

STAVNINGSKONTROLL

VIGGO KANN, NADA

- RÄTTSTAVNINGSPROBLEMET
- PROGRAMMET STAVA
 - BLOOMFILTER
 - HANTERING AV SAMMANSÄTTNINGAR
 - HANTERING AV BÖJNINGAR
 - RÄTTELSEFÖRSLAG
 - OPTIMERING
 - KONSTRUKTION AV SUFFIXREGLER
- SKYDD AV ORDLISTAN

RÄTTSTAVNINGSMETODER

LINGVISTISKA

- FÖRSÖK FÖRSTÅ TEXTEN
- LOKALISERA ORIMLIGHETER
- ÅTGÄRDA DESSA SÅ ATT TEXTEN BLIR RIKTIG

MED HJÄLP AV T.EX.

- ORDLISTOR MED ORDKLASS-MÄRKNING
- BÖJNINGSGREGLER
- GRAMMATIK FÖR SPRÅKET

STATISTISKA

- FÖRSÖK INTE FÖRSTÅ TEXTEN
- HITTA ORD SOM INTE FINNS
- BYT UT DESSA ORD PÅ TROLIGASTE SÄTT

MED HJÄLP AV T.EX.

- ORDLISTOR
- FREKVENSORDLISTOR
- STATISTISK INFORMATION OM STAVFEL

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.

STAVA: BLOOMFILTER

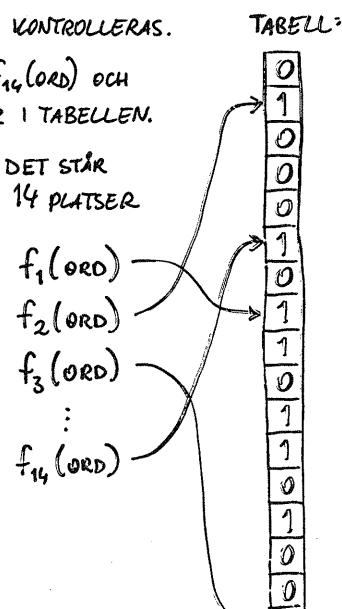
ORDLISTAN LAGRAS SOM ETT BLOOMFILTER, DVS MED HASHNING I EN TABELL MED ENDAST NOLLOR OCH ETTOR.

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



PROBLEM VID SAMMANSÄTTNING

TID + NÖD → TIDSNÖD } BINDEBOKSTAV
TID + VARV → TIDEVARV }
BINDA + BOKSTÄVER → BINDEBOKSTÄVER }
STUGA + GEMAK → STUGUGEMAK } VOKAL-BYTE
ÄGA + DEL → ÄGODEL }
ÄTA + PINNE → ÄTPINNE } UTESLUTET A
RÄTT + STAVNING + PROGRAM → RÄTTSTAVNINGSPROGRAM
BINDE-S VID TRELEDSSAMMANSÄTTNING
MEDIUM + ARKIV → MEDIARKIV } SPECIALFALL

FÖRLED I SAMMANSÄTTNINGAR ÄR I GRUNDFORM,
MED NÅGRA FÅ UNDANTAG, TEX BOENDEFORM

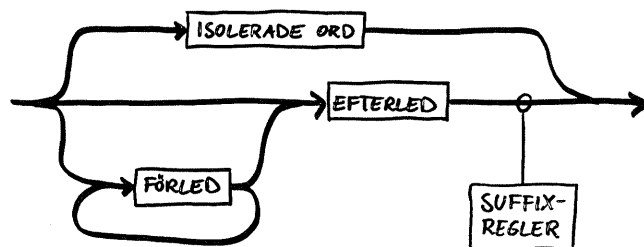
VISSA ORD KAN BARA STÅ SOM FÖRLED, TEX
AUTO-, HYPER-, INFRA-, META-, PSEUDO-

LÖSNING: HA EN SEPARAT DATABAS FÖR
FÖRLED

STAVA: HANTERING AV SAMMANSÄTTNINGAR

TRE ORDLISTOR:

- ISOLERADE ORD (TEX ÖMSÖM, ELLER)
- FÖRLED (TEX FOTBOLLS, MEDIE, STAVNINGSS)
- EFTERLED (TEX KUNSKAP, STÖRT, LÄTA, MEDIUM)



