
Subject: Re: Sorting and comparing 2 files
Posted by [Craig Markwardt](#) on Wed, 01 Mar 2006 06:47:47 GMT
[View Forum Message](#) <> [Reply to Message](#)

"sanam" <ajay.pillai@wachovia.com> writes:

> I have a scripting problem and hope you could help me.
>
> Perl, Python or shell scripting is fine.
>
> I have 2 files fileA is a list of assets owned by us and fileB is a
> list of assets held by the whole industry which includes our assets
> too.
>
> I tried fgrep but it is painfully slow
>
> /usr/xpg4/bin/fgrep -f sorted_car_mark_history.csv sorted_duedates.csv
>> /tmp/output.csv
>
> FileA is 2456320 bytes (153520 lines)
> FileB is 100028430 bytes (3334281 lines)
>
> FileA has 2 columns separated by comma where as fileB has 4 columns
> separated by commas.
>
> I need to match the 2 columns in fileA with the 2 columns in fileB (BTW
> the column position is 1 and 2) and for every matched record in FileB
> spool it out to an output file.

This really sounds like a database problem, not an IDL problem. It's possible to download a very nifty database program like sqlite3 (from sqlite.org), ingest these two databases, and then perform a JOIN operation. sqlite3 is completely public domain.

Since you are doing only a very simple match between the first 13 character columns of the two tables, it is possible to make the match in IDL, with a function like CMSET_OP() and set-intersection.
Example,

```
:: Read data into IDL
spawn, 'cat a.txt', A
spawn, 'cat b.txt', B
;; Extract matching columns (characters 0-12)
B2 = strmid(B, 0, 13)
;; Locate the matches
I1 = cmset_op(B2, 'AND', A, /indices)

;; Make the output
```

```
openw, 50, 'output.txt'  
printf, 50, B(II), format='(A)'  
close, 50
```

Of course this needs error checking, etc. And it **assumes** that the two sets can be matched by their first 13 columns.

CMSET_OP() can be found on my web page.
<http://cow.physics.wisc.edu/~craigm/idl/idl.html>

Good luck!
Craig

--

Craig B. Markwardt, Ph.D. EMAIL: craigmnet@REMOVEcow.physics.wisc.edu
Astrophysics, IDL, Finance, Derivatives | Remove "net" for better response
