tsv-utils之文件列进行注释：annotation

一、tsv-utils annotation介绍

功能描述：

`tsv-utils 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；

`注意事项:` 主键设定为第一列字段，注释信息支持多列注释注释信息，注释文件尽量保留表头信息，并以`#`开头表示注释信息。

命令行接口：

```
$ tsv-utils  annotation

Usage: tsv-utils annotation [options] <db> <text>
Options:
  -p  STR string for missing data. default: [-]
  -c  INT column for annotation. default: [1]
  -r      remove unmapped target. default: [NOT]
```

可选参数：

```
 -p  指定    指定字符用来补齐缺失的注释信息，默认为“ -”；
 -c  整数    指定需要注释的列，默认为第一列；
 -r          删除没有注释信息的列，默认为NOT;
```

二、使用场景实例及其用法

1. 16S扩增子数据分析: `Tax4fun` 功能预测, `megablast` 比对结果关联物种注释。

示例演示

示例文件： `megablast.txt，SILVA_123_SSURef.txt`

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

```
ZOTU_1  JQ088343.1.1479 100.000 404     0       0       1       404     355
  758     0.0     747
ZOTU_2  JF138741.1.1356 100.000 429     0       0       1       429     323
  751     0.0     793
ZOTU_3  LN568336.1.1321 98.765  405     4       1       1       404     298
  702     0.0     719
ZOTU_4  FJ672447.1.1441 99.764  424     1       0       1       424     355
  778     0.0     778
ZOTU_5  EU542474.1.1489 100.000 424     0       0       1       424     353
  776     0.0     784
ZOTU_6  JQ072613.1.1343 100.000 404     0       0       1       404     270
  673     0.0     747
```

```
$ cat SILVA_123_SSURef.txt | head -n6
```

```
X92134.1.1483
Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;
Pseudomonas;bacterium 52N3
EF442993.1.1422
Bacteria;Proteobacteria;Deltaproteobacteria;Desulfobacterales;Desulfobulbacea
e;Desulfobulbus;Desulfobulbus sp. DSM 2033
JQ972882.1.1517
Bacteria;Actinobacteria;Actinobacteria;Streptosporangiales;Thermomonosporacea
e;Actinomadura;Actinomadura bangladeshensis
FJ799133.1.1480
Bacteria;Tenericutes;Mollicutes;Acholeplasmatales;Acholeplasmataceae;Acholepl
asma;bacterium enrichment culture clone BA75
AY554420.1.1452
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Proteinip
hilum;Bacteroides sp. 22C
AY741401.1.1512 Bacteria;Proteobacteria;Gammaproteobacteria;NKB5;Legionella-
like amoebal pathogen HT99
```

运行命令：

参数使用**1**：：默认参数，若注释信息缺失，则默认用 `-` 填充。

```
$ cut -f1,2  megablast.txt | tsv-utils annotation -c 2 SILVA_123_SSURef.txt -
| head -n6
```

```
ZOTU_1  JQ088343.1.1479
Bacteria;Proteobacteria;Epsilonproteobacteria;Campylobacterales;Campylobacter
aceae;Arcobacter;uncultured bacterium
ZOTU_2  JF138741.1.1356 -
ZOTU_3  LN568336.1.1321
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Acetobacteraceae
;Roseomonas;uncultured bacterium
ZOTU_4  FJ672447.1.1441
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Proteinip
hilum;uncultured bacterium
ZOTU_5  EU542474.1.1489 Bacteria;Bacteroidetes;Bacteroidetes
vadinHA17;uncultured bacterium
ZOTU_6  JQ072613.1.1343
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Rhizobiu
m;uncultured bacterium
```

参数使用**2**： 设置参数 `-p`，指定对应的符号填充缺失注释。

```
$ cut -f1,2  megablast.txt | tsv-utils annotation  -p "NA" -c 2
SILVA_123_SSURef.txt - | head -n6
```

```
ZOTU_1  JQ088343.1.1479 Bacteria;Proteobacteria;Epsilonproteobacteria;Campylobacterales;Campylobacteraceae;Arcobacter;uncultured bacterium
ZOTU_2  JF138741.1.1356 NA
ZOTU_3  LN568336.1.1321 Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Acetobacteraceae;Roseomonas;uncultured bacterium
ZOTU_4  FJ672447.1.1441 Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Proteiniphilum;uncultured bacterium
ZOTU_5  EU542474.1.1489 Bacteria;Bacteroidetes;Bacteroidetes vadinHA17;uncultured bacterium
ZOTU_6  JQ072613.1.1343 Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Rhizobium;uncultured bacterium
```

参数使用**3**：：设置参数`-r`，若注释信息缺失，删除对应的行。

```
$ cut -f1,2  megablast.txt | tsv-utils annotation  -r  -c 2 SILVA_123_SSURef.txt - | head -n6
```

```
ZOTU_1  JQ088343.1.1479 Bacteria;Proteobacteria;Epsilonproteobacteria;Campylobacterales;Campylobacteraceae;Arcobacter;uncultured bacterium
ZOTU_3  LN568336.1.1321 Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Acetobacteraceae;Roseomonas;uncultured bacterium
ZOTU_4  FJ672447.1.1441 Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Proteiniphilum;uncultured bacterium
ZOTU_5  EU542474.1.1489 Bacteria;Bacteroidetes;Bacteroidetes vadinHA17;uncultured bacterium
ZOTU_6  JQ072613.1.1343 Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Rhizobium;uncultured bacterium
ZOTU_7  CU926767.1.1359 Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Paludibacter;uncultured bacterium
```


Last Update: 8/30/2020 5:58:14 PM