Compilers and Formal Languages

| Email: | christian.urban at kcl.ac.uk |
|-----------------|---|
| Office Hour: | Friday 12 – 14 |
| Location: | N7.07 (North Wing, Bush House) |
| Slides & Progs: | KEATS |
| Pollev: | <pre>https://pollev.com/cfltutoratki576</pre> |
| | |

| 1 Introduction, Languages | 6 While-Language |
|------------------------------------|----------------------------------|
| 2 Regular Expressions, Derivatives | 7 Compilation, JVM |
| 3 Automata, Regular Languages | 8 Compiling Functional Languages |
| 4 Lexing, Tokenising | 9 Optimisations |
| 5 Grammars, Parsing | 10 LLVM |

For Installation Problems

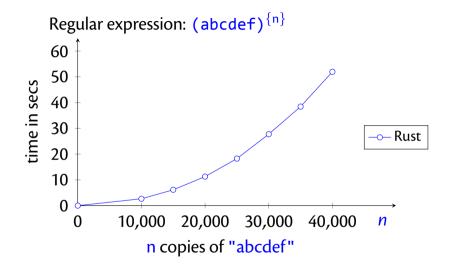
- Harry Dilnot (harry.dilnot@kcl.ac.uk) Windows expert
- Oliver Iliffe (oliver.iliffe@kcl.ac.uk)

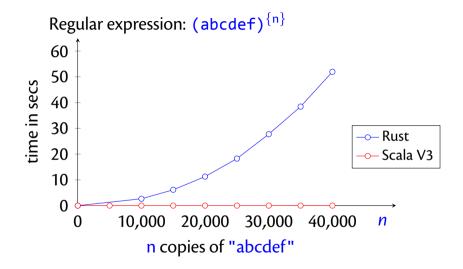
From Pollev last week

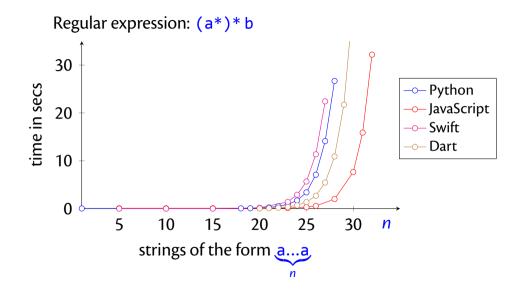
Is the equivalence of two regexes belong in the P or NP class of problems?

From Pollev last week

If state machines are not efficient, then how/why do many lexer packages like the logos crate in rust compile down a lexer definition down to a jump table driven state machine? Could we achieve quicker lexing with things like SIMD instructions?



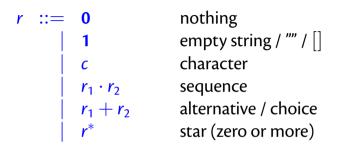




From Pollev last week

For a regular expression $r = r_1 \cdot r_2$, to prove that der c $r = (der c r) \cdot r^{\{n-1\}}$, is there a way to prove it in the general case instead of how you do the calculations for each n in the videos?

(Basic) Regular Expressions



How about ranges [a-z], r^+ and $\sim r$? Do they increase the set of languages we can recognise?



Assume you have an alphabet consisting of the letters *a*, *b* and *c* only. Find a (basic!) regular expression that matches all strings *except ab* and *ac*!

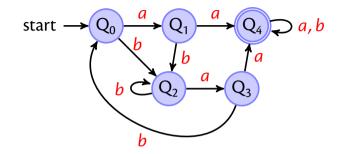
Automata

A deterministic finite automaton, DFA, consists of:

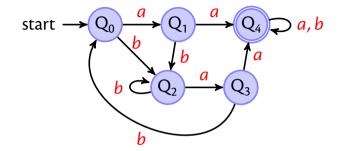
- an alphabet Σ
- a set of states **Qs**
- one of these states is the start state Q_0
- some states are accepting states *F*, and
- there is transition function δ

which takes a state as argument and a character and produces a new state; this function might not be everywhere defined \Rightarrow partial function

 $\mathsf{A}(\boldsymbol{\Sigma},\mathsf{Qs},\mathsf{Q_0},\mathsf{F},\boldsymbol{\delta})$



