Azzolini Riccardo 2020-11-17

Pumping lemma per i linguaggi regolari

1 Enunciato

Teorema (Pumping lemma[1]): Sia L un linguaggio regolare. Esiste allora una costante n (dipendente da L) tale che ogni stringa $w \in L$ con lunghezza $|w| \geq n$ possa essere scomposta in tre stringhe, $w = xyz$, in modo che:

1. $y \neq \epsilon$;
2. $|xy| \leq n$;
3. per ogni $k \geq 0$, anche xy^kz è una stringa di L ($xy^kz \in L$).

Di queste tre proprietà, la più importante è la terza: essa afferma sostanzialmente che a partire dalla stringa xyz si possono costruire un'infinità di stringhe del linguaggio L, sostituendo y con una potenza k-esima di y. La potenza di una stringa è definita come

$$y^k = \underbrace{y \dots y}_{k}$$

con il caso base $y^0 = \epsilon$.

Il nome "pumping lemma" deriva proprio dal fatto di poter "pompare", "iniettare" quante copie si vogliono di y, ottenendo ancora stringhe del linguaggio.

2 Dimostrazione

Per dimostrare questo risultato, bisogna innanzitutto individuare la costante n che sta alla base delle proprietà espresse dal pumping lemma. A tale scopo, si usa la definizione di linguaggio regolare: un linguaggio L è regolare se e solo se esiste un DFA $A = \langle Q, \Sigma, \delta, q_0, F\rangle$ che lo riconosca, cioè tale che $L(A) = L$. La costante n viene posta uguale al numero di stati dell'automa A: $n = |Q|$. Questo è solo un candidato di valore per la costante: adesso bisogna dimostrare che, con tale scelta, le proprietà asserite dal pumping lemma sono verificate.

[1] Questo teorema è chiamato "lemma" per motivi storici (è stato introdotto inizialmente come lemma finalizzato alla dimostrazione di un altro risultato).

Si consideri una stringa $w = a_1 \dots a_m \in L$, avente lunghezza $|w| = m \geq n$. Siccome $w \in L$ e $L = L(A)$, esiste una computazione dell'automa A che accetta w; sia essa

$$p_0 \xrightarrow{a_1} p_1 \cdots p_{m-1} \xrightarrow{a_m} p_m$$

con $p_0 = q_0$ e $p_m \in F$ (perché una computazione accettante parte dallo stato iniziale e arriva in uno stato finale). Formalmente, ogni generico stato p_h che compare in questa computazione è il risultato dell'applicazione della funzione di transizione estesa, a partire da $p_0 = q_0$, sulla stringa $a_1 \dots a_h$:

$$\forall h = 1, \dots, m \quad p_h = \hat{\delta}(q_0, a_1 \dots a_h)$$

Un'osservazione importante è che nella computazione compaiono $m + 1$ stati; siccome $m \geq n$, allora $m + 1 \geq n + 1 > n = |Q|$, quindi deve esserci almeno uno stato che compare due o più volte. In particolare, deve esserci almeno una ripetizione già solo nei primi $n + 1$ stati della computazione, $p_0, \dots, p_n$. Formalmente, esistono due indici i e j tali che $0 \leq i < j \leq n$ (quindi anche $i \neq j$) e $p_i = p_j$.

Usando gli indici i e j, si spezza la stringa $w = a_1 \dots a_m$ nelle tre stringhe x, y e z così fatte:

$$x = a_1 \dots a_i \qquad y = a_{i+1} \dots a_j \qquad z = a_{j+1} \dots a_m$$

In pratica, queste stringhe hanno le seguenti proprietà:

- (T1): x è la stringa che porta dallo stato $p_0 = q_0$ allo stato p_i,

  $$p_0 \xrightarrow{a_1} p_1 \cdots p_{i-1} \xrightarrow{a_i} p_i$$

  cioè $p_i = \hat{\delta}(q_0, x)$;

