained\_models.html#semantic-similarity-models)

\* [Bitext Mining](../../../../docs/sentence\_transformer/pretrained\_models.html#bitext-mining)

\* [Image & Text-Models](../../../../docs/sentence\_transformer/pretrained\_models.html#image-text-models)

\* [INSTRUCTOR models](../../../../docs/sentence\_transformer/pretrained\_models.html#instructor-models)

\* [Scientific Similarity Models](../../../../docs/sentence\_transformer/pretrained\_models.html#scientific-similarity-models)

\* [Training Overview](../../../../docs/sentence\_transformer/training\_overview.html)

\* [Why Finetune?](../../../../docs/sentence\_transformer/training\_overview.html#why-finetune)

\* [Training Components](../../../../docs/sentence\_transformer/training\_overview.html#training-components)

\* [Dataset](../../../../docs/sentence\_transformer/training\_overview.html#dataset)

- \* [Dataset Format](../../docs/sentence\_transformer/training\_overview.html#dataset-format)
- \* [Loss Function](../../docs/sentence\_transformer/training\_overview.html#loss-function)
- \* [Training Arguments](../../docs/sentence\_transformer/training\_overview.html#training-arguments)
- \* [Evaluator](../../docs/sentence\_transformer/training\_overview.html#evaluator)
- \* [Trainer](../../docs/sentence\_transformer/training\_overview.html#trainer)
- \* [Callbacks](../../docs/sentence\_transformer/training\_overview.html#callbacks)
- \* [Multi-Dataset Training](../../docs/sentence\_transformer/training\_overview.html#multi-dataset-training)
- \* [Deprecated Training](../../docs/sentence\_transformer/training\_overview.html#deprecated-training)
- \* [Best Base Embedding Models](../../docs/sentence\_transformer/training\_overview.html#best-base-embedding-models)
- \* [Dataset Overview](../../docs/sentence\_transformer/dataset\_overview.html)
  - \* [Datasets on the Hugging Face Hub](../../docs/sentence\_transformer/dataset\_overview.html#datasets-on-the-hugging-face-hub)
  - \* [Pre-existing Datasets](../../docs/sentence\_transformer/dataset\_overview.html#pre-existing-datasets)
- \* [Loss Overview](../../docs/sentence\_transformer/loss\_overview.html)
  - \* [Loss modifiers](../../docs/sentence\_transformer/loss\_overview.html#loss-modifiers)
  - \* [Distillation](../../docs/sentence\_transformer/loss\_overview.html#distillation)
  - \* [Commonly used Loss Functions](../../docs/sentence\_transformer/loss\_overview.html#commonly-used-loss-functions)
  - \* [Custom Loss Functions](../../docs/sentence\_transformer/loss\_overview.html#custom-loss-functions)
- \* [Training Examples](../../docs/sentence\_transformer/training/examples.html)
- \* [Semantic Textual Similarity](../../training/sts/README.html)

- \* [Training data](../../training/sts/README.html#training-data)
- \* [Loss Function](../../training/sts/README.html#loss-function)
- \* [Natural Language Inference](../../training/nli/README.html)
- \* [Data](../../training/nli/README.html#data)
- \* [SoftmaxLoss](../../training/nli/README.html#softmaxloss)
- \* [MultipleNegativesRankingLoss](../../training/nli/README.html#multiplenegativesrankingloss)
- \* [Paraphrase Data](../../training/paraphrases/README.html)
- \* [Pre-Trained Models](../../training/paraphrases/README.html#pre-trained-models)
- \* [Quora Duplicate Questions](../../training/quora\_duplicate\_questions/README.html)
- \* [Training](../../training/quora\_duplicate\_questions/README.html#training)

\*

[MultipleNegativesRankingLoss](../../training/quora\_duplicate\_questions/README.html#multiplenegativesrankingloss)

- \* [Pretrained Models](../../training/quora\_duplicate\_questions/README.html#pretrained-models)
- \* [MS MARCO](../../training/ms\_marco/README.html)
- \* [Bi-Encoder](../../training/ms\_marco/README.html#bi-encoder)
- \* [Matryoshka Embeddings](../../training/matryoshka/README.html)
- \* [Use Cases](../../training/matryoshka/README.html#use-cases)
- \* [Results](../../training/matryoshka/README.html#results)
- \* [Training](../../training/matryoshka/README.html#training)
- \* [Inference](../../training/matryoshka/README.html#inference)
- \* [Code Examples](../../training/matryoshka/README.html#code-examples)
- \* [Adaptive Layers](../../training/adaptive\_layer/README.html)
- \* [Use Cases](../../training/adaptive\_layer/README.html#use-cases)
- \* [Results](../../training/adaptive\_layer/README.html#results)
- \* [Training](../../training/adaptive\_layer/README.html#training)
- \* [Inference](../../training/adaptive\_layer/README.html#inference)

