Uppgift 3: Teckensträngssökning

- Instruktioner för rutfältet – Nere på denna sida finns korta instruktioner för att använda svarsrutfältet.

Beteckningar

Att utföra teckensträngssökning är en allmän uppgift inom datavetenskapen. När teckensträngarna för mönstret p och texten t, har givits, är uppgiften att från texten t söka alla sådana index j, varifrån mönstret p i texten börjar. Om exempelvis $p =$ "uubu" och $t =$ "ubwubuuubuu", ska man hitta en förekomst vid $j = 7$, eftersom $t[7..10] =$ "uubu" $= p$ (se figur 1). De ovan använda beteckningarna beskrivs noggrannare nedan.

Vi hänvisar till teckensträngens tecken genom att ange tecknets index inom hakparenteser efter teckensträngsvariabeln. Till exempel $x[1]$ betyder första tecknet i teckensträngen x. Dessutom betyder beteckningen $x[i..j]$ en delsträng inom teckensträngen x med början vid index i och slut vid index j. Beteckningen $|x|$ betyder längden på teckensträngen x. Delsträngar av formen $x[1..j]$ som består av början på teckensträngen x kallas prefix och delsträngar av formen $x[i..|x|]$ som består av slutet på teckensträngen x kallas suffix. Vi använder operatorerna $=$ och $\neq$ på vanligt sätt för att ange teckens eller teckensträngars likhet och olikhet.

Exempel 1. Om teckensträngen x är "university", gäller bl.a. att:

- $|x| = 10$.
- $x[1] =$ "u", $x[3] = x[8] =$ "i" och $x[6] =$ "r". Då gäller t.ex. $x[6] \neq x[8]$ och $x[3] \neq x[1]$.
- $x[1..3] =$ "uni", $x[3..3] =$ "i" och $x[4..10] =$ "versity". Då gäller $x[3..4] =$ "iv" $\neq x[8..9] =$ "it".
- $x[1..6] =$ "univer" är ett prefix till teckensträngen x och $x[7..10] =$ "sity" är ett av dess suffix.

Naiv sökning

Figur 1 förklarar den enkla s.k. naiva sökmetoden, som undersöker villkoret $p = t[j..j+|p|-1]$ enskilt för varje möjlig förekomst av mönstret vid positionerna $j = 1, \ldots, |t| - |p| + 1$. Figuren använder mönstret $p =$ "uubu" och texten $t =$ "ubwubuuubuu" som nämndes i början av uppgiften. Mönstret passas först in med textens början, alltså vid position $j = 1$, sedan vid $j = 2$, och så vidare, så att mönstret flyttas ett steg i taget mot textens slut. Vid varje dylikt jämförelsesteg undersöks om mönstret förekommer vid det aktuella indexet j. Undersökningen av det här villkoret $p = t[j..j + |p| - 1]$ görs genom att jämföra varje enskilt tecken från mönstrets början till slut (från vänster till höger): först jämförs tecknen $p[1]$ och $t[j]$, och om de överensstämmer, så jämförs följande tecken $p[2]$ och $t[j+1]$, och så vidare. Mera formellt uttryckt jämförs tecknen $p[i]$ och $t[j + i - 1]$ i ordningsföljd $i = 1, \ldots, |p|$. Teckenjämförelserna avbryts, om en olikhet upptäcks mellan tecknen $p[i]$ och $t[j + i - 1]$, för då är $p \neq t[j..j + |p| - 1]$. Om däremot tecknen överensstämmer ända till och med det sista teckenparet $p[|p|]$ och $t[j+|p|-1]$, gäller $p = t[j..j + |p| - 1]$ och denna förekomst av mönstret från index j sparas.

