tsv-utils之空元素设定成指定元素：placehold

一、tsv-utils placehold 介绍

功能描述：

`tsv-utils placehold` 将文件中的空元素设定成指定元素。

命令行接口：

```
$ tsv-utils placehold

Usage: tsv-utils placehold [options] <tsv>

Options:
  -p STR replace emtpy cell with specified STR, default: [-]
```

可选参数：

```
-p  指定  用指定字符替换空元素，默认用 "-" 替换；
```

二、使用场景实例及其用法

经典使用场景：

`emapper` 是使用比较广泛的基因功能注释工具, 其输出格式,针对缺值情况,使用的是空位置, 一些制表符分割接口会合并连续制表符, 给数据提取带来一定麻烦, 使用`-` 取代空位置是一个比较好的习惯。
`placehold` 可以将空位置替换成`-`, 或者指定字符串。

示例演示:

输入文件： `emapper.txt`

```
$ cat emapper.txt | head -n6
```

```
# emapper version: emapper-2.0.1 emapper DB: 2.0
# command: emapper.py  -i Y1.faa --target_orthologs one2one --scratch_dir scratch --output_dir mapper --output Y1 --keep_mapping_files --cpu 72 -m diamond
# time: Sun Aug  9 17:28:10 2020
#query_name     seed_eggNOG_ortholog    seed_ortholog_evalue    seed_ortholog_score     best_tax_level  Preferred_name  GOs     EC      KEGG_ko KEGG_Pathway    KEGG_Module     KEGG_Reaction   KEGG_rclass  BRITE  KEGG_TC CAZy    BiGG_Reaction   taxonomic scope eggNOG OGs      best eggNOG OG  COG Functional cat.     eggNOG free text desc.
Y1_g_00001      1231336.L248_1956       9.4e-37 159.8   Lactobacillaceae                        Bacteria        1TSRV@1239,3F4EQ@33958,4HCIZ@91061,COG1132@1,COG1132@2        NA|NA|NA        V       ABC transporter transmembrane region
Y1_g_00002      543734.LCABL_17770      9.3e-193        679.5   Lactobacillaceae        dnaJ                    ko:K03686               ko00000,ko03029,ko03110                 Bacteria        1TP00@1239,3F490@33958,4H9KA@91061,COG0484@1,COG0484@2  NA|NA|NA        O       ATP binding to DnaK triggers the release of the substrate protein, thus completing the reaction cycle. Several rounds of ATP-dependent interactions between DnaJ, DnaK and GrpE are required for fully efficient folding. Also involved, together with DnaK and GrpE, in the DNA replication of plasmids through activation of initiation proteins
```

运行命令： 填充空白元素, 使用 `-p` 设定填充字符串。

```
$ tsv-utils  placehold -p '-' emapper.txt | head -n6
```

```
# emapper version: emapper-2.0.1 emapper DB: 2.0
# command: emapper.py  -i Y1.faa --target_orthologs one2one --scratch_dir scratch --output_dir mapper --output Y1 --keep_mapping_files --cpu 72 -m diamond
# time: Sun Aug  9 17:28:10 2020
#query_name     seed_eggNOG_ortholog    seed_ortholog_evalue    seed_ortholog_score     best_tax_level  Preferred_name  GOs     EC      KEGG_ko KEGG_Pathway    KEGG_Module     KEGG_Reaction   KEGG_rclass  BRITE  KEGG_TC CAZy    BiGG_Reaction   taxonomic scope eggNOG OGs      best eggNOG OG  COG Functional cat.     eggNOG free text desc.
Y1_g_00001      1231336.L248_1956       9.4e-37 159.8   Lactobacillaceae        -       -       -       -       -       -       -       -       -       -       -       -       -       -       -       -       -       Bacteria        1TSRV@1239,3F4EQ@33958,4HCIZ@91061,COG1132@1,COG1132@2        NA|NA|NA        V       ABC transporter transmembrane region
Y1_g_00002      543734.LCABL_17770      9.3e-193        679.5   Lactobacillaceae        dnaJ    -       -       ko:K03686       -       -       -       -       ko00000,ko03029,ko03110 -       -       -    Bacteria        1TP00@1239,3F490@33958,4H9KA@91061,COG0484@1,COG0484@2  NA|NA|NA        O       ATP binding to DnaK triggers the release of the substrate protein, thus completing the reaction cycle. Several rounds of ATP-dependent interactions between DnaJ, DnaK and GrpE are required for fully efficient folding. Also involved, together with DnaK and GrpE, in the DNA replication of plasmids through activation of initiation proteins
```

注意事项：一般数值使用`0`, 字符串使用`-`填充。

Last Update: 8/30/2020 3:57:49 PM