- \* [Code Examples](../../training/adaptive\_layer/README.html#code-examples)
- \* [Multilingual Models](../../training/multilingual/README.html)
- \* [Extend your own models](../../training/multilingual/README.html#extend-your-own-models)
- \* [Training](../../training/multilingual/README.html#training)
- \* [Datasets](../../training/multilingual/README.html#datasets)
- \* [Sources for Training Data](../../training/multilingual/README.html#sources-for-training-data)
- \* [Evaluation](../../training/multilingual/README.html#evaluation)

\* [Available Pre-trained Models](../../training/multilingual/README.html#available-pre-trained-models)

- \* [Usage](../../training/multilingual/README.html#usage)
- \* [Performance](../../training/multilingual/README.html#performance)
- \* [Citation](../../training/multilingual/README.html#citation)
- \* [Model Distillation](../../training/distillation/README.html)
- \* [Knowledge Distillation](../../training/distillation/README.html#knowledge-distillation)

\* [Speed - Performance Trade-Off](../../training/distillation/README.html#speed-performance-trade-off)

- \* [Dimensionality Reduction](../../training/distillation/README.html#dimensionality-reduction)
- \* [Quantization](../../training/distillation/README.html#quantization)
- \* [Augmented SBERT](../../training/data\_augmentation/README.html)
- \* [Motivation](../../training/data\_augmentation/README.html#motivation)

\* [Extend to your own datasets](../../training/data\_augmentation/README.html#extend-to-your-own-datasets)

- \* [Methodology](../../training/data\_augmentation/README.html#methodology)

\* [Scenario 1: Limited or small annotated datasets (few labeled sentence-pairs)](../../training/data\_augmentation/README.html#scenario-1-limited-or-small-annotated-datasets-few-labeled-sentence-pairs)

\* [Scenario 2: No annotated datasets (Only unlabeled

sentence-pairs)](../../training/data\_augmentation/README.html#scenario-2-no-annotated-datasets-only-unlabeled-sentence-pairs)

- \* [Training](../../training/data\_augmentation/README.html#training)
- \* [Citation](../../training/data\_augmentation/README.html#citation)
- \* [Training with Prompts](../../training/prompts/README.html)
- \* [What are Prompts?](../../training/prompts/README.html#what-are-prompts)
- \* [Why would we train with Prompts?](../../training/prompts/README.html#why-would-we-train-with-prompts)
- \* [How do we train with Prompts?](../../training/prompts/README.html#how-do-we-train-with-prompts)
- \* [Training with PEFT Adapters](../../training/peft/README.html)
- \* [Compatibility Methods](../../training/peft/README.html#compatibility-methods)
- \* [Adding a New Adapter](../../training/peft/README.html#adding-a-new-adapter)
- \* [Loading a Pretrained Adapter](../../training/peft/README.html#loading-a-pretrained-adapter)
- \* [Training Script](../../training/peft/README.html#training-script)
- \* [Unsupervised Learning](../../unsupervised\_learning/README.html)
- \* [TSDAE](../../unsupervised\_learning/README.html#tsdae)
- \* [SimCSE](../../unsupervised\_learning/README.html#simcse)
- \* [CT](../../unsupervised\_learning/README.html#ct)
- \* [CT (In-Batch Negative Sampling)](../../unsupervised\_learning/README.html#ct-in-batch-negative-sampling)
- \* [Masked Language Model (MLM)](../../unsupervised\_learning/README.html#masked-language-model-mlm)
- \* [GenQ](../../unsupervised\_learning/README.html#genq)
- \* [GPL](../../unsupervised\_learning/README.html#gpl)
- \* [Performance Comparison](../../unsupervised\_learning/README.html#performance-comparison)



\* [Domain Adaptation](../../domain\_adaptation/README.html)

\* [Domain Adaptation vs. Unsupervised Learning](../../domain\_adaptation/README.html#domain-adaptation-vs-unsupervised-learning)

\* [Adaptive Pre-Training](../../domain\_adaptation/README.html#adaptive-pre-training)

\* [GPL: Generative Pseudo-Labeling](../../domain\_adaptation/README.html#gpl-generative-pseudo-labeling)

\* [Hyperparameter Optimization](../../training/hpo/README.html)

\* [HPO Components](../../training/hpo/README.html#hpo-components)

\* [Putting It All Together](../../training/hpo/README.html#putting-it-all-together)

\* [Example Scripts](../../training/hpo/README.html#example-scripts)

\* [Distributed Training](../../docs/sentence\_transformer/training/distributed.html)

\* [Comparison](../../docs/sentence\_transformer/training/distributed.html#comparison)

\* [FSDP](../../docs/sentence\_transformer/training/distributed.html#fsdp)

## Cross Encoder

\* [Usage](../../docs/cross\_encoder/usage/usage.html)

\* [Retrieve & Re-Rank](../retrieve\_rerank/README.html)

\* [Retrieve & Re-Rank Pipeline](../retrieve\_rerank/README.html#retrieve-re-rank-pipeline)

\* [Retrieval: Bi-Encoder](../retrieve\_rerank/README.html#retrieval-bi-encoder)

\* [Re-Ranker: Cross-Encoder](../retrieve\_rerank/README.html#re-ranker-cross-encoder)