I figur 1 är mönstrets matchande tecken märkta med understrykning och icke matchande tecken med överstrykning. Till exempel vid $j = 1$ matchar första teckenjämförelsen $p[1] = t[1] =$ "u", medan följande jämförelse inte matchar $p[2] =$ "u" $\neq t[2] =$ "b". Teckenjämförelsen avslutas

	1	2	3	4	5	6	7	8	9	10	11
	u	b	w	u	b	u	u	u	b	u	u
j=1	u	~~u~~	b	u							
j=2		~~u~~	u	b	u						
j=3			~~u~~	u	b	u					
j=4				u	~~u~~	b	u				
j=5					~~u~~	u	b	u			
j=6						u	u	~~b~~	u		
j=7							u	u	b	u	
j=8								u	~~u~~	b	u

*Figur 1:** Naiv sökning.

efter icke matchande tecken och sökningen flyttar till nästa position $j = 2$. Vid den ifrågavarande positionen är jämförelsen genast i början icke-matchande mellan tecknen $p[1] =$ ”u” och $t[2] =$ ”b”. Märk, att senare vid $j = 7$ matchar alla teckenjämförelser och mönstret hittas i texten.

Heuristiken för matchande suffix

Naiva metoden flyttar mönstret ett steg i taget i relation till texten (d.v.s. index j ökas alltid med ett). Heuristiken för matchande suffix försöker göra längre än ett steg långa förskjutningar. En väsentlig förändring är att vända om undersökningen av villkoret $p = t[j..j + |p| - 1]$, så att teckenjämförelserna görs från slutet till början (från höger till vänster): först jämförs tecknen $p[|p|]$ och $t[j + |p| - 1]$, och om de matchar, jämförs tecknen $p[|p| - 1]$ och $t[j + |p| - 2]$ och så vidare, möjligen ända till tecknen $p[1]$ och $t[j]$. Mera formellt uttryckt, jämförs tecknen $p[i]$ och $t[j + i - 1]$ enligt ordningsföljden $i = |p|, \dots, 1$. Jämförelserna avbryts, om ett icke-matchande teckenpar $p[i]$ och $t[j + i - 1]$ påträffas.

Om det första icke-matchande teckenparet $p[i] \neq t[j+i-1]$ hittas vid positionen $i < |p|$, är det genast klart på basen av de tidigare matchningarna, att $p[i+1..|p|] = t[j+i..j+|p|-1]$. Nu är det viktigt att notera, att om mönstret flyttas från nuvarande positionen j till en sådan position $j + k$, där $k < |p|$, så ingår den tidigare nämnda delen $t[j + i..j + |p| - 1] = p[i + 1..|p|]$ helt eller delvis även i den följande delsträngen $t[j + k..j + k + |p| - 1]$ som ska undersökas i texten. Detta betyder, att i position $j + k$ gäller $p = t[j + k..j + |p| - 1]$ endast om det ifrågavarande tidigare matchade mönstrets suffix $p[i + 1..|p|]$ stämmer överens med den del av mönstret, som är sida vid sida med den motsvarande textdelen $t[j + i..j + |p| - 1] = p[i + 1..|p|]$. Detta villkor kan undersökas lätt, om det före den egentliga sökningen i teckensträngen inleds, har skapats en förskjutningstabell S vars alla värden $S[h]$, där $h = 1, \dots, |p|$, anger den minsta längden för att flytta fram mönstret, så att det matchande suffixet $p[h..|p|]$ inte leder till att mönstret inte matchar i den nya positionen. När det första icke-matchande teckenparet $p[i] \neq t[j + i - 1]$ hittas i position $i < |p|$, kan mönstret härnäst flyttas till position $j + S[i + 1]$. Om teckenparet genast i position $i = |p|$ inte matchar, flyttas mönstret bara ett steg till $j + 1$. Om mönstret å andra sidan hittas (alltså det matchar med texten) är det matchande suffixet $p[1..|p|]$ och

mönstret kan flyttas till position $j + S[1]$.

Figur 2: Förskjutningstabellen S för mönstret p = “uubu”.

h	1	2	3	4
$S[h]$	3	3	3	2

S

	1	2	3	4	5	6	7	8	9	10	11
	u	b	w	u	b	u	u	u	b	u	u
j=1	u	u	~~b~~	u							
		u	u	~~b~~	u						
j=3			~~u~~	u	b	u					
				u	~~u~~	~~b~~	u				
					~~u~~	u	b	u			
j=6						u	u	b	~~u~~		
j=7							u	u	b	u	
								u	~~u~~	~~b~~	u

Figur 3: Heuristiken för matchande suffix.

