

FÖRELÄSNING

TILLÄMPNING: STAVNINGSKONTROLL

- BLOOMFILTER FÖR ORDUPPSLAGNING
- HANTERING AV SAMMANSÄTTNINGAR
- HANTERING AV BÖJNINGSFORMER
- GENERERING AV RÄTTELSEFÖRSLAG
- OPTIMERING

BLOOMFILTER

DATASTRUKTUR FÖR SNABB KOLL AV MÄNGDTILLHÖRIGHET I STORA MÄNGDER.

BARA TVÅ OPERATIONER:

- INSERT(x) STOPPAR IN x I MÄNGDEN
- ISIN(x) KOLLAR OM x ÄR I MÄNGDEN

EGENSKAPER:

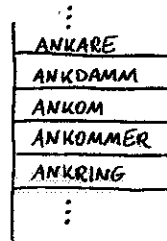
- BÅDA OPERATIONERNA HAR TIDSKOMPLEXITET $O(1)$
- LITET MINNESUTRYMME KRÄVS
- ELEMENTEN LAGRAS INTE I KLARTEXT
- LITEN SANNOLIKHET FÖR ATT ISIN(x) ÄR SANT TROTS ATT x INTE INGÅR I MÄNGDEN.

TILLÄMPNINGSEXEMPEL:

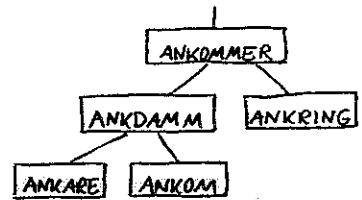
- LAGRING AV ORDLISTAN I STAVNINGSKONTROLL
- LAGRING AV BESÖKTA LÄNKAR I WEBBLÄSARE.

DATASTRUKTURER FÖR ORDLISTAN

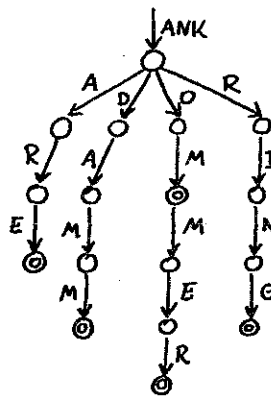
1. SORTERAD VEKTOR



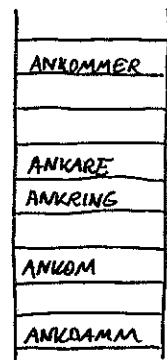
2. BINÄRT TRÄD



3. AUTOMAT



4. HASHTABELL



STAVA: BLOOMFILTER

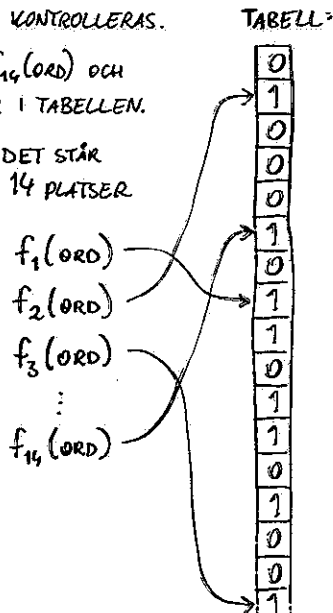
ORDLISTAN LAGRAS SOM ETT BLOOMFILTER, DVS MED HASHNING I EN TABELL MED ENDAST NOLLOR OCH ETTOR.

Använd 14 HASHFUNKTIONER

$f_i: \text{ORD} \rightarrow \text{PLATS I HASHTABELL}$
 $1 \leq i \leq 14$

GIVET ORD SOM SKA KONTROLLERAS.

- BERÄKNA $f_1(\text{ORD}), \dots, f_{14}(\text{ORD})$ OCH UNDERSÖK DESSA PLATSER I TABELLEN.
- ACCEPTERA ORDET OM DET STÅR ETTOR PÅ ALLA DESSA 14 PLATSER



SANNOLIKHET FÖR FEL

LÅT n VARA ANTALET ORD I ORDLISTAN
 m — " — BITAR I VEKTORN
 k — " — HASHFUNKTIONER

$$P[\text{EN VISS BIT SÄTTS TILL SANT} \mid \text{VID EN VISS HASHNING}] = \frac{1}{m}$$

$$P[\text{EN VISS BIT SÄTTS INTE TILL SANT} \mid \text{VID EN VISS HASHNING}] = 1 - \frac{1}{m}$$

$$P[\text{EN VISS BIT ÄR FORTFARANDE FALSK EFTER } kn \text{ HASHNINGAR}] = \left(1 - \frac{1}{m}\right)^{kn}$$

$$P[\text{EN VISS BIT ÄR SANN EFTER } kn \text{ HASHNINGAR}] = 1 - \left(1 - \frac{1}{m}\right)^{kn}$$

$$P[\text{K SLUMPVISA UPSLAGNINGAR I VEKTORN GER ALLA SVARET SANT}] = \underbrace{\left[1 - \left(1 - \frac{1}{m}\right)^{kn}\right]^k}_{f(k)}$$

$f(k)$ MINIMERAS AV

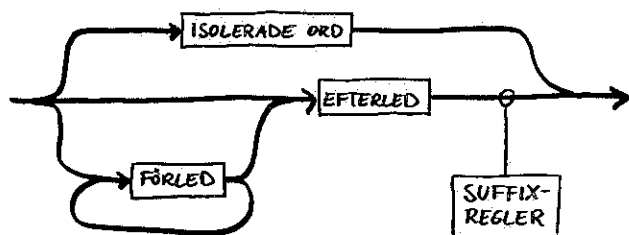
$$k = -\frac{\ln 2}{n \cdot \ln\left(1 - \frac{1}{m}\right)} \approx \ln 2 \cdot \frac{m}{n} \approx 0,69 \frac{m}{n}$$

