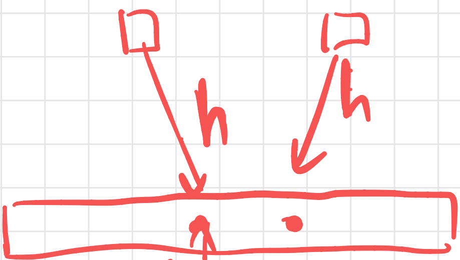
Хеш-Таблицы

```
a = dict()
a[5] = 6
a[3] = 1
print(a[3])
```

$O(1)$

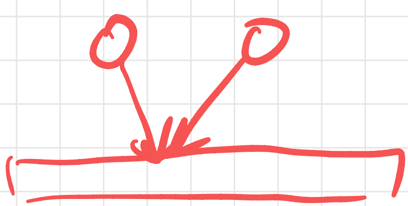


n ключей

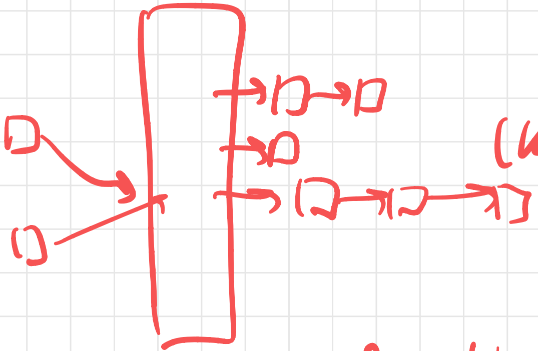
$(key_i, value_i)$ m ячеек

$h: keys \to \{0, \dots, m-1\}$

Коллизия – это делать



Separate Chaining



(key, value)

find(key)

put(key, value)

```python
def find(key):
    i = Hash(key)

    for (k, v) in Lists[i]:
        if k == key:
            return v

    raise "Not Found"

def put(key, new_value):
    i = Hash(key)

    for (j, (k, v)) in enumerate(Lists[i]):
        if k == key:
            Lists[i] = (k, new_value)
            return

    Lists[i].append((key, new_value))
```

время работы:

$$\alpha = n/m$$

1) пусть h ведёт себя дост. случ. $\Rightarrow$

$$T(\text{find}) = O(1+\alpha)$$
$$T(\text{put}) = O(1+\alpha)$$

2) $\alpha = n/m$ put $n{+}{+}$

$$m = \frac{n}{2} \to m' = n$$

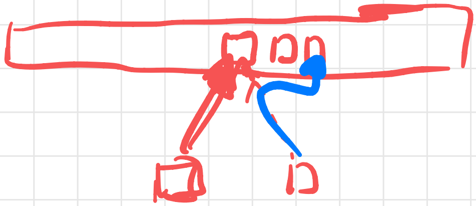
n

$| \; | \; | \; | \; | \; | \; | \; | \; O(n)$

$n \quad +n \quad 2n$

$m = \frac{n}{2} \qquad m' = n$

Open Addressing



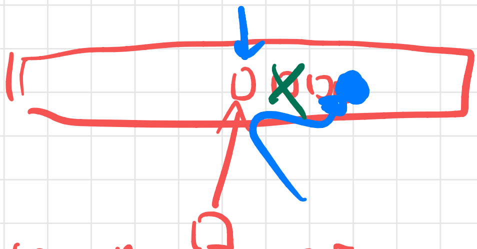
(key, value)

положить вперёд куда не занято.

```
put(k, v)
    i = hash(k)
    while keys[i] ≠ None:
        if keys[i] = k
            value[i] = v
            return
        else
            i++
    keys[i] = k
    values[i] = v
```

```
find(k)
    i = hash(k)
    while keys[i] ≠ None
        if keys[i] == k:
            return values[i]
        i++
    return Not found
```

delete(k)



удалим значение нужно

не просто keys[i] = None

values[i] = None

keys[i] = k

values[i] = "deleted"

особое значение.

Заметим: $\textcircled{1} \to$ true $\quad p$

$\quad\quad\quad\quad\;\; \to$ false $\quad 1-p$

Вопрос: сколько бросков в среднем нужно сделать до первого true

Д-во: $E[X] = p\cdot 1 + (1-p)(1+E[X])$

$$pE[x] = 1$$

$$E[x] = 1/p.$$

$$p = 1-\alpha.$$

В open addressing

$$T(find) = O(1 + 1/(1-\alpha))$$

$$T(put) = O(1 + 1/(1-\alpha))$$

примерно

Переполнение.