- the start state can be an accepting state
- it is possible that there is no accepting state
- all states might be accepting (but this does not necessarily mean all strings are accepted)



for this automaton δ is the function

$$\begin{array}{ccc} (\mathbf{Q}_0,a) \to \mathbf{Q}_1 & (\mathbf{Q}_1,a) \to \mathbf{Q}_4 & (\mathbf{Q}_4,a) \to \mathbf{Q}_4 \\ (\mathbf{Q}_0,b) \to \mathbf{Q}_2 & (\mathbf{Q}_1,b) \to \mathbf{Q}_2 & (\mathbf{Q}_4,b) \to \mathbf{Q}_4 \end{array} \cdots$$

Accepting a String

Given

 $\mathsf{A}(\varSigma,\mathsf{Qs},\mathsf{Q_0},\mathit{F},\delta)$

you can define

$$\widehat{\delta}(Q, []) \stackrel{\text{def}}{=} Q$$
$$\widehat{\delta}(Q, c :: s) \stackrel{\text{def}}{=} \widehat{\delta}(\delta(Q, c), s)$$

Accepting a String

Given

 $\mathsf{A}(\boldsymbol{\Sigma}, \mathbf{Q}\mathbf{s}, \mathbf{Q}_\mathbf{0}, \mathbf{F}, \boldsymbol{\delta})$

you can define

$$\widehat{\delta}(\mathbf{Q}, []) \stackrel{\text{def}}{=} \mathbf{Q}$$
$$\widehat{\delta}(\mathbf{Q}, \mathbf{c} :: \mathbf{s}) \stackrel{\text{def}}{=} \widehat{\delta}(\delta(\mathbf{Q}, \mathbf{c}), \mathbf{s})$$

Whether a string s is accepted by A?

$$\widehat{\delta}(\mathbf{Q_0}, \mathbf{s}) \in \mathbf{F}$$

Regular Languages

A language is a set of strings.

A regular expression specifies a language.

A language is **regular** iff there exists a regular expression that recognises all its strings.

Regular Languages

A language is a set of strings.

A regular expression specifies a language.

A language is **regular** iff there exists a regular expression that recognises all its strings.

not all languages are regular, e.g. $a^n b^n$ is not

Regular Languages (2)

A language is **regular** iff there exists a regular expression that recognises all its strings.

or **equivalently**

A language is **regular** iff there exists a deterministic finite automaton that recognises all its strings.

Non-Deterministic Finite Automata

 $N(\Sigma, \mathrm{Qs}, \mathrm{Qs}_{\mathrm{0}}, \mathrm{F}, \rho)$

A non-deterministic finite automaton (NFA) consists of:

- a finite set of states, Qs
- <u>some</u> these states are the start states, Qs₀
- some states are accepting states, and
- there is transition relation, ρ

 $\begin{array}{c} (\mathsf{Q}_1,a) \to \mathsf{Q}_2 \\ (\mathsf{Q}_1,a) \to \mathsf{Q}_3 \end{array} \cdots$

Non-Deterministic Finite Automata

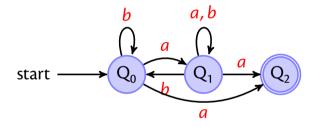
 $N(\Sigma, \mathrm{Qs}, \mathrm{Qs}_{\mathrm{0}}, \mathrm{F}, \rho)$

A non-deterministic finite automaton (NFA) consists of:

- a finite set of states, Qs
- <u>some</u> these states are the start states, Qs₀
- some states are accepting states, and
- there is transition relation, ρ

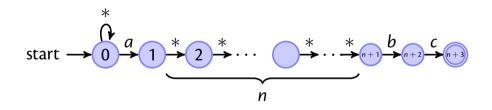
$$\begin{array}{ll} (\mathbf{Q}_1, a) \to \mathbf{Q}_2 \\ (\mathbf{Q}_1, a) \to \mathbf{Q}_3 \end{array} \dots \qquad (\mathbf{Q}_1, a) \to \{\mathbf{Q}_2, \mathbf{Q}_3\} \end{array}$$

An NFA Example



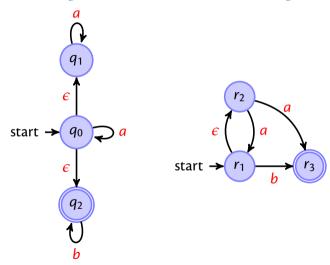
Another Example

For the regular expression $(.^*)a(.^{\{n\}})bc$

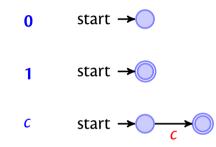


Note the star-transitions: accept any character.

