Cairo University
Faculty of Engineering
Computer Engineering Dept.

Sheet 1: Basic Text Processing

1) Write regular expressions for the following:

   a. the set of all alphabetic strings.
   b. the set of all lower case alphabetic strings ending in *b*.
   c. the set of all strings from the alphabet *a,b* such that each *a* is immediately preceded by and immediately followed by *b*.
   d. the set of all binary strings with at least four ones.
   e. the set of all binary strings where the number of zeros is a multiple of 3.

2) Write regular expressions for the following languages. By "word", we mean an **alphabetic** string separated from other words by whitespace, any relevant punctuation, line breaks, and so forth.

   a. the set of all strings with two consecutive repeated words in the same case (e.g., "Humbert Humbert" and "the the" but not "the bug" or "the big bug").
   b. all strings that start at the beginning of the line with an integer and that end at the end of the line with a word.
   c. all strings that have both the word *grotto* and the word *raven* in them (but not, e.g., words like grottos that merely contain the word grotto).

3) Write a regular expression that matches responses to this question: "*What are blue, grey and red?*" The following 6 responses should be matched:

   colours
   colors
   they're colours
   they're colors
   they are colours
   they are colors

4) Write a python code for implementing the “Byte-pair Encoding” tokenization algorithm.

5) Mention a pair of words having:
   a. Same lemmas and same stems
   b. Same lemmas and different stems
   c. Different lemmas and same stems
   d. Different lemmas and different stems