Text searching

$T[1..n]$ — text

$P[1..m]$ — pattern

Is P a substring of T? If so, where?

```
AlmostBruteForce(T[1..n], P[1..m]):
  for s ← 1 to n − m + 1
      equal ← True
      i ← 1
      while equal and i ≤ m
          if T[s + i − 1] ≠ P[i]
              equal ← False
          else
              i ← i + 1
      if equal
          return s
  return None
```

shift

$O(mn)$ time

T = AAAA.---AAAA

P = AAA--AB

Treat strings as integers

$$P = \sum_{i=0}^{m-1} 10^i P[i]$$

$$t_s = \text{num value of } T[s \ldots s+m-1] = \sum_{i=0}^{m-1} 10^i \cdot T[s+i]$$

$$t_{s+1} = 10\left(t_s - 10^{m-1} T[s]\right) + T[s+m]$$

```
NUMBERSEARCH(T[1..n], P[1..m]):
  σ ← 10^(m−1) mod q
  p ← 0
  t_1 ← 0
  for i ← 1 to m
      p ← (10·p + P[i]) mod q
      t_1 ← (10·t_1 + T[i]) mod q
  for s ← 1 to n − m + 1
      if p = t_s          ← IF P = T[s...s+m-1]
          return s
      t_{s+1} ← (10·(t_s − σ·T[s]) + T[s+m]) mod q
  return NONE
```

Time: $O(n + Fm) = O(n + nm/q) = O(n+m)$

F ↑ #False matches

If the mod q values are "random" — They aren't.

$E[F] = n \cdot \frac{1}{q}$ $\Rightarrow$ choose $q > n$

```
hp ← h(P)    // O(m) time
for s ← 1 to n-m
    hts ← h(T[s..s+m-1])    // O(m) time
    if hp = hts
        compare brute force
```

Rolling hash / Sliding hash

Rabin/Karp

Zobrist hashing

hash chess positions

$hashpiece_1 \oplus hashpiece_2 \oplus \cdots$
$\oplus hashpiece_1$
$\oplus hashpiece_1'$

```
KarpRabin(T[1..n], P[1..m]:
  q ← a random prime number between 2 and ⌈m² lg m⌉
  σ ← 10^(m−1) mod q
  p̃ ← 0
  t̃_1 ← 0
  for i ← 1 to m
    p̃ ← (10 · p̃ mod q) + P[i] mod q
    t̃_1 ← (10 · t̃_1 mod q) + T[i] mod q

  for s ← 1 to n − m + 1
    if p̃ = t̃_s
      if P = T_s        ⟨⟨brute-force O(m)-time comparison⟩⟩
        return s
    t̃_{s+1} ← (10 · (t̃_s − (σ · T[s] mod q) mod q) mod q) + T[s + m] mod q
  return NONE
```

$\Theta(m^2)$ choices

Prime # Theorem: there are $\Theta(N/\log N)$ primes less than N.

Lemma: Any integer x has $O(\log x)$ prime factors.

$\Pr[\tilde{p} = \tilde{t}_s] \leq O(1/m)$

$p < 10^m$ so p has $O(m)$ prime factors

$t_s < 10^m$ —— $O(m)$ prime factors

$|p - t_s| < 10^m$ —— " " "

$O(m)$ of the $\Theta(m^2)$ primes q cause a collision

$E[F] \leq \frac{n}{m}$ $\rightarrow E[\text{Time}] = O(n + E[F]m) = O(n+m)$

```
CarterWegmanKarpRabin(T[1..n], P[1..m]:
  q ← an arbitrary prime number larger than m^2
  b ← Random(q) − 1          ⟨⟨uniform between 0 and q − 1⟩⟩
  σ ← b^(m−1) mod q
  p̃ ← 0
  t̃_1 ← 0
  for i ← 1 to m
      p̃ ← (b · p̃ mod q) + P[i] mod q
      t̃_1 ← (b · t̃_1 mod q) + T[i] mod q

  for s ← 1 to n − m + 1
      if p̃ = t̃_s
          if P = T_s          ⟨⟨brute-force O(m)-time comparison⟩⟩
              return s
      t̃_{s+1} ← (b · (t̃_s − (σ · T[s] mod q) mod q) mod q) + T[s + m] mod q
  return None
```

Treat $P[1..m]$ as coeffs of poly of degree $m-1$ evaluate at random b.

$\Pr[\tilde{p} = \tilde{t}_s] < \frac{m}{m^2} = 1/m$

Any poly of degree m has $\leq m$ roots.

$\longrightarrow$ $O(m+n)$ exp. time