T-(538|725)-MALV, Málvinnsla
Tilreiðing texta

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

Ágúst 2007

Outline

Outline

Hvenær er hentugt að nota Perl?

"Quick and dirty"

- Þegar skrifa þarf "quick and dirty" forrit á skömmum tíma.
- Í flestum tilvikum er Perl notað til að þróa forrit sem tekur minna en eina klukkustund að hanna, skrifa og prófa.

Þegar meðhöndla þarf texta

- Innbyggð meðhöndlun á reglulegum segðum.
- Einföld skráarmeðhöndlun.
- Gagnleg gagnaskipan.

Hvenær er hentugt að nota Perl?

"Quick and dirty"

- Þegar skrifa þarf "quick and dirty" forrit á skömmum tíma.
- Í flestum tilvikum er Perl notað til að þróa forrit sem tekur minna en eina klukkustund að hanna, skrifa og prófa.

Þegar meðhöndla þarf texta

- Innbyggð meðhöndlun á reglulegum segðum.
- Einföld skráarmeðhöndlun.
- Gagnleg gagnaskipan.

Orðstöðulykill (e. concordance)

- Orðstöðulykill er skrá yfir orðmyndir sem koma fyrir í tilteknum texta eða textum, ásamt upplýsingum um nánasta samhengi þeirra.
- Algengastir eru svonefndir KWIC-lyklar (e. Key Word In Context) þar sem hvert dæmi um lykilorðið stendur í miðri línu ásamt orðum sem standa næst á undan því og eftir í textanum.
- `http://en.wikipedia.org/wiki/Concordance_%28publishing%29`
- `http://www.lexis.hi.is/corpus/leit.pl`

Dæmi: Bygging orðstöðulykils í Perl – concordance.pl

```
($file_name, $pattern, $width) = @ARGV;
open(FILE, "$file_name") || die "Could not open $file_name."
while ($line = <FILE>) {
    $text .= $line;
}

# new lines are replaced by spaces
$text =~ s/\n/ /g;

# matches the pattern with 0..width to the right and left
while ($text =~ m/(.{0,$width}$pattern.{0,$width})/g) {
    print "$1\n";   # $1 contains the match
}
```

Dæmi: Bygging orðstöðulykils í Perl

- perl -w concordance.pl visindavefur2.txt vegna 15 > concordance.out
- perl -w concordance.pl visindavefur2.txt einnig 20 > concordance.out

Dæmi um hentuga gagnaskipan

Tætitafla (e. hash)

- Notuð t.d. þegar telja þarf fjölda af tilteknum orðum, mörkum, o.s.frv. í texta.
- Upplagt að nota í dæmi 4.2 í Skiladæmi I.
- Sjá kafla 4.4.3 í kennslubók og t.d.

  `http://www.perl.com/doc/manual/html/pod/perlfunc/sort.html`

  varðandi röðun í tengslum við tætitöflu.

Outline

Tilreiðing (e. tokenisation)

- Texti brotinn upp í einstakar merkingarlegar máleiningar.
- Í flestum tilvikum upp í einstök orð.
- Gert með því að finna mörk orða, þ.e. staði þar sem einu orði lýkur og annað orð hefst.
- Tókar/lesmálsorð: orðin sem verða til við uppbrot textans.
- "Word segmentation"
  - Tilreiðing í málum þar sem mörk á milli orða eru ekki skýr.
  - T.d. þegar bil (e. whitespace) er ekki notað á milli orða.
  - Kínverska, Thai
- Við einbeitum okkur að tilreiðingu fyrir "Space-delimited languages".

Tilreiðing

Forritunarmál

- Hluti af lesgreiningu við þýðingu forritunarmála.
- Forritunarmál eru hins vegar hönnuð með það í huga að koma í veg fyrir margræðni – bæði varðandi einstök lesmálsorð og setningaskipan.

Náttúruleg mál

- Sami stafurinn getur þjónað mismunandi hlutverkum.
- Setningaskipan er ekki jafn ströng og í forritunarmálum.

Tilreiðing

Forritunarmál

- Hluti af lesgreiningu við þýðingu forritunarmála.
- Forritunarmál eru hins vegar hönnuð með það í huga að koma í veg fyrir margræðni – bæði varðandi einstök lesmálsorð og setningaskipan.

Náttúruleg mál

- Sami stafurinn getur þjónað mismunandi hlutverkum.
- Setningaskipan er ekki jafn ströng og í forritunarmálum.

Er tilreiðing ekkert mál?

