T-(538|725)-MALV, Málvinnsla
Reglulegar segðir og Perl

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

Ágúst 2007

Outline

Outline

Strengir

Stafróf

- Endanlegt mengi stafa.
- Dæmi: $\{0,1\}$ er tvíundarstafrófið.

Strengur

- Strengur s úr stafrófi Σ er endanleg runa stafa sem dregnir eru úr Σ.
- $|s|$ táknar lengdina á s.
- ϵ táknar tóma strenginn; lengd hans er 0.

Strengir

Stafróf

- Endanlegt mengi stafa.
- Dæmi: $\{0,1\}$ er tvíundarstafrófið.

Strengur

- Strengur s úr stafrófi Σ er endanleg runa stafa sem dregnir eru úr Σ.
- $|s|$ táknar lengdina á s.
- ϵ táknar tóma strenginn; lengd hans er 0.

Mál

Skilgreining

- Mengi strengja.
- Dæmi: $\emptyset$, $\{\epsilon\}$, {ab,ba}, {011,101,111}.

Samtenging og margföldun

- Ef x og y eru strengir þá er samtenging þeirra, xy, strengur sem fæst með því að bæta y við x.
- $s\epsilon = \epsilon s = s$
- $s^0 = \epsilon$, $s^1 = s$, $s^2 = ss$,
- $s^i = ss^{i-1}$, $i > 0$

Mál

Skilgreining

- Mengi strengja.
- Dæmi: $\emptyset$, $\{\epsilon\}$, {ab,ba}, {011,101,111}.

Samtenging og margföldun

- Ef x og y eru strengir þá er samtenging þeirra, xy, strengur sem fæst með því að bæta y við x.
- $s\epsilon = \epsilon s = s$
- $s^0 = \epsilon$, $s^1 = s$, $s^2 = ss$,
- $s^i = ss^{i-1}$, $i > 0$

Aðgerðir á málum

- $L \cup M = \{s \mid s \in L \text{ eða } s \in M\}$
- $LM = \{st \mid s \in L \text{ og } t \in M\}$
- Kleene closure: 0 eða fleiri samtengingar af L
  - $L^* = \cup_{i=0}^{\infty} L^i$
- Positive closure: 1 eða fleiri samtengingar af L
  - $L^+ = \cup_{i=1}^{\infty} L^i$

Dæmi um mál

$L = \{A, B, \ldots, Z, a, b, \ldots, z\}$ og $D = \{0, 1, \ldots, 9\}$.
Hvaða mál (mengi strengja) eru þá:

- $L \cup D$
- LD
- L^4
- L^*
- $L(L \cup D)^*$
- D^+

Outline

Reglulegar segðir (e. regular expressions)

- Mál sem notað er til að lýsa mengi strengja.
- Sérstaklega öflugt til að lýsa mynstrum (e. patterns) sem leita skal að í texta.
- Sérhver regluleg segð r stendur fyrir mál $L(r)$.
- Samanstanda af venjulegum stöfum (t.d. *abc*) ásamt stöfum sem hafa sérstaka merkingu (þ.e. "metacharacters") eins og "*" og "+".
- Hægt að búa til flóknari reglulegar segðir úr einfaldri segðum með því að nota sérstakar reglur.

Reglulegar segðir

Fyrir stafróf Σ:

1. ϵ er regluleg segð (RS) sem stendur fyrir $\{\epsilon\}$.
2. Ef $a \in \Sigma$, þá er a RS sem stendur fyrir $\{a\}$.
3. G.r.f. að r og s séu RS sem standa fyrir málin $L(r)$ og $L(s)$. Þá gildir:
   - $(r)|(s)$ er RS sem táknar $L(r) \cup L(s)$.
   - $(r)(s)$ er RS sem táknar $L(r)L(s)$.
   - $(r)^*$ er RS sem táknar $(L(r))^*$.
   - (r) er RS sem táknar $L(r)$.

