TP 3 : Automates en C-Grep

09 octobre

*Ce sujet écrit par Jean Baptiste Bianquis est assez directement inspiré d'une série d'articles écrits par Russ Cox et disponibles en ligne : https://swtch.com/ rsc/regexp/. Russ Cox est l'un des créateurs du langage Go, et l'auteur de la bibliothèque d'expressions régulières https://github.com/google/re2*RE2.

1 Introduction

L'objectif de cette séance est de programmer une version rudimentaire (mais utilisable) de `grep` en C, utilisant l'automate de Thompson associé à une expression. Les limitations que nous allons accepter :

- le programme prendra en entrée (en plus de l'expression régulière) un éventuel fichier passé en argument, ou l'entrée standard sinon ;
- le programme traitera le flux d'entrée ligne par ligne, et décidera simplement pour chaque ligne si elle est acceptée (dans son ensemble) par l'expression ou non : les lignes acceptées seront envoyées sur la sortie standard ;
- la seule " classe de caractères " valide sera `.` (qui accepte un caractère quelconque, autre que `\n`), il n'y aura donc pas de `\w`, de `[aeiou]...`
- les seuls opérateurs seront la concaténation (que nous noterons explicitement `@`), l'alternative (ou) `|`, l'étoile `*` et le « zéro ou une fois » `?` : pas de `{5}` par exemple ;
- les caractères `*`, `@`, `|`, `.` et `?` seront réservés pour les opérateurs, et ne pourront donc jamais être *matchés* (pas de `\?` pour accepter un caractère `?`) ;
- le plus important, et de loin : l'expression sera donnée en notation postfixe, ce qui simplifiera énormément son analyse syntaxique.

Exemples 1. *En supposant que notre programme a été compilé vers un exécutable* `mygrep`*, on aurait les équivalences suivantes :*

- ```
  grep -E '^ab*$' entree.txt
  mygrep 'ab*@' entree.txt
  ```
- ```
  grep -E '^(ab)*$' entree.txt
  mygrep 'ab@*' entree.txt
  ```
- ```
  grep -E '(ab)*c$'
  mygrep '.*ab@*@c@'
  ```
- ```
  grep -E '([ab].)*'
  mygrep '.*ab|.@*.*@@
  ```

2 Construction de l'automate de Thompson

▷ **Question 1.**

1. Rappeler, sous forme de schémas, les constructions de Thompson pour des expressions de la forme :
   (a) a, où $a \in \Sigma$;
   (b) ef, où e et f sont des expressions régulières ;

(c) $e|f$, où e et f sont des expressions régulières ;

(d) e^*, où e est une expression régulière.

2. Proposer une construction pour $e?$, où e est une expression régulière et le $?$ signifie " zéro ou une fois ".

   Remarques 1.

   Cette construction devra posséder un unique état initial, sans transition entrante, et un unique état final, sans transition sortante.

   Cette construction est inutile en théorie puisque l'on dispose d'un automate pour ϵ et donc pour $e? \equiv e|\epsilon$. Elle nous évite cependant de traiter le cas ϵ, et utilise moins d'états.

3. Combien de transitions sortantes étiquetées par une lettre un état de l'automate de Thompson peut-il posséder ? et combien de transitions sortantes étiquetées par ϵ ? Peut-il posséder à la fois les deux types de transition sortante ?

4. Justifier que si la représentation postfixe de e est de longueur n (en tant que chaîne de caractères), alors l'automate de Thompson associé possède au plus $2n$ états.

◁

Voici la représentation d'automate que nous allons ici utiliser :
Chaque état de l'automate sera représenté par une structure de ce type :

```
struct state {
    int c;
    struct state *out1;
    struct state *out2;
    int last_set;
  };

typedef struct state state_t;
```

- On définit trois constantes globales `MATCH`, `EPS` et `ALL` distinctes et strictement supérieures à 255.
- Si `c` est entre 0 et 255, l'état possède une unique transition sortante, étiquetée par le caractère `c`. Dans ce cas, `out1` pointe vers l'état d'arrivée de cette transition, et `out2` doit valoir `NULL`.
- Si `c` a la valeur particulière `MATCH`, alors il n'a aucune transition sortante, et les pointeurs `out1` et `out2` sont nuls.
- Si `c` a la valeur `EPS`, alors il possède une ou deux transitions sortantes étiquetées par ϵ, et `out1` et `out2` pointent vers les états d'arrivée de ces transitions (`out2` sera nul s'il n'y a qu'une seule transition).
- Si `c` a la valeur `ALL`, alors l'état possède une transition sortante pour chaque caractère de l'alphabet, et ces transitions pointent toutes vers l'état désigné par `out1`. Le pointeur `out2` doit être nul dans ce cas.
- Le champ `last_set` sera expliqué plus tard : pour l'instant, il faut juste savoir qu'il devra être initialisé à `-1` à la création de l'état.

