Dicionários / Tabelas de Símbolos I

Joaquim Madeira 25/05/2021

Ficheiro ZIP

- Está disponível no Moodle um ficheiro ZIP de suporte aos tópicos de hoje
- O tipo abstrato Hash Table usando Open Addressing
- Versões "simples", que permitem trabalho autónomo de desenvolvimento e teste

Sumário

- Recap
- Motivação
- Digital Search Trees Breve referência
- Prefix Trees Breve referência
- Hash Tables Tabelas de Dispersão
- Funções de Hashing
- Representação usando um array Open Addressing
- O TAD Hash Table (String, String)

Recapitulação



TAD Árvore Binária de Procura

- Conjunto de elementos do mesmo tipo
- Armazenados em-ordem
- Procura / inserção / remoção / substituição
- Pertença
- search() / insert() / remove() / replace()
- size() / isEmpty() / contains()
- create() / destroy()

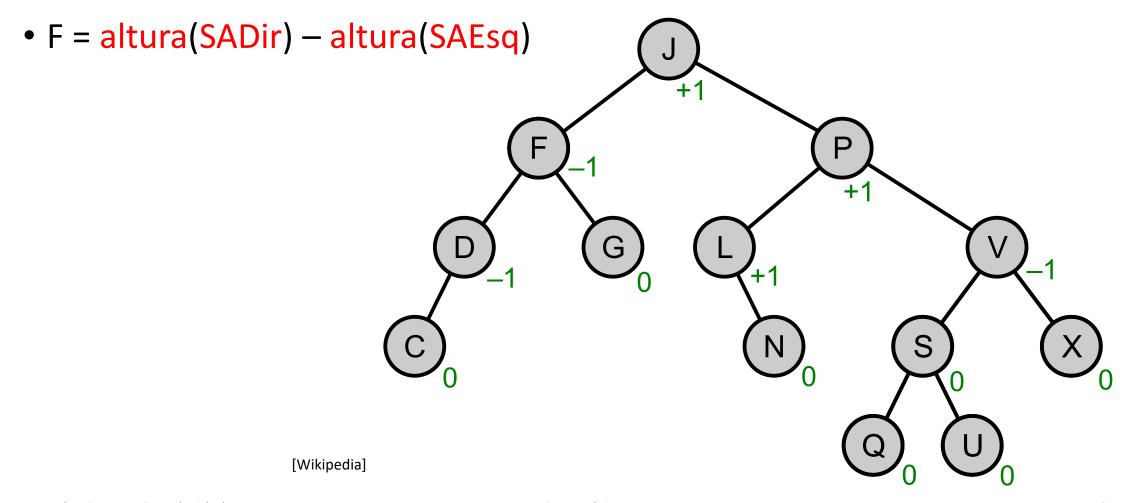
Lista ligada / Array ordenado / ABP

search	N	$\lg N$	h
insert	N	N	h = height of BST
min / max	N	1	(proportional to log N if keys inserted in random order)
floor / ceiling	N	$\lg N$	h
rank	N	$\lg N$	h
select	N	1	h
ordered iteration	$N \log N$	N	N [Sedgewick & Wayne]

Árvores equilibradas em altura

- Esforço computacional das operações habituais sobre ABPs depende do comprimento do caminho a partir da raiz da árvore
- Evitar que uma ABP tenha uma altura "exagerada", para assegurar um bom "comportamento" – Altura ε O(log n)
- O que fazer ?
- Assegurar que, para cada nó, a altura das suas duas subárvores não é "muito diferente" – Critério de equilíbrio

Fator de equilíbrio de um nó



Quando fazer? / Como fazer?