Reglulegar segðir

Forgangur:

- * hefur hæstan forgang.
- Samtenging hefur næst hæstan forgang.
- | hefur lægsta forgang.
- Því gildir t.d. að: $(a)|((b)^*(c)) = a|b^*c$

Dæmi um reglulegar segðir

Hvaða mál tákna reglulegu segðirnar:

- $a|b$
- $(a|b)(a|b)$
- a^*
- $a|b^*c$

Meira um reglulegar segðir

Aðrir stafir með sérstaka merkingu

Í mörgum tólum sem bjóða upp á reglulegar segðir hafa þessir stafir einnig sérstaka merkingu:

- ? + . {n}
- Sjá skýringar í töflu 2.9 bls. 37

Meira um reglulegar segðir

"Character classes"

- Listi af stöfum innan í hornklofum stemmir við hvaða staf sem er í listanum.
- Reglulega segðin $[abc]$ merkir eitt tilvik af a, eða b eða c $(a|b|c)$.

"Complement and range"

- `[^a]` merkir hvaða stafur sem er ekki `a`.
- `[a-zA-Z]` merkir a, b, ..., z, A, B, ..., Z.

Meira um reglulegar segðir

"Character classes"

- Listi af stöfum innan í hornklofum stemmir við hvaða staf sem er í listanum.
- Reglulega segðin $[abc]$ merkir eitt tilvik af a, eða b eða c $(a|b|c)$.

"Complement and range"

- `[^a]` `merkir hvaða stafur sem er ekki a.`
- [a-zA-Z] merkir a, b, ..., z, A, B, ..., Z.

“Longest match”

Margræðni

- “String matching” getur verið margræð.
- T.d. strengurinn $s =$“aabbc” og reglulega segðin a^+b^*
- Reglulega segðin passar við þessa hlutstrengi úr s: a, aa, ab, aab, abb, $aabb$

Einræðing - tvær reglur

Tól sem styðja reglulegar segðir:

- Byrja að “matcha” eins snemma og hægt er í strengnum.
- “Matcha” eins marga stafi í einu og mögulegt er.

“Longest match”

Margræðni

- “String matching” getur verið margræð.
- T.d. strengurinn $s =$“aabbc” og reglulega segðin a^+b^*
- Reglulega segðin passar við þessa hlutstrengi úr s: a, aa, ab, aab, abb, $aabb$

Einræðing - tvær reglur

Tól sem styðja reglulegar segðir:

- Byrja að “matcha” eins snemma og hægt er í strengnum.
- “Matcha” eins marga stafi í einu og mögulegt er.

Tengsl við endanlegar stöðuvélar

- Hægt að breyta reglulegri segð í endanlega stöðuvél á vélrænan hátt.
  - T.d. tekið fyrir í námskeiðinu *Þýðendur*.
- Endanleg stöðuvél getur þá borið kennsl á þá strengi sem tiltekin regluleg segð stendur fyrir.

Ýmis tól og forritunarmál

- grep/egrep (Unix/Linux-tól)
  - grep 'ab*c' myFile
  - Prentar út allar línur úr myFile sem innihalda strengina *ac*, *abc*, *abbc*, *abbbc*, o.s.frv.
  - Undir Windows getið þið sett upp *Cygwin* `http://www.cygwin.com/` sem er "Linux-like environment for Windows".
- Reglulegar segðir eru líka notaðar í Perl, Python, Java, C#.

Outline

Forritunarmálið Perl

Uppruni

- PERL = "Practical Extraction and Report Language"
- Búið til árið 1987 af Larry Wall
- Allrahandatól fyrir UNIX notendur
- Öflugra en skeljaforrit og einfaldar í notkun en C
- Sérstaklega hröð textavinnsla

Halló Heimur

```
#Reading a string and printing it back out
print("What is your name? ");
$name = <STDIN>;
print "Pleased to meet you ", $name;
```

Forritunarmálið Perl

Útlit forrits

- Venjuleg textaskrá (*.pl) með röð Perl skipana
- Auðbil hunsuð í upphafi línu
- Hver skipun endar á semikommu (";")
- Fjöldi auðbila utan strengja skiptir ekki máli
- Allt á eftir "#" er hunsað (athugasemd/skjölun)

Stutt forrit

```
#Prentar smá texta
print("Þetta er texti\n",
      "sem nær yfir tvær línur");
```

