Formal Languages and Parsing

CS 462

Jeffrey Shallit

LaTeXed by *Sibelius Peng*

Preface

Disclaimer Much of the information on this set of notes is transcribed directly/indirectly from the lectures of CS 462 during Winter 2022 as well as other related resources. I do not make any warranties about the completeness, reliability and accuracy of this set of notes. Use at your own risk.

For any questions, send me an email via `https://notes.sibeliusp.com/contact`.

You can find my notes for other courses on `https://notes.sibeliusp.com/`.

Sibelius Peng

Contents

1

CS 462 notation

- Natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$ and we use letters $i, j, k, \ell, m, n \in \mathbb{N}$.
- Finite string/word: a map from $[0, n-1]$ (an interval) to Σ (a finite alphabet of symbols)

  $w[i]$ is ith symbol of w
- infinite strings/words: a map from $\mathbb{N}$ to Σ. We denote infinite strings by bold-face:

$$\mathbf{w} = \mathbf{w}[0]\mathbf{w}[1]\mathbf{w}[2]\cdots$$

- Σ^* is the set of all finite words over Σ.
- Σ^ω is the set of all infinite words over Σ. Also written $\Sigma^{\mathbb{N}}$.
- $\Sigma^\infty = \Sigma^* \cup \Sigma^{\mathbb{N}}$.

Finite words typically denote by s, t, u, v, w, x, y, z

1.1 Some refreshers from CS 360/365

- x is a **prefix** of z if there exists y such that $z = xy$
- x is a **suffix** of z if there exists y such that $z = yx$
- x is a **subword** (factor) of z if there exists w, y such that $z = wxy$.
- x is a **subsequence** of z if x can be obtained from z by striking out zero or more symbols.

> Remark:
> Does substring mean contiguous (like subword)? or noncontiguous (like subsequence)? This definition depends the author of the book.
>
> Empty string ϵ is a first-class string like any other string and is not ruled out unless done so explicitly.

Then we have "proper" prefix, suffix, etc. If $z = xy$ and $x \neq z$, then x is a **proper prefix** of z.

1.2 Other notations

A shorthand for subword:

$$w[a..b] = w[a]w[a+1]\cdots w[b]$$

Concatenation of strings:

which is not commutative in general. Because we write concatenation in a multiplicative way, we can raise strings to powers: $x^n = \underbrace{xx\cdots x}_{n \text{ times}}$, or formally

$$\begin{aligned} x^0 &= \epsilon \\ x^n &= x \cdot x^{n-1} \qquad n \geq 1 \\ x^{m+n} &= x^m x^n \end{aligned}$$

A word is not of the form z^n, $n \geq 2$, $z \neq \epsilon$ is called **primitive**. The set of binary primitive words are denoted

$$P_2 = \{0, 1, 01, 10, 001, 010, 011, \ldots\}$$

One open question: is P_2 context-free? Probably not! But no one knows a proof.