package com.twitter.simclusters\_v2.scalding.embedding.common

import com.twitter.simclusters\_v2.thriftscala.\_

import java.net.InetAddress

import java.net.UnknownHostException

object EmbeddingUtil {

type UserId = Long

type ClusterId = Int

type ProducerId = Long

type EmbeddingScore = Double

type SemanticCoreEntityId = Long

type HashtagId = String

type Language = String

implicit val internalIdOrdering: Ordering[InternalId] = Ordering.by {

case InternalId.EntityId(id) => id.toString

case InternalId.Hashtag(strId) => strId

case InternalId.ClusterId(iid) => iid.toString

case InternalId.LocaleEntityId(LocaleEntityId(entityId, lang)) => lang + entityId.toString

}

implicit val embeddingTypeOrdering: Ordering[EmbeddingType] = Ordering.by(\_.getValue)

/\*\*

\* We do not need to group by model version since we are making the

\* This ordering holds the assumption that we would NEVER generate embeddings for two separate

\* SimClusters KnownFor versions under the same dataset.

\*/

implicit val SimClustersEmbeddingIdOrdering: Ordering[SimClustersEmbeddingId] = Ordering.by {

case SimClustersEmbeddingId(embeddingType, \_, internalId) => (embeddingType, internalId)

}

val ModelVersionPathMap: Map[ModelVersion, String] = Map(

ModelVersion.Model20m145kDec11 -> "model\_20m\_145k\_dec11",

ModelVersion.Model20m145kUpdated -> "model\_20m\_145k\_updated",

ModelVersion.Model20m145k2020 -> "model\_20m\_145k\_2020"

)

/\*\*

\* Generates the HDFS output path in order to consolidate the offline embeddings datasets under

\* a common directory pattern.

\* Prepends "/gcs" if the detected data center is qus1.

\*

\* @param isAdhoc Whether the dataset was generated from an adhoc run

\* @param isManhattanKeyVal Whether the dataset is written as KeyVal and is intended to be imported to Manhattan

\* @param modelVersion The model version of SimClusters KnownFor that is used to generate the embedding

\* @param pathSuffix Any additional path structure suffixed at the end of the path

\* @return The consolidated HDFS path, for example:

\* /user/cassowary/adhoc/manhattan\_sequence\_files/simclusters\_embeddings/model\_20m\_145k\_updated/...

\*/

def getHdfsPath(

isAdhoc: Boolean,

isManhattanKeyVal: Boolean,

modelVersion: ModelVersion,

pathSuffix: String

): String = {

val adhoc = if (isAdhoc) "adhoc/" else ""

val user = System.getenv("USER")

val gcs: String =

try {

InetAddress.getAllByName("metadata.google.internal") // throws Exception if not in GCP.

"/gcs"

} catch {

case \_: UnknownHostException => ""

}

val datasetType = if (isManhattanKeyVal) "manhattan\_sequence\_files" else "processed"

val path = s"/user/$user/$adhoc$datasetType/simclusters\_embeddings"

s"$gcs${path}\_${ModelVersionPathMap(modelVersion)}\_$pathSuffix"

}

def favScoreExtractor(u: UserToInterestedInClusterScores): (Double, ScoreType.ScoreType) = {

(u.favScoreClusterNormalizedOnly.getOrElse(0.0), ScoreType.FavScore)

}

def followScoreExtractor(u: UserToInterestedInClusterScores): (Double, ScoreType.ScoreType) = {

(u.followScoreClusterNormalizedOnly.getOrElse(0.0), ScoreType.FollowScore)

}

def logFavScoreExtractor(u: UserToInterestedInClusterScores): (Double, ScoreType.ScoreType) = {

(u.logFavScoreClusterNormalizedOnly.getOrElse(0.0), ScoreType.LogFavScore)

}

// Define all scores to extract from the SimCluster InterestedIn source

val scoreExtractors: Seq[UserToInterestedInClusterScores => (Double, ScoreType.ScoreType)] =

Seq(

favScoreExtractor,

followScoreExtractor

)

object ScoreType extends Enumeration {

type ScoreType = Value

val FavScore: Value = Value(1)

val FollowScore: Value = Value(2)

val LogFavScore: Value = Value(3)

}

@deprecated("Use 'common/ModelVersions'", "2019-09-04")

final val ModelVersion20M145KDec11: String = "20M\_145K\_dec11"

@deprecated("Use 'common/ModelVersions'", "2019-09-04")

final val ModelVersion20M145KUpdated: String = "20M\_145K\_updated"

@deprecated("Use 'common/ModelVersions'", "2019-09-04")

final val ModelVersionMap: Map[String, ModelVersion] = Map(

ModelVersion20M145KDec11 -> ModelVersion.Model20m145kDec11,

ModelVersion20M145KUpdated -> ModelVersion.Model20m145kUpdated

)

}