L'automate lui-même sera représenté par la structure suivante :

```
struct nfa {
    state_t *start;
    state_t *final;
    int n;
};

typedef struct nfa nfa_t;
```

- Le champ `start` pointe vers l'état initial de l'automate.
- Le champ `final` pointe vers l'état final.
- Le champ `n` indique le nombre total d'états de l'automate.

▷ **Question 2.**

1. Écrire la fonction `new_state` renvoyant un pointeur vers un nouvel état (alloué sur le tas). Comme dit plus haut, on initialisera `last_set` à `-1`.
2. Écrire la fonction `character` qui renvoie un `nfa_t` reconnaissant le caractère donné. Attention, on renvoie bien un `nfa_t`, par valeur, et pas un `nfa_t*`.
3. Écrire une fonction `all` qui renvoie un `nfa_t` reconnaissant n'importe quel mot de longueur 1. On utilisera le même automate que pour `character`, sauf que le champ `c` de l'état initial sera mis à la valeur `ALL`.
4. Écrire les fonctions `concat`, `alternative`, `star` et `maybe` qui correspondent aux différentes constructions du dernier exercice. Autrement dit, dans l'appel `concat(a, b)`, on suppose que a et b sont deux automates de Thompson, d'ensembles d'états disjoints, reconnaissant deux expressions régulières e et f, et l'on demande de renvoyer l'automate de Thompson pour ef.

◁

La construction de l'automate de Thompson suit directement la structure de l'expression (en tant qu'arbre binaire/unaire), ce qui permet de la réaliser très naturellement à partir de la version postfixe de l'expression, à l'aide d'une pile.

▷ **Question 3.**

1. Rappeler le principe de l'évaluation d'une expression arithmétique en notation postfixe en traitant l'exemple suivant : $12\ 4 + 33! * -$ (où $+$ et $-$ sont binaires et ! unaire).
2. À l'aide du type `stack_t` et des fonctions associées fournis, écrire une fonction `build` qui prend en entrée une *regex* en notation postfixe (sous forme d'une chaîne de caractères) et renvoie l'automate de Thompson correspondant.
3. Déterminer la complexité temporelle de `build` en fonction de la longueur m de la chaîne donnant l'écriture postfixe de l'expression régulière.

◁

3 Exécution de l'automate

L'automate de Thompson que nous venons de construire est un ϵ-AFND, et il n'est donc pas possible de tester l'appartenance d'un mot à son langage en se déplaçant simplement d'état en état à chaque caractère. Deux options se présentent à nous :

- éliminer les ϵ-transitions puis déterminiser l'automate, pour finalement effectuer la reconnaissance sur l'automate déterministe ;
- décider directement l'appartenance en exécutant l'automate de Thompson, en gérant les ϵ-transitions et le non déterminisme.

L'élimination des ϵ-transitions n'est pas complètement immédiate à programmer, mais elle se ramène essentiellement à un parcours de graphe, et peut être effectuée de manière efficace. Le calcul de l'automate des parties pour déterminiser, en revanche, peut provoquer une explosion combinatoire.

Pour l'éviter, nous allons choisir la deuxième approche. Deux variantes sont possibles :

- procéder par *backtracking*, en essayant une par une les transitions possibles à partir d'un état jusqu'à les épuiser ou en trouver une permettant d'accepter le mot ;
- garder trace de l'ensemble des états dans lesquels on *peut* se trouver à un moment donné de la lecture du mot, ce qui revient essentiellement à explorer un chemin dans l'automate des parties, sans calculer explicitement cet automate.

▷ **Question 4.** On se place ici dans le cas particulier des automates non-déterministes du type de l'automate de Thompson :

- un seul état initial ;

- un seul état final ;
- entre une et deux transitions sortantes par état, avec le cas à deux transitions sortantes réservé aux ϵ-transitions.

