- Given a sequence of strings $S_1 \dots S_n$ (maybe duplicate)
  => filter to keep just the distinct one (1 copy per string)
when memory size M; relation between n, M, total length of the strings

1) Bloom Filter: in this case $n = \Theta(M)$

$m = c \cdot n < M$
↳ size of Bloom F

$k = \frac{m}{n} \ln 2 = \frac{c \cdot n}{n} \ln 2 = c \ln 2$

=> - SCAN THE STRINGS S_i and ∀ of them set

// take a string and check, if all pos are 1 strings => belongs to set => we don't do anything

check if $BI[h_j(S_i)] = 1 \quad \forall j = 1, k$

$\forall j = 1, \dots, k$

$\forall i = 1, \dots n$

else print S_i

od

$\forall j = 1 \dots k$ set

// in groups $BI[h_j(S_i)] = 1$

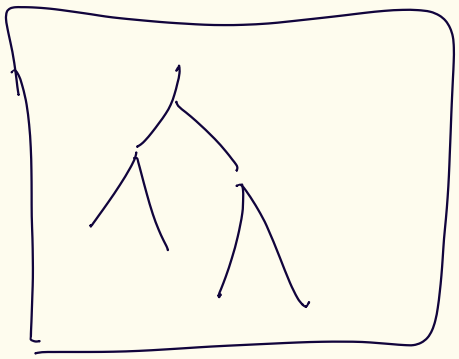
Cost: total lenght of the strings

$O\left(\frac{N}{B}\right)$ Los => It's a scan and all checks are done in M

← insertion ← scan

time $O(n \cdot k + N)$

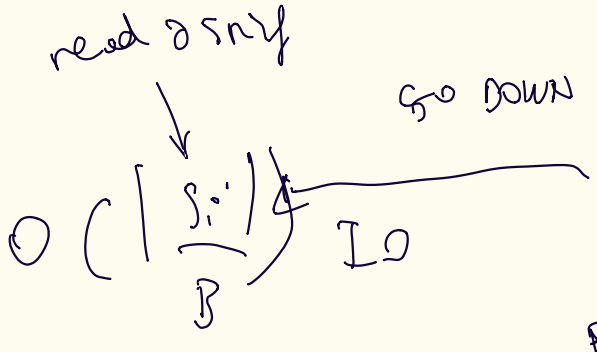
TRIE if the total length of the distinct strings
fits in M => < M



scan on disk and check if it in the TRIE

BETTER => PATRICIA TRIE: depends on the # of strings + their total length

string S_i
=> do a patricia tree search to know if
$S_i \in PT$ if present, do nothing
=> if is not present
=> print
insert S_i in PT



BUT $n < M$ => partition the rest Atts in M

BEST SOLUTION => hash

$\forall s_i \dots \rightarrow h(s_i) \xrightarrow{\text{create pair}} \langle h(s_i), i \rangle$

and use an external memory sorter

=> SORT

=> keep i of the first after sort

SPECIAL BF

7 celle $m = 7$

insert (1 2 3 4)

$h_1 : x \bmod 7$

$h_2 : 3x \bmod 7$

Key	h_1	h_2
1	2	3
2	4	6
3	6	2
4	1	5

0	1	2	3	4	5	6
0	1	2	3	1	1	2

METTO UN 1 ogni volta che entro in una cella

Query (5)

$h_1(5) = 10 \bmod 7 = 3$

$h_2(5) = 15 \bmod 7 = 1$

=> questo è errore

ex 5 is NOT $\in$ Set of keys

Il numero ottimale di funzioni di hash per il Bloom filter di size m and n keys

$$\Rightarrow k = \frac{m}{n} \ln 2$$

probabilità di avere una cella = 0

$$e^{-\frac{kn}{m}} \simeq e^{-\frac{\frac{m}{n} \ln 2 \cdot n}{m}} = e^{-\ln 2} =$$

$$= \frac{1}{e^{\ln 2}} = \frac{1}{2}$$