- Assegurar o critério de equilíbrio sempre que se adiciona ou remove um nó : F = -1, 0, +1
- Reposicionar nós / subárvores quando falha : F = -2, +2
- MAS, manter o critério de ordem da ABP !!
- 4 tipos de operações de rotação
- Basta fazer a verificação / rotações ao longo do caminho entre a raiz e o nó – traceback

Árvore AVL – Inserir + Equilibrar, se necessário

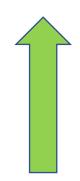
[Wikipedia]

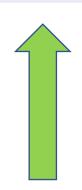
1ª experiência computacional

- Criar uma árvore vazia
- Inserir ordenadamente sucessivos números pares: 2, 4, 6, ...
- Procurar cada um desses números pares na árvore
- Contar o número de comparações efetuadas em cada nó
 - 1 ou 2 comparações por nó visitado

Procurar os sucessivos números pares

nós	Altura ABP	Nº médio comps	Altura AVL	Nº médio comps
5000	4999	5001	12	17,69
10000	9999	10001	13	19,19
20000	19999	20001	14	20,69
40000	39999	40001	15	22,19



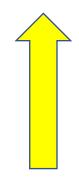


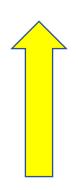
2ª experiência computacional

- Criar uma árvore vazia
- Inserir uma sequência de números aleatórios
- Procurar cada um desses números na árvore
- Contar o número de comparações efetuadas em cada nó
 - 1 ou 2 comparações por nó visitado

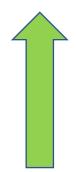
Procurar os sucessivos números aleatórios

nós	Altura ABP	Nº médio comps	Altura AVL	Nº médio comps
2500	27	19,64	12	16,18
5000	25	22,10	14	17,66
10000	30	25,72	15	18,85
20000	28	25,83	16	19,83
40000	32	26,73	16	20,91









Dicionários – Motivação

TAD Dicionário / Tabela de Símbolos

- Usar chaves para aceder a itens / valores
- Chaves e itens / valores podem ser de qualquer tipo
- Chaves são comparáveis
- MAS, não há duas chaves iguais !!
- Sem limite de tamanho / do número de pares (chave, valor)
- Chaves não existentes são associadas a um VALOR_NULO
- API simples / Código cliente simples

Aplicações

application	key	value		
contacts	name	phone number, address		
credit card	account number	transaction details		
file share	name of song	computer ID		
dictionary	word	definition		
web search	keyword	list of web pages		
book index	word	list of page numbers		
cloud storage	file name	file contents		
domain name service	domain name	IP address		
reverse DNS	IP address	domain name		
compiler	variable name	value and type		
internet routing	destination	best route		

[Sedgewick & Wayne]

Operações básicas

- Criar uma tabela vazia
- Registar um par (chave, valor) put
 - Se chave ainda não existe, adicionar (chave, valor)
 - Se já existe, alterar o valor
- Consultar o valor associado a uma chave get
- Verificar se uma chave pertence à tabela contains
- Limpar / destruir
- EXTRA: iterar sobre todas as chaves (em ordem)

Java

- HashMap<>
- TreeMap<>
- LinkedHashMap<>

- Diferenças ?
- System.out.println(myMap); // O que acontece ?

Algumas estruturas de dados – Fazer melhor?

implementation	guarantee		average case			ordered	key	
	search	insert	delete	search hit	insert	delete	ops?	interface
sequential search (unordered list)	N	N	N	½ N	N	½ N		equals()
binary search (ordered array)	1g <i>N</i>	N	N	1g <i>N</i>	½ N	½ N	~	compareTo()
BST	N	N	N	1.39 lg <i>N</i>	1.39 lg <i>N</i>	\sqrt{N}	~	compareTo()
red-black BST	2 1g N	2 1g N	2 1g N	1.0 lg <i>N</i>	1.0 lg <i>N</i>	1.0 lg <i>N</i>	V	compareTo()

[Sedgewick & Wayne]

Digital Search Trees

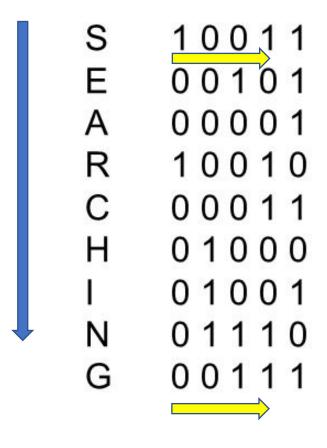
Ideias principais

- Árvore binária
- Armazenar um par (chave, valor) em cada nó, sem repetições
- Cada elemento é inserido como folha da árvore
- Diferença para as ABPs :
- Se a chave procurada for a chave registada num nó -> procura bem sucedida
- Caso contrário, examinar um bit da chave e procurar recursivamente ou na SAEsq ou na SADir

