

L2 Informatique

TP Algorithmique – Table de hachage

On reprend ici la problématique de la première partie du TP 1 sur la fréquence d'apparition des mots dans un texte, mais en utilisant une structure de données différente.

1. Définir comme structure de données une table de hachage avec chaînage coalescent, permettant de stocker des `occMot`¹. La table de hachage utilisera une zone primaire et une zone de réserve. Cette structure représentera un ensemble de mots.
2. Écrire une opération d'initialisation de la table de hachage, étant donnée une taille entrée en paramètre. La zone de réserve représentera 15% de la taille totale de la table.
3. Écrire la fonction de hachage de Bernstein H , telle que pour un mot $s \in [A..Z]^L$ de longueur L et un facteur de compression m :
 - $H(s, m) = h_L$
 - $h_i = (33 \cdot h_{i-1} + \text{code}(s[i])) \bmod m$, pour $i \in [1..L]$
 - $h_0 = 5381$code($s[i]$) dénote le code ASCII du i -ème caractère du mot s .
4. Écrire une opération d'ajout d'une occurrence d'un mot dans un ensemble de mots. Si le mot est déjà présent dans l'ensemble, on ajoutera une occurrence sans ajouter de nouvel élément ; sinon, on ajoutera un nouvel élément dans la table de hachage.
5. Écrire un sous-programme permettant de stocker dans l'ensemble toutes les occurrences de mots d'un fichier texte.
6. Écrire une fonction qui retourne le nombre d'occurrences d'un mot dans un ensemble de mots (table de hachage). Tester cette fonction dans le programme principal, après avoir construit l'ensemble correspondant aux mots du fichier `bouledesuif.txt`.
7. Écrire un sous-programme permettant d'afficher les informations suivantes :
 - pour chaque alvéole occupée dans la table de hachage : numéro de l'alvéole, mot et nombre d'occurrences ;
 - facteur de remplissage de la zone primaire ;
 - facteur de remplissage de la zone de réserve ;
 - facteur de remplissage total (de la table) ;
 - taille du plus grand paquet (plus long chaînage).

Tester ce sous-programme en comparant les informations calculées pour différentes tailles de table de hachage, par exemple 50 000 000, 1 000 000, 200 000, 50 000, 10 000, 5 000 et 3 586 (nombre de mots distincts dans le texte du fichier `bouledesuif.txt`).

1. `occMot` est défini comme un enregistrement composé d'un mot (chaîne de caractères) et d'un nombre d'occurrences (entier).