taxon-utils之过滤掉指定的节点：filter

一、taxon-utils filter介绍

功能描述：

`taxon-utils filter`,过滤掉指定的节点（包含其所有子节点），使用场景：Kraken序列分类中去除特定分类的比对结果，比如病毒，或者真菌类。

命令行接口：

```
$ taxon-utils filter

Usage: taxon-utils filter [options] <taxon> <taxonId:list>

Options:
  -c  INT column for taxonId lineage. default: [1]

Usage: taxon-utils filter [options] 1.txt 3,2759
```

可选参数：

```
  -c     整型        指定要过滤的列，默认为第一列；
```

二、使用场景实例及其用法

使用场景经典案例：

宏基因组序列分类， 去除一些污染序列。

示例演示：

示例文件在: `data` 目录

示例文件： `classify.txt.gz`, `taxon.map.gz`

```
$ zcat classify.txt.gz | head -n  6
```

```
C	A01050:204:HF7FGDSXY:4:1101:12943:1016	25457	150|149 0:22 25457:4 0:7 25458:5 0:78 |:| 3:4 3343:5 8109:2 25457:5 25460:1 25457:6 25693:5 0:48 25458:2 0:1 3343:2 14694:7 3343:10 3:5 0:12
C	A01050:204:HF7FGDSXY:4:1101:12337:1031	18331	150|149 0:60 31294:5 0:25 18331:3 0:23 |:| 0:93 18331:3 0:19
C	A01050:204:HF7FGDSXY:4:1101:16866:1047	8364	150|150 0:17 20367:2 35825:1 0:8 3:4 20616:5 3:1 4523:5 3:3 0:9 3:5 0:19 27173:2 0:12 27173:2 0:21 |:| 0:2 2878:3 24660:5 28576:3 30576:2 3:13 0:9 2483:17 27170:3 2483:2 27170:5 3:6 8364:3 3:7 8364:3 27172:2 8364:5 27172:6 171:5 0:15
C	A01050:204:HF7FGDSXY:4:1101:22742:1047	23158	150|150 0:82 23158:4 0:30 |:| 0:116
C	A01050:204:HF7FGDSXY:4:1101:28664:1063	559	150|150 0:22 559:3 0:7 171:5 0:79 |:| 0:116
C	A01050:204:HF7FGDSXY:4:1101:20473:1094	25986	150|150 0:45 25986:3 0:68 |:| 0:116
```

注意事项：`classify`文件为`Kraken2`分类的结果，第一列为标识符：`C`：可以分类的序列，`U`：不能分类的序列，第三列为分类的`Taxonomy ID`，可以为`NCBI`分类号或者使用`GTDB`的自定义分类号，正常我们只需要第二列和第三列。

```
$ zcat taxon.map.gz | head -n 6
```

```
1	1	no rank root	root
2	1	superkingdom	Archaea root
3	1	superkingdom	Bacteria	root
4	3	phylum	4572-55 Bacteria
5	3	phylum	AABM5-125-24	Bacteria
6	3	phylum	AB1-6	Bacteria
```

运行命令：

运行命令：

```
$ zcat classify.txt.gz | grep -P "^C" | cut -f2,3 | head -n 6
```

```
A01050:204:HF7FGDSXY:4:1101:12943:1016  25457  k:Bacteria,p:Eremiobacterota,c:Eremiobacteria,o:UBP12,f:UBA5184,g:Palsa-1478,s:Palsa-1478 sp003140215   3,56,265,1390,3343,8109,25457
A01050:204:HF7FGDSXY:4:1101:12337:1031  18331  k:Bacteria,p:Gemmatimonadota,c:Gemmatimonadetes,o:Gemmatimonadales,f:Gemmatimonadaceae,g:Fen-1231,s:Fen-1231 sp003171215  3,78,283,806,2237,5826,18331
A01050:204:HF7FGDSXY:4:1101:16866:1047  8364  k:Bacteria,p:Actinobacteriota,c:Actinobacteria,o:Actinomycetales,f:Micrococcaceae,g:Pseudarthrobacter   3,9,171,559,2483,8364
A01050:204:HF7FGDSXY:4:1101:22742:1047  23158  k:Bacteria,p:Proteobacteria,c:Alphaproteobacteria,o:Rhizobiales,f:Beijerinckiaceae,g:Methylobacterium,s:Methylobacterium sp003173775  3,110,175,1021,1696,7507,23158
A01050:204:HF7FGDSXY:4:1101:28664:1063  559  k:Bacteria,p:Actinobacteriota,c:Actinobacteria,o:Actinomycetales  3,9,171,559
A01050:204:HF7FGDSXY:4:1101:20473:1094  25986  k:Bacteria,p:Actinobacteriota,c:Actinobacteria,o:Actinomycetales,f:Dermatophilaceae,g:Pedococcus,s:Pedococcus sp001426245  3,9,171,559,1919,8203,25986
```

将第二例翻译成物种世系格式：

```
$ zcat classify.txt.gz | grep -P "^C" | cut -f2,3 | head -n 6 | taxon-utils lineage   -c 2  taxon.map.gz - > lineage.txt
$ cat lineage.txt
```

```
A01050:204:HF7FGDSXY:4:1101:12943:1016  25457  root,Bacteria,Eremiobacterota,Eremiobacteria,UBP12,UBA5184,Palsa-1478,Palsa-1478 sp003140215    1,3,56,265,1390,3343,8109,25457
A01050:204:HF7FGDSXY:4:1101:12337:1031  18331  root,Bacteria,Gemmatimonadota,Gemmatimonadetes,Gemmatimonadales,Gemmatimonadaceae,Fen-1231,Fen-1231 sp003171215 1,3,78,283,806,2237,5826,18331
A01050:204:HF7FGDSXY:4:1101:16866:1047  8364  root,Bacteria,Actinobacteriota,Actinobacteria,Actinomycetales,Micrococcaceae,Pseudarthrobacter  1,3,9,171,559,2483,8364
A01050:204:HF7FGDSXY:4:1101:22742:1047  23158  root,Bacteria,Proteobacteria,Alphaproteobacteria,Rhizobiales,Beijerinckiaceae,Methylobacterium,Methylobacterium sp003173775  1,3,110,175,1021,1696,7507,23158
A01050:204:HF7FGDSXY:4:1101:28664:1063  559  root,Bacteria,Actinobacteriota,Actinobacteria,Actinomycetales   1,3,9,171,559
A01050:204:HF7FGDSXY:4:1101:20473:1094  25986  root,Bacteria,Actinobacteriota,Actinobacteria,Actinomycetales,Dermatophilaceae,Pedococcus,Pedococcus sp001426245        1,3,9,171,559,1919,8203,25986
```

基于`lineage`的输出结果，过滤掉一些序列。

```
$ taxon-utils  filter  lineage.txt  25457,559
```

```
A01050:204:HF7FGDSXY:4:1101:12337:1031  18331
root,Bacteria,Gemmatimonadota,Gemmatimonadetes,Gemmatimonadales,Gemmatimonadaceae,Fen-1231,Fen-1231 sp003171215 1,3,78,283,806,2237,5826,18331
A01050:204:HF7FGDSXY:4:1101:22742:1047  23158
root,Bacteria,Proteobacteria,Alphaproteobacteria,Rhizobiales,Beijerinckiaceae,Methylobacterium,Methylobacterium sp003173775
1,3,110,175,1021,1696,7507,23158
```


Last Update: 2020-08-10 11:56 AM