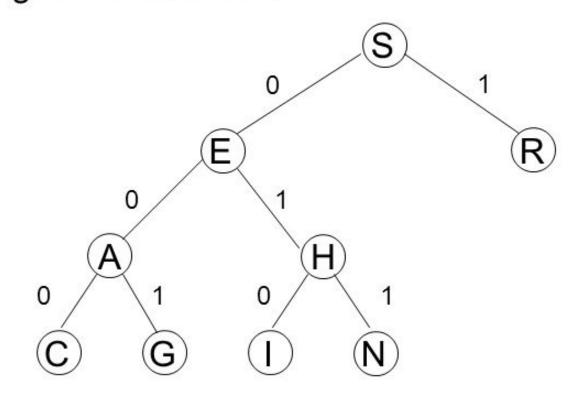
Ideias principais

- O caminho, a partir da raiz, para um dado nó é definido por sucessivos bits da sua chave
 -> procura + inserção
- bit = 0 -> avançar para a SAEsquerda
- bit = 1 -> avançar para a SADireita
- A organização de uma árvore depende da sequência de inserção dos seus elementos

Exemplo – Fazer a sequência de inserção



Digital Search Tree



Procura recursiva

```
Item* search(Node* p, Key searchKey, int i) {
  if (p == NULL) return NULL;
  if (equals(searchKey, p->item->key)) return p->item;
  if (bit(searchKey, i) == 0)
     return search(p->left, searchKey, i+1);
  return search(p->right, searchKey, i+1);
```

Procura recursiva

- O nº de nós visitados é, quando muito, igual ao nº de bits mais significativos necessários para distinguir a chave de procura das outras chaves
- MAS, a visita de cada nó implica a comparação usando toda a chave
- Pior Caso : O(n), n é o nº de bits da chave
- Caso Médio : O(log n), para chaves aleatórias

Tries — Prefix Trees

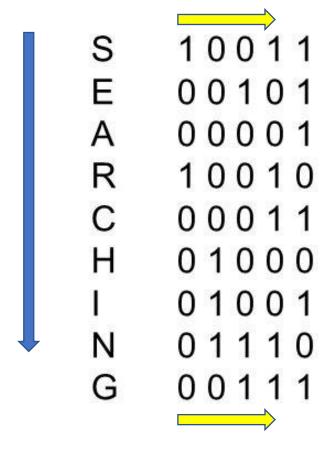
Ideias principais

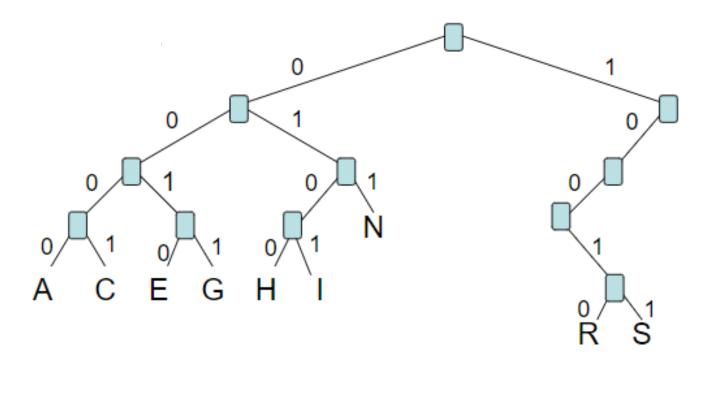
- Árvore binária
- Armazenar um par (chave, valor) em cada FOLHA, sem repetições
- Diferenças para as ABPs :
- Os nós não-terminais não contêm itens!!
- Em cada nó não-terminal, examinar um bit da chave e procurar recursivamente ou na SAEsq ou na SADir
- Se a chave procurada for a chave registada numa folha -> procura bem sucedida

