SAPIENZA
UNIVERSITÀ DI ROMA
DIPARTIMENTO DI INFORMATICA

FOUNDATIONS OF COMPUTER SCIENCE
LECTURE 5: Non-Regular Languages

Prof. Daniele Gorla

Not all languages are regular

- By now, a regular language can be characterized as
  - One that can be recognized to a deterministic/non-deterministic Finite Automaton
  - One that is associated to a regular expression
  - One that can be generated by a regular grammar

- The question now is: can any language be characterized by at least one of these 3 ways?

- The answer is NO

- Consider the language $B = \{0^n1^n \mid n \geq 0\} = \{\varepsilon, 01, 0011, 000111, \dots\}$ (the same number of 0s and 1s)

- Intuitively, a DFA that recognizes B should remember how many 0s have been seen so far (one state for every natural number)

this DFA has an infinite number of states, not possible by definition

→ But it cannot, since it has a **finite** number of states.

Need for a formal proof

Notice that just the idea of «counting» something a possibily unbounded number of times doesn't necessarily imply not being regular

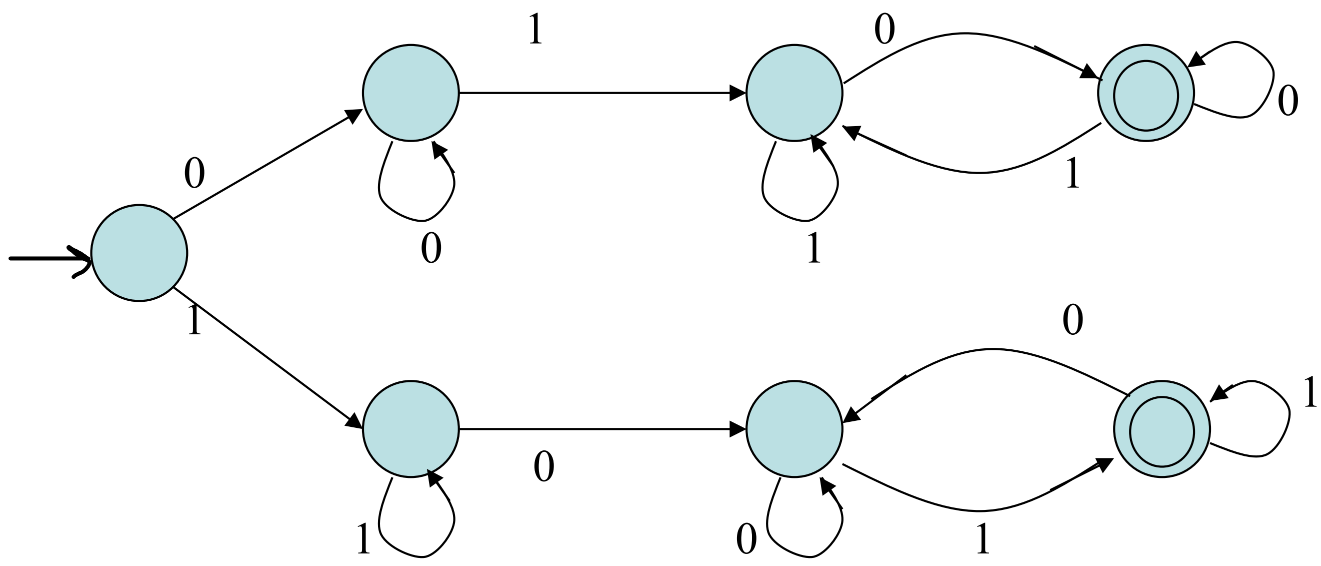
EXAMPLE:

The language $C = \{w \mid w$ has an equal number of 0s and 1s$\}$ is NOT regular

whereas

The language $D = \{w \mid w$ has an equal number of occurrences of 01 and 10 as substrings$\}$ IS!

Indeed, D is accepted by the followign DFA:



prove that this is equivalent to strings starting and ending with the same character as exercise

if this is true, the regex is
0* U 1* U 0(0 U 1)*0 U 1(0 U 1)*1

The Pumping Lemma

ONLY for infinite languages

Pumping lemma If A is a regular language, then there is a number p (the pumping length) where if s is any string in A of length at least p, then s may be divided into three pieces, $s = xyz$, satisfying the following conditions:

1. for each $i \geq 0$, $xy^iz \in A$,
2. $|y| > 0$, and
3. $|xy| \leq p$.

you can have y infinitely many times and this will still be in the regular language (y^i CAN be epsilon, with i=0, however y CANNOT)

- When s is divided into xyz, either x or z may be ε, but not y
  (without this condition the theorem would be trivially true!)

this is because if y is epsilon, the 1st condition is uninformative

- Condition 3 states that the pieces x and y together have length at most p; this is an extra technical condition that we occasionally find useful when proving certain languages to be nonregular.

