NAME

concord – build concordance of input document

SYNOPSIS

concord [**–i** *ignored-words-file-name*] [**–c** [**–pn** *proper-names-file*]] [*input-file-name*]

MS-DOS SYNOPSIS

concord [*input-file-name*] [**/i** *ignored-words-file-name*] [**/c** [**/pn** *proper-names-file*]]

DESCRIPTION

If no *input-file-name* is specified, the input is taken from the standard input; otherwise the contents of the file named by *input-file-name* is the input.

A concordance is an alphabetical list of the words in a document, with references to the places in the document in which each word appears. Here, alphabetical order means ASCII collating order, except that if case is not significant, then the upper-case letters are considered as being in the same position in the collating order as their lower-case correspondents.

A document is a list of characters organized as follows:

A word is

a list of any non-delimeter characters (including, but not restricted to, letters, digits, hyphens, and apostrophes) surrounded by word delimiters, i.e., white space (beginning- and end-of-file, control-L, newline, carriage-return, tab, space) or punctuation (comma, period, question mark, exclamation mark, quotation mark).

A line is

a list of words surrounded by line delimiters (beginning- and end-of-file, control-L, newline, carriage return).

A page is

a list of lines surrounded by page delimiters (beginning- and end-of-file, control-L).

A document is

a list of pages surrounded by document delimiters (beginning- and end-of-file).

Note that if an *X* is built of a list of *Y*s, then *X* delimiters are also *Y* delimiters.

Hyphenated words may cross line and page boundaries. If the word preceding a line delimiter ends with a hyphen, then the first word of the next line, which may be on the next page, is considered to be part of the word from the previous line. For example:

```
                              ... num-
ber
```

is considered the word “number”.

Lines and pages are numbered only logically; that is, there are no line numbers and page numbers in the input document. The first *X*, *X* being a line or page, is numbered 1. The *X* that follows an *X* delimeter, *d* is of number $i+1$ when *i* is the number of the *X* that preceeds *d*. In the output listing, the logical line and page numbers are printed.

Only the first 72 characters of a word are considered significant in building the concordance; that is words that differ in the 73rd character are considered identical to the concordance list. Any word whose full length is not significant is flagged both in the output of the document and in the concordance list.

Documents, pages, and lines may be empty. That is, each may contain no characters. This is manifested by having two consecutive delimiters of the right kind in the file.

A reference for a word consists of a page-number, line-number pair, the numbers of the page and line containing the beginning of the word. If a word appears more than once on a give line, then the list of references for a word will contain multiple copies of the same page-number, line-number pair.

When building the concordance for a document, case is significant, that is, “Concordance” is different from “concordance”.

The program accepts a document as the input and outputs to standard output a concordance of that document. The output consists of

1. a listing of the document with the lines numbered on the left and the pages divided by rows of equal signs and numbered both at the top and the bottom and

2. the concordance table with a header announcing that the two columns of the table are the words and the list of references.

Normally, all words are listed in the concordance. If the −**i** option is present in the command line, the following word in the command line is taken as the name of a file containing words not to be counted. The words in this file are separated by white space.

Normally words that differ in the case (upper and lower) of their letters are counted as different in the concordance. If the −**c** option is present in the command line, then case distinctions are ignored, not only in building the concordance, but also in comparing for ignored words. Moreover, if case distinctions are ignored, but the −**pn** option is present in the command line, then the following word is taken as the name of a file containing proper names. The words in this file are separated by white space. A proper name is a word that is to be distinguished in the concordance from other versions of the same word that differ only in the case of its letters. Thus, if case distinctions are ignored but there is a proper names file containing the word “Dov”, then it is considered differerent from “dov” “dOv”, “doV”, “dOV”, “DOv”, and “DOV”, all of which are considered the same. Obviously, in order to be meaningful, each word in the proper names file must have at least one upper-case letter.