Forritunarmálið Perl

Útlit forrits

- Venjuleg textaskrá (*.pl) með röð Perl skipana
- Auðbil hunsuð í upphafi línu
- Hver skipun endar á semikommu (";")
- Fjöldi auðbila utan strengja skiptir ekki máli
- Allt á eftir "#" er hunsað (athugasemd/skjölun)

Stutt forrit

```
#Prentar smá texta
print("Þetta er texti\n",
      "sem nær yfir tvær línur");
```

Forritunarmálið Perl

Kverður (Scalars)

- Allir strengir og tölur eru kverður
- Kverðubreytur byrja alltaf á "$" (dollar)
- Perl breytir sjálfkrafa á milli strengja og talna

Dæmi

- `$a=2; $b=6; $c=$a.$b; $d=$c/2; print $d;`
- `$nafn='jón'; print "Ég heiti :\t$nafn\n";`

Forritunarmálið Perl

Kverður (Scalars)

- Allir strengir og tölur eru kverður
- Kverðubreytur byrja alltaf á "$" (dollar)
- Perl breytir sjálfkrafa á milli strengja og talna

Dæmi

- `$a=2; $b=6; $c=$a.$b; $d=$c/2; print $d;`
- `$nafn='jón'; print "Ég heiti :\t$nafn\n";`

Forritunarmálið Perl

Fylki (Arrays)

- Fylki er safn kverða
- Fylkjabreytur byrja alltaf á "@" (at)
- Vísað er á stak með tölu í hornklofa "[númer]"

Dæmi

- `@g=('jón',25); print "$g[0] er $g[1] ára";`
- `@a=(1,2); @b=(3,4); @c=(@a,@b);`

Forritunarmálið Perl

Fylki (Arrays)

- Fylki er safn kverða
- Fylkjabreytur byrja alltaf á "@" (at)
- Vísað er á stak með tölu í hornklofa "[númer]"

Dæmi

- `@g=('jón',25); print "$g[0] er $g[1] ára";`
- `@a=(1,2); @b=(3,4); @c=(@a,@b);`

Forritunarmálið Perl

Tengifylki (Associative Array)

- Sama og tætitafla (hash table) eða kort (map)
- Fylki þar sem stök eru sótt með lyklastrengjum

Dæmi

- `%enska=('epli','aple','pera','pear');`
- `print enska{'epli'};`
- `print keys(%enska);`
- `print values(%enska);`

Forritunarmálið Perl

Skráarhald (File Handle)

- Strengur (upphafsstafastrengur) sem bendir á skrá
- Fyrst tengt við skrá og svo notað við lestur og skrif
- Perl er þegar búið að tengja STDIN, STDOUT og STDERR

Dæmi

```
open(INNSKRA, "text.txt");
open(UTSKRA, ">nidurstodur.txt");
while(<INNSKRA>) { #Ein lína í einu geymd í $_
    print UTSKRA $_,"\n---\n";
}
close(INNSKRA); close(UTSKRA);
```

Forritunarmálið Perl

Skráarhald (File Handle)

- Strengur (upphafsstafastrengur) sem bendir á skrá
- Fyrst tengt við skrá og svo notað við lestur og skrif
- Perl er þegar búið að tengja STDIN, STDOUT og STDERR

Dæmi

```
open(INNSKRA, "text.txt");
open(UTSKRA, ">nidurstodur.txt");
while(<INNSKRA>) { #Ein lína í einu geymd í $_
    print UTSKRA $_,"\n---\n";
}
close(INNSKRA); close(UTSKRA);
```

Sjálfgefið skráarhald

- Ef skráarhaldið er tómi strengurinn, þ.e. <> þá
  - Nota skrár tilgreindar á skipanalínu
  - Nota STDIN ef skrár ekki tilgreindar

Dæmi

```
while(<>) {
    print $_; #Ath: má sleppa $_
}
```

Forritunarmálið Perl

Sjálfgefið skráarhald

- Ef skráarhaldið er tómi strengurinn, þ.e. <> þá
  - Nota skrár tilgreindar á skipanalínu
  - Nota STDIN ef skrár ekki tilgreindar

