scala\_library(

sources = [

"\*.scala",

],

platform = "java8",

tags = [

"bazel-compatible",

"bazel-compatible:migrated",

"bazel-only",

],

dependencies = [

"3rdparty/jvm/com/twitter/algebird:core",

"3rdparty/jvm/com/twitter/algebird:util",

"3rdparty/jvm/com/twitter/storehaus:algebra",

"3rdparty/jvm/com/twitter/storehaus:core",

"3rdparty/src/jvm/com/twitter/scalding:args",

"3rdparty/src/jvm/com/twitter/scalding:commons",

"3rdparty/src/jvm/com/twitter/scalding:core",

"3rdparty/src/jvm/com/twitter/scalding:date",

"3rdparty/src/jvm/com/twitter/scalding:db",

"3rdparty/src/jvm/com/twitter/scalding:parquet",

"ann/src/main/scala/com/twitter/ann/hnsw",

"ann/src/main/scala/com/twitter/ann/scalding/offline",

"ann/src/main/scala/com/twitter/ann/util",

"geoduck/hadoop/scalding/datasets:userlocation-scala",

"iesource/common/src/main/scala/com/twitter/iesource/common/util",

"iesource/processing/events/src/main/scala/com/twitter/iesource/processing/events/batch" +

":server\_engagements-scala",

"iesource/thrift",

"src/java/com/twitter/ml/api/constant",

"src/scala/com/twitter/ml/api/util",

"src/scala/com/twitter/ml/featurestore/catalog/entities/core",

"src/scala/com/twitter/ml/featurestore/catalog/features/geo",

"src/scala/com/twitter/ml/featurestore/lib/batch",

"src/scala/com/twitter/scalding\_internal/dalv2",

"src/scala/com/twitter/scalding\_internal/dalv2/dataset",

"src/scala/com/twitter/scalding\_internal/db",

"src/scala/com/twitter/scalding\_internal/db/jdbc",

"src/scala/com/twitter/scalding\_internal/error\_handling",

"src/scala/com/twitter/scalding\_internal/job",

"src/scala/com/twitter/scalding\_internal/job/analytics\_batch",

"src/scala/com/twitter/scalding\_internal/multiformat",

"src/scala/com/twitter/scalding\_internal/source",

"src/scala/com/twitter/scalding\_internal/source/lzo\_scrooge",

"src/scala/com/twitter/scalding\_internal/typed",

"src/scala/com/twitter/simclusters\_v2/hdfs\_sources",

"src/scala/com/twitter/simclusters\_v2/scalding/common",

"src/thrift/com/twitter/ml/api:data-java",

"src/thrift/com/twitter/ml/api:interpretable-model-java",

"tweetsource/public\_tweets/src/main/scala/com/twitter/tweetsource/public\_tweets:public\_tweets-scala",

"twml/runtime/src/main/scala/com/twitter/twml/runtime/scalding",

"util/util-core:scala",

"util/util-stats/src/main/scala",

],

)

scalding\_job(

name = "tweet-embedding-generation-adhoc-job",

main = "com.twitter.simclusters\_v2.scalding.mbcg.TweetEmbeddingGenerationAdhocJob",

args = [

"--dateRange 2021-10-30T00 2021-10-30T01",

"--model\_name model",

"--model\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/models/tfx\_model\_1104/1635973177/tweet\_tower\_with\_signature",

"--concurrency\_level 60",

"--embedding\_dimension 128",

"--expected\_elements 30000000",

"--max\_M 20",

"--ef\_construction 200",

"--tweet\_embedding\_name output",

"--ann\_output\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/ann\_index/test\_11\_04\_adhoc",

],

config = [

("hadoop.submitter.cpu", 60),

("hadoop.submitter.jvm.total-memory", "256g"),

("submitter.tier", "preemptible"),

("hadoop.map.jvm.total-memory", "6144m"),

],

hadoop\_cluster = "atla-proc3",

platform = "java8",

role = "cassowary",

runtime\_platform = "java8",

tags = [

"bazel-compatible:migrated",

"bazel-only",

],

dependencies = [":mbcg"],

)

scalding\_job(

name = "tweet-embedding-generation-batch-job",

main = "com.twitter.simclusters\_v2.scalding.mbcg.TweetEmbeddingGenerationBatchJob",

args = [

"--model\_name model",

"--model\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/models/tfx\_0119\_1day\_0110\_3l\_5e\_f2v\_gpu\_resave/tweet\_tower\_with\_signature",