Two Epsilon NFA Examples

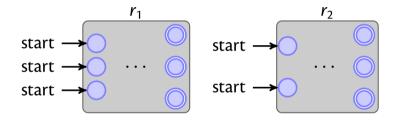


Thompson: Rexp to ϵ **NFA**





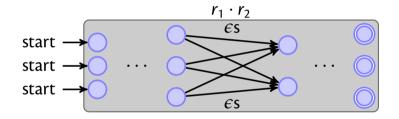
By recursion we are given two automata:



We need to (1) change the accepting nodes of the first automaton into non-accepting nodes, and (2) connect them via ϵ -transitions to the starting state of the second automaton.



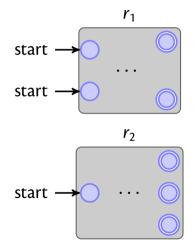
By recursion we are given two automata:



We need to (1) change the accepting nodes of the first automaton into non-accepting nodes, and (2) connect them via ϵ -transitions to the starting state of the second automaton.

Case $r_1 + r_2$

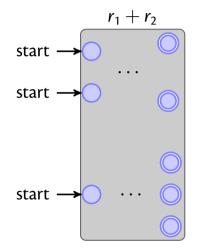
By recursion we are given two automata:



We can just put both automata together.

Case $r_1 + r_2$

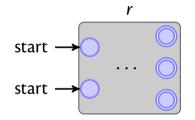
By recursion we are given two automata:



We can just put both automata together.

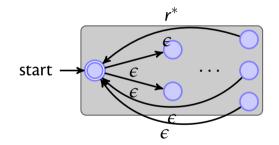


By recursion we are given an automaton for *r*:



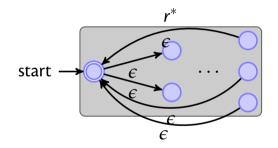


By recursion we are given an automaton for *r*:

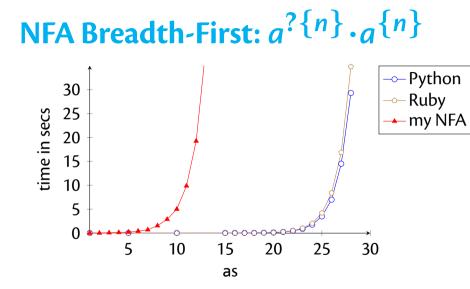




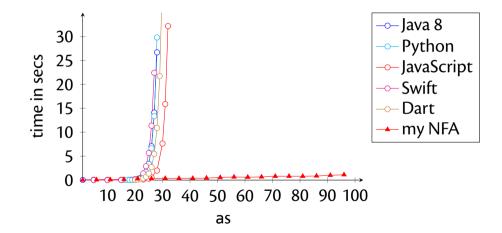
By recursion we are given an automaton for *r*:

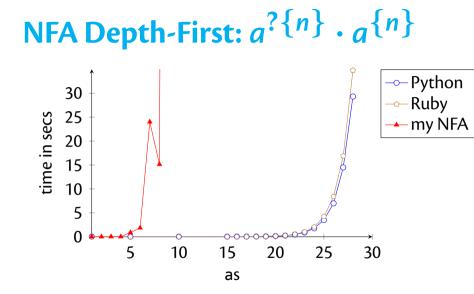


Why can't we just have an epsilon transition from the accepting states to the starting state?

