fastx-utils之根据标识符去除重复序列：dedup

一、fastx-utils dedup 介绍

功能描述：

`fastx-utils dedup` 根据序列去除重复。

命令行接口：

```
$ fastx-utils  dedup

Usage: fastx-utils dedup <in.fa/fq>
```

二、使用场景实例及其用法

示例演示：

示例文件： `protein.faa`

```
$ cat protein.faa | head -n 10
```

```
>AGE80385.1
MKKIIIISATTIVIGITSFAYFGSKTPLHNEAKAVESQKHNNHKKEEIPAFPKADHNAKKIDNDFSVVTNPKSNLVLINK
HRKLPDGYIPEDLTRPNVPFISPKDKEKTLLRKDAAEALENMFKAAKKEGLDLTAVSGYRSYKRQKSLHDTYVRRQGKAE
ANSVSAIPGTSEHQTGLAMDISSKSAKFQLEPIFGETAEGKWVAEHAHEFGFVIRYLEDKTDTTEYAYEPWHLRYVGNPY
ATYLYKHHLTLEEAMEDKK
>AGE79558.1
MINHELVERINFLAKKAKAEGLTEEEQRERQSLREQYLKGFRQNMLNELKGIKVVNEEGTDVTPTKLKALKKQDNAKLN
>AGE81073.1
MDKVISNEILQQFKDRMRLGDDEDANLRRILFASNKDLTRVCGNYDLNIDEVFKELVFERSRYVYNDALEYFDKNFLSQI
NSLSIGKALEAIKLDGD
>AGE80385.1
MQNGKVKWFNSEKGFGFIEVEGGEDVFVHFSAIQGEGFKTLEEGQEVTFEVEQGNRGPQATNVNKK
>AGE81522.1
MKKTILTTTAALTMMGTGMGINVDHIKPAEVKADTISFYDVPNNHWATKAITNLANRNIVVGYGNGQFGFGDNVTRGQVA
RMIYNYLKPADAGNFKNPFSDIKGHMFEKEILALAKVGIIKGYGEGKFGPDDILTREQMAQVLTNAFKFEGTKKTSFVDV
DKNSWSYKAIGALEEKGVTIGTGGNMYSPTSVVTREQYSQFLFNSINVIEKETKPEEKPNTGGEVKPEEKPNTGGEVKPE
EKPNTGEETKPVNIPEWLETSLATNDFTFTQAWYDGSEAINKAASTNAQQIVKNINSKYGTNLKYSEVGAIVQLVDGARE
QLWLAGMNVNDFRVTFRVSNNAMIELTKELVTLVNSDLNLDQEIQEIPSAPMKIKNVEKGDYKIRISPAMADQMITIIIE
KK
```

运行命令： 去除`protein.faa`文件中的重复序列.

```
$ fastx-utils  dedup  protein.faa | head -n  10
```

```
>AGE80385.1
MKKIIIISATTIVIGITSFAYFGSKTPLHNEAKAVESQKHNNHKKEEIPAFPKADHNAKKIDNDFSVVTNPKSNLVLINK
HRKLPDGYIPEDLTRPNVPFISPKDKEKTLLRKDAAEALENMFKAAKKEGLDLTAVSGYRSYKRQKSLHDTYVRRQGKAE
ANSVSAIPGTSEHQTGLAMDISSKSAKFQLEPIFGETAEGKWVAEHAHEFGFVIRYLEDKTDTTEYAYEPWHLRYVGNPY
ATYLYKHHLTLEEAMEDKK
>AGE79558.1
MINHELVERINFLAKKAKAEGLTEEEQRERQSLREQYLKGFRQNMLNELKGIKVVNEEGTDVTPTKLKALKKQDNAKLN
>AGE81073.1
MDKVISNEILQQFKDRMRLGDDEDANLRRILFASNKDLTRVCGNYDLNIDEVFKELVFERSRYVYNDALEYFDKNFLSQI
NSLSIGKALEAIKLDGD
>AGE81522.1
MKKTILTTTAALTMMGTGMGINVDHIKPAEVKADTISFYDVPNNHWATKAITNLANRNIVVGYGNGQFGFGDNVTRGQVA
RMIYNYLKPADAGNFKNPFSDIKGHMFEKEILALAKVGIIKGYGEGKFGPDDILTREQMAQVLTNAFKFEGTKKTSFVDV
DKNSWSYKAIGALEEKGVTIGTGGNMYSPTSVVTREQYSQFLFNSINVIEKETKPEEKPNTGGEVKPEEKPNTGGEVKPE
EKPNTGEETKPVNIPEWLETSLATNDFTFTQAWYDGSEAINKAASTNAQQIVKNINSKYGTNLKYSEVGAIVQLVDGARE
QLWLAGMNVNDFRVTFRVSNNAMIELTKELVTLVNSDLNLDQEIQEIPSAPMKIKNVEKGDYKIRISPAMADQMITIIIE
KK
>AGE80137.1
MKLLDLLSKGIVIGDGAVGTLLHSHGLQSSFEELNVSDPDLIISIHKQYVAAGADVIQTNTYGANEAKLRMYGLENQVTK
INRAAVKLAKASVTDKNAILGTIGGMKHIGAVTTTDMEREFMLLEQAGALLEEQVDGLLLETFYDEFELLHAVKVLRKQT
NIPIVAQLALHEAGTTQNGNDVNEILKQFIDYGANVVGLNCQLGPLHMTEAFKMISIPQNGYLSAYPNAGLPNYVEGRYV
YEGSPAYFEAMTPNFIEQGIRLLGGCCGTTPEHIQSMKRAVANITPVIEKETIQRPKVVHTHEKRSKAHVTLAEKAKKQT
TVVVELDPPKTLDTQRFFEGARALKRAGADAITLADNSLASPRVSNMAMGALLTKHDIPVLTHLTCRDHNVIGLQSHLLG
LSALGMEEVLALTGDPARVGDFPGATSVYDLSSIELIKMIKEMNDGRSILGKSIGPATRFSVGGAFNPHVRHLKAAVKRM
ERKIDAGAEYFLTQPIYDIKLIEEVYEATKHLEQPIFIGIMPLVSKRNADFLHFEVPGITLPEEIRERMDGHETKEAAIE
EGIRISQELVDAAMKYFNGIYLITPFLKYEITEHLVKYVREKQEVKEGIN
```

本文材料为 **BASE** (**B**iostack **A**pplied bioinformatic **SE**ies) 课程 **Linux Command Line Tools for Life Scientists** 材料，

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