Como remover valores duplicados em um arquivo de texto no Linux usando scripts

Quando lidamos com grandes conjuntos de dados em arquivos de texto no Linux, é comum encontrarmos valores duplicados que podem dificultar a análise ou processamento posterior. Felizmente, existem várias maneiras de remover esses valores duplicados usando scripts. Neste artigo, iremos explorar alguns exemplos de como realizar essa tarefa de forma eficiente.

Exemplos:

1. Utilizando o comando sort e uniq: O comando sort é amplamente utilizado para classificar linhas de texto em ordem alfabética ou numérica. Ao combiná-lo com o comando uniq, que remove linhas duplicadas consecutivas, podemos remover valores duplicados em um arquivo de texto. Por exemplo:

```
sort arquivo.txt | uniq > arquivo_sem_duplicatas.txt
```

Nesse exemplo, o arquivo.txt é classificado alfabeticamente pelo comando sort e, em seguida, o comando uniq remove as linhas duplicadas consecutivas. O resultado é redirecionado para o arquivo arquivo_sem_duplicatas.txt.

2. Utilizando o comando awk: O awk é uma ferramenta poderosa para manipulação de texto no Linux. Podemos utilizá-lo para remover valores duplicados em um arquivo de texto usando um script simples. Veja o exemplo abaixo:

```
awk '!a[$0]++' arquivo.txt > arquivo_sem_duplicatas.txt
```

Nesse caso, o script awk verifica se a linha atual ($0) já foi vista antes. Se não, ele imprime a linha e armazena em um array associativo a[$0]. Dessa forma, apenas as linhas únicas são impressas no arquivo_sem_duplicatas.txt.

Conclusão: Remover valores duplicados em um arquivo de texto no Linux pode ser feito de várias maneiras, mas os exemplos apresentados neste artigo demonstram duas abordagens eficientes. O uso do comando sort em conjunto com o comando uniq é simples e amplamente suportado. Por outro lado, o uso do awk permite maior flexibilidade e controle na manipulação do texto. Escolha a opção que melhor se adequar às suas necessidades e aproveite a facilidade de trabalhar com arquivos de texto sem valores duplicados.