UNIVERSITY OF EDINBURGH
Edinburgh Regional Computing Centre

User Note 70

(October 1985)

Title:

Searching for Patterns in EMAS Character Files

Author: Neil Hamilton-Smith

Contact: Advisory service

Software Support Category: See Note 15

Synopsis

This Note is about commands which are available on EMAS for scanning text to look for particular words or phrases.

Keywords

CHEF, CHEFL, CONCORD, EM, GREP, KWDSCAN, LOCATE, LOOK, MATCH ADDR, NLINES, OCP, pattern matching, PDTOTEXT, SAMEBYTES, scanning text, SCREED, SHOW, STARTSWITH, SUPERSNAP, TSEARCH, TSEARCHALL, YSEARCH

Edinburgh Regional Computing Centre

James Clerk Maxwell Building, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ. Telephone 031-667 1081

Introduction

This Note is about commands which are available on EMAS 2900, and in some cases on EMAS-3, for scanning text to look for particular words or phrases. If you want to write your own programs for searching text, the final section of this Note may be relevant and you should also see User Note 27.

In this Note a pattern is any sequence of characters which can be typed in at a terminal. The illustrations below use words from literary texts as patterns but the commands and procedures described are not restricted to such patterns. All the commands and procedures described can handle character files, some can handle partitioned files some or all of whose members are character files, and some can handle other types of file.

The two files used for the illustrations are one containing 'Pericles, Prince of Tyre' and a partitioned file containing a random selection of Shakespeare's sonnets. The command ANALYSE gives a measure of their size.

Command: analyse pericles
*File: *PERICLES Type: CHARACTER Length: 106351 Bytes*
Last altered: 14/12/84 at 17.25.11

Command: analyse sonnets
File: SONNETS Type: PARTITIONED Length: 33728 Bytes
Last altered: 04/01/85 at 12.16.29
Members:

S001	*S002*	*S003*	*S004*	*S005*	*S006*	*S007*
S008	*S009*	*S012*	*S013*	*S014*	*S015*	*S016*
S017	*S018*	*S019*	*S030*	*S031*	*S044*	*S050*
S051	*S052*	*S053*	*S057*	*S060*	*S061*	*S062*
S063	*S070*	*S071*	*S080*	*S081*	*S090*	*S091*
S100	*S101*	*S107*	*S110*	*S111*	*S120*	*S121*
S130	*S131*	*S140*	*S141*	*S150*	*S151*	*S154*

An editor is probably the type of command that first comes to mind for looking for patterns, so the use of editors will be explored and then other commands and procedures will be described.

Editors

The EMAS Subsystem Editor (EDIT) and the Edinburgh Compatible Context Editor (ECCE) are automatically available to users and are described in Chapter 8 of the EMAS 2900 Users's Guide and on information cards. If the aim is simply to search for patterns without making alterations then LOOK or SHOW may be used instead of EDIT or ECCE. Suppose we wish to find out whether the pattern 'flowers' appears in the text of Pericles. Then the following sequence of commands would produce the output shown.

Command: look pericles
Look: (a/flowers/p1)*
On sweetest flowers^, yet they poison breed.
[Enter Marina, with a basket of flowers^.]
To strew thy green with flowers^; the yellows, blues,
Come, give me your flowers^. On the sea-margent
*^*B**
Look: e
LOOK PERICLES finished.

Command: show pericles
Show
> (f/flowers/p)*
On sweetest ^flowers, yet they poison breed.
[Enter Marina, with a basket of ^flowers.]
To strew thy green with ^flowers; the yellows, blues,
Come, give me your ^flowers. On the sea-margent
end
> %c

Command:

By default these editors display only the line of text containing the pattern just found: they can easily be instructed to display several lines of context. Additionally their pattern matching is case independent so 'Flowers' will not be overlooked when the pattern being searched for is 'flowers': it is easy to switch to case dependent matching (use command Z with LOOK, or %L with SHOW).

User Note 46 describes several editors: here is a brief outline of the use of three of them for pattern searching.

CHEF

The Christchurch Edit Facility (CHEF) is accessible on EMAS 2900 via directory KNTLIB.GENERAL. By default it will perform a case dependent search for a given pattern and will display only the line of text containing the pattern just found. It is easy to make the search case independent (command LC) and to display several lines of context (command V). If the text being scanned is not to be altered then the command CHEFL may be used, cf LOOK and EDIT. The following sequence of commands would produce the output shown.