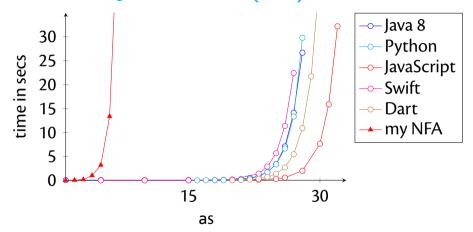


NFA Breadth-First: $(a^*)^* \cdot b$

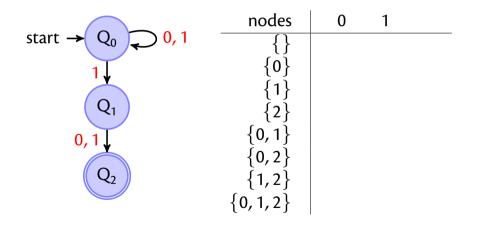


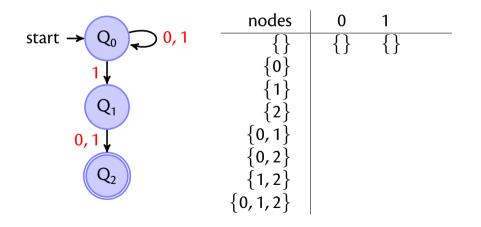


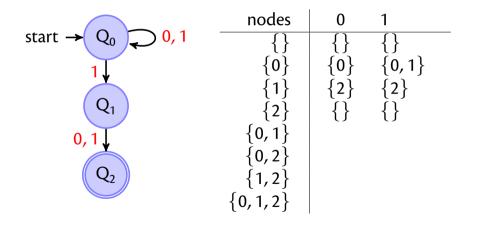
NFA Depth-First: $(a^*)^* \cdot b$

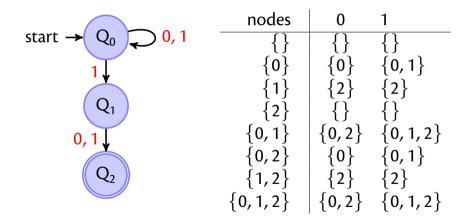


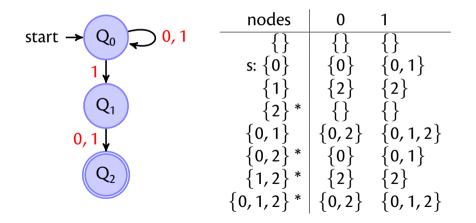
The punchline is that many existing libraries do depth-first search in NFAs (with backtracking).