OCH GER DÅ $f(k) = 2^{-k}$

STAVA: HANTERING AV SAMMANSÄTTNINGAR

TRE ORDLISTOR:

- ISOLERADE ORD (TEX ÖMSOM, ELLER)
- FÖRLED (TEX FOTBOLLS, MEDIE, STAVNING)
- EFTERLED (TEX KUNSKAP, STÖRT, LITA, MEDIUM)



STAVA: SUFFIXREGLER

VI LAGRAR VISSA BESTÄMDA FORMER AV ORDEN I ORDLISTAN OCH INFÖR SUFFIXREGLER AV TYPEN

-ORNA ← -A, -AN, -OR

VILKET BETYDER "OM ORDEN X_A, X_{AN} OCH X_{OR} FINNS I ORDLISTAN SÅ SKA OCKSÅ X_{ORNA} GODKÄNNAS"

EXEMPEL: DOCKA, DOCKAN, DOCKOR ⇒ DOCKORNA

BARA TVÅ AV 1585 ORD SOM PASSAR IN PÅ DENNA REGEL ÄR FEL.

EXEMPEL (FEL): DEKA, DEKAN, DEKOR ⇒ DEKORNA

REGLERNA UTFORMAS SÅ ATT DENNA TYP AV FEL MINIMERAS.

EXEMPEL PÅ RÄTTSTAVNING

FELSTAVAT ORD: STRUTN

- KASTA OM INTILLIGGANDE BOKSTÄVER:
TSRUTN, SRIUTN, STURTN, STRUVN, STRUNT
- SÄTT IN EN EXTRA BOKSTAV:
... ESTRUTN, SETRUTN, STERUTN, STREUTN,
STRUETN, STRUTEN, STRUTNE, ...
- TA BORT EN BOKSTAV:
JRUTN, SRUTN, STUTN, STRJN, STRUN, STEUT.
- ERSÄTT EN BOKSTAV:
... YTRUTN, ... SXRUTN, ... STWUTN,
... STRVTN, ... STRUUM, ... STRUTS, ...

PROBLEM: VÄLDIGT MÅNGA ORD MÅSTE GENERERAS, VARAV MÅNGA ÄR HELTOKIGA.

EXEMPEL: FELSTAVAT 10-BOKSTAVSORD GER 618 ORD

OPTIMERING AV HASHNING

OBSERVATION:

NÄSTAN ALL TID AV RÄTTSTAVNINGEN TILLBRINGAS I HASHNINGSFUNKTIONEN.

FÖRSÖK OPTIMERA HASHNINGEN!

URSPRUNGLIGA HASHFUNKTIONER:

$$f_i(w) = \sum_j k_{ij} \cdot w[j] \text{ mod } p_i$$

FÖR ETT ORD MED n BOKSTÄVER GÖRS

- n MULTIPLIKATIONER
- n MODULOBERÄKNINGAR
- $n-1$ ADDITIONER
- $3n$ INDEXERINGAR

VID VARJE HASHFUNKTIONSBERÄKNING.

MODULOBERÄKNING KRÄVER DIVISION OCH TAR LÄNGST TID AV DESSA OPERATIONER.

GRAFOTAKTISK TABELL

TABELL SOM INNEHÅLLER ALLA 4-BOKSTAVS-FÖLJDER (FYRGRAM) SOM FÖREKOMMER I SPRÅKET.

EXEMPEL: STRUET, RUTNET, SXRUET

TA OCKSÅ HÄNSYN TILL ORDBÖRJAN OCH ORDSLUT.

I SVENSKA SPRÅKET FINNS 6% AV ALLA FYRGRAM

TVÅ ANVÄNDNINGSMÖJLIGHETER:

1. ALTERNATIV TILL BLOOMFILTER FÖR ATT AVGÖRA OM ETT ORD FINNS I SPRÅKET
2. HITTA VAR I ETT ORD EN FELSTAVNING FÖREKOMMER

EXEMPEL: STRUTN

ALLA FYRGRAM UTOM UTN ■ FINNS I T.

ALLTSÅ MÅSTE FELSTAVNINGEN VARA BLAND DOM TRE SISTA BOKSTÄVERNA I ORDET.

MODULOBERÄKNING MED FLYTTAL

$$\begin{aligned} x \text{ mod } p & \text{ RESTEN DÅ } x \text{ DIVIDERAS MED } p. \\ & = x - \left\lfloor \frac{x}{p} \right\rfloor \cdot p \quad \text{Låt } q = \frac{x}{p}. \\ & = x - \lfloor x \cdot q \rfloor \cdot p \end{aligned}$$

FÖRBEHANDLA GENOM ATT BERÄKNA $\frac{1}{p_i}$ FÖR ALLA HASHFUNKTIONER.

DÄREFTER KAN EN MODULOBERÄKNING ERSÄTTAS AV

- EN OMVANDLING HETAL \rightarrow FLYTTAL
- EN FLYTTALS MULTIPLIKATION
- EN OMVANDLING FLYTTAL \rightarrow HETAL
- EN HETALS MULTIPLIKATION
- EN HETALS SUBTRAKTION