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

import com.twitter.dal.client.dataset.KeyValDALDataset

import com.twitter.dal.client.dataset.SnapshotDALDatasetBase

import com.twitter.scalding.\_

import com.twitter.scalding\_internal.dalv2.DAL

import com.twitter.scalding\_internal.dalv2.DALWrite.D

import com.twitter.scalding\_internal.dalv2.DALWrite.WriteExtension

import com.twitter.scalding\_internal.dalv2.remote\_access.AllowCrossClusterSameDC

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

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

import com.twitter.simclusters\_v2.scalding.embedding.common.EmbeddingUtil

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

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

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

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

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

import com.twitter.simclusters\_v2.thriftscala.{SimClustersEmbedding => ThriftSimClustersEmbedding}

import com.twitter.wtf.scalding.jobs.common.AdhocExecutionApp

import com.twitter.wtf.scalding.jobs.common.ScheduledExecutionApp

import java.util.TimeZone

/\*\*

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

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

\*/

/\*\*

./bazel bundle src/scala/com/twitter/simclusters\_v2/scalding/embedding/tfg:fav\_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.FavTfgTopicEmbeddingsAdhocApp \

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

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

-- --date 2020-12-08

\*/

object FavTfgTopicEmbeddingsAdhocApp extends TfgBasedTopicEmbeddingsBaseApp with AdhocExecutionApp {

override val isAdhoc: Boolean = true

override val embeddingType: EmbeddingType = EmbeddingType.FavTfgTopic

override val embeddingSource: KeyValDALDataset[

KeyVal[SimClustersEmbeddingId, ThriftSimClustersEmbedding]

] = EntityEmbeddingsSources.FavTfgTopicEmbeddingsDataset

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

override val modelVersion: ModelVersion = ModelVersion.Model20m145kUpdated

override val parquetDataSource: SnapshotDALDatasetBase[TfgTopicEmbeddings] =

EntityEmbeddingsSources.FavTfgTopicEmbeddingsParquetDataset

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

scores.favScore.getOrElse(0.0)

}

/\*\*

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

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

\*/

object FavTfgTopicEmbeddingsScheduledApp

extends TfgBasedTopicEmbeddingsBaseApp

with ScheduledExecutionApp {

override val isAdhoc: Boolean = false

override val embeddingType: EmbeddingType = EmbeddingType.FavTfgTopic

override val embeddingSource: KeyValDALDataset[

KeyVal[SimClustersEmbeddingId, ThriftSimClustersEmbedding]

] = EntityEmbeddingsSources.FavTfgTopicEmbeddingsDataset

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

override val modelVersion: ModelVersion = ModelVersion.Model20m145kUpdated

override val parquetDataSource: SnapshotDALDatasetBase[TfgTopicEmbeddings] =

EntityEmbeddingsSources.FavTfgTopicEmbeddingsParquetDataset

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

scores.favScore.getOrElse(0.0)

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

override val batchIncrement: Duration = Days(1)

}

/\*\*

./bazel bundle src/scala/com/twitter/simclusters\_v2/scalding/embedding/tfg:fav\_tfg\_topic\_embeddings\_2020-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.FavTfgTopicEmbeddings2020AdhocApp \

--target src/scala/com/twitter/simclusters\_v2/scalding/embedding/tfg:fav\_tfg\_topic\_embeddings\_2020-adhoc \

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

-- --date 2020-12-08

\*/

object FavTfgTopicEmbeddings2020AdhocApp

extends TfgBasedTopicEmbeddingsBaseApp

with AdhocExecutionApp {

override val isAdhoc: Boolean = true

override val embeddingType: EmbeddingType = EmbeddingType.FavTfgTopic

override val embeddingSource: KeyValDALDataset[

KeyVal[SimClustersEmbeddingId, ThriftSimClustersEmbedding]