\* [Example Scripts](../retrieve\_rerank/README.html#example-scripts)

\* [Pre-trained Bi-Encoders (Retrieval)](../retrieve\_rerank/README.html#pre-trained-bi-encoders-retrieval)

\* [Pre-trained Cross-Encoders (Re-Ranker)](../retrieve\_rerank/README.html#pre-trained-cross-encoders-re-ranker)

\* [Pretrained Models](../../docs/cross\_encoder/pretrained\_models.html)

- \* [MS MARCO](../../docs/cross\_encoder/pretrained\_models.html#ms-marco)
- \* [SQuAD (QNLI)](../../docs/cross\_encoder/pretrained\_models.html#squad-qnli)
- \* [STSbenchmark](../../docs/cross\_encoder/pretrained\_models.html#stsbenchmark)

\* [Quora Duplicate

Questions](../../docs/cross\_encoder/pretrained\_models.html#quora-duplicate-questions)

- \* [NLI](../../docs/cross\_encoder/pretrained\_models.html#nli)
- \* [Community Models](../../docs/cross\_encoder/pretrained\_models.html#community-models)
- \* [Training Overview](../../docs/cross\_encoder/training\_overview.html)
- \* [Training Examples](../../docs/cross\_encoder/training/examples.html)
- \* [MS MARCO](../../training/ms\_marco/cross\_encoder\_README.html)
- \* [Cross-Encoder](../../training/ms\_marco/cross\_encoder\_README.html#cross-encoder)

\* [Cross-Encoder Knowledge

Distillation](../../training/ms\_marco/cross\_encoder\_README.html#cross-encoder-knowledge-distillation)

## Package Reference

- \* [Sentence Transformer](../../docs/package\_reference/sentence\_transformer/index.html)

\*

[SentenceTransformer](../../docs/package\_reference/sentence\_transformer/SentenceTransformer.html)

\*

[SentenceTransformer](../../docs/package\_reference/sentence\_transformer/SentenceTransformer.html#id1)

\*

[SentenceTransformerModelCardData](../../docs/package\_reference/sentence\_transformer/SentenceTransformer.html#sentencetransformermodelcarddata)

\*

[SimilarityFunction](../../docs/package\_reference/sentence\_transformer/SentenceTransformer.html#similarityfunction)

\* [Trainer](../../docs/package\_reference/sentence\_transformer/trainer.html)

\*

[SentenceTransformerTrainer](../../docs/package\_reference/sentence\_transformer/trainer.html#sentencetransformertrainer)

\* [Training Arguments](../../docs/package\_reference/sentence\_transformer/training\_args.html)

\*

[SentenceTransformerTrainingArguments](../../docs/package\_reference/sentence\_transformer/training\_args.html#sentencetransformertrainingarguments)

\* [Losses](../../docs/package\_reference/sentence\_transformer/losses.html)

\*

[BatchAllTripletLoss](../../docs/package\_reference/sentence\_transformer/losses.html#batchalltripletloss)

\*

[BatchHardSoftMarginTripletLoss](../../docs/package\_reference/sentence\_transformer/losses.html#batchhardsoftmargintripletloss)

\*

[BatchHardTripletLoss](../../docs/package\_reference/sentence\_transformer/losses.html#batchhardtripletloss)

\*

[BatchSemiHardTripletLoss](../../docs/package\_reference/sentence\_transformer/losses.html#batchsemihardtripletloss)

\*

[ContrastiveLoss](../../docs/package\_reference/sentence\_transformer/losses.html#contrastiveloss)

\*

[OnlineContrastiveLoss](../../docs/package\_reference/sentence\_transformer/losses.html#onlinecontrastiveloss)

\*

[ContrastiveTensionLoss](../../docs/package\_reference/sentence\_transformer/losses.html#contrastivetensionloss)

\*

[ContrastiveTensionLossInBatchNegatives](../../docs/package\_reference/sentence\_transformer/losses.html#contrastivetensionlossinbatchnegatives)

\* [CoSENTLoss](../../docs/package\_reference/sentence\_transformer/losses.html#cosentloss)

\* [AngleLoss](../../docs/package\_reference/sentence\_transformer/losses.html#angleloss)

\*

[CosineSimilarityLoss](../../docs/package\_reference/sentence\_transformer/losses.html#cosinesimilarityloss)

\*

[DenoisingAutoEncoderLoss](../../docs/package\_reference/sentence\_transformer/losses.html#denoisingautoencoderloss)

\*

[GISTEmbedLoss](../../docs/package\_reference/sentence\_transformer/losses.html#gistembedloss)

\*

[CachedGISTEmbedLoss](../../docs/package\_reference/sentence\_transformer/losses.html#cachedgistembedloss)

\* [MSELoss](../../docs/package\_reference/sentence\_transformer/losses.html#mseloss)

\*

[MarginMSELoss](../../docs/package\_reference/sentence\_transformer/losses.html#marginmseloss)

\*

[MatryoshkaLoss](../../docs/package\_reference/sentence\_transformer/losses.html#matryoshkaloss)