STAVA: SUFFIXREGLER

VILAGRAR VISSA BESTÄMDA FORMER AV ORDEN
I ORDLISTAN OCH INFÖR SUFFIXREGLER AV TYPEN

-ORNA ← -A, -AN, -OR

VILKET BETYDER "OM ORDEN XA, XAN OCH XOR
FINNS I ORDLISTAN SÅ SKA OCKSÅ XORNA GODKÄNNAS"

EXEMPEL: DOCKA, DOCKAN, DOCKOR ⇒ DOCKORNA

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

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

REGLERNA UTFORMAS SÅ ATT DENNA TYP AV FEL
MINIMERAS.

GENERERING AV RÄTTELSEFÖRSLAG

GIVET ETT FELSTAVAT ORD

- GENERERA ALLA ORD SOM LIGGER NÄRA ORDET I NÅGON LÄMPLIG METRIK,
- FÖRKASTA GENERERADE ORD SOM INTE FINNS I ORDLISTAN
- LÄMNA RESTERANDE ORD SOM RÄTTELSEFÖRSLAG

OLIKA METRIK FÖR OLIKA TYPER AV TEXT

EXEMPEL:

- TEXT INMATAD MED HJÄLP AV TANGENTBORD
- MASKINSKRIVEN TEXT INLÄST MED OCR
- HANDSKRIVEN TEXT (PENNDATOR)

FELSTAVNINGAR I TANGENTBORDSINMATAD

TEXT BERÖR TILL 80-90% PÅ ATT

- TVÅ INTILLIGGANDE BOKSTÄVER KASTATS OM,
- EN BOKSTAV FÖRSVUNNIT,
- EN EXTRA BOKSTAV STOPPATS IN,
- EN BOKSTAV BYTTES UT MOT EN ANNAN.

DESSA PUNKTER GÄLLER FÖR SÅVÄL SKRIVFEL SOM VANLIGA STAVFEL SOM BERÖR PÅ OKUNSKAP OM STAVNINGEN.

EXEMPEL PÅ RÄTTSTAVNING

FELSTAVAT ORD: STRUTN

- KASTA OM INTILLIGGANDE BOKSTÄVER:
TSRUTN, SRIUTN, STURTN, STRUN, STRUNT
- SÄTT IN EN EXTRA BOKSTAV:
...ESTRUTN, SETRUTN, STERUTN, STREUTN,
STRUETN, STRUTEN, STRUTNE, ...
- TA BORT EN BOKSTAV:
.TRUTN, .SRUTN, .STUTN, .STRUN, STRUT.
- ERSÄTT EN BOKSTAV:
... YTRUTN, ... SXTRUTN, ... STWUTN,
... STRVTN, ... STRUUN, ... STRUTS, ...

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

EXEMPEL: FELSTAVAT 10-BOKSTAVSORD GER 618 ORD

GRAFOTAKTISK TABELL

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

EXEMPEL: STRUET, RUTNET, SXRU#T

TA OCKSÅ HÄNSYN TILL ORDBÖRJAN OCH ORDSLUT.

I SVENSKA SPRÅKET FINNS 6% AV ALLA FYRGRAM

TVÅ ANVÄNDNINGSGRÄNSER:

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.

RANGORDNADE GENERERING AV RÄTTELSEFÖRSLAG

GIVET ETT FELSTAVAT ORD

- GENERERA ALLA ORD SOM LIGGER NÄRA ORDET I NÅGON LÄMPLIG METRIK,
- FÖRKASTA GENERERADE ORD SOM INTE FINNS I ORDLISTAN
- LÄMNA RESTERANDE ORD SOM RÄTTELSEFÖRSLAG

RANGORDNA RESTERANDE ORD
SÅ ATT DET TROLIGASTE
RÄTTELSEFÖRSLAGET KOMMER FÖRST

RANGORDNING AV RÄTTELSEFÖRSLAG

METOD A: FÖRFINAD METRIK

EXEMPEL: MINST AVSTÅND HAR ORD SOM SKILJER PÅ

- ENKEL/DUBBELSKRIVNING AV KONSONANT
- E OCH Ä BYTTA, G OCH J BYTTA
- BOKSTAV BYTT MOT NÄRLIGGANDE BOKSTAV PÅ TANGENTBORDE