Ideias principais

- Chaves de tamanho fixo; nenhuma chave é prefixo de outra
- O caminho, a partir da raiz, para cada folha é definido por sucessivos bits da sua chave
 -> procura + inserção
- bit = 0 -> avançar para a SAEsquerda
- bit = 1 -> avançar para a SADireita
- A organização de uma árvore NÃO depende da sequência de inserção dos seus elementos

Exemplo – Fazer a sequência de inserção





30

Procura recursiva

```
Item* search(Node* p, Key searchKey, int i) {
  if (p == NULL) return NULL;
  if (p->left == NULL && p->right == NULL )
    if (equals(searchKey, p->item->key)) return p->item;
    else return NULL;
  if (bit(searchKey, i) == 0)
     return search(p->left, searchKey, i+1);
  return search(p->right, searchKey, i+1);
```

Procura recursiva

- O nº de nós visitados é, quando muito, igual ao nº de bits mais significativos necessários para distinguir a chave de procura das outras chaves
- MAS, a visita de uma folha implica a comparação usando toda a chave
- Pior Caso : O(n), n é o nº de bits da chave
- Caso Médio : O(log n), para chaves aleatórias

Hash Tables — Tabelas de Dispersão

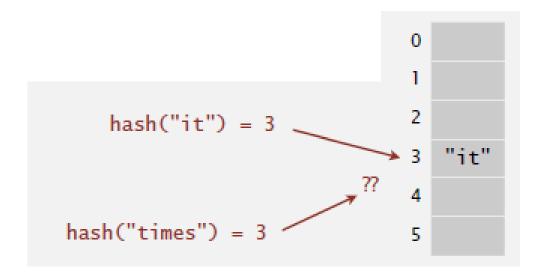
Tabelas de Dispersão

- Estrutura de dados para armazenar pares (chave, valor)
- Sem chaves duplicadas
- Sem uma ordem implícita !!
- MAS, com acesso rápido!!



Tabelas de Dispersão

- Armazenar itens numa tabela/array indexada pela chave
 - Índice é função da chave
- Função de Hashing: para calcular o índice a partir da chave
 - Rapidez !!
- Colisão: 2 chaves diferentes originam o mesmo resultado da função de hashing



[Sedgewick & Wayne]

35

Tabelas de Dispersão – Problemas

- Como calcular a função de hashing ?
 - Rapidez + simplicidade
- Como verificar se duas chaves são iguais ?
- Como resolver colisões ?
 - Método / estrutura de dados para armazenar itens com o mesmo valor de hashing
 - Rapidez !!
 - Memória adicional ?

Espaço de memória vs Tempo de execução

- Não há limitações memória : usar a chave diretamente como índice!!
- Não há restrições temporais : colisões resolvidas com procura sequencial
- MAS, o espaço de memória é limitado!!
- E pretendemos operações em tempo quase-constante, qualquer que seja a chave !!
- Como fazer ?

- Requisito : se x == y, então hash(x) = hash(y)
- Desejável : se x ≠ y, então hash(x) ≠ hash(y)
- Exemplos simples
- int hash(int x) { return x; }
- int hash(double x) { long bits = doubleToLongBits(x); // 32 to 64 bits return (int) (bits ^ logicalShiftRight(bits, 32)); }

```
int hash(char* s) {
                         int hash = 0;
                         for (int i = 0; i < strlen(s); i++)
                                hash = s[i] + (31 * hash);
                         return hash;
hash("call") = 3045982
                = 108 + 31 \times (108 + 31 \times (97 + 31 \times (99)))
                                                                        M. Horner
                = 99 \times 31^{3} + 97 \times 31^{2} + 108 \times 31^{1} + 108 \times 31^{0}
```

- Há muitas funções de hashing para diferentes aplicações
 - Que outras aplicações conhecem?
- Diferentes graus de complexidade
- Diferenças no desempenho computacional

Tabelas de Hashing : privilegiar a rapidez e o nº reduzido de colisões

Conversão para índices da tabela

- Índices da tabela entre 0 e M 1
- M é um número primo ou uma potência de 2
- Como fazer ?