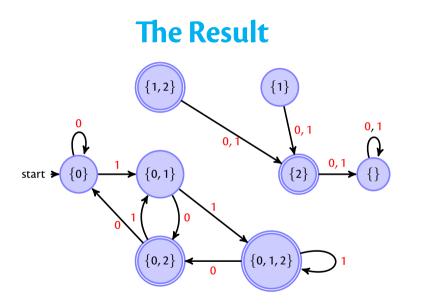




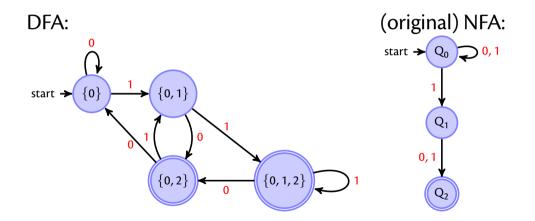


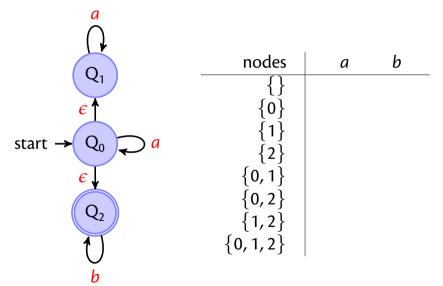


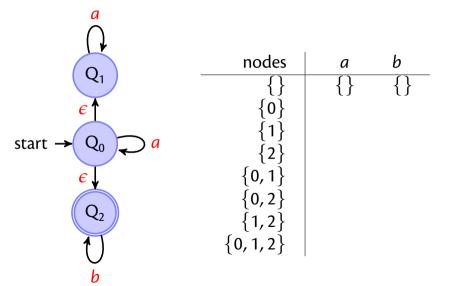


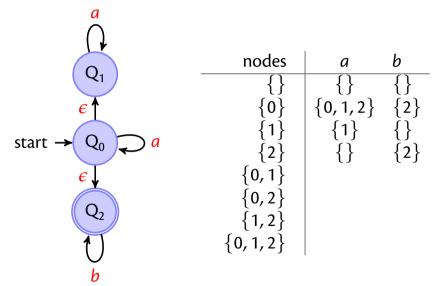


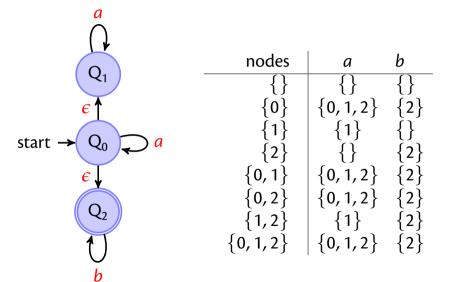
Removing Dead States

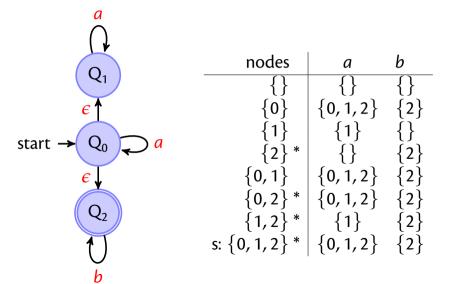




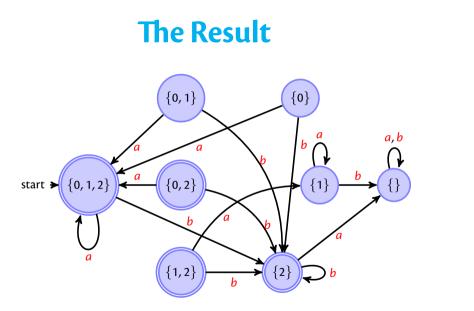




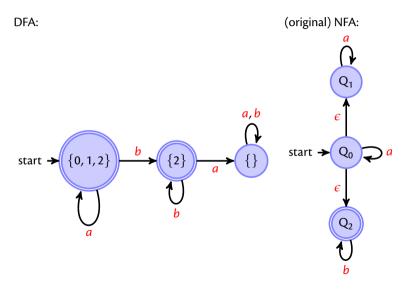




CFL 03, King's College London – p. 31/6



Removing Dead States



Regexps and Automata

Thompson's subset construction

Regexps → NFAs → DFAs

Regexps and Automata

Thompson's subset construction

minimisation

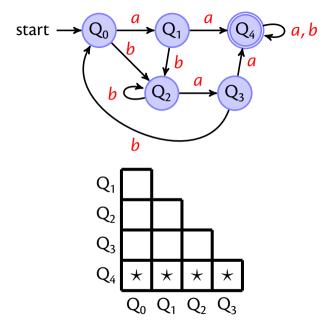
DFA Minimisation

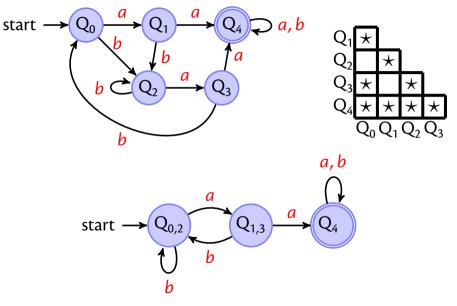
- 1. Take all pairs (q, p) with $q \neq p$
- 2. Mark all pairs that accepting and non-accepting states
- 3. For all unmarked pairs (q, p) and all characters c test whether

 $(\delta(q,c),\delta(p,c))$

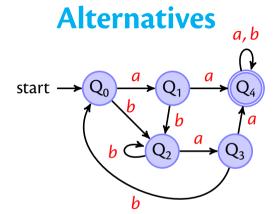
are marked. If yes in at least one case, then also mark (q, p).

- 4. Repeat last step until no change.
- 5. All unmarked pairs can be merged.

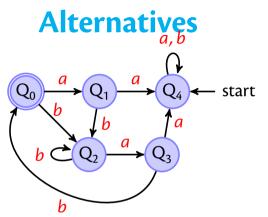




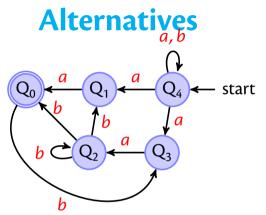
CFL 03, King's College London – p. 37/6



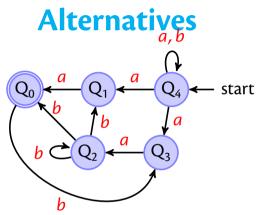
• exchange initial / accepting states



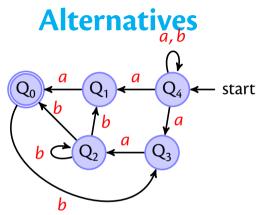
- exchange initial / accepting states
- reverse all edges



