NAME

wstem – 文字列を形態素解析器で分割し, 頻度を数える

SYNOPSIS

```
#include <geta/ae.h>

syminfo *
wstem(char *text, WAM *w, int dir, char *path, int *ndp)
```

DESCRIPTION

wstem は, *text* で与えられた文字列を *path* で示された形態素解析器により単語に分解し, さらにそれらが name であるとして, wam_name2id(3) により *w* の *dir* で示される CW の ID に対応付けを行う. 対応する ID があった単語は, 集められ, 配列に詰め合わされて返される(戻り配列のメンバ id に ID がセットされる). 各 ID は *text* 中での出現回数が数えられ, その頻度が数えられる. この数は戻り配列のメンバ TF_d にセットされる. (全ての重複している ID は1エントリにまとめられ, その結果, 戻り値である配列には同じ ID(単語)はたかだか1回しか現れないことが保証される) メンバ TF には, その ID(単語) のTF (wam_freq_sum(w, dir, id) で取得)がセットされる. 結果の配列のための領域は **ws** によって確保される. この領域は free(3) により解放することができる.

ID の出現順と *text* でのその ID に対応する単語(name)の出現順は関係が無い.

形態素解析器は文字列 -1 を唯一の引き数として呼び出され, さらに解析すべき *text* が標準入力から与えられる. そして, 解析結果のうち WAM の行(列)の name に相当する情報(通常, 単語)のみを, 標準出力に 1 行に 1 単語の割りで書き出さなければならない.

SEE ALSO

wam(3), ws(3), syminfo(3), dump_syminfo(3)

BUGS

形態素解析器が異常終了した場合でも, **wstem** は正常終了することがある.

形態素解析器の出力する単語は *text* の長さ+512 バイトより長くなってはならない.

形態素解析後の単語の個数は " *text* の(バイト数で数えたときの)長さ" 個より多くなってはならない.