$\alpha \geq \frac{1}{2} \qquad m \to 2m$

Построение Хеш функций

Идея: $h(k)$ не зависит от $h(k_1), \ldots h(k_t)$

$$h(x) = x \,\%\, m$$

— хорошая или нет?

10, 15, 200, 42

$$\{0, m, 2m, 3m, \ldots\}$$

Универсальное Семейство

$$h \in U(\mathcal{H})$$

↑ семейство хеш функций.

Выбирать h случайно из $\mathcal{H}$.

$h \in \mathcal{H}:$

$$h: \underset{\{0, \ldots, p-1\}}{\underset{\shortparallel}{Keys}} \to \{0, \ldots, m-1\}$$

<u>пример</u>

$$\mathcal{H} = \left\{ h_{a,b} \;\middle|\; \begin{array}{l} a \in \{1, \ldots, p-1\}, \\ b \in \{0, \ldots, p-1\} \end{array} \right\}$$

$$h_{a,b}(k) = [(a \cdot k + b) \,\%\, p] \,\%\, m.$$

<u>Def</u> $\mathcal{H}$ — универсальное

$\forall k, l \in Keys$, $k \neq l$

$$\underset{h \in \mathcal{H}}{P}[h(k) = h(l)] \leq \frac{1}{m}.$$

Рассмотриваем Separate Chaining.

Утв: $\mathop{E}_{h}[len(L_{h(k)})] \le 1 + \alpha$ — ключ k

если k живёт в Х.Т.

$\mathop{E}_{h}[len(L_{h(k)})] \le \alpha$

иначе

Д-во:

$$E[len(L_{h(k)} \setminus \{k\})] \le \alpha$$

$$E[len(L_{h(k)} \setminus \{k\})] = \sum_{\substack{l \in HashTable \\ l \ne k}} P[h(l) = h(k)] \le$$

$$\le \underbrace{(n-1)}_{\text{кол-во эл-тов в Х.Т.}} \cdot \frac{1}{m} \le \alpha$$

□

Keys $\in \{0, \dots, p-1\}$

$$\mathcal{H} = \{h_{a,b} \mid a \in \{1, \dots, p-1\},\ b \in \{0, \dots, p-1\}\}$$

$$h_{a,b}(k) = [(a \cdot k + b) \% p] \% m$$

$\forall k, l \in$ Keys, $k \neq l$: $\quad P_{h \in \mathcal{H}}[h(k) = h(l)] \leq \frac{1}{m}$

Д-во. Универсальности.

Пусть k, l — два разл. ключа.

$$x = (ak + b) \% p$$
$$y = (al + b) \% p$$

Несовпадение: $x \neq y$.

Пусть $x = y$.

Тогда $ak + b \equiv al + b \mod p$

$a(k - l) \equiv 0 \mod p. \Rightarrow k = l$

Пусть k и l фикс.

$(a, b) \to (x, y)$

$x \neq y$

$(x, y) \to (a, b)$

$x \neq y$

$x = (ak + b) \,\%\, p$

$y = (al + b) \,\%\, p$

$x - y = a(k - l) \mod p$

$a = (x - y) \cdot (k - l)^{-1} \mod p$

$b = x - ak \mod p$

$(a, b) \leftrightarrow (x, y)$

(a, b) — выбрали случайно

$\Updownarrow$

(x, y) выбрали случайно

$$\underset{a,b}{P}\left[h(k)=h(l)\right] = \underset{\substack{x \in U\{0,\dots,p-1\} \\ y \in U(\{0,\dots,p-1\}\setminus x)}}{P}\left[h(k)=h(l)\right] =$$

$$h(k) = x \,\%\, m$$
$$h(l) = y \,\%\, m$$

Давайте x - фикс.

$$\underset{y \neq x}{P}\left[y \,\%\, m = x \,\%\, m\right] \leqslant$$

$$\leqslant \frac{\lceil p/m \rceil - 1}{p-1} \leqslant$$

$$\le \frac{\frac{p+m-1}{m} - 1}{p-1}$$

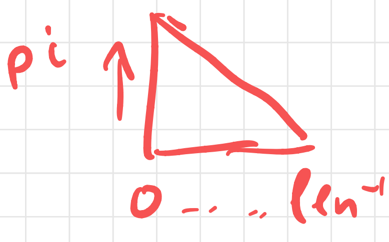
$$= 1/m \qquad \square.$$

Хеширование строк

(Строковые алгоритмы)

$S: \; S_0 \ldots S_{len-1}$

$$hash(S) = S_0 p^{len-1} + S_1 p^{len-2} + \ldots + S_{len-1} \cdot p^0$$



p - некоторое фикс. число

или выбрать случайно
или какое-то конст.

