

tsv-utils之对两列进行汇总统计-tsv-utils bins

一、tsv-utils bins介绍

功能描述：

`tsv-utils 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。

命令行接口：

```
1 $ tsv-utils bins
2
3 Usage: tsv-utils bins [options] [table ...]
4 Options:
5   -t INT  target to uniq and counts, default: [2]
6   -s INT  target to summary, default: [1]
7   -d STR  delimiter between elements, default: [,]
```

可选参数：

```
1 -t 整数 目标列取唯一值并统计其对应值的频数，默认为第二列；
2 -s 整数 指定被汇总的列，默认为第一列；
3 -d 指定 被汇总元素之间的间隔符，默认为' , '；
```

二、使用场景实例及其用法

示例演示

示例文件： `kofams.txt`

```
1 $ cat kofams.txt | head -n 6
```

```
1 #seqid ko      threshold      score  eval   domain_coverage
   query_coverage definition      type
2 Y1_g_00002    K03686  415.90  500.5  1.4e-150      0.50      0.98
   molecular chaperone DnaJ
3 Y1_g_00003    K04043  830.13  895.1  7.1e-270      0.79      0.99
   molecular chaperone DnaK
4 Y1_g_00004    K03687  30.90   186.6  1.1e-55 0.43      0.92      molecular
   chaperone GrpE
5 Y1_g_00005    K03705  67.20   310.9  2.3e-93 0.79      0.98      heat-
   inducible transcriptional repressor
6 Y1_g_00006    K02495  263.57  311.1  2.2e-93 0.71      0.97      oxygen-
   independent coproporphyrinogen III oxidase [EC:1.3.98.3]
```

运行命令：统计第二列 `-t 2` 的唯一值与其频数，并将其对应第一列 `-s 1` 的值汇总，设置 `-d` 参数，将汇总元素之间的间隔符设为 `;`

```
1 | $ tsv-utils bins -t 2 -s 1 -d ';' kofams.txt | head -n 7
```

```
1 | #catalog      number  members
2 | K13052  1      Y1_g_01376
3 | K13292  1      Y1_g_00935
4 | K03495  1      Y1_g_01211
5 | K03496  1      Y1_g_01823
6 | K01071  1      Y1_g_01732
7 | K03497  2      Y1_g_01822;Y1_g_01824
```

本文材料为 **BASE (Biostack Applied bioinformatic SEies)** 课程 **Linux Command Line Tools for Life Scientists** 材料，版权归 上海逻捷信息科技有限公司 所有。

Last Update: 8/30/2020 6:48:15 PM