tsv-utils之指定的文件列进行注释：definition

一、tsv-utils definition介绍

功能描述：

`tsv-utils definition` 根据字典：`key/value` 对指定的文件列进行注释；

注意事情1：`key/value` 至选择2列，第一列为`key`，第二列为`value`，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。

注意事情2：针对一对多的 `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，可指定合并元素之间的分割符。

命令行接口：

```
$ tsv-utils  definition

Usage: tsv-utils definition [options] <db> <tab>
Options:
  -c INT     target col for annotation, default: [1]
  -d char*   delimitor between key and definition, default: ['  ']
  -s char    delimitor for duplicated values elements in db, default, [',']
  -t char*   title name, default: 'definition'
  -p char*   placehold, default: [-]
```

可选参数：

```
-c  整数    添加注释的列，默认为第一列；
-d  字符串  关键词和注释之间的间隔符，默认为'    ';
-s  字符串  注释文件中重复注释之间的间隔符，默认为',';
-t  字符串  注释列的表头，默认为definition;
-p  字符串  若不存在注释则用指定的字符填充，默认为'-';
```

二、使用场景实例及其用法###

示例演示：

示例文件：`ko.txt, annotation.txt`

```
$ cat ko.txt  | head -n 6
```

```
#KO     A-1     A-2     B-1     B-2     C-1     C-2
K00001  0.000808980004315947    0.000819109861945028    0.000612948680922169    0.000635933480682086    0.000559136968670103    0.000542486474351325
K00002  7.72038380845697e-06    7.22688422533568e-06    2.49766518305712e-05    1.57502121695051e-05    9.56550887982675e-05    9.858089243293e-05
K00003  0.000587762479037045    0.000592135220229698    0.000519999878818898    0.000549497183764971    0.000666059282177164    0.000677436021110664
K00004  1.49014375653928e-05    1.46178664834576e-05    1.85089288332373e-05    2.31926009709905e-05    8.68979017822108e-06    8.51275477401169e-06
K00005  1.96858905532703e-05    1.86794815743808e-05    9.00123097627357e-05    9.40179323138053e-05    0.00011454456970013     0.000116547766601784
```

```
$ cat annotation.txt  | head -n 6
```

```
K00001  alcohol dehydrogenase [EC:1.1.1.1]
K00002  alcohol dehydrogenase (NADP+) [EC:1.1.1.2]
K00003  homoserine dehydrogenase [EC:1.1.1.3]
K00004  (R,R)-butanediol dehydrogenase / meso-butanediol dehydrogenase / diacetyl reductase [EC:1.1.1.4 1.1.1.- 1.1.1.303]
K00005  glycerol dehydrogenase [EC:1.1.1.6]
K00006  glycerol-3-phosphate dehydrogenase (NAD+) [EC:1.1.1.8]
```

运行命令：

示例参数**1:** 使用默认参数，运行后在第一列后面添加注释, 设定列表 `title` 为 `description`, 如未指定, 使用 `definition`

```
$ tsv-utils definition -t "description"  annotation.txt  ko.txt | head -n 6
```

```
#KO     description     A-1     A-2     B-1     B-2     C-1     C-2
K00001  alcohol dehydrogenase [EC:1.1.1.1]      0.000808980004315947    0.000819109861945028    0.000612948680922169    0.000635933480682086    0.000559136968670103    0.000542486474351325
K00002  alcohol dehydrogenase (NADP+) [EC:1.1.1.2]      7.72038380845697e-06    7.22688422533568e-06    2.49766518305712e-05    1.57502121695051e-05    9.56550887982675e-05    9.858089243293e-05
K00003  homoserine dehydrogenase [EC:1.1.1.3]   0.000587762479037045    0.000592135220229698    0.000519999878818898    0.000549497183764971    0.000666059282177164    0.000677436021110664
K00004  (R,R)-butanediol dehydrogenase / meso-butanediol dehydrogenase / diacetyl reductase [EC:1.1.1.4 1.1.1.- 1.1.1.303]      1.49014375653928e-05    1.46178664834576e-05    1.85089288332373e-05    2.31926009709905e-05    8.68979017822108e-06    8.51275477401169e-06
K00005  glycerol dehydrogenase [EC:1.1.1.6]     1.96858905532703e-05    1.86794815743808e-05    9.00123097627357e-05    9.40179323138053e-05    0.00011454456970013     0.000116547766601784
```

示例参数**2:** 设置-d参数更改列修饰模式,可以直接追加在列字符串后。

```
$ tsv-utils definition -d"; "  annotation.txt  ko.txt | head -n 6
```

```
#KO     A-1     A-2     B-1     B-2     C-1     C-2
KO0001; alcohol dehydrogenase [EC:1.1.1.1]      0.000808980004315947
0.000819109861945028    0.000612948680922169    0.000635933480682086
0.000559136968670103    0.000542486474351325
KO0002; alcohol dehydrogenase (NADP+) [EC:1.1.1.2]      7.72038380845697e-06
   7.22688422533568e-06    2.49766518305712e-05    1.57502121695051e-05
9.56550887982675e-05    9.858089243293e-05
KO0003; homoserine dehydrogenase [EC:1.1.1.3]   0.000587762479037045
0.000592135220229698    0.000519999878818898    0.000549497183764971
0.000666059282177164    0.000677436021110664
KO0004; (R,R)-butanediol dehydrogenase / meso-butanediol dehydrogenase /
diacetyl reductase [EC:1.1.1.4 1.1.1.- 1.1.1.303]      1.49014375653928e-05
1.46178664834576e-05    1.85089288332373e-05    2.31926009709905e-05
8.68979017822108e-06    8.51275477401169e-06
KO0005; glycerol dehydrogenase [EC:1.1.1.6]     1.96858905532703e-05
1.86794815743808e-05    9.00123097627357e-05    9.40179323138053e-05
0.00011454456970013     0.000116547766601784
```


Last Update: 8/30/2020 5:27:04 PM