Glossary

Array concatenation

Extension of array $[a_1, ..., a_n]$ with array $[b_1, ..., b_m]$ to produce a new array $[a_1, ..., a_n, b_1, ..., b_m]$.

Bag-of-words (BoW) Word2Vec model

A neural network model which predicts a center word from the context words in a window of text. See Mikolov's paper.

Boolean mask of an array A

An array M of the same dimensions as A, where elements of M are Boolean values (true/false or, equivalently, 1/0) to indicate whether the corresponding elements of A should be considered or ignored for some operation.

DataFrame

A tabular data structure with (possibly labeled) rows and columns, where columns can be different data types. This is different from a matrix, where rows and columns are numerically indexed (but, typically, do not assume string labels) and all values are of the same type. A Pandas DataFrame uses (possibly multiple) NumPy arrays, which are used to represent matrices.

Diagonal elements

Elements in a matrix that lie along the diagonal set of values going from the top-left to bottom- right direction (in a square matrix). See numpy.diag().

Document-term matrix (DTM)

A matrix where the numeric value (typically, a count or non-negative weight) in the position (i, j) represents the degree of importance of a word in column j in a document in column i.

Dot product

Also known as "inner" product or "sum" product. The sum of the products of row values and corresponding column values for two matrices or vectors. For matrices (shape #, #), the number of columns in the first matrix must match the number of rows in the second matrix. For vectors with only one shape value (shape #, <blank>), the number of elements in each vector must match. See numpy.dot().

FastText model

A word-embedding model introduced by Tomas Mikolov and Facebook in 2016 that can provide a vector for any word that has a subword in the model's training vocabulary. See Gensim's FastText and Facebook's FastText.

Long tail

The rightmost region of a statistical distribution that includes many occurrences far away from the central portion of the distribution.

N-gram

A sequence of contiguous N tokens from some text, which can be characters, subwords, words, sentences, and even documents.

NumPy array

A data structure to store indexed values as vectors, matrices, or higher formats (as tensors, in general.)

Off-diagonal elements of a matrix

All elements of a matrix which do not lie on the diagonal of the matrix.

Orthogonal vectors

Two vectors which result in zero dot product. Same as perpendicular vectors. Geometrically, these are vectors with 90° angle between them; for example, $[0, 1] \cdot [2, 0] = 0 \cdot 2 + 1 \cdot 0 = 0$

Similarity

A merit measure of likeness among two structures, such as vectors or matrices.

Skip-gram Word2Vec model

A neural network model which predicts context words from the center word in a window of text. See Mikolov's paper.

Sparse matrix format

A compressed way of storing sparse matrices in order to save computer storage, memory, and computation. SKLearn stores such sparse matrices in several data structures, including CSR array.

Subword

A sequence of contiguous characters from a larger word.

Sum product

See Dot product.

Term frequency-inverse document frequency (TF-IDF)

A formula that automatically assigns low weights to generic and infrequent words or high weights to frequent terms in specific documents. It is used to define the importance of terms in a document. So, low-importance terms can be considered as stopwords and high-importance terms can be used to identify, compare, and relate documents.

User-defined function (UDF)

A function designed and written by the user using this function, not by the third party (such as an external or internal library).

Vector

A point $x := (x_1, ..., x_d)$ in *d*-dimensional vector space. Geometrically, it is an arrow from the origin $0 := (0_1, ..., 0_d)$ to x. Every vector has its head at 0 and its tail at x.

Vector and matrix operations

Dot product $C := AB = A \cdot B$

A matrix with each element (i, j) is dot product of the i th row of A and jth column of B. The result is a compact way to compute and represent all possible dot products of two matrices. See numpy.matmul().

Element-wise math operations

Operations on two objects (vectors or matrices) of the same size and shape. The result is the same shape as the original inputs. Includes matrix addition, subtraction, and Hadamard product multiplication. See numpy.multiply().

Operations with a scalar

Each vector or matrix element has the

specified math operation performed by the same scalar.

Transpose

Replaces elements a_{ij} with elements a_{ji}. See Matrix types: Transpose.

Matrix types

Dense

A matrix with *mostly* non-zero values. Related: csr_matrix.todense().

Diagonal

A matrix with zero off-diagonal values (i.e., those that are not on diagonal).

Identity

A square matrix with 1s on diagonal and 0s off-diagonal (i.e., everywhere else). This is also a diagonal matrix, by definition. Typically labeled as I or I_n or $I_{n\times n}$ to indicate $n \times n$ identity matrix. It is similar to a scalar 1, because (left or right) multiplication by an identity matrix doesn't change the subject matrix; i.e., $IA = A$ and $BI = B$, whenever such multiplication makes sense. See numpy.identity().

Matrix

A collection of same-sized vectors formed by stacking (vertically or horizontally) that looks like a table or box of numbers.

Sparse

A matrix with *few* non-zero values. Complementary values are typically either zeros or NaNs.

Square

A matrix with the same number of rows and columns.

Symmetric

A matrix that is unchanged if transposed so the elements reflected about diagonal are the same; i.e., $a_{ij} = a_{ji}$. By such definition, a symmetric matrix is always square.

Transpose

A matrix that is flipped about its diagonal; i.e., element a_{ij} takes position (j, i). Double transposition returns the original matrix; i.e., $A'' = A$, where $A' = A^T$ indicates transposition of the matrix $A = [a_{ij}]$.

Zero

A matrix with all zero values. Similarly to a scalar zero number, such a matrix zeroes out the product (whenever the product makes sense), i.e., $0A = 0, B0 = 0$ for any matrices A, B and zero matrix, labeled as 0 herein, whenever such multiplication makes sense. See numpy.zeros().

Vector space

A mathematical structure of the same-dimensional vectors, which respect two operations: a vector addition and multiplication by a scalar.

Word2Vec

A neural network model that associates words with dense vector representations, which capture the syntactic and semantic relationships of words in a corpus across many dimensions. See Gensim's Word2Vec.

Word embedding

Learned vector representation of a word that associates that word to other words with similar meanings, forms, morphology, etc. See Mikolov's paper.