Languages, Regular Expressions
Required reading: FC Sections 3.1, 3.2

1. Languages
   a. An alphabet is a finite, non-empty set: e.g., $\Sigma = \{0,1\}$.
   b. The elements of an alphabet are called symbols: e.g., 0,1.
   c. Finite sequence of elements are called strings or words: e.g., 101, 111.
   d. In a word, order matters: e.g., 10 $\neq$ 01.
   e. These are some of the most important operations on words:
      i. Length: Let $x = 101$, then $|x| = 3$.
      ii. Concatenation: Let $x = 101$ and $y = 10$, then $xy = 10110$.
      iii. Reverse: Let $x = 110$, then $x^R = 011$.
   f. The set of all words of length $N$ over an alphabet: $\Sigma^N$.
      i. E.g., Let $\Sigma = \{0,1\}$
         1. $\Sigma^1 = \{0,1\}$
         2. $\Sigma^2 = \{00,01,10,11\}$
         3. $\Sigma^3 = \{000,001,010,011,100,101,110,111\}$
         4. ...
         5. $\Sigma^0 = \{\epsilon\}$ (The empty word)
   g. The set of all possible words, of any length, over an alphabet: $\Sigma^*$ (also called the universe).
      i. Let $\Sigma = \{a,b\}$, then $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, aab, ...\}$.
   h. A language is a set of words over an alphabet.
      i. A language over alphabet $\Sigma$ is thus a subset of $\Sigma^*$.
         1. $L_1 = \{011, 1010, 111\}$
         2. $L_2 = \{0,10,110,1110,11110,\ldots\}$
         3. $L_3 = \{x \in \Sigma^* | n_0(x) = n_1(x)\}$, where $n_0(x)$ is the number of 0s in the word $x$, and $n_1(x)$ is the number of 1s in the word $x$.
         4. $L_4 = \{x | x$ represents a multiple of 5 in binary$\}$
   i. Languages can be finite or infinite.
   j. There are also operations over languages.
      i. Because languages are sets, we can talk about the union, intersection, and difference between languages.
      ii. We can also concatenate languages to produce a new language.
         1. Let $L=\{aa,bb\}$ and $M=\{\epsilon,00,11\}$, then language $LM = \{aa, aa00, aa11, bb, bb00, bb11\}$.
      iii. You can also concatenate a language to itself, any amount of times $L^N$ (same thinking that applies for the set of all words).

2. Regular Expressions
   a. Formal languages may contain random collections of words.
   b. In more interesting languages, the words share a common structure.
      i. $L_1 = \{x \in \{a,b\}^* | na(x) = nb(x)\}$
         1. $\{\epsilon, ab, ba, aabb, abab, abba, baba, bbab, ...\}$
         2. In English, “the set of all words with the same number of a’s and b’s.”
      ii. Let $\Sigma = \{a,b\}$ and $L_2 = \{x \in \Sigma^* | x$ starts and ends with $a\}$
         1. $\{a, aa, aaaa, aba, aaaa, aabaa, abaa, abaa, ...\}$
2. More concisely, we can write it as a regular expression: a(a|b)*a

C. Operators:
   i. Or: a|b (one or the other, not both simultaneously).
   ii. Concatenation: ab
   iii. Length: *
   iv. There’s an order of precedence. Highest is *, then concatenation, then the or.

D. Examples
   i. Let \( \Sigma = \{a, b\} \)
   ii. Consider the regular expression: \( r = a^*b^* \)
      1. \( \{\epsilon, a, b, ab, aa, bb, aab, abb, bbb, ...\} \)
      2. In English, “the set of all words that have 0 or any amount of a’s at the beginning followed by 0 or any amount of b’s at the end.”
   iii. Consider the regular expression: \( r = (a|aa|aaa)(bb)^* \)
      1. \( \{a, abb, abbbb, ..., aa, aabb, aabbbb, ..., aaa, aaabb, aaabbbb, ...\} \)
      2. In English, “the set of all words that have 1, 2, or 3 a’s at the beginning, and an even number of b’s at the end.”
   iv. Consider the regular expression: \( r = (a(a|b)^*)((a|b)^*a) \)
      1. In English, the set of all words with “an a followed by zero or more a’s or b’s or both” or “zero or more a’s or b’s or both followed by an a.”

E. A regular expression generates a language.
F. A language is regular if it can be expressed by a regular expression.
G. In other words, a regular language is one whose words’ structure can be described in a formal, mathematical way.
H. This means that the language can be mechanically described.

I. More examples:
   i. Give English-language descriptions of the languages generated by:
      1. \( (a|b)^* \)
         a.
      2. \( a^*|b^* \)
         a.
      3. \( b^*(ab^*ab^*)^* \)
         a.
      4. \( b^*(abb^*) \)
         a.
   ii. Give regular expressions over \( \Sigma = \{a, b\} \) that generate the following language:
      1. \( L_1 = \{x \mid x \text{ contains 3 consecutive a’s}\} \)
         a.
      2. \( L_2 = \{x \mid x \text{ has even length}\} \)
         a.
      3. \( L_3 = \{x \mid \text{nb}(x) = 2 \mod 3\} \)
         a.
      4. \( L_4 = \{x \mid x \text{ contains the substring aaba}\} \)
         a.
      5. \( L_5 = \{x \mid \text{nb}(x) < 2\} \)
         a.
      6. \( L_6 = \{x \mid x \text{ doesn’t end in aa}\} \)
         a.