NAME

STRSIM − Measure the similarity of two character strings.

SYNOPSIS

SCORE=STRSIM(STR1,L1,STR2,L2)

SCORE	is the REAL*8 similarity score in [0,1]
STR1(L1)	is the CHARACTER*1 first string
STR2(L2)	is the CHARACTER*1 second string

DESCRIPTION

Two strings are similar to the extent that they contain the same letters in the same order. Thus differences in the letters and differences in their order both decrease the similarity between the strings. This routine uses a heuristic to measure the difference in this sense between two strings that might not be of the same length. It identifies the longer string and the shorter one, and initializes to zeros a vector ORDER as long as the longer string. Then it sets NO=0. Next it searches the longer string for each character in the shorter one; if the character at index KS of the shorter string is matched by the character at index KL of the longer string and ORDER(KL) is still zero, ORDER(KL)=KS and NO=NO+1. Then it sorts the indices in ORDER into ascending order counting the number of swaps that are needed; if either element in a comparison is zero (indicating no match) a swap is counted but not performed. Finally it returns

$$\mathrm{STRSIM} = 1 - \frac{\mathrm{SWAPS}}{(\mathrm{MAXSWP}+\mathrm{NO})}.$$

The factor NO is present to ensure that STRSIM has a reasonable value when both strings are short.

LINKAGE

gfortran source.f −L${HOME}/lib −lmisc

AUTHOR

Michael Kupferschmid

EXAMPLE

```
      CHARACTER*24 STR1,STR2
      REAL*8 S,STRSIM
    2 CALL PROMPT('STR1=',5)
      READ(5,*,END=1) STR1
      L1=LENGTH(STR1,24)
      CALL PROMPT('STR2=',5)
      READ(5,*,END=1) STR2
      L2=LENGTH(STR2,24)
      S=STRSIM(STR1,L1,STR2,L2)
      PRINT *,S
      GO TO 2
    1 STOP
      END
```

This example produced the following output:

```
unix[1] a.out
STR1= a
STR2= a
   1.0000000000000000
STR1= a
STR2= ab
  0.50000000000000000
STR1= a
STR2= abc
  0.25000000000000000
STR1= abcd
STR2= abcde
  0.71428571428571430
STR1= fellow
STR2= fallow
  0.75000000000000000
STR1= string
STR2= longstring
  0.25490196078431371
STR1= abcd
STR2= xyzw
   0.0000000000000000
unix[2]
```