\*

[Matryoshka2dLoss](../../docs/package\_reference/sentence\_transformer/losses.html#matryoshka2dloss)

\*

[AdaptiveLayerLoss](../../docs/package\_reference/sentence\_transformer/losses.html#adaptivelayerloss)

\*

[MegaBatchMarginLoss](../../docs/package\_reference/sentence\_transformer/losses.html#megabatchmarginloss)

\*

[MultipleNegativesRankingLoss](../../docs/package\_reference/sentence\_transformer/losses.html#multiplenegativesrankingloss)

\*

[CachedMultipleNegativesRankingLoss](../../docs/package\_reference/sentence\_transformer/losses.html#cachedmultiplenegativesrankingloss)

\*

[MultipleNegativesSymmetricRankingLoss](../../docs/package\_reference/sentence\_transformer/losses.html#multiplenegativessymmetricrankingloss)

\*

[CachedMultipleNegativesSymmetricRankingLoss](../../docs/package\_reference/sentence\_transformer/losses.html#cachedmultiplenegativessymmetricrankingloss)

\*

\* [SoftmaxLoss](../../docs/package\_reference/sentence\_transformer/losses.html#softmaxloss)

\* [TripletLoss](../../docs/package\_reference/sentence\_transformer/losses.html#tripletloss)

\* [Samplers](../../docs/package\_reference/sentence\_transformer/sampler.html)

\*

[BatchSamplers](../../../../docs/package\_reference/sentence\_transformer/sampler.html#batchsamplers  
)

\*

[MultiDatasetBatchSamplers](../../../../docs/package\_reference/sentence\_transformer/sampler.html#m  
ultidatasetbatchsamplers)

\* [Evaluation](../../../../docs/package\_reference/sentence\_transformer/evaluation.html)

\*

[BinaryClassificationEvaluator](../../../../docs/package\_reference/sentence\_transformer/evaluation.html  
#binaryclassificationevaluator)

\*

[EmbeddingSimilarityEvaluator](../../../../docs/package\_reference/sentence\_transformer/evaluation.ht  
ml#embeddingsimilarityevaluator)

\*

[InformationRetrievalEvaluator](../../../../docs/package\_reference/sentence\_transformer/evaluation.htm  
l#informationretrievalevaluator)

\*

[NanoBEIREvaluator](../../../../docs/package\_reference/sentence\_transformer/evaluation.html#nanobe  
irevaluator)

\*

[MSEEvaluator](../../../../docs/package\_reference/sentence\_transformer/evaluation.html#mseevaluator  
)

\*

[ParaphraseMiningEvaluator](../../../../docs/package\_reference/sentence\_transformer/evaluation.html#  
paraphraseminingevaluator)

\*

[RerankingEvaluator](../../../../docs/package\_reference/sentence\_transformer/evaluation.html#rerankin  
gevaluator)

\*

[SentenceEvaluator](../../docs/package\_reference/sentence\_transformer/evaluation.html#sentenceevaluator)

\*

[SequentialEvaluator](../../docs/package\_reference/sentence\_transformer/evaluation.html#sequentialevaluator)

\*

[TranslationEvaluator](../../docs/package\_reference/sentence\_transformer/evaluation.html#translationevaluator)

\*

[TripletEvaluator](../../docs/package\_reference/sentence\_transformer/evaluation.html#tripletevaluator)

\* [Datasets](../../docs/package\_reference/sentence\_transformer/datasets.html)

\*

[ParallelSentencesDataset](../../docs/package\_reference/sentence\_transformer/datasets.html#parallelsentencesdataset)

\*

[SentenceLabelDataset](../../docs/package\_reference/sentence\_transformer/datasets.html#sentencelabeldataset)

\*

[DenoisingAutoEncoderDataset](../../docs/package\_reference/sentence\_transformer/datasets.html#denoisingautoencoderdataset)

\*

[NoDuplicatesDataLoader](../../docs/package\_reference/sentence\_transformer/datasets.html#noduplicatesdataloader)

\* [Models](../../docs/package\_reference/sentence\_transformer/models.html)

\*

[Main

Classes](../../docs/package\_reference/sentence\_transformer/models.html#main-classes)

\* [Further

Classes](../../docs/package\_reference/sentence\_transformer/models.html#further-classes)

\* [quantization](../../docs/package\_reference/sentence\_transformer/quantization.html)

\*

[`quantize\_embeddings()`](../../docs/package\_reference/sentence\_transformer/quantization.html#sentence\_transformers.quantization.quantize\_embeddings)

\*

[`semantic\_search\_faiss()`](../../docs/package\_reference/sentence\_transformer/quantization.html#sentence\_transformers.quantization.semantic\_search\_faiss)

\*

[`semantic\_search\_usearch()`](../../docs/package\_reference/sentence\_transformer/quantization.html#sentence\_transformers.quantization.semantic\_search\_usearch)

\* [Cross Encoder](../../docs/package\_reference/cross\_encoder/index.html)