"--concurrency\_level 60",

"--embedding\_dimension 128",

"--expected\_elements 5000000",

"--max\_M 40",

"--ef\_construction 800",

"--tweet\_embedding\_name output",

"--f2v\_input.feature\_store\_embedding Follow2VecProducerEmbedding200Dataset",

"--f2v\_input.feature\_store\_major\_version 20210708",

"--minFavCount 32",

"--ann\_output\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/ann\_index/0125\_batch\_index\_f2v\_minfav",

],

config = [

("hadoop.submitter.cpu", 60),

("hadoop.submitter.jvm.total-memory", "256g"),

("hadoop.map.jvm.total-memory", "6144m"),

("hadoop.submitter.disk", "100g"),

],

cron = "\*/5 \* \* \* \*",

hadoop\_cluster = "atla-proc3",

platform = "java8",

role = "cassowary",

runtime\_platform = "java8",

tags = [

"bazel-compatible:migrated",

"bazel-only",

],

dependencies = [":mbcg"],

)

scalding\_job(

name = "tweet-embedding-generation-batch-job-alternate",

main = "com.twitter.simclusters\_v2.scalding.mbcg.TweetEmbeddingGenerationBatchJobAlternate",

args = [

"--model\_name model",

"--model\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/models/tfx\_331\_329\_1e\_128em\_b128\_hn10\_all\_gpu/tweet\_tower\_with\_signature",

"--concurrency\_level 60",

"--embedding\_dimension 128",

"--expected\_elements 5000000",

"--max\_M 40",

"--ef\_construction 800",

"--tweet\_embedding\_name output",

"--f2v\_input.feature\_store\_embedding Follow2VecProducerEmbedding200Dataset",

"--f2v\_input.feature\_store\_major\_version 20210708",

"--minFavCount 100",

"--indexAllTweets",

"--ann\_output\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/ann\_index/0401\_batch\_index\_f2v\_cosine\_all\_tweets",

],

config = [

("hadoop.submitter.cpu", 60),

("hadoop.submitter.jvm.total-memory", "256g"),

("hadoop.map.jvm.total-memory", "6144m"),

("hadoop.submitter.disk", "100g"),

],

contact = "no-reply@twitter.com",

cron = "\*/5 \* \* \* \*",

hadoop\_cluster = "atla-proc3",

platform = "java8",

role = "cassowary",

runtime\_platform = "java8",

tags = [

"bazel-compatible:migrated",

"bazel-only",

],

dependencies = [":mbcg"],

)

scalding\_job(

name = "tweet-embedding-generation-batch-job-experimental",

main = "com.twitter.simclusters\_v2.scalding.mbcg.TweetEmbeddingGenerationBatchJobExperimental",

args = [

"--model\_name model",

"--model\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/models/tfx\_0127\_1day\_0110\_3l\_10e\_128e\_normf2v\_nocosine\_gpu/tweet\_tower\_with\_signature",

"--concurrency\_level 60",

"--embedding\_dimension 128",

"--expected\_elements 5000000",

"--max\_M 40",

"--ef\_construction 800",

"--tweet\_embedding\_name output",

"--f2v\_input.feature\_store\_embedding Follow2VecProducerEmbedding200Dataset",

"--f2v\_input.feature\_store\_major\_version 20210708",

"--minFavCount 32",

"--ann\_output\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/ann\_index/0128\_f2v\_1week\_batch\_index",

],

config = [

("hadoop.submitter.cpu", 60),

("hadoop.submitter.jvm.total-memory", "256g"),

("hadoop.map.jvm.total-memory", "6144m"),

("hadoop.submitter.disk", "100g"),

],

contact = "no-reply@twitter.com",

cron = "\*/5 \* \* \* \*",

hadoop\_cluster = "atla-proc3",

platform = "java8",

role = "cassowary",

runtime\_platform = "java8",

tags = [

"bazel-compatible:migrated",

"bazel-only",

],

dependencies = [":mbcg"],

)

scalding\_job(

name = "user-embedding-generation-adhoc-job",

main = "com.twitter.simclusters\_v2.scalding.mbcg.UserEmbeddingGenerationAdhocJob",

args = [

"--dateRange 2021-12-01T00 2021-12-01T01",

"--model\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/models/tfx\_1202\_logs\_100m\_b64\_hn10\_1127\_video\_persistent/user\_tower\_with\_signature",

"--embedding\_dimension 128",

"--user\_embedding\_name output",

"--kvs\_output\_path /user/cassowary/explore\_mbcg/user\_kvs\_store/1207\_adhoc\_model\_store",

