package com.twitter.simclusters\_v2.common.clustering

/\*\*

\* Groups entities by a single embedding dimension with the largest score.

\*/

class LargestDimensionClusteringMethod extends ClusteringMethod {

/\*\*

\* @param embeddings map of entity IDs and corresponding embeddings

\* @param similarityFn function that outputs discrete value (0.0 or 1.0).

\* 1.0 if the dimensions of the highest score (weight) from two given embeddings match.

\* 0.0 otherwise.

\* e.g.

\* case 1: E1=[0.0, 0.1, 0.6, 0.2], E2=[0.1, 0.3, 0.8, 0.0]. similarityFn(E1, E2)=1.0

\* case 2: E1=[0.0, 0.1, 0.6, 0.2], E2=[0.1, 0.4, 0.2, 0.0]. similarityFn(E1, E2)=0.0

\* @tparam T embedding type. e.g. SimClustersEmbedding

\*

\* @return A set of sets of entity IDs, each set representing a distinct cluster.

\*/

override def cluster[T](

embeddings: Map[Long, T],

similarityFn: (T, T) => Double,

recordStatCallback: (String, Long) => Unit

): Set[Set[Long]] = {

// rely on clustering by connected component.

// similarityThreshold=0.1 because it's larger than 0.0 (similarityFn returns 0.0 if two embeddings

// don't share the largest dimension.

new ConnectedComponentsClusteringMethod(similarityThreshold = 0.1)

.cluster(embeddings, similarityFn, recordStatCallback)

}

}