- (T2): y è la stringa che porta dallo stato p_i allo stato $p_j = p_i$,

  $$p_i \xrightarrow{a_{i+1}} p_{i+1} \cdots p_{j-1} \xrightarrow{a_j} p_j$$

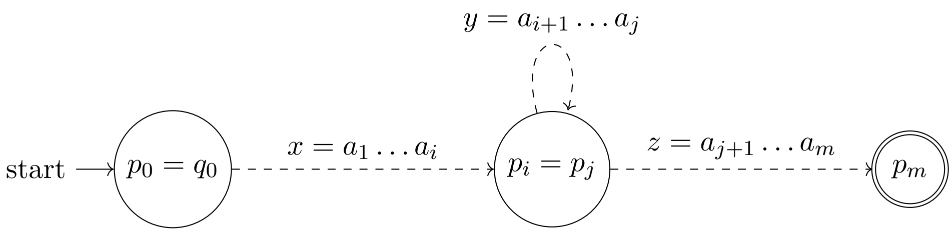
  cioè $p_j = \hat{\delta}(q_0, xy) = \hat{\delta}(p_i, y)$;

- (T3): z è la stringa che porta dallo stato $p_j = p_i$ allo stato finale p_m,

  $$p_j \xrightarrow{a_{j+1}} p_{j+1} \cdots p_{m-1} \xrightarrow{a_m} p_m$$

  cioè $p_m = \hat{\delta}(q_0, xyz) = \hat{\delta}(p_j, z)$.

La situazione può essere rappresentata graficamente:



Adesso bisogna verificare che le proprietà (1)–(3) del lemma effettivamente valgono:

1. $y \neq \epsilon$

   Siccome $i < j$, la stringa $y = a_{i+1} \dots a_j$ contiene almeno un simbolo, quindi $y \neq \epsilon$.

2. $|xy| \leq n$

   Da $j \leq n$ segue immediatamente che

   $$|xy| = |a_1 \dots a_i a_{i+1} \dots a_j| = j \leq n$$

3. Per ogni $k \geq 0$, $xy^k z \in L$.

   Prima di dimostrare la proprietà vera e propria, si dimostra per induzione su $k \geq 0$ che $\hat{\delta}(q_0, xy^k) = p_i$.

   - *Caso base*: $k = 0$.

     $$\begin{aligned} \hat{\delta}(q_0, xy^0) &= \hat{\delta}(q_0, x\epsilon) \\ &= \hat{\delta}(q_0, x) \\ &= p_i && \text{[per (T1)]} \end{aligned}$$

   - *Caso induttivo*: $k = h + 1$, con ipotesi induttiva $\hat{\delta}(q_0, xy^h) = p_i$.

     $$\begin{aligned} \hat{\delta}(q_0, xy^{h+1}) &= \hat{\delta}(q_0, xy^h y) \\ &= \hat{\delta}(\hat{\delta}(xy^h), y) && \text{[per definizione di } \hat{\delta}] \\ &= \hat{\delta}(p_i, y) && \text{[per ipotesi induttiva]} \\ &= p_j = p_i && \text{[per (T2)]} \end{aligned}$$

   Adesso si può effettivamente dimostrare la proprietà (3):

   $$\begin{aligned} \hat{\delta}(q_0, xy^k z) &= \hat{\delta}(\hat{\delta}(q_0, xy^k), z) && \text{[per deifnizione di } \hat{\delta}] \\ &= \hat{\delta}(p_i, z) && \text{[per il fatto appena dimostrato]} \\ &= \hat{\delta}(p_j, z) && \text{[per } p_i = p_j] \\ &= p_m && \text{[per (T3)]} \end{aligned}$$

   dato che $p_m \in F$, A accetta $xy^k z$, ovvero $xy^k z \in L$.

3 Osservazioni sulla costante n

In generale, dato un linguaggio regolare L, la costante n per cui vale il pumping lemma non è unica. Si chiama **costante di pumping** di L, indicata con N_L, la più piccola costante n per cui è verificato il pumping lemma.

Se il linguaggio L contiene almeno una stringa $w \in L$ di lunghezza $|w| \geq N_L$, allora L contiene infinite stringhe. Infatti, considerando la scomposizione $w = xyz$, siccome $y \neq \epsilon$ si ha che

$$xy^0z \neq xyz \neq xy^2z \neq xy^3z \neq \cdots$$

e per la proprietà (3) del pumping lemma tutte queste infinite stringhe diverse appartengono a L:

$$xy^0z, xyz, xy^2z, xy^3z, \dots xy^kz, \dots \in L$$

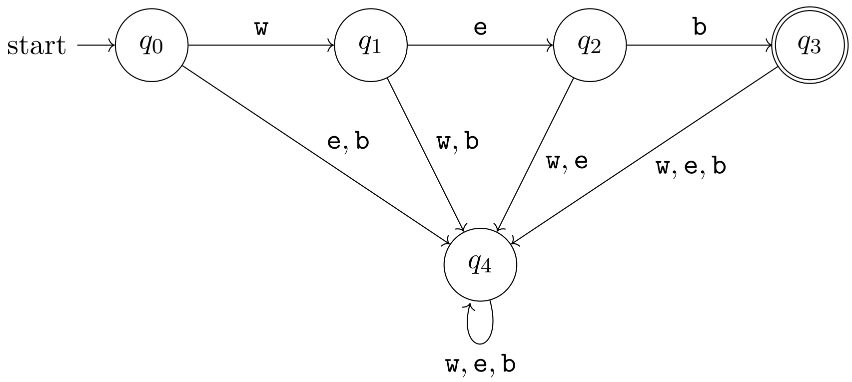
Viceversa, se L è un linguaggio finito, il lemma vale banalmente, semplicemente prendendo come costante di pumping n un numero che sia maggiore della lunghezza di tutte le stringhe in L:

$$N_L = \max\{|w| \mid w \in L\} + 1$$

Così, non esiste nessuna stringa $w \in L$ tale che $|w| \geq n$, ma queste sono le uniche stringhe considerate dall'asserto del pumping lemma, che quindi risulta vuotamente verificato (si ha una quantificazione universale su un dominio vuoto).

3.1 Esempio su un linguaggio finito

Si consideri, ad esempio, il caso del linguaggio $L = \{\texttt{web}\}$ sull'alfabeto $\Sigma = \{\texttt{w}, \texttt{e}, \texttt{b}\}$. L è riconosciuto dal seguente DFA:

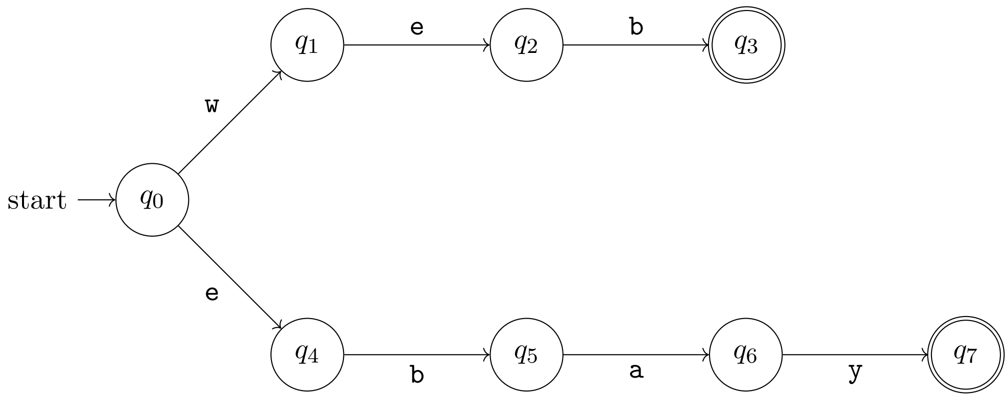


Si potrebbe dimostrare che non esiste alcun DFA che accetti L e abbia meno stati di questo.

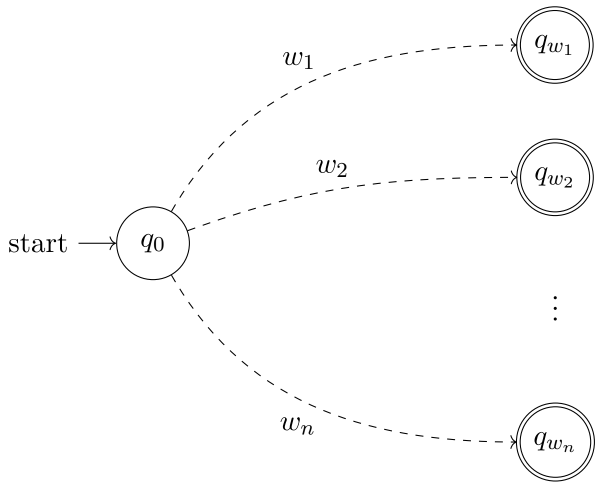
Ragionando come nella dimostrazione generale del pumping lemma, si sceglie la costante $n = |Q| = 5$. Allora, la proprietà espressa dal lemma deve valere per ogni $w \in L$ tale che $|w| \geq 5$, ma l'unica stringa in L è `web`, che ha lunghezza $|\texttt{web}| = 3 < 5$, cioè nessuna stringa in L ha lunghezza maggiore o uguale a 5, quindi il lemma è vuotamente verificato.