METOD B: FREKVENSLISTA

ORDEN SORTERADE EFTER FREKVENSI I TIO KLASSER
VANLIGT ORD FÖREDRAS FRAMFÖR OVANLIGARE

UTVÄRDERING AV DESSA METODER

GJORD SOM EXJÖBB AV MIKAEL TILLENUS 1995/96

METOD	ANDEL KORREKTA RÄTTELSE
SLUMP	32%
A	61%
B	56%
A+B	70%

90% AV DOM 632 TESTADE FELSTAVNINGARNA
SKULLE IDEALISKT KUNNA RÄTTAS. ÖVRIGA ORD
FANNS INTE MED I ORDLISTAN.

OPTIMERING AV HASHNING

OBSERVATION:

NÄSTAN ALL TID AV RÄTTSTAVNINGEN
TILLBRINGAS I HASHNINGSFUNKTIONEN.

FÖRSÖK OPTIMERA HASHNINGEN!

URSPRUNGLIGA HASHFUNKTIONER:

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

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.

TEST AV HASHNINGSOPTIMERING

INDATA: ORDLISTA MED 200000 ORD VARAV
KNAPPT HÄLFTEN FINNS I STAVAS
ORDLISTA

TID UTAN OPTIMERING: 27.3 s

TID MED KOMPILATORNS OPTIMERING - O2
PÅSLAGEN: 23.0 s

TID OM DESSUTOM BARA EN AVSLUTANDE
MODULOBERÄKNING UTFÖRS: 11.9 s

TID OM PROCESSORNS INBYGGDA MODULO-
FUNKTION ANVÄNDS: 6.4 s

EFFEKTIVITET

FUNKTION IMPLEMENTERAD SOM ARRAY:

```
bool IsVowel(char c)
```

IMPLEMENTERAS SOM

```
bool IsVowel[256];
```

SÅ ATT TEX `IsVowel['e'] == true`

FYRGRAMSTABELL SOM BITARRAY:

```
bool Is4gram(char c[4])
```

IMPLEMENTERAS SOM

```
unsigned char gtable[30*30*30*30/8];
```

FREKVENSOROLISTA SOM BLOOMFILTER:

```
int FrequencyClass(char *word)
{
    char c, tmpword[MAXWORDSIZE];
    strcpy(tmpword + 1, word);
    for (c = '9'; c > '0'; c--) {
        tmpword[0] = c;
        if (InXList(tmpword)) return c - '0';
    }
    return 0; /* FINNS INTE MED I FREKVENSLISTAN */
}
```

KONSTRUKTION AV SUFFIXREGLER

ALLA SVENSKA BÖJNINGSMÖNSTER FINNS BESKRIVNA

I THE MORPHOLOGY OF PRESENT-DAY SWEDISH AV STAFFAN HELLBERG.

EXEMPEL: 727 SÖK/A VB SÖKA(S)
SÖKER
SÖKES
SÖKS
SÖKTE(S)
SÖKT(S)
SÖKANDE(S)
SÖKT(S)
SÖKT
SÖKTA(S)
SÖKTA(S)
SÖKTE(S)
SÖK(S)
SÖK-

I SAOL FINNS ORDFORMERNA SÖKA OCH SÖKTE.

STAVA BEHÖVER DÅ REGLERNA

ER	A, TE	AS	A, TE
ES	A, TE	S	A, TE
		TES	A, TE
T	A, TE	TS	A, TE
TA	A, TE	TAS	A, TE
@	A, TE		

↑ TOMT SUFFIX

SUFFIXREGLUPPSLAGNING

EXEMPEL: SLÅ UPP ORDET **KLOCKORNA**

- Kolla för vart och ett av suffixen E, -A, -NA, -RNA, -ORNA. OM SUFFIXET FINNS SOM INGÅNGSSUFFIX (MINST EN STAVELSE SKA ÅTERSTÅ)
- Kolla för varje regel med matchande ingångssuffix
 - OM ÖVRIGA KRAV PÅ ORDET ÄR UPPFYLDA
 - OM ALLA ORD SOM BILDAS MED REGELNS UPPSLAGS-SUFFIX FINNS I ORDLISTAN (ELLER INTE FINNS)

