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

import com.twitter.dal.client.dataset.{KeyValDALDataset, SnapshotDALDatasetBase}

import com.twitter.scalding.\_

import com.twitter.scalding\_internal.multiformat.format.keyval.KeyVal

import com.twitter.simclusters\_v2.hdfs\_sources.EntityEmbeddingsSources

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

EmbeddingType,

ModelVersion,

SimClustersEmbeddingId,

TfgTopicEmbeddings,

UserToInterestedInClusterScores,

SimClustersEmbedding => ThriftSimClustersEmbedding

}

import com.twitter.wtf.scalding.jobs.common.{AdhocExecutionApp, ScheduledExecutionApp}

/\*\*

\* Jobs to generate Logfav-based Topic-Follow-Graph (TFG) topic embeddings

\* A topic's logfav-based TFG embedding is the sum of its followers' logfav-based InterestedIn

\*/

/\*\*

./bazel bundle src/scala/com/twitter/simclusters\_v2/scalding/embedding/tfg:logfav\_tfg\_topic\_embeddings-adhoc

scalding remote run \

--user cassowary \

--keytab /var/lib/tss/keys/fluffy/keytabs/client/cassowary.keytab \

--principal service\_acoount@TWITTER.BIZ \

--cluster bluebird-qus1 \

--main-class com.twitter.simclusters\_v2.scalding.embedding.tfg.LogFavTfgTopicEmbeddingsAdhocApp \

--target src/scala/com/twitter/simclusters\_v2/scalding/embedding/tfg:logfav\_tfg\_topic\_embeddings-adhoc \

--hadoop-properties "scalding.with.reducers.set.explicitly=true mapreduce.job.reduces=4000" \

-- --date 2020-12-08

\*/

object LogFavTfgTopicEmbeddingsAdhocApp

extends TfgBasedTopicEmbeddingsBaseApp

with AdhocExecutionApp {

override val isAdhoc: Boolean = true

override val embeddingType: EmbeddingType = EmbeddingType.LogFavTfgTopic

override val embeddingSource: KeyValDALDataset[

KeyVal[SimClustersEmbeddingId, ThriftSimClustersEmbedding]

] = EntityEmbeddingsSources.LogFavTfgTopicEmbeddingsDataset

override val pathSuffix: String = "logfav\_tfg\_topic\_embedding"

override val modelVersion: ModelVersion = ModelVersion.Model20m145kUpdated

override val parquetDataSource: SnapshotDALDatasetBase[TfgTopicEmbeddings] =

EntityEmbeddingsSources.LogFavTfgTopicEmbeddingsParquetDataset

override def scoreExtractor: UserToInterestedInClusterScores => Double = scores =>

scores.logFavScore.getOrElse(0.0)

}

/\*\*

./bazel bundle src/scala/com/twitter/simclusters\_v2/scalding/embedding/tfg:logfav\_tfg\_topic\_embeddings

capesospy-v2 update --build\_locally --start\_cron logfav\_tfg\_topic\_embeddings src/scala/com/twitter/simclusters\_v2/capesos\_config/atla\_proc3.yaml

\*/

object LogFavTfgTopicEmbeddingsScheduledApp

extends TfgBasedTopicEmbeddingsBaseApp

with ScheduledExecutionApp {

override val isAdhoc: Boolean = false

override val embeddingType: EmbeddingType = EmbeddingType.LogFavTfgTopic

override val embeddingSource: KeyValDALDataset[

KeyVal[SimClustersEmbeddingId, ThriftSimClustersEmbedding]

] = EntityEmbeddingsSources.LogFavTfgTopicEmbeddingsDataset

override val pathSuffix: String = "logfav\_tfg\_topic\_embedding"

override val modelVersion: ModelVersion = ModelVersion.Model20m145kUpdated

override def scoreExtractor: UserToInterestedInClusterScores => Double = scores =>

scores.logFavScore.getOrElse(0.0)

override val parquetDataSource: SnapshotDALDatasetBase[TfgTopicEmbeddings] =

EntityEmbeddingsSources.LogFavTfgTopicEmbeddingsParquetDataset

override val firstTime: RichDate = RichDate("2020-05-25")

override val batchIncrement: Duration = Days(1)

}