Figur 3 illustrerar heuristiken för matchande suffix med samma teckensträngar p = ”uubu” och t = ”ubwubuuubuu” som tidigare. Den på förhand beräknade förskjutningstabellen S för mönstret p = ”uubu” presenteras i figur 2. Vi kommer senare att beskriva mera i detalj hur värdena i förskjutningstabellen S bestäms.

Till exempel i figur 3 hittas i begynnelsepositionen $j = 1$ först en matchning $p[4] = t[4] =$ ”u” och sedan ett icke-matchande par $p[3] =$ ”b” $\neq t[3] =$ ”w”. De matchande tecknen har märkts med understrykning och de icke-matchande med överstrykning. Det matchande suffixet var alltså $p[4] = t[4] =$ ”u” och jämförelsen var icke-matchande i position $i = 3$. Mönstret flyttas därefter $S[i + 1] = S[4] = 2$ steg framåt till position $j = 1 + S[4] = 1 + 2 = 3$. Figur 3 beskriver principen för att bestämma förskjutningen genom att omge det matchande suffixet med en rektangel, inom vilken mönstrets tecken skall matcha också i nästa förskjutningsposition. Figuren visar de överhoppade, ogiltiga förskjutningspositionerna i mönstret med grå färg, och de icke-matchande tecknen är överstrukna. Till exempel position $j = 2$ hoppas över, för där skulle det matchande suffixet inte matcha $p[4] = t[4] =$ ”u” $\neq p[3] =$ ”b”. Däremot hoppas position $j = 3$ inte över, för där överensstämmer det matchande suffixet: $p[4] = t[4] =$ ”u” $= p[2]$.

I den ovannämnda positionen $j = 3$ matchas suffixet $p[2..4] = t[4..6] =$ ”ubu”, på vilken en icke-matchande del $p[1] =$ ”u” $\neq t[3] =$ ”w” i position $j = 1$ följer. Därpå flyttar vi oss till positionen $j = 3 + S[i + 1] = 3 + S[2] = 3 + 3 = 6$. Positionerna $j = 4$ och $j = 5$ kan vi hoppa över, för i position $j = 4$ skulle vi ha en icke-matchande del $p[2..4] = t[4..6] =$ ”ubu” $\neq p[1..3] =$

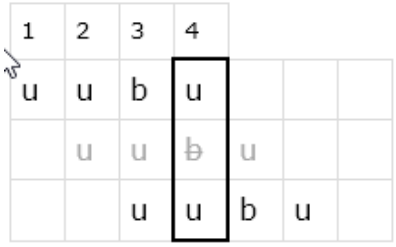
”uub” och i position $j = 5$ en icke-matchande del $p[3..4] = t[5..6] =$ ”bu” $\neq p[1..2] =$ ”uu”. Däremot hoppas inte position $j = 6$ över, för där matchar $p[4] = t[6] =$ ”u” $= p[1]$.

I position $j = 7$ matchar alla tecken och det matchande suffixet är $p[1..4]$. På basen av detta skulle mönstret härnäst flyttas till position $j = 7 + S[1] = 7 + 3 = 10$, som redan skulle placera en del av mönstret utanför texten. Därför avslutas exemplets sökning efter position $j = 7$.

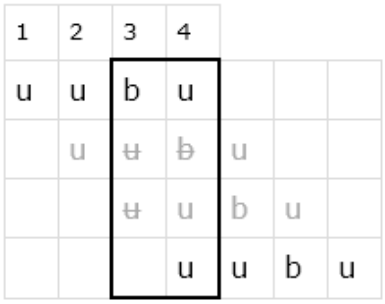
Beräkning av värden i förskjutningstabellen S

Värdena i förskjutningstabellen S kan enkelt bestämmas genom att undersöka hur mönstrets alla suffix matchar med mönstrets olika förskjutningar i relation till sig själv. Figur 4 visar beräkningen av värdena $S[4]$, $S[3]$, $S[2]$ och $S[1]$ för exempelmönstret $p =$ ”uubu” från exemplet i figur 2. Allmänt taget kan värdet $S[h]$ beräknas genom att steg för steg genom testning söka mönstrets första förskjutning i relation till sig själv, där alla tecken i suffixet $p[h..|p|]$ och mönstrets förskjutning matchar.