Command: chefl pericles
106351
Enter H for help (Q for quit)
> lc
On
> ,x/flowers/;pa
210 /On sweetest flowers, yet they poison breed.
1905 /[Enter Marina, with a basket of flowers.]
1908 /To strew thy green with flowers; the yellows, blues,
1921 /Come, give me your flowers. On the sea-margent
> q

Note how the output begins with the size of the text in characters (106351) and then confirms that case independence is On.

EM

The editor EM is accessible on EMAS 2900 via directory KNTLIB.GENERAL. By default it will perform a case dependent search for a given pattern and will display only the line of text containing the pattern just found. It is easy to display either several lines of context (command %) or the number of the line containing the pattern just found (command =). The following sequence of commands would produce the output shown.

Command: em pericles
Editor
3301
> 1,$g/flowers/p
On sweetest flowers, yet they poison breed.
[Enter Marina, with a basket of flowers.]
To strew thy green with flowers; the yellows, blues,
Come, give me your flowers. On the sea-margent
> q

Note how the output begins with the size of the text in lines (3301).

SCREED

A screen editor which is automatically available to all users and is described in SCREED: A Screen Editor. It performs a case dependent match and displays the line found, the preceding line and the following line (or additional adjacent lines, by selecting a larger REPORT number). For example:

And both like serpents are, who though they feed
On sweetest flowers, yet they poison breed.
Antioch, farewell! for wisdom sees, those men

None of these editors can search through a complete partitioned file. Thus if we want to know whether Shakespeare uses the word 'flowers' in any of his sonnets it would appear that either the sonnets must be inspected severally or PDTOTEXT pdfile, character file
must first be called and then the sonnets can be inspected together.
However, there are other commands which avoid this snag. One such command is SUPERSNAP, which is accessed on EMAS 2900 via directory CONLIB.GENERAL. It is fully described in User Note 36 and produces output not unlike that provided by YSEARCH which is described later.

TSEARCHALL

This command, which is accessed on EMAS-3 via directory ERCLIB:GENERAL and on EMAS 2900 via directory CONLIB.GENERAL, allows one or more character files to be searched for a particular pattern. If the file is a partitioned file then any non-character members are totally ignored.

If the pattern and file name are given as parameters to the command, for example

Command: TSEARCHALL flowers,pericles

then there are the following consequences:

- unless the pattern parameter is enclosed in double quotes, any spaces in the pattern will be removed and the pattern will be converted to upper case, i.e. the pattern will change;
- only one file can be nominated for inspection, control returning to *Command:* at the end of the search.

If the command TSEARCHALL is given without parameters then spaces in the pattern are preserved. When the search through the nominated file is complete TSEARCHALL issues the prompt *File/.END:* which allows either for a further file to be searched for the same pattern or for control to be returned to *Command:* level.

Whichever way the pattern and file are specified, if the pattern is not found then the filename, and the member name if a partitioned file, followed by the message *Not found* are printed. The search is case dependent.

If the pattern is found then the filename, and the member name if a partitioned file, are printed. This is followed by a three line display of each occurrence of the pattern found in the file with the pattern in the middle line of its context.

The following example shows how two files may be searched for all occurrences of the word 'flowers'.

Command: tsearchall
Text: flowers
File/.END: pericles

PERICLES

And both like serpents are, who though they feed
On sweetest flowers, yet they poison breed.
Antioch, farewell! for wisdom sees, those men

I am resolv'd.
[Enter Marina, with a basket of flowers.]
<Q MAR>

No, I will rob Tellus of her weed,
To strew thy green with flowers; the yellows, blues,
The purple violets, and marigolds,

Chang'd with this unprofitable woe!
Come, give me your flowers. On the sea-margent
Walk with Leonine; the air is quick there,

File/.END: sonnets

SONNETS
Member S001 *Not found*

Member S002 *Not found*

Member S003 *Not found*

Member S004 *Not found*

Member S005
Nor it, nor no remembrance what it was:
But flowers distill'd, though they with winter meet,
Leese but their show; their substance still lives sweet.

