

tsv-utils之列去除重复字符串: uniq

一. tsv-utils uniq

功能描述:

`tsv-utils uniq` 程序为了解决组合列进行 `uniq` 操作的实现, 并可以通过 `-c` 可选项指示是否显示频数, `-f` 选项指定fields 信息

命令行接口:

```
1 $ tsv-utils uniq
2
3 Usage: tsv-utils uniq [options] <text>
4
5 Options:
6   -f STR fields pattern: 1:2:3, [1];
7   -c      display the counts;
8   -i      flag to ignore lines start with '#';
```

可选参数:

```
1 -f 指定 参数格式1:2:3;
2 -c 显示频数;
3 -i 忽略以#开头的行;
```

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:

参数示例1: 统计被注释到的 `ko` 个数(第二列), 忽略一行的注释行.

```
1 | $ tsv-utils uniq -i -f2 kofams.txt | wc -l
```

```
1 | 1118
```

参数示例2: 统计被注释到的 **ko** 个数(第二列), 忽略一行的注释行, 以及显示每个 **ko** 的个数

```
1 | $ tsv-utils uniq -i -c -f2 kofams.txt | head -n 6
```

```
1 | k13052 1
2 | k13292 1
3 | k03495 1
4 | k03496 1
5 | k01071 1
6 | k03497 2
```

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 Upate: 8/30/2020 6:41:01 PM