\* [CrossEncoder](../../docs/package\_reference/cross\_encoder/cross\_encoder.html)

\* [CrossEncoder](../../docs/package\_reference/cross\_encoder/cross\_encoder.html#id1)

\* [Training

Inputs](../../docs/package\_reference/cross\_encoder/cross\_encoder.html#training-inputs)

\* [Evaluation](../../docs/package\_reference/cross\_encoder/evaluation.html)

\*

[CEBinaryAccuracyEvaluator](../../docs/package\_reference/cross\_encoder/evaluation.html#cebinaryaccuracyevaluator)

\*

[CEBinaryClassificationEvaluator](../../docs/package\_reference/cross\_encoder/evaluation.html#cebinaryclassificationevaluator)

\*

[CECorrelationEvaluator](../../docs/package\_reference/cross\_encoder/evaluation.html#cecorrelationevaluator)



nevaluator)

\* [CEF1Evaluator](../../docs/package\_reference/cross\_encoder/evaluation.html#cef1evaluator)

\*

[CESoftmaxAccuracyEvaluator](../../docs/package\_reference/cross\_encoder/evaluation.html#cesoftmaxaccuracyevaluator)

\*

[CERerankingEvaluator](../../docs/package\_reference/cross\_encoder/evaluation.html#cererankingevaluator)

\* [util](../../docs/package\_reference/util.html)

\* [Helper Functions](../../docs/package\_reference/util.html#module-sentence\_transformers.util)

\*

[`community\_detection()`](../../docs/package\_reference/util.html#sentence\_transformers.util.community\_detection)

\* [`http\_get()`](../../docs/package\_reference/util.html#sentence\_transformers.util.http\_get)

\*

[`is\_training\_available()`](../../docs/package\_reference/util.html#sentence\_transformers.util.is\_training\_available)

\*

[`mine\_hard\_negatives()`](../../docs/package\_reference/util.html#sentence\_transformers.util.mine\_hard\_negatives)

\*

[`normalize\_embeddings()`](../../docs/package\_reference/util.html#sentence\_transformers.util.normalize\_embeddings)

\*

[`paraphrase\_mining()`](../../docs/package\_reference/util.html#sentence\_transformers.util.paraphrase\_mining)

\*

[`semantic\_search()`](../../docs/package\_reference/util.html#sentence\_transformers.util.semantic\_search)

\*

[`truncate\_embeddings()`](../../docs/package\_reference/util.html#sentence\_transformers.util.truncate\_embeddings)

\* [Model Optimization](../../docs/package\_reference/util.html#module-sentence\_transformers.backend)

\*

[`export\_dynamic\_quantized\_onnx\_model()`](../../docs/package\_reference/util.html#sentence\_transformers.backend.export\_dynamic\_quantized\_onnx\_model)

\*

[`export\_optimized\_onnx\_model()`](../../docs/package\_reference/util.html#sentence\_transformers.backend.export\_optimized\_onnx\_model)

\*

[`export\_static\_quantized\_openvino\_model()`](../../docs/package\_reference/util.html#sentence\_transformers.backend.export\_static\_quantized\_openvino\_model)

\*

\* [Similarity Metrics](../../docs/package\_reference/util.html#module-sentence\_transformers.util)

\*

\* [`cos\_sim()`](../../docs/package\_reference/util.html#sentence\_transformers.util.cos\_sim)

\*

\* [`dot\_score()`](../../docs/package\_reference/util.html#sentence\_transformers.util.dot\_score)

\*

[`euclidean\_sim()`](../../docs/package\_reference/util.html#sentence\_transformers.util.euclidean\_sim)

\*

[`manhattan\_sim()`](../../docs/package\_reference/util.html#sentence\_transformers.util.manhattan\_sim)

\*

[`pairwise\_cos\_sim()`](../../docs/package\_reference/util.html#sentence\_transformers.util.pairwise\_

cos\_sim)

\*

[ pairwise\_dot\_score() ](../../docs/package\_reference/util.html#sentence\_transformers.util.pairwise\_dot\_score)

\*

[ pairwise\_euclidean\_sim() ](../../docs/package\_reference/util.html#sentence\_transformers.util.pairwise\_euclidean\_sim)

\*

[ pairwise\_manhattan\_sim() ](../../docs/package\_reference/util.html#sentence\_transformers.util.pairwise\_manhattan\_sim)

\_\_[Sentence Transformers](../../index.html)

\* [(../../index.html)

\* [Usage](../../docs/sentence\_transformer/usage/usage.html)

\* Semantic Search

\*

[

Edit

on

GitHub](https://github.com/UKPLab/sentence-transformers/blob/master/examples/applications/semantic-search/README.md)

\* \* \*

# Semantic Search

Semantic search seeks to improve search accuracy by understanding the semantic meaning of the search query and the corpus to search over. Semantic search can also perform well given synonyms, abbreviations, and misspellings, unlike

keyword search engines that can only find documents based on lexical matches.