Member S006 *Not found*

Member S007 *Not found*
:
Member S014 *Not found*

Member S015 *Not found*

Member S016
And many maiden gardens, yet unset,
With virtuous wish would bear your living flowers,
Much liker than your painted counterfeit:

Member S017 *Not found*

Member S018 *Not found*
:
Member S151 *Not found*

Member S154 *Not found*

File/.END: .end

If you are only interested in searching for at most one occurrence of a pattern then the command TSEARCH could be used. It is similar to TSEARCHALL in all ways except that if the pattern is found in the nominated file then the pattern is displayed in context and control is returned to *Command:* level.

Command: tsearch
Text: flowers
File/.END: pericles

PERICLES

And both like serpents are, who though they feed
On sweetest flowers, yet they poison breed.
Antioch, farewell! for wisdom sees, those men

Found

Command:

YSEARCH

This command can be used to inspect a file of any type. It is accessed on EMAS-3 via directory ERCLIB:GENERAL and on EMAS 2900 via directory CONLIB.GENERAL. It takes a single optional parameter, the file name, but prompts for this and other necessary information, thus:

File: The name of the file to be inspected.

Relstart: The number of the byte within the file at which searching is to start. Any value greater than or equal to zero but not greater than the size of the file is valid. This value may be given in decimal or hexadecimal form (in the latter form preceded by X), for example 32 or X20. The maximum value of Relstart could be found from the output of ANALYSE (see Introduction, above). Thus by a careful choice of Relstart all or part of the file may be inspected.

STR/SHORT/INT: The type of pattern that is to be searched for.

STR means a pattern consisting of one or more ISO-coded characters, i.e. a standard text string;
SHORT means a halfword (16-bit) pattern;
INT means a fullword (32-bit) pattern.

The next prompt depends upon which of these three is selected.

STRING: follows STR. A non-null string of characters terminated by typing Return is expected. Thus the string cannot include a newline character. The pattern matching is case dependent.

Search for: follows either SHORT or INT. A 16 or 32-bit integer is expected, which may be given in either decimal or hexadecimal representation, for example 26220 or X666C, 1718382455 or X666C6F77. If an invalid value is supplied then it is ignored and the prompt is reissued. The pattern is searched for only on two-byte boundaries starting from Relstart (whether Relstart is odd or even). This is because YSEARCH is intended primarily for binary (non-text) searches, in which distinguishing even and odd boundaries is sometimes desirable.

If the pattern is not found, then the message *NOT FOUND* is printed and control returns to *Command:* level.

If the pattern is found then the message *FOUND* is printed followed by two or three lines of output which give the address in store of the information, 32 bytes in hexadecimal form and then the same in text form. This is followed by the prompt *Continue?*, to which the response Y or N (Yes or No) may be given.

The following example shows the dialogue that occurred when the sonnets were inspected for the word 'flowers'.

Command: ysearch
File: sonnets
Relstart: 1
STR/SHORT/INT: str
STRING: flowers

FOUND:

```
(00FC09F0)                                    77686174          what
(00000A00)  20697420  7761733A  0A427574  20666C6F  it was: But flo
(00000A10)  77657273  20646973  74696C6C            wers distill'
```

Continue? y

FOUND:

```
(00FC2010)                                    20626561          bea
(00002020)  7220796F  7572206C  6976696E  6720666C  r your living fl
(00002030)  6F776572  732C0A4D  75636820            owers, Much l
```

Continue? y

NOT FOUND

Command:

YSEARCH made it very easy to determine whether the word 'flowers' was used in any sonnet. We have established that it is used twice in the collection but we do not know the particular sonnets in which it was used.

GREP

If you want to know the number of the line of text containing a particular pattern then the editors CHEF and EM mentioned earlier could be used. However, these two commands cannot operate on partitioned files. The command GREP, which is accessed on EMAS 2900 via directory KNTLIB.GENERAL, will search a file of any type for a given pattern. This command takes two parameters, pattern and file, of which the first is optional.

If pattern is supplied as the first parameter it must be enclosed in double quotes thus,

Command: GREP "flowers",pericles

If the pattern is not supplied as the first parameter then it is prompted for; in this case the quotes are unnecessary.

The following example shows how GREP might be used.