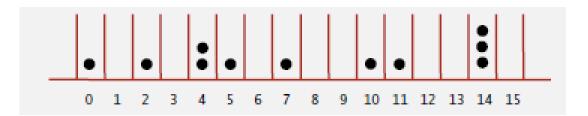
abs(hash(x)) % M

42

Equiprobabilidade

- Assume-se a equiprobabilidade !!
- Cada chave tem a mesma probabilidade de ser mapeada num dos índices (0 a M 1)

• O que acontece na prática?



• Lembram-se do Paradoxo do Aniversário?

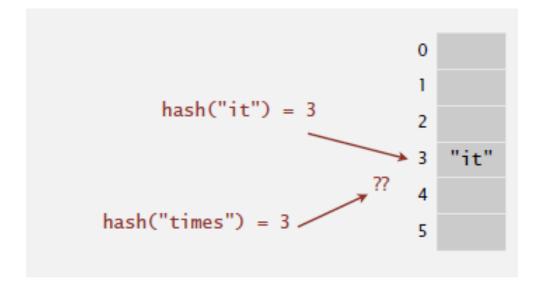
Hash Tables - Open Addressing

44

Colisões – Como proceder ?

- Duas chaves distintas são mapeadas no mesmo índice da tabela!!
- Colisões são "evitadas" usando tabelas de muito grande dimensão!!

- Como gerir de modo eficiente ?
- Sem usar "demasiada" memória!!



Open Addressing (IBM, 1953)

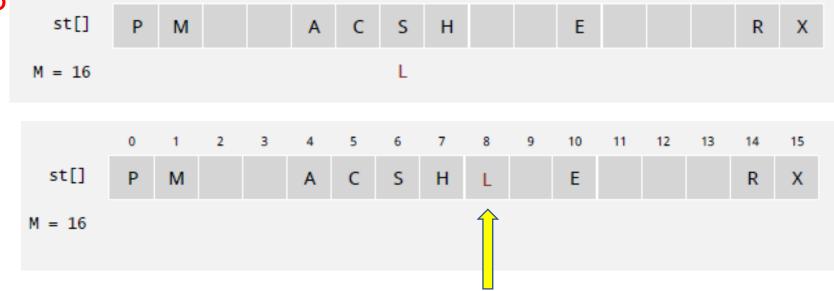
- Quando há uma colisão, procurar o espaço vago seguinte e armazenar o item – (chave, valor)
- Linear Probing Sondagem Linear
- O tamanho da tabela (M) tem de ser maior do que o número de itens (N) !!
- Quantas vezes maior ??

st[0] jocularly mullst[1] st[2] listen st[3] suburban mullst[30000] browsing linear probing (M = 30001, N = 15000)

Inserir na tabela – Linear Probing

- Guardar na posição i, se estiver disponível
- Caso contrário, tentar (i + 1) % M, (i + 2) % M, etc.

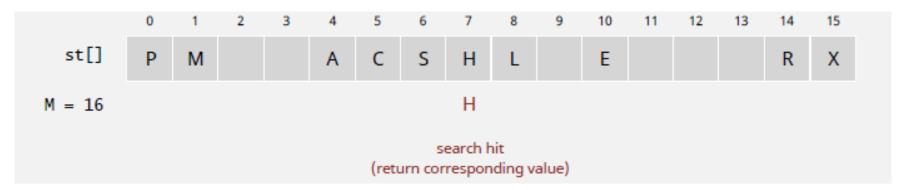
- Inserir L -> índice = 6
- Colisão!!
- ...