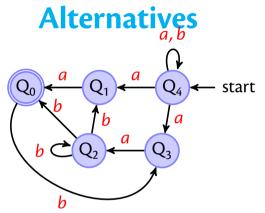
- exchange initial / accepting states
- reverse all edges
- subset construction \Rightarrow DFA



- exchange initial / accepting states
- reverse all edges
- subset construction \Rightarrow DFA
- remove dead states



- exchange initial / accepting states
- reverse all edges
- subset construction \Rightarrow DFA
- remove dead states
- repeat once more



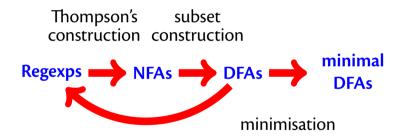
- exchange initial / accepting states
- reverse all edges
- subset construction \Rightarrow DFA
- remove dead states
- repeat once more \Rightarrow minimal DFA

Regexps and Automata

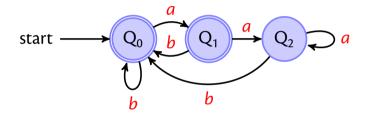
Thompson's subset construction

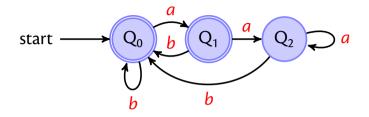
minimisation

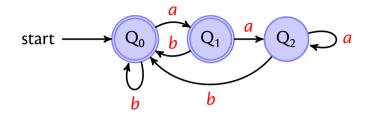
Regexps and Automata



DFA to Rexp

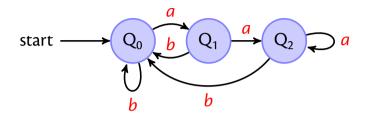


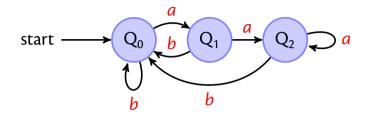




You know how to solve since school days, no?

$$\begin{array}{l} Q_0 \,=\, 2\,Q_0 + 3\,Q_1 + 4\,Q_2 \\ Q_1 \,=\, 2\,Q_0 + 3\,Q_1 + 1\,Q_2 \\ Q_2 \,=\, 1\,Q_0 + 5\,Q_1 + 2\,Q_2 \end{array}$$

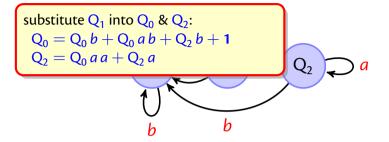




$$Q_{0} = Q_{0} b + Q_{1} b + Q_{2} b + 1$$

$$Q_{1} = Q_{0} a$$

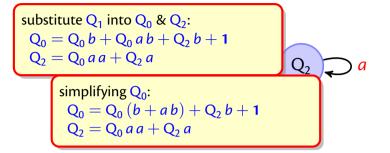
$$Q_{2} = Q_{1} a + Q_{2} a$$



$$Q_{0} = Q_{0} b + Q_{1} b + Q_{2} b + 1$$

$$Q_{1} = Q_{0} a$$

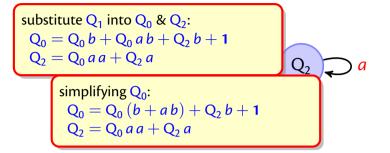
$$Q_{2} = Q_{1} a + Q_{2} a$$



$$Q_{0} = Q_{0} b + Q_{1} b + Q_{2} b + 1$$

$$Q_{1} = Q_{0} a$$

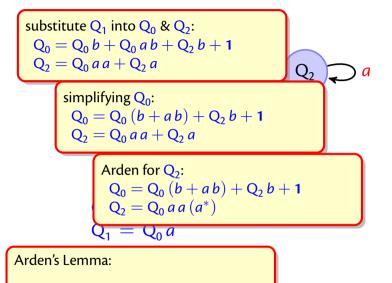
$$Q_{2} = Q_{1} a + Q_{2} a$$



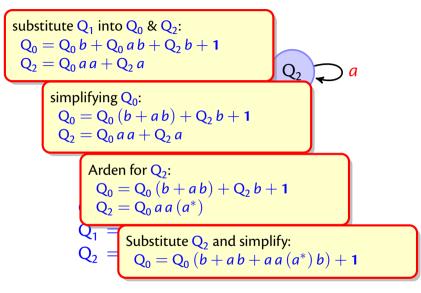
$$Q_0 = Q_0 b + Q_1 b + Q_2 b + 1 Q_1 = Q_0 a$$