Наблюдение: p достаточно велико, то $hash(s) = hash(t) \Rightarrow$

$$\Rightarrow s = t$$

В реальности $hash(s) \,\%\, M$.
$\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad ||$
нек-ая большая конст

$s = t \Rightarrow hash(s) = hash(t) \mod M$

$s \neq t \Rightarrow$ скорее всего $hash(s) \neq hash(t) \mod M$.

$$hash(s) = s_0\, p^{len-1} + s_1\, p^{len-2} + \ldots + s_{len-1} \cdot p^0 \ \underline{\bmod M}$$

Нововведение

$$hash(s\,c) = p \cdot hash(s) + c \cdot p^0$$

пусть $s = s_0 \ldots s_{n-1}$

пусть $h_i = hash(s_0 \ldots s_{i-1})$

вычислим $h_0 \ldots h_n$

$h_0 = 0$

for $i = 0..n-1$

$$h_{i+1} = (p\, h_i + s_i) \,\%\, M$$

$hash(s_l \ldots s_r) = ?$

$h_{r+1} = hash(s_0 \ldots s_r)$

$h_l = hash(s_0 \ldots s_{l-1})$

Напоминание

$$hash(s_l \ldots s_r) = hash(s_0 \ldots s_r) - hash(s_0 \ldots s_{l-1}) \cdot p^{r-(l-1)}$$

$$= h_r - h_l \cdot p^{r+1-l}$$

Давайте считать, что $p^0 \ldots p^n$ тоже предподсчитаем

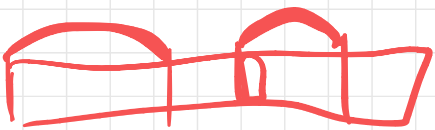
```
pows[0] = 1
for i = 1..n
    pows[i] = p · pows[i-1] % M
```

Задача:

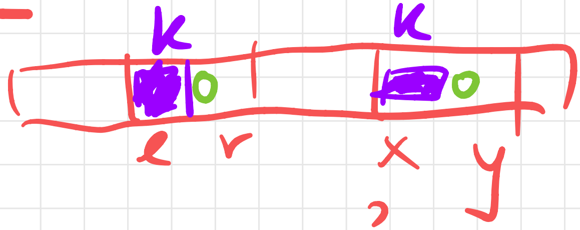
$$S_l \ldots S_r \overset{?}{=} S_x \ldots S_y$$

$$\Rightarrow O(1)$$

Например Z-функция за $O(n \log n)$



Задача



$$S_l \ldots S_r \;\overset{?}{\lesseqgtr}\; S_x \ldots S_y$$

$k = lcp(S_l..S_r, S_x..S_y)$

$k =$ бин. поиск.

С какой вероятностью коллизия хеша?

утв: $S_0 .. S_{n-1}$

$S_i \in$ случайные, то

$hash(S_0 .. S_{n-1}) \in$ случ. величина равномерно в $0..M-1$.

$$\underset{s,t}{P}[hash(s) = hash(t)] = 1/M.$$

s, t - случ. строки

<u>Утв:</u> $hash(s) = \sum s_i p^{len-1-i}$ —

— полином над p.

Пусть s и t две фикс. строки, $s \neq t$.

$hash(s) = hash(t) \Leftrightarrow$

$$\Leftrightarrow \underbrace{hash(s) - hash(t)} \equiv 0 \mod M$$

$f = \sum a_i p^i$ — полином. над p

$$hash(s) - hash(t) \equiv 0 \mod M$$

$\Leftrightarrow p$ — корень мод. многочлена f.

Пусть p — выбрано случ. в $[0..M-1]$

$$P_p[p \text{ корень } f] \leq \frac{\deg f}{M}$$

$$\Rightarrow P[s \text{ коллизится с } t] \leq \frac{\max(|s|, |t|) - 1}{M}$$

Лемма: Кол-во корней у многочлена f над $\mathbb{F}_p$ не более чем $\deg f$.

Резюме

$n = 10^5$ $\qquad M = 10^{18}+3$

$P[\text{коллизии}] \le n/M$.

Perfect Hashing.

$x_1, \ldots, x_n$

Н.Т. $x_i \to i$

$x_i \to value_i = data[i]$

$\Theta(n)$ памяти получить статическую хэш-таблицу работающую за $O(1)$.

0 -- 10^9

dict["abc"]

hash ("abc") → 213