15

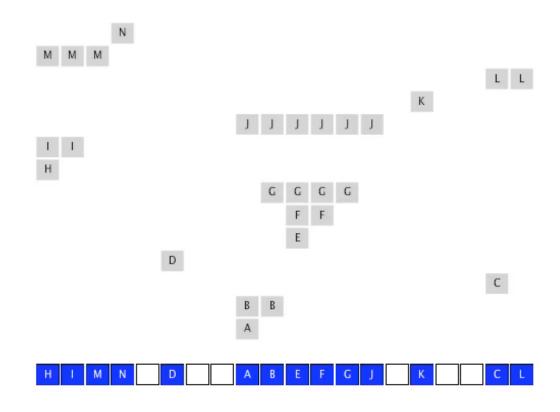
Procurar na tabela – Linear Probing

- Procurar na posição i
- Se estiver ocupada, verificar se as chaves são iguais
- Se forem diferentes, tentar em (i + 1) % M, (i + 2) % M, etc.
- Até encontrar a chave procurada ou chegar a um espaço vago
- Procurar H -> índice = 4
 -> 4 comparações



Problema – Clustering

- Cluster : bloco de itens contíguos
- Novas chaves são indexadas no meio de "grandes" clusters
- E os itens colocados no final dos clusters



[Sedgewick & Wayne]

49

Análise – Linear Probing – Knuth, 1963

- Fator de carga Load Factor $\lambda = N / M$
- Nº médio de tentativas para encontrar um item

$$1/2 \times (1 + 1/(1 - \lambda))$$
 -> 1.5, se $\lambda = 50\%$

$$-> 1.5$$
, se $\lambda = 50\%$

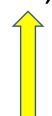
$$-> 3$$
, se $\lambda = 80\%$

• Nº médio de tentativas para inserir um item ou concluir que não existe

$$1/2 \times (1 + 1/(1 - \lambda)^2)$$
 -> 2.5, se $\lambda = 50\%$

$$-> 2.5$$
, se $\lambda = 50\%$

$$-> 13$$
, se $\lambda = 80\%$



Análise – Linear Probing

- M muito grande -> demasiados espaços vagos !!
- M "pequeno" -> tempo de procura aumenta muito!!
- Limiar habitual para o fator de carga : 50%
- Nº médio de tentativas para encontrar um item : 1,5 hit
- Nº médio de tentativas para inserir um item : 2,5 miss
- Como controlar ? RESIZING + REHASHING !!

51

Inserir na tabela – Quadratic Probing

- Guardar na posição i, se estiver disponível
- Caso contrário, tentar (i + 1) % M, (i + 4) % M, (i + 9) % M, etc.

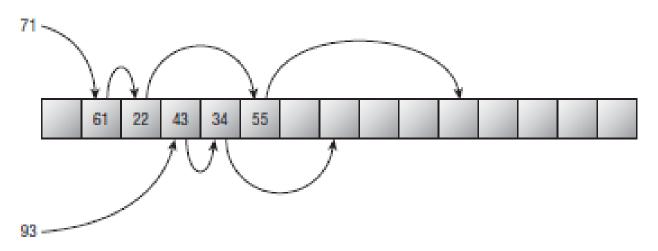


Figure 8-4: Quadratic probing reduces primary clustering.

[Stephens]

Resizing + Rehashing

- Objetivo : fator de carga ≤ 1/2
- Duplicar o tamanho do array quando fator de carga ≥ 1/2
- Reduzir para metade o tamanho do array quando fator de carga ≤ 1/8
- Criar a nova tabela e adicionar, um a um, todos os itens



Apagar um item (chave, valor)?



Lazy Deletion

- Marcar inicialmente todos elementos da tabela como livres
- Ao inserir um item, o correspondente elemento fica ocupado
- Ao apagar um item, marcar esse elemento da tabela como apagado
- Para que qualquer cadeia que o use não seja quebrada!!
- E se possa continuar a procurar uma chave usando probing
- Quando termina uma procura ?
- Ao encontrar a chave procurada ou um elemento marcado como livre

Eficiência

 A complexidade temporal de uma procura é limitada inferiormente por O(1) e superiormente por O(N)

- Pior Caso ?
- Sequência de colisões
- Toda a tabela tem de ser percorrida e cada elemento consultado para encontrar a chave procurada!!
- Ou concluir que não existe na tabela!!

