# 重複行を見つけるuniqコマンド

重複のないメールアドレスのリストを作成する場合、ファイルをエディタで開いて、目で確認しながら1行ずつ削除するのでしょうか？もちろんその必要はありません。重複している行を削除するuniq(ユニーク)というコマンドを使えば簡単に行えます。

●テキストファイル「maillist.org.txt」(メールアドレスに重複がある)

|  |
| --- |
| makotoo2@sora.example.com  makotoo2@sora.example.com  yoneta@knet.example.com  sakura@knet.example.com  makotoo2@sora.example.com  sasaki@peachpress.example.com  aoki@sora.example.com  naoko@turob.example.com  namida@sunnet.example.com  naoko@turob.example.com  kamida@sora.example.com  makotoo2@sora.example.com |

●uniqコマンドの書式

|  |
| --- |
| uniq 【ファイルパス】 |

**■uniqコマンドの使い方と注意点**

uniqコマンドには一つ注意点があります。それは、uniqコマンドは隣接する重複行のみ削除するということです。そのため、同じ行が離れた位置にある場合は削除されません。前述の「maillist.org.txt」に対して、uniqコマンドをそのまま実行した結果を下示します。

|  |
| --- |
| [ec2-user@ip-172-31-8-35 ~]$ uniq maillist.org.txt  makotoo2@sora.example.com←重複  yoneta@knet.example.com  sakura@knet.example.com  makotoo2@sora.example.com←重複  sasaki@peachpress.example.com  aoki@sora.example.com  naoko@turob.example.com←重複  namida@sunnet.example.com  naoko@turob.example.com←重複  kamida@sora.example.com  makotoo2@sora.example.com←重複 |

元のファイルと見比べてみると分かりますが、隣接する重複行以外はそのまま残っています。なので、sortコマンドを利用して並び替えます。下記のようにsortコマンドの結果をパイプ「|」を介してuniqコマンドに渡します。

|  |
| --- |
| [ec2-user@ip-172-31-8-35 ~]$ sort maillist.org.txt | uniq  aoki@sora.example.com  kamida@sora.example.com  makotoo2@sora.example.com  namida@sunnet.example.com  naoko@turob.example.com  sakura@knet.example.com  sasaki@peachpress.example.com  yoneta@knet.example.com |

これで全ての重複行が削除されました。

**■sortコマンドで重複行を取り除く**

最近のsortコマンドでは重複行を表示しないという「-u」オプションが追加されています。したがって、「sort maillist.org.txt | uniq」は下記のようにコマンドを実行しても同じ結果になります。

|  |
| --- |
| [ec2-user@ip-172-31-8-35 ~]$ sort -u maillist.org.txt  aoki@sora.example.com  kamida@sora.example.com  makotoo2@sora.example.com  namida@sunnet.example.com  naoko@turob.example.com  sakura@knet.example.com  sasaki@peachpress.example.com  yoneta@knet.example.com |

**■uniqコマンドを使いこなす**

uniqコマンドはもう不要かというと、そんなことはありません。例えば、前述の「maillist.org.txt」から、メールアドレスを複数回登録した人のアドレスを抜き出したいといったケースです。そのためには、重複した行のみを表示する「-d」オプションを付けてuniqコマンドを実行します。

|  |
| --- |
| [ec2-user@ip-172-31-8-35 ~]$ sort maillist.org.txt | uniq -d ←重複行のみ表示  makotoo2@sora.example.com  naoko@turob.example.com |

この場合も、あらかじめsortコマンドで並び替えておく必要がある点に注意してください。

次に同じメールアドレスが何回登録されていたかを調べたいといったケースを考えます。

それには、uniqコマンドを各行の先頭に重複していた回数を表示する「-c」オプションを付けて実行します。

|  |
| --- |
| [ec2-user@ip-172-31-8-35 ~]$ sort maillist.org.txt | uniq -c ←重複回数を表示  1 aoki@sora.example.com  1 kamida@sora.example.com  4 makotoo2@sora.example.com  1 namida@sunnet.example.com  2 naoko@turob.example.com  1 sakura@knet.example.com  1 sasaki@peachpress.example.com  1 yoneta@knet.example.com |

さらに上記の結果を登録回数の多い順に並び替えるには、結果をもう一度ソートします。

下記のように、sortコマンドにパイプ「|」で渡します。

|  |
| --- |
| [ec2-user@ip-172-31-8-35 ~]$ sort maillist.org.txt | uniq -c | sort -nrb  4 makotoo2@sora.example.com  2 naoko@turob.example.com  1 yoneta@knet.example.com  1 sasaki@peachpress.example.com  1 sakura@knet.example.com  1 namida@sunnet.example.com  1 kamida@sora.example.com  1 aoki@sora.example.com |

ここで使用したsortコマンドの「-b」オプションは、先頭のスペースを無視するオプションです。