

# Documentation – LANGSTAT

## NOM

langstat – compteur de lettres

## SYNOPSIS

`./langstat.sh [OPTION]... FILE`

## DESCRIPTION

Analyse un fichier de type texte et indique, pour chacune des lettres de l'alphabet A à Z, le nombre de mots dans lesquelles la lettre est contenue au moins une fois. La recherche s'effectue nativement sur les majuscules uniquement, les minuscules doivent être incluses via l'option -l ou -i. Si le fichier n'est pas formaté un mot par ligne, l'option -t est requise.

## OPTIONS

- h, --help  
Affiche un message d'aide rapide, rappel de syntaxe et exemples. Cette option outrepassse toute autre option
- i Effectue la recherche des minuscules et majuscules combinées (c'est-à-dire minuscule et majuscule correspondantes sont combinées sous la même statistique). Cette option inclut l'option -l et l'outrepasse
- l Inclut les minuscules a à z
- L, --lower-only  
Inclut les minuscules et exclut les majuscules. Cette option outrepassse l'option -l
- o Traite le fichier lettre par lettre, et renvoie le nombre total d'occurrences de chacune, indépendamment du nombre de mots. Cette option outrepassse l'option -t
- plus-lower  
Inclut les caractères accentués en minuscules : àâäçèéêëïîôöùûüæœ
- plus-upper  
Inclut les caractères accentués en majuscules : ÀÂÄÇÈÉÊËÏÎÔÖÙÛÜÆË
- plus-spec  
Inclut les caractères spéciaux de ponctuation: ?!';,:;
- t Indique que le fichier à traiter n'est pas formaté un mot par ligne
- y Inclut les traits d'union

## NOTES

L'option -i ralentit très perceptiblement l'exécution du script sur les gros fichiers.

Aussi avec cette option -i, en sortie ce sont les caractères majuscules qui sont utilisés par défaut pour l'affichage, on peut combiner -i et -L|--lower-only pour avoir les minuscules affichées en lieu et place.

Le script n'inclut pas la possibilité de réunir les caractères accentués sous la même statistique que leurs pendants non accentués. Pour obtenir ceci, il est nécessaire d'effectuer un traitement préalable du fichier à analyser, remplaçant les caractères accentués par les non accentués (avec les commandes sed ou tr, par exemple).

Dans le cadre de l'utilisation de l'option -t, les mots composés (contenant un trait d'union) ne sont pas reconnus en tant que tels et sont traités comme plusieurs mots différents.

Ex : 'franc-tireur' est traité comme 'franc-' et 'tireur' séparés.