Eficiência

implementation	worst-case cost			average case cost (after N random inserts)		key
	search	insert	delete	search hit	insert	interface
sequential search (unordered list)	N	N	N	½ N	N	equals()
binary search (ordered array)	$\lg N$	N	N	$\lg N$	½ N	compareTo()
BST	N	N	N	1.4 lg N	1.4 lg <i>N</i>	compareTo()
linear probing	N	N	N	3–5*	3–5 *	equals() hashCode()

* under the uniform hashing assumption

[Sedgewick & Wayne]

Joaquim Madeira

Exemplo

Hash Table (String, String)

TAD Hash Table

```
HashTable* HashTableCreate(unsigned int capacity, hashFunction hashF,
                           probeFunction probeF, unsigned int resizeIsEnabled);
void HashTableDestroy(HashTable** p);
int HashTableContains(const HashTable* hashT, const char* key);
char* HashTableGet(HashTable* hashT, const char* key);
int HashTablePut(HashTable* hashT, const char* key, const char* value);
int HashTableReplace(const HashTable* hashT, const char* key,
                     const char* value);
int HashTableRemove(HashTable* hashT, const char* key);
```

Estrutura de dados

```
struct _HashTableHeader {
   unsigned int size;
   unsigned int numActive;
   unsigned int numUsed;
   hashFunction hashF;
   probeFunction probeF;
   unsigned int resizeIsEnabled;
   struct _HashTableBin* table;
};
```

```
struct _HashTableBin {
  char* key;
  char* value;
  unsigned int isDeleted;
  unsigned int isFree;
};
```

Funções auxiliares para testes

```
unsigned int hash1(const char* key) {
   assert(strlen(key) > 0);
   return key[0];
}

unsigned int hash2(const char* key) {
   assert(strlen(key) > 0);
   if (strlen(key) == 1) return key[0];
   return key[0] + key[1];
}
```

61

Procura de uma chave

```
for (unsigned int i = 0; i < hashT->size; i++) {
 index = hashT->probeF(hashKey, i, hashT->size);
 bin = &(hashT->table[index]);
  if (bin->isFree) {
    // Not in the table !
   return index;
  if ((bin->isDeleted == 0) && (strcmp(bin->key, key) == 0)) {
    // Found it!
    return index;
```

Exemplo – M = 17 - N = 12

```
size = 17 | Used = 12 | Active = 12
 0 - Free = 0 - Deleted = 0 - Hash = 68, 1st index =
                                                        0, (December, The last month of the year)
 1 - Free = 1 - Deleted = 0 -
 2 - Free = 0 - Deleted = 0 - Hash = 70, 1st index = 2, (February, The second month of the year)
 3 - Free = 1 - Deleted = 0 -
 4 - Free = 1 - Deleted = 0 -
 5 - Free = 1 - Deleted = 0 -
 6 - Free = 0 - Deleted = 0 - Hash = 74, 1st index = 6, (January, 1st month of the year)
 7 - Free = 0 - Deleted = 0 - Hash = 74, 1st index =
                                                        6, (June, 6th month)
 8 - Free = 0 - Deleted = 0 - Hash = 74, 1st index =
                                                        6, (July, 7th month)
                                                        9, (March, 3rd month)
 9 - Free = 0 - Deleted = 0 - Hash = 77, 1st index =
10 - Free = 0 - Deleted = 0 - Hash = 77, 1st index = 9, (May, 5th month)
11 - Free = 0 - Deleted = 0 - Hash = 79, 1st index = 11, (October, 10th month)
12 - Free = 0 - Deleted = 0 - Hash =
                                      78, 1st index = 10, (November, Almost at the end of the year)
13 - Free = 1 - Deleted = 0 -
14 - Free = 0 - Deleted = 0 - Hash = 65, 1st index = 14, (April, 4th month)
15 - Free = 0 - Deleted = 0 - Hash = 65, 1st index = 14, (August, 8th month)
16 - Free = 0 - Deleted = 0 - Hash = 83, 1st index = 15, (September, 9th month)
```

Tarefas

- Analisar as funções desenvolvidas
- E o simples programa de teste