- "Clairson International Corp. said it expects to report a net loss for its second quarter ended March 26 and doesn't expect to meet analysts' profit estimates of $3.9 to $4 million, or 76 cents a share to 79 cents a share, for its year ending Sept. 24."
- Punkturinn er hér notaður á þrjá mismunandi vegu. Hvenær er punktur hluti af tóka og hvenær ekki?
- ' notað á tvo mismunandi vegu.

Tilreiðing

Skammstafanir

- Bera þarf kennsl á skammstafanir.
- "Leitarvefurinn dohop.com hefur sett nýjustu lausn sína á ferðaáætlunum á markað í Bandaríkjunum. En veflausn íslenska hugbúnaðarfyrirtækisins dohop ehf. auðveldar fólki að gera ferðaáætlanir á netinu." (mbl.is, 08.02.2006)
- `http://is.wikipedia.org/wiki/Listi_yfir_algengar_skammstafanir_%C3%AD_%C3%ADslensku`

Fleiryrt orð (e. multiword expressions)

- Í einhverjum tilvikum gæti þurft að meðhöndla runu tóka sem einn tóka.
- *in spite of, aftur á móti, að auki, 26. mars*

Tilreiðing

Skammstafanir

- Bera þarf kennsl á skammstafanir.
- "Leitarvefurinn dohop.com hefur sett nýjustu lausn sína á ferðaáætlunum á markað í Bandaríkjunum. En veflausn íslenska hugbúnaðarfyrirtækisins dohop ehf. auðveldar fólki að gera ferðaáætlanir á netinu." (mbl.is, 08.02.2006)
- `http://is.wikipedia.org/wiki/Listi_yfir_algengar_skammstafanir_%C3%AD_%C3%ADslensku`

Fleiryrt orð (e. multiword expressions)

- Í einhverjum tilvikum gæti þurft að meðhöndla runu tóka sem einn tóka.
- *in spite of*, *aftur á móti*, *að auki*, *26. mars*

Outline

Setningaskipting (e. sentence segmentation)

- Texti brotinn upp í einstakar setningar.
- Ákvarða þarf mörk setninga.
  - Mörkin koma fyrir á milli einstakra orða.
  - "Sentence boundary detection"
- Við fyrstu sýn virðist þetta vera einfalt.
- Er ekki nægjanlegt að leita að ".", "?", "!"
- Og stundum ":", ";"
- Hvað með: "Ertu frá þér maður, og sjálfur sjómannadagurinn framundan!", segir prestsfrúin ...

Setningaskipting

- Dugar ekki einföld regla?
- delim = “.” | “!” | “?”

```
IF (right context = delim + space + capital letter OR
                    delim + quote + space + capital letter OR
                    delim + space + quote + capital letter)
THEN sentence boundary
```

- Skammstafanir geta gert setningaskiptingu erfiða:
  - “The contemporary viewer may simply ogle the vast wooded vistas rising up from the Saguenay River and Lac St. Jean, standing in for the St. Lawrence River.”
  - “The firm said it plans to sublease its current headquarters at 55 Water St. A spokesman declined to elaborate.”

Einföld setningagaskipting

- Ef t.d. lití er á punkt á undan bili sem enda setningar þá tekst að bera kennsl á um 90% af punktum sem enda setningar í Brown málheildinni (`http://en.wikipedia.org/wiki/Brown_Corpus`).
- Hægt að komast langt með einföldum reglulegum segðum án þess að nota lista af skammstöfunum.
- G.r.f. þremur tegundum af skammstöfunum (dæmi fyrir ensku):

```
A., B., C.          [A-Za-z]\.
U.S., m.p.h.        [A-Za-z]\.([A-Za-z]\.)+
Mr., St., Assn.     [A-Z][bcdfghj-np-tvxz]+\.
```

- Með þessu (ásamt leiðinni að ofan) tekst að bera kennsl á um 98% af setningaskiptingum í Brown málheildinni.

Greinar um tilreiðingu

- Sjá vefsíðu námskeiðs undir “Annað efni - Tilreiðing”.

Outline

Lesgreinir

- Lesgreinir (e. lexical analyser) er forrit sem greinir les (tóka) í texta.
- Forrit sem býr til lesgreini er kallað **lesgreinissmiður** (e. lexical analyser generator)
  - Dæmi: Lex/Flex/JFlex (`http://jflex.de/`)
  - Notandi skilgreinir reglulegar segðir.
  - Forritið býr til endanlegar stöðuvélar.
  - Stöðuvélarnar notaðar til að bera kennsl á tóka.