] = EntityEmbeddingsSources.FavTfgTopicEmbeddings2020Dataset

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

override val modelVersion: ModelVersion = ModelVersion.Model20m145k2020

override val parquetDataSource: SnapshotDALDatasetBase[TfgTopicEmbeddings] =

EntityEmbeddingsSources.FavTfgTopicEmbeddings2020ParquetDataset

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

scores.favScore.getOrElse(0.0)

}

/\*\*

./bazel bundle src/scala/com/twitter/simclusters\_v2/scalding/embedding/tfg:fav\_tfg\_topic\_embeddings\_2020

capesospy-v2 update --build\_locally --start\_cron fav\_tfg\_topic\_embeddings\_2020 src/scala/com/twitter/simclusters\_v2/capesos\_config/atla\_proc3.yaml

\*/

object FavTfgTopicEmbeddings2020ScheduledApp

extends TfgBasedTopicEmbeddingsBaseApp

with ScheduledExecutionApp {

override val isAdhoc: Boolean = false

override val embeddingType: EmbeddingType = EmbeddingType.FavTfgTopic

override val embeddingSource: KeyValDALDataset[

KeyVal[SimClustersEmbeddingId, ThriftSimClustersEmbedding]

] = EntityEmbeddingsSources.FavTfgTopicEmbeddings2020Dataset

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

override val modelVersion: ModelVersion = ModelVersion.Model20m145k2020

override val parquetDataSource: SnapshotDALDatasetBase[TfgTopicEmbeddings] =

EntityEmbeddingsSources.FavTfgTopicEmbeddings2020ParquetDataset

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

scores.favScore.getOrElse(0.0)

override val firstTime: RichDate = RichDate("2021-03-10")

override val batchIncrement: Duration = Days(1)

}

/\*\*

./bazel bundle src/scala/com/twitter/simclusters\_v2/scalding/embedding/tfg:fav\_tfg\_topic\_embeddings\_2020\_copy

scalding scalding remote run --target src/scala/com/twitter/simclusters\_v2/scalding/embedding/tfg:fav\_tfg\_topic\_embeddings\_2020\_copy

\*/

/\*\*

\* This is a copy job where we copy the previous version of TFG and write to a new one.

\* The dependent dataset for TFG has been deleted.

\* Instead of restarting the entire job, we create this temp hacky solution to keep TFG dataset alive until we deprecate topics.

\* Having a table TFG doesn't lead to a big quality concern b/c TFG is built from topic follows, which is relative stable

\* and we don't have new topics anymore.

\*/

object FavTfgTopicEmbeddings2020CopyScheduledApp extends ScheduledExecutionApp {

val isAdhoc: Boolean = false

val embeddingType: EmbeddingType = EmbeddingType.FavTfgTopic

val embeddingSource: KeyValDALDataset[

KeyVal[SimClustersEmbeddingId, ThriftSimClustersEmbedding]

] = EntityEmbeddingsSources.FavTfgTopicEmbeddings2020Dataset

val pathSuffix: String = "fav\_tfg\_topic\_embedding"

val modelVersion: ModelVersion = ModelVersion.Model20m145k2020

override val firstTime: RichDate = RichDate("2023-01-20")

override val batchIncrement: Duration = Days(3)

def runOnDateRange(

args: Args

)(

implicit dateRange: DateRange,

timeZone: TimeZone,

uniqueID: UniqueID

): Execution[Unit] = {

DAL

.readMostRecentSnapshotNoOlderThan(

EntityEmbeddingsSources.FavTfgTopicEmbeddings2020Dataset,

Days(21))

.withRemoteReadPolicy(AllowCrossClusterSameDC)

.toTypedPipe

.writeDALVersionedKeyValExecution(

EntityEmbeddingsSources.FavTfgTopicEmbeddings2020Dataset,

D.Suffix(

EmbeddingUtil

.getHdfsPath(isAdhoc = isAdhoc, isManhattanKeyVal = true, modelVersion, pathSuffix))

)

}

}