EFFEKTIVITET:

- LAGRA REGLERNA I EN ARRAY SORTERADE EFTER INGÅNGSSUFFIXET SKRIVET BAKLÄNGES. DÄRMED FÖRENKLAS BINÄRSÖKKNINGEN.
- FÖRBERÄKNA INDEX I ARRAYEN FÖR SISTA BOKSTAVEN I INGÅNGSSUFFIXEN OCH LAGRA I EN EGEN ARRAY. DÄRMED FÅR BINÄRSÖKKNINGEN BRA STARTERÄNSET.
- VID EN SÖKNING GÖRS OFTA FLERA UPPSLAGNINGAR AV SAMMA ORD I ORDLISTAN. SPARA DÄRFÖR UPPSLAGNINGENS UTGÅNG I ETT CACHEMINNE FÖRSTA GÅNGEN OCH ANVÄND CACHEMINNET VID FORTSATTA UPPSLAGNINGAR.

GENERERING FRÅN REGLER

KOMMER FELAKTIGA ORD ATT GENERERAS FRÅN EN SUFFIXREGL?

PROGRAMMET **GENERERA** VISAR ALLA ORD SOM KAN PRODUCERAS FRÅN EN REGEL.

EXEMPEL: ER A, TE GENERERAR 289 ORD, BLAND ANDRA LÄER, MATER, NOLLER SOM ÄR FEL.

IDE: FÖRBJUD SUFFIXET -ADE

ER A, TE, ~ADE

↑ NEGERAR, DVS GODKÄNNER BARA SÖKER OM SÖKADE INTE FINNS.

MER PROBLEM: S A, TE, ~ADE GENERERAR ORD SOM KLÖSS OCH VÄXS.

ÅTGÄRD: REGEN FÅR INTE TILLÄMPAS PÅ ORD SOM SLUTAR PÅ S, X ELLER Z.

[^SXZ]S A, TE, ~ADE

DET FINNS OCKSÅ ANDRA KRAV MAN KAN INFÖRA I REGLERNA, SE

<http://www.nada.kth.se/stava/suffixregler.html>

SKYDD AV ORDLISTAN

NÄR MAN KÖPER EN ORDLISTA MÅSTE MAN FÖRBINDA SIG ATT ORDLISTAN ÄR SKYDDAD SÅ ATT DEN INTE KAN UTVINNAS UR PROGRAMMET.

ETT STAVNINGSPROGRAM SOM ALDRIG SVARAR FEL KAN INTE SKYDDA SIN ORDLISTA!

ALGORITM FÖR UTVINNING AV ORDLISTA:

- GENERERA ALLA TÄNKBARA ORD (UTNYTTJA FÖRGRAMMEN SÅ BLIR DET INTE SÅ MÅNGA)
- NOTERA VILKA ORD SOM GODKÄNNES AV STAVNINGSPROGRAMMET.

OM MAN GÖR DETTA MED STAVA FÅR MAN EN ORDLISTA MED 2,3 % SKÄPFORD FÖR ORD MED ≤ 7 BOKSTÄVER.

EN SÄN ORDLISTA ÄR OFTAST OBRUKBAR!

PROBLEM

VI FICK TILLGÅNG TILL SAOL I TEXTFORM.

STÖRSTA PROBLEMET: FEL I SAOL

HITTILLS HAR VI HITTAT MER ÄN 475 FEL.

EXEMPEL:

CHILE|NARE -AR[E]N PL. -ARE

DRÄNK|A -TE V. -BAR -T ADJ. -NING

FOKUS -EN EL. -

LIV -RÄDNING SE D:O

ANNAT PROBLEM: TECKNET | FÖREKOMMER INTE BARA MELLAN SAMMANSATTA LED OCH FÖRE BÖJNING-ÄNDELSE.

EXEMPEL: GRAMM|ATIKER

JAMAI|CAN EL. -KAN

SLUTSATS: STAVA KAN ANVÄNDAS NÄR MAN VILL UPPTÄCKA FEL I ORDLISTOR!