Methods in Computational Linguistics

Lab: Regular Expressions

Diego Frassinelli

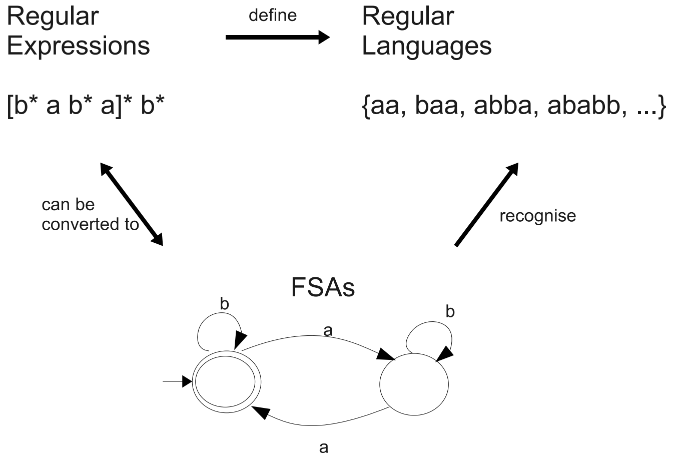
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

December 04 2019

Today's slides are based on the materials provided by:

- Roman Klinger
- Christian Scheible

Regular Expressions



Regular Expressions

[b* a b* a]* b*

define

Regular Languages

{aa, baa, abba, ababb, ...}

can be converted to

recognise

FSAs

- An algebraic notation for characterizing a set of strings (J&M)
- We use them as search patterns in text

Regular Expressions: Correctness

You can test the correctness of your Regular Expression using the terminal

```
> echo ”Look looook!” | grep −E ”Look”
> Look looook!
```

```
> echo ”Look looook!” | grep −E ”o+”
> Look looook!
```

Alternatives

Pattern	Matches	Example
[Ww]oodchuck	Woodchuck woodchuck	the woodchuck is a bird
[0123456789]	any digit	my number is 0170012345

Pattern	Matches	Example
[A-Z]	Uppercase characters	from London
[a-z]	Lowercase characters	mY KEY BOARD IS BROKEN
[0-9]	all digits	Chapter 1: Down the Rabbit Hole
.	all characters	Chapter 1: Down the Rabbit Hole

Alternatives: Exercises

- Fish or Dish
- All vowels
- All lower case letters
- All lower case letters and numbers
- Three arbitrary characters after an uppercase letter

Negations

Pattern	Matches	Example
[^A-Z]	Not uppercase letters	from London
[^Ss]	Neither “s” nor “S”	SSSssSssSo...
[e^]	either “e” or “^”	Look here ^^
a^b	the pattern “a carat b”	Look up a^b now

Negations: Exercises

- Not v
- Not vowels
- Neither lowercase nor number

Disjunctions

Pattern	Matches	Example
a|b|c	“a”, “b” or “c”	a black cat
cat|dog	“cat” or “dog”	concatenate strings about dogs
party|ies	“party” or “ies”	a party in the seventies
part(y|ies)	“party” or “parties”	a party in the seventies

Disjunctions: Exercises

- fox or foxes

Repetitions

Pattern	Matches	Example
colou?r	the previous item is optional	You can spell it color or colour
coo*l!	0 or more of previous chars	col! cool! coool! cooool!
coo+l!	1 or more of previous chars	col! cool! coool! cooool!
coo{3}l!	3 of previous chars	col! coool! cooool!
coo{3,5}l!	3 to 5 of previous chars	cool! cooool! cooooool!
coo{3,}l!	3 or more of previous chars	coool! cooool!, coooooool!

Repetitions: Exercises

- An a followed by any number of b
- any number of a followed by any number of b
- Course or courses
- Any sequence of letters ending with an x
- At least one a and at least two b

Anchors

Pattern	Matches	Example
^[A-Z]	Initial uppercase letter	London - Paris
^[^A-Za-z]	Initial non-alphabetic char	”Hello”
\.$	Final full-stop	the end.
.$	Any final character	the end

Escaping: if special symbols should be found as character, put a “\” in front of it

Anchors: Exercises

- A string starting with an uppercase and ending with a question mark or an exclamation mark

Order of Application

1. Grouping: ()
2. Repetitions: * + ? {}
3. Anchors: ^$
4. Disjunctions: |