- Written in a more precise way, the P.L. is:

$$A \text{ regular} \Rightarrow \exists\, p \in \mathbb{N} : \forall\, s \in A \; (|s| \geq p \;\Rightarrow\; \exists\, x,y,z \text{ s.t. } (s = xyz \;\wedge\; |y| > 0 \;\wedge\; |xy| \leq p \;\wedge\; \forall i \in \mathbb{N}.\, xy^iz \in A))$$

if

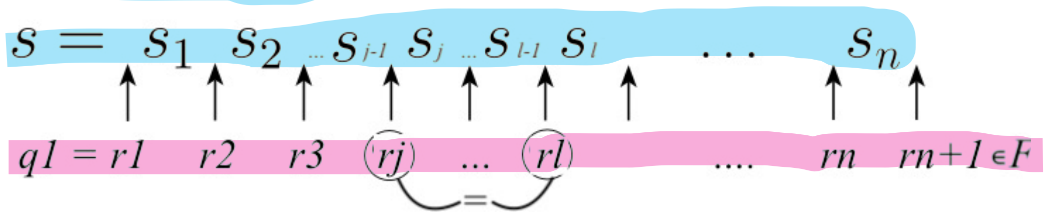
Proof of the Pumping Lemma

PROOF Let $M = (Q, \Sigma, \delta, q_1, F)$ be a DFA recognizing A and p be the number of states of M.

Let $s = s_1 s_2 \cdots s_n$ be a string in A of length n, where $n \geq p$. a string longer than the # of states

Let $r_1, \ldots, r_{n+1}$ be the sequence of states that M enters while processing s, so $r_{i+1} = \delta(r_i, s_i)$ for $1 \leq i \leq n$.
This sequence has length $n + 1$, which is at least $p + 1$.
Among the first $p + 1$ elements in the sequence, two must be the same state
We call the first of these r_j and the second r_l.

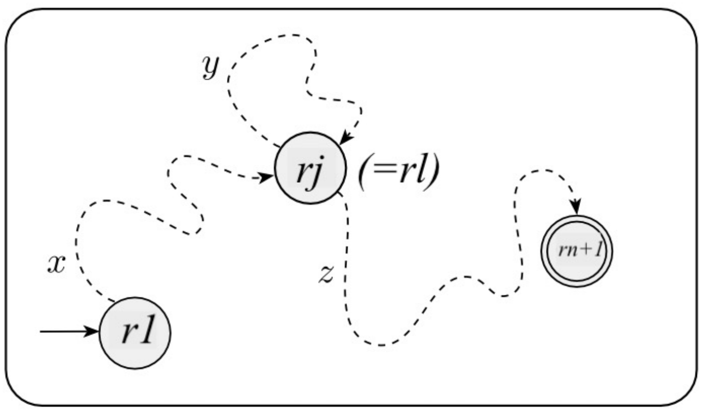
this is true by definition of the string. length is at least p and I have at least p+1 states traversed, so two states are equal.



Now let $x = s_1 \cdots s_{j-1}$, $y = s_j \cdots s_{l-1}$, and $z = s_l \cdots s_n$. so $|xy| \leq p$.
As $j \neq l$, so $|y| > 0$;
M must accept $xy^i z$ for $i \geq 0$.

pigeons-hole principle and injective functions

M



Q.E.D.

Usage of the Pumping Lemma

P.L.: A regular $\Rightarrow \exists\, p \in \mathbb{N}\ \forall\, s \in A\ (\, |s| \geq p \ \Rightarrow\ \exists\, x,y,z$ s.t. $(\, s = xyz\ \wedge$
$|y| > 0\ \wedge\ |xy| \leq p\ \wedge$
$\forall i \in \mathbb{N}.\ xy^i z \in A)\,)$

Hence, the contrapositive of this statement is

$\forall\, p \in \mathbb{N}\ \exists\, s \in A\ (\, |s| \geq p \wedge \forall\, x,y,z\ (s \neq xyz \vee |y| = 0 \vee |xy| > p \vee \exists\, i \in \mathbb{N}.\ xy^i z \notin A)$

$\Rightarrow A$ is not regular

can't split the string s.t. the properties are met (y is empty or xy are longer than p or xy^iz does not belong to A for some i)

Equivalently:

$\forall\, p \in \mathbb{N}\ \exists\, s \in A\ (\, |s| \geq p \wedge \forall\, x,y,z\ ((s = xyz \wedge |y| > 0 \wedge |xy| \leq p) \Rightarrow \exists\, i \in \mathbb{N}.\ xy^i z \notin A)$

$\Rightarrow A$ is not regular

basically for every decomposition, I can find an index s.t. there exists an i that makes xy^iz not part of A. In the event that some decompositions are part of A and some not, this is inconclusive (ex. 0^2n).

