Glossary

Accuracy score

A metric for evaluating clustering quality that is equal to the proportion of correctly matched labels.

Adjusted Rand Index (ARI)

This index rescales the Rand index to be an integral of [0, 1] where zero indicates no pattern and one indicates perfect clustering.

Agglomerative approach

A bottom-up approach to building hierarchical clustering trees. This approach begins with every point in its own cluster, then merges the closest clusters at each iteration until all clusters are combined into one.

BERT (Bidirectional Encoder Representations from Transformers)

A new method for encoding entire sentences or documents as numeric vectors. BERT is a neural network based language model which can be extended for a dozen of key NLP tasks, including question and answering, language translation and the generation of meaningful sentence embeddings.

Centroid

The average vector of a set of vectors. For the scalar values 1, 2, and 3, the centroid is just the average of these numbers. In the vector space, a centroid is an element-wise sum of vectors divided by the number of vectors. It's just the mean vector.

Clustering

An unsupervised method for unlabeled documents where clusters, or groups of similar documents, are identified based on their intra-cluster similarity and inter-cluster dissimilarity.

Dendrograms

Or hierarchical clustering trees. A tree which results from grouping points in the hierarchical clustering algorithm. A dendrogram contains one root at its top and branches down one split at a time. The bottom nodes containing single observations are leaf nodes; the remaining nodes are interior nodes. You can use dendrograms to evaluate the quality of clustering and the appropriate number of clusters.

Divisive approach

A top-down approach to hierarchical clustering. This approach begins with older document vectors in a single cluster and then recursively breaks them up until each vector is in its own cluster.

Edit distance

See Levenshtein distance.

Hamming distance

A stream metric which counts element-wise mismatches in the two strings of equal length. In other words, it is the number of substitutions needed to make two strings equal.

Hierarchical clustering

A clustering algorithm used to group documents based on similarity of their vector representations. The result of hierarchical clustering is a hierarchical clustering tree, or dendrogram, which represents similar documents as the bottom leaf nodes.

Hyperparameters

Parameters controlled by human experts.

Interior nodes

All non-leaf nodes in a dendrogram.

K-means clustering

A popular clustering algorithm, in which the user first must specify *k* as the number of clusters to find. The algorithm then initializes with *k* random points in the original feature space, which eventually converge to *k* centroids representing *k* respective clusters. At each iteration, the algorithm uniquely assigns all points to their closest centroid and then re-estimates *k* centroids from the groups of assigned points.

Leaf nodes

The bottom nodes in a dendrogram, containing single observations.

Levenshtein distance

A distance metric equal to the minimal count of character additions, deletions, and substitutions needed to change one string into another.

> **Example 1**: Changing “Cornell” to” eCornell” takes one edit operation of addition or insertion of the letter “e”.
>
> **Example 2**: Changing “cat” to “dog” takes three substitution edits.

Lexical similarity

Relates entities on the basis of syntax, structure, and content of the text. Lexical similarity is prominently applied in autocomplete algorithms, spell checkers, and spell correctors.

Linkages

Inter-cluster metrics used to measure the distance between clusters. Since a cluster is not a single point, but spread out over space, linkages measure the distance of two points, A and B, in different clusters. Outlined below are several linkage methods, which each have different approaches for selecting points A and B.

Minimum linkage

Or single linkage. Uses points A and B that are the closest together while still remaining in separate clusters. As a side effect, this tends to combine close intermediate points, which makes thresholding of distinct clusters difficult.

Maximum linkage

Or complete linkage. Uses points A and B which are furthest from one another. This is opposite to the minimum linkage approach.The maximum linkage approach tends to produce many small clusters, which have observations that are similar to observations in other clusters. It is not ideal for a stable clustering method.

Centroid linkage

Computes points A and B as the cluster centers. This linkage can result in dendrogram inversions, where the edges of the dendrogram would cross each other. This complicates the interpretation and threshold.

Group average linkage

This approach uses the average distance among all pair combinations of points into clusters, and is a compromise between single and complete linkages.

Ward's linkage

This approach is used to minimize the inter-cluster sum of squared distances of points to their centroids. `scikit-learn` uses Ward's linkage as its default, and it works fairly well in general.

Medoid

The central representative point of a cluster, which must be one of the cluster points. To compute the medoid, first compute the centroid and then identify the existing point closest to the centroid.

Principal component analysis (PCA)

An unsupervised learning algorithm that uses singular value decomposition (SVD) to reduce vector dimensionality. In this course, you use it to compress 768-dimensional vectors to two-dimensional vectors.

Rand index

A metric for evaluating clustering quality that counts all pairs of points assigned to the correct clusters, as well as those assigned to the incorrect clusters. However, this metric relies on true labels. which are not always accessible. In that case, human experts can look at the symbols of paired points.

SBERT (Sentence BERT)

An extension of BERT, which is easy to use and is only a few hundred megabytes compared to the eight gigabyte FastText model. Sentence BERT does not store static word vectors. Instead, each word vector, if needed, is generated dynamically from its semantic context.

Semantic similarity

Relates entities on the basis of semantic meaning, census and context of the text. To compute semantic similarity, represent words, phrases, and sentences as numerical vectors and compute their dot products, Euclidean distances, and cosine similarities.

Silhouette score

A metric that measures the quality of clustering of documents across topics. It is calculated as the mean intra cluster distance divided by the mean inter cluster distance from documents to their nearest neighboring topics.