Command: grep ,pericles
Pattern: flowers
210 On sweetest flowers, yet they poison breed.
1905 [Enter Marina, with a basket of flowers.]
1908 To strew thy green with flowers; the yellows, blues,
1921 Come, give me your flowers. On the sea-margent

Command: grep ,sonnets
Pattern: flowers
74 But flowers distill'd, though they with winter meet,
203 With virtuous wish would bear your living flowers,

The line numbers given are cumulative from the start of the partitioned file. To find out which sonnets contain these lines the command NLINES, accessed on EMAS-3 via directory ERCLIB:GENERAL and on EMAS 2900 via directory CONLIB.GENERAL, could be used:

Command: nlines pericles, sonnets
PERICLES - 3301 lines [min 0(line 1),max 67(line 2730)]
SONNETS is a PDfile with 49 CHARACTER members as follows:
SONNETS_S001 - 15 lines [min 5(line 1),max 53(line 7)]
SONNETS_S002 - 15 lines [min 5(line 1),max 49(line 15)]
SONNETS_S003 - 15 lines [min 5(line 1),max 49(line 2)]
SONNETS_S004 - 15 lines [min 5(line 1),max 46(line 4)]
SONNETS_S005 - 15 lines [min 5(line 1),max 56(line 15)]
SONNETS_S006 - 15 lines [min 5(line 1),max 50(line 12)]
SONNETS_S007 - 15 lines [min 5(line 1),max 46(line 6)]
SONNETS_S008 - 15 lines [min 5(line 1),max 53(line 4)]
SONNETS_S009 - 15 lines [min 5(line 1),max 52(line 11)]
SONNETS_S012 - 15 lines [min 6(line 1),max 50(line 14)]
SONNETS_S013 - 15 lines [min 6(line 1),max 50(line 9)]
SONNETS_S014 - 15 lines [min 6(line 1),max 46(line 13)]
SONNETS_S015 - 15 lines [min 6(line 1),max 48(line 4)]
SONNETS_S016 - 15 lines [min 6(line 1),max 50(line 8)]
SONNETS_S017 - 15 lines [min 6(line 1),max 53(line 5)]
SONNETS_S018 - 17 lines [min 0(line 2),max 50(line 10)]
SONNETS_S019 - 17 lines [min 0(line 2),max 50(line 5)]
SONNETS_S030 - 16 lines [min 0(line 2),max 51(line 8)]
SONNETS_S031 - 17 lines [min 0(line 2),max 51(line 5)]
SONNETS_S044 - 15 lines [min 6(line 1),max 50(line 11)]

```
SONNETS_S050 - 15 lines      [min 6(line 1),max 50(line 5)]
SONNETS_S051 - 15 lines      [min 6(line 1),max 49(line 4)]
SONNETS_S052 - 15 lines      [min 6(line 1),max 47(line 5)]
SONNETS_S053 - 15 lines      [min 6(line 1),max 48(line 15)]
SONNETS_S057 - 15 lines      [min 6(line 1),max 48(line 7)]
SONNETS_S060 - 15 lines      [min 6(line 1),max 49(line 2)]
SONNETS_S061 - 15 lines      [min 6(line 1),max 49(line 14)]
SONNETS_S062 - 15 lines      [min 6(line 1),max 44(line 14)]
SONNETS_S063 - 15 lines      [min 6(line 1),max 54(line 4)]
SONNETS_S070 - 15 lines      [min 6(line 1),max 48(line 15)]
SONNETS_S071 - 15 lines      [min 6(line 1),max 49(line 5)]
SONNETS_S080 - 15 lines      [min 6(line 1),max 47(line 4)]
SONNETS_S081 - 15 lines      [min 6(line 1),max 54(line 15)]
SONNETS_S090 - 15 lines      [min 6(line 1),max 50(line 6)]
SONNETS_S091 - 15 lines      [min 6(line 1),max 52(line 5)]
SONNETS_S100 - 15 lines      [min 7(line 1),max 50(line 2)]
SONNETS_S101 - 15 lines      [min 7(line 1),max 46(line 10)]
SONNETS_S107 - 15 lines      [min 7(line 1),max 50(line 15)]
SONNETS_S110 - 15 lines      [min 7(line 1),max 54(line 4)]
SONNETS_S111 - 15 lines      [min 7(line 1),max 46(line 5)]
SONNETS_S120 - 15 lines      [min 7(line 1),max 46(line 10)]
SONNETS_S121 - 15 lines      [min 7(line 1),max 51(line 12)]
SONNETS_S130 - 15 lines      [min 7(line 1),max 50(line 13)]
SONNETS_S131 - 15 lines      [min 7(line 1),max 48(line 3)]
SONNETS_S140 - 15 lines      [min 7(line 1),max 57(line 15)]
SONNETS_S141 - 15 lines      [min 7(line 1),max 51(line 6)]
SONNETS_S150 - 15 lines      [min 7(line 1),max 49(line 5)]
SONNETS_S151 - 15 lines      [min 7(line 1),max 47(line 9)]
SONNETS_S154 - 15 lines      [min 7(line 1),max 49(line 4)]
  ( 49 CHARACTER members totalling 742 lines).
```

Then a small sum shows that lines 74 and 203 lie within sonnets 5 and 16.

