Understanding Regular Language and Regular Expression

The following questions can help you understand the working of regular expressions.

Ex. 1: Find the shortest string that is not in the language represented by the regular expression $\mathbf{a^*(ab)^*b^*}$.

Solution: It can easily be seen that Λ, a, b, which are strings in the language with length 1 or less. Of the strings with length 2 aa, bb and ab are in the language. However, ba is not in it. Thus the answer is ba.

Ex. 2: For the two regular expressions given below,
(a) find a string corresponding to $\mathbf{r_2}$ but not to $\mathbf{r_1}$ and
(b) find a string corresponding to both $\mathbf{r_1}$ and $\mathbf{r_2}$.

$\mathbf{r_1 = a^* + b^*}$ $\quad \mathbf{r_2 = ab^* + ba^* + b^*a + (a^*b)^*}$

Solution: (a) Any string consisting of only a's or only b's and the empty string are in $\mathbf{r_1}$. So we need to find strings of $\mathbf{r_2}$ which contain at least one a and at least one b. For example ab and ba are such strings.
(b) A string corresponding to $\mathbf{r_1}$ consists of only a's or only b's or the empty string. The only strings corresponding to $\mathbf{r_2}$ which consist of only a's or b's are a, b and the strings consiting of only b's (from $\mathbf{(a^*b)^*}$).

Ex. 3: Let $\mathbf{r_1}$ and $\mathbf{r_2}$ be arbitrary regular expressions over some alphabet. Find a simple (the shortest and with the smallest nesting of * and +) regular expression which is equal to each of the following regular expressions.
(a) $\mathbf{(r_1 + r_2 + r_1r_2 + r_2r_1)^*}$
(b) $\mathbf{(r_1(r_1 + r_2)^*)^+}$

Solution: One general strategy to approach this type of question is to try to see whether or not they are equal to simple regular expressions that are familiar to us such as $\mathbf{a}$, $\mathbf{a^*}$, $\mathbf{a^+}$, $\mathbf{(a + b)^*}$, $\mathbf{(a + b)^+}$ etc.
(a) Since $\mathbf{(r_1 + r_2)^*}$ represents all strings consisting of strings of $\mathbf{r_1}$ and/or $\mathbf{r_2}$, $\mathbf{r_1r_2 + r_2r_1}$ in the given regular expression is redundant, that is, they do not produce any strings that are not represented by $\mathbf{(r_1 + r_2)^*}$. Thus $\mathbf{(r_1 + r_2 + r_1r_2 + r_2r_1)^*}$ is reduced to $\mathbf{(r_1 + r_2)^*}$.
(b) $\mathbf{(r_1(r_1 + r_2)^*)^+}$ means that all the strings represented by it must consist of one or more strings of $\mathbf{(r_1(r_1 + r_2)^*)}$. However, the strings of $\mathbf{(r_1(r_1 + r_2)^*)}$ start with a string of $\mathbf{r_1}$ followed by any number of strings taken arbitrarily from $\mathbf{r_1}$ and/or $\mathbf{r_2}$. Thus anything that comes after the first $\mathbf{r_1}$ in $\mathbf{(r_1(r_1 + r_2)^*)^+}$ is represented by $\mathbf{(r_1 + r_2)^*}$. Hence $\mathbf{(r_1(r_1 + r_2)^*)}$ also represents the strings of $\mathbf{(r_1(r_1 + r_2)^*)^+}$, and conversely $\mathbf{(r_1(r_1 + r_2)^*)^+}$ represents the strings represented by $\mathbf{(r_1(r_1 + r_2)^*)}$. Hence $\mathbf{(r_1(r_1 + r_2)^*)^+}$ is reduced to $\mathbf{(r_1(r_1 + r_2)^*)}$.

Ex. 4: Find a regular expression corresponding to the language of all strings over the alphabet { a, b } that contain exactly two a's.

Solution: A string in this language must have at least two a's. Since any string of b's can be placed in front of the first a, behind the second a and between the two a's, and since an arbitrary string of b's can be represented by the regular expression $\mathbf{b}^*$, $\mathbf{b}^*\mathrm{a}\,\mathbf{b}^*\mathrm{a}\,\mathbf{b}^*$ is a regular expression for this language.

Ex. 5: Find a regular expression corresponding to the language of all strings over the alphabet { a, b } that do not end with ab.
Solution: Any string in a language over { a , b } must end in a or b. Hence if a string does not end with ab then it ends with a or if it ends with b the last b must be preceded by a symbol b. Since it can have any string in front of the last a or bb, $\mathbf{(a + b)^*(a + bb)}$ is a regular expression for the language.

Ex. 6: Find a regular expression corresponding to the language of all strings over the alphabet { a, b } that contain no more than one occurrence of the string aa.
Solution: If there is one substring aa in a string of the language, then that aa can be followed by any number of b. If an a comes after that aa, then that a must be preceded by b because otherwise there are two occurrences of aa. Hence any string that follows aa is represented by $\mathbf{(b + ba)^*}$. On the other hand if an a precedes the aa, then it must be followed by b. Hence a string preceding the aa can be represented by $\mathbf{(b + ab)^*}$. Hence if a string of the language contains aa then it corresponds to the regular expression $\mathbf{(b + ab)^*aa(b + ba)^*}$.
If there is no aa but at least one a exists in a string of the language, then applying the same argument as for aa to a, $\mathbf{(b + ab)^*a(b + ba)^*}$ is obtained as a regular expression corresponding to such strings.
If there may not be any a in a string of the language, then applying the same argument as for aa to Λ, $\mathbf{(b + ab)^*(b + ba)^*}$ is obtained as a regular expression corresponding to such strings.
Altogether $\mathbf{(b + ab)^*(\Lambda + a + aa)(b + ba)^*}$ is a regular expression for the language.

Ex. 7: Find a regular expression corresponding to the language of strings of even lengths over the alphabet of { a, b }.
Solution: Since any string of even length can be expressed as the concatenation of strings of length 2 and since the strings of length 2 are aa, ab, ba, bb, a regular expression corresponding to the language is $\mathbf{(aa + ab + ba + bb)^*}$. Note that 0 is an even number. Hence the string Λ is in this language.

Ex. 8: Describe as simply as possible in English the language corresponding to the regular expression $\mathbf{a^*b(a^*ba^*b)^*a^*}$.
Solution: A string in the language can start and end with a or b, it has at least one b, and after the first b all the b's in the string appear in pairs. Any numbe of a's can appear any place in the string. Thus simply put, it is the set of strings over the alphabet { a, b } that contain an odd number of b's

Ex. 9: Describe as simply as possible in English the language corresponding to the regular expression $\mathbf{((a + b)^3)^*(\Lambda + a + b)}$.
Solution: $\mathbf{((a + b)^3)}$ represents the strings of length 3. Hence $\mathbf{((a + b)^3)^*}$ represents the strings of length a multiple of 3. Since $\mathbf{((a + b)^3)^*(a + b)}$ represents the strings of length 3n + 1, where n is a natural number, the given regular expression represents the strings of length 3n and 3n + 1, where n is a natural number.

Ex. 10: Describe as simply as possible in English the language corresponding to the regular expression **$(b + ab)^*(a + ab)^*$**.

Solution: $(b + ab)^*$ represents strings which do not contain any substring aa and which end in b, and **$(a + ab)^*$** represents strings which do not contain any substring bb. Hence altogether it represents any string consisting of a substring with no aa followed by one b followed by a substring with no bb.