1. Proposer une fonction très simple `backtrack` ayant le prototype suivant :

   ```
   bool backtrack(state_t *state, char *s);
   ```

   Cette fonction renverra `true` si la lecture du mot s depuis l'état pointé par `state` nous amène dans un état final, `false` sinon. On considérera que le mot s'arrête au premier caractère nul ou `'\n'` rencontré, exclu.

On se donne ensuite les deux fonctions suivantes :

```
bool accept_backtrack(nfa_t a, char *s){
  return backtrack(a.start, s);
}

void match_stream_backtrack(nfa_t a, FILE *in){
  char *line = malloc((MAX_LINE_LENGTH + 1) * sizeof(char));
  while (true) {
    if (fgets(line, MAX_LINE_LENGTH, in) == NULL) break;
    if (accept_backtrack(a, line)) {
      printf("%s", line);
    }
  }
  free(line);
}
```

Remarque 1. *L'appel* `fgets(line, MAX_LINE_LENGTH, in)` *lit un maximum de* `MAX_LINE_LENGTH` *caractères depuis le flux* `in` *jusqu'à tomber sur un caractère* `'\n'` *ou sur la fin du flux. Ces caractères sont recopiés sur* `line`*, y compris le retour à la ligne s'il y en avait un, et un caractère nul est placé ensuite (on peut donc écrire jusqu'à* `MAX_LINE_LENGTH` *1+ caractères). L'appel renvoie un pointeur nul si aucun caractère n'a été lu (si l'on était déjà à la fin du flux, donc).*
On suppose la constante `MAX_LINE_LENGTH` *préalablement définie, et suffisamment grande pour traiter toutes les lignes du flux.*

2. Ecrire un programme ayant le comportement suivant :
   - le premier argument en ligne de commande (obligatoire) est la *regex* sur laquelle on travaille, en notation postfixe ;
   - s'il y a un deuxième argument, il fournit le fichier d'entrée – sinon, l'entrée est l'entrée standard ;
   - le programme traite les lignes de l'entrée une par une, dans l'ordre, en affichant celles qui sont acceptées par l'expression régulière sur la sortie standard.

   Remarque 2. *On ne se préoccupera pas pour l'instant de libérer proprement la mémoire.*

3. Utiliser ce programme pour chercher tous les mots de la langue française contenant à la fois un q et un w. Comparer le temps d'exécution de cette requête avec une requête équivalente traitée par `grep` (on utilisera l'utilitaire `time` pour ce faire : `time grep regex file`).

4. Que se passe-t-il si on exécute notre programme avec la *regex* suivante : $(a?)^*$ (et une entrée non vide quelconque) ?

   Remarque 3. *C'est évidemment un problème qu'il faudrait régler si on voulait réellement utiliser cette solution, mais nous allons en choisir une autre...*

◁

L'autre idée possible pour exécuter un automate non déterministe (avec ou sans ϵ-transitions) est de maintenir à jour l'ensemble des états dans lesquels on peut se trouver à un moment donné de la lecture de l'entrée. Un tel ensemble correspond exactement à un état de l'automate des parties : essentiellement, on explore et construit uniquement le chemin de l'automate des parties qui correspond à la lecture du mot.

Pour représenter un ensemble d'états, on utilise la structure suivante :

```
struct set {
    int length;
    int id;
    state_t **states;
};

typedef struct set set_t;
```

On fournit une fonction pour créer un ensemble vide de capacité et `id` données, et une pour libérer la mémoire associée :

```
set_t *empty_set(int capacity, int id){
    state_t **arr = malloc(capacity * sizeof(state_t*));
    set_t *s = malloc(sizeof(set_t));
    s->length = 0;
    s->id = id;
    s->states = arr;
    return s;
}
```

- Le champ `length` indique le cardinal de l'ensemble.
- Le champ `states` est un tableau de pointeurs vers des états. Sa longueur sera au moins égale à `length` mais peut être supérieure : dans ce cas, seules les valeurs des cases d'indice `0` à `length - 1` ont un sens, les autres peuvent être ignorées.
- Le champ `id` permet d'identifier l'ensemble de manière unique. Ce champ sera utilisé en combinaison avec le champ `last_set` de la structure `state` : pour un état `s`, la valeur de `s.last_set` sera égale au champ `id` de l'ensemble construit le plus récemment qui contient `s`. Si `s` n'appartient à aucun des ensembles construits jusqu'à maintenant, son champ `last_set` sera égal à `-1` (ce sera donc le cas initialement, en particulier).