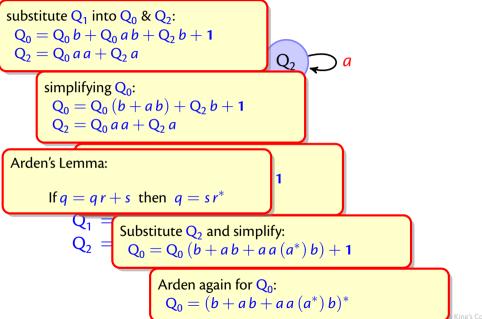
Arden's Lemma:

If
$$q = qr + s$$
 then $q = sr^*$

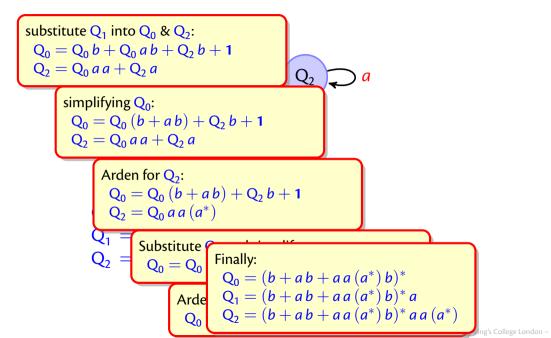


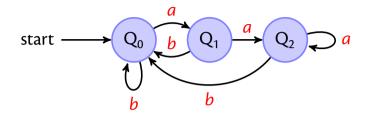
If
$$q = qr + s$$
 then $q = sr^*$





King's College London – p. 42/6





$$Q_{0} = Q_{0} b + Q_{1} b + Q_{2} b + 1$$

$$Q_{1} = Q_{0} a$$

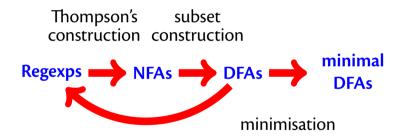
$$Q_{2} = Q_{1} a + 0$$
Finally:
$$Q_{0} = (b + a b + a a (a^{*}) b)^{*}$$

$$Q_{1} = (b + a b + a a (a^{*}) b)^{*} a$$

$$Q_{2} = (b + a b + a a (a^{*}) b)^{*} a a (a^{*})$$

ing's College London – p. 42/6

Regexps and Automata



Regular Languages (3)

A language is **regular** iff there exists a regular expression that recognises all its strings.

or equivalently

A language is **regular** iff there exists a deterministic finite automaton that recognises all its strings.

Regular Languages (3)

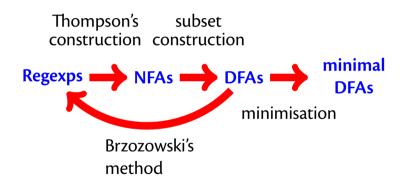
A language is **regular** iff there exists a regular expression that recognises all its strings.

or equivalently

A language is **regular** iff there exists a deterministic finite automaton that recognises all its strings.

Why is every finite set of strings a regular language?

Regexps and Automata



Regular Languages

Two equivalent definitions:

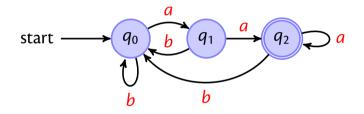
A language is regular iff there exists a regular expression that recognises all its strings.