Practical use (for proving that A is not regular):

in that case i should try to find regularity. of course if i find regularity before (and that includes all the 5 ways to do so), I can omit the use of the PL

- Consider a generic p
- Find a string $s \in A$ long at least p and decompose it in all possible xyz, with $|y|>0$ and $|xy| \leq p$
- For each such decomposition, find an i such that $xy^i z \notin A$
- Then, A is not regular

Example of Usage (1)

SAPIENZA
Università di Roma
Dipartimento di Informatica

Let us prove that $B = \{0^n 1^n \mid n \geq 0\}$ is not regular

Let's fix a generic p and choose s to be the string $0^p 1^p$. 0^p has len p

Consider all possible decompositions into three pieces, $s = xyz$, and show that the string xy^iz is not in B for some $i \geq 0$.

1. y consists only of 0s: In this case, $xyyz$ has more 0s than 1s and so is not a member of B
2. y consists only of 1s: as before.
3. y consists of both 0s and 1s: In this case, $xyyz$ may have the same number of 0s and 1s, but they will be out of order with some 1s before 0s. Hence it is not a member of B. the structure is broken. e.g. y=01, xyyz=0....0101....11, violating B.

REMARK: by using Condition 3 of the P.L., the only possible case to consider is 1
xy terminates at most at 0^p and xy need to be at most p, as we need to have space for z
in this case z cannot be epsilon, because the whole string length violates condition 3

Example of Usage (2)

Let us prove that $C = \{w \mid w$ has an equal number of 0s and 1s$\}$ is not regular.

Fix p and choose s to be the string 0^p1^p.

In considering all possible decompositions, $s = xyz$, remember that $|xy| \leq p$.

Hence, y must be made up only of 0s and so $xyyz$ doesn't belong to C.

REMARK: Here the choice of s is not obvious and another choice of s could have made the proof not working

EXAMPLE: choose $s = (01)^p$ and consider the decomposition

$$x = \varepsilon \qquad y = 01 \qquad z = (01)^{p-1}$$

Then $xy^iz = (01)^i(01)^{p-1} \in C$ for every value of i.

REMARK: Alternative proof that C is nonregular:

- We know that $B = \{0^n1^n\}$ is nonregular.
- If C were regular, $C \cap 0^*1^*$ also would be regular (since 0^*1^* is regular and regular languages are closed under intersection).
- But $C \cap 0^*1^* = B$: CONTRADICTION!

c tells you that #0 = #1
0*1* tells you that some 0s are followed by some 1s

their intersection tells you that a sequence of all 0s is followed by a sequence of all 1s

Example of Usage (3)

SAPIENZA
UNIVERSITÀ DI ROMA
DIPARTIMENTO DI INFORMATICA

Show that $E = \{0^i 1^j \mid i > j\}$ is not regular.

Let p be any natural number and $s = 0^{p+1} 1^p$.

Then s can be split into xyz, with $|y| > 0$ and y consisting only of 0s (since $|xy| \leq p$)

REMARK: any $i > 0$ would NOT yield a contradiction: indeed,

$$xy^iz = 0^k 1^p \in E, \text{ since } k > p+1 > p$$

However, $xy^0z = 0^k 1^p \notin E$, since $k \leq p$

the number of 1s is more than the number of 0s
we just need one number to prove a contradiction

One last Example (not trivial)

Prove that the language of all strings of 1s whose length is a perfect square is not regular:

$$D = \{1^{n^2} \mid n \geq 0\}.$$

remember to always check the 3 conditions

Fix any p and consider $s = 1^{p^2}$. taking a string belonging to D

Let $s = xyz$, with $|y| > 0$ and $|xy| \leq p$. Hence, also $|y| \leq p$.

Thus,

$$|xyyz| = |xyz| + |y| \begin{cases} > |xyz| = p^2 \\ = p^2 + |y| \leq p^2 + p < p^2 + 2p + 1 = (p+1)^2 \end{cases}$$

the len of xyz=len s= p^2, thus len xyyz > p^2

maggioriamo con (p+1)^2 perchè potrebbe esserci un quadrato perfetto in (p+2)^2

Hence, $p^2 < |xyyz| < (p+1)^2$ and so $xyyz \notin D$.

there is no perfect square between p^2 and (p+1)^2 !!!

esempio p=2, p^2=4, (p+2)^2=(2+2)^2=4^2=16, è un quadrato perfetto