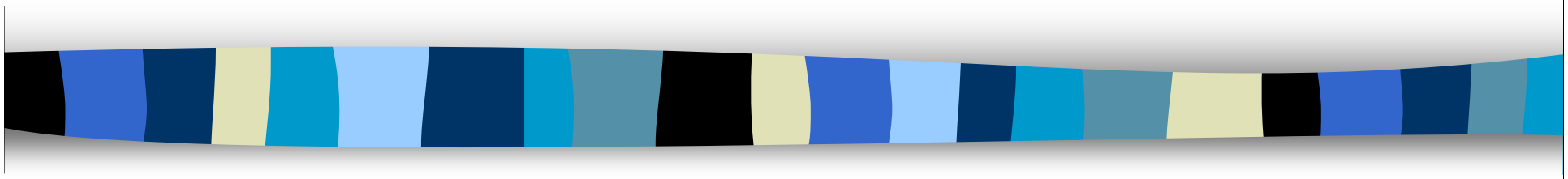
# Reguljära uttryck, *regular expressions*, regexp, RE



Peter Dalenius
petda@ida.liu.se
Institutionen för datavetenskap
Linköpings universitet
2008-01-21

# Vad är poängen?

- Den enklaste formen av formellt språk
- Används oftast för att söka i textsträngar
- Exempel:
  - *wildcards* är en enklare form
  - *grep* m.fl. kommandon tar reguljära uttryck
  - scriptspråk, t.ex. Perl
- Det finns några olika uppsättningar, olika kraftfulla, en del till och med mer än ”riktiga” reguljära uttryck

# Reguljära uttryck i Ruby

```
irb(main):001:0> /be/ =~ "To be, or not to be..."
=> 3
irb(main):002:0> /o+/ =~ "Moo mooo moooo!"
=> 1
irb(main):003:0> "23, 18, 45" =~ /1[0-9]/
=> 4
irb(main):004:0> /\w\d/ =~ "ABC123"
=> 2
irb(main):005:0> /kul/ =~ "Inget roligt här!"
=> nil
```

# Enskilda tecken

- Alla tecken utom \/^$|.+*?()[]{} matchar sig själva. Dessa måste föregås av \ om man vill matcha dem.
- . matchar ett godtyckligt tecken.
- [*characters*] matchar ett av de uppräknade tecknen.
  - [aeiouyåääö] matchar en vokal
  - [a-zA-Z] matchar en (engelsk) bokstav
- [^*characters*] matchar ett tecken som inte ingår i uppräkningen
- Det finns genvägar för några klasser av tecken:
  - \w matchar ett alfanumeriskt tecken
  - \d matchar en siffra
  - \s matchar ett *white space*

# Sammansättning

- Låt **a** och **b** vara två reguljära uttryck:
  - **ab** matchar de två strängarna efter varann
  - **a|b** matchar endera av **a** eller **b**
  - **a*** matchar noll eller fler förekomster av **a**
  - **a+** matchar en eller fler förekomster av **a**
  - **a**{*m*} matchar exakt *m* förekomster av **a**
  - **a**{*m*,} matchar minst *m* förekomster av **a**
  - **a**{*m*,*n*} matchar *m* till *n* förekomster av **a**

# Förankring

- ^ och $ matchar början/slutet av en rad
- \A och \z matchar början/slutet av strängen
- \b matchar ordgränser (\B tvärtom)

# Funktioner som använder regexp

```
irb(main):010:0> s = "The stars, like dust"
=> "The stars, like dust"
irb(main):011:0> s.sub(/e/,'$')
=> "Th$ stars, like dust"
irb(main):012:0> s.gsub(/[aeiouy]/,'*')
=> "Th* st*rs, l*k* d*st"
irb(main):013:0> s.gsub(/(^|,|\s)\w/)
                        { |m| m.upcase }
=> "The Stars, Like Dust"
irb(main):014:0> s.split(/[\s,]+/)
=> ["The", "stars", "like", "dust"]
irb(main):015:0> s.scan(/[aeiouy]\w/)
=> ["ar", "ik", "us"]
```

# Resultat av matchning

```
irb(main):021:0> /[aeiouy]{2,}/ =~
                                  "the moon is a cheese"
=> 5
irb(main):022:0> [$`, $&, $']
=> ["the m", "oo", "n is a cheese"]
irb(main):023:0> /(\d\d):(\d\d):(\d\d)/ =~
                                "Klockan är 12:35:40 nu"
=> 12
irb(main):024:0> [$1, $2, $3]
=> ["12", "35", "40"]
```

# Resultat av matchning

```
irb(main):025:0> s = "the moon is a cheese"
=> "the moon is a cheese"
irb(main):026:0> re = /[aeiouy]{2,}/
=> /[aeiouy]{2,}/
irb(main):027:0> md = re.match(s)
=> #<MatchData:0xdf65ed4>
irb(main):028:0> [md.pre_match, md[0],
                                  md.post_match]
=> ["the m", "oo", "n is a cheese"]
irb(main):029:0> md2 =
                    /(\d\d):(\d\d)/.match("14:45")
=> #<MatchData:0xdb94f94>
irb(main):030:0> md2.captures
=> ["14", "45"]
```