3.2 Ogni linguaggio finito è regolare

Il fatto che il pumping lemma valga per ogni linguaggio finito poteva essere previsto osservando che *ogni linguaggio finito è regolare*. Ad esempio, il linguaggio finito $L = \{\texttt{web}, \texttt{ebay}\}$ è intuitivamente riconosciuto dal seguente NFA:



Questa costruzione, simile a quella usata per la ricerca di parole chiave nei documenti, può facilmente essere estesa a qualunque linguaggio finito $L = \{w_1, w_2, \ldots, w_n\}$:



Allora, per ogni linguaggio finito esiste un NFA che lo riconosca, cioè ogni linguaggio finito è appunto regolare.

4 Applicazione

Il pumping lemma esprime una *condizione necessaria per i linguaggi regolari*: siccome esso vale per ogni linguaggio regolare, un linguaggio L per cui il lemma non vale non può essere regolare. Quindi, se si verifica che *non esiste* alcun $n \geq 0$ che possa fungere da costante di pumping per L, si è dimostrato che L *non* è un linguaggio regolare.

4.1 Esempio

Si vuole dimostrare che il linguaggio

$$L = \{0^n 1^n \mid n \geq 1\} = \{01, 0011, 000111, 00001111, \dots\}$$

non è regolare. Si suppone per assurdo che L sia invece regolare, e che N_L sia la sua costante di pumping.

Dato un qualunque $m \geq N_L$, si consideri la stringa

$$w = 0^m 1^m = \underbrace{0 \dots 0}_{m} \underbrace{1 \dots 1}_{m} \in L$$

che ha lunghezza $|w| > N_L$. Per il pumping lemma, dovrebbe esistere una scomposizione $w = xyz$ che soddisfi le condizioni (1)–(3), ma adesso si dimostrerà che, se sono soddisfatte la (1) e la (2), allora non può valere la (3), cioè non possono essere verificate insieme tutte e tre le condizioni.

La proprietà (2) afferma che $|xy| \leq N_L$; avendo scelto $m \geq N_L$, si deduce che la stringa xy deve occorrere entro i primi m simboli di w, che sono tutti 0: in sintesi, xy consiste solo di zeri. Per la proprietà (1), poi, deve essere $y \neq \epsilon$, ovvero y deve contenere almeno un simbolo, e, come appena detto, tale simbolo è sicuramente uno 0. Dunque, in generale, si ha $y = 0^h$ con $h \geq 1$, e complessivamente w ha la forma:

$$w = xyz = \overbrace{0 \dots 0}^{x} \overbrace{\underbrace{0 \dots 0}_{h}}^{y} \overbrace{0 \dots 0 \underbrace{1 \dots 1}_{m}}^{z}$$

Per mostrare che non vale la proprietà (3),

$$\forall k \geq 0 \quad xy^k z \in L$$

è sufficiente considerare un controesempio, come il caso in cui $k = 0$:

$$xy^0 z = xz = \overbrace{0 \dots 0}^{x} \overbrace{0 \dots 0 \underbrace{1 \dots 1}_{m}}^{z}$$

Se xyz conteneva, per definizione, m zeri, "cancellando" la stringa y di h zeri rimangono $m-h$ zeri, cioè $h \geq 1$ zeri in meno rispetto agli uno: siccome il numero di simboli 0 non è uguale al numero di 1, $xy^0z \notin L$. Ciò contraddice la proprietà (3) del pumping lemma, quindi si deduce che l'assunzione originale che il linguaggio fosse regolare è sbagliata.

La dimostrazione è conclusa, ma a scopo illustrativo può essere utile considerare un altro controesempio, il caso in cui $k = 2$:

$$xy^2z = xyyz = \overbrace{0\ldots0}^{x}\overbrace{0\ldots0}^{y}\overbrace{0\ldots0}^{y}\overbrace{0\ldots0\underbrace{1\ldots1}_{m}}^{z}$$

qui il numero di zeri è $m + h$, sicuramente superiore al numero di uno, perciò anche $xy^2z \notin L$.