TP 6 – Concordance

0 Introduction

Une concordance est une liste de toutes les occurrences de tous les mots d'un texte (voir section 5 pour un exemple). Dans ce tp, on utilise un dictionnaire pour créer une concordance d'un fichier texte. Dans un premier temps, on implante le dictionnaire par une liste chaînée et dans un deuxième temps par une table de hachage.

Seules les opérations les plus importantes sont mentionnées dans les questions. Il y en a d'autres (par exemple, libération de mémoire, affichage, etc.) qu'il est également une bonne idée d'implanter.

1 Dictionnaire (liste chaînée)

▸ **Exercice 1 :** Dans cet exercice, un dictionnaire est implanté par une liste de couples (mot, nombre d'occurrences). On utilise la structure suivante.

```
typedef struct _link {
    char *word;
    int occurrences;     /* nombre d'occurrences */
    struct _link *next;
} link;
```

L'archive `Base.zip` contient un programme qui lit un fichier de texte et insère tous les mots de ce texte (en ordre inverse) dans une liste chaînée.

(a) Modifier le programme pour qu'il cherche chaque mot dans la liste avant de l'insérer. (Utiliser la fonction `find_list`.) Si le mot est déjà présent dans la liste, on incrémente le champs `occurrence` par 1. Sinon, on ajoute le mot au début de la liste grâce à la fonction `insert_first_list`.

(b) Faire en sorte que le programme, à la fin de l'exécution, affiche le nombre de mots distincts ainsi que le nombre de mots total.

(c) Tester votre programme sur les fichiers `old.txt` (petit), `2city11.txt` (grand) ou sur un autre fichier texte de taille importante, par exemple du site `http://www.gutenberg.org/`.

2 Occurrences (liste de listes)

▸ **Exercice 2 :** Dans cet exercice, on remplace le simple compte d'occurrences de chaque mot par une liste chaînée qui stocke toutes les positions où un mot apparait dans le texte. La liste devient donc une liste de listes. On utilise (et modifie) les structures suivantes.

```
typedef struct _olink {
        int pos;
        struct _olink *next;
} olink;

typedef struct _link {
    char *word;
    olink *occurrences;  /* liste des occurrences */
    struct _link *next;
} link;
```

(a) Écrire une fonction `void add_occurrence(link *lnk, int pos)` qui ajoute une occurrence (dans une nouvelle cellule `olink`) à la liste `lnk->occurrences`.

(b) Modifier vos fonctions et programme principal pour qu'ils fonctionnent avec cette nouvelle représentation des occurrences.

3 Table de hachage

▸ **Exercice 3 :** Dans cet exercice, on réimplante le dictionnaire par une table de hachage. On utilise la structure suivante.

```
typedef struct _table {
    link **buckets;
    int  M;       /* nombre de seaux */
    int  size;    /* nombre de mots dans la table */
} table;
```

(a) Écrire une fonction `table *create_table(int M)` qui crée une nouvelle table de hachage à `M` seaux. Tous les seaux sont initialement vides.

(b) Écrire une fonction `void add_occ_table(table *tab, char word[], int pos)` qui insère dans la table de hachage une occurrence du mot `word` avec sa position `pos` dans le texte.

1. La fonction calcule d'abord le seau correspondant au mot `word`. Codez une fonction de hachage que vous pensez adaptée.

2. Si le mot est déjà présent dans le seau, alors la fonction insère une nouvelle occurrence dans la liste d'occurrences grâce à la fonction `add_occurrence` de l'exercice précédent.
3. Sinon, la fonction insère le nouveau mot au début du seau.

(c) Modifier le programme principal pour qu'il fonctionne avec cette nouvelle représentation du dictionnaire.
- Tester le programme sur le fichier `2city11.txt` avec 1, 50 et 2000 seaux. Comparer le temps d'exécution.
- Tester le programme sur le fichier http://norvig.com/big.txt.

(d) Écrire une fonction `int size_table(table *tab)` qui renvoie le nombre de mots présents dans la table. Vérifier que le compte est bon.

▸ **Exercice 4 :** Modifier le programme principal pour qu'il prenne deux arguments sur la ligne de commande : le nom d'un fichier d'entrée et le nom d'un fichier de sortie sur lequel la concordance sera écrite.

4 Pour aller plus loin

▸ **Exercice 5 :** Les questions suivantes sont des suggestions pour explorer davantage la table de hachage et la concordance. Certaines sont plus laborieuses que d'autres. Vous pouvez les faire (ou pas) dans l'ordre que vous voulez.

(a) Faire en sorte que l'affichage de la concordance liste les mots en ordre lexicographique et les occurrences en ordre croissant.

(b) Créer une visualisation des tailles des seaux (longueur des listes) pour vérifier que la fonction de hachage distribue les mots de façon uniforme.

(c) Permettre au nombre de seaux d'augmenter si la table devient trop pleine.

(d) Nettoyer les mots du texte d'entrée. Par exemple, vous pouvez supprimer tout caractère sauf les alphanumériques et les convertir en minuscules.

(e) Écrire un programme qui lit une concordance et recrée le texte. (Il est possible que le résultat ne soit pas exactement le texte d'origine, par exemple si vous avez nettoyé les mots.)

(f) (★) Remplacer la fonction de hachage par une fonction de hachage cryptographique, par exemple MD5 : https://en.wikipedia.org/wiki/MD5. Comparer le temps d'exécution du programme sur un fichier volumineux, en utilisant l'ancienne fonction de hachage et la nouvelle.

5 Exemple

Pour le texte suivant :

```
Old MACDONALD had a farm
E-I-E-I-O
And on his farm he had a cow
E-I-E-I-O
With a moo moo here
And a moo moo there
Here a moo, there a moo
Everywhere a moo moo
Old MacDonald had a farm
E-I-E-I-O
```

on a la concordance (les positions commencent par 0) :

```
a : 3 12 16 21 26 29 32 38
And : 6 20
cow : 13
E-I-E-I-O : 5 14 40
Everywhere : 31
farm : 4 9 39
had : 2 11 37
he : 10
here : 19
Here : 25
his : 8
MACDONALD : 1
MacDonald : 36
moo : 17 18 22 23 30 33 34
moo, : 27
Old : 0 35
on : 7
there : 24 28
With : 15
```