## ## Background

The idea behind semantic search is to embed all entries in your corpus, whether they be sentences, paragraphs, or documents, into a vector space. At search time, the query is embedded into the same vector space and the closest embeddings from your corpus are found. These entries should have a high semantic similarity with the query.

![[SemanticSearch]](<https://raw.githubusercontent.com/UKPLab/sentence-transformers/master/docs/img/SemanticSearch.png>)

## ## Symmetric vs. Asymmetric Semantic Search

A **critical distinction** for your setup is `_symmetric_` vs. `_asymmetric semantic search_`:

- \* For **symmetric semantic search** your query and the entries in your corpus are of about the same length and have the same amount of content. An example would be searching for similar questions: Your query could for example be `_âœ“How to learn Python online?âœ•_` and you want to find an entry like `_âœ“How to learn Python on the web?âœ•_`. For symmetric tasks, you could potentially flip the query and the entries in your corpus.

- \* Related training example: [Quora Duplicate Questions](../training/quora\_duplicate\_questions/README.html).

\* Suitable models: [Pre-Trained Sentence Embedding

Models](../../docs/sentence\_transformer/pretrained\_models.html)

\* For **asymmetric semantic search**, you usually have a **short query** (like a question or some keywords) and you want to find a longer paragraph answering the query. An example would be a query like `“What is Python?”` and you want to find the paragraph `“Python is an interpreted, high-level and general-purpose programming language. Python’s design philosophy is…”`. For asymmetric tasks, flipping the query and the entries in your corpus usually does not make sense.

\* Related training example: [MS MARCO](../../training/ms\_marco/README.html)

\* Suitable models: [Pre-Trained MS MARCO

Models](../../docs/pretrained-models/msmarco-v5.html)

It is critical **that you choose the right model** for your type of task.

### ## Manual Implementation

For small corpora (up to about 1 million entries), we can perform semantic search with a manual implementation by computing the embeddings for the corpus as well as for our query, and then calculating the [semantic textual similarity](../../docs/sentence\_transformer/usage/semantic\_textual\_similarity.html) using

```
[SentenceTransformer.similarity](../../docs/package_reference/sentence_transformer/SentenceTransformer.html#sentence_transformers.SentenceTransformer.similarity  
"sentence_transformers.SentenceTransformer.similarity").
```

For a simple example, see

[semantic\_search.py](https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/semantic\_search.py):

## Output

Query: A man is eating pasta.

Top 5 most similar sentences in corpus:

A man is eating food. (Score: 0.7035)

A man is eating a piece of bread. (Score: 0.5272)

A man is riding a horse. (Score: 0.1889)

A man is riding a white horse on an enclosed ground. (Score: 0.1047)

A cheetah is running behind its prey. (Score: 0.0980)

Query: Someone in a gorilla costume is playing a set of drums.

Top 5 most similar sentences in corpus:

A monkey is playing drums. (Score: 0.6433)

A woman is playing violin. (Score: 0.2564)

A man is riding a horse. (Score: 0.1389)

A man is riding a white horse on an enclosed ground. (Score: 0.1191)

A cheetah is running behind its prey. (Score: 0.1080)

Query: A cheetah chases prey on across a field.

Top 5 most similar sentences in corpus:

A cheetah is running behind its prey. (Score: 0.8253)

A man is eating food. (Score: 0.1399)

A monkey is playing drums. (Score: 0.1292)

A man is riding a white horse on an enclosed ground. (Score: 0.1097)

A man is riding a horse. (Score: 0.0650)

"""

This is a simple application for sentence embeddings: semantic search

We have a corpus with various sentences. Then, for a given query sentence,  
we want to find the most similar sentence in this corpus.

This script outputs for various queries the top 5 most similar sentences in the corpus.

"""

```
import torch
```

```
from sentence_transformers import SentenceTransformer
```

```
embedder = SentenceTransformer("all-MiniLM-L6-v2")
```

```
# Corpus with example sentences
```

```
corpus = [
```

```
    "A man is eating food.",
```

```

"A man is eating a piece of bread.",
"The girl is carrying a baby.",
"A man is riding a horse.",
"A woman is playing violin.",
"Two men pushed carts through the woods.",
"A man is riding a white horse on an enclosed ground.",
"A monkey is playing drums.",
"A cheetah is running behind its prey.",
]

# Use "convert_to_tensor=True" to keep the tensors on GPU (if available)
corpus_embeddings = embedder.encode(corpus, convert_to_tensor=True)

# Query sentences:
queries = [
    "A man is eating pasta.",
    "Someone in a gorilla costume is playing a set of drums.",
    "A cheetah chases prey on across a field.",
]

# Find the closest 5 sentences of the corpus for each query sentence based on cosine similarity
top_k = min(5, len(corpus))

for query in queries:
    query_embedding = embedder.encode(query, convert_to_tensor=True)

    # We use cosine-similarity and torch.topk to find the highest 5 scores
    similarity_scores = embedder.similarity(query_embedding, corpus_embeddings)[0]
    scores, indices = torch.topk(similarity_scores, k=top_k)

```



```

print("\nQuery:", query)

print("Top 5 most similar sentences in corpus:")

for score, idx in zip(scores, indices):

    print(corpus[idx], f"(Score: {score:.4f})")

"""

# Alternatively, we can also use util.semantic_search to perform cosine similarty + topk

hits = util.semantic_search(query_embedding, corpus_embeddings, top_k=5)

hits = hits[0]    #Get the hits for the first query

for hit in hits:

    print(corpus[hit['corpus_id']], "(Score: {:.4f})".format(hit['score']))

"""

```

## ## Optimized Implementation

Instead of implementing semantic search by yourself, you can use the  
`util.semantic\_search` function.