▷ **Question 5.**

1. Ecrire une fonction `add_state` ayant la spécification suivante :

   Prototype :

   ```
   void add_state(set_t *set, state_t *s);
   ```

   Préconditions :

   - `set` est un pointeur valide vers une structure de type `set` ;
   - `set.length >= 0` ;
   - `set.states` est suffisamment grand pour contenir tous les états ;
   - `s` est soit le pointeur nul, soit un pointeur vers un état de l'automate de Thompson ;
   - les états présents dans `set` sont exactement les états `x` dont le champ `last_set` est égal au champ `id` de `set` (ce qui peut inclure l'état `s`).

   Postconditions :

   - l'état `s` a été ajouté à `set` (s'il n'y était pas déjà) ;
   - tous les états accessibles depuis `s` en n'utilisant que des ϵ-transitions l'ont également été (à nouveau, s'ils n'y étaient pas déjà) ;
   - les champs des différents objets ont été mis à jour pour conserver les invariants.

2. Ecrire une fonction `step` ayant la spécification suivante :

   Prototype :

   ```
   void step(set_t *old_set, char c, set_t *new_set);
   ```

   Préconditions :

   - `old_set` et `new_set` sont deux pointeurs valides (et non aliasés);
   - les tableaux `states` des deux ensembles sont suffisamment grands pour recevoir tous les états nécessaires.

Postconditions :

- `new_set` contient l'ensemble des états accessibles depuis les états de `old_set` en effectuant une transition étiquetée par `c`, plus éventuellement des ϵ-transitions ;
- l'identifiant de `new_set` est incrémenté de une unité par rapport à celui de `old_set` (et le champ `last_set` a été mis à jour en conséquence dans les états concernés).

3. Déterminer la complexité en temps de la fonction `step`.
4. Écrire une fonction `accept` qui prend en entrée un automate et une chaîne, et renvoie `true` ou `false` suivant que le mot (la chaîne) est reconnu ou pas. À nouveau, on considérera que le mot se termine juste avant le premier caractère `'\n` ou `'\0'` rencontré.

```
bool accept(nfa_t a, char *s, set_t *s1, set_t *s2);
```

Remarques 2.

Les deux arguments supplémentaires `s1` *et* `s2` *sont fournis pour éviter d'avoir à allouer des ensembles.*

On pourra supposer que tous les états de l'automate ont une valeur de `last_set` *strictement inférieure à* `s1->id` *au début de l'appel, et il faudra garantir que c'est toujours le cas à la fin de l'appel.*

5. Écrire une fonction `match_stream` ayant le même comportement que `match_stream_backtrack` mais utilisant la nouvelle méthode d'exécution de l'automate. On ne créera que deux `set_t` au total.

```
void match_stream(nfa_t a, FILE *in);
```

◁

Il reste une question à régler : jusqu'à présent, nous avons toujours fait en sorte, lorsque nous écrivions du C, de respecter une politique " zéro déchet ", c'est-à-dire de libérer toute la mémoire allouée avant de quitter le programme. Ici, ce n'est pas le cas : nous avons alloué chaque état de l'automate individuellement à l'aide d'un `malloc`, mais nous n'avons jamais fait les `free` qui y correspondent. D'un certain côté, ce n'est pas grave du tout : nous ne voulons libérer les états qu'immédiatement avant de quitter le programme. Dans ce cas précis, en production, on laisserait le système d'exploitation s'en charger (de manière automatique en sortie de programme), ce qui serait plus simple et marginalement plus efficace. Cependant, il y a quelque chose de problématique : si nous *voulions* libérer proprement la mémoire (par exemple parce qu'on souhaitait construire plusieurs automates successivement), nous aurions beaucoup de mal à le faire !

▷ **Question 6.**

1. On propose le code suivant pour libérer les états associés à un automate fini :

```
void free_accessible_states(state_t *q){
  if (q == NULL) return;
  free_accessible(q->out1);
  free_accessible(q->out2);
  free(q);
}

void free_automaton(nfa_t a){
  free_accessible(a.start);
}
```

Expliquer pourquoi cela ne fonctionne pas.

2. Proposer un algorithme permettant de réaliser cette opération, sans changer notre manière d'allouer les états (on ne demande pas de le programmer).
3. Une solution beaucoup plus simple est de ne pas utiliser du tout de `malloc`. Proposer une solution utilisant deux variables globales :

```
state_t* State_array;
int Nb_states;
```

Les seules modifications à apporter sont dans `new_state` et `main`.

◁