A language is regular iff there exists an automaton that recognises all its strings.

for example $a^n b^n$ is not regular



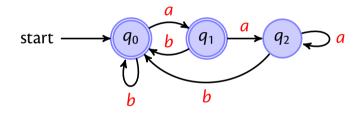
Regular languages are closed under negation:



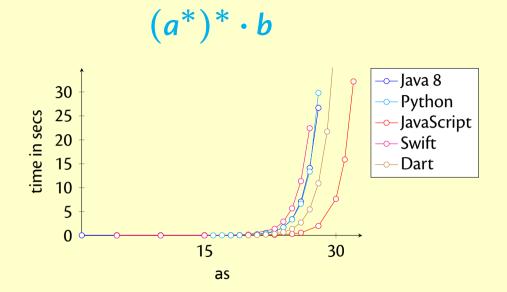
But requires that the automaton is completed!

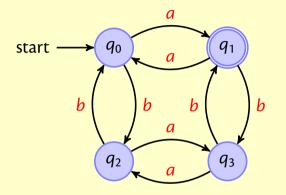


Regular languages are closed under negation:



But requires that the automaton is completed!





Which language?

CW1: Regexes and L-function Given

| <i>r</i> + | $L(r^+)$ | $\stackrel{\text{\tiny def}}{=} \bigcup_{1 \leq i} . L(r)^i$ |
|------------------|-----------------|---|
| <i>r</i> ? | $L(r^{?})$ | $\stackrel{\text{\tiny def}}{=} L(r) \cup \{[]\}$ |
| $r_1 \& r_2$ | $L(r_1\&r_2)$ | $\stackrel{\text{\tiny def}}{=} L(r_1) \cap L(r_2)$ |
| r {n} | $L(r^{\{n\}})$ | $\stackrel{\text{\tiny def}}{=} L(r)^n$ |
| r{m} | $L(r^{\{m\}})$ | $\stackrel{\text{\tiny def}}{=} \bigcup_{0 \le i \le m} . L(r)^i$ |
| r ^{n} | $L(r^{\{n\}})$ | $\stackrel{\text{\tiny def}}{=} \bigcup_{n \leq i} L(r)^i$ |
| $r^{\{nm\}}$ | $L(r^{\{nm\}})$ | $\stackrel{\text{\tiny def}}{=} \bigcup_{n \leq i \leq m} . L(r)^i$ |
| \sim r | $L(\sim r)$ | $\stackrel{	ext{def}}{=} \Sigma^* - L(r)$ |

Nullable

 $nullable(r^+) \stackrel{\text{def}}{=} nullable(r)$ $nullable(r^?) \stackrel{\text{def}}{=} true$ nullable($r^{\{n\}}$) $nullable(r^{\{..m\}})$ $nullable(\sim r) \stackrel{\text{def}}{=} ! nullable(r)$

nullable $(r_1 \& r_2) \stackrel{\text{def}}{=} \text{nullable}(r_1) \land \text{nullable}(r_2)$ $\stackrel{\text{def}}{=}$ if n = 0 then true else nullable(r) $\stackrel{\text{\tiny def}}{=}$ true nullable $(r^{\{n..\}}) \stackrel{\text{def}}{=} if n = 0$ then true else nullable(r)nullable $(r^{\{n..m\}}) \stackrel{\text{def}}{=} if n = 0$ then true else nullable(r)

Derivative

```
\stackrel{\text{\tiny def}}{=} (\operatorname{der} \operatorname{c} r) \cdot r^*
der c (r^+)
der c (r^?) \stackrel{\text{def}}{=} der c r
der c (r_1 \& r_2) \stackrel{\text{def}}{=} (der c r_1) \& (der c r_2)
der c(r^{\{n\}}) \stackrel{\text{def}}{=} if n = 0 then 0 else (der c r) \cdot r^{\{n-1\}}
der c(r^{\{..m\}}) \stackrel{\text{def}}{=} if m = 0 then 0 else (der c r) \cdot r^{\{..m-1\}}
der c(r^{\{n..\}}) \stackrel{\text{def}}{=} if n = 0 then (der c r) \cdot r^* else (der c r) \cdot r^{\{n-1..\}}
der c (r^{\{n..m\}})
                             \stackrel{\text{\tiny def}}{=} if n = 0 \land m = 0 then 0 else
                                    if n = 0 then (\operatorname{der} c r) \cdot r^{\{\dots m-1\}} else (\operatorname{der} c r) \cdot r^{\{n-1,\dots m-1\}}
                        \stackrel{\text{\tiny def}}{=} \sim (\operatorname{der} \operatorname{cr})
der c (\sim r)
```