Dæmi

```
while(<>) {
    print $_; #Ath: má sleppa $_
}
```

Forritunarmálið Perl

Núverandi lína

- Sérstaka breytan $_ geymir oft þá línu sem verið er að lesa frá inntaki, þ.e. núverandi línu
- Ef þessa línu þarf að geyma, má nota aðra breytu:
  $sidast = $_;
- Ýmsar aðgerðir gera ráð fyrir að unnið sé með þessa breytu, nema annað sé tekið fram

Forritunarmálið Perl

Skilyrtar skipanir

- `print "Tómt" if ! $fullt ;`
- `print "Tókst!" unless $villa >2;`
- `$n=100; print "$n\n" while $n-- > 0;`
- `$n=0; print "$n\n" until $n++ > 100;`

Skilyrtar blokkir

```
if($n>100) {
    print "sigur!";
} elsif($s>50) {
    print "samt sigur!";
} else {
    print "tap";
}
```

Forritunarmálið Perl

Skilyrtar skipanir

- `print "Tómt" if ! $fullt ;`
- `print "Tókst!" unless $villa >2;`
- `$n=100; print "$n\n" while $n-- > 0;`
- `$n=0; print "$n\n" until $n++ > 100;`

Skilyrtar blokkir

```
if($n>100) {
    print "sigur!";
} elsif($s>50) {
    print "samt sigur!";
} else {
    print "tap";
}
```

Forritunarmálið Perl

foreach lykkjan

- Mjög öflug lykkjuskipun til að vinna með stök fylkja
- `@tolur = ('einn', 'tveir', 'thrir');`
- `foreach $tala (@tolur) { print "Talan $tala ...\n"; }`

Forritunarmálið Perl

Notkun reglulegra segða í Perl

- Mátun (match): **m**/regex/modifiers
- Skipting (substitution): **s**/regex/replacement/modifiers
- Þýðing (translation):
  tr/search_list/replacement_list/modifiers

Dæmi

```
while(<>) {
    print if(m/ab*c/i);
}
```

Forritunarmálið Perl

Notkun reglulegra segða í Perl

- Mátun (match): **m**/regex/modifiers
- Skipting (substitution): **s**/regex/replacement/modifiers
- Þýðing (translation):
  tr/search_list/replacement_list/modifiers

Dæmi

```
while(<>) {
    print if(m/ab*c/i);
}
```

Forritunarmálið Perl

Annað dæmi

```
while($line = <>) {
    if($line =~ m/ab*c/i) {
        $line =~ s/ab*c/ABC/g;
    } else {
        $line =~ tr/A-Z/a-z/;
    }
}
```

Forritunarmálið Perl

Vísað í mátaðan streng

- Oft er þörf á að vísa aftur á mátaðan strenghluta, t.d. til að sjá hvort hann endurtekur sig eða til að prenta hann
- Strenghluti er settur í biðminni með svigum () og vísað á hann með \númer innan sömu segðar eða með $númer utan hennar

Dæmi

```
while($line = <>) {
    while($line=~ m/\$ *([0-9]+)\.?([0-9]*)/g) {
        print "Dollars: ",$1," Cents: ", $2, "\n";
    }
}
```

Forritunarmálið Perl

Vísað í mátaðan streng

- Oft er þörf á að vísa aftur á mátaðan strenghluta, t.d. til að sjá hvort hann endurtekur sig eða til að prenta hann
- Strenghluti er settur í biðminni með svigum () og vísað á hann með \númer innan sömu segðar eða með $númer utan hennar

Dæmi

```
while($line = <>) {
    while($line=~ m/\$ *([0-9]+)\.?([0-9]*)/g) {
        print "Dollars: ",$1, " Cents: ", $2, "\n";
    }
}
```

Annað áhugavert

Dæmi

```
perl -i.old -p -e "s/foo/bar/g" wiffle.bat
```

- -e keyrir þær Perl skipanir sem eru í strengnum fyrir aftan
- Skipunin sem er keyrð hér er s/foo/bar/g sem skiptir út öllum strengjum "foo" í sjálfgefnu inntaki (hér er það hver lína í skránni "wiffle.bat") fyrir strenginn "bar"
- -p prentar út niðurstöðuna, en vegna -i er það ekki gert á skjáinn, heldur í sömu skrá og inntakið er úr, en upphaflega skráin er vistuð með endingunni ".old"