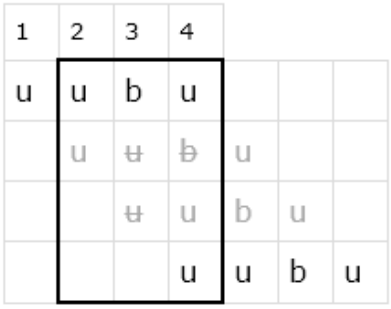
Figur 4: Beräkning av värdena i förskjutningstabellen S för mönstret p = ”uubu”.



S[4] = 2



S[3] = 3



S[2] = 3

1	2	3	4			
u	u	b	u			
	u	~~u~~	~~b~~	u		
		~~u~~	u	b	u	
			u	u	b	u

S[1] = 4

Till exempel i figur 4 för värdet $S[4]$ avgränsar en rektangel suffixet $p[4..4]$ = ”u”. Under mönstret har vi först placerat ett mönster som har förskjutits ett steg framåt, och för den positionen har vi undersökt om den del av det förskjutna mönstret som hamnar inne i rektangeln matchar suffixet $p[4..4]$. I det här fallet matchade det inte: i den första förskjutningspositionen i rektangeln matchar inte tecknen $p[4]$ = ”u” $\neq$ ”b” = $p[3]$. Raden i fråga är på grund av detta gråfärgad: det är fråga om en förskjutning som leder till att tecknen inte matchar. I raden under har ett steg lagts till, det vill säga mönstret har flyttat två steg, och matchningen har igen undersökts i rektangelområdet. Nu matchar tecknen $p[4]$ = ”u” = $p[2]$, på basen av vilket $S[4] = 2$, för 2 var den minsta förskjutning, som inte ledde till icke-matchande tecken.

På motsvarande sätt finns det till exempel vid värdet för $S[1]$ en rektangel, som avgränsar suffixet $p[1..4]$ = ”uubu”. Under den finns först en förskjutning på ett steg, som leder till icke-matchande tecken inne i rektangeln $p[2..4]$ = ”ubu” $\neq$ ”uub” = $p[1..3]$. Därefter är det en förskjutning på två steg, som leder till icke-matchande tecken $p[3..4]$ = ”bu” $\neq$ ”uu” = $p[1..2]$. Till sist är det en förskjutning på tre steg, för vilken $p[4]$ = ”u” = $p[1]$ matchar i rektangeln. På det här sättet är $S[1] = 3$, för 3 är den minsta förskjutningen, som inte ledde till icke-matchande tecken.

Det är bra att notera, att förskjutningstabellens värden beror mycket på mönstrets struktur. Ett extremfall är mönstret där alla tecken är lika: då är $S[h] = 1$ för alla index h. Figur 5 visar ett exempel på beräkningen av värdet på $S[2]$ för mönstret p = ”uuu”. Ett annat extremfall är mönstret där det sista tecknet inte finns någon annanstans i mönstret: då är $S[h] = |p|$ för alla index h. Figur 6 visar beräknandet av värdet på $S[2]$ för ett sådant mönster p = ”uub”.

Figur 5: S[2] för mönstret p = ”uuu”.

1	2	3				
u	u	u				
	u	u	u			

S[2] = 1

Figur 6: S[2] för mönstret p = ”uub”.

Frågorna

I frågorna 3.1 – 3.3 används teckensträngen x = ”mississippi”.

1	2	3				
u	u	b				
	u	~~u~~	b			
		~~u~~	u	b		
			u	u	b	

S[2] = 3

Fråga 3.1. (1 p.)

Bestäm värdet på $|x|$.

Fråga 3.2. (1 p.)

Ange teckensträngarna:

a) $x[2..5]$ (0.5 p.)

b) $x[7..9]$ (0.5 p.)

Fråga 3.3. (1 p.)

Ange alla index j, för vilka det gäller att $x[j] = x[j+3]$.

Det kan finnas noll eller flera rätta alternativ. Genom att välja rätta alternativ får man poäng. Genom att välja felaktiga alternativ dras poäng av. Frågans minimipoäng är 0.

Svaret på frågan 3.3

Fråga 3.4. (3 p.)

