TP2 Organisation de données

Véronique Demianenko et Emilie Caillerie

21 Mars 2024

1 Introduction

L'objectif de ce TP est de trouver les mots en double dans deux fichiers .txt en utilisant les tables de hachage, ici en l'occurence nous utiliserons le document $texte_{S}hakespeareetcorncob_{l}owercase$.

2 Choix de M

Pour les fonctions de hachage par division, par multiplication ainsi que le hachage par adressage fermé, on prend M = 58111, qui est le nombre premier le plus proche du nombre de mots le plus grand entre les deux textes.

Pour le hachage par adressage ouvert par sondage linéaire et double hachage, on prend M = 243073. En effet, vu qu'on ajoute tous les élements à la liste, on a au pire tout le temps des collisions, donc on fait : len(shake)+len(corn). On multiplie par 3 pour avoir un taux de remplissage de 30%, et on prend le premier nombre premier qui suit ce résultat.

3 Fonctions de hachage

On va réutiliser en priorité les fonctions de hachage par division, puisque nous avions obtenu de meilleurs résultats avec celle-ci dans le premier TP. Pour le hachage par adressage ouvert en double hachage, on utilisera donc la fonction de hachage par multiplication en plus.

4 Hachage par adressage ouvert et fermé

Pour l'adressage fermé, l'idée est de créer une liste de listes qui contiennent les mots de même indice.

Pour l'adressage ouvert par sondage linéaire, l'idée est d'implémenter les mots dans une seule et même liste. Dans le cas où un mot possède le même indice qu'un autre, on recalcule son indice avec la fonction suivante : h(x,i)= (h'(x)+i)modM, jusqu'à ce qu'il puisse être intégré à la liste. Pour le double

hachage, l'idée reste la même. Cette fois, la fonction est la suivante : h(x,i) = (h1(x)+i*h2(x))modM.

On remarque que le temps d'exécution du programme de hachage par adressage ouvert en sondage linéaire est presque 2 fois plus rapide que par double hachage, ce qui n'est pas normal. Cela est dû au fait que nous avons réutilisé les deux fonctions de hachage que nous avions trouvé pour le tp précédent, et que notre fonction de hachage par multiplication était bien moins efficace que le hachage par division.

5 Intersection

Pour trouver les mots qui sont en double, pour l'adressage ouvert, il suffit de parcourir la liste et de voir s'ils ont déjà été lus ou non. Pour l'adressage fermé, vu que les mots en double devraient se trouver dans une seule et même liste de la table de hachage, et non pas dans deux différentes, il suffit de parcourir les listes de la table de hachage.

Nous obtenons pour les 3 méthodes (fermé, ouvert par sondage linéaire, ouvert par double hachage), 15759 mots en commun dans les deux textes.