NAME

RDKitRemoveDuplicateMolecules.py - Remove duplicate molecules

SYNOPSIS

RDKitRemoveDuplicateMolecules.py [--infileParams <Name,Value,...>] [--mode <remove or count>] [--outfileParams <Name,Value,...>] [--overwrite] [--useChirality <yes or no>] [-w <dir>] [-o <outfile>] -i <infile>

RDKitRemoveDuplicateMolecules.py -h | --help | -e | --examples

DESCRIPTION

Identify and remove duplicate molecules based on canonical SMILES strings or simply count the number of duplicate molecules.

The supported input file formats are: SD (.sdf, .sd), SMILES (.smi., csv, .tsv, .txt)

The supported output file formats are: SD (.sdf, .sd), SMILES (.smi)

OPTIONS

-e, --examples

Print examples.

-h, --help

Print this help message.

-i, --infile <infile>

Input file name.

--infileParams <Name,Value,...> [default: auto]

A comma delimited list of parameter name and value pairs for reading molecules from files. The supported parameter names for different file formats, along with their default values, are shown below:

```
SD: removeHydrogens,yes,sanitize,yes,strictParsing,yes
SMILES: smilesColumn,1,smilesNameColumn,2,smilesDelimiter,space,
    smilesTitleLine,auto,sanitize,yes
```

Possible values for smilesDelimiter: space, comma or tab.

-m, --mode <remove or count> [default: remove]

Specify whether to remove duplicate molecules and write out filtered molecules to output files or or simply count the number of duplicate molecules.

-o, --outfile <outfile>

Output file name.

--outfileParams <Name,Value,...> [default: auto]

A comma delimited list of parameter name and value pairs for writing molecules to files. The supported parameter names for different file formats, along with their default values, are shown below:

```
SD: compute2DCoords,auto,kekulize,no
SMILES: kekulize,no,smilesDelimiter,space, smilesIsomeric,yes,
    smilesTitleLine,yes,smilesMolName,yes,smilesMolProps,no
```

Default value for compute2DCoords: yes for SMILES input file; no for all other file types.

--overwrite

Overwrite existing files.

-u, --useChirality <yes or no> [default: yes]

Use stereochemistry information for generation of canonical SMILES strings to identify duplicate molecules.

-w, --workingdir <dir>

Location of working directory which defaults to the current directory.

EXAMPLES

To remove duplicate molecules and generate output files containing unique and duplicate SMILES strings, type:

```
% RDKitRemoveDuplicateMolecules.py -i Sample.smi -o SampleOut.smi
```

To remove duplicate molecules without using stereochemistry information for generation of canonical SMILES and generate output files containing unique and duplicate SMILES strings, type:

```
% RDKitRemoveDuplicateMolecules.py -u no -i Sample.sdf -o SampleOut.sdf
```

To count number of unique and duplicate molecules without generating any output files, type:

```
% RDKitRemoveDuplicateMolecules.py -m count -i Sample.sdf
```

To remove duplicate molecules from a CSV SMILES file, SMILES strings in column 1, name in column 2, and generate output SD files containing unique and duplicate molecules, type:

```
% RDKitRemoveDuplicateMolecules.py --infileParams
  "smilesDelimiter,comma,smilesTitleLine,yes,smilesColumn,1,
  smilesNameColumn,2" --outfileParams "compute2DCoords,yes"
  -i SampleSMILES.csv -o SampleOut.sdf
```

AUTHOR

Manish Sud(msud@san.rr.com)

SEE ALSO

RDKitConvertFileFormat.py, RDKitRemoveInvalidMolecules.py, RDKitRemoveSalts, RDKitSearchFunctionalGroups.py, RDKitSearchSMARTS.py

COPYRIGHT

The functionality available in this script is implemented using RDKit, an open source toolkit for cheminformatics developed by Greg Landrum.

This file is part of MayaChemTools.

MayaChemTools is free software; you can redistribute it and/or modify it under the terms of the GNU Lesser General Public License as published by the Free Software Foundation; either version 3 of the License, or (at your option) any later version.