JFlex

Java kóði búinn til

- Tól sem býr til lesgreini (endanlega stöðuvél) út frá gefnum reglulegum segðum.
- Býr til Java kóða sem inniheldur stöðuvél (stöðuskiptatöflu).
- Inntak: JFlex frumforrit (t.d. Simple.flex)
- Úttak: Java kóði (t.d. Simple.java)

Java kóðinn þýddur og keyrður

- javac Simple.java (úttak Simple.class)
- java Simple textaskra

JFlex

Java kóði búinn til

- Tól sem býr til lesgreini (endanlega stöðuvél) út frá gefnum reglulegum segðum.
- Býr til Java kóða sem inniheldur stöðuvél (stöðuskiptatöflu).
- Inntak: JFlex frumforrit (t.d. Simple.flex)
- Úttak: Java kóði (t.d. Simple.java)

Java kóðinn þýddur og keyrður

- javac Simple.java (úttak Simple.class)
- java Simple textaskra

JFlex

Til að fá JFlex til að keyra

- Setja c:\jflex\bin í path.
- Breyta c:\jflex\bin\jflex.bat skránni í:
  - set JFLEX_HOME="C:\JFLEX"
  - REM for JDK 1.2
  - java -Xmx128m -jar %JFLEX_HOME%\lib\JFlex.jar

JFlex dæmi

```
%% Stöðuvél sem ber kennsl á (a|b)*abb

%public
%class Simple
%standalone
%unicode

%{
  String str = "Fann: ";
%}

Pattern = (a|b)*abb

%%
{Pattern}   { System.out.println(str + " " + yytext());}
.           { ;}
```

JFlex dæmi

```
%% Góður tilreiðari fyrir íslensku?

%public
%class Tokeniser1
%standalone
%unicode

%{
%}

WhiteSpace = [ \t\f\n]
Lower = [a-záéðíóúýþæö]
Upper = [A-ZÁÉÐÍÓÚÝÞÆÖ]
IceChar = {Upper}|{Lower}
IceWord = {IceChar}+

%%
{WhiteSpace}    {;}
{IceWord}       { System.out.println(yytext());}
.               { System.out.println(yytext());}
```

Outline

Unix/Linux tól

Ýmis konar Unix tól eru til sem auðvelda tilreiðingu á texta:

- **grep** (general regular expression parser)
- **tr** (translate characters)
- **sed** (string/stream edit)
- Og fleiri tól sem við skoðum síðar.

tc

- “translate characters”
- `http://en.wikipedia.org/wiki/Tr_(Unix)`
- tr set1 set2 < inputfile > outputfile
- Dæmi (breytir lágstöfum í hástafi):

```
tr '[a-z]' '[A-Z]' < inputfile > outputfile
```

- Íslenskum stöfum bætt við:

```
tr   '[a-z\341\346\351\355\360\363\366\372\375\376]'
     '[A-Z\301\306\311\315\320\323\326\332\335\336]'
     < inputfile > outputfile
```

  - “Octal values” : `http://en.wikipedia.org/wiki/Octal`

tc

- tr -d 'set1'
  - Eyðir stöfum í mengi 'set1'.
- tr -c 'set1' 'char2'
  - Breytir þeim stöfum sem ekki eru í mengi 'set1' yfir í 'char2'.
- tr -s set1 set2
  - Skiptir út stöfum í *set1* fyrir stafi í set2 og "suppress the output" (sérhver röð af endurteknum staf verður einn stafur).
- Dæmi (tilreiðari?):

```
tr -s ' ' '\012' < inputfile > outputfile

tr -cs '[a-z\341\346\351\355\360\363\366\372\375\376
        A-Z\301\306\311\315\320\323\326\332\335\336
        0-9.,!?]' '\012' < inputfile > outputfile
```

sed

- String/Stream editor:
  `http://www.grymoire.com/Unix/Sed.html`
- Vinnur á einni línu í einu í inntaksskrá.
- Nýtist vel þegar breyta þarf texta línu miðað við gefna reglulega segð.
- 's' for substitution:

```
sed 's/abc/(abc)/'  < inntak > uttak
sed 's/[a-z]*/(&)/' < inntak > uttak
```

  - & stendur fyrir strenginn sem finnst.

```
sed 's/[a-z]*/(&)/g' < inntak > uttak
```

  - 'g' fyrir "global replacement", ef breyta skal sérhverju mynstri í línunni (en ekki einungis því fyrsta).

sed

- Í sed stendur \n fyrir “newline”
- Hvað gerir eftirfarandi sed skipun:

  ```
  sed 's/\.$/\n\./' inntak.txt > uttak.txt
  ```
- Hægt að nota sed fyrir ýmislegt annað en breytingu á texta:
- sed 5q < inntak.txt > uttak.txt
- Prentar út fyrstu 5 línurnar og hættir ('q')
- sed '/^$/d' < inntak.txt > uttak.txt
- Eyðir tómum línum