Lecture #8: Nonregular Languages, Part One
Proof of the Pumping Lemma

Note: This supplemental document is for interest only (for students wishing to know how the Pumping Lemma can be proved): Students will not be expected to understand the proof of the Pumping Lemma in order to apply it, or to do well in this course.

Claim (Pumping Lemma)**.** *Let Σ be an alphabet and let $A \subseteq \Sigma^\star$.*

If A is a regular language, then there is a number $p \geq 1$ (called the **pumping length** *for A) — which only depends on A — such that if s is any string in A with length at least p, then s can be divided into three pieces $s = xyz$ (for $x, y, z \in \Sigma^\star$), satisfying the following three conditions.*

1. *$xy^iz \in A$ for every integer i such that $i \geq 0$.*
2. *$|y| > 0$ (so that $y \neq \lambda$).*
3. *$|xy| \leq p$.*

Proof. Let $A \subseteq \Sigma^\star$ be a regular language.

Then there exists a ***deterministic finite automaton***

$$M = (Q, \Sigma, \delta, q_0, F)$$

with language A.

Let $p = |Q|$ — the number of states in M — so that p is a positive integer that depends on A (but not on anything else that is introduced, after this, in this proof).

Either A does not include any strings $S \in \Sigma^\star$ with length at least p, or A includes at least one such string. These cases are considered next.

- *Case:* A does not include any strings $s \in \Sigma^\star$ with length at least p.

  In this case there is nothing more that we need to prove — because the claim only said something strings $s \in A$ such that $|s| \geq p$ (and no such strings exist).

- *Case:* A includes at least one string $s \in \Sigma^\star$ with length at least p.

  Let s be some string in $\Sigma^\star$ such that $s \in A$ and $|s| \geq p$. It is necessary (and sufficient) to show that it is possible to write s as xyz (for $x, y, z \in \Sigma^\star$) such that

1. $xy^iz \in A$ for every positive integer i.
2. $|y| > 0$ (so that $y \neq \lambda$).
3. $|xy| \leq p$.

Let $m = |s|$, so that $m \geq p$, and suppose that

$$s = \alpha_1\alpha_2 \dots \alpha_m$$

for $\alpha_1, \alpha_2, \dots, \alpha_m \in \Sigma$.

Let $r_0, r_1, r_2, \dots, r_m$ be the sequences of states visited as s is processed — so that $r_0 = q_0 = \delta^\star(q_0, \lambda)$, and

$$r_i = \delta^\star(q_0, \alpha_1\alpha_2 \dots \alpha_i)$$

for $1 \leq i \leq m$.

Consider the *first* $p + 1$ states in this sequence,

$$r_0, r_1, r_2, \dots r_p,$$

which are visited as the prefix $\alpha_1\alpha_2 \dots \alpha_p$ of s, with length p, is processed.

Since $|Q| = m = p$ and the above sequence of states has length $p + 1$, these states cannot all be distinct — so that *at least* one state $\widehat{q} \in Q$ must appear ***at least twice*** in the above sequence.

Now let $\widehat{q} \in Q$ be a state that *does* appear at least twice in the sequence $r_0, r_1, r_2, \dots, r_p$. Suppose i and j are integers such that $\widehat{q}$ first appears as r_i in this sequence and then appears for the second time in the sequence as r_j — so that $0 \leq i < j \leq p$.

- Let $x = \alpha_1\alpha_2 \dots \alpha_i \in \Sigma^\star$. Then x is the prefix of s with length i and

  $$\delta^\star(q_0, x) = \delta^\star(q_0, \alpha_1\alpha_2 \dots \alpha_i) = r_i = \widehat{q},$$

  since $\widehat{q}$ is the state that is reached after processing the first i symbols in s.
- Let $y = \alpha_{i+1}\alpha_{i+2} \dots \alpha_j$, the substring of s including the next $j - i$ symbols after the prefix x. Then, $\delta^\star(q_0, x) = \widehat{q}$ — as noted above — and since

  $$\delta^\star(q_0, xy) = \delta^\star(q_0, \alpha_1\alpha_2 \dots \alpha_j) = r_j = \widehat{q}$$

  as well,

  $$\delta^\star(\widehat{q}, y) = \delta^\star(r_i, \alpha_{i+1}\alpha_{i+2} \dots \alpha_j) = r_j = \widehat{q}$$

  as well: Processing the next $j - i$ symbols in s moves M from state $\widehat{q}$ back to itself.

– Finally, set $z = \alpha_{j+1}\alpha_{j+2}\dots\alpha_m$ — so that $x, y, z \in \Sigma^\star$ and $s = xyz$. Since $s \in A$,

$$\delta^\star(q_0, s) = \delta^\star(q_0, xyz) = q_F$$

for some *accepting* state $q_F \in F$. Now, since $\delta^\star(q_0, xy) = \widehat{q}$, as noted above,

$$\delta^\star(\widehat{q}, z) = \delta^\star(\widehat{q}, \alpha_{j+1}\alpha_{j+2}\dots\alpha_m) = q_F,$$

that is, processing the final $m-s$ symbols in s takes M from state $\widehat{q}$ to the accepting state q_F.

Once again consider the above properties 1, 2, and 3.

1. Since $\delta^\star(\widehat{q}, y) = \widehat{q}$, as noted above, it is easily proved by induction on i that

   $$\delta^\star(\widehat{q}, y^i) = \widehat{q}$$

   for every integer i, such that $i \geq 0$, as well. Consequently, if i is a non-negative integer then

   $$\begin{aligned} \delta^\star(q_0, xy^iz) &= \delta^\star(\widehat{q}, y^iz) && \text{(since } \delta^\star(q_0, x) = \widehat{q}\text{)} \\ &= \delta^\star(\widehat{q}z) && \text{(since } \delta^\star(\widehat{q}, y^i) = \widehat{q}\text{)} \\ &= q_F \in F && \text{(as noted above).} \end{aligned}$$

   Thus $xy^iz \in A$, since M accepts this string.
   Since i was an arbitrarily chosen non-negative integer it follows that $xy^iz \in A$ for *every* non-negative integer i. That is, property 1 is satisfied.

2. Since $y = \alpha_{i+1}\alpha_{i+2}\dots,\alpha_j$, $|y| = j - i > 0$ (and $y \neq \lambda$). That is, property 2 is also satisfied.

3. Finally, since $xy = \alpha_1\alpha_2\dots\alpha_i \cdot \alpha_{i+1}\alpha_{i+2}\dots\alpha_j = \alpha_1\alpha_2\dots\alpha_j$, $|xy| = j \leq p$: Property 3 is satisfied as well.

Since A was an arbitrarily chosen regular language, this establishes the Pumping Lemma.

□