Package 'GeneralPretrainedModelTools'

December 19, 2023

Type Package

Title Tools to Support General Pre-Trained Models from CDM Data

Version 0.1.0

Description Tools to help create large general pre-
trained models from data in the OMOP Common Data Model.

Depends DatabaseConnector (>= 6.1.0)

Imports SqlRender (>= 1.13.0),
rlang,
dplyr,
ParallelLogger,
arrow,
Andromeda,
bit64,
checkmate,
readr

Suggests FeatureExtraction,
testthat,
Eunomia

Remotes ohdsi/FeatureExtraction,
ohdsi/Eunomia

License Apache License

RoxygenNote 7.2.3

Roxygen list(markdown = TRUE)

Encoding UTF-8

NeedsCompilation no

R topics documented:

`computeParquetDescriptives` *Compute descriptive statistics for Parquet files*

Description

Computes descriptive statistics for data extracted using the `extractCdmToParquet()` function. It will produce two CSV tables in the `folder`:

- **TableDescriptives.csv** will contain the row count and, where applicable, the person count per table.
- **ConceptDescriptives.csv** will contain, for each concept, the number of occurrences, and the number of persons having that concept.

Usage

```
computeParquetDescriptives(folder)
```

Arguments

`folder`	The folder on the local file system where the Parquet files were written.

Value

Does not return anything. Is called for the side-effect of generating the two CSV files.

`createCdmCovariateSettings` *Create CDM covariate settings*

Description

Create covariate settings for extracting verbatim data from a subset of fields of a subset of tables in the OMOP Common Data Model

Usage

```
createCdmCovariateSettings(
  folder,
  windowStart = -365,
  windowEnd = 0,
  partitions = 10,
  analysisId = 999
)
```

Arguments

folder	The folder on the local file system where the Parquet files will be written.
windowStart	The start of the window relative to the cohort start date where CDM data will be extracted.
windowEnd	The end of the window relative to the cohort start date where CDM data will be extracted.
partitions	The number of partitions to divide the data in.
analysisId	The covariate analysis ID.

Value

An object of type covariateSettings, to be used with the FeatureExtraction package.

---

extractCdmToParquet *Extract data from the database*

Description

Extract data from the server for a random sample of persons, and stores them in the local file system as Parquet files. Has the following features:

- Extracts the subset of CDM tables and fields listed here: https://github.com/OHDSI/GeneralPretrainedModelTools/b
- Can restrict to a sample of person_ids, as specified with the sampleSize argument.
- Loads and saves the tables in as many partitions as the user specifies (see partitions argument). The partitioning is done by person_id (or concept_id for the concept and concept_ancestor table), in a way that the n-th partition of each domain table refers to the same person_ids.
- Restricts the concept table to standard concepts only (ie. those concepts that are allowed to be used in the CDM), to save space.
- Loading can be done with multiple threads (see maxCores argument) for speedup.
- If the process is interrupted for some reason (e.g. the server drops the connection) you can just restart it and it will pick up where it left off. (unless forceRestart = TRUE) .

Usage

```
extractCdmToParquet(
  connectionDetails,
  cdmDatabaseSchema,
  workDatabaseSchema,
  partitionTablePrefix = "GPM_",
  folder,
  sampleSize = 1e+06,
  partitions = 200,
  maxCores = 3,
  forceRestart = FALSE,
  dropPartitionTablesWhenDone = FALSE
)
```

Arguments

`connectionDetails`	An R object of type `connectionDetails` created using the `DatabaseConnector::createConnectio` function.
`cdmDatabaseSchema`	The name of the database schema that contains the OMOP CDM instance. Requires read permissions to this database. On SQL Server, this should specify both the database and the schema, so for example 'cdm_instance.dbo'.
`workDatabaseSchema`	The name of the database schema where work tables can be created.
`partitionTablePrefix`	The prefix to use when creating table names in the `workDatabaseSChema` for storing the person ID and concept ID partition tables.
`folder`	The folder on the local file system where the Parquet files will be written.
`sampleSize`	The number of persons to be included in the sample.
`partitions`	The number of partitions. Fewer partitions may lead to memory issues.
`maxCores`	The maximum number of parallel threads to use.
`forceRestart`	If `FALSE`, when data is already found in the `folder` the process will continue where it left off. If `TRUE`, any existing data files will be deleted, and the process will start from scratch.
`dropPartitionTablesWhenDone`	Drop the partition tables when done? If not, they could be reused for a future data pull.

Index