I denna fråga kontrolleras värdena i förskjutningstabellen S i heuristiken för matchande suffix för mönstret $p =$ ”bbububu”. Ange i varje delfråga hur det enskilda värdet beräknas i samma form som i figur 4. I svarsrutfältet finns redan mönstret p ifyllt på första raden. Observera följande i ditt svar:

- I svarsrutfältet ska du ange en rektangel som avgränsar suffixet.
- På raderna där förskjutningarna leder till icke-matchande teckenpar är mönstret grått och dess icke-matchande tecken stryks över.

- Märk inte ut något överflödigt i rutfältet: poängsättningen görs på basen av hur exakt ditt rutfält är rätt ifyllt.

Observera att genom att dubbelklicka på en cell får du upp en "minieditor". Minieditorn visar inte överstrykning, understrykning eller gråa bokstäver. De syns först när du går ut från minieditorn genom att trycka på `Esc`-tangenten. Du kan skriva in bokstäverna "b" och "u" i cellen också med hjälp av verktygsbalken utan att använda minieditorn.

a) Ange beräkningen av värdet $S[7]$ i samma form som i figur 4. (1 p.)

Svaret på frågan 3.4a

S[7] =

b) Ange beräkningen av värdet $S[4]$ i samma form som i figur 4. (1 p.)

Svaret för 3.4b kommer här

S[4] =

c) Ange beräkningen av värdet $S[1]$ i samma form som i figur 4. (1 p.)

Svar för frågan 3.4c här.

S[1] =

Fråga 3.5. (3 p.)

Ange de steg som heuristiken för matchande suffix tar vid sökningen av mönstret $p =$ "bbububu" från texten $t =$ "ubbuubbubububuubub". Ange ditt svar i samma form som i figur 3. I rutfältet är texten och mönstret i position $j = 1$ färdigt ifyllda. Observera följande när du svarar:

- Mönstret anges i varje möjlig position j (även i de överhoppade).
- Position j skrivs in i den första kolumnen på de rader, där mönstret jämförs med texten. På de ifrågavarande raderna understryks dessutom varje matchande tecken i mönstret, medan icke-matchande tecken stryks över.
  - Även i den på förhand givna raden för $j = 1$ ska berörda tecken i mönstret under- och överstrykas!
- På raderna med överhoppade positioner:
  - position j skrivs inte ut i vänstra kolumnen,
  - mönstret anges med grå färg och
  - tecken som man vet att inte matchar stryks över.
- Det är frivilligt att rita rektanglarna som avgränsar suffixen som i figur 3; man förlorar inga poäng i denna fråga för att de saknas eller är felaktiga.
- Märk inte ut något överflödigt i rutfältet: poängsättningen görs på basen av hur exakt ditt rutfält är rätt ifyllt.

Svaret på fråga 3.5 kommer här.

Fråga 3.6. (6 p.)

I frågan finns två delfrågor a) och b), i vilka det för var och en är givet en tabell S. Ange skilt för båda delfrågorna en sådan teckensträng p, vars förskjutningstabell enligt heuristiken för matchande suffix motsvarar den tabell S som givits i delfrågan. Teckensträngen p får **endast innehålla tecknen** ”u” och ”b”.

a)

h	1	2	3	4	5	6	7	8
S[h]	8	8	8	8	3	3	3	1

(3 p.)

b)

h	1	2	3	4	5	6	7	8	9	10	11	12
S[h]	11	11	11	11	11	11	5	5	5	5	3	3

(3 p.)

Fråga 3.7. (5 p.)

Formulera ett 5 tecken långt mönster p och en 12 tecken lång textsträng t innehållande **endast tecknen** ”u” och ”b”, så att heuristiken för matchande suffix utför **exakt** 16 teckenjämförelser när den söker mönstrets förekomster i texten. Här räknas endast teckenjämförelser som utförts under det egentliga sökskedet (jämförelserna som gjorts under beräkningen av förskjutningstabellen S räknas inte). Som jämförelser räknas både matchande och icke-matchande jämförelser. Till exempel i sökningen i figur 3 utförs allt som allt 11 teckenjämförelser: 2 i position $j = 1$, 4 i position $j = 3$, 1 i position $j = 6$ och 4 i position $j = 7$.

Strängar

5 tecken långt mönster p: ○

12 tecken lång text t: ○