The function accepts the following parameters:

```

sentence_transformers.util.semantic_search(_query_embeddings: ~torch.Tensor,
corpus_embeddings: ~torch.Tensor, query_chunk_size: int = 100, corpus_chunk_size: int = 500000,
top_k: int = 10, score_function: ~typing.Callable[[_torch.Tensor, ~torch.Tensor], ~torch.Tensor] =
<function cos_sim>) -> list[list[dict[str, int] |

```

float]]][[source]]([https://github.com/UKPLab/sentence-transformers/blob/master/sentence\\_transformers\\util.py#L440-L517](https://github.com/UKPLab/sentence-transformers/blob/master/sentence_transformers\\util.py#L440-L517))if•

This function performs a cosine similarity search between a list of query embeddings and a list of corpus embeddings. It can be used for Information Retrieval / Semantic Search for corpora up to about 1 Million entries.

Parameters:

\* \*\*query\_embeddings\*\* (\_Tensor\_) â€“ A 2 dimensional tensor with the query embeddings.

\* \*\*corpus\_embeddings\*\* (\_Tensor\_) â€“ A 2 dimensional tensor with the corpus embeddings.

\* \*\*query\_chunk\_size\*\* (\_int\_ \_\_, \_\_optional\_\_) â€“ Process 100 queries simultaneously. Increasing that value increases the speed, but requires more memory. Defaults to 100.

\* \*\*corpus\_chunk\_size\*\* (\_int\_ \_\_, \_\_optional\_\_) â€“ Scans the corpus 100k entries at a time. Increasing that value increases the speed, but requires more memory. Defaults to 500000.

\* \*\*top\_k\*\* (\_int\_ \_\_, \_\_optional\_\_) â€“ Retrieve top k matching entries. Defaults to 10.

\* \*\*score\_function\*\* (\_Callable\_ \_[[[\_Tensor\_ \_\_, \_\_Tensor\_ \_]]], \_\_Tensor\_ \_], \_\_optional\_\_) â€“ Function for computing scores. By default, cosine similarity.

Returns:

A list with one entry for each query. Each entry is a list of dictionaries with the keys `corpus_id` and `score`, sorted by decreasing cosine similarity scores.

Return type:

```
List[List[Dict[str, Union[int, float]]]]
```

By default, up to 100 queries are processed in parallel. Further, the corpus is chunked into set of up to 500k entries. You can increase `query_chunk_size` and `corpus_chunk_size`, which leads to increased speed for large corpora, but also increases the memory requirement.

## Speed Optimization

To get the optimal speed for the `util.semantic_search` method, it is advisable to have the `query_embeddings` as well as the `corpus_embeddings` on the same GPU-device. This significantly boost the performance. Further, we can normalize the corpus embeddings so that each corpus embeddings is of length 1. In that case, we can use dot-product for computing scores.

```

corpus_embeddings = corpus_embeddings.to("cuda")

corpus_embeddings = util.normalize_embeddings(corpus_embeddings)

query_embeddings = query_embeddings.to("cuda")

query_embeddings = util.normalize_embeddings(query_embeddings)

hits = util.semantic_search(query_embeddings, corpus_embeddings,
score_function=util.dot_score)

```

## ## Elasticsearch

[Elasticsearch](https://www.elastic.co/elasticsearch/) has the possibility to [index dense vectors](https://www.elastic.co/what-is/vector-search) and to use them for document scoring. We can easily index embedding vectors, store other data alongside our vectors and, most importantly, efficiently retrieve relevant entries using [approximate nearest neighbor search](https://www.elastic.co/blog/introducing-approximate-nearest-neighbor-search-in-elasticsearch-8-0) (HNSW, see also below) on the embeddings.

For further details, see

[semantic\_search\_quora\_elasticsearch.py](https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/semantic\_search\_quora\_elasticsearch.py).

## ## Approximate Nearest Neighbor

Searching a large corpus with millions of embeddings can be time-consuming if exact nearest neighbor search is used (like it is used by `util.semantic_search``).

In that case, Approximate Nearest Neighbor (ANN) can be helpful. Here, the data is partitioned into smaller fractions of similar embeddings. This index can be searched efficiently and the embeddings with the highest similarity (the nearest neighbors) can be retrieved within milliseconds, even if you have millions of vectors. However, the results are not necessarily exact. It is possible that some vectors with high similarity will be missed.