],

config = [

("hadoop.submitter.cpu", 60),

("hadoop.submitter.jvm.total-memory", "256g"),

("submitter.tier", "preemptible"),

("hadoop.map.jvm.total-memory", "6144m"),

],

contact = "no-reply@twitter.com",

hadoop\_cluster = "atla-proc3",

platform = "java8",

role = "cassowary",

runtime\_platform = "java8",

tags = [

"bazel-compatible:migrated",

"bazel-only",

"known-to-fail-jira:SD-20253",

],

dependencies = [":mbcg"],

)

scalding\_job(

name = "user-embedding-generation-batch-job",

main = "com.twitter.simclusters\_v2.scalding.mbcg.UserEmbeddingGenerationBatchJob",

args = [

"--model\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/models/tfx\_0119\_1day\_0110\_3l\_5e\_f2v\_gpu\_resave/user\_tower\_with\_signature",

"--embedding\_dimension 128",

"--user\_embedding\_name output",

"--f2v\_input.feature\_store\_embedding FollowBasedConsumerFollow2VecAvgEmbedding200Dataset",

"--f2v\_input.feature\_store\_major\_version 20210708",

"--kvs\_output\_path /user/cassowary/explore\_mbcg/user\_kvs\_store/0125\_refreshed\_model\_store\_f2v",

],

config = [

("hadoop.submitter.cpu", 60),

("hadoop.submitter.jvm.total-memory", "256g"),

("submitter.tier", "preemptible"),

("hadoop.map.jvm.total-memory", "6144m"),

],

contact = "no-reply@twitter.com",

cron = "\*/30 \* \* \* \*",

hadoop\_cluster = "atla-proc3",

platform = "java8",

role = "cassowary",

runtime\_platform = "java8",

tags = [

"bazel-compatible:migrated",

"bazel-only",

],

dependencies = [":mbcg"],

)

scalding\_job(

name = "user-embedding-generation-batch-job-alternate",

main = "com.twitter.simclusters\_v2.scalding.mbcg.UserEmbeddingGenerationBatchJobAlternate",

args = [

"--model\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/models/tfx\_331\_329\_1e\_128em\_b128\_hn10\_all\_gpu/user\_tower\_with\_signature",

"--embedding\_dimension 128",

"--user\_embedding\_name output",

"--f2v\_input.feature\_store\_embedding FollowBasedConsumerFollow2VecAvgEmbedding200Dataset",

"--f2v\_input.feature\_store\_major\_version 20210708",

"--kvs\_output\_path /user/cassowary/explore\_mbcg/user\_kvs\_store/0401\_refreshed\_model\_store\_all",

],

config = [

("hadoop.submitter.cpu", 60),

("hadoop.submitter.jvm.total-memory", "256g"),

("submitter.tier", "preemptible"),

("hadoop.map.jvm.total-memory", "6144m"),

],

contact = "no-reply@twitter.com",

cron = "\*/30 \* \* \* \*",

hadoop\_cluster = "atla-proc3",

platform = "java8",

role = "cassowary",

runtime\_platform = "java8",

tags = [

"bazel-compatible:migrated",

"bazel-only",

],

dependencies = [":mbcg"],

)

scalding\_job(

name = "user-embedding-generation-batch-job-experimental",

main = "com.twitter.simclusters\_v2.scalding.mbcg.UserEmbeddingGenerationBatchJobExperimental",

args = [

"--model\_path hdfs:///atla/proc/user/cassowary/explore\_mbcg/models/tfx\_0127\_1day\_0110\_3l\_10e\_128e\_normf2v\_nocosine\_gpu/user\_tower\_with\_signature",

"--embedding\_dimension 128",

"--user\_embedding\_name output",

"--f2v\_input.feature\_store\_embedding FollowBasedConsumerFollow2VecAvgEmbedding200Dataset",

"--f2v\_input.feature\_store\_major\_version 20210708",

"--kvs\_output\_path /user/cassowary/explore\_mbcg/user\_kvs\_store/0328\_f2v\_cosine\_all\_tweets\_model\_store",

],

config = [

("hadoop.submitter.cpu", 60),

("hadoop.submitter.jvm.total-memory", "256g"),

("submitter.tier", "preemptible"),

("hadoop.map.jvm.total-memory", "6144m"),

],

contact = "no-reply@twitter.com",

cron = "\*/30 \* \* \* \*",

hadoop\_cluster = "atla-proc3",

platform = "java8",

role = "cassowary",

runtime\_platform = "java8",

tags = [

"bazel-compatible:migrated",

"bazel-only",

],

dependencies = [":mbcg"],

)