tsv-utils: 制表符操作通用程序

一、程序介绍

> `tsv-utils` 为一组处理制表符格式文件的程序集合，主要使用`klib` 进行开发，包括`khash, kvect, kstring`等。

二、主程序接口

当前释放版本：`version: 0.0.1-r2`

```
$ tsv-utils
```

```
Usage:   tsv-utils <command> <arguments>
Version: 0.0.1-r2

Command:
      -- Combination
         agg            combinate multi-file.
         join           join tables with primery key.
         tsv2xlsx       convert (multi-)tsv file to Excel file.

      -- Numeric data frame
         rank           rank/merge for numeric table.
         abundance      calculate relative abundance for numeric table.
         norm           normalization with counts map and normalization
                        factor for numeric table.
         stand          standardization for numeric table.
         melt           merge values with bin table file.
         distribution   calculate bins feature distribution.
         trim           trim rows using cutoff.(sum operation).
         nfilter        filter using value with specied collum. op: >= | <=
.

      -- Editing
         annotation     annotating specify collum with key/value(s) db.
         links          transform annotation with links map and definitions.
         associate      associate with links map.
         definition     adding definition collum with key/value(s) db.
         replace        replace specify collum elements with key/value(s)
db.
         reorder        reorder rows by specify key in specify collum.
         subset         retrieve ids in/not in list file [row].
         subcolumn      retrieve ids in/not in list file [collum].
         collapse       collapse '\t' separator to specify delim.
         add_headline   add headline.
         groupline      add groupline.
         placehold      fill empty cell with specify STR.
         reshape        reshape and bin using map file.

      -- Matrix Operation
```

```
         transpose      matrix transpose.
         submatrix      submatrix by id.
         matrix2tab     binary format.
         matrix2melt    elements in submatrix using metadata.

      -- Summary
         cut            print selected parts of lines.
         bins           uniq/bin/summary.
         uniq           unique specify collum and counts.
         nlines         calculate lines of file.
         stats          calculate stats for selected collum.
         unpack         unpack the bins files.

      -- auxiliary utils.
         view             View text file, ignor comments and blank lines.

Licenced:
(c) 2018-2020 - LEI ZHANG
Logic Informatics Co.,Ltd.
zhanglei@logicinformatics.com
```

三、主要子命令功能介绍

主要子命令功能介绍：

合并操作：

1. agg： 对多个制表符文件按照指定列进行合并，可以指定主键（可以组合多列），以及合并的多列；
2. join： 合并多个制表符文件，主键为第一列，其余列合并在一个文件的不同列；
3. tsv2xlsx：转换多个文件到EXCE文件的不同sheet；

数值操作：

4. rank： 对数值型制表符文件的数值所有列进行合并，并抽取不重复的指定前N行；
5. abundance：对数值型制表符文件计算相对丰度；
6. norm： 对数值型制表符文件标准化，将数值标准化一个固定的数值，比如：OTU表，或者 宏基因组丰度数值表； new = (old * factor) / total；
7. stand：对数值型制表符文件标准化，计算标准： z = (x – μ) / σ；
8. melt：根据bin文件以及元素（elements）的数值映射表合并新的数值；
9. distribution：根据bin文件以及多样本元素（elements）的分类以及数值映射表合并新的数值表，新的数值表关联bin文件的主键和elements表中的特征。
10. trim：对数值表进行过滤行操作，操作逻辑: 数值表的行数值加和过滤低于一定阈值;
11. nfilter：对数值表进行过滤行操作，根据指定列的数值大于等于或者小于等于；

编辑操作：

12. annotation： 根据字典：key/values 对指定的文件列进行注释，注释信息在行尾；
13. associate：根据字典：key/value元素，将指定的制表符文件的列进行关联，将key/value的value元素转移至主键， 形成key/values对，对于重复的对于关系形成一对多的关系，比如gene:ko 的注释关系， 可以通过ko:pathway 字典，构建gene:pathway的一对多关系；
14. links： 在associate的基础上添加了对新的value的文字描述；
15. definition：根据字典：key/values 对指定的文件列进行注释，一对多的key/value关系的value进行合并，经将注释信息合并在指定列的后面，可指定分割符；
16. replace： 根据字典：key/value对替换指定列中可以匹配的key的value；
17. reorder： 根据列表的顺序，调整指定文件的顺序；
18. subset： 根据列表中的元素，对指定列进行行操作（子集和补集）；

19. subcolumn： 根据指定的文件或者字符串，提取对应的列，并合并新的文件；
20. collapse： 将指定的多列元素进行合并，可指定合并的元素的分隔符；
21. add_headline： 对制表符添加表头；
22. groupline： 根据key/value对文件添加分组表头；
23. placehold： 将文件中的空元素设定成指定元素；
24. reshape： 根据key/value对，合并属于相同value的字符串；

矩阵操作：

25. transpose： 对矩阵文件转置操作；
26. submatrix： 针对矩阵（方阵）进行抽取子阵；
27. matrix2tab： 将矩阵格式的文件转换成成对的文件（三列）；
28. matrix2melt： 根据ID映射表抽取矩阵中的数值，可用于估计分组中的ID之间的距离，比如Anosim作图；

抽取/统计操作：

29. cut： 根据指定的列数值抽取或者删除对应的列，抽取列按照指定顺序重新组合；
30. bins: 根据指定的2列进行key/value汇总，合并相同key的value，并按照指定分隔符进行合并新字符串。统计元素个数。
31. uniq： 根据指定的列或者多列，统计Unique元素的个数。
32. nlines： 计算文件的行数，可添加输出结果的注释信息；
33. stats： 计算指定数值列的数值统计信息，包括：最大值，最小值，均值，方差等；
34. unpack： bin子命令的反操作，将bin的文件转换成key/value关系，拆分指定分隔符的values数值；

辅助操作

35. view： 辅助操作子命令，可以将double型转换成int类型，去除文件中的空行以及注释行；

Last Update: 8/31/2020 9:09:06 PM