For all ANN methods, there are usually one or more parameters to tune that determine the recall-speed trade-off. If you want the highest speed, you have a high chance of missing hits. If you want high recall, the search speed decreases.

Three popular libraries for approximate nearest neighbor are

[Annoy](<https://github.com/spotify/annoy>),  
[FAISS](<https://github.com/facebookresearch/faiss>), and  
[hnswlib](<https://github.com/nmslib/hnswlib/>).

Examples:

[semantic\_search\_quora\_hnswlib.py]([https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/semantic\\_search\\_quora\\_hnswlib.py](https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/semantic_search_quora_hnswlib.py))

\*

[semantic\_search\_quora\_annoy.py](https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/semantic\_search\_quora\_annoy.py)

\*

[semantic\_search\_quora\_faiss.py](https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/semantic\_search\_quora\_faiss.py)

## ## Retrieve & Re-Rank

For complex semantic search scenarios, a two-stage retrieve & re-rank pipeline is advisable:

![[Information Retrieval]](https://raw.githubusercontent.com/UKPLab/sentence-transformers/master/docs/img/InformationRetrieval.png)

For further details, see [Retrieve & Re-rank](../retrieve\_rerank/README.html).

## ## Examples

We list a handful of common use cases:

### ### Similar Questions Retrieval

[semantic\_search\_quora\_pytorch.py](https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/semantic\_search\_quora\_pytorch.py) [ [Colab

version](https://colab.research.google.com/drive/12cn5Oo0v3HfQQ8Tv6-ukgxXSmT3zl35A?usp=sharing)

] shows an example based on the [Quora duplicate questions](https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs) dataset. The user can enter a question, and the code retrieves the most similar questions from the dataset using `util.semantic\_search`. As model, we use [distilbert-multilingual-nli-stsb-quora-ranking](https://huggingface.co/sentence-transformers/distilbert-multilingual-nli-stsb-quora-ranking), which was trained to identify similar questions and supports 50+ languages. Hence, the user can input the question in any of the 50+ languages. This is a **\*\*symmetric search task\*\***, as the search queries have the same length and content as the questions in the corpus.

### ### Similar Publication Retrieval

[semantic\_search\_publications.py](https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/semantic\_search\_publications.py) [ [Colab version](https://colab.research.google.com/drive/12hfBveGHRsxhPIUMmJYrII2IFU4fOX06?usp=sharing)

] shows an example how to find similar scientific publications. As corpus, we use all publications that have been presented at the EMNLP 2016 - 2018 conferences. As search query, we input the title and abstract of more recent publications and find related publications from our corpus. We use the [SPECTER](https://huggingface.co/sentence-transformers/allenai-specter) model. This is a **\*\*symmetric search task\*\***, as the paper in the corpus consists of title & abstract and we search for title & abstract.

### ### Question & Answer Retrieval

[semantic\_search\_wikipedia\_qa.py](https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/semantic\_search\_wikipedia\_qa.py) [ [Colab Version](https://colab.research.google.com/drive/11GunvCqJuebfeTlgbJWkIMT0xJH6PWF1?usp=sharing)

]: This example uses a model that was trained on the [Natural Questions dataset](https://huggingface.co/datasets/sentence-transformers/natural-questions). It consists of about 100k real Google search queries, together with an annotated passage from Wikipedia that provides the answer. It is an example of an **asymmetric search task**. As corpus, we use the smaller [Simple English Wikipedia](https://simple.wikipedia.org/wiki/Main\_Page) so that it fits easily into memory.

[retrieve\_rerank\_simple\_wikipedia.ipynb](https://github.com/UKPLab/sentence-transformers/tree/master/examples/applications/semantic-search/./retrieve\_rerank/retrieve\_rerank\_simple\_wikipedia.ipynb) [ [Colab Version](https://colab.research.google.com/github/UKPLab/sentence-transformers/blob/master/examples/applications/retrieve\_rerank/retrieve\_rerank\_simple\_wikipedia.ipynb)

]: This script uses the [Retrieve & Re-rank](./retrieve\_rerank/README.html) strategy and is an example for an **asymmetric search task**. We split all Wikipedia articles into paragraphs and encode them with a bi-encoder. If a new query / question is entered, it is encoded by the same bi-encoder and the paragraphs with the highest cosine-similarity are retrieved. Next, the



retrieved candidates are scored by a Cross-Encoder re-ranker and the 5 passages with the highest score from the Cross-Encoder are presented to the user. We use models that were trained on the [MS Marco Passage Reranking](https://github.com/microsoft/MSMARCO-Passage-Ranking/) dataset, a dataset with about 500k real queries from Bing search.

[  
Previous](../docs/sentence\_transformer/usage/semantic\_textual\_similarity.html  
"Semantic Textual Similarity") [Next ](../retrieve\_rerank/README.html  
"Retrieve & Re-Rank")

\* \* \*

(C) Copyright 2025.

Built with [Sphinx](https://www.sphinx-doc.org/) using a  
[theme](https://github.com/readthedocs/sphinx\_rtd\_theme) provided by [Read the  
Docs](https://readthedocs.org).