NLINES will print an error message for any file in its parameter list which is either not available or of type other than character or partitioned.

Two remarks could be made about all the commands described so far. The first concerns the immediate context of the patterns that might be matched. For instance, if the pattern being searched for were 'lowers' then any of the words lowers, blowers, flowers, glowers, followers etc. could be picked up. The second concerns the identification of a line of context in the source text. Some of the commands give line numbers for this purpose. In literary applications it is often more useful to have these references in some form of Chapter, Verse, Page, Act, Scene and Line. Many literary works have been prepared in machine readable form with markers for these divisions. The programs CONCORD, accessed on EMAS-3 via directory ERCLIB:GENERAL and on EMAS 2900 via directory CONLIB.GENERAL, and OCP, accessed via directory ERCLIB.GENERAL, can be used to search for lists of any particular words and present those found in context with references more complete than just line numbers. These two programs can only search character files, and their output from a search of Pericles is as follows:

```
                                                    4   flowers
  I: I: 142                    On sweetest flowers, yet they poison breed.
 IV: I:  16    [Enter Marina, with a basket of flowers.]
 IV: I:  18          To strew thy green with flowers; the yellows, blues,
 IV: I:  30             Come, give me your flowers. On the sea-margent
```

Here CONCORD has been instructed to provide references of the form Act: Scene: and line number within that scene. Alternatively the line numbering could have been counted over the whole play (see CHEF, EM and GREP above).

```
                                                    flowers 4

 I  I  186                        On sweetest flowers, yet they poison breed
IV  I 1559          [Enter Marina, with a basket of flowers
IV  I 1561                  To strew thy green with flowers; the yellows, blues
IV  I 1573                    Come, give me your flowers. On the sea-margent
```

Here OCP has been instructed to provide references of the form Act, Scene and line number.

CONCORD is described in the ERCC CONCORD manual, and OCP is described in the OUCS Oxford Concordance Program Users' Manual.

Timing

The cpu times taken to perform the examples given in this Note were measured 5 times on each EMAS 2900 mainframe and then averaged for the table shown below. They should be used only as a guide to the relative speed of each command rather than as a statement of an invariable quantity.

	Pericles	Sonnets
ANALYSE	0.05	0.15
LOOK: case independent	0.23	-
case dependent	0.14	-
SHOW: case independent	0.19	-
case dependent	0.14	-
CHEFL: case independent	16.50	-
case independent	2.74	-
EM	3.10	-
YSEARCH	2.05	0.71
TSEARCHALL	0.33	0.48
GREP	5.10	1.64
NLINES	1.24	0.49
CONCORD	13.10	-
OCP	103.22	-

Other Procedures

If you want to write your own program for pattern matching, then the IMP function MATCH ADDR (described in User Note 27) may be useful. That Note also describes SAMEBYTES, STARTSWITH and KWDSCAN which may be relevant.

Another function to consider is LOCATE which is accessible on EMAS-3 via directory ERCLIB:GENERAL and on EMAS 2900 via directory CONLIB.GENERAL:

```
%EXTERNALINTEGERFNSPEC locate(%STRING(255) s, %INTEGERNAME curp,
                              %INTEGER lastb)
```

This performs a fast search for the text in string 's' commencing from the virtual address given in 'curp' and up to (but not including) the virtual address given in 'lastb' ('s' need not contain only printable characters). To assist in getting good paging characteristics in certain applications, the function returns a result after about 4Kbytes have been searched; the search may be immediately restarted if required using the returned value in 'curp'.

Results returned:

1	string 's' found; 'curp' contains the virtual address of the first byte.
0	string 's' not found at all; curp=lastb.
-1	string 's' not found in about 4